Andante: reducing side-chain rotamer search space during comparative modeling using environment-specific substitution probabilities

Richard E. Smith¹,⁎, Simon C. Lovell¹,†, David F. Burke¹, Rinaldo W. Montalvao¹ and Tom L. Blundell¹

¹Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK

Received on July 10, 2006; revised on February 21, 2007; accepted on February 25, 2007

Advance Access publication March 6, 2007

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: The accurate placement of side chains in computational protein modeling and design involves the searching of vast numbers of rotamer combinations.

Results: We have applied the information contained within structurally aligned homologous families, in the form of conserved χ angle conservation rules, to the problem of the comparative modeling. This allows the accurate borrowing of entire side-chain conformations and/or the restriction to high probability rotamer bins. The application of these rules consistently reduces the number of rotamer combinations that need to be searched to trivial values and also reduces the overall side-chain root mean square deviation (rmsd) of the final model. The approach is complementary to current side-chain placement algorithms that use the decomposition of interacting clusters to increase the speed of the placement process.

Contact: res50@mole.bio.cam.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

The accurate selection of amino acid side-chain conformations is an ever-present challenge in the computer modeling of protein structures. The major limiting factor is that the number of rotamer combinations increases rapidly, resulting in an NP hard optimization problem that quickly makes exhaustive searches computationally intractable (Pierce and Winfree, 2002). A variety of dedicated search algorithms has been devised to explore side-chain conformational space. These algorithms assume a correct, fixed backbone conformation and then try to solve the combinational optimization problem to find an accurate ensemble of side-chain conformations for the defined backbone by means of some test function. Assuming the backbone topology is accurate, the placement of side chains then becomes a search to identify the global minimum energy conformation for a set of side-chain conformations defined by the target protein’s sequence. Approximating the global minimum energy conformation is made possible by the use of knowledge-based, representative libraries of statistically significant, low energy conformations side-chain conformations (De Maeyer et al., 1997; Dunbrack and Cohen, 1997; Dunbrack and Karplus, 1993; Lovell et al., 2000; Ponder and Richards, 1987) commonly referred to as rotamers. There are numerous different search strategies that make use of these rotamer libraries, the most common of which are Monte Carlo methods (Holm and Sander, 1991), genetic algorithms (Tuffery et al., 1991), simulated annealing (Lee and Subbiah, 1991), mean field optimization (Delarue and Koehl, 1997; Vasquez, 1995), dead-end elimination algorithms (Goldstein, 1994; Lasters and Desmet, 1993; Lasters et al., 1995; Looger and Hellinga, 2001; Pierce et al., 2000), integer programming approaches (Kingsford et al., 2005), tree-searching algorithms (Bower et al., 1997; Canutescu et al., 2003; Gordon and Mayo, 1999) and graph theory approaches (Canutescu et al., 2003; Xie and Sahinidis, 2006; Xu, 2005). These approaches have various completion times, but a reduction in run time usually has a corresponding tradeoff in accuracy (Voigt et al., 2000).

In the context of comparative modeling, structural information from homologous templates can also be exploited in the side-chain placement problem. Sutcliffe et al. (Sutcliffe et al., 1987) derived 20 × 20 rules for building side chains from template conformations based on an analysis of positions of atoms in substituted amino acids in families of homologous proteins. Summers et al. (Summers and Karplus, 1989; Summers et al., 1987) analyzed the conservation of χ angles at topologically equivalent positions in a small set of proteins and used the information to model side chains in rhizopus-pepsin (Summers and Karplus, 1989). Ogata and Umeayama (Ogata and Umeayama, 1998) applied a similar approach to tightly clustered positions of superimposed homologues and investigated the influence of local space homology on the effects of χ angle conservation. Sali and Blundell (Sali and Blundell, 1993) modeled protein side chains by satisfaction of spatial restraints making use of probability density functions of rotamer conformations derived from the structures of a large number of homologous families (Sali and Blundell, 1993). In the work presented here, we utilize the structural
environment information for amino acid residues contained in the HOMSTRAD (Mizuguchi et al., 1998; Stebbings and Mizuguchi, 2004) database to extend the $\chi$ angle conservation approach to substitution of residues classified in six structural environments and four percentage identity (PID) ranges for $\chi_1$, $\chi_1 + 2$ and $\chi_1 + 2 + 3$ angles. The usefulness of environment-specific propensities (Rice and Eisenberg, 1997; Shi et al., 2001) and substitution tables (Overington et al., 1990; Overington et al., 1992) has been successfully demonstrated in the areas of sequence-structure homology fold recognition (Rice and Eisenberg, 1997; Shi et al., 2001). Substitution probabilities have also been used for fragment ranking in homology modeling (Topham et al., 1993). The goal of this work is to make extensive use of the $\chi$ angle conservation probabilities in order to contain the explosion of rotamer combinations in two separate ways. The program Andante systematically applies these probabilities as preliminary filters to ‘borrow’ (use exact parent $\chi$ angles) entire, high probability, side-chain conformations and/or then restrict possible rotamer solutions for non-borrowed positions to high probability $\chi$ angle bins. This results in a significant reduction in the rotamer search space during the rest of the side-chain placement process, principally because the size of the interacting clusters of amino acid side chains that must be solved is much smaller. The application of homologous template information results in models with reduced side-chain root mean square deviation (rmsd) to the original target that can be seen by comparison of the borrowed conformations with the equivalent rotameric solutions provided by SCWRL3.0 in the absence of using information from homologues. The stand-alone SCWRL programs do have the facility to retain desired side-chain conformations (Bower et al., 1997; Canutescu et al., 2003). However, they do not have an automated, pre-processing step to evaluate multiple homologous template information. In order to allow comparison, we also present results obtained by SCWRL3.0 when it is provided with borrowed side-chain information extracted by the Andante program.

2 METHODS

2.1 Environment-specific $\chi$ angle conservation analysis

In order to derive borrowing/restricting rules for amino acid side-chain conformations, the structural alignments of 637 families contained in the HOMSTRAD database were analyzed for substitutions of residue A in a specific environment, E, to residue B in any environment. Six structural environments were defined by considering all combinations of three secondary structure states (helix, strand and coil) and two measures of solvent accessibility (buried or exposed). The structural environments were assigned using the Joy suite of programs (Mizuguchi et al., 1998). The analysis of $\chi$ angle conservation was carried out for the following residue types: all residues (excluding Ala and Gly) for $\chi_1$, $\chi_1 + 2$ or $\chi_1 + 2 + 3$. Pro has four $\chi$ angles, but $\chi_1$ defines the other three. For Pro residues, there were no Pro–Pro–non-Pro substitutions observed with a high probability of $\chi_1 + 2$ or $\chi_1$ alone being conserved, therefore, building/restricting information is limited to Pro–Pro substitutions only. The analysis of $\chi_1 + 2 + 3$ for Glu, Gln and Met was performed to investigate whether entire conformations could be borrowed in certain circumstances. Arg and Lys were included in this group in order to establish if restricting out to the $\chi_3$ angles could be performed. Only crystal structures with a resolution of 2.5Å better were examined. Substitutions to and from incomplete side chains and side chains with atoms having occupancy <1.0 were ignored. Only side chains with all atoms having a B-factor <40.0 were used. $\chi$ angles were considered conserved if found to be in the range $+\text{30}^\circ$ to $-\text{30}^\circ$ of each other. The results were partitioned into four percentage identity ranges, 0–19, 20–39, 40–59, >60%, respectively, based on the structural alignments contained in HOMSTRAD. The limitation of a maximum of 90% PID for HOMSTRAD family members means each family can contribute to the statistics in each of the four PID ranges. This means larger multimember families do not heavily bias the statistic in any single PID range. Corrections for two-fold, symmetric side chains and chemically equivalent positions were carried out in an analogous manner to that outlined in Summers et al. (Summers and Karplus, 1989) and Ogata and Umeyama (Ogata and Umeyama, 1998). No corrections were made for Leu or Val residues. The environment-specific $\chi$ angle conservation probabilities were calculated from the HOMSTRAD structural alignments as follows. For each alignment position, $i$, $\chi$ angle conservation ($\chi_1$, $\chi_1 + 2$, $\chi_1 + 2 + 3$ where applicable) substitutions were counted for residue type A in environment E, going to residue type B in any environment, $A_i^E - B$. The reverse substitution was also counted, $B_j^E - A_i$, as substitution patterns are not necessarily symmetric. The counts were tabulated for four PID ranges (0–19, 20–39, 40–59, >60% PID). The probability of $\chi$ angle conservation, $P_c$, (Equation 1) was given by total number of conserved $\chi$ angles counted for a given substitution, $(A_i^E - B_j)$, divided by the raw total of substitutions observed for that given substitution, $(A_i^E - B_j)$.

$$P_c = \frac{A_i^E - B_j}{(A_i^E - B_j)}$$

A total of 2046 crystal structures was used giving rise to 831,616 observed substitutions. Substitutions with a low-total observed count (<50 data points) were discarded. This produced 1832 $\chi$ conservation rules where the probability that the $\chi$ angle conservation was over 50%. There are 56 rules for $\chi_1 + 2 + 3$ (41 for buried template environment, 15 for exposed), 424 for $\chi_1 + 2$ only (217 for buried, 210 for exposed) and 1352 for $\chi_1$ only (586 for buried, 766 for exposed positions). The energy function used for van der Waals overlap clash checks is that used by Bower and Dunbrack (Bower and Dunbrack, 1997, 2003). The total fixed energy, $E_{rot}$, for any single rotamer, $j$, of residue $i$, is given by Equation 2.

$$E_{rot,i} = E_{Ebb,i} + E_{Econs,i} + E \log P_{c,i}$$

$E_{bb,i}$ is the clash energy of the rotamer atoms with the backbone atoms of the target. This is calculated for all non $i$ residue atoms. C-$\beta$ atoms are considered to be part of the backbone. The threshold for tolerating backbone clashes was set to 12.0 kcal/mol. $E_{Econs,i}$ is the clash energy with ‘borrowed’ (locked) rotamer conformations. The default threshold for tolerated clashes between borrowed conformations was set to 10.0 kcal/mol. $E_{logP_{c,i}}$ is a pseudo-energy term based on the log of the rotamer library probability of observing that rotamer conformation. These values are normalized against the probability of the highest library rotamer for a given residue.

2.2 Program flow

2.2.1 Disulphide bond placement Andante uses the Grade A definition of Sowdhhamini et al. (Sowdhhamini et al., 1989) in order to build possible disulphide bonds. By definition, a Grade A disulphide bond meets the following criteria: for any pair of Cys residues, Cys$_i$ and Cys$_j$, the C$_{\gamma}$–C$_{\gamma}$ and C$_{C}–C_{C}$ distances must be less than 7.0 and 4.0 Å, respectively; the dihedral angle across C$_{\gamma}$–S$_{\gamma}$–S$_{\gamma}$–C$_{\gamma}$ must be in the range $60–120^\circ$ with a $S_i$ to $S_j$ distance between 1.8 and 2.2Å; finally, both the Cys residues must also have a $\chi$ in the range of $[30–90]^\circ$ or $[150–180]^\circ$. In order to allow for errors in the modeling of the backbones, some flexibility in the Cys rotamers is allowed. For each
Cys rotamer, an additional rotamer with a χ1 +/− 15° is created from each Penultimate Rotamer Library (Lovell et al., 2000) rotamer, giving a total of nine Cys rotamers. All rotamer combinations for all Cys–Cys pairs are searched. If the configuration is considered Grade A, a score (Equation 3) is assigned to that rotamer pair and that Cys–Cys pair.

\[
score = (2.04 - R_n) + (90 - X_n)/10 + E \log P(i) + E \log P(j) + E_{bb}(i)/10.0 + E_{bb}(j)/10.0
\]

where \( R_n \) is the \( S_i - S_j \) distance, \( X_n \) is the dihedral angle across \( C_{\beta} - S_i - S_j - C_{\beta} \). ElogP is the normalized pseudoenergy attributed to each Cys rotamer based on the probability of it being observed. \( E_{bb}(i) \) is the backbone clash energy for rotamers \( i \) and \( j \). All rotamer combinations are tried for each Cys–Cys pair within the allowed \( C_n \) and \( C_{\beta} \) cutoffs. The lowest scoring (closest to ideal) disulphide bond found from all pairs is constructed if it fits the Grade A criteria. The search is then repeated until no more disulphide bonds are found.

### 2.2.2 Construction of high probability side-chain conformations (Borrowing phase)

The application of this type of homologous template information can be performed using any combination of five steps in the order described below. The first three steps involve the ‘borrowing’ of entire rotamer conformations from templates at positions having a high probability that enough \( \chi \) angles to define the entire side-chain are conserved. The final two steps involve imposing the possible rotamer solutions to the non-borrowed positions to library rotamers that have \( \chi \) angles that are within 30° of a suitably high-scoring template position as defined by the conservation probabilities. The numbers of \( \chi \) angles that are used as restrictions depend on which rules are applied. The order that the borrowing/restricting rules are applied is such that the larger, hydrophobic residues (set A: DFHILNYWP) are considered first, followed by set B (set B: EQM), which is much less populated and has considerably fewer borrowing rules (see Table S1). Finally, the smaller residues of set C (set C: STVC) are considered last. The objective of using set A first is to correctly fill out a large volume of the core, as this has numerous downstream benefits (see Results section).

#### 2.2.2.1 \( \chi_1 + 2 \) definable positions

The first borrowing step is to build side-chain conformations definable by two \( \chi \) angles (set A). For each sequence-structure alignment position, \( i \), the following occurs. Working in decreasing order of PID to the target, the observed substitution between template and target is determined. If both the template residue and the target residue are members of set A, and the environment-specific probability of that entire rotamer configuration is conserved and exceeds a cutoff threshold, a substitution weight (Equation 4) is generated for that pair of residues.

\[
\text{Weight} = 0.6 \times \text{PID} + 0.4 \times P_{12}
\]

where, \( P_{12} \) is the probability that both \( \chi_1 \) and \( \chi_2 \) angles are conserved for the observed substitution of template residue, A, in environment, \( E \), going to target residue, \( B \). The substitution weight is variable depending on the substitution and the PID between the target and template sequences. Arbitrary weights of 0.6 and 0.4 are assigned to the PID, and probability portions of the substitution weight calculated by Equations 4–6. This gives preference to higher PID templates when multiple templates are used and identical substitutions exceed the cut-off probability. The entire rotamer configuration is then built on to the backbone template of the target using the necessary \( \chi \) angles obtained from the template residue. If the conformation is compatible with the backbone and does not clash with another higher-scoring borrowed position, then it is retained. If the conformation does clash with another higher scoring borrowed position, then the position with the lower weight is discarded. For each residue, the single highest scoring conformation from all corresponding template positions is retained.

#### 2.2.2.2 \( \chi_1 + 2 + 3 \) definable positions (Borrowing phase)

The above procedure is then repeated for residue positions definable by three \( \chi \) angles (set B). These positions are assigned a weight in a similar fashion (Equation 5) to those positions defined by two \( \chi \) angles.

\[
\text{Weight} = 0.6 \times \text{PID} + 0.4 \times P_{123}
\]

where, \( P_{123} \) is the probability that all \( \chi_1 + 2 + 3 \) angles are conserved for the observed substitution of template residue, \( A \), in environment, \( E \), going to target residue, \( B \).

#### 2.2.2.3 \( \chi_1 \) definable positions (Borrowing phase)

The above procedure (2.2.2.1) is then repeated for residue positions definable by one \( \chi \) angle (set C). Here the substitution weight is given by Equation 6.

\[
\text{Weight} = (0.6 \times \text{PID} + 0.4 \times P_1)/10.0
\]

where, \( P_1 \) is the probability that the \( \chi_1 \) angle is conserved for the observed substitution. The highest scoring positions are built and checked for structural compatibility as described above. The scaling factor (10.0) ensures that residues with predicted \( \chi_1 + 2 + 3 \) are retained over those where only \( \chi_1 \) is predicted. After the borrowing phase only the highest scoring side-chain configuration for any given residue is retained, though it can be selected from any template structure.

#### 2.2.2.4 Restricting rotamer solutions to \( \chi \) bins

A further use of the homologous template information is the application of the conservation probabilities to any non-borrowed positions. These are positions where the probability that all the \( \chi \) angles necessary to define an entire side-chain configuration are conserved is below the cut-off threshold. However, there may be information that \( \chi_1 \) or \( \chi_1 + 2 \) may be conserved for an observed substitution. Instead of considering all possible rotamers solutions for these positions, the substitution probabilities are then reapplied and used to restrict possible library rotamer solutions to \( \chi \) angle bins defined by the highest scoring template side-chain conformations. For non-borrowed or non-disulphide bonded positions, initial side-chain conformations and interacting clusters are defined in a manner analogous to that outlined by Bower et al. (Bower et al., 1997). These clusters are searched using a standard simulated annealing protocol to determine the lowest energy ensemble for each cluster. For the jack-knife testing two rotamer libraries were used: the secondary structure-dependent library of Lovell (Lovell et al., 2000) and the backbone-dependent library of Dunbrack (Bower et al., 1997).

### 3 RESULTS AND DISCUSSION

As would be expected the highest probabilities for conserved \( \chi \)-angles are for substitutions involving identical or chemical similar residues at PID >60% in buried environments. Specifically, FY/FY, W/W and H/H make up the majority of rules showing conserved probabilities >90% in this PID range. The lowest probabilities are observed for substitutions from exposed positions. For example, the lowest recorded probabilities were observed at \( \chi_1 \) were seen for P (exposed, helix)/E and P (exposed, coil)/E at <11% both in the PID range 0–20%. At \( \chi_1 + 2 \), the lowest observed probabilities were for D (coil, exposed)/L (10.1%, PID range 0–19%) and N (coil, exposed)/R (10.2%, PID range 40–59%). An interesting point resulting from the way the rules were derived is that some rules could be considered family specific. For example, considering \( \chi_1 \) only, N (buried, helix)/H has a very high probability (~95%) of being
conserved in the lowest PID range (0–20%). However, this is a rarely observed substitution only seen in three HOMSTRAD families. Extended listings (down to 10% probability) of all the calculated probabilities used for the jack-knife test are given in the Supplementary Material (see S_chi1.txt, S_chi12.txt, S_chi123.txt). A jack-knife test was performed on 14 HOMSTRAD families (Table 1). The families were selected on the basis that the target structure did not contain any co-factors, ligands or non-standard amino acids. All template information was taken from crystal structures with resolutions of 2.50 Å or better. Also, a minimum of three templates was required in order to make use of the information contained in multiple template structures. Borrowing and restricting thresholds (conserved χ probability) were set at 65 and 70%, and both the Dunbrack and Lovell rotamer libraries were used. Setting the cut-off probability threshold at 70% reduced the overall number of borrowing/restricting rules enforceable to 828 from the original 1832 (Table S1). Twelve different combinations of the borrowing/restricting procedures were used to validate the borrowing rules and to investigate the overall effect the method has on the number of combinations of rotamers that need to be searched during the final selection process. Summarized results from the 70% cutoff and the Lovell rotamer library are listed below (Table 2). Full results for both 65 and 70% runs with both the Lovell and Dunbrack rotamer libraries are listed in the Supplementary Material (see Tables S2a and b). Side chains were positioned on the backbone coordinates of the original target structure. The results presented are divided into two categories. The first category was selected to estimate the effect of borrowing entire side-chain conformations on the final side-chain rmsd. The second category was chosen to estimate the reduction in the number of rotamer combinations that need to be searched during the final selection process. 

Table 1. The 14 HOMSTRAD families used for jack-knife test. The full HOMSTRAD family name, plus HOMSTRAD abbreviation. Target = HOMSTRAD structure modeled. Length = number of standard amino acids found in the target. Parent-target PID is listed after the template name

<table>
<thead>
<tr>
<th>Full homstrad family</th>
<th>Target</th>
<th>Length</th>
<th>Parents + PID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aldo/keto reductase</td>
<td>(aldosered)</td>
<td>1ah4</td>
<td>314</td>
</tr>
<tr>
<td>Aspartic proteinase</td>
<td>(asp)</td>
<td>3app</td>
<td>323</td>
</tr>
<tr>
<td>Beta-lactamase</td>
<td>(blm)</td>
<td>3blm</td>
<td>257</td>
</tr>
<tr>
<td>Epidermal growth factor-like domain</td>
<td>(egf)</td>
<td>hqcg</td>
<td>51</td>
</tr>
<tr>
<td>Glycosyl hydrolase family 22</td>
<td>(gh22)</td>
<td>liz2</td>
<td>129</td>
</tr>
<tr>
<td>Glutathione</td>
<td>(gluts)</td>
<td>lgta</td>
<td>218</td>
</tr>
<tr>
<td>S-transferase</td>
<td>GTP-binding protein</td>
<td>(gtp)</td>
<td>1a4ra</td>
</tr>
<tr>
<td>Immunoglobulin domain - V set</td>
<td>(igvar-h)</td>
<td>6fahb</td>
<td>126</td>
</tr>
<tr>
<td>Proflin</td>
<td>(proflin)</td>
<td>1aef</td>
<td>125</td>
</tr>
<tr>
<td>Short-chain dehydrogenases/reductases</td>
<td>(sdr)</td>
<td>2aei</td>
<td>252</td>
</tr>
<tr>
<td>Src homology</td>
<td>(sh3)</td>
<td>1shg</td>
<td>57</td>
</tr>
<tr>
<td>Triose phosphate isomerase</td>
<td>(tim)</td>
<td>1tima</td>
<td>247</td>
</tr>
<tr>
<td>Tumor necrosis factor</td>
<td>(TNF)</td>
<td>4tsva</td>
<td>147</td>
</tr>
<tr>
<td>Ubiquitin conjugating enzyme</td>
<td>(ucc)</td>
<td>1ubh</td>
<td>159</td>
</tr>
</tbody>
</table>

Table 2. Side-chain rmsd for Andante runs in various operating modes using the Penultimate Rotamer Library

<table>
<thead>
<tr>
<th>Family Andante mode</th>
<th>aldosered</th>
<th>asp</th>
<th>blm</th>
<th>egf</th>
<th>gft22</th>
<th>gluts</th>
<th>gtp</th>
<th>igvarh</th>
<th>proflin</th>
<th>sdr</th>
<th>sh3</th>
<th>tim</th>
<th>TNF</th>
<th>uce</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 No borrowing</td>
<td>1.98</td>
<td>1.42</td>
<td>1.72</td>
<td>1.91</td>
<td>1.80</td>
<td>2.15</td>
<td>1.93</td>
<td>2.62</td>
<td>1.84</td>
<td>1.73</td>
<td>1.74</td>
<td>2.10</td>
<td>2.12</td>
<td>1.89</td>
<td>1.92</td>
</tr>
<tr>
<td>2 +B</td>
<td>1.52</td>
<td>1.14</td>
<td>1.55</td>
<td>1.89</td>
<td>1.58</td>
<td>1.75</td>
<td>2.33</td>
<td>1.74</td>
<td>1.64</td>
<td>1.68</td>
<td>1.96</td>
<td>1.60</td>
<td>1.70</td>
<td>1.70</td>
<td>1.70</td>
</tr>
<tr>
<td>3 +B+R1</td>
<td>1.42</td>
<td>1.06</td>
<td>1.56</td>
<td>1.75</td>
<td>1.56</td>
<td>1.69</td>
<td>1.75</td>
<td>2.36</td>
<td>1.70</td>
<td>1.55</td>
<td>1.70</td>
<td>1.97</td>
<td>1.56</td>
<td>1.75</td>
<td>1.67</td>
</tr>
<tr>
<td>4 +B+R1+R12</td>
<td>1.37</td>
<td>1.07</td>
<td>1.56</td>
<td>1.75</td>
<td>1.56</td>
<td>1.62</td>
<td>1.79</td>
<td>2.35</td>
<td>1.69</td>
<td>1.53</td>
<td>1.70</td>
<td>1.97</td>
<td>1.59</td>
<td>1.76</td>
<td>1.66</td>
</tr>
<tr>
<td>5 +R1</td>
<td>1.58</td>
<td>1.07</td>
<td>1.56</td>
<td>1.76</td>
<td>1.63</td>
<td>1.84</td>
<td>1.85</td>
<td>2.59</td>
<td>1.68</td>
<td>1.57</td>
<td>1.67</td>
<td>2.01</td>
<td>1.74</td>
<td>1.80</td>
<td>1.74</td>
</tr>
<tr>
<td>6 +R1+R12</td>
<td>1.51</td>
<td>1.08</td>
<td>1.54</td>
<td>1.75</td>
<td>1.68</td>
<td>1.76</td>
<td>1.87</td>
<td>2.57</td>
<td>1.64</td>
<td>1.56</td>
<td>1.67</td>
<td>1.96</td>
<td>1.76</td>
<td>1.81</td>
<td>1.73</td>
</tr>
<tr>
<td>7 SCWRL3 (alone)</td>
<td>1.89</td>
<td>1.13</td>
<td>1.56</td>
<td>1.81</td>
<td>1.62</td>
<td>1.98</td>
<td>1.64</td>
<td>2.22</td>
<td>1.62</td>
<td>1.45</td>
<td>1.58</td>
<td>2.08</td>
<td>1.91</td>
<td>1.74</td>
<td>1.73</td>
</tr>
<tr>
<td>8 SCWRL3+B</td>
<td>1.46</td>
<td>0.96</td>
<td>1.53</td>
<td>1.77</td>
<td>1.60</td>
<td>1.57</td>
<td>1.53</td>
<td>2.27</td>
<td>1.71</td>
<td>1.36</td>
<td>1.61</td>
<td>1.98</td>
<td>1.64</td>
<td>1.65</td>
<td>1.62</td>
</tr>
<tr>
<td>9 TEMPLATE</td>
<td>0.80</td>
<td>0.75</td>
<td>1.02</td>
<td>0.75</td>
<td>0.75</td>
<td>0.92</td>
<td>0.86</td>
<td>1.87</td>
<td>1.06</td>
<td>0.93</td>
<td>0.93</td>
<td>1.31</td>
<td>0.93</td>
<td>0.66</td>
<td>0.97</td>
</tr>
</tbody>
</table>

+B, full borrowing at χ1 +1 and χ2 +1 performed. +R1 means restricting at χ1 was performed at all suitable positions (non Ala, Gly, borrowed or restricted at χ1 +1). +R12 means restricting was performed at all suitable residue positions (non Ala, Gly or borrowed). ‘TEMPLATE’ results are for all completely borrowed side-chain conformations selected by Andante-derived borrowing rules.
be searched as a result of using information derived from the template structures. Table 2 shows the effects of addition of homologous template information to the overall side-chain rmsd. For Andante, when only borrowing is enforced (Table 2, row 1 versus row 2) the overall side-chain rmsds produced by Andante are in all 14 cases reduced. Similar trends can be observed when any combination of borrowing or restricting using template information is applied. SCWRL3 results showed a reduction in overall side-chain rmsd in 11/14 cases (rows 7 and 8) when supplied with Andante borrowed side-chain conformations. Row 9 of Table 2 lists the side-chain rmsds of the borrowed positions (TEMPLATE) relative to the observed conformation in the target structure. In all cases, these are considerably lower than the overall side-chain rmsd. These values are further analysed in Tables S3a and S3b. Here, the rmsds of the borrowed conformations are compared to the equivalent rotameric solutions supplied by SCWRL3. For some cases, there is a considerable decrease in the rmsd of the TEMPLATE conformations to the observed conformations in both buried and exposed environments. For ald/keto reductase family target, 1a4h, the rmsd of the TEMPLATE side chains is reduced by 0.6 Å over 165 positions. A similar improvement is noted for the glutathione S-transferase family. Figure 1a shows the packing of a cluster of several large hydrophobic residues in 1a4h. This example shows that borrowing can lead to a more accurate result compared to Andante or SCWRL3 using different rotamer libraries. A similar example for the glutathione S-transferase family is shown in Figure 1b. Table S3a shows that for 11/14 test sets, the borrowed conformations found in buried positions showed a decreased rmsd compared to rotamer library solutions placed by SCWRL3 for the corresponding side-chain conformation observed in the actual target structure file. Also, 12/14 cases showed decreased rmsd for the equivalent exposed positions (Table S3b). This is encouraging as one would hope to see a reduction of the incorrect packing of pockets of buried, hydrophobic residues using template information. This improvement might come as a result of borrowing an entire set of correct conformations or the selection of one correct conformation that forces other neighbouring conformation into a more accurate orientation. However, an increase in the accuracy of prediction of exposed residues without adding hydrogen bonding or solvation terms to the test function is a bonus. In four cases, the epidermal growth factor-like domain (1hcgb), glutathione S-transferase (1gta), SH3 domain (1shg) and ubiquitin conjugating enzyme (1u9a) families, borrowing of template conformations was very low (<20% of total). This can be attributed to a combination of three factors. First, the rules for borrowing decrease with PID (Table S1). For these four cases, the majority of templates have PIDs in the range 20–39%. Second, as to be expected there are more borrowing rules for entire side-chain conformations from template residues in buried environments (Table S1). For epidermal growth factor-like domain, glutathione S-transferase, sh3 and ubiquitin conjugating enzyme the target structures have 88, 71, 82 and 75% residues that are exposed, respectively. This compares to the aspatic proteinase and ald/keto reductase cases, which have 65 and 59% exposed residues, respectively, but had borrowed conformation percentages of 52 and 32%, respectively. The final factor is the amino acid composition of the target structure. A high percentage of Ala, Gly, Arg and Lys will reduce the amount of borrowing possible. There are no borrowing rules for complete conformations of Arg and Lys and no rules for Gln, Glu and Met (1+2+3) at lower PIDs. The targets for epidermal growth factor-like domain (1hcgb), sh3 (1shg) and ubiquitin conjugating enzyme (1u9a) have 24, 27 and 28% composition of Gly, Ala, Lys and Arg, respectively. This compares to the asp (3app) and ald/keto reductase family (1a4h) cases that have GAKR compositions of 20 and 22%, respectively. While no major improvement in the overall side-chain rmsd is seen in these cases, the general trend for lower rmsd for the borrowed conformations is still observed (Table S3a and b). Table S5 explores the use of different templates within the family for modeling the target. The trend shown here is an expected one in that the majority of the borrowing comes from the highest PID template structure, although borrowing is still performed from more distant homologues underlining the importance of using multiple family members as templates. A final point regarding the results from the 12 runs using Andante in various borrowing/restricting modes is that in 8/14 cases (see Tables S2a and b) Andante was able to provide the lowest rmsd model if all
models produced and results using the Dunbrack rotamer library are considered. The other issue of interest is the extent to which borrowing and restricting influences the size of interacting clusters. Typically, DEE calculations are used to eliminate rotamers and reduce the total number of rotamer combinations that need to be searched. Considering that no DEE is implemented in this version, the reduction in the total number of combinations searched can be quite dramatic while still improving rmsd values. Table 3 lists the values of the total combinations, total backbone compatible combinations and the total number of combinations searched for four targets when various borrowing options are enforced (results for all families are given in Table S5). Compared to cases where no borrowing/restricting rules are applied, simply enabling the borrowing mode reduces the number of backbone compatible rotamer combinations by an average of ~10^{35} combinations overall families. The values designated 'S' in Table 3 are the total number of combinations searched for all identified clusters. In 13 of the 14 cases, the cluster combination totals are reduced to a trivial size that could be searched exhaustively. For the aldo/keto reductase and aspartic proteinase cases, massive reductions (10^96 and 10^55) in the number of combinations are observed. For three families; epidermal growth factor-like domain, profilin and tumor necrosis factor, the number of combinations is reduced to zero for some runs. Additionally, application of the restricting rules in concert with the borrowing rules further reduces the number of combinations searched (Table S5). If the restricting mode, without borrowing, is applied, similar reductions in the numbers of combinations searched are observed.

4 CONCLUSIONS

In the work presented here, we have derived environment-specific χ angle conversion probabilities for use in comparative modeling. These probabilities were then applied in the form of borrowing/restricting rules during the side-chain selection portion of the process. There are several advantages to implementing the borrowing and restricting rules when using the approach of defining interacting clusters and then determining low energy conformations for those clusters. The most striking is the large reduction in the number of combinations that need to be searched. The appropriate application of these rules is shown to select accurate conformations from homologous templates giving a decrease in rmsd at the borrowed positions relative equivalent rotameric solutions. Building the borrowed conformations also has a number of knock-on effects if applied to the graph theory-cluster solving approach used by SCWRL3 and other programs by means of reducing the number of combinations that need to be search to resolve the defined clusters. The borrowed/restricted positions prevent the growth of large clusters in three ways. First, borrowing entire side-chain conformations automatically reduces the total numbers of rotamer combinations possible. Second, borrowed positions disrupt the interaction graph; large numbers of rotamers from non-borrowed positions can be eliminated because they are incompatible with the extended backbone, thus restricting the growth of interacting clusters when the initial lowest energy, clashing residues are ‘spun-out’ to define interacting clusters. Third, restricting solutions to defined χ bins reduces the number of rotamers that are ‘spun out’, again reducing the number of new interactions that will be identified thus limiting the growth of such clusters. This will result in an increase in speed, as large numbers of smaller clusters can be resolved more quickly than a smaller number of larger clusters. The χ angle conservation probabilities may have some application in the area of protein redesign. Any sequence/rotamer combination evaluated on a fixed backbone will have a corresponding PID to the sequence of the protein that adopts the fold in question. Instead of randomly changing rotamers during the repacking phase, it may be possible to use information from the original template. One could attempt to apply the low range (0–39% PID) restricting rules at χ1 to limit potential rotamer solutions to similar χ1 bins as that observed in the original template. This may improve sampling by directing potential side-chain conformations into the same area of space occupied by the original template residue. Andante is part of the Orchestral suite of comparative modeling programs (Tripos Inc., USA).

<table>
<thead>
<tr>
<th>Mode</th>
<th>aldosed</th>
<th>asp</th>
<th>egf</th>
<th>profilin</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TC</td>
<td>BC</td>
<td>S</td>
<td>TC</td>
</tr>
<tr>
<td>No rules</td>
<td>219.5</td>
<td>194.3</td>
<td>36.1</td>
<td>178.5</td>
</tr>
<tr>
<td>+B</td>
<td>98.6</td>
<td>2.8</td>
<td></td>
<td>101.1</td>
</tr>
<tr>
<td>+B + R1</td>
<td>69.3</td>
<td>1.7</td>
<td></td>
<td>86.6</td>
</tr>
<tr>
<td>+B + R1 + R12</td>
<td>65.1</td>
<td>1.8</td>
<td></td>
<td>85.3</td>
</tr>
<tr>
<td>+R1</td>
<td>108.3</td>
<td>7.3</td>
<td></td>
<td>110.1</td>
</tr>
<tr>
<td>+R1 + R12</td>
<td>87.2</td>
<td>5.0</td>
<td></td>
<td>101.7</td>
</tr>
</tbody>
</table>

No rule denotes no borrowing or restricting applied. +B, full borrowing at χ1, χ1 + 2 and χ1 + 2 + 3 performed. +R1 means restricting at χ1 was performed at all suitable positions (non Ala, Gly, borrowed or restricted at χ1 + 2). +R12 means restricting was performed at all suitable residue positions (non Ala, Gly or borrowed). TC denotes total rotamer combinations for all residues in the target protein. BC denotes total number of rotamers combinations compatible with the backbone. S denotes number of combinations searched. All values are in Log10. Full names for HOMSTRAD family names are list in Table 2.
ACKNOWLEDGEMENTS

Many thanks to Matthias Keil of Tripos Inc., for technical assistance and to Tripos Inc., for funding.

Conflict of Interest: none declared.

REFERENCES


