Refinement of protein cores and protein–peptide interfaces using a potential scaling approach

Ralph Nico Riemann1,2 and Martin Zacharias3,4

1International University Bremen, School of Engineering and Science, D-28759 Bremen, Germany and 2Institut für Molekulare Biotechnologie, Beutenbergstrasse 11, D-07745 Jena, Germany

Abstract

Refinement of side chain conformations in protein model structures and at the interface of predicted protein–protein or protein–peptide complexes is an important step during protein structural modelling and docking. A common approach for side chain prediction is to assume a rigid protein main chain for both docking partners and search for an optimal set of side chain rotamers to optimize the steric fit. However, depending on the target–template similarity in the case of comparative protein modelling and on the accuracy of an initially docked complex, the main chain template structure is only an approximation of a realistic target main chain. An inaccurate rigid main chain conformation can in turn interfere with the prediction of side chain conformations. In the present study, a potential scaling approach (PS-MD) during a molecular dynamics (MD) simulation that also allows the inclusion of explicit solvent has been used to predict side chain conformations on semi-flexible protein main chains. The PS-MD method converges much faster to realistic protein–peptide interface structures or protein core structures than standard MD simulations. Depending on the accuracy of the protein main chain, it also gives significantly better results compared with the standard rotamer search method.

Keywords: molecular dynamics simulation/peptide–protein docking/potential smoothing/protein design/side chain prediction

Introduction

The prediction of amino acid side chain conformations is an important step in comparative protein modelling, protein design and for generating atomic coordinates from low-resolution protein folding simulations. Most of the current side chain prediction methods are based on elegant and rapid searches of side chain rotamer libraries to identify an optimal set of side chain conformations for a given rigid protein main chain structure (Summers and Karplus, 1989; Holm and Sander, 1991; Lee and Subbiah, 1991; Dunbrack and Karplus, 1993; Koehl and Delarue, 1994; Dahiyat and Mayo, 1997; Tuffery et al., 1997; Samudrala and Moult, 1998; Swain and Kemp, 2001; Xiang and Honig, 2001; reviewed by Dunbrack, 2002; Canutescu et al., 2003; Källblad and Dean, 2003). However, in the case of comparative protein modelling, depending on the target–template similarity, the main chain structure of the template is only an approximation of a realistic target main chain. In the case of protein design one is interested in predicting the structure of a protein mutation before the synthesis. X-ray crystallography studies indicate that the protein main chain conformation can change in response to point mutations (Arcus et al., 1995; Wright and Lim, 2001). An inaccurate rigid main chain conformation can in turn interfere with the prediction of side chain conformations (Tuffery et al., 1997). Another limitation of current side chain prediction methods is due to the application of relatively simple energy functions that in many cases consider only the steric packing of generated side chain conformations and completely neglect the influence of surrounding solvent molecules (Canutescu et al., 2003).

The situation is similar in the case of predicted protein–protein or protein–peptide complexes. A typical protein–protein docking simulation to systematically search for putative interaction geometries results in a set of possible complexes that can still deviate from a realistic structure owing to approximations used during the initial docking approach (e.g. rigid protein partners). The refinement, in particular of the interface between the binding partners, is an important step in order to achieve an accurate complex structure (Najmanovich et al., 2000; Smith and Sternberg, 2002). As for the prediction of protein core structures, a straightforward and rapid approach for interface refinement is to assume that the protein main chain structure at the interface is approximatively correct and to predict side chain conformations based on the rotamer search and optimal steric fit (Gray et al., 2003). The success of this approach may significantly depend on the accuracy of the main chain structure at the interface close to the native structure, after the initial docking attempt. In addition, water molecules, not present during standard side chain rotamer searches, might be of importance for the protein–protein interface structure.

In principle, molecular dynamics (MD) simulations allow for simultaneous adjustment of both side chain and backbone atoms and can be performed in the presence of water molecules. However, repacking of side chain conformations during an MD simulation starting, for example, from a random initial side chain placement involves crossing of many high-energy barriers. The simulation can easily be trapped in unfavourable states that are separated from low-energy states by large energy barriers. A standard method to enhance the convergence of MD simulations is to use the simulated annealing (SA) approach, i.e. to initially use a high simulation temperature to allow for barrier crossing followed by cooling of the simulation to identify favourable low-energy states (Brünger et al., 1997). The high initial temperatures used in SA approaches may, however, interfere with the presence of explicit water molecules during the simulation and can also distort the overall protein fold. In the present study, we have developed and tested a potential scaling method that allows the specific lowering of energy barriers for residues located in the protein core or at the interface of protein–peptide complexes during an MD simulation with a minimum perturbation of other residues and surrounding structures.
water molecules. The smooth rescaling of the potential during MD simulations allows a rapid convergence to an optimal interface structure with simultaneous adjustment of both side chain and backbone protein core or interface structure. The approach was tested on three test proteins and two protein–peptide complexes with known structures, including a systematic evaluation and comparison with standard MD simulations and a rotamer search based method. In general, the potential scaling (PS-MD) approach was found to be much more efficient than standard MD simulations to obtain realistic side chain conformations from arbitrary initial placements. Depending on the accuracy of the protein main chain, the approach also performed better than search strategies based on rotamer libraries.

**Materials and methods**

The performance of the potential scaling approach for the prediction of protein core side chains was systematically evaluated on three test structures that represent different protein folds. One structure consisted of four \( \beta \)-strands enclosed between two \( \alpha \)-helices in a sandwich conformation (SH2 domain; PDB entry: 1CWE; Mikol et al., 1995). The second test case was an all-\( \beta \)-sheet protein (SH3 domain; PDB entry: 1GFC; Kohda et al., 1994). This structure corresponds to an energy-minimized average structure obtained by NMR spectroscopy (the other two test cases have been determined by X-ray crystallography). The third test protein was an all-\( \alpha \)-helical protein (DNA-binding fragment of the cro repressor protein of *Escherichia coli* phage lambda; PDB entry: 1R69; Mondragon et al., 1989). The three cases (test set) were used to systematically optimize the generation of start structures and to test simulation parameters. In addition, the potential scaling method was further evaluated on a set of 12 proteins with between 56 and 124 amino acid residues (evaluation set), for which high-resolution X-ray structures were available for the comparison of predicted and experimentally observed side chain conformations (the PDB code of the protein structures is given in Table I). To test the potential scaling method for protein–protein and protein–peptide interface refinement two protein–peptide complexes were used as model test systems. One complex consisted of the protein MDM2 from *Xenopus laevis* in complex with an \( \alpha \)-helical trans-activation domain (residues 17–27) of the protein p53 (PDB entry: 1YCQ; Kussie et al., 1996). The second complex corresponds to the protein binding domain (residues 218–277) taken of the chaperone protein GroEL and a peptide consisting of 11 residues (residues 602–612; mPDB entry: 1DKD; Chen and Sigler, 1999).

**Potential scaling of side chain interactions**

All energy minimizations (EMs) and MD simulations were performed using a modified version of the sander module of the Amber6 package (Case et al., 1999) with the Cornell et al. (1995) force field parameter file (parm94). During this potential scaling MD run, all pairwise Lennard–Jones and electrostatic interactions of selected side chain atoms beyond \( C_{\beta} \) were scaled. A soft-core scaling scheme of the following form was employed (Zacharias et al., 1994) and implemented in the sander code:

\[
V(r_{ij}) = (1 - \lambda) \left( e_{ij} \left( \frac{r_{ij}^6}{C_{16}/C_{17}} \right) - \left( \frac{r_{ij}^2}{r_{ij}^2 + \lambda \cdot \delta} \right)^6 + \frac{q_{ij} (4\pi \varepsilon_0)}{r_{ij}^2 + \lambda \cdot \delta} \right)^{1/2}
\]

Here, \( \lambda \) corresponds to a scaling parameter that can vary between 0 (original interaction energy) and 1 (no interaction), whilst \( \delta \) is the shift parameter and \( e_{ij} \) is the minimum Lennard–Jones energy for the interaction between atoms \( i \) and \( j \), respectively. This functional form allows a smooth transition between zero and full potential and preserves the position of the energy minimum of the Lennard–Jones term (see Figure 1). For \( \delta \) several values from 10 to 30 \( \text{Å}^2 \) were tested and, owing to its performance, \( \delta = 20 \text{Å}^2 \) was finally chosen for the simulations using the ddd model, whereas \( \delta = 15 \text{Å}^2 \) was used for simulations in the presence of explicit water molecules. During the potential scaling run the scaling parameter \( \lambda \) was varied linearly in 21 steps.

Since the PS-MD approach is expected to especially improve the sampling of buried and sterically constrained side chains, it was applied only to buried side chains and side chains located at the interface of protein–peptide complexes (85% buried solvent-accessible surface area as judged by

<table>
<thead>
<tr>
<th>Table I. Evaluation of the PS-MD approach on a set of proteins with known (X-ray) structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-ray structure</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>1PDO</td>
</tr>
<tr>
<td>1WHO</td>
</tr>
<tr>
<td>1MSI</td>
</tr>
<tr>
<td>1H5</td>
</tr>
<tr>
<td>1CY5</td>
</tr>
<tr>
<td>1FR</td>
</tr>
<tr>
<td>1TUL</td>
</tr>
<tr>
<td>1AWD</td>
</tr>
<tr>
<td>1PGX</td>
</tr>
</tbody>
</table>

The second column indicates the side chain RMSD (SC-RMSD) of predicted buried side chains (<85% solvent accessible) from the experiment using SCWRL3.0 (Canutescu et al., 2003) built onto the experimental main chain structure. The third column indicates the range of main chain deviations: \( (C_{\beta} \text{-RMSD}) \) of a set of generated eight start structures for each protein. Column 4 gives the buried side chain RMSD range (average in parentheses) using SCWRL3.0 on the eight start structures and column 5 indicates the results of using PS-MD on the same set of start structures.

466
were subsequently energy minimized (2000 steps) with the full experimental structure for the rest of the protein. The structures of 5.0 kcal mol\(^{-1}\)/C0 were used for the evaluation set of proteins (12 protein structures, see Table I) except that in this case only eight start structures per protein were used.

In the case of the peptide–protein complexes the peptide side chains were first randomized using the MD simulations, with \(\lambda\) values close to 1 for the peptide side chains (see above). In addition, ligand peptides were randomly misplaced by up to 1.3 Å (RMSD of the \(C_{\alpha}\) atoms) from their placement in the experimental structure to mimic an imperfect docking result. The randomized initial structures were energy minimized using the full potential and by applying window restraints (allowing free motion within \(\pm 2\) Å from the start structure). This step also leads to a relaxed protein main chain in part compatible with the initial side chain placements. This step was necessary to compare the PS-MD method with standard and ‘temperature SA’ MD simulations using the same start structures and the full potential during the entire simulation.

Simulations with a distance-dependent dielectric constant
Simulations in the absence of solvent were performed using a distance-dependent dielectric constant of \(\varepsilon = 4r\) (ddd model), a 12.0 Å cut-off and a time step of 1 fs. Bonds involving hydrogen were restrained by applying the RATTLE (Anderson, 1983) algorithm. In addition to regular MD simulations (at 300 K with a time constant of 2.0 ps) and PS-MD simulations, the performance of SA (Kirkpatrick et al., 1983; Brunger et al., 1997) simulations was also tested starting at a simulation temperature of 1000 K cooled down to 300 K (within 11 steps) for the same simulation time as used for the PS-MD and regular MD runs. The start structures also served as restraining reference structures during the PS-MD, SA-MD and standard MD simulations. The total simulation length was 252 ps. All final structures were energy minimized using the full energy function.

Explicit solvent simulations
In order to systematically test the potential scaling scheme for protein–peptide interface refinement in the presence of explicit solvent, the solvated part of the protein–peptide complexes was restricted to a water cap around the binding site. In the case of the MDM2–p53 complex, a water cap with radius 18.5 Å and 421 TIP3 water molecules (Jorgensen et al., 1983) centred at the \(C_{\alpha}\) atom of Leu22 was added. Beyond the radius of 18.5 Å a harmonic potential with a force constant of 1.5 kcal mol\(^{-1}\)/Å\(^2\) was applied to prevent solvent molecules from escaping out of the water cap.

Restraint MD simulation and generation of start structures
In order to allow for conformational relaxation of the protein main chain during potential scaling and standard MD simulations, positional window restraints were applied that allow for free motion within a preset range with respect to the start structure and application of a soft quadratic penalty (force constant of 5.0 kcal mol\(^{-1}\)/Å\(^2\)) for deviations beyond the preset range. In general, all side chains involved in the potential scaling were completely free to move throughout all stages of the simulations.

Start structures (24) for each test case were generated using high-temperature MD simulations (900 K; 100 ps), with \(\lambda\) values close to 1 for the selected fully mobile side chains (means with reduced interaction of the selected side chains) and positional window restraints (\(\pm 1\) Å) with respect to the experimental structure for the rest of the protein. The structures were subsequently energy minimized (2000 steps) with the full force field followed by a short MD simulation (300 K, 3 ps) and another EM (2000 steps; both with the same positional window restraints as above). Generated structures with high steric strain and structures with an average root mean square deviation (RMSD) of <2.0 Å for buried side chains with respect to the experiment were not considered. This protocol leads to start structures with randomly misplaced buried side chains that are sterically possible and average main chain deviations from the experiment of 0.4–0.8 Å (RMSD of the \(C_{\alpha}\) atoms). The same conditions were used to generate start structures for the evaluation set of proteins (12 protein structures, see Table I) except that in this case only eight start structures per protein were used.

The solvent-accessible surface of the side chain in the experimental reference structure). The respective residue numbers for the set of test structures are as follows. (i) In the case of 1R69 (10 residues): 6, 20, 24, 31, 34, 35, 44, 48, 52 and 59; (ii) in the case of 1CWE (18 residues): 19, 29, 30, 31, 32, 42, 43, 44, 45, 56, 61, 71, 77, 80, 83, 85, 87 and 98; and (iii) in the case of 1GFC (10 residues): 5, 11, 18, 19, 21, 27, 29, 38, 49 and 54. The same selection criteria for the definition of buried residues were applied by applying the RATTLE (Anderson, 1983) algorithm. In addition to regular MD simulations (at 300 K with a time constant of 2.0 ps) and PS-MD simulations, the performance of SA (Kirkpatrick et al., 1983; Brunger et al., 1997) simulations was also tested starting at a simulation temperature of 1000 K cooled down to 300 K (within 11 steps) for the same simulation time as used for the PS-MD and regular MD runs. The start structures also served as restraining reference structures during the PS-MD, SA-MD and standard MD simulations. The total simulation length was 252 ps. All final structures were energy minimized using the full energy function.

Explicit solvent simulations
In order to systematically test the potential scaling scheme for protein–peptide interface refinement in the presence of explicit solvent, the solvated part of the protein–peptide complexes was restricted to a water cap around the binding site. In the case of the MDM2–p53 complex, a water cap with radius 18.5 Å and 421 TIP3 water molecules (Jorgensen et al., 1983) centred at the \(C_{\alpha}\) atom of Leu22 was added. Beyond the radius of 18.5 Å a harmonic potential with a force constant of 1.5 kcal mol\(^{-1}\)/Å\(^2\) was applied to prevent solvent molecules from escaping out of the water cap. The cap completely embedded the peptide and a large part of the receptor protein. A similar set-up was used in the case of the GroEL fragment and the bound peptide, with a solvent cap radius of 19.5 Å (531 TIP3 water molecules) centred at the \(C_{\alpha}\) atom of Thr605 of the ligand peptide. All receptor and peptide atoms were mobile during the simulations; however, the motion of all non-scaled atoms was restricted to
$\pm 2$ Å of the corresponding atom position in the start structure (window restraints). All simulations with explicit solvent were performed using a dielectric constant of $\varepsilon = 1.0$. An 11.0 Å cut-off distance for non-bonded interactions was also used and the system was coupled to a heat bath of 300 K. The time step was 1 fs for all MD runs and bonds involving hydrogen have been restrained by applying the RATTLE algorithm (Andersen, 1983).

**Comparison with a side chain rotamer search method and experiment**

For comparison the systematic side chain rotamer search method SCWRL3.0 (Canutescu et al., 2003) based on the dead-end elimination theorem, a rigid backbone and a simple steric packing energy function was applied on the same set of start structures. The SCWRL3.0 approach is a deterministic side chain prediction method. Therefore, differences in the final results are solely due to the differences in protein main chains of the various start structures. To allow direct comparison with the PS-MD results the same set of buried or interface side chains as used during PS-MD was rebuilt [nearly experimental conformations for the other side chains employing the $\rightarrow$ (side chain fixing) option of SCWRL3.0].

**Comparison with experimental structures**

The PS-MD method for predicting the conformation of buried or interface side chain allows for conformational changes of the protein backbone within preset boundaries (typically free backbone movements of up to $\pm 1$ Å were allowed). However, this complicates the comparison of predicted side chain placements with experimental reference structures in terms of a RMSD of buried side chains from the experiment, since it is not only influenced by the side chain but also by the backbone conformation. In order to allow a direct comparison of the side chain placements with experimental reference structures, the final structures of the PS-MD and MD simulations as well as the SCWRL3.0 results were energy minimized using the experimental backbone structures as restraining reference structures. This step basically eliminates the contribution of the protein main chain. At the same time the side chains stay nearly in the same orientation/conformation. Independently, final structures were also compared with the experimental structures in terms of internal coordinates (side chain dihedral torsion angles).

**Smoothing of RMSD distribution curves**

To allow better comparison, RMSD distribution curves were smoothed using a step-wise error function $F_i$(RMSD) representing a single RMSD result and the graph $G$(RMSD) is then the normalized sum over all $N$ final structures, with

$$F_i$(RMSD) = $\sum_{n=1}^{N} \delta$(RMSD $- R_i - n \cdot \Delta$) $- \delta$(RMSD $- R_i + n \cdot \Delta$)

and

$$G$(RMSD) = $\frac{1}{N} \sum_{i=1}^{N} F_i$(RMSD)

where $R_i$ is the RMSD value of the $i$-th structure, $N$ the normalizing pre-factor and $\Delta$ is the step width ($=0.04$ Å). The function $\delta$($r - r_0$) equals zero if the distance $r$ is smaller than $r_0$ and 1 if it is greater.

**Results**

**Core side chain prediction with a semi-flexible main chain**

In a first set of simulations the PS-MD approach was used to predict the core side chain conformations of three test proteins. These test structures represent different protein folds. One structure (PDB entry: 1CWE; Mikol et al., 1995), termed 1CWE in the following, consists of four $\beta$-strands enclosed between two $\alpha$-helices in a sandwich conformation ($\alpha$/$\beta$-protein with a SH2 domain fold). The second test case, termed 1GFC, was an all $\beta$-sheet protein (NMR structure of an SH3 domain; PDB entry: 1GFC; Kohda et al., 1994) and the third test structure (PDB entry: 1R69; Mondragon et al., 1989), termed 1R69, corresponds to an all $\alpha$-helical protein (DNA-binding fragment of the cro repressor protein of _E.coli_ phage lambda, see Figure 2). The approach was compared with standard MD simulations, (temperature) simulated annealing (T-SA-MD) runs and a systematic rotamer search method based on the dead-end elimination theorem and using a fixed protein main chain (SCWRL3.0). In order to test the sensitivity of the simulation results on the side chain start placement and small deviations of the protein main chain in each case, 24 start structures with random but sterically allowed initial placement of buried side chains were generated (see Materials and methods). An overview of the simulation steps of the PS-MD method is given in Figure 3. The protein main chain of these start structures deviated from the corresponding experimental main chain structure on average by 0.3–0.8 Å but also included larger local deviations at positions of initially misplaced side chains (of 1–1.5 Å). To allow conformational relaxation of the complete protein structures during the side chain building process, the protein main chain and solvent-exposed residues were allowed to move from the respective start positions by up to $\pm 1.0$ Å (positional window constraints). Larger displacements were penalized by a soft harmonic potential (force constant 5.0 kcal mol$^{-1}$ Å$^{-2}$). Initial tests indicated that the PS-MD approach is not useful for the structure prediction of solvent-exposed side chains. For example, in the case of the 1R69 test protein a set of 24 random start structures with all side chains included in the PS-MD approach was generated and showed an average side chain RMSD of $\sim$2.1 Å from the experiment. Application of the PS-MD method resulted in final structures with only slightly improved solvent-exposed side chain conformations and an average RMSD of $\sim$1.7 Å from the experiment (average over all side chains with >50% solvent accessibility). In contrast, the buried side chains (<15% solvent accessible) were on average predicted to better than 0.8 Å (Figure 4, 1r69). Also, in the absence of explicit solvent and employing a distance-dependent dielectric constant charged solvent-exposed residues are promoted to move towards the protein interior that interfere with the placement of buried non-polar residues. Therefore, the above positional window restraints were applied to both atom positions of the main chain as well as solvent-exposed residues. The purpose was to keep these residues in an extended solvent-exposed conformation. All buried side chains subject to the potential scaling were, however, completely mobile. A selected start structure and a final structure superimposed on the corresponding experimental structures are illustrated in Figure 2 (1R69 case). Except for the second test case the PS-MD method resulted in buried side chain placements closer to the experiment compared with standard MD simulations.
SCWRL3.0 rotamer search method applied on the same set of start structures performed the same as the PS-MD method for two test cases (1GFC and 1CWE; Figure 4B), but worse in the case of the third α-helical test case (1R69).

Each of the various start structures resulted in final structures of different force field energy. In order to check if the final energy (obtained after energy minimizing the final structure, including the positional window restraints to the start structure; see Materials and methods) could be used to select realistic predictions, the 12 lowest energy structures (out of 24) from the PS-MD and the SCWRL3.0 predictions were compared (Figure 5A). This subset has an on average lower side chain RMSD with respect to the experiment and further improves the prediction results (Figure 5A). However, it does so for both PS-MD and SCWRL3.0 approaches and results in the same trends, as seen in the comparison of all final structures (Figure 5A). To find out if the PS-MD method can improve the SCWRL3.0 predictions, the latter predicted structures were used as start structures for the PS-MD approach. A slight improvement was observed for test cases 1 (1GFC) and 2 (1CWE) and a significant improvement in the case of test structure 3 (1R69). The RMSD distributions appear very similar to those obtained starting from the original start structures (compare Figures 4 and 5B). Hence, starting from the SCWRL3.0 predictions does not further improve the results and indicates that protein main chain deviations of each start structure may limit the quality of the prediction. It is important to note that the start structures used for the above systematic test showed overall close agreement to the protein main chain of the experimental structure. However, the generation of start structures with initially misplaced side chains can lead to protein main chain deviations that are larger than the average main chain deviation. It is likely that the significant local deviation of the main chain is one reason for the relatively poor performance of both the SCWRL3.0 and the PS-MD methods for some start structures.

Core side chain prediction on an evaluation set of proteins

For a more reliable validation of the present PS-MD method, 12 additional high-resolution X-ray protein structures (>2.0 Å resolution) were chosen from the PDB consisting of 56–124 amino acids (evaluation set, see Table I). For comparison the SCWRL3.0 side chain prediction method was applied.
Fig. 3. Schematic overview of the PS-MD simulation approach.

Fig. 4. RMSD distributions of predicted buried side chain core residues (see Materials and methods) with respect to the corresponding experimental structure. (A) The buried side chain RMSD of start structures (dashed curve), final structures after standard MD (thin line) and after applying PS-MD (solid line) are shown. The panels show the results for start structures with a main chain deviation between 0.4 and 0.8 Å from the experiment. All RMSD values for the buried side chains have been obtained after performing energy minimizations, including harmonic restraints with respect to the corresponding experimental main chain structure (see Materials and methods). In (B) the dashed line indicates the results of rotamer-based SCWRL3.0 predictions (main chain of the corresponding start structure). Results of ‘temperature simulated annealing’-MD simulations are indicated using a thin line, whereas the distribution obtained with the PS-MD method are indicated using solid lines.

to each target using the experimental main chain structure. As indicated in the second column of Table I, in this case SCWRL3.0 gives excellent predictions, with RMSDs of buried side chains of <1 Å (in most cases). However, this represents an ideal case and in practical applications of side chain prediction one often relies on main chain template structures that deviate from the real main chain structure (e.g. homology modelling or prediction of the structure of mutations based on the wild-type...
In order to generate template structures with small but significant main chain deviations from the experimental reference in each case, eight start structures were generated with random initial placement of buried side chains and protein main chain deviations of 0.5–0.8 Å from the experiment (see start structure generation in Materials and methods and Table I). The deterministic side chain prediction using SCWRL3.0 and these slightly ‘incorrect’ main chains as (rigid) templates performed quite well, but significantly worse than when using the experimental main chain structure as the template (compare columns 4 and 2 in Table I). The PS-MD approach starting from the same start structures and using the same conditions as described in the previous benchmark (and the same procedure to calculate the deviation from the experiment, see above) resulted in overall more realistic side chain placement in 10 out of 12 cases compared with SCWRL3.0 (compare columns 5 and 4 in Table I).

**Comparative protein modelling using PS-MD**

Comparison of homologous protein structures in the range of 30–50% sequence identity in most cases translates to a structural deviation of the main chain in the range of 0.5–2 Å (Martí-Renom *et al*., 2000). In the case of a relatively even distribution of sequence variation (Figure 6), it is expected that the main chain deviation between homologous protein structures may also be evenly distributed along the protein chain (except for long sequence insertions or deletions). The PS-MD approach was applied to predict the side chain conformations of buried residues in 1R69 based on the protein main chain of the two homologous proteins 2CRO and 1ADR (both bacterial DNA-binding domains). The sequence identity of the two proteins with respect to 1R69 is 57% (2CRO) and 30% (1ADR), respectively. The overall main chain RMSD of these two protein domains with respect to the known structure of 1R69 was 0.8 Å (2CRO) and 1.5 Å (1ADR), respectively. These structures have also been used previously to benchmark protein side chain predictions (Chung and Subbiah, 1995; Desjarlais and Handel, 1999; Schueler-Furman and Baker, 2003). In the case of the 1GFC structure, a protein main chain of the homologous protein 1QWE (viral C-Src SH3 domain) was used for side chain prediction. The sequence identity of the two latter proteins is 34% (main chain RMSD ~1.4 Å).
The sequence of each template was initially replaced by the corresponding target sequence and the alignment given in Figure 6. Loop segments were built using the SPDBV program (GueX and Peitsch, 1997). The PS-MD approach was started in each case from 24 start structures (12 in the case of 1R69 and 1GFC) with random initial placement of the buried side chains (see Materials and methods and the previous paragraph). In contrast to the previous benchmark, in the case of comparative modelling the known template structure (2CRO or 1ADR in the case of modelling 1R69 or 1QWE in the case of modelling 1GFC) was used as the restraining reference structure (with window restraints, see Materials and methods). For comparison, the SCWRL3.0 method was applied (using the template structure main chains, respectively, as templates). Note that since the SCWRL3.0 approach is a deterministic side chain prediction method it results in only one final structure. All final structures were compared with the experimental 1R69 and 1GFC structures, respectively, by evaluation of the deviation of the side chain torsion angles $\chi_1$ and $\chi_2$ (from those in the known target structure; see Tables II and III). In addition, to separate main chain and side chain RMSD the final structures were shortly energy minimized using the respective main chain (1R69 with window restraints, 1GFC with harmonic restraints because of the high backbone deviation) as a restraining reference (this eliminates the main chain RMSD of the final structures with very little effect on the side chain placement). The same procedure was applied to the predictions using SCWRL3.0. For all cases the buried side chain RMSD of final structures obtained after applying the PS-MD approach was significantly lower than in the case of the SCWRL3.0 prediction (Figure 7). In addition, the predicted $\chi_1$ and $\chi_2$ side chain dihedral angles using PS-MD were overall in closer agreement with the known target structures compared with the SCWRL3.0 results (Tables II and III).

Table II. Side chain prediction for target sequence 1R69 modelled on the main chains of template 2CRO and 1ADR

<table>
<thead>
<tr>
<th>Residue</th>
<th>Val6</th>
<th>Leu20</th>
<th>Val24</th>
<th>Ile31</th>
<th>Leu34</th>
<th>Gln35</th>
<th>Phe44</th>
<th>Leu48</th>
<th>Leu52</th>
<th>Leu59</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRO, SCWRL ($\chi_1$)</td>
<td>7.2$^\circ$</td>
<td>6.0$^\circ$</td>
<td>3.9$^\circ$</td>
<td>12.7$^\circ$</td>
<td>135.3$^\circ$</td>
<td>2.6$^\circ$</td>
<td>118.1$^\circ$</td>
<td>1.3$^\circ$</td>
<td>4.8$^\circ$</td>
<td>95.0$^\circ$</td>
</tr>
<tr>
<td>CRO, PS ($\chi_2 &lt; 30^\circ$)</td>
<td>3/8</td>
<td>8/8</td>
<td>3/8</td>
<td>6/8</td>
<td>8/8</td>
<td>8/8</td>
<td>8/8</td>
<td>7/8</td>
<td>7/8</td>
<td>5/8</td>
</tr>
<tr>
<td>ADR, SCWRL ($\chi_1$)</td>
<td>6.3$^\circ$</td>
<td>6.0$^\circ$</td>
<td>0.7$^\circ$</td>
<td>9.5$^\circ$</td>
<td>20.9$^\circ$</td>
<td>3.5$^\circ$</td>
<td>129.8$^\circ$</td>
<td>108.4$^\circ$</td>
<td>5.3$^\circ$</td>
<td>96.7$^\circ$</td>
</tr>
<tr>
<td>ADR, PS ($\chi_2 &lt; 30^\circ$)</td>
<td>8/8</td>
<td>8/8</td>
<td>5/8</td>
<td>3/8</td>
<td>6/8</td>
<td>8/8</td>
<td>8/8</td>
<td>7/8</td>
<td>7/8</td>
<td>7/8</td>
</tr>
<tr>
<td>CRO, SCWRL ($\chi_2$)</td>
<td>2.7$^\circ$</td>
<td>1.0$^\circ$</td>
<td>1.0$^\circ$</td>
<td>1.0$^\circ$</td>
<td>1.0$^\circ$</td>
<td>1.0$^\circ$</td>
<td>1.0$^\circ$</td>
<td>1.0$^\circ$</td>
<td>1.0$^\circ$</td>
<td>1.0$^\circ$</td>
</tr>
<tr>
<td>CRO, PS ($\chi_2 &lt; 30^\circ$)</td>
<td>8/8</td>
<td>8/8</td>
<td>5/8</td>
<td>3/8</td>
<td>6/8</td>
<td>8/8</td>
<td>8/8</td>
<td>7/8</td>
<td>7/8</td>
<td>7/8</td>
</tr>
<tr>
<td>ADR, SCWRL ($\chi_2$)</td>
<td>2.7$^\circ$</td>
<td>1.0$^\circ$</td>
<td>1.0$^\circ$</td>
<td>1.0$^\circ$</td>
<td>1.0$^\circ$</td>
<td>1.0$^\circ$</td>
<td>1.0$^\circ$</td>
<td>1.0$^\circ$</td>
<td>1.0$^\circ$</td>
<td>1.0$^\circ$</td>
</tr>
<tr>
<td>ADR, PS ($\chi_2 &lt; 30^\circ$)</td>
<td>8/8</td>
<td>8/8</td>
<td>5/8</td>
<td>3/8</td>
<td>6/8</td>
<td>8/8</td>
<td>8/8</td>
<td>7/8</td>
<td>7/8</td>
<td>7/8</td>
</tr>
</tbody>
</table>

The first column indicates the template (1CRO or 1ADR) and the method (rotamer search program SCWRL3.0 or PS-MD) used for comparative modelling and buried side chain predictions. In the case of SCWRL predictions the deviation of the predicted versus experimental $\chi_1$ and $\chi_2$ dihedral angles is given, whereas in the case of the PS-MD method the fraction of correctly predicted $\chi_1$/$\chi_2$ dihedral angles (within 30$^\circ$ of experiment) of the eight energetically most favourable predicted structures (out of 24) is given. The conformation of all surface-exposed (non-buried) residues was the same in all cases.

Table III. Side chain prediction for the target sequence 1GFC modelled on the main chains of the template 1QWE

<table>
<thead>
<tr>
<th>Residue</th>
<th>Val5</th>
<th>Phe11</th>
<th>Glu18</th>
<th>Leu19</th>
<th>Phe21</th>
<th>Ile27</th>
<th>Val29</th>
<th>Trp38</th>
<th>Phe49</th>
<th>Val54</th>
</tr>
</thead>
<tbody>
<tr>
<td>QWE, SCWRL ($\chi_1$)</td>
<td>5.7$^\circ$</td>
<td>7.9$^\circ$</td>
<td>159.9$^\circ$</td>
<td>40.6$^\circ$</td>
<td>114.7$^\circ$</td>
<td>69.1$^\circ$</td>
<td>119.1$^\circ$</td>
<td>134.2$^\circ$</td>
<td>115.2$^\circ$</td>
<td>3.2$^\circ$</td>
</tr>
<tr>
<td>QWE, PS ($\chi_2 &lt; 30^\circ$)</td>
<td>3/4</td>
<td>3/4</td>
<td>0/4</td>
<td>1/4</td>
<td>2/4</td>
<td>0/4</td>
<td>2/4</td>
<td>4/4</td>
<td>4/4</td>
<td>0/4</td>
</tr>
<tr>
<td>QWE, SCWRL ($\chi_2$)</td>
<td>11.0$^\circ$</td>
<td>10.6$^\circ$</td>
<td>9.4$^\circ$</td>
<td>4.4$^\circ$</td>
<td>93.6$^\circ$</td>
<td>4.4$^\circ$</td>
<td>93.6$^\circ$</td>
<td>64.3$^\circ$</td>
<td>26.3$^\circ$</td>
<td>–</td>
</tr>
<tr>
<td>QWE, PS ($\chi_2 &lt; 30^\circ$)</td>
<td>4/4</td>
<td>3/4</td>
<td>2/4</td>
<td>4/4</td>
<td>2/4</td>
<td>–</td>
<td>4/4</td>
<td>2/4</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

See the footnote to Table II. In the case of the PS-MD method the fraction of correctly predicted $\chi_1$/$\chi_2$ dihedral angles (within 30$^\circ$ of experiment) of the four energetically most favourable predicted structures (out of 12) is given.

Refinement of predicted protein–peptide interfaces

The refinement of side chain placements at protein–protein and protein–peptide interfaces is an important step during docking simulations or refinement of low-resolution structures and complexes obtained by homology modelling. Such protein–protein interfaces are at least in part close to the solvent and it is expected that the inclusion of explicit solvent molecules and readjustment of the protein main chain are critical for a realistic interface side chain prediction. The performance of the side chain PS-MD method for interface refinement was tested on two protein–peptide complexes with known 3D structures (PDB-entries: 1YCQ and 1DKD; Figures 8 and 9). One case consisted of the protein MDM2 from X.laevis in complex with an $\alpha$-helical trans-activation domain (residues 17–27 of the protein p53 (PDB entry: 1YCQ). The second complex corresponds to the protein-binding domain of the chaperone protein GroEL (residues 218–277) and a peptide consisting of 11 residues (PDB entry: 1DKD; residues 602–612).

These relatively small test cases with 4–5 critical amino acids at the interface were chosen since they allow a systematic test with several start structures of various degrees of protein main chain and side chain deviations from the known experimental structure. In both test cases the larger protein partner (receptor) of the complex forms a concave surface at the interface with little possible variation of side chain placements. In contrast, the convex surface of the smaller peptide ligand allows many sterically possible side chain conformations. For the present test calculations, all side chains of the ligands were completely mobile, whereas the peptide main chain and all residues of the receptor protein were allowed to move freely within a window of $\pm 2 \, \text{Å}$ with respect to the start structure. This still allows for considerable adjustment of the complex geometry during the refinement simulations. All side chains of the peptide ligand were subject to potential scaling.
To mimic an initial docking of the peptide–protein complexes the start structures contained random placements of the ligand side chains, and the complete ligand peptide was initially placed near the receptor binding region with overall main chain misplacements (ligand) of between 0.2 and 1.3 Å from the experimental placement (Tables III and IV). Such displacements are typical for successful initial protein–protein and protein–peptide docking placements (Halperin et al., 2002). The RMSD of the initially placed interface side chains was in the range of 1–4 Å from the experiment. Between 420 and 530 TIP3P water molecules were added in a cap completely encompassing the ligand and interface region (see Materials and methods). For both test cases the PS-MD approach performed significantly better than both standard MD simulations and application of the SCWRL3.0 approach. Except for the ligand placements with small deviation from the experiment, neither SCWRL3.0 nor standard MD simulations (of 315 ps) gave any significant improvement of the side chain placement compared with the random starting placement. Even an extension of the standard MD simulations to a total simulation length of ~1 ns did not improve the results significantly, indicating that the side chains are trapped in conformations separated by energy barriers from conformations close to the experiment. In contrast, the PS-MD approach resulted in significant improvement of the side chain placement in ~5 (1DKD) to 6 (1YCQ) out of 7–9 start structures, respectively, during relatively short simulation times (315 ps; Figure 10). Interestingly, similar results were obtained in the absence of solvent (ddd model), indicating that at least in the present two cases steric interactions have a more dominant influence on the side chain placement than explicit solvent molecules.

**Discussion**

Most current methods to predict amino acid side chains in protein cores or at protein–protein interfaces are based on systematic searches of discrete possible rotamers based on the fixed main chain (reviewed by Dunbrack, 2002). Deterministic and extremely rapid algorithms based on the dead-end-elimination theorem have been developed that allow side chain prediction of medium-sized proteins within a few seconds (Canutescu et al., 2003). However, such approaches are limited due to the neglect of conformational adjustment of the protein main chain (Harbury et al., 1995). In particular, homology modelling of proteins can involve side chain placements on template protein main chains which, depending on the template–target homology, may significantly deviate from the true main chain. The purpose of the present study was to test and evaluate an alternative side chain prediction scheme that allows protein main chain adjustment during side chain generation within preset limits and the inclusion of explicit solvent using MD simulations. To overcome the sampling limitations...
of conventional MD simulations we employed a potential scaling scheme of the selected amino acid side chains. The approach is related to SA-MD methods where one starts with a very high simulation temperature that allows energy barriers to be overcome, followed by a gradual decrease of the temperature and final EM (Kirkpatrick et al., 1983; Brunger et al., 1997). In contrast to SA, the PS-MD approach allows the barriers involved for side chain readjustments to be selectively lowered with a minimum perturbation of other parts of the simulation system. Increased temperature simulations always affect the complete simulation systems and can for example strongly disturb the solvent atmosphere in explicit solvent simulations.

Potential scaling approaches have been used previously to tackle the problem of sampling limitations of conventional MD simulations (Piela et al., 1989; van Schaik et al., 1992; Mierke and Kessler, 1993; Dahiyat and Mayo, 1997; Huber et al., 1997; Tappura et al., 2000; Tappura, 2001; Riemann and Zacharias, 2004). However, the application of potential scaling during MD simulations to predict and refine side chain conformations in protein cores or at protein–protein interfaces has so far not been evaluated.

In the case of protein side chain prediction of protein cores (in the absence of solvent) with slightly perturbed protein main chains, PS-MD showed similar prediction accuracy to SA-MD and only for one test system better performance than the dead-end elimination-theorem based rotamer search approach SQWRL3.0. The results depended on the protein type and the best results were obtained with an α-helical protein domain (1R69). The approach was further tested on 12 proteins for which high-resolution X-ray structures have been determined (Table I). Generation of start structures with random initial buried side chain placements resulted in structures with main chain deviations of 0.5–0.8 Å from the experiment. The SQWRL3.0 approach gave excellent prediction results for buried side chains when using the experimental main chain structure. Still reasonable performance was achieved when using the generated start structures (with ‘incorrect’ main chain) as templates. However, using the same start structures for the PS-MD approach and allowing main chain flexibility during PS-MD of up to ±1 Å resulted in overall better buried side chain predictions in 10 out of 12 cases compared with the deterministic approach using a rigid main chain structure.

In the case of comparative protein modelling of the 1R69 or 1GFC test cases with protein templates that deviated by 0.8–1.6 Å from the known target main chain, the PS-MD approach performed significantly better than the SQWRL3.0 approach (and conventional MD simulations). It is important to note that although the start structures used in the first test of the method had an overall main chain deviation of ~0.4–0.8 Å from the experiment, considerably larger deviations of the main chain near initially misplaced side chains were present. This can strongly affect the performance of the approach and might be an explanation as to why the PS-MD method performed quite well in the case of comparative modelling and less well for some of the start structures used in the first test.

The present PS-MD method performed significantly better than conventional MD or the deterministic side chain prediction in the case of protein–peptide interface refinement. In this case the protein main chain of the peptide was allowed to deviate from the experiment by up to 1.3 Å, including overall translational and rotational shifts of the ligand. An additional difficulty is due to the further stabilization of the misplaced peptide conformations by EM of the incorrect side chain placements before the side chain refinement. However, this corresponds to a typical docking scenario where one energy minimizes initially docked complexes to reduce steric overlap between the receptor and the ligand. Apparently, the

Fig. 9. (A) Superposition of one GroEL-bound peptide start structure (residues 602–612, light grey) onto the experimental peptide structure (black) in complex with the chaperone GroEL-binding domain (residues 218–277; PDB entry: 1DKD). The GroEL-binding domain is shown as surface representation and the peptides as stick model. (B) Same as in (A) but showing the final peptide structure (light grey) obtained using SCWRL3.0. (C) Same as in (A) showing the final peptide structure (light grey) obtained using standard MD simulations (see Materials and methods). (D) Same as in (A) with the final structure from the PS-MD approach (light grey). For comparison the experimental bound peptide is shown in black in each panel.
The performance of the deterministic side chain prediction approach, which does not allow overall conformational relaxation of the protein and the peptide during side chain prediction, depends sensitively on the main chain conformation of the start structure. However, conventional MD simulations that allow overall conformational adjustment of both the protein and the peptide fail because the initial side chain placement at the interface is separated from the realistic placement by high-energy barriers. The PS-MD approach allows easy crossing of the barriers during the potential scaling and final settlement of side chains close to the realistic structure for most start conformations. A disadvantage of the present approach is that it is much more time consuming than methods based on a fixed protein main chain and a rotamer search. However, even in the case of larger proteins or protein–protein complexes it is often possible to limit the region of interest to only the interface between two proteins in the case of a protein–protein complex or to part of the protein structure (e.g. region around a protein mutation). Especially in the case of introducing several mutations in a protein, it has been shown experimentally that conformational adjustments of the protein are important and side chain prediction based on a fixed side chain (of the wild-type structure) can fail to accurately predict the side chain conformations owing to the mutation (Baldwin et al., 1993).

Fig. 10. (A) Sorted RMSD of p53 peptide interface side chains of nine final PS-MD structures (315 ps; open bars) in complex 1YCQ in comparison with the starting structures (closed bars) and results from standard MD simulations (315 ps; light grey bars). All standard and PS-MD simulations were performed, including 421 TIP3P water molecules and with ‘window restraints’ (allowing full mobility of all peptide side chains and free motion within ±2 Å of all other solute atoms with respect to the corresponding start structure). (B) Same as in (A) except that the results are presented for seven structures for the GroEL-binding peptide in complex with the GroEL protein-binding domain (PDB entry: 1DKD). (C) Same as in (A) but for vacuum runs using the ddd model. The SCWRL3.0 predictions (dark grey bars) are also included. (D) Same as in (C) but for the 1DKD case including the SCWRL3.0 predictions (dark grey bars).

See the footnote to Table III.

Table IV. Cα RMSDs of peptide–receptor start structures (1–7) with respect to the experiment (1DKD complex)

<table>
<thead>
<tr>
<th>Structure number (1DKD)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cα RMSD total (Å)</td>
<td>0.55</td>
<td>0.30</td>
<td>0.23</td>
<td>0.19</td>
<td>0.29</td>
<td>0.17</td>
<td>0.21</td>
</tr>
<tr>
<td>Cα RMSD protein (Å)</td>
<td>0.46</td>
<td>0.24</td>
<td>0.21</td>
<td>0.17</td>
<td>0.23</td>
<td>0.16</td>
<td>0.17</td>
</tr>
<tr>
<td>Cα RMSD peptide (Å)</td>
<td>1.10</td>
<td>0.48</td>
<td>0.35</td>
<td>0.27</td>
<td>0.50</td>
<td>0.22</td>
<td>0.32</td>
</tr>
<tr>
<td>RMSD peptide (Å; free residues)</td>
<td>3.88</td>
<td>3.45</td>
<td>3.53</td>
<td>3.08</td>
<td>3.35</td>
<td>3.36</td>
<td>2.48</td>
</tr>
</tbody>
</table>

Table V. Cα RMSDs of peptide–receptor start structures (1–9) with respect to the experiment (1YCQ complex)

<table>
<thead>
<tr>
<th>Structure number (1YCQ)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cα RMSD total (Å)</td>
<td>0.23</td>
<td>0.30</td>
<td>0.56</td>
<td>0.54</td>
<td>0.55</td>
<td>0.45</td>
<td>0.38</td>
<td>0.42</td>
<td>0.24</td>
</tr>
<tr>
<td>Cα RMSD protein (Å)</td>
<td>0.22</td>
<td>0.22</td>
<td>0.51</td>
<td>0.51</td>
<td>0.47</td>
<td>0.40</td>
<td>0.14</td>
<td>0.14</td>
<td>0.13</td>
</tr>
<tr>
<td>Cα RMSD peptide (Å)</td>
<td>0.33</td>
<td>0.66</td>
<td>0.89</td>
<td>0.71</td>
<td>1.05</td>
<td>0.78</td>
<td>1.09</td>
<td>1.25</td>
<td>0.65</td>
</tr>
<tr>
<td>RMSD peptide (Å; free residues)</td>
<td>3.64</td>
<td>3.96</td>
<td>3.39</td>
<td>2.94</td>
<td>2.44</td>
<td>2.78</td>
<td>2.89</td>
<td>2.70</td>
<td>1.06</td>
</tr>
</tbody>
</table>

Cα RMSD total/protein/peptide refers to the RMSD of the complex/protein/peptide with respect to the experimental structure, respectively. The last row indicates the RMSD of the three hydrophobic residues 19, 23 and 26 of the peptide at the interface to the receptor protein.
Acknowledgements

The authors thank A. Barthel and D. Roccatano for helpful discussions. This work was supported by a grant to the Jenaer Center of Bioinformatics (JCB) of the Bundesministerium für Bildung und Forschung (BMBF), Germany.

References

Case, D.A. et al. (1999) AMBER 6, University of California, San Francisco.

Received January 17, 2005; revised April 21, 2005; accepted July 11, 2005

Edited by Fred Cohen