AMB ALGORITHMS FOR MOLECULAR BIOLOGY

RESEARCH Open Access

Protein Structure Idealization: How accurately is it possible to model protein structures with dihedral angles?

Xuefeng Cui[1], Shuai Cheng Li[2], Dongbo Bu[3], Babak Alipanahi[1] and Ming Li[1*]

Abstract

Previous studies show that the same type of bond lengths and angles fit Gaussian distributions well with small standard deviations on high resolution protein structure data. The mean values of these Gaussian distributions have been widely used as ideal bond lengths and angles in bioinformatics. However, we are not aware of any research done to evaluate how accurately we can model protein structures with dihedral angles and ideal bond lengths and angles. Here, we introduce the protein structure idealization problem. We focus on the protein backbone structure idealization. We describe a fast $O(nm/\epsilon)$ dynamic programming algorithm to find an idealized protein backbone structure that is approximately optimal according to our scoring function. The scoring function evaluates not only the free energy, but also the similarity with the target structure. Thus, the idealized protein structures found by our algorithm are guaranteed to be protein-like and close to the target protein structure.

We have implemented our protein structure idealization algorithm and idealized the high resolution protein structures with low sequence identities of the CULLPDB_PC30_RES1.6_R0.25 data set. We demonstrate that idealized backbone structures always exist with small changes and significantly better free energy. We also applied our algorithm to refine protein pseudo-structures determined in NMR experiments.

Keywords: Protein structure idealization, Ideal bond length and angle, Dihedral angle space

Background

When studying the functions of a protein, it is crucial to know the three-dimensional structure consisting of the Cartesian coordinates of all the atoms of the protein. These atoms are bonded together by inter-atomic forces called chemical bonds. It has been observed that the bond lengths and angles of the same type assume a Gaussian distribution with a small standard deviation (STDEV) in high resolution protein structure data. Typically, the bond lengths on protein backbones have STDEVs between 0.019Å and 0.033Å and the bond angles on protein backbones have STDEVs between 1.5° and 2.7° [1,2]. These results suggest the possibility of modeling protein structures with the mean values of bond lengths and angles, which are often referred to as *ideal values*.

Ideal bond lengths and angles have been widely used in nuclear magnetic resonance (NMR) protein structure determination [3,4] and in protein structure prediction [5-9]. Moreover, stereochemical restraints are also used in X-ray protein structure determination [10,11]. In protein structure prediction, the main advantage of using ideal bond lengths and angles is a reduction in the search space for the target protein structure [12,13]. Specifically, if the target protein has n amino acids, the number of N, C_α and C atoms on the backbone is $3n$, and thus the Cartesian search space for the idealized backbone structure has a degree of freedom of $9n$ [12,13]. However, if all bond lengths and angles have ideal values, the protein backbone structure can be represented by a series of bond torsion angles in the feasible bond torsion angle space. In this case, the degree of freedom is reduced to approximately one tenth of that in the Cartesian space [12,13].

Although ideal bond lengths and angles have been widely used and accepted, we are not aware of any research done to evaluate how accurately it is possible

*Correspondence: mli@cs.uwaterloo.ca
[1] University of Waterloo, Ontario, Canada
Full list of author information is available at the end of the article

to model protein structures with dihedral angles. This motivates us to solve what we call the *protein structure idealization problem*: Given the coordinates of the target protein structure, find the coordinates of the optimal idealized protein structure. Here, an idealized protein structure is a protein structure with bond lengths and angles that are ideal with respect to a given scoring function; the function depends on the resultant structure's free energy, as well as its similarity with the target structure. Thus, the idealized protein structure is taken to be a protein-like structure that is close to the target protein structure.

First, we solve the protein structure idealization problem by idealizing the backbone structure and then idealizing the side-chain structure. This approach is widely accepted because previous research suggests that the backbone conformation is archived before the side-chain conformations are archived [14]. In our work, Ω dihedral angles are rounded to be either 0° or 180°. Some discussions on the properness of idealizing Ω dihedral angles can be found in [15,16].

We introduce a novel dynamic programming algorithm with a run-time complexity of $O(n/\epsilon^8)$, where ϵ is a small constant, to find the optimal idealized protein backbone structure according to our scoring function. In practice, we observed that it is unnecessary to remember the entire dynamic programming table. Thus, with a filtering technique, the run-time complexity is further reduced to $O(nm/\epsilon)$, where m is a constant integer.

In our initial study on the protein structure idealization problem, side-chain structures are determined using an exhaustive search which assumes that side-chain structures of different residues are independent from each other. The scoring function is similar to the one we used for backbone structure idealization. In practice, we observe that it is fast to regenerate idealized structures that are similar to a given idealized structure. We also refine the idealized backbone and side-chain structures according to our scoring functions iteratively.

We use our algorithm to evaluate how accurately it is possible to model protein structures with dihedral angles. We idealize all the X-ray protein structures from PDB [17] which satisfy the high resolution and the low sequence identity constraints downloaded on June 6, 2008 [18,19]. The results show that such idealized structures always exist and that they are very similar to the target structures in terms of the root mean square deviation (RMSD) of C_α or all atoms. Moreover, the idealized backbone structures tend to have dDFIRE free energy scores [20,21], which are significantly better than the target structures. The results support our conclusion that it is possible to model protein structures accurately with dihedral angles on all high resolution protein backbone structures.

One application of the protein structure idealization algorithm is to refine protein pseudo-structures either determined in experiments or predicted by computers. We have demonstrated one such case to improve poor (Φ, Ψ) dihedral angles of protein structures determined by NMR. The experiment result is also consistent with the previous experiment showing that the idealized structure has a small RMSD and better backbone free energy. We discuss several potential applications for our protein structure idealization algorithm in the conclusion.

Protein backbone structure idealization

Given the target protein backbone structure, we would like to find the optimal idealized backbone structure. For an idealized protein backbone structure, the coordinates of O, H and C_β backbone atoms can be calculated from the coordinates of N, C_α and C backbone atoms. Thus, we specifically describe how to generate coordinates of N, C_α and C atoms in this section. For simplicity, a structure is always referred to as a protein backbone structure unless strictly specified.

Idealized backbone structure generation

Given the target structure, we would like to generate idealized structures fulfilling two generation goals. First, the idealized structures should be similar to the target structure. Second, each pair of idealized structures should be at least some distance away to avoid redundant computation. Furthermore, we are interested in generating as many of these idealized structures as possible.

Before describing how we fulfill the generation goals, we describe a simple distance metric to measure the distance between two sets of coordinates representing the target protein. Let P_i be a set of coordinates representing the target protein, and $P_i^j \in P_i$ be the coordinate of the j-th atom of the target protein. Thus, there is $P_i = \{P_i^1, P_i^2, ..., P_i^{3n}\}$, where n is the number of amino acids of the target protein. For simplicity, let P_0 always represent the target structure, and P_i represent a generated idealized structure for $i > 0$. Let $D(P_i^k, P_j^k)$ be the Euclidean distance between P_i^k and P_j^k. We describe the distance between P_i and P_j as the bottleneck distance:

$$D(P_i, P_j) = \max_k D(P_i^k, P_j^k). \tag{1}$$

Using this distance metric, we fulfill both generation goals by satisfying the following generation constraints:

$$\begin{cases} D(P_0, P_i) \leq r \;\; \forall i > 0 \\ D(P_i, P_j) \geq \epsilon \;\; \forall i, j > 0 \end{cases}. \tag{2}$$

The first generation constraint assumes that the accuracy of the coordinates of the target structure is reasonably good, and no-worse than r. If this constraint is satisfied, the distance between the target coordinate and any generated coordinate representing the same atom is upper bounded by r. Thus, it is reasonable for any generated idealized structure P_i to be considered similar to target structure P_0. If the second generation constraint is satisfied, for each pair of generated idealized structures, there exists a pair of coordinates, one from each structure representing the same atom, such that they are at least ϵ distance away from each other. Therefore, both generation goals are achieved.

These generation constraints suggest limiting the search space inside a sphere with radius r, and discreting the search space with grids of size ϵ. When $\epsilon = 0.001$Å, the accuracy of X-ray crystallography [22] and PDB (protein database) format [23] is reached. Thus, this method is capable of generating all possible idealized structures at the accuracy of X-ray crystallography and PDB format.

Given the limited and discrete search space of each atom, one can generate idealized structure coordinates from the first atom to the last atom. For the first atom, an idealized coordinate lies within a sphere. Thus, the number of generated coordinates is bounded by $O(1/\epsilon^3)$. For each generated coordinate P_i^1 of the first atom, an idealized coordinate of the second atom lies on a ball surface with a constant distance to P_i^1. Thus, the number of generated coordinates is bounded by $O(1/\epsilon^2)$. For each generated coordinate pair (P_i^1, P_i^2) of the first two atoms, an idealized coordinate of the third atom lies on a circle with a constant distances to P_i^1 and P_i^2. Thus, the number of generated coordinates is bounded by $O(1/\epsilon)$. Similarly, the number of generated coordinates for any of the following atoms is also bounded by $O(1/\epsilon)$. Moreover, since we round Ω dihedral angles to either 0° or 180°, the coordinate of any C_α atom is unique and can be calculated from the coordinates of the previous three atoms.

Therefore, the total number of coordinates generated for all atoms is bounded by $O(1/\epsilon^{2n+4})$ by induction. Here, it is acceptable to assume that r is a constant because it is only related to the first atom. For subsequent atoms, we did not limit the search space to be inside the sphere with radius r as described above, and thus the actual number of generated coordinates should be much smaller in practice.

Idealized backbone structure scoring function

Given the generated idealized structures $\{P_i\}$, we need a scoring function $S_{BB}(P_i)$ to find the optimal idealized structure. The scoring function should evaluate not only the similarity between generated idealized structure P_i and target structure P_0, but should also evaluate the free energy of P_i, to ensure that P_i is protein-like. Thus, we define our scoring function as follows:

$$\begin{aligned} S_{BB}(P_i) =& S_f(P_i) - w_1 D_\alpha(P_i, P_0) - w_2 D_\beta(P_i, P_0) \\ & - w_3 D_H(P_i, P_0) - w_4 D_{\Phi,\Psi}(P_i, P_0), \end{aligned} \tag{3}$$

where w_a are the weighting parameters, $S_f(P_i)$ is the free energy score, $D_\alpha(P_i, P_0)$ is the root mean square divergence (RMSD) of C_α atoms, $D_\beta(P_i, P_0)$ is the RMSD of C_β atoms, $D_H(P_i, P_0)$ is the RMSD of the hydrogen and oxygen atoms participating in hydrogen bonds, and $D_{\Phi,\Psi}(P_i, P_0)$ is the RMSD of (Φ, Ψ) dihedral angles.

In our scoring function, the free energy is evaluated by a (Φ, Ψ) dihedral angle log-odd score as the free energy score $S_f(P_i)$. Specifically, we discrete the Ramachandran plot into grids of 360 by 360, and draw one plot for each type of amino acid. Then, we calculate the log-odd score $S_f(P_i^{1,t})$ of idealized structure $P_i^{1,t}$ of the first t atoms:

$$S_f(P_i^{1,t}) = \sum_{5 \le i \le t, A_i = C_\alpha} \log \frac{P_{AA_{i-3}}(\Phi_{i-3}, \Psi_{i-3})}{P_{null}(\Phi_{i-3}, \Psi_{i-3})}, \tag{4}$$

where one log-odd score is calculated at each C_α atom (by checking that atom type A_i is C_α) for the previous amino acid (represented by the previous C_α atom at $i-3$), $P_{AA_{i-3}}(\Phi_{i-3}, \Psi_{i-3})$ is the probability of the grid containing (Φ_{i-3}, Ψ_{i-3}) on the Ramachandran plot of amino acid type AA_{i-3}, and $P_{null}(\Phi_{i-3}, \Psi_{i-3})$ is the probability of the null model with a uniform distribution such that $P_{null}(\Phi_{i-3}, \Psi_{i-3}) = \frac{1}{360}\frac{1}{360}$.

Structure similarity is evaluated by other distance matrices in our scoring function. We use $D_\alpha(P_i, P_0)$ and $D_{\Phi,\Psi}(P_i, P_0)$ to serve as distance metrics to conserve the backbone structures, and $D_\beta(P_i, P_0)$ to serve as a distance metric to conserve the side-chain structure compatibilities; we also use $D_H(P_i, P_0)$ to serve as a distance metric to conserve the hydrogen bonds. Thus, some global dependencies are addressed implicitly by distance matrices $D_\beta(P_i^{1,t}, P_0^{1,t})$ and $D_H(P_i, P_0)$.

Dynamic programming algorithm

Theoretically, one can calculate scores for all generated idealized structures and find the optimal one with the maximum score. This method works well as long as similar structures always have similar scores. More formally, the method requires the assumption that $D(P_i, P_j) \le \epsilon \implies |S_{BB}(P_i) - S_{BB}(P_j)| \le \epsilon_s$, which is reasonable for small ϵ. Note that, since the total number of generated idealized structures is bounded by $O(1/\epsilon^{2n+4})$, this method is computationally expensive. Thus, we introduce a dynamic programming algorithm with a filtering technique to find the optimal idealized structure efficiently.

The dynamic programming algorithm has two assumptions. One assumption is that given two generated idealized structures $P_i^{1,t-1}$ and $P_j^{1,t-1}$ of the first $t-1$ atoms, such that $D(P_i^{t-k,t-1}, P_j^{t-k,t-1}) \le \epsilon$, for any generated coordinate P_i^t of the t'th atom, there always exists a generated coordinate P_j^t, such that $D(P_i^t, P_j^t) \le \epsilon$. The other assumption is that the scoring function satisfies the additive property, such that $S_{BB}(P_i^{1,t}) = S_{BB}(P_i^{1,t-k}) \oplus S_{BB}(P_i^{t-k+1,t})$, under some addition operator $\oplus$.

We observed that counter examples of the first assumption when $k \geq 5$ are rare, though counter examples do exist theoretically. The second assumption holds for our scoring function. Distance matrices $D_\alpha(P_i^{1,t}, P_0^{1,t})$, $D_\beta(P_i^{1,t}, P_0^{1,t})$, $D_H(P_i^{1,t}, P_0^{1,t})$ and $D_{\Phi,\Psi}(P_i^{1,t}, P_0^{1,t})$ satisfy the additive property because RMSD $D_{RMS}(P_i^{1,t}, P_0^{1,t})$ satisfies the additive property:

$$\begin{aligned} & D_{RMS}(P_i^{1,t}, P_0^{1,t}) \\ & = D_{RMS}(P_i^{1,t-k}, P_0^{1,t-k}) \oplus D_{RMS}(P_i^{t-k+1,t}, P_0^{t-k+1,t}) \\ & = \sqrt{\frac{D_{RMS}^2(P_i^{1,t-k}, P_0^{1,t-k})(t-k) + D_{RMS}^2(P_i^{t-k+1,t}, P_0^{t-k+1,t})k}{t}}. \end{aligned} \tag{5}$$

Moreover, the free energy score $S_f(P_i^{1,t})$ satisfies the additive property as follows:

$$\begin{aligned} S_f(P_i^{1,t}) &= S_f(P_i^{1,t-k}) \oplus S_f(P_i^{t-k+1,t}) \\ &= S_f(P_i^{1,t-k}) + S_f(P_i^{t-k+1,t}). \end{aligned} \tag{6}$$

The second assumption is fundamental to our dynamic programming algorithm. By induction, the first assumption implies that if $D(P_i^{t-k,t-1}, P_j^{t-k,t-1}) \leq \epsilon$, for any generated idealized structure $P_i^{t,n}$, there always exists a generated idealized structure $P_j^{t,n}$ such that $D(P_i^{t,n}, P_j^{t,n}) \leq \epsilon$. Recall that the scoring function assumes that $D(P_i^{t,n}, P_j^{t,n}) \leq \epsilon \implies |S_{BB}(P_i^{t,n}) - S_{BB}(P_j^{t,n})| \leq \epsilon_s$, and thus there is $S_{BB}(P_i^{t,n}) \approx S_{BB}(P_j^{t,n})$. If $S_{BB}(P_i^{1,t-1}) \geq S_{BB}(P_j^{1,t-1})$, there is approximately $S_{BB}(P_i) = S_{BB}(P_i^{1,t-1}) \oplus S_{BB}(P_i^{t,n}) \geq S_{BB}(P_j^{1,t-1}) \oplus S_{BB}(P_j^{t,n}) = S_{BB}(P_j)$. Therefore, if $D(P_i^{t-k,t-1}, P_j^{t-k,t-1}) \leq \epsilon$ and $S_{BB}(P_i^{1,t-1}) \geq S_{BB}(P_j^{1,t-1})$, there is no need to generate $P_j^{t,n}$ to find an approximately optimal solution.

Based on this observation, we developed a novel dynamic programming algorithm. Idealized structures are still generated as previously described, but the generation process is stopped for some idealized structures if we know it cannot lead us to the optimal one. First, the search space for each atom of the target protein is discretized to grids of size ϵ. When generating coordinates for atom t, if $P_i^{t-k+1,t}$ and $P_j^{t-k+1,t}$ are located in the same grid set $G_g^{t-k+1,t}$, we know that there is no need to continue the generation process on the lower scoring one of $P_i^{1,t}$ and $P_j^{1,t}$. Thus, we define the dynamic programming table $T_{BB}(t, G_g^{t-k+1,t})$ to be the optimal idealized structure for each observed tail grid set $G_g^{t-k+1,t}$ as follows:

$$\begin{cases} T_{BB}(t, G_g^{t-k+1,t}) = \max_{i,j} T_{BB}(t-1, G_i^{t-k,t-1}) \\ \quad \oplus S_{BB}(P_j^t) \\ T_{BB}(k, G_g^{1,k}) = \max_i S_{BB}(P_i^{1,k}) \end{cases}, \tag{7}$$

where $G_g^{t-k+1,t-1} = G_i^{t-k+1,t-1}$, $P_j^{t-k+1,t} \in G_g^{t-k+1,t}$ and $S_{BB}(P_j^{1,t-1}) \oplus S_{BB}(P_j^t) = S_{BB}(P_j^{1,t})$. Thus, the dynamic programming table can be calculated from the first atom to the last atom. Finally, the optimal idealized structure is the one with the highest score $\max_g G_g^{3n-k+1,3n}$.

The run-time complexity of our dynamic programming algorithm depends on the value of k. To keep all possible (Φ, Ψ) dihedral angles of the previous residue when generating C_α atoms, we have to choose $k \geq 5$. For speed, we choose $k = 5$ in our implementation. In this case, the number of score calculations required to calculate $T_{BB}(t, G_g^{t-4,t})$ is no more than the maximum number of coordinates sampled for six consecutive backbone atoms. Recall that there are exactly two C_α atoms in six consecutive backbone atoms, and the Ω dihedral angle is rounded. Thus, the coordinate of one C_α atom can be calculated from the coordinates of the other C_α atom and the two atoms between them. For this reason, the maximum number of sampled coordinates is bounded by $O(1/\epsilon^8)$. Moreover, the number of score calculations required to calculate $T_{BB}(k, G_g^{1,k})$ is no more than the maximum number of possible coordinates sampled for five consecutive backbone atoms, which is also $O(1/\epsilon^8)$. Therefore, the run-time complexity of our dynamic programming algorithm is $O(n/\epsilon^8)$.

To increase the speed for the dynamic programming algorithm, we applied an additional filtering technique to remember only the highly scored idealized structures. Specifically, the algorithm only remembers the optimal idealized structure for the top m scored tail configurations instead of all possible conformations. Thus, the run-time complexity is reduced to $O(nm/\epsilon)$. This approach works well in practice because an optimal idealized structure with a long poorly scored fragment is rare. Thus, we assumed that the local quality of the idealized structure should be reasonably high (in the top m score list).

Protein side-chain structure idealization

After the backbone structure of the target protein has been idealized, we begin to idealize the side-chain structures. When doing this, the idealized backbone structure is considered to be rigid. This approach is widely accepted because previous research suggests that the backbone conformation is archived before the side-chain conformations are archived [14]. After the side-chain idealization, we should have a complete idealized protein structure with all of the backbone and the side-chain structures idealized.

Protein side-chains suffer from low quality problems when determining protein structures. This is mainly because side-chains are not as stable as backbones, and they are more likely to have disorder problems than are backbones in crystals [22]. Thus, the target side-chain structure might be a poor reference for defining the search space and for evaluating the structure similarity score for generated idealized side-chain structures. To address this, we perform an exhaustive search on the entire feasible torsion angle space, instead of the limited torsion angle space, around the target side-chain structure.

Our side-chain idealization method assumes that the side-chain conformations of different residues are independent of each other. Otherwise, all residues with dependencies have to be generated together and the run-time complexity increases exponentially to the number of atoms involved. Moreover, the $N_{\eta 1} - C_{\zeta} - N_{\epsilon} - C_{\delta}$ and the $N_{\eta 2} - C_{\zeta} - N_{\epsilon} - C_{\delta}$ torsion angles of arginine residues are rounded to be either 0° or 180°. Then, the degree of freedom of the search space for each residue is at most four and it is now practical to perform an exhaustive search for each residue independently.

To find the optimal idealized side-chain structure, we design a new scoring function involving the similarity among the generated idealized side-chain structures and the target side-chain structures, and the free energy of the generated idealized side-chain structures. Let P_0 be the target side-chain structure of some residue, and P_i for all $i > 0$ be a generated idealized side-chain structure of the same residue. Then, the scoring function $S_{SC}(P_i)$ is defined:

$$S_{SC}(P_i) = S_f(P_i) - w_1 D_{H'}(P_i, P_0) - w_2 D_{\chi}(P_i, P_0),$$

where w_k are the weighting parameters, $S_f(P_i)$ is the free energy score, $D_{H'}(P_i, P_0)$ is the root mean square divergence (RMSD) of all non-hydrogen atoms, and $D_{\chi}(P_i, P_0)$ is the RMSD of χ torsion angles.

In our scoring function, the free energy score $S_f(P_i)$ is defined as a simple χ torsion angle log-odd score, which is similar to the free energy score of our backbone scoring function. Moreover, the log-odd score is based on the popular backbone dependent rotamer library downloaded from Dunbrack's lab [24]. Certainly, other local free energy scores can be adopted here. Similar to the backbone scoring function, $D_{H'}(P_i, P_0)$ and $D_{\chi}(P_i, P_0)$ serve as distance metrics to conserve the side-chain structure.

Result

To study the protein structure idealization problem and its applications, we implemente our protein structure idealization algorithm. In our implementation, we use the mean bond lengths and angles that had been reported in [2] as the ideal bond lengths and angles, respectively. When idealizing the protein backbone structure, we set the search space radius of an atom as $r = 1.6$Å and the discrete grid size as $\epsilon = r/5$. We find that $m = 50,000$ had a reasonable balance between speed and accuracy. When idealizing the protein side-chain structure, we set the search space of a rotamer dihedral angle to be within 3σ distance from the mean value, where σ is the STDEV of the rotamer dihedral angle, and we set the discrete grid size to be 10°. We also refine the idealized structure by iteratively reducing the search space and the discrete grid size by a constant factor of 0.5. Since finding the best scoring function for the protein structure idealization is out of the scope of this paper, we set all weights $w_a = 1.0$ for all a in our scoring function.

PDB protein structure idealization

In this experiment, we addressed how accurately it is possible to model protein structures with dihedral angles. We idealized high resolution protein structures with low sequence identities of the CULLPDB_PC30_RE

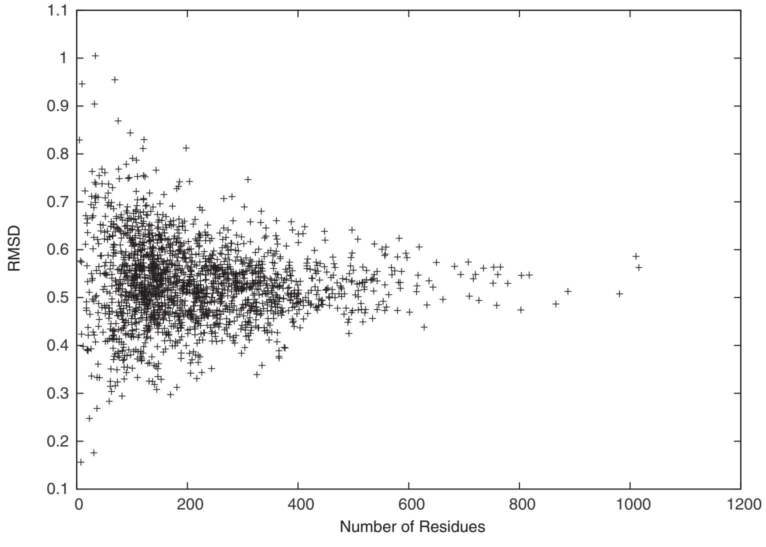


Figure 1 C_{α}-RMSD.

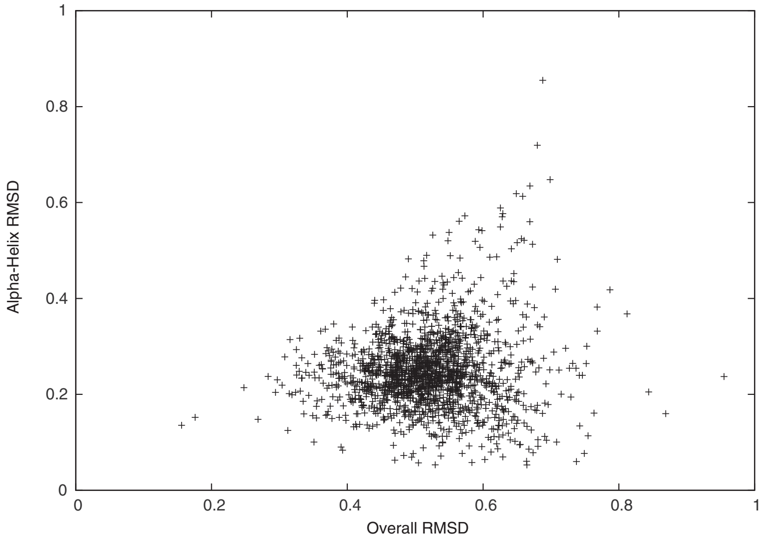


Figure 2 C_α-RMSD of all regions v.s. α-helix regions.

S1.6_R0.25 data set [18,19]. In fact, the CULLPDB_PC 30_RES1.6_R0.25 data set is the complete set of X-ray protein structures in PDB [17] with a sequence identity cutoff of 30%, a resolution cutoff of 1.6Å, and an *R* factor cutoff of 0.25. In summary, the data set contains 1898 proteins with an average length of 227 residues, as downloaded on June 6, 2008.

To show that the idealized and the target backbone structures are very similar, we calculated the C_α-RMSD as shown in Figure 1. The C_α-RMSD is a popular distance metric to evaluate the backbone distance between two protein backbone structures. The result shows that most distances between the idealized and the target backbone structures are small with mean 0.53Å and STDEV 0.08Å. Specifically, the smallest C_α-RMSD reaches 0.16Å, and 90% of the C_α-RMSDs are smaller than 0.63Å. Moreover, the C_α-RMSD is upper bounded by 1.00Å, although the search space radius for each atom is set to be 1.6Å. This result is consistent with the result of checking (Φ, Ψ) dihedral angles, where the average difference between the idealized and the target (Φ, Ψ) dihedral angles is as small as 0.08°. Therefore, it is possible to model protein backbone structures in CULLPDB_PC30_RES1.6_R0.25 accurately using only Φ and Ψ dihedral angles.

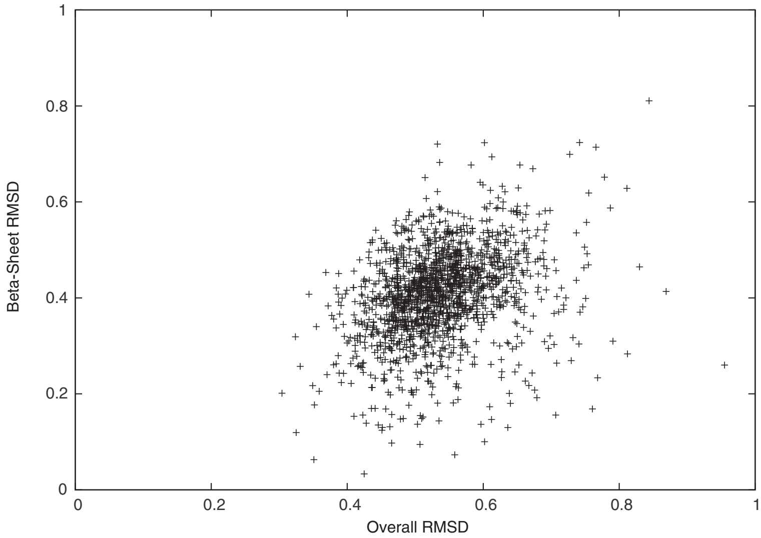


Figure 3 C_α-RMSD of all regions v.s. β-sheet regions.

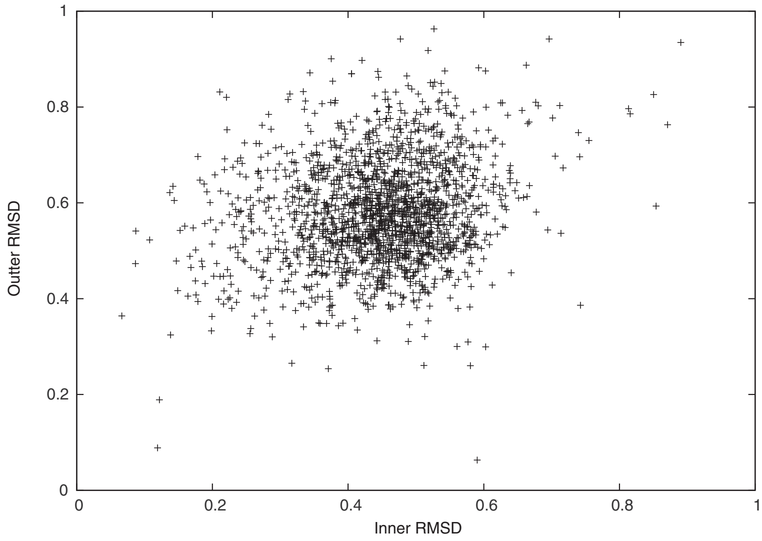


Figure 4 C_α-RMSD of inner regions v.s. outer regions.

We studied the C_α-RMSD further in different regions of the target protein structures. In Figures 2 and 3, we see that the C_α-RMSD of the α-helix and the β-sheet regions are smaller than that of the complete protein by 0.28Å and 0.12Å, respectively. Indeed, these regions are more restricted because of using $D_H(P_i, P_0)$ to conserve hydrogen bonds of α-helices and β-sheets in our scoring function. We also observe that the C_α-RMSD of residues that are closer to the geometric center of a target protein structure is 0.13Å smaller on average than the C_α-RMSD of the other residues that are farther, as shown in Figure 4. Thus, the inner residues tend to be closer to the idealization state than are the outer residues. We did not observe any significant differences on the C_α-RMSD between the buried and the exposed regions.

We also calculated the all-atom RMSD to show that the idealized and the target structures are very similar. In Figure 5, we see that most distances between the idealized and the target structures are small, with mean 0.79Å and STDEV 0.13Å. Moreover, the smallest all-atom RMSD reaches 0.45Å, and 90% of the all-atom RMSDs are smaller than 0.94Å. Note that both the C_α-RMSD and the all-atom RMSD between the idealized and the target structures tend to be stable when the target protein is long. Therefore, it is possible to model protein

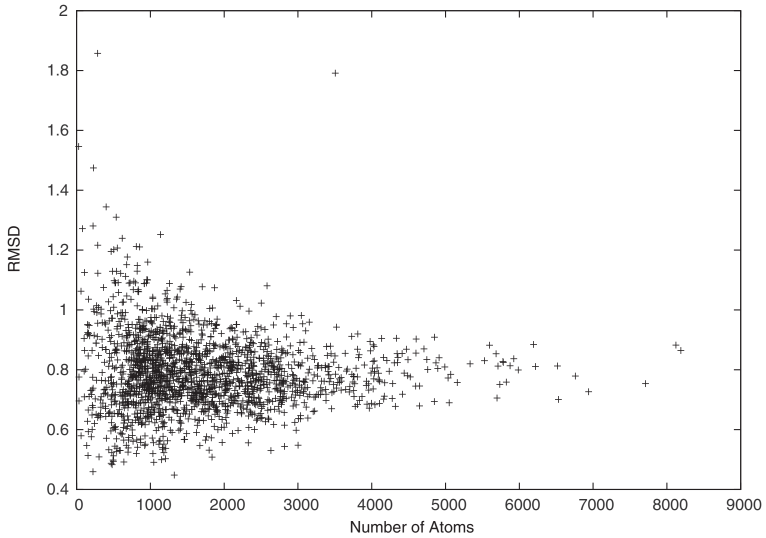


Figure 5 All-atom RMSD.

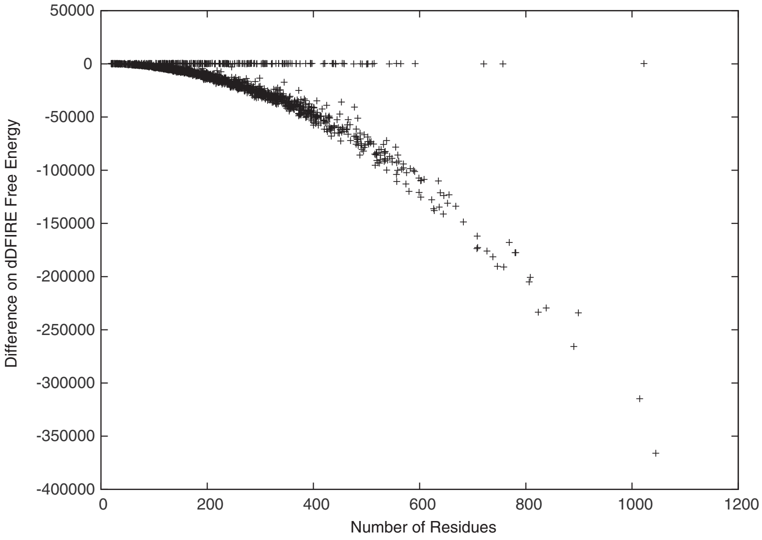


Figure 6 Protein backbone free energy (calculated by dDFIRE).

structures accurately with only Φ, Ψ, and χ dihedral angles.

The idealized backbone structures are also favored in terms of free energy. This is shown by checking the free energy differences between the idealized and the target protein backbone structures in Figure 6. Here, we calculate the free energy using dDFIRE [20,21], and observe that the dDFIRE free energy of most idealized backbone structures are significantly better than are those of the target backbone structures. For the rest without significant improvements, the difference is close to zero. This may be the result of some tight thereochemical restraints used in existing X-ray structure refinement programs [15,16]. It is also interesting that the observed free energy improvements are clearly not independent from the protein length. The figure suggests that the free energy difference has a square dependence on the protein length.

After idealizing the side-chain structures, the free energy is either improved by a relatively biger amount or worsened by a relatively smaller amount as shown in Figure 7. Unfortunately, in most cases, the free energy is worsened slightly but is still in a stable state with negative values. Again, here we used dDFIRE [20,21] to calculate the free energy. We observed that the dDFIRE free energy is improved for 90 or 4.74% of the idealized protein structures and is worsened slightly by 44 on average. Moreover,

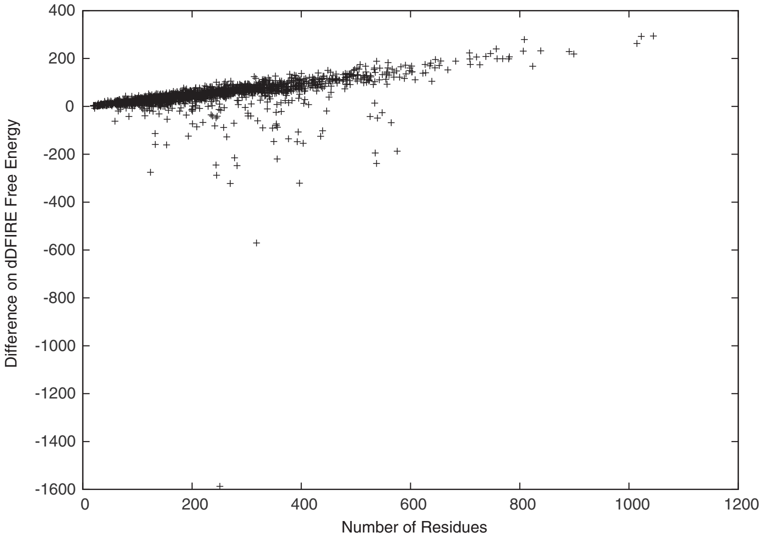


Figure 7 Protein all-atom free energy (calculated by dDFIRE).

Table 1 DSSP hydrogen bond differences before and after idealization

Type	Count difference	Percent difference
Parallel Bridge	9	0.04%
Antiparallel Bridge	-211	-0.37%
2_7 Helix	7080	26.46%
3_{10} Helix	-1018	-2.35%
α Helix	-1644	-1.48%
π Helix	-82	-1.27%
All	5183	1.85%

the dDFIRE free energy is improved by 1585 in the best case, and worsened by 293 in the worst case. The figure also suggests that the free energy difference has a linear dependence on the protein length.

Several side-chain prediction tools have been proven to predict accurate side-chain structures from native backbone structures [8,9,25,26]. However, these tools does not perform well when predicting side-chain structures from predicted backbone structures. To address this, we compared the predicted side-chain structures given the native backbone structures and those given the predicted backbone structures in terms of free energy. Here, we treat the idealized backbone structures of the CULLPDB_PC30_RES1.6_R0.25 data set as those which are best possibly predicted. Moreover, we used SCWRL4 [9] to predict side-chain structures and dDFIRE [20,21] to calculate free energies. The result shows that the free energy is worsened slightly by 43 if the predicted backbone structures are used. We do not think this difference is significant to side-chain prediction. Certainly, more experiments will show if this is conclusive.

Finally, we study the effects of idealization on hydrogen bonds. As shown in Table 1, we compare the number of hydrogen bonds detected by the DSSP program [27,28]. Here, only differences of the most popular types of hydrogen bonds are included. We note that the total number of hydrogen bonds is increased by 1.59% or 0.012 per residue after idealization. Specifically, the effects of idealization on hydrogen bonds of β bridges is minor, and the loss of the hydrogen bonds on α helices is reasonably controlled under 1.48%. Interestingly, the idealized backbone structures have significantly more 2_7 ribbons. The reason behind this observation remains open.

In summary, we demonstrate that using dihedral angles with ideal bond lengths and angles is capable of modeling protein structures that are highly similar to the ones in CULLPDB_PC30_RES1.6_R0.25 [18,19]. Since CULLPDB_PC30_RES1.6_R0.25 is the complete set of PDB protein structures satisfying the high resolution and the low sequence identity constraints, it is reasonable to extend the conclusion to all protein backbone structures. A positive side effect is that idealization improves backbone free energy, while most hydrogen bonds are conserved.

NMR protein structure refinement

In this experiment, we demonstrate an application of the protein structure idealization problem in NMR by idealizing 32 NMR protein structures. The NMR protein structures were randomly chosen from PDB [17]

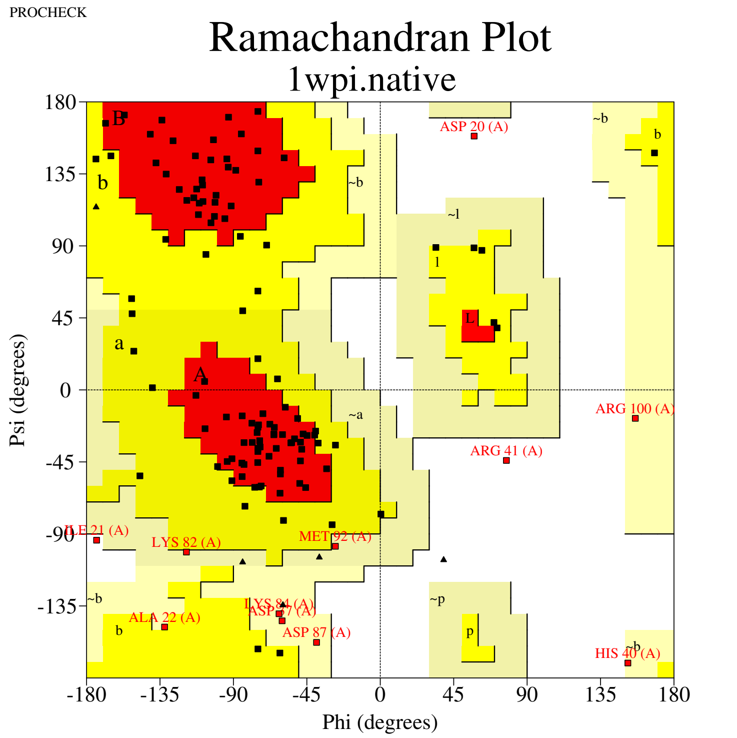
Table 2 The percentages of the favored (Φ, Ψ) dihedral angles of *32* NMR protein structures before and after idealization

PDB	Native	Ideal	Diff	PDB	Native	Ideal	Diff
1SSK	44.6%	71.9%	27.3%	2LBN	59.7%	77.6%	17.9%
2KQP	62.9%	80.0%	17.1%	1WPI	64.4%	81.4%	17.0%
1EXE	60.5%	76.7%	16.2%	2LNV	58.6%	72.4%	13.8%
1X6F	64.1%	73.1%	9.0%	2L6B	72.2%	81.1%	8.9%
2GFU	72.3%	80.4%	8.1%	1PC2	79.3%	87.4%	8.1%
2LMR	79.7%	87.0%	7.3%	2KA0	72.6%	78.3%	5.7%
2L3O	71.3%	76.9%	5.6%	1O1W	67.2%	72.1%	4.9%
2CQ9	78.3%	82.6%	4.3%	2RQA	72.0%	75.4%	3.4%
2D86	89.0%	92.1%	3.1%	1NTC	80.5%	83.1%	2.6%
2JZT	76.6%	79.0%	2.4%	2CZN	76.5%	76.5%	0.0%
1RCH	75.4%	74.6%	-0.8%	2JU1	77.1%	75.9%	-1.2%
2KV7	85.5%	84.2%	-1.3%	2JT2	83.6%	81.5%	-2.1%
2KYW	83.8%	81.1%	-2.7%	2OSR	82.7%	80.0%	-2.7%
2L6M	81.7%	78.5%	-3.2%	2CU1	81.1%	77.8%	-3.3%
1AJ3	93.3%	88.8%	-4.5%	1WI5	84.0%	78.0%	-6.0%
1NMW	85.0%	78.0%	-7.0%	2LBV	83.9%	74.8%	-9.1%

with a sequence identity cutoff of 30% and a gapless fragment length cutoff of 80 residues. In cases of multiple chains or models of some NMR protein structures, only the first chain from the first model is used in this experiment. This addition to the conclusion of the previous experiment shows that poor (Φ, Ψ) dihedral angles of the NMR protein structures are improved by idealizing them.

To demonstrate this, we compared the percentage of favored (Φ, Ψ) dihedral angles calculated by PROCHECK [29] in Table 2. After idealization, we see that 19 out of 32 NMR protein structures have more favored (Φ, Ψ) dihedral angles. Overall, the percentage is increased by 4.34% on average and 27.30% in the best case, which is closer to the minimum percentage of 90% expected in a good quality model [29].

Note that for those NMR protein structures that already have more than approximately 75% of favored (Φ, Ψ) dihedral angles, idealization harms the percentage by -0.85% on average. There are at least two reasons for this. First, our free energy score $S_f(P_i)$ is calculated from a data set that is different from the one used by PROCHECK. In fact, we used 1898 protein structures of the CULLPDB_PC30_RES1.6_R0.25 data set [18,19], while PROCHECK used 118 protein structures, with a resolution cutoff of 2.0Å and an R factor cutoff of 0.20 [29]. Although the percentages of favored (Φ, Ψ) dihedral angles are decreased in Table 2, our free energy scores of proteins 1WI5, 1NMW, and 2LBV are increased by 0.22, 1.35, and 0.31, respectively, after idealization. Second, our implementation is trying to optimize our scoring function $S_{BB}(P_i)$, instead of optimizing only the free



Plot statistics

Residues in most favoured regions [A,B,L]	76	64.4%
Residues in additional allowed regions [a,b,l,p]	31	26.3%
Residues in generously allowed regions [~a,~b,~l,~p]	8	6.8%
Residues in disallowed regions	3	2.5%
	----	------
Number of non-glycine and non-proline residues	118	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	5	
Number of proline residues	8	

Total number of residues	133	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms
and R-factor no greater than 20%, a good quality model would be expected
to have over 90% in the most favoured regions.

Figure 8 **Ramachandran plot of the native NMR protein structure 1WPI.**

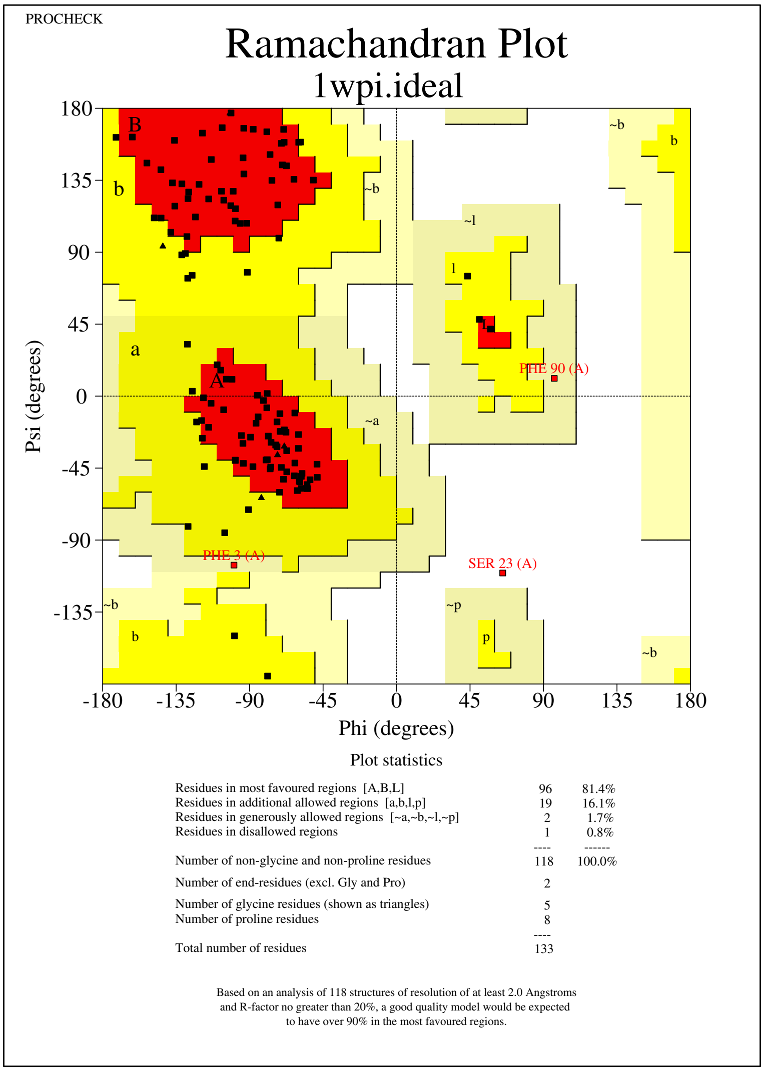
energy score. Thus, it is possible to see decreased free energy scores after idealization, especially when the target protein structure has a high percentage of favored (Φ, Ψ) dihedral angles.

Our conclusion is further supported by the case study of the NMR structure with PDB ID 1WPI [30]. From the Ramachandran plots drawn by PROCHECK [29] in Figures 8 and 9, we find that (Φ, Ψ) dihedral angles tend to move towards favored regions. Specifically, the native structure contains only 64.4% of (Φ, Ψ) dihedral angles in favored regions, while the idealized structure contains a significantly improved percentage of 81.4% of (Φ, Ψ) dihedral angles in favored regions. Moreover, the native structure contains three (Φ, Ψ) dihedral angles that are not in any feasible areas of the Ramachandran plot. However, there is only one such case found in the idealized structure. Thus, two infeasible (Φ, Ψ) dihedral angles are fixed by the (Φ, Ψ) dihedral angle log-odd score. Here, we did not, but certainly can, implement a hard constraint to disallow any infeasible (Φ, Ψ) dihedral angles.

In summary, we have demonstrated that protein structure idealization can be used to improve poor (Φ, Ψ) dihedral angles of protein pseudo-structures. These protein pseudo-structures can either be predicted or be experimentally determined. More applications of the protein structure idealization problem will be studied.

Conclusion

We have introduced the protein structure idealization problem and performed our first attempt to solve it. The experiment results show that idealized structures always exist with small changes on the coordinates. Furthermore, the idealized backbone structures have significantly



Plot statistics

Residues in most favoured regions [A,B,L]	96	81.4%
Residues in additional allowed regions [a,b,l,p]	19	16.1%
Residues in generously allowed regions [~a,~b,~l,~p]	2	1.7%
Residues in disallowed regions	1	0.8%
	----	------
Number of non-glycine and non-proline residues	118	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	5	
Number of proline residues	8	

Total number of residues	133	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms
and R-factor no greater than 20%, a good quality model would be expected
to have over 90% in the most favoured regions.

Figure 9 Ramachandran plot of the idealized NMR protein structure 1WPI.

better free energy and (Φ, Ψ) dihedral angle distributions. Therefore, protein structures can be modeled accurately with dihedral angles and ideal bond lengths and angles, and it is feasible to predict protein backbone and side-chain structures by searching the dihedral angle space.

Our protein structure idealization algorithm may be improved in several ways. Since our scoring functions are very simple with all weights $w_a = 1.0$ in the current implementation, there is space for improvements. We are also looking forward to adding protein-ligand interaction energy to our scoring function and to studying the effect of idealization on protein-ligand interactions. Moreover, since some atoms are more flexible than others, we can also set different search spaces for different atoms in our algorithm. For example, when idealizing X-ray protein structures, the search space of each atom could be selected according to its B-factor. We can also adopt a divide-and-conquer algorithm in our algorithm to find the global, rather than local, optimal idealized structure. Specifically, we can divide the protein structure into small fragments, idealize each fragment separately, and merge idealized fragments. The key is to divide the protein structure by a tree decomposition of the interaction graph and to remember the optimal idealized fragment for each possible configuration of atoms with interactions to external atoms. Similar ideas have already been used successfully to improve the speed and the accuracy of backbone and side-chain structure predictions [8,9,25,26,31,32].

Our protein structure idealization algorithm can also correct modelling errors of protein structures in PDB [17]. In fact, previous research indicates that many bond conformations and side-chain rotamers are likely incorrect in PDB, and it is useful to have an automated mechanism to fix these problems [33,34]. Thus, we can address these problems by idealizing all protein structures in PDB with our protein structure idealization algorithm and using our specially tuned scoring functions.

The idealized version of the PDB [17] provides new protein structure references to study protein structures and functions. For example, we can rebuild fragment and rotamer libraries based on the idealized PDB. It would then be more intuitive to use the idealized fragment or rotamer libraries in the protein backbone or side-chain structure prediction algorithms searching the dihedral angle space. Thus, we expect to see some improvements of the accuracy of these algorithms with the idealized fragment and rotamer libraries. Therefore, we also provide a new approach for discovering unusual atoms and bonds by comparing the idealized and the original PDB structures. Although most of these unusual atoms and bonds are due to errors, we expect to discover some biochemical insights that assist in understanding protein functions.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
The algorithm implementation and all experiments were done by XC; the protein structure idealization algorithm was designed by XC and SCL; the CULLPDB_PC30_RES1.6_R0.25 experiment was designed by XC, SCL and DB; the NMR experiment was designed by XC and BA; the project was directed by ML; All authors read and approved the final manuscript.

Acknowledgements
This work was supported by a Startup Grant at City University of Hong Kong [7002731], National Basic Research Program of China [2012CB316500], an NSERC Grant [OGP0046506], Canada Research Chair program, an NSERC Collaborative Grant, OCRiT, Premier's Discovery Award, Killam Prize and SHARCNET.

Author details
[1]University of Waterloo, Ontario, Canada. [2]City University of Hong Kong, Hong Kong, China. [3]Chinese Academy of Sciences, Beijing, China.

Received: 20 December 20121 Accepted: 5 February 2013
Published: 25 February 2013

References
1. Engh RA, Huber R: **Accurate bond and angle parameters for X-ray protein structure refinement.** *Acta Crystallogr Sect A* 1991, **47:**392–400.
2. Engh RA, Huber R: **Structure quality and target parameters.** *Vol F Int Tables Crystallogr* 2006:382–392. chap. 18.3.
3. Güntert P, Wüthrich K: **Improved efficiency of protein structure calculations from NMR data using the program DIANA with redundant dihedral angle constraints.** *J Biomol NMR* 1991, **1**(4):447–456.
4. Güntert P, Mumenthaler C, Wüthrich K: **Torsion angle dynamics for NMR structure calculation with the new program DYANA.** *J of Mol Biol* 1997, **273:**283–98. [http://www.ncbi.nlm.nih.gov/pubmed/9367762]
5. 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.
6. Simons KT, Strauss C, Baker D: **Prospects for ab initio protein structural genomics.** *J Mol Biol* 2001, **306:**1191–1199.
7. Li SC, Bu D, Xu J, Li M: **Fragment-HMM: a new approach to protein structure prediction.** *Protein Sci* 2008, **17**(11):1925–1934.
8. Canutescu AA, Shelenkov AA, Dunbrack RL: **A graph-theory algorithm for rapid protein side-chain prediction.** *Protein Sci* 2003, **12**(9):2001–2014. [http://dx.doi.org/10.1110/ps.03154503]
9. Krivov GG, Shapovalov MV, Dunbrack RL: **Improved prediction of protein side-chain conformations with SCWRL4.** *Proteins: Struct Funct Bioinformatics* 2009, **77**(4):778–795. [http://dx.doi.org/10.1002/prot.22488]
10. Kuszewski J, Gronenborn AM, Clore GM: **Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases.** *Protein Sci* 1996, **5**(6):1067–1080.
11. Kuszewski J, Gronenborn AM, MClore G: **Improvements and extensions in the conformational database potential for the refinement of NMR and X-ray structures of proteins and nucleic acids.** *J Magn Reson* 1997, **125:**171–177.
12. Rice LM, Brünger AT: **Torsion angle dynamics: Reduced variable conformational sampling enhances crystallographic structure refinement.** *Proteins: Struct Funct Genet* 1994, **19**(4):277–290. [http://dx.doi.org/10.1002/prot.340190403]
13. Stein EG, Rice LM, Brünger AT: **Torsion-angle molecular dynamics as a new efficient tool for NMR structure calculation.** *J Magn Reson* 1997, **124:**154–164.
14. Dunbrack RL, Cohen FE: **Bayesian statistical analysis of protein side-chain rotamer preferences.** *Protein Sci* 1997, **6**(8):1661–1681. [http://www.proteinscience.org/cgi/content/abstract/6/8/1661]
15. Evans PR: **An introduction to stereochemical restraints.** *Acta Crystallogr D* 2007, **63:**58–61.

16. Jaskolski M, Gilski M, Dauter Z, Wlodawer A: **Stereochemical restraints revisited: how accurate are refinement targets and how much should protein structures be allowed to deviate from them?.** *Acta Crystallogr D* 2007, **63**(5):611–620.
17. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28:**235–242.
18. Wang G, Dunbrack RL: **PISCES: a protein sequence culling server.** *Bioinformatics* 2003, **19**(12):1589–1591.
19. Wang G, Dunbrack RL: **PISCES: recent improvements to a PDB sequence culling server.** *Nucleic Acids Res* 2005, **33**(Web-Server-Issue):94–98.
20. Yang Y, Zhou Y: **Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions.** *Protein Sci* 2008, **17**(7):1212–1219.
21. Yang Y, Zhou Y: **Specific interactions for ab initio folding of protein terminal regions with secondary structures.** *Proteins* 2008, **72**(2):793–803.
22. Lattman E, Loll PJ, Loll P: *Protein Crystallography: A Concise Guide*: Johns Hopkins University Press; 2008.
23. Rutgers UCSD: **Protein data bank contents guide.** 2011. http://www.wwpdb.org/documentation/format33/v3.3.html.
24. Shapovalov MV, Dunbrack RL: **A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions.** *Structure London England 1993* 2011, **19**(6):844—858. [http://www.ncbi.nlm.nih.gov/pubmed/21645855]
25. Xu J: *Rapid protein side-chain packing via tree decomposition*: Springer; 2005.
26. Xu J, Jiao F, Berger B: **A tree-decomposition approach to protein structure prediction.** *Fourth Int IEEE Comput Soc Comput Syst Bioinformatics Conf* 2005:247–256. [http://view.ncbi.nlm.nih.gov/pubmed/16447982]
27. Kabsch W, Sander C: **Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features.** *Biopolymers* 1983, **22**(12):2577–2637.
28. Joosten RP, te Beek TAH, Krieger E, Hekkelman ML, Hooft RWW, Schneider R, Sander C, Vriend G: **A series of PDB related databases for everyday needs.** *Nucleic Acids Res* 2011, **39**(suppl 1):D411–D419.
29. Laskowski RA, MacArthur MW, Moss DS, Thornton JM: **PROCHECK: a program to check the stereochemical quality of protein structures.** *J of Appl Crystallogr* 1993, **26**(2):283–291. [http://dx.doi.org/10.1107/S0021889892009944]
30. Jung JW, Yee A, Wu B, Arrowsmith CH, Lee W: **Solution structure of YKR049C, a putative redox protein from Saccharomyces cerevisiae.** *J Biochem Mol Biol* 2005, **38**(5):500–504.
31. Xu Y, Xu D: **Protein threading using PROSPECT: Design and evaluation.** *Proteins: Struct Funct Genet* 2000, **40**(3):343–354. [http://dx.doi.org/10.1002/1097-0134(20000815)40:3]
32. Kim D, Xu D, Guo JT, Ellrott K, Xu Y: **PROSPECT II: protein structure prediction program for genome-scale applications.** *Protein Eng* 2003, **16**(9):641–650. [http://view.ncbi.nlm.nih.gov/pubmed/14560049]
33. Hooft RWW, Vriend G, Sander C, Abola EE: **Errors in protein structures.** *NATURE* 1996, **381**(6580):272.
34. Joosten RP, Joosten K, Cohen SX, Vriend G, Perrakis A: **Automatic rebuilding and optimization of crystallographic structures in the Protein Data Bank.** *Bioinformatics* 2011, **27**(24):3392–3398. [http://dx.doi.org/10.1093/bioinformatics/btr590]

doi:10.1186/1748-7188-8-5
Cite this article as: Cui *et al.*: Protein Structure Idealization: How accurately is it possible to model protein structures with dihedral angles?. *Algorithms for Molecular Biology* 2013 **8**:5.