Grow to Fit Molecular Dynamics (G2FMD): an ab initio method for protein side-chain assignment and refinement

Wei Zhang$^{1,2}$ and Yong Duan$^{2,3}$

$^1$Department of Chemistry and Biochemistry, University of Delaware, Newark, DE 19716 and $^2$UC Davis Genome Center and Department of Applied Science, University of California, Davis, CA 95616, USA

Abstract

The rough energy landscapes and tight packing of protein interiors are two of the critical factors that have prevented the wide application of physics-based models in protein side-chain assignment and protein structure prediction in general. Complementing the rotamer-based methods, we propose an ab initio method that utilizes molecular mechanics simulations for protein side-chain assignment and refinement. By reducing the side-chain size, a smooth energy landscape was obtained owing to the increased distances between the side chains. The side chains then gradually grow back during molecular dynamics simulations while adjusting to their surrounding driven by the interaction energies. The method overcomes the barriers due to tight packing that limit conformational sampling of physics-based models. A key feature of this approach is that the resulting structures are free from steric collisions allowing the application of all-atom models in the subsequent refinement. Tests on a small set of proteins showed nearly 100% accuracy on both $\chi_1$ and $\chi_2$ of buried residues and 94% of them were within 20° from the native conformation, 79% were within 10° and 42% were within 5°. However, the accuracy decreased when exposed side chains were involved. Further improvement and application of the method and the possible reasons that affect the accuracy on the exposed side chains are discussed. Keywords: G2FMD/Grow to Fit Molecular Dynamics/protein structure/side-chain assignment

Introduction

Despite the promise and potential of accuracy, all-atom physics-based models have so far played only minor roles in protein structure prediction (Moult et al., 2003; Moult, 2005). The most common applications have been energy minimization to remove steric clashes in the final steps of structure refinement and scoring to rank the perspective structures. One of the challenges arises from the prerequisite of correct assignment of protein side chains for successful applications of molecular mechanics force fields, in particular an assignment that is free from steric collisions. The rough free energy landscape renders the overall molecular energy sensitive to the detailed side-chain packing. The tightly packed protein interiors typically seen in well-folded proteins further exacerbate the problem; even slight mis-assignment of the side chains can easily lead to collisions and cause high energy and instability of the structures that complicate the efforts at subsequent refinement. Therefore, a method that can generate collision-free side chains can be beneficial.

A large number of excellent and powerful conformational sampling methods have been proposed and have been used in protein structure prediction and side-chain assignment and refinement. These methods can be classified into three main categories. In the first category, enhanced sampling ability comes mainly from increased temperature and the entropic force drives systems to cross energy barriers. Examples include simulated annealing and replica exchange molecular dynamics methods. The second category is to examine the conformational space systematically (either exhaustively or preferentially). These methods have been widely used in the rotamer-based side-chain assignment. The third category is to remove the (high) energy barriers. An example is the ‘functional smoothing’ (Andricioaei and Straub, 1998) approach in which the original energy function is transformed into a smoother function with fewer local minima. The method has been successfully applied to the optimization of atomic clusters (Andricioaei and Straub, 1998), tetrapeptide (Andricioaei and Straub, 1998) and crystals of rigid molecules (Wawak et al., 1998). For complex systems such as proteins, small conformation deviations may cause serious property changes and the design of effective smooth functions is often a challenging endeavor. Other similar methods include the convex global underestimator algorithm in which the energy hyper-surface is first approximated by a convex (quadratic) function (Maranas and Floudas, 1994; Foreman et al., 1999). The method proposed in this paper falls into the third general category that seeks to remove the high energy barriers due to side-chain packing.

Correct assignment of side chains is an important step in protein structure prediction because the ultimate goal is to provide a biochemical framework for the physical interpretation of protein functions. Starting from main-chain conformations, side chains are assigned to allow further refinement and close scrutiny of the structures. However, because of the rough energy landscape and combinatorial nature, optimal assignment of protein side chains remains a challenging task. A typical approach is to use knowledge-based potential to assign the side chains from a library of low-energy conformations called rotamers (Ponder and Richards, 1987; Dunbrack and Karplus, 1993, 1994). Encouraging progress has been made in this area (Vasquez, 1996; Bower et al., 1997; DeMaeyer et al., 1997; Kuhlman and Baker, 2000; Xiang and Honig, 2001; Liang and Grishin, 2002; Canutescu et al., 2003; Tramontano and Morea, 2003). Large rotamer libraries, complex scoring functions and advanced search algorithms have helped to improve the sampling accuracy and efficiency (Tuffery et al., 1991; Desmet et al., 1992; Koehl and Delarue, 1994; Hwang and Liao, 1995; DeMaeyer et al., 1997; Desjarlais and Clarke, 1998; Mendes et al., 1999a,b; Looger
and Hellinga, 2001; Xiang and Honig, 2001; Desmet et al., 2002; Fernandez-Recio et al., 2002; Liang and Grishin, 2002; Liu et al., 2003; Tsai et al., 2003; Peterson et al., 2004).

Soft energy functional forms (Levitt, 1992; Koehl and Delarue, 1994; Huang et al., 1998; Fernandez-Recio et al., 2002; Tsai et al., 2003; Peterson et al., 2004) and flexible side chains (Mendes et al., 1999a) have also been contemplated in many side-chain prediction methods to reduce the effect of the exquisite sensitivity to steric clashes and to allow placement of rotamers into the approximate positions in an environment that would otherwise be inhospitable (Huang et al., 1998). Baker and co-workers used a two-step refinement protocol in their Rosetta algorithm (Bradley et al., 2003) to increase the efficiency of conformation sampling around the native structures. Similarly, Fang and Shortle recently used a statistical potential for local side-chain and backbone interactions to improve the speed and accuracy of finding near-native conformations (Fang and Shortle, 2005a,b).

In addition to the statistical potentials, all-atom force field has been utilized in side-chain assignment (Wilson et al., 1993; Petrella et al., 1998), typically in energy minimization and scoring. However, a particularly acute problem confronting \textit{ab initio} methods in protein structure prediction and side-chain assignment is the mispacking. When mispacking occurs, in most cases, it is impossible to rearrange the mispacked side chains without unfolding backbones because they will come into collision with neighboring side chains very easily. Hence proteins often need to unfold from the mispacked state to continue to search the conformational space and to find the correctly packed state. However, unfolding–refolding further complicates the conformational search. Therefore, the application of all-atom models has been limited to only local refinement. Furthermore, energy minimization usually cannot correct mispacking. Rather, it often moves the structures away from the initial position because of the side-chain collision and the resulting high energy due to collision.

In this paper, we present an alternative method that is complementary to the rotamer-based methods by utilizing an \textit{ab initio} all-atom molecular mechanics force field to assign and to refine protein side chains. Our Grow to Fit Molecular Dynamics (G2FMD) side-chain assignment and refinement method overcomes the steric collision problem by decreasing the size of side-chain groups first and letting the side chains adjust to the local energy landscape, when they slowly grow back. Meanwhile, the reduction in side chains reduces the energy barriers and accelerates the conformational search process. The method generates a smoother energy landscape which greatly improves the conformational sampling efficiency. Initial tests show that the method can find the correct side-chain positions with high accuracy. Results of application to protein side-chain assignments are presented. The potential applications of this method are also discussed.

**Methods**

**Grow to Fit Molecular Dynamics (G2FMD)**

The side-chain groups are first reduced by reassigning the coordinates of side-chain atoms and scaling down side-chain bonds. The basic AMBER energy function has the following form:

\[
U(R) = \sum_{\text{bonds}} K_r (r - r_{eq})^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} V_\phi [1 + \cos(\phi - r)] + \sum_{\text{atoms}} \left( \frac{A_i}{R_{ij}^l} - \frac{B_i}{R_{ij}^s} \right) + \sum_{\text{atoms}} q_i q_j \frac{1}{\epsilon R_{ij}}
\]

The angle and dihedral terms were kept unchanged during rescaling (discussed later). The bond, van der Waals and electrostatic terms of atom \(i\) were scaled according to the scaling parameter \(\lambda_i\) (0 \(\leq\) \(\lambda_i\) \(\leq\) 1). For atoms that were not subjected to scaling, the scaling parameter \(\lambda\) was a constant (\(\lambda_i = 1\)); for those that were subjected to scaling \(\lambda_i = \lambda\) which was varied during the simulation (0 \(\leq\) \(\lambda\) \(\leq\) 1).

In the bond terms, the bond force constant, \(K_r\), was kept unchanged. For the bond between atoms \(i\) and \(j\), \(r_{eq}\) was varied by scaling to \(r_{eq}^\prime = \lambda_i r_{eq}\), while \(\theta_{eq}\) was \(\lambda_i \theta_{eq}\). The bond length \(r_{eq}\) was the unscaled (standard) equilibrium bond length in the AMBER force field. Thus, \(\lambda_{ij} = \lambda\), \(1/2(1 + \lambda)\), 1, if the scaling was applied to, respectively, both bonded atoms, only one of the two or none of the two atoms. For example, bonds between \(C\alpha\) and \(C\beta\) were scaled as \(1/2(1 + \lambda)\) since only \(C\beta\) was subjected to scaling and \(C\alpha\) is part of the main-chain atom and was not subjected to scaling.

The charges were scaled as \(q_i^\prime = \sqrt{\lambda_i} q_i\), while \(q_i^0\) was the unscaled AMBER charge of atom \(i\). Thus, the electrostatic interaction between two scaled atoms was scaled by \(\lambda\). With this scaling scheme, the electrostatic interaction energy between the scaled atoms within the same side-chain group were kept unchanged during the rescaling because both distance and charges were scaled at the same rate.

The van der Waals parameters \(A_{ij}\) and \(B_{ij}\) were scaled by adjusting the van der Waals radii of atoms \(i\) and \(j\). This was accomplished by scaling \(\sigma_{ij} = \lambda_{ij} \sigma_{ij}^0\) in the Lennard–Jones potential \(\phi_{ij}^r = 4\epsilon_{ij} \left( \frac{\sigma_{ij}^0}{r_{ij}} \right)^6 - \left( \frac{\sigma_{ij}^0}{r_{ij}} \right)^12\), where \(\sigma_{ij}^0\) is the van der Waals parameter without scaling. Again, \(\lambda_{ij} = 1/2(\lambda_i + \lambda_j)\) was the combined scaling factor. In this way, the rescaled van der Waals potentials were also kept unchanged for atoms of the same side-chain groups because their radius and distance were scaled with the same rate. However, van der Waals interactions with other parts of the protein were scaled (reduced) accordingly. An implication of these scaling schemes of van der Waals and charges is that the 1–4 electrostatic and 1–4 van der Waals terms are also kept constant within the same side-chain groups. This further implies that the dihedral terms can be kept unscaled because of the coupling between 1–4 terms and dihedral energy, which, in combination, serve to achieve a balanced energy profile in the dihedral space.

Series trial simulations were launched to test the scaling parameters. Results showed that the sampled conformational space decreases noticeably when the time of the grow cycle was less than 1 ps. Because backbone atoms were not rescaled, the shrunk side chains tended to collapse with backbones when
the scaling parameter \( \lambda \) became smaller than 0.4. Balancing among simulation time, sampling efficiency and the possibility of collision, we chose to use 10 and 1 ps for the growth and shrink steps (described in next paragraph), respectively, and \( \lambda \) was set to 0.6 \( \leq \lambda \leq 1 \). The simulations were stopped arbitrarily after 200 growth and shrink cycles (2.2 ns), which took about 1 day to complete on a dual-CPU 2.4 GHz Intel Xeon PC.

After rescaling the side-chain groups, molecular dynamics simulation was launched with two iterative steps: shrink and growth. First, the side chains were reduced by scaling the coordinates. In this way, collisions due to incorrect assignment (e.g. random assignment) were removed. Then, the scaled groups gradually grew back to their normal size in 10 ps while the relevant parameters were updated every 10 steps. Energy minimization was performed at the end of each MD cycle to remove the energy fluctuation inherent in MD simulations. The energy was compared against that saved from the previous cycle and a conformation was selected by the Monte Carlo procedure (at 300 K). The selected conformation was saved and became the starting structure in the next cycle which typically started by gradually reducing the side chains to 60% of the normal size in 1.0 ps.

The backbones were restrained by a harmonic force (5.0 kcal/mol Å\(^2\)) in all simulations. In the single side-chain assignment, the target side chain was first assigned randomly. During the G2FMD simulation, the shrink-and-grow cycle was applied only to the selected side-chain group, while all other atoms were restrained to the native conformation by harmonic forces (5.0 kcal/mol Å\(^2\)). In the multiple side-chain assignment, the target side chain was first assigned randomly and the shrink-and-grow cycle was simultaneously applied to all the selected side-chain groups, while other atoms were restrained. Similarly, for the assignment of the entire protein side chains, all the side-chain groups were initially assigned randomly and all the side-chain groups underwent shrink-and-grow cycles simultaneously while only the backbone was restrained.

Assignment tests were performed separately on the exposed and buried side chains. A side chain was considered buried if the solvent-accessible surface area, calculated using NACCESS (Hubbard and Thornton, 1993), was less than 20% of its total surface area. Otherwise, it was considered exposed.

**Simulation protocols**

The proteins and peptides were represented using the Duan et al. force field (Duan et al., 2003) (also known as the AMBER ff03 force field). A modified version of the AMBER 8 simulation package was used in all simulations. All bonds were constrained by the SHAKE algorithm (Ryckaert et al., 1977) to allow an integration time step of 1.0 fs. A Generalized Born (GB) implicit solvent model (Onufriev et al., 2000, 2002) was applied to model the solvation effect, with an effective 0.2 M salt concentration. Surface area terms were not explicitly modeled in the GB model as their contribution is typically small. The interior and solvent dielectric constants were set to 1.0 and 78.5, respectively. Simulations were done at constant temperature ensemble and the temperature was maintained at 300 K by a Berendsen thermostat (Berendsen et al., 1984).

**Results and discussion**

In this section, we will first examine the impact of the scaled side chains by calculating the energy profiles of individual side chains modeled as dipeptides. A potential of mean force (PMF) of a selected side chain was calculated in the context of a protein. The accuracy of side-chain assignment and refinement was tested on a number of proteins. Here, we first present the detailed results on the human cell adhesion protein (PDB code 1fna). We will first test its ability to assign single side chains. This is considered relatively simple as it is designed to give best-case results since it avoids the combinatorial problems. Assignment tests of multiple buried and exposed side chains are presented next. Finally, test assignments for the entire protein and for other proteins are presented.

**Energy profiles with scaled side chains**

Interactions between peptide backbone and side chains are some of the most important determinants of both side-chain and main-chain conformations. In molecular mechanics, the interactions are modeled by a combination of electrostatic, van der Waals, torsion potentials and, to a less extent, bond and bond angle interactions. The last two are hard degrees of freedom that play minor roles to determine side-chain conformations at room temperature. These parameters might be correlated to some extent and were developed for the full-size side chains.

Care was taken when the method was developed such that all components were reduced in a consistent manner to minimize the difference. To verify this, we compared the energy profiles of the scaled dipeptides (Ace-X-Nhe) with the normal energy profiles for 17 amino acids. Amino acids Ala and Gly were excluded from the tests because they both have trivial side chains. Pro was kept at normal size in this study because the ring structure is relatively rigid and its small side chain typically poses no significant problems in side chain assignment.

Figure 1A and B illustrate the normal and reduced leucine dipeptides (Ace-Leu-Nhe), respectively. The leucine side chain is represented by sticks and balls and the reduced side chain is 60% of its normal size. Figure 1C shows the energy profiles of the dipeptides, calculated by energy minimization while restraining the \( \chi_1 \) torsion angle. Overall, the energy profiles are highly similar and the main features of the full-size side chain energy profile were well preserved after the side chain was scaled. The overall root-mean-square difference between the energy profiles was less than 1.0 kcal/mol. The similarity between the energy surfaces indicates that the change in side-chain size did not dramatically alter the energy profiles.

The energy profiles of other 16 dipeptides (Ace-X-Nhe), except Ala, Gly and Pro, were also calculated by restrained energy minimization for the normal and reduced (\( \lambda = 0.6 \)) side chains and are shown in Figure 2. The high degree of similarity is evident. Although the energy minima have been well maintained, for some amino acids (e.g. Cys, Lys and Trp), the relative ordering of minima was changed by a small amount. For Val, although the relative ordering of the three energy minima was preserved, the energy gaps were increased. In the case of Trp, the energy minimum at 60° became close to the minimum at 300°. These differences will gradually disappear when the side chains grow back to the normal size and,
therefore, should not pose significant difficulty to the method. Overall, the energy profiles of all amino acids were well preserved after the reduction of side-chain sizes.

**Free energy profile with scaled side chain**

A key motivation of the work was to develop a method that can effectively reduce the free energy barriers while still retaining the basic features (e.g. minima) of the free energy surface. Because of the scaled side chains, interactions with the neighboring amino acids are expected to change. This can also change the energy and free energy profiles. Intuitively, one may expect that this can reduce the energy barriers for buried side chains. Yet, ideally, scaling the side chains should only reduce the energy barriers but should retain the positions of the lowest free energy minimum.

The potential of mean force (PMF) was calculated using umbrella sampling and the weighted histogram analysis method (WHAM) (Kumar et al., 1992, 1995) to examine the side-chain scaling. In these calculations, restrained simulations were conducted at a set of side-chain dihedral angles. The PMF was obtained by post-processing using a program prepared by Alan Grossfield (http://dasher.wustl.edu/alan). For qualitative comparison, two sets of PMF results of human cell adhesion protein (PDB id 1fna) were generated, one for the normal protein and the other with all side chains scaled by $\lambda = 0.6$. With the backbone fixed, the side-chain $\chi_1$ angle of Ser50 was varied from $5^\circ$ to $360^\circ$ with a $5^\circ$ increment. A 100 ps restrained MD simulation was performed at each $\chi_1$ angle. For each set of PMF, 72 trajectories and a total of 7.2 ns simulation were collected for analysis.

The results are illustrated in Figure 3, which shows the calculated PMF for a partially exposed Ser50 side chain of protein 1fna. Unlike the dipeptide energy profiles that mainly reflect the interaction between side chain and main chain, the PMF profile is influenced also by the local environment. Since the side-chain shrinking increased the distances and changed the interactions with other neighboring side-chain groups, differences in the PMF profiles after shrinking are expected. However, one may wish to retain the free energy minima while reducing the barrier. This was indeed the case. The two free energy minima and the lowest free energy minimum were well retained and the barrier separating these two minima was reduced by about 2.0 kcal/mol, equivalent to a >20-fold enhancement in terms of barrier-crossing ability at room temperature (300 K). Thus, the side chain conformational sampling has been significantly enhanced because of the smoother free energy landscape. In the reduced-size state, the side chain exhibits an additional energy minimum near $200^\circ$.

Since Ser50 is a partially exposed side chain in 1fna with 63.5% of its surface exposed, greater reduction in the free energy barriers is expected for the buried and long side chains.

**Single buried side chains**

Single residue assignment and refinement are considered the simplest case since it avoids the combinatorial problem. The results represent the best possible assignment achievable by the method. We tested the method on 17 buried residues of protein 1fna and the results are summarized in Tables I and II. For the 17 buried residues, the final assignment was 100% correct. Here a ‘correct’ assignment is such that the torsion
angle is less than 40° away from the native conformation which is a typical criterion used in most other side-chain assignment studies (Dunbrack and Karplus, 1993, 1994; Bower et al., 1997; Canutescu et al., 2003). In spite of the poorly assigned initial (random) conformations, the refinement was very effective and efficient. More than half of the residues reached the native conformations after the first cycle (10 ps). Twelve of the 17 residues (70%) were correctly assigned within 10 cycles (110 ps). The slowest one was Ile15 which took 52 cycles (572 ps). In all cases, the native conformations were correctly maintained after they were selected from the Monte Carlo procedure. In addition to the high efficiency, the accuracy was also encouraging. All of the assigned conformations were within 20° of the native conformations; more than 90% were within 10° and ~80% were within 5°.

Figure 4 illustrates the detailed sampling and Monte Carlo procedure of a representative side chain, Ile54. Starting from a random conformation, the side chain reached its native conformation within 40 ps (four refinement cycles), as judged by the \( \chi_1 \) and \( \chi_2 \) torsion angles. Interestingly, although it frequently sampled other conformations later (Figure 4A), the native conformation (Figure 4B) was successfully selected during the Monte Carlo process. The results suggested that the underlying Duan et al. (2003) AMBER ff03 force field is reasonably accurate when it is applied to the buried residues. Furthermore, the side-chain torsion parameters were derived mainly by fitting to a set of representative organic compounds (Wang and Kollman, 2001) based on Cornell et al.’s charge set (Cornell et al., 1995) whereas the Duan et al. charge set was derived based on condensed phase quantum chemical calculations. The accuracy observed here suggests that the Duan et al. charges are reasonably compatible with the side-chain torsion parameters.

**Multiple buried side chains**

Simultaneous refinement of multiple buried side chains is the next level of challenge. Formally, this is a combinatorial problem because one may potentially try all possible combinations to find the optimum conformation when the side chains are correctly assigned. We first tested the method against a case
### Table I. Statistical results of single side-chain assignment on each buried residue

<table>
<thead>
<tr>
<th>Residue</th>
<th>Residue index</th>
<th>Initial $\delta_{X_1}$ (°)</th>
<th>Initial $\delta_{X_2}$ (°)</th>
<th>Success ($\chi_{1,2}$) (cycles)$^b$</th>
<th>Final ($\delta_{X_1}$) (°)</th>
<th>SD ($\delta_{X_1}$) (°)</th>
<th>Final ($\delta_{X_2}$) (°)</th>
<th>SD ($\delta_{X_2}$) (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leu</td>
<td>13</td>
<td>119.0</td>
<td>17.2</td>
<td>4</td>
<td>3.1</td>
<td>2.7</td>
<td>4.9</td>
<td>2.8</td>
</tr>
<tr>
<td>Ile</td>
<td>15</td>
<td>134.5</td>
<td>15.3</td>
<td>52</td>
<td>4.7</td>
<td>4.2</td>
<td>3.2</td>
<td>3.0</td>
</tr>
<tr>
<td>Trp</td>
<td>17</td>
<td>120.6</td>
<td>95.8</td>
<td>1</td>
<td>2.5</td>
<td>1.7</td>
<td>4.1</td>
<td>4.0</td>
</tr>
<tr>
<td>Val</td>
<td>24</td>
<td>122.3</td>
<td>27</td>
<td>9</td>
<td>4.8</td>
<td>1.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tyr</td>
<td>27</td>
<td>122.2</td>
<td>93.3</td>
<td>1</td>
<td>2.8</td>
<td>2.2</td>
<td>3.7</td>
<td>2.8</td>
</tr>
<tr>
<td>Ile</td>
<td>29</td>
<td>127.0</td>
<td>4.5</td>
<td>9</td>
<td>2.6</td>
<td>1.5</td>
<td>4.9</td>
<td>2.8</td>
</tr>
<tr>
<td>Thr</td>
<td>30</td>
<td>4.8</td>
<td>1</td>
<td>2.4</td>
<td>3.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ile</td>
<td>54</td>
<td>130.0</td>
<td>3.4</td>
<td>4</td>
<td>6.8</td>
<td>3.8</td>
<td>3.3</td>
<td>2.6</td>
</tr>
<tr>
<td>Leu</td>
<td>57</td>
<td>124.0</td>
<td>117.7</td>
<td>4</td>
<td>2.4</td>
<td>1.8</td>
<td>5.6</td>
<td>4.5</td>
</tr>
<tr>
<td>Tyr</td>
<td>63</td>
<td>108.1</td>
<td>93.3</td>
<td>1</td>
<td>4.8</td>
<td>0.0</td>
<td>16.3</td>
<td>0.0</td>
</tr>
<tr>
<td>Ile</td>
<td>65</td>
<td>127.4</td>
<td>7.0</td>
<td>11</td>
<td>3.9</td>
<td>1.8</td>
<td>6.3</td>
<td>2.4</td>
</tr>
<tr>
<td>Thr</td>
<td>66</td>
<td>116.7</td>
<td></td>
<td>1</td>
<td>0.8</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Val</td>
<td>67</td>
<td>123.1</td>
<td></td>
<td>1</td>
<td>4.6</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Val</td>
<td>70</td>
<td>128.1</td>
<td></td>
<td>13</td>
<td>4.8</td>
<td>1.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thr</td>
<td>71</td>
<td>9.7</td>
<td></td>
<td>11.3</td>
<td>5.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ser</td>
<td>79</td>
<td>112.9</td>
<td></td>
<td>18 (76)</td>
<td>6.0</td>
<td>2.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ile</td>
<td>83</td>
<td>119.1</td>
<td>0.5</td>
<td>1</td>
<td>2.9</td>
<td>2.3</td>
<td>2.9</td>
<td>2.2</td>
</tr>
</tbody>
</table>

$^a$ $\delta_{X_1}$ and $\delta_{X_2}$ represent the deviations from native $\chi_1$ and $\chi_2$, respectively. $^b$ Success (cycles) represents the first time when both sampled the ‘correct’ conformations (deviation < 40°). The numbers in parentheses indicate that the side chains moved to other conformations later and the numbers represent the final cycle when they moved back to the correct conformation.

### Table II. Statistical results of multi-side-chain assignment on all buried residues

<table>
<thead>
<tr>
<th>Residue</th>
<th>Residue index</th>
<th>Initial $\delta_{X_1}$ (°)</th>
<th>Initial $\delta_{X_2}$ (°)</th>
<th>Success ($\chi_{1,2}$) (cycles)$^b$</th>
<th>Final ($\delta_{X_1}$) (°)</th>
<th>SD ($\delta_{X_1}$) (°)</th>
<th>Final ($\delta_{X_2}$) (°)</th>
<th>SD ($\delta_{X_2}$) (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ile</td>
<td>15</td>
<td>134.5</td>
<td>15.3</td>
<td>1</td>
<td>3.5</td>
<td>4.0</td>
<td>2.0</td>
<td>1.7</td>
</tr>
<tr>
<td>Trp</td>
<td>17</td>
<td>120.6</td>
<td>95.8</td>
<td>1</td>
<td>2.2</td>
<td>1.0</td>
<td>5.4</td>
<td>3.2</td>
</tr>
<tr>
<td>Val</td>
<td>24</td>
<td>121.4</td>
<td></td>
<td>11</td>
<td>3.3</td>
<td>2.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tyr</td>
<td>27</td>
<td>122.2</td>
<td>93.3</td>
<td>20</td>
<td>1.9</td>
<td>0.2</td>
<td>4.1</td>
<td>1.6</td>
</tr>
<tr>
<td>Ile</td>
<td>29</td>
<td>127.0</td>
<td>4.5</td>
<td>7 (104)</td>
<td>4.0</td>
<td>2.1</td>
<td>2.4</td>
<td>2.4</td>
</tr>
<tr>
<td>Thr</td>
<td>30</td>
<td>4.8</td>
<td></td>
<td>1</td>
<td>4.5</td>
<td>2.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ile</td>
<td>54</td>
<td>129.0</td>
<td>3.4</td>
<td>1 (148)</td>
<td>12.2</td>
<td>7.7</td>
<td>3.6</td>
<td>6.5</td>
</tr>
<tr>
<td>Leu</td>
<td>57</td>
<td>124.0</td>
<td>117.7</td>
<td>1 (74)</td>
<td>5.1</td>
<td>4.5</td>
<td>5.6</td>
<td>3.4</td>
</tr>
<tr>
<td>Tyr</td>
<td>63</td>
<td>108.1</td>
<td>93.3</td>
<td>1</td>
<td>3.4</td>
<td>1.3</td>
<td>3.9</td>
<td>4.4</td>
</tr>
<tr>
<td>Ile</td>
<td>65</td>
<td>127.4</td>
<td>7</td>
<td>46</td>
<td>8.5</td>
<td>5.8</td>
<td>6.7</td>
<td>2.4</td>
</tr>
<tr>
<td>Thr</td>
<td>66</td>
<td>116.7</td>
<td></td>
<td>138</td>
<td>3.9</td>
<td>2.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Val</td>
<td>67</td>
<td>123.1</td>
<td></td>
<td>43</td>
<td>5.1</td>
<td>2.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Val</td>
<td>70</td>
<td>128.1</td>
<td></td>
<td>7</td>
<td>2.9</td>
<td>1.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thr</td>
<td>71</td>
<td>9.7</td>
<td></td>
<td>1</td>
<td>5.2</td>
<td>2.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ser</td>
<td>79</td>
<td>112.9</td>
<td></td>
<td>1</td>
<td>5.9</td>
<td>2.7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$^a$ $\delta_{X_1}$ and $\delta_{X_2}$ represent the deviations from native $\chi_1$ and $\chi_2$, respectively. $^b$ Success (cycles) represents the first time when both sampled the ‘correct’ conformations (deviation < 40°). The numbers in parentheses indicate that the side chains moved to other conformations later and the numbers represent the final cycle when they moved back to the correct conformation.

**Fig. 4.** Sampling of a single buried side chain Ile54 of protein 1fna during the simulation as measured by the deviations of $\chi_1$ (black) and $\chi_2$ (gray) angles from the native position (dchi) at the end of each cycle (A) and after a Monte Carlo step (B).
of five buried residues. These five residuals were close in space with typical distance of about 3.0–4.0 Å. The results are shown in Figure 5. Both $\chi_1$ and $\chi_2$ of all five residues were correctly assigned after $\sim 1.2$ ns. In comparison with single residue assignment, a longer time was needed to find the native conformations. Even though all five residues found their respective correct conformations individually much earlier than 1.2 ns (Figure 5A), the correct conformations were not kept in the Monte Carlo steps. Nevertheless, when all five residues reached the correct side-chain conformations, they were selected by the Monte Carlo procedure because the correct conformation has the lowest energy.

We further tested the method on all buried residues of the protein 1fna. The terminal residues and Gly and Ala residues were excluded and a total of 15 buried residues were chosen which constituted the entire core of the protein and were closely packed. Figure 6 shows the actual samplings and final assignments of the 15 residues. The results are also summarized in Table II. Among the 15 side chains, the $\chi_1$ torsion angles of 13 side chains were initially poorly assigned owing to the random assignment and the initial deviations were $>100^\circ$. Similarly, half of the $\chi_2$ torsion angles were also poorly assigned. Within $\sim 1.5$ ns, all $\chi_1$ and $\chi_2$ were corrected by the Monte Carlo procedure. Even though almost all of the side chains started from completely wrong initial conformations, 13 of the 15 side chains (except Ile65 and Val43) found their native conformations almost immediately. After having found its native conformation, Thr66 moved to a non-native position later and became the last residue that eventually went back to the correct conformation. When it finally moved back to its native position at $\sim 1.5$ ns, all the side chains were in their native conformations and all the native side-chain conformations were kept for the rest of the simulation time.

One challenge for side-chain assignment is that the problem is inherently combinatorial. Because the search space grows exponentially as the number of side chains increases, a much longer time is needed to find the optimal packing when the number of residues increases. This appeared to be less of an issue because a similar time was needed to find the native assignment for both five residues ($\sim 1.2$ ns) and 15 residues ($\sim 1.5$ ns). This was perhaps because the method relies on the physical interactions among residues and allows them to adjust to the optimal conformations gradually and to adjust to the favorable positions dictated by the energy profile. The reduced side chains allow them to rotate relatively freely. As a
consequence, the side chains experience a smooth energy landscape which funnels them to the optimal position dictated by the force field. This scenario is analogous to what has been envisaged by the folding funnel theory in which the funnel-shaped protein folding free energy landscape helped to reduce an inherently combinatorial problem. It is noted that although these results are encouraging, they are insufficient to demonstrate the elimination of the combinatorial problem.

It is also noteworthy that nearly all the refined side chains were within a 10° deviation from their native conformations. The largest deviation was for Ile54 whose average $\chi_1$ was about 12.2° from its native conformation. This high level of accuracy is encouraging.

To examine the accuracy further, we conducted the method on five other proteins selected from the PDB. The results were compared with SCWRL3 (Canutescu et al., 2003) (http://dunbrack.fccc.edu/SCWRL3.php) assignments (Table III). With the G2FMD method, all the $\chi_1$ and $\chi_2$ of each protein were accurately assigned within 1.6 ns of the 2.2 ns simulations and the assignment accuracy was nearly 100%. The only exception was one Lys side chain of protein 1vie, whose assigned $\chi_1$ angle was 53° away from its native position. This particular residue was 19.6% exposed, slightly below the cutoff (20%) of a buried residue. Therefore, its conformation could have been under the influence of solvent. Nevertheless, similarly to the 1fna case, most of the residues were assigned quickly within only a few Monte Carlo cycles.

The average refinement result ($\chi_1$ and $\chi_2$) for each protein showed less than a 10° deviation from their native conformations. The successful assignments on different proteins with such high accuracy were a clear validation of the method on buried residues. Overall, 94% of the torsion angles were within 20° of the native conformation, 79% were within 10° and 42% were within 5°. Although SCWRL3 also showed an excellent result of 93.5% accuracy, nine out of the 139 dihedral angles were more than 40° away from native conformations. Because of those incorrectly assigned residues, the average deviation from the native conformation was systematically larger than in our results. The successful assignment for different proteins with such high accuracy was a clear validation of the G2FMD method for buried residues. This comparison illustrates that the G2FMD method has a slight accuracy advantage over SCWRL3; the latter represents one of the best rotamer-based side-chain assignment methods.

**Exposed side chains**

As an example, Figure 7 shows the results for an exposed residue Glu41 of protein 1fna in a single side-chain assignment test. The differences between the initial random assignment and the native conformations were 116.9° and 132.6° for $\chi_1$ and $\chi_2$, respectively. Figure 7A shows that the conformational sampling of a regular restrained MD simulation. Figure 7B shows the sampling results of G2FMD simulation, which presented the actual conformation of each fully re-grown snapshot. Figure 7C is the saved conformation after Monte Carlo step in G2FMD (C).
representation and solvent models. Another factor that might have an important contribution is the crystal packing effect, since the experimental structures were obtained in a crystalline environment and our assignment was done in the solution phase. Similar results were obtained on other exposed residues (Lys49, Lys58, Lys81) (data not shown). Despite this, comparison between G2FMD (Figure 7A) and MD (Figure 7C) reinforced the notion that the G2FMD method significantly enhances the sampling efficiency: MD tends to stay in a certain conformation for a long time whereas G2FMD was able to sample among different conformations much more frequently because of the smoother energy landscape.

Side-chain assignments of the entire proteins

We tested the G2FMD on a set of 21 proteins, including the six that were used in buried side-chain tests discussed earlier. In this test, the assignments were performed on the whole proteins. The final assignment results are presented in Figure 8. The side-chain angle was considered correctly assigned if it was within 40° of its experimental value, which is a typical criterion used in most side-chain assignment studies.

Among the assignments, the buried side chains were systematically better (73 and 59% for \( \chi_1 \) and \( \chi_{1+2} \), respectively) than the exposed residues (56 and 42%). This suggests that the solvent model and, perhaps, crystal packing played important roles. In comparison, nearly 100% accuracy was achieved in the assignments of the buried side chains when the exposed side chains were kept near their native conformations. This was probably an indication of a cooperative effect in the sense that the interactions between the exposed and the buried side chains can influence the assignment; a lower assignment accuracy of the exposed side chains can propagate and affect the accuracy of the buried ones. Nevertheless, overall, the average assignment accuracy of the 21 proteins was 64 and 51% for \( \chi_1 \) and \( \chi_{1+2} \), respectively, including both buried and exposed side chains.

Discussion

In the G2FMD method, accurate side-chain assignment requires both the ability to locate the low-energy conformations and the accurate representation of the energy surface, particularly in terms of side-chain energies. Hence, to a certain extent, our method puts the accuracy of the force field and solvent models to a serious test. Although the accuracy of the force field is a very important issue, our limited goal in this work was to examine whether or not the proposed method can enhance the conformational sampling by making the energy surface smoother and can do so while preserving the important features of the free energy landscape. The large steric energy barriers at the protein interior make it very difficult to search the conformational space with ab initio methods. From this perspective, tests on the buried side chains clearly demonstrated that the proposed method indeed can significantly enhance sampling.

A common problem confronting all side-chain assignment methods is that small errors in the backbone coordinates can force an incorrect choice of side-chain rotamers (Wilson et al., 1993). Samudrala et al. (2000) compared the accuracy of different side-chain assignment methods by constructing side chains on near-native backbones (C\(_\alpha\) r.m.s.d. \(<4.0\) Å). They found that a simple naive method can achieve an accuracy comparable to more sophisticated methods. Furthermore, according to the results of CASP, the all-atom r.m.s.d. of the best structures is rarely better than 2.0 Å even for the buried side chains (Tramontano and Morea, 2003). With this level of main-chain structure accuracy, incorrect packing and misassignment of side chains can occur that may prevent further refinement. Therefore, methods that can combine side-chain assignment and main-chain structure refinement are very much needed. Furthermore, separation of the main-chain (protein structure) prediction and side-chain assignment has limited the accuracy of both steps. A natural synergy is to allow main chains to adjust during side-chain assignment. This is possible only when a method becomes available that reduces the energy barriers due to side-chain packing and represents the side chains with sufficient detail and accuracy. Gray et al. tried simultaneous refinement of the backbone and side-chain conformations with a statistical-based energy function and achieved a 50% success rate (Gray et al., 2003). The proposed G2FMD method may help by producing collision-free structures.
Many computational methods have been devised to facilitate the search for the global energy minimum of complex systems. The G2FMD method smoothed the energy landscape by decreasing the size of side-chain groups and enables proteins to transfer among different conformations without unfolding the backbone structure. Therefore, we were able to apply molecular mechanics simulation to the side-chain assignment. The method showed encouraging results on buried side-chain assignments and demonstrated nearly 100% accuracy on all six randomly picked proteins. Although other side-chain assignment methods also successfully assigned buried residues with better than 90% accuracy (Xiang and Honig, 2001; Liang and Grishin, 2002; Canutescu et al., 2003; Peterson et al., 2004), simultaneous assignment of all the buried side chains correctly and the notably small deviation from the native conformations, especially for both \( \chi_1 \) and \( \chi_2 \), were the best results.

Finding the optimal side-chain packing is a common issue in structure prediction, protein design and protein docking. Hence the encouraging results for the side-chain assignment suggest that a similar strategy can be applied in those fields. Camacho and Vajda (2001) developed a tunable protein–protein docking method which slowly tunes in the contribution of the van der Waals energy in their ranking free energy. The methods successfully refined 10 rigid body receptor–ligand complexes from 10 to 2 Å r.m.s.d. The idea of smooth energy by turning down the van der Waals term is similar to our G2FMD method. Here we did that by actually changing the size of side-chain groups. One advantage of our method is that it solves the steric collision problem and enables side chains to rearrange more easily.

It is common that the core residues are assigned more accurately than the exposed ones. The exposed side chains interact with the solvent directly and are considerably more mobile than the buried side chains. As a consequence, their orientation is much less well determined than that of buried ones. In addition, some of the exposed side chains are involved in the crystal packing and removal of the crystal packing interactions could have an adverse effect on their native conformations. Jacobson et al. (2002) studied the crystal packing effect on protein side-chain conformations and found that inclusion of the crystal environments could significantly improve the accuracy of their side-chain assignment. In our case, the continuum solvent model might also have played a role. These problems could be corrected by further refinement of the energetic terms, which is beyond the scope of the present study.

Accurate force field and sufficient conformational search are the two essential parts to predicting good protein structures. The current force fields are still imperfect and insufficient conformational search might still be a problem, especially in the case when the side chains of whole proteins are included. To improve the sampling efficiency further, we also tried to utilize the temperature annealing method in G2FMD. In those simulations (data not shown), we applied 100 ps simulated annealing by raising the temperature to 800 K before side-chain growth. Although simulated annealing helped to locate some side chains at the correct conformation in the shrunk states, most of the correct assignments were found after the side-chain growth process. Therefore, we did not observe a significant improvement in final assignment.

One potential problem with the G2FMD method was that the hydrogen bonds among side-chain groups could break when side chains were reduced, causing difficulty for those side-chain groups to find their native conformations. Fortunately, their conformations are influenced by the physical driving forces when the side-chain groups grow back to their normal size. In our tests using 1fna (Tables I and II), six of the buried side chains were polar and formed hydrogen bonds, including Tyr27, Thr30, Tyr63, Thr66, Thr71 and Ser79. In all cases, we observed that these side chains were correctly assigned. Furthermore, there was no clear indication of any difficulty specifically related to these side chains. Hence it appears that the problems, if there are any, are not significant. This was probably because of the particular procedure that allows the side chains to adjust gradually to the physical environment when they grow back to their normal size.

**Conclusion**

An ab initio side-chain assignment and refinement method has been presented. The method overcomes the steric collision problem, which is the main reason for insufficient conformational sampling of physics-based models and enabled molecular mechanics simulations to be applied to protein side-chain assignment and refinement. The method was tested on a small set of proteins and ~100% accuracy was achieved on both \( \chi_1 \) and \( \chi_2 \) of buried residues. Overall, more than 94% of the torsion angles were within 20° of the native conformation, 79% were within 10° and 42% were within 5°. The method has the potential to be applied further to protein structure refinement.

**Acknowledgements**

Thanks are due to Drs Hongxing Lei and Zhixiang Wang and other members of Duan group for helpful discussions. This work was supported by research grants from the NIH (GM64458 and GM67168 to Y.D.).

**References**


Received March 18, 2005; revised October 15, 2005; accepted October 24, 2005

Edited by Richard Goldstein