IRECS: A new algorithm for the selection of most probable ensembles of side-chain conformations in protein models

CHRISTOPH HARTMANN, IRIS ANTES, AND THOMAS LENGAUER

Max-Planck-Institute for Informatics, 66123 Saarbrücken, Germany

(RECEIVED November 29, 2006; FINAL REVISION March 1, 2007; ACCEPTED April 3, 2007)

Abstract

We introduce a new algorithm, IRECS (Iterative REDuction of Conformational Space), for identifying ensembles of most probable side-chain conformations for homology modeling. On the basis of a given rotamer library, IRECS ranks all side-chain rotamers of a protein according to the probability with which each side chain adopts the respective rotamer conformation. This ranking enables the user to select small rotamer sets that are most likely to contain a near-native rotamer for each side chain. IRECS can therefore act as a fast heuristic alternative to the Dead-End-Elimination algorithm (DEE). In contrast to DEE, IRECS allows for the selection of rotamer subsets of arbitrary size, thus being able to define structure ensembles for a protein. We show that the selection of more than one rotamer per side chain is generally meaningful, since the selected rotamers represent the conformational space of flexible side chains. A knowledge-based statistical potential ROTA was constructed for the IRECS algorithm. The potential was optimized to discriminate between side-chain conformations of native and rotameric decoys of protein structures. By restricting the number of rotamers per side chain to one, IRECS can optimize side chains for a single conformation model. The average accuracy of IRECS for the χ_1 and $\chi_1 + \chi_2$ dihedral angles amounts to 84.7% and 71.6%, respectively, using a 40° cutoff. When we compared IRECS with SCWRL and SCAP, the performance of IRECS was comparable to that of both methods. IRECS and the ROTA potential are available for download from the URL <http://irecs.bioinf.mpi-inf.mpg.de>.

Keywords: new methods; protein structure prediction, IRECS, side-chain prediction; structural ensembles

Supplemental material: [see www.proteinscience.org](http://www.proteinscience.org)

Computational methods for structural analysis and molecular interaction studies require high-quality three-dimensional (3D) structures of the target protein. In cases in which no experimental structure is available, homology modeling techniques can be used to build a computational model of the protein if a template protein structure with sufficient sequence similarity is at hand. At a sequence

identity level above 50%, for instance, the conformation of the peptide backbone is highly conserved in most cases. However, the side-chain conformations are more variable. Owing to insufficient information about the protein environment, side-chain conformations must be computed based on the backbone conformation of the model and the limited information on side-chain conformations and environment that can be deduced from template structures. The conformational space of the side chains is usually limited to a discrete set of so-called rotamers during modeling. The classical side-chain prediction problem is to assign the correct rotamer to each residue of the protein backbone. This problem is known

Reprint requests to: Christoph Hartmann, Max-Planck-Institute for Informatics, 66123 Saarbrücken, Germany; e-mail: Hartmann@mpi-inf.mpg.de; fax: 49-681-9325399.

Article published online ahead of print. Article and publication date are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.062658307>.

to be NP-hard if pairwise potentials are used for scoring energy (Pierce and Winfree 2002). Many algorithmic heuristics have been applied to this problem, among them Monte Carlo simulations (Holm and Sander 1992; Liang and Grishin 2002), steepest-descent minimization with random restarts (Xiang and Honig 2001), and self-consistent mean-field approaches (Koehl and Delarue 1994). The tool SCWRL combines the Dead-End Elimination (DEE) algorithm, which restricts the conformational space, with an exhaustive but fast search of the remaining conformational space (Bower et al. 1997; Dunbrack and Cohen 1997; Canutescu et al. 2003). In addition to the combinatorial problem, other issues like imperfect scoring schemes, incomplete modeling of the protein environment (e.g., solvent, cofactors, ligands, attached macromolecules), and side-chain flexibility hamper the generation of a high-quality protein model and also complicate the evaluation of side-chain predictions. Because of these challenges, the achieved prediction accuracies reported in the literature for χ_1 and $\chi_1 + 2$ are just below 90% and 80%, respectively, considering all side chains and a rotational deviation of $\pm 40^\circ$ (Mendes et al. 1999; Liang and Grishin 2002; Peterson et al. 2004).

In fields such as structure-based drug design or protein-protein docking, it is quite important to take protein flexibility into account. Hence a representation of the protein consisting of only a single conformation is insufficient (Leach and Lemon 1998; Claussen et al. 2001; Teodoro and Kavraki 2003; Gray 2006). In computational approaches protein flexibility is often represented via a set of protein conformations, also known as a conformational ensemble. This representation is motivated by the observation that a protein thermodynamically adopts no unique but one of many conformations. There are many low-energy conformations of the protein with the global minimum-energy conformation (GMEC) as the most probable conformation. In the conformational ensemble, each of the side chains has a limited set of allowed rotamers that are either attached to a fixed backbone conformation or may be combined with different backbone conformations. These rotamer sets usually comprise the side-chain conformations observed in different crystal structures, multiple molecular simulation snapshots, or from differently parameterized side-chain predictions. In the latter case, the obtained rotamer sets usually represent subsets of the full rotamer sets that could be assigned to the respective side chains from an underlying rotamer library.

One of the most prominent approaches for generating rotamer subsets for protein ensembles is the Dead-End-Elimination (DEE) algorithm (Desmet et al. 1992). This algorithm removes rotamers that cannot be part of the GMEC. There are some performance extensions that apply tighter removal criteria than Desmet et al. (1992)

did, like the Goldstein DEE (Goldstein 1994) or Split DEE (Gordon et al. 2003). There are also several algorithms that speed up the application of the DEE criteria by preprocessing the graph representing all pairwise side-chain interactions, like the articulation-point reduction technique of SCWRL (Canutescu et al. 2003). The R3 algorithm of Xie and Sahinidis (2006) combines several of these techniques. As Gordon and coworkers found, the performance of the DEE algorithm in eliminating rotamers strongly depends on the rotamer library and scoring function used (Gordon et al. 2003). DEE works best if there are few rotamers per side chain and if the scoring function assigns a higher weight to single-residue contributions than to pairwise residue-interaction energies.

All potential functions developed so far introduce some (mostly not quantified) errors in the absolute and relative values of the computed energies. Even if these error ranges were known, it would not be possible to include them in the DEE criteria without drastically decreasing the DEE performance. Moreover, the existence of these errors raises the question if the theoretical GMEC is also the most probable conformation in reality.

In the present study, we propose the IRECS algorithm as a heuristic alternative to the DEE algorithm for rotamer ensemble selection. The algorithmic nature of IRECS easily allows for considerations of error ranges as well as handling the flexibility of side chains by presetting the average size of the selected rotamer sets right from the start. First, we formulate the probabilistic rotamer-subset problem. Second, we describe the IRECS algorithm for solving this problem heuristically. Third, we introduce a new scoring function that we calibrate to be applicable to rotamer scoring. Fourth, the algorithm is evaluated on a large protein test set, and its performance is compared to the other side-chain prediction programs, SCWRL and SCAP. Finally, we discuss the pros and cons of using IRECS instead of DEE or single side-chain conformation prediction algorithms and point to possible improvements and extensions to our method.

Probabilistic rotamer-subset problem

First, we formally define the problem of finding the most probable rotamer subset and introduce the notation used below in the manuscript. $L = \{L_1, \dots, L_n\}$ is the set of the initial sets of rotamers that are assigned to each of the n side chains of a target protein. For simplicity, we assume that these sets are maximal, that is, any set L_i contains all rotamers from the underlying rotamer library for the amino acid at position i of the target protein. $X = \{X_1, \dots, X_n\}$ denotes any possible collection of subsets that can be created by eliminating rotamers from the sets $L_1, \dots, L_n$. The set L is the rotamer universe of the target protein, and the set X is the considered reduced rotamer

universe of the target protein. X represents a selection on L and is a solution candidate of the IRECS algorithm. In the special case that all $X_1, \dots, X_n$ contain exactly one rotamer, X corresponds to a solution candidate for the traditional side-chain prediction problem for a single protein conformation.

$G(L)$ is the discrete conformational space available to the target protein created from a fixed backbone conformation and exactly one rotamer per side chain i out of L_i . By restriction of all sets L_i to the smaller subsets X_i , the conformational space $G(L)$ is restricted to the conformational space $G(X)$. We assign a probability to the restricted rotamer universe X and denote it $P_G(X)$. This probability can be interpreted as the probability of observing a conformation of a protein from $G(L)$ that contains only rotamers from X . By this definition, $P_G(L)$ is one. The probability $P_G(X)$ is the sum of the probabilities of all conformations C of a protein that can be constructed using the fixed-backbone conformation and rotamers from X :

$$P_G(X) = \sum_{C \in G(X)} P(C) \quad (1)$$

The probability $P_G(X)$ thus involves two sets of terms, (1) the individual probabilities of the conformations and (2) the total number of conformations in $G(X)$. Desired solutions X for a usual protein-modeling purpose would have a maximal conformational probability and a minimal number of conformations. Since both cannot be optimized simultaneously, we restrict the number of conformations of $G(X)$ to lie within a user-defined range and then maximize $P_G(X)$. The number of conformations of $G(X)$ is

$$\prod_i^n |X_i|$$

and thus depends on the number of side chains of the target protein and the sizes of the rotamer subsets X_i . We use the rotamer density

$$\rho_{\text{rot}}(X) = \frac{1}{n} \sum_i^n |X_i| \quad (2)$$

as an estimate for the conformational variety. The division by n allows for comparing the conformational complexity of target proteins of different sizes. This measure allows us to fix the overall number of rotamers but allows for choosing the individual number of rotamers in each rotamer subset X_i .

The probabilistic rotamer-subset problem can now be defined as follows: Given the full rotamer universe L for n side chains and a fixed-backbone conformation of a

target protein, find a selection X that meets the following criteria:

$$\begin{aligned} P_G(X) &\geq P_G(Y) \text{ and } b_{\text{low}} \leq \rho_{\text{rot}}(X), \\ \rho_{\text{rot}}(X) &\leq b_{\text{high}} \text{ and } 1 \leq b_{\text{low}} \leq b_{\text{high}} \end{aligned} \quad (3)$$

with Y being any possible selection from L . Here b_{low} and b_{high} are user-defined constants. The subset probabilities $P_G(X)$ are given by Boltzmann's law, restricted to conformations C of $G(L)$. For a single conformation, $P(C)$ is given by

$$P(C) = \frac{e^{-\frac{E(C)}{kT}}}{\sum_{C \in G(L)} e^{-\frac{E(C)}{kT}}} \quad (4)$$

where $E(C)$ is the energy of conformation C , k is the Boltzmann constant, and T is the temperature. Insertion of Equation 4 into the definition of the conformation probability gives

$$P_G(X) = \frac{\sum_{C \in G(X)} e^{-\frac{E(C)}{kT}}}{\sum_{C \in G(L)} e^{-\frac{E(C)}{kT}}} \quad (5)$$

This equation relies on the inaccurate but useful assumption that the full conformational space of the protein can be approximated by $G(L)$. The computation of the subset probability $P_G(X)$ requires enumerating all conformations in $G(X)$ and exactly determining their energies. Since neither is possible, we must rely on heuristic scoring schemes and algorithms for computing $P_G(X)$. If we fix the rotamer density ρ_{rot} to one, the probability $P_G(X)$ of the subset X that contains all rotamers of the GMEC is greater than that of any other possible subset selection with the rotamer density of one. As noted above, the probabilistic rotamer-subset problem with a rotamer density of one is equivalent to the side-chain prediction problem. Thus, in this case, the probabilistic rotamer-subset problem is also NP-hard. On the other side, the problem becomes trivial if the rotamer density has no upper limit, since then the full rotamer universe L is a legal solution candidate.

Materials and Methods

Structural analysis

If the rotamer subset contains only one rotamer per side chain, the IRECS algorithm creates a side-chain conformation assignment, which can be evaluated with the traditional measures of root-mean-square distance (RMSD) of side-chain atoms and χ_1 , χ_{1+2} dihedral angle matching. Before distance and coordinate comparisons are performed, the coordinates of symmetric atoms in the

residues arginine, aspartic acid, glutamic acid, phenylalanine, and tyrosine are selected to yield to the optimal atom matching. Hydrogen atoms are not considered. Two side-chain dihedral angles are considered as matching if they differ by $<40^\circ$, as it is common for the evaluation of many side-chain prediction methods using such coarse-grained rotamer libraries as we do. A predicted side-chain conformation is considered as being correct if all of its dihedral angles match those of the conformation of the corresponding side chain inside a reference high-quality protein structure (usually determined by X-ray crystallography). A selection of multiple rotamers for a single side chain is considered correct if it contains at least one correct rotamer.

For the comparison of pairs of protein conformations by their RMSD, we use only coordinates of heavy side-chain atoms, excluding coordinates of C_β atoms. The discrimination between core and surface residues is performed using relative solvent accessibility (RSA) as a burial indicator. RSA values are calculated with the Naccess tool (version 2.1.1, <http://wolf.bi.umist.ac.uk/naccess/>; Hubbard et al. 1991). A residue is defined as belonging to the protein core if its relative accessibility is $<20\%$. This marks nearly half of all residues of the proteins in the test set as belonging to the core.

The IRECS algorithm: Building rotamer subsets

We developed a new heuristic algorithm IRECS (Iterative REDuction of Conformational Space) for the selection of rotamer subsets. The rotamer library used, the scoring function, and the final average number of rotamers assigned to the side chains can be configured freely. The algorithm combines the effective energy approach of the SCMF algorithm (Koehl and Delarue 1996) with a heuristic greedy iteration scheme. The general workflow of the algorithm is sketched in Figure 1.

First, the full rotamer universe L of all conformations for all side chains of the target protein is constructed: Starting from the backbone conformation, the atom coordinates of one initial conformation per side chain are built using the standard bond lengths and angles from the CHARMM (Brooks et al. 1983) force field. Then, all rotamers are constructed by repeated rotations of the initial side-chain conformation according to the dihedral angles of the rotamers found in the BBDEP rotamer library. Subsequently, interactions with the fixed backbone are computed. A simple filtering procedure removes all rotamers that are likely not to represent the correct conformation of their side chain: For every rotamer the difference between its own interaction score with the backbone and the score of the best-scoring rotamer of the respective side chain is computed. If this difference is larger than some threshold (10, by definition without a

Figure 1. Program flow diagram of IRECS.

unit here, determined by test runs on the training set), the rotamer is removed directly.

After this filtering step, pairwise interaction scores between all rotamers of different side chains are computed. This is one of the computationally most elaborate steps because of the still large number of active rotamers at this stage. This computation is accelerated by the definition of a local residue neighborhood: For a given residue, its local neighborhood comprises all other residues of the protein that can possibly interact with it. The neighborhood can be computed by checking for each candidate residue if the sum of the radii of the enclosing spheres of the target residue and the candidate residue is smaller than the distance between the sphere centers plus the distance cutoff of the applied scoring function (ROTA: 10 Å). Next, a probability is assigned to each rotamer. These probabilities assess how likely it is that the side chain actually is in the respective rotameric state. Initially, for each side chain, rotamer probabilities are distributed evenly. The effective energy E_{eff} of a rotamer x_i is defined as:

$$E_{\text{eff}}(x_i) = U_{\text{self}}(x_i) + U_{\text{inter}}(x_i, b) + \sum_{j=1, j \neq i}^s \sum_{y_j}^{c_j} p(y_j) U_{\text{inter}}(x_i, y_j) \quad (6)$$

The calculation of the terms U_{self} and U_{inter} is described below. $p(y_i)$ is the probability of the rotamer y_i , S denotes the set of all side chains, b is the protein backbone, and C_j is the set of rotamers of some side chain j . This definition is patterned after the application of the self-consistent mean-field (SCMF) theory to the side-chain prediction problem (Koehl and Delarue 1994).

After this initialization procedure, IRECS enters an iterative optimization scheme that tries to minimize the effective energy of the whole system. For this purpose, the range of effective rotamer energies is calculated for each side chain:

$$\Delta E_{\text{eff}}(s) = \max_{y_i \in C_i} (E_{\text{eff}}(y_i)) - \min_{x_i \in C_i} (E_{\text{eff}}(x_i)) \quad (7)$$

where x_i and y_i are the side-chain rotamers with the largest and smallest effective energy, respectively. The side chain with the largest $\Delta E_{\text{eff}}(s)$ is identified, and for this side chain the probability of the rotamer with the highest energy is set to zero and the rotamer is labeled inactive. This leads effectively to the elimination of energetically unfavorable rotamers. In addition, side chains are preferably chosen in which the energy gap between different conformations is large, that is, which have only a low number of energetically favorable conformations and thus limited flexibility. Subsequently, all probabilities of the remaining active rotamers for this side chain are set to $1/m$, where m is the number of active rotamers of this side chain. We also performed test runs in which we calculated the probabilities by Boltzmann's law using the effective rotamer energies of all remaining rotamers of a side chain. This, however, did not lead to an improvement of the overall performance of IRECS.

The algorithm described above has several advantages. First, the algorithm is certain to terminate, since the probability of each rotamer increases monotonically until termination or until it is set to zero for the rest of the optimization. Second, the number of iterations required for system calibration is independent from the scoring function used. Since exactly one rotamer is removed per iteration, the number of iterations is determined by the number of side chains, the starting rotamer density, and the user-defined rotamer density at the end of the optimization. Third, the size of the conformational space of all side chains also decreases monotonically by removing available rotamers. Fourth, the one-by-one removal of rotamers defines a unique elimination order of rotamers. This order is used to define meaningful rotamer subsets of arbitrary size. The algorithm can be stopped at any rotamer density, $\rho_{\text{rot}}(X)$ (see Equation 2), that is, any average number of rotamers per side chain. For a given rotamer density, the actual number of rotamers per side chain in the final structure can vary and is correlated to the flexibility of a specific residue.

Specifically, such a set comprises all rotamers that are eliminated in the latest IRECS steps. It has to be stressed that the process of rotamer removal is heuristic. Rather than performing a complete re-computation of all rotamer probabilities as in the SCMF algorithm, only the probabilities of the rotamers of the considered side chain are readjusted.

Finally, the protein model is written to a text file in PDB format. If a side chain has multiple rotamers, each of these rotamers is assigned an uppercase letter, and the atoms of these rotamers are labeled with this letter in their alternative location field. Also, the effective energy of the rotamers is written to the temperature field of each atom of a rotamer. Using the Boltzmann law and all rotamers of a side chain, rotamer probabilities are calculated that fit to the final effective energy. These probabilities are written to the occupancy field of each atom of a rotamer. Since only one or a few rotamers of a side chain are left after an IRECS optimization and appear in the protein model file, their probabilities do not add to one. This helps the user to estimate for each side chain the loss in conformational representation that occurred through the application of IRECS on the full rotamer sets from the rotamer library.

The performance of IRECS is crucially dependent on the scoring function used; thus we derived a special scoring function (IRECS_{score}) based on a new knowledge-based scoring term ROTA, which is used to calculate the second and third terms of Equation 6.

IRECS_{score} scoring function

The overall functional form of the IRECS_{score} is given in Equation 6. In this section we define the terms U_{self} and U_{inter} of the effective rotamer energy. U_{self} calculates self-scores of the rotamers and interaction scores of rotamers with the local backbone, and U_{inter} calculates nonbonded interactions of a rotamer with other rotamers and interactions with the whole protein backbone. Both terms are weighted with respect to each other using the training set of single protein structures (set A) mentioned below.

The self-score, U_{self} , is calculated using the BBDEP (backbone-dependent rotamer library) frequency scores. It is calculated analogously to the definition of the scoring function used in SCWRL (Canutescu et al. 2003). The conditional rotamer probabilities are taken from the BBDEP using the local backbone torsion angles Φ_i and Ψ_i .

$$U_{\text{self}}(x_i) = -w_1 \log \frac{p(x_i | \Phi_i, \Psi_i)}{\max_{y_i \in C_i} p(y_i | \Phi_i, \Psi_i)}. \quad (8)$$

The interaction term, U_{inter} , for a rotamer x_i with another rotamer or the protein backbone is computed by summing up all ROTA scores of all pairwise distances

between all atoms a of this rotamer and all atoms b of the other entity z , which is either another rotamer or the protein backbone as defined in Equation 6.

$$U_{\text{inter}}(x_i, z) = w_2 \text{ROTA}(a, b) \quad (9)$$

The weights w_1 and w_2 were fitted using the training part of protein set A defined below. Their values are 1.0 and 0.14, respectively.

The ROTA function

We follow the formalism for conditional probabilities of pairwise atom–atom preferences in proteins using statistical observations on native structures introduced by Samudrala and Moulton (1998a,b) to derive the RAPDF (residue-atom specific probability density function) scoring function. This function proved to be useful in model quality assessment programs (MQAP) (Tosatto 2005) as well as for side-chain prediction (Samudrala and Moulton 1998a,b). Our knowledge-based scoring function ROTA is derived from a “correct” set of crystal structures and a reference set of “incorrect” (also: decoy) protein models with the same backbone conformation as the structures from the correct set but enumerating multiple and possibly wrong rotameric states of the side chains.

The aim is to analyze a set of correct $\{C\}$ and incorrect $\{I\}$ structures on the basis of certain structural features f , which in our case are atom–atom distances. The scoring function sums over all negative log-conditional probabilities for observing a given set of features in the set of correct structures divided by the probability of observing the same feature in the overall set of combined correct and incorrect structures:

$$S(\{f\}) = - \sum_f \ln \frac{P(f|C)}{P(f|(C+I))} \quad (10)$$

This sum is proportional to $-\ln P(C|\{f\})$, where $P(C|\{f\})$ is the probability of a structure being correct if features $\{f\}$ are observed. The derivation of the scoring function requires calculating the probability of observed features $\{f\}$ both in a set of correct structures and a reference set comprising correct and incorrect structures. Since no a priori information about the correctness of a structure must be included in $P(\{f\})$, an appropriate composition of a reference set from correct and incorrect structures is hard to find. In order to solve this problem, we assume that the set of incorrect structures contains examples for all relevant error cases, in our case structural motives that originate from side-chain prediction errors. Usually it is not known what exactly characterizes all incorrect protein structures. Since side-chain prediction is a quite straightforward process, though, we show that we can construct artificial sets of incorrect structures that contain

the most relevant errors occurring in side-chain modeling. Given such a representative set of incorrect structures, we can derive a second scoring function:

$$\bar{S}(\{f\}) = - \sum_f \ln \frac{P(f|I)}{P(f|(C+I))} \quad (11)$$

Analogous to S , $\bar{S}$ is proportional to $-\ln P(I|\{f\})$, where $P(I|\{f\})$ is the probability of a structure being incorrect if features $\{f\}$ are observed. In combination, both yield a difference function ΔS :

$$\Delta S(\{f\}) = S(\{f\}) - \bar{S}(\{f\}) = - \sum_f \ln \frac{P(f|C)}{P(f|I)} \quad (12)$$

The required conditional probabilities $P(f|C)$ and $P(f|I)$ are obtained by counting the individual frequency of each feature f in the correct and incorrect set. The features used here and also for the RAPDF are different distance intervals of atom pairs that all have one of the unique 167 PDB atom types (residue name plus atom identifier). The probability of observing two atoms with type a and b within distance range d is calculated by

$$P(d_{ab}|C) = \text{frequency}(d_{ab}) = \frac{N(d_{ab})}{\sum_{d'_{ab}} N(d'_{ab})} \quad (13)$$

where d_{ab} is a specific distance bin for atom types a and b , and $N(d_{ab})$ is the number of atoms observed in the correct set to be in the distance range d and that have atom types a and b . The denominator contains the sum of all counts of all distance ranges for these two atom types. Substitution of $P(d_{ab}|C)$ and $P(d_{ab}|I)$ (computed analog) into Equation 12 yields the ROTA function. We used distance bins of width 0.25 Å, within the range 0–10 Å. In cases in which there are no instances of a certain atom pair within a distance range in the correct set, the score for this distance bin is set to the score of the distance bin that has the smallest correct/incorrect counts ratio. This correction was performed especially for small-distance ranges, which account for steric clashes; the score for this distance range was rounded up to 4.0. Given a query atom distance, we interpolate the scores of the two neighboring distance bins linearly.

Side-chain decoy sets for the derivation of the ROTA function

Since ROTA must discriminate between correct and incorrect side-chain conformations, we created the decoy sets AllLib, AllRot, TripleLib, and TripleRot with different distortion techniques for side-chain conformations. For the set of correct protein structures, we chose crystal structures from the high-quality top-500 set (Lovell et al.

2003), downloaded from the server at <http://kinemage.biochem.duke.edu/> (Richardson Laboratory). We randomly split the set into 490 structures for training and 10 structures for testing. The PDB codes of the test set comprise 1edg, 1jet, 1lmb4, 1pen, 1qb7, 1qq4, 1vfy, 1wab, 3ebx, and 3ezm. All of the 500 structures were used as base structures for generating decoy sets. Only the decoys of the training set were used to derive ROTA. The decoy sets were generated using four different decoy procedures. Each of them was designed to capture some aspect of fallacious side-chain prediction. For the AllLib and AllRot sets, all side chains were randomly rotated simultaneously. This procedure was performed 10 times for each protein structure. The decoys of the sets TripleLib and TripleRot were created by rotating the side chains of only three neighboring residues simultaneously. We slid a window comprising three residues along the peptide chain, creating one decoy per shift. This procedure is supposed to create decoy distances between atom pairs of native and disturbed side chains and for atom pairs of close pairs of disturbed side chains. The sets AllLib and TripleLib draw their new side-chain dihedral angles from the backbone-dependent rotamer library, with a selection probability among all available rotamers as defined in the library. This causes the decoys to contain mostly the rotamers favored by the BBDEP, but ignoring any long-range forces caused by atomic interactions. When distance distributions of atom pairs are drawn from the correct set and this kind of decoy set, the ROTA scores receive two important properties: First, ROTA favors rotamers that are less probable in the BBDEP; and second, ROTA favors those rotamers that establish meaningful atomic interactions with their environment. Rotamers that have a high frequency in the library but are not represented proportionally in the correct structure set are penalized. This should be the case for most of the false rotamer assignments from the BBDEP. This setup was chosen to adjust the ROTA scores to the BBDEP frequency scores.

The side-chain dihedral angles of the decoys in the AllRot and TripleRot sets are distributed uniformly over all rotational angles from 0° to 360°. Steric clashes were not removed from the generated conformations, since clashes should also be penalized by ROTA without help of another scoring function. As a result, most of the decoys are clearly identifiable as decoys by quite simple methods. Only in the case of small perturbations, which occurred often in the TripleLib and TripleRot sets, are decoys hard to distinguish from the starting crystal structures. See Table 1 for details on the rotamer sets.

We also used the backbone decoy set 4-state-reduced, downloaded from the “Decoys ‘R’ Us” Web site (<http://dd.compbio.washington.edu>), for comparing the performance of the four ROTA versions and RAPDF on discrim-

Table 1. Properties of the decoy sets for testing

Test decoy sets	Number of proteins	Number of decoys	Side-chain RMSD range (Å)
4state	7	4656	2.55–13.25
AllLib	10	5000	0.76–3.68
AllRot	10	5000	1.81–4.34
TripleLib	10	1842	0.01–3.06
TripleRot	10	1842	0.06–3.15

inating decoys sets with disturbed backbone and side-chain conformations.

Structure sets for testing the performance of IRECS

The PISCES Web server at <http://dunbrack.fccc.edu/piscs> (Wang and Dunbrack 2003, 2005) was used to compile a list of high-quality protein structures. The list holds 641 structures with a resolution better than or equal to 1.5 Å, an R and R_{free} under 0.3, and a pairwise sequence identity lower than 25%. The list was divided into two main sets: The first set A contains 194 crystal structures with single conformations for all side chains. Set A was split again into a test subset of 160 structures and a training subset of 34 structures. The second set B contains 447 crystal structures that have at least one side chain with multiple conformations.

This selection does not take the structural training set of ROTA and the BBDEP into account. The Top-500 set has 31 overlaps with the test subset of set A and 59 overlaps with the set B. The training set of the BBDEP has 83 overlaps with the test subset of set A and 146 overlaps with the set B. These overlaps imply that the evaluation of IRECS and SCWRL will be biased in so far as a single side chain in the test set can benefit from its membership in the training sets. As it would be easy to reduce these overlaps by exchanging structures with those of close homologous proteins, such a replacement would more conceal the bias than reduce it. Since all protein sets used here aim at representing the conformational space of the proteins currently deposited in the PDB, it is not possible to eliminate this bias without reducing the represented conformational space of the training sets. As SCWRL and IRECS both make use of the BBDEP, the influence of this bias remains controllable in a comparison of these programs.

Construction of conformations

We calculated atom coordinates of the rotamers by using the standard values for bond angles and lengths of the CHARMM force field (Brooks et al. 1983; MacKerell et al. 1998; MacKerell et al. 2004). Only the coordinates of heavy atoms were calculated. Dihedral angles were drawn from the BBDEP (Dunbrack and Cohen 1997) and

therefore depend on the local conformation of the backbone. The BBDEP (version of May 2002) was downloaded from <http://dunbrack.fccc.edu/bbdep/>. We experienced limited accuracy improvements of IRECS when using native bond lengths and angles, but they were not used in the evaluation shown here.

Comparison setup

The RAPDF was downloaded from the Decoys 'R' Us Web site (<http://dd.compbio.washington.edu/>; Samudrala and Levitt 2000). All MQAP scores were calculated using an in-house Python-script implementation. SCWRL and SCAP were started using default parameters. We used SCWRL version 3.0. All side-chain prediction tools were tested on the same machine (AMD Opteron V20z).

Results

Performance of ROTA: Discriminating native from decoy structures

All versions of the scoring function ROTA, the BBDEP scores, and the RAPDF were tested for their ability to discriminate a native structure from a set of decoy structures. The purpose of this evaluation was to motivate our selection of a special ROTA version for the side-chain prediction. For the evaluation, we used the test set part of the generated side-chain decoy sets and the 4-state reduced decoy set. The results are summarized in Tables 2 and 3. The achieved Z-scores of the native structures show that the concept of using side-chain decoys as references for the scoring function increases the ability of discriminating between native structures and decoy structures with disturbed side chains but also decreases the ability of discriminating between backbone decoys and the native structure: RAPDF can identify all native structures in the 4-state set (putting it to rank one), whereas all different versions of the ROTA show a decreased performance on this set. In contrast, the number of native structures found at rank one among all the side-chain decoy sets is higher for the AllLib and

AllRot version of ROTA. The decoy sets TripleLib and TripleRot are more challenging, since they contain a larger fraction of decoys that are quite similar to the respective native structures. These sets are also handled quite well by all ROTA versions, since the rank of the native structure is 1 in more than 70% of all test sets. The RAPDF has very low discriminatory power for these sets, identifying the native structure from the decoys in only 25% of all test cases. Also the observed Z-scores show that the ROTA versions have a higher discriminatory power on the side-chain decoy sets than RAPDF, whereas the Z-scores are often quite similar. On the AllRot test set, the Z-score achieved by the RAPDF is greater than that of three out of four ROTA versions. Nevertheless, the ROTA function derived from the AllRot decoy set achieves greater or equal Z-score results than any of the non-ROTA scoring functions. The BBDEP scores show a low performance in the MQAP scenario in general, with a better performance on the decoys sets AllLib and AllRot. The correlation coefficients between the RMSD and score of the inspected structures generally support the above analysis, with two exceptions. First, RAPDF performs better than all other functions on the AllRot decoy set. Second, contrary to the prior analysis, the different versions of ROTA show nearly equal results.

Based on this analysis, we selected the ROTA version derived from the AllLib decoy set for the rotamer subset selection. We expected the rotamer reduction process to generate conformations through wrong rotamer assignments that are most similar to decoy structures from the TripleLib, where many side chains have correct conformations and only few are wrong. Owing to its high performance on this decoy set, the ROTA version derived from AllLib is the best candidate for guiding the rotamer removal process of IRECS.

Prediction of final structures with single side-chain conformations

By predicting protein models with a rotamer density of one, we can compare the performance of IRECS directly

Table 2. Number of native structures at rank one and their average Z-score among different sets of decoy structures

	4state (7)	AllLib (10)	AllRot (10)	TripleLib (10)	TripleRot (10)
RAPDF	7/-3.189	9/-6.744	10/-9.160	1/-1.516	4/-1.385
BBDEP	2/-2.569	6/-3.190	9/-9.391	0/-0.298	0/-0.953
ROTA-AllLib	5/-3.196	10/-7.850	10/-9.026	9/-1.531	8/-1.413
ROTA-AllRot	6/-3.547	10/-7.773	10/-9.391	7/-1.524	9/-1.453
ROTA-TripleLib	3/-1.439	10/-6.965	10/-8.243	6/-1.328	7/-1.266
ROTA-TripleRot	3/-1.445	10/-6.975	10/-8.254	7/-1.333	7/-1.272

The small deviation between the Z-score found here for the RAPDF on the 4state decoy set and that computed by S. Tosatto (Tosatto 2005) results from our usage of the scoring function form of the side-chain prediction algorithm in Samudrala and Moulton (1998b), since we do not calculate scores for intra-side-chain interactions.

Table 3. Correlation coefficients between computed score and native-decoy side-chain RMSD

	4state	AllLib	AllRot	TripleLib	TripleRot
RAPDF	0.666	0.556	0.621	0.579	0.684
BBDEP	0.224	0.197	0.468	−0.042	0.271
ROTA-AllLib	0.572	0.571	0.582	0.639	0.714
ROTA-AllRot	0.552	0.569	0.594	0.625	0.721
ROTA-TripleLib	0.525	0.536	0.555	0.602	0.710
ROTA-TripleRot	0.528	0.536	0.556	0.601	0.711

with that of other side-chain prediction tools. The tools SCWRL and SCAP were selected for comparison because of their popularity, availability, and high accuracy. The optimization algorithms used by these tools are quite different. All tools were used to predict models of all 160 proteins of the test part of set A. None of the tools included the knowledge of the true side-chain conformations from the native structure, except that some of the proteins were part of the training sets of the BBDep and ROTA. The performance of the tools was measured by the average similarity of the predicted model to the reference structure, once for all side chains and once for all side chains belonging to the core. In Table 4 the average values for all similarity measures are shown. In general, IRECS performs at best with respect to most of the applied similarity measures, except for the $\chi_1 + 2$ accuracy of side chains in the core. However, the differences observed here are all quite small and may be dependent on the selection of the test set. It should also be noted that dependent of the protein target each of the three tools is sometimes able to outperform the other tools by far. Figure 2 shows the average prediction accuracy of IRECS for each amino acid. The achieved prediction accuracy is comparable to that of other methods and does not show much difference (>10% increase or decrease) to the average prediction accuracy reported for other methods (Liang and Grishin 2002; Peterson et al. 2004).

Since it is hard to decide the respective influence of the optimization algorithm or the scoring function on the prediction accuracy, a simulation was performed that

aimed at eliminating the influence of the IRECS algorithm and to measure the performance of the IRECS_{score}. In this simulation the prediction of a side chain is simplified by choosing the native conformation for all other side chains of the protein. Each rotamer of the query side chain is then scored using only the native conformation of its neighboring side chains. This simulation was carried out for all side chains of the test set A. As expected, the performance of this simulation was slightly higher than the original IRECS program. Considering all side chains, the overall RMSD was lowered to 1.5 Å (1.551 Å for the IRECS program) (see Table 4), the average RMSD to 0.753 Å (0.775 Å), the χ_1 accuracy increased to 85.4% (84.7%), and the $\chi_1 + 2$ accuracy to 72.9% (71.6%). If only the side chains in the core of the proteins are considered, the overall RMSD was lowered to 0.999 Å (1.046 Å for the IRECS program), the average RMSD to 0.483 Å (0.502 Å), the χ_1 accuracy was slightly lower at 91.9% (92.0%), and the $\chi_1 + 2$ accuracy increased to 82.8% (82.2%). These results show that the IRECS algorithm quite accurately approximates the combinatorial problem of side-chain prediction and implies that the scoring function and modeling of the molecular environment are the accuracy-limiting components.

Relation between the selected rotamer subsets and side-chain flexibility

IRECS selects those residues for rotamer reduction that have the largest range of effective rotamer energies. This

Table 4. Comparison of side-chain prediction performance on the target backbone

	Runtime (sec) ^a	All				Core ^b			
		Overall RMSD (Å)	Average RMSD (Å)	χ_1 (%)	$\chi_1 + 2$ (%)	Overall RMSD (Å)	Average RMSD (Å)	χ_1 (%)	$\chi_1 + 2$ (%)
SCWRL	4.7	1.677	0.849	82.3	68.0	1.191	0.572	90.2	80.1
SCAP	161.0	1.605	0.823	84.0	70.6	1.120	0.529	91.6	83.9
IRECS	23.1	1.551	0.775	84.7	71.6	1.046	0.502	92.0	82.2

Runtime and RMSD values are averaged over all 160 predicted protein structures; the accuracy of χ_1 and $\chi_1 + 2$ predictions is averaged over all predicted residues.

^aThe runtime is averaged without the runtime on PDB code 1gd0, since SCWRL required >16 h for it.

^bTarget protein 1wfb was excluded since none of its residues were labeled as core.

Figure 2. Prediction accuracy of IRECS (Y-axis, percentage of matched dihedral angles) of amino acids for χ_1 (black), $\chi_1 + 2$ (dark gray), $\chi_1 + 2 + 3$ (light gray), and $\chi_1 + 2 + 3 + 4$ (white) for different amino acids (X-axis).

implies that, at the start of the optimization, a selected side chain has a pair of rotamers in which one rotamer dominates the other by some energy difference ΔE_{eff} . As the optimization continues, ΔE_{eff} decreases monotonically toward zero, for all side chains, and ΔE_{eff} finally becomes zero by definition for all side chains that only have one rotamer left. This also implies that in later stages the rotamers left in the ensembles have nearly the same interaction energies with their environment. This is the condition for a side chain to be flexible, since no rotameric state can dominate the others; conformational changes of flexible side chains have low energy barriers between the available rotamers. Therefore, we can interpret the final ensemble of rotamers of a side chain as its discrete conformational space and use the differences in effective rotamer energies as a measure for the degree of side-chain flexibility. We can also compare the flexibility of different side chains by defining a threshold value on the range of effective rotamer energies ΔE_{eff} and simply counting the still reachable rotameric states. Since IRECS aims at minimizing the energy difference ΔE_{eff} inside the rotamer ensembles of each side chain, the ensembles created by IRECS maximize the overall conformational flexibility of the side chains for a given rotamer density. IRECS assigns different numbers of rotamers to each side chain and thus can weight their conformational flexibility. In Figure 3, the average number of rotamers per remaining side chain is shown, resulting from 160 protein models built with IRECS on the native backbone and a rotamer density of two.

Residues at the protein surface are assigned more rotamers than residues in the protein core. Amino acids that are known to be highly flexible like arginine, lysine, or glutamic acid are assigned a much higher conformational variability than amino acids known to be more rigid like tryptophan, tyrosine, cysteine, or histidine. These

averaged observations can also be made by analyzing a single prediction run, as is shown in Figure 4. Here, we predicted side chains with a rotamer density of two on the backbone of the human UDP-galactose 4-epimerase known from the crystal structure 1ek6 (Thoden et al. 2000). This crystal structure was chosen as an example because of its high quality (1.5 Å resolution, R -value 0.169, and R_{free} 0.198) and the presence of 16 side chains with alternative conformations. Figure 4A shows a correctly predicted arginine; it directly points toward the protein core and therefore is highly restricted in its conformational space. Since IRECS assigns only one rotamer to this residue, Figure 4B displays a glutamic acid at the protein surface with two alternative conformations in the crystal structure, which both were predicted by the IRECS algorithm. Figure 4C shows the helix between residues 69 and 80 of the B-chain. The conformational space of the residues on the helix surface ranges between four and nine rotamers, whereas all side chains buried have only one rotamer left. This figure is also intended to visualize some obstacles of side-chain prediction. For example, Gln 74 has a χ_1 dihedral angle that is not represented by the rotamer library at all. IRECS selects the closest possible rotamers here, but the χ_1 differ all $\sim 60^\circ$. Arg 75 has three rotamers assigned, with two of them being far from the native conformation and only one rotamer with a similar χ_1 . This rotamer lies quite close to the native side-chain conformation, which gives the rotamer a similar ROTA interaction score, but its other dihedral angles are far away from those of the crystal structure. Lys 78 has two alternative conformations in the crystal structure, with an occupancy of 0.5 each. IRECS assigns eight rotamers to this residue, matching one of the alternative conformations quite well and missing the second alternative conformation completely. All buried residues are predicted optimally, considering the granularity of the backbone-dependent rotamer library.

Figure 3. Assignment of different rotamer numbers to specific residues (on X-axis). Average number of rotamers (on Y-axis) in protein models generated with IRECS and a rotamer density of two.

Figure 4. Rotamer ensembles with rotamer density two, predicted on 1EK6 (chain B), crystal structure of the human UDP-galactose 4-epimerase (Thoden et al. 2000). (Red) The predicted side-chain conformations; (blue) the crystal structure. Images were created with PyMOL (DeLano 2002). (A) Correct and unambiguous prediction of Arg 61. This side chain points toward the core of the protein. (B) Prediction of Glu 63 at the end of a β -strand. Both ambiguous conformations of the crystal structure were reproduced. (C) Helix on chain B, residues from *left* (69) to *right* (80): Asp, Gln, Gly, Ala, Leu, Gln, Arg, Leu, Phe, Lys, Lys, Tyr. The helix has been cut out, and only the surface of the remaining crystal structure is drawn.

Prediction of final structures with native-like rotamer subsets

Although the IRECS algorithm is an alternative to using the DEE algorithm for reducing the complexity of the side-chain prediction task, the performance of both algorithms is not directly comparable. The quality of the different DEE criteria is measured in terms of the degrees of freedom remaining after their application, asserting that no rotamer was removed that can belong to the GMEC. The IRECS algorithm does not meet this strict criterion but tries to maintain a most probable subset of rotamers during iteration. That is, IRECS can remove rotamers that are part of the GMEC, but it is unlikely to do so in early iterations. We therefore measure the performance of the IRECS algorithm by its ability to maintain a maximum number of native rotamers given a lower bound for the rotamer density (ρ_{rot}). The ability of the IRECS algorithm to generate rotamer ensembles of sizes that reflect the flexibility of the respective side chains enables us to predict side-chain conformations with higher accuracy at the cost of obtaining ambiguous conformations. The rotamer ensembles were predicted for all 160 structures of the test part of the protein set A, using the native backbone. First, we tested the efficiency of IRECS to predict rotamer ensembles that contain at

least one rotamer with a similar χ_1 angle to that of the respective reference side chain. In Figure 5 the average percentage of rotamer ensembles fulfilling this criterion is shown for different stages of the rotamer removal process, measured by the current overall density of rotamers, ρ_{rot} . Each of the curves describe this average percentage, once using the IRECS scoring function (ROTA + BBDEP), using only components of the scoring function or using randomized rotamer interaction scores (the shaded area under the curves was just added for better visibility).

IRECS performs quite well using ROTA and BBDEP independently. In the range between rotamer densities three and one, the performance of IRECS using BBDEP drops significantly, whereas IRECS using ROTA has a rather stable performance also in this region. When IRECS uses both the ROTA and BBDEP components of its scoring function, the performance is highest for all levels of rotamer density. However, the average percentage of side chains without a rotamer having a correct χ_1 angle in the ensemble more than doubles in the last steps of the algorithm, going from 6% for a rotamer density of two up to 15% for a rotamer density of one. This high chance of removing the correct rotamer in the last steps of the algorithm has three reasons. First, at the start, the rotamer ensembles contain many rotamers that can be clearly identified as useless because of their high effective rotamer energy. In later iteration steps, all rotamers of an ensemble have almost the same effective energy; thus, the correct rotamer is hard to identify. Second, the error range of the scoring function leads to incorrect removals of rotamers. Although we have not measured the error range of ROTA and BBDEP, the differences of the effective rotamer energy can decrease below the error range of the IRECS scoring function. The removal of the last rotamers from the ensembles thus may become nearly

Figure 5. Average percentage of residues with a correct χ_1 rotamer (Y-axis) for different rotamer densities (X-axis).

random. Last, this behavior can be explained with the flexibility of some side chains that require a whole ensemble of rotamers for an adequate representation of their conformation. Since the performance of IRECS is evaluated with only single conformations for all side chains here, the chance of selecting the same conformation as observed in the crystal structure is quite low for these side chains. This also shows that reducing the rotamer density beyond the amount of flexibility that some side chains require helps little in raising the accuracy of the generated protein conformation but worsens it by neglecting any conformational side-chain flexibility.

In a second analysis, these characteristics were examined by comparing the rotamer ensembles generated using IRECS with rotamer ensembles sparsely found in the 447 crystal structures of set B. Such sets are formed whenever multiple side-chain conformations can be assigned to the observed electron density during the crystallographic refinement, leading to alternative coordinates for a given side chain in the PDB file. We chose only those side chains whose conformations differ by $>40^\circ$ in one of their χ_1 and χ_2 dihedral angles for this evaluation. We did not use NMR ensembles for this analysis since these ensembles usually also contain multiple conformations of the backbone. Since IRECS can only handle a single conformation of the backbone, it is not meaningful to compare the output of IRECS with such NMR ensembles.

The protein models computed with IRECS for this evaluation all have a rotamer density of two. The number of rotamer ensembles that were compared between crystal structures and IRECS models and their respective sizes are listed in Table 5. The results of the comparison of rotamer ensembles computed with IRECS and those rotamer ensembles found in crystal structures are listed in Table 6. In this table the accuracy of IRECS is measured for all ensemble pairs with certain sizes by dividing the number of rotamers in the crystal ensemble having one or more matching rotamers in the ensemble of the IRECS model, respectively, by the number of all ensemble pairs of such sizes. Two rotamers are considered as matching here if their χ_1 and χ_2 angles differ by $<40^\circ$. For example, the upper-right cell of Table 6 shows

the percentage (92.3%) of ensemble pairs in which the ensemble derived from the crystal structure has three rotamers, the ensemble computed with IRECS has one rotamer, and this rotamer matches at least to one of the rotamers from the ensemble of the crystal structure. A visualization of such pairs of rotamer ensembles can be found in Figure 4. The overall results support the strategy of representing the conformation of flexible side chains with an ensemble of rotamers. IRECS maintains a high degree of accuracy by assigning more rotamers to those side chains that have multiple conformations also in the crystal structure. Whenever IRECS assigns only a single rotamer to a side chain, the prediction of this side chain is much more accurate (88.7%, 81.2%, and 92.3%) on average than for models with rotamer density one (71.6%) (see Table 4). If some side chains are harder to predict, IRECS adapts by assigning more rotamers to these side chains. In this way an overall good accuracy is maintained.

Runtime

The runtime of the IRECS algorithm primarily depends on the number of residues of the target protein, the average number of rotamers per side chain at the start, and the scoring scheme used for rotamer interactions, since this determines the number of interacting rotamer pairs. Our scoring scheme uses a 10 Å atom distance cutoff, which is in the lower/middle of other approaches that use cutoffs starting with 3.4 Å (SCWRL steric clash score) to 20 Å (RAPDF). Given that the average starting number of rotamers per side chain is n and the number of side chains to be predicted is p , a single IRECS iteration requires $O(n^2p)$. After removal of a rotamer from side chain s , $O(np)$ other rotamers can interact with the $O(n)$ rotamers of s and need to update their effective rotamer energies. If the minimization is continued until rotamer density one, $O(np)$ IRECS iterations must be performed. The computational complexity of a complete IRECS optimization is therefore $O(n^3p^2)$. This is the same complexity as for a single round of the Goldstein DEE where it tries to remove each rotamer (Gordon et al. 2003). In practice, the runtimes are not comparable since IRECS can remove many more rotamers than the DEE algorithm but does not protect the GMEC of the protein.

Discussion

The algorithm and the scoring function are presented in the Materials and Methods section, and their performance is discussed in the Results section. Thus here we focus on the discussion of potential areas of application for which IRECS will be especially useful and mention some factors limiting its performance.

Table 5. Total numbers of compared rotamer ensembles of different sizes

Number of rotamers in IRECS model	Number of rotamers in crystal structure		
	1	2	3
1	79,790	1947	13
2	30,761	2051	37
≥ 3	10,113	825	22

Table 6. χ_{1+2} accuracy of rotamer ensembles predicted with IRECS

Number of rotamers in ensemble computed with IRECS	(Number of matches between both ensembles)/(Number of rotamers in ensemble derived from crystal structure) (%)							
	1/1	1/2	2/2	$\geq 1/2$	1/3	2/3	3/3	$\geq 1/3$
1	88.7	81.2	3.9	85.1	92.3	0.0	0.0	92.3
2	87.1	38.7	54.7	93.4	21.6	75.7	0.0	97.3
≥ 3	85.0	28.7	64.5	93.2	4.5	9.1	81.8	95.4

The ability of IRECS to create rotamer ensembles that model the flexibility of each side chain is of special interest in all applications in which the number of rotamers per side chain has a large impact on the runtime of the modeling procedure. Such effects are commonly observed in docking applications that handle the flexibility of the ligand and receptor simultaneously (Leach and Lemon 1998; Claussen et al. 2001; Teodoro and Kavraki 2003; Gray 2006). One limiting factor of IRECS is the granularity of the rotamer library used. It has been observed that the accuracy of side-chain prediction often improves when using more detailed rotamer libraries than we have done (Xiang and Honig 2001; Shetty et al. 2003). However, using such libraries requires very large runtimes. We therefore decided to use the more coarse-grained backbone-dependent rotamer library (Dunbrack and Cohen 1997), which represents a suitable tradeoff between runtime and detail of representation of the conformational space. Using the rotamers of the BBDEP and their associated conditional probabilities, we are able to curb the combinatorial explosion effectively and to find a good first guess for all side-chain conformations in the discrete rotamer space. Further refinement techniques that work on single protein conformations like standard energy minimization by steepest gradient descent are then free to leave the discrete rotamer conformation space and to optimize the protein conformation in continuous space. The quality of the final structure then relies first on the ability of the side-chain prediction tool to sample discrete conformational space and second on the accurate modeling of the energetic landscape around the discrete solution by the refinement method. As Xiang and Honig (2001) have already stated, we also conclude that the quality of the energy functions remains the dominant factor for side-chain prediction.

In deriving ROTA, we followed the strategy of enforcing the detection of structural features that are highly discriminatory between correct structures and explicit side-chain decoys. Structural features that are equally common to both correct and decoy structures are automatically filtered out and do not have any influence on the score of a query structure. Using side-chain decoys as a structural basis for the reference state instead of using a theoretical reference state renders ROTA much more

sensitive toward any errors arising from false side-chain prediction, at the cost of lower performance in identifying other modeling errors. The procedure of generating the reference state shown here is applicable to other scoring and modeling applications, whenever the space of relevant modeling errors can be simulated realistically.

Since our and all other approaches for side-chain prediction published so far still have insufficient accuracy, both scoring techniques based on knowledge-based statistical potentials and force fields have to be improved further. Whereas the accuracy of force-field scoring can be improved primarily through more detailed and complex modeling of the protein environment (especially the solvent), knowledge-based scoring schemes allow for treating many complex structural features implicitly. We therefore regard the knowledge-based scoring schemes as the tools of choice for further efforts toward increasing the accuracy of side-chain prediction. Possible improvements include the use of an extended feature space used for scoring. In our view, the publicly available data on protein structure offer such a dense description of protein conformational space that it should be possible to derive scoring functions with much more detailed features than only pairwise atom distances. To further improve the prediction accuracy of IRECS, we plan to include conformational information about side-chain conformations in homologous proteins.

Acknowledgments

This work has been performed in the context of the BioSapiens Network of Excellence (EU contract no. LSHG-CT-2003-503265) and the DFG project KA 1804/1-2. We thank Francesco Domingues for sharing his knowledge on quality measurement of crystallographic models. We also thank the research groups of R.L. Dunbrack Jr., B. Honig, M. Karplus, M. Levitt, J. Moult, D.C. Richardson, and R. Samudrala for publicly sharing and maintaining all the data and computational services required for this study.

References

- Bower, M.J., Cohen, F.E., and Dunbrack, R.L. 1997. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: A new homology modeling tool. *J. Mol. Biol.* **267**: 1268–1282.
- Brooks, B.R., Brucoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M. 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**: 187–217.

- Canutescu, A.A., Shelenkov, A.A., and Dunbrack, R.L. 2003. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* **12**: 2001–2014.
- Claussen, H., Buning, C., Rarey, M., and Lengauer, T. 2001. FlexE: Efficient molecular docking considering protein structure variations. *J. Mol. Biol.* **308**: 377–395.
- DeLano, W.L. 2002. *The PyMOL molecular graphics system*. DeLano Scientific, San Carlos, CA.
- Desmet, J., Demaeyer, M., Hazes, B., and Lasters, I. 1992. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **356**: 539–542.
- Dunbrack, R.L. and Cohen, F.E. 1997. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* **6**: 1661–1681.
- Goldstein, R.F. 1994. Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys. J.* **66**: 1335–1340.
- Gordon, D.B., Hom, G.K., Mayo, S.L., and Pierce, N.A. 2003. Exact rotamer optimization for protein design. *J. Comput. Chem.* **24**: 232–243.
- Gray, J.J. 2006. High-resolution protein–protein docking. *Curr. Opin. Struct. Biol.* **16**: 183–193.
- Holm, L. and Sander, C. 1992. Fast and simple Monte Carlo algorithm for side chain optimization in proteins: Application to model building by homology. *Proteins* **14**: 213–223.
- Hubbard, S.J., Campbell, S.F., and Thornton, J.M. 1991. Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. *J. Mol. Biol.* **220**: 507–530.
- Koehl, P. and Delarue, M. 1994. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.* **239**: 249–275.
- Koehl, P. and Delarue, M. 1996. Mean-field minimization methods for biological macromolecules. *Curr. Opin. Struct. Biol.* **6**: 222–226.
- Leach, A.R. and Lemon, A.P. 1998. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins* **33**: 227–239.
- Liang, S. and Grishin, N.V. 2002. Side-chain modeling with an optimized scoring function. *Protein Sci.* **11**: 322–331.
- Lovell, S.C., Davis, I.W., Arendall, W.B., Bakker, P.I.W.D., Word, J.M., Prisant, M.G., Richardson, J.S., and Richardson, D.C. 2003. Structure validation by C_{α} geometry: ϕ, ψ and C_{β} deviation. *Proteins* **50**: 437–450.
- MacKerell, A.D., Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., et al. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**: 3586–3616.
- MacKerell, A.D., Feig, M., and Brooks, C.L. 2004. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* **25**: 1400–1415.
- Mendes, J., Baptista, A.M., Carrondo, M.A., and Soares, C.M. 1999. Improved modeling of side-chains in proteins with rotamer-based methods: A flexible rotamer model. *Proteins* **37**: 530–543.
- Peterson, R.W., Dutton, P.L., and Wand, A.J. 2004. Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library. *Protein Sci.* **13**: 735–751.
- Pierce, N.A. and Winfree, E. 2002. Protein design is NP-hard. *Protein Eng.* **15**: 779–782.
- Samudrala, R. and Levitt, M. 2000. Decoys ‘R’ Us: A database of incorrect conformations to improve protein structure prediction. *Protein Sci.* **9**: 1399–1401.
- Samudrala, R. and Moul, J. 1998a. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.* **275**: 895–916.
- Samudrala, R. and Moul, J. 1998b. Determinants of side chain conformational preferences in protein structures. *Protein Eng.* **11**: 991–997.
- Shetty, R.P., Bakker, P.I.W.D., DePristo, M.A., and Blundell, T.L. 2003. Advantages of fine-grained side chain conformer libraries. *Protein Eng.* **16**: 963–969.
- Teodoro, M.L. and Kavraki, L.E. 2003. Conformational flexibility models for the receptor in structure based drug design. *Curr. Pharm. Des.* **9**: 1635–1648.
- Thoden, J.B., Wohlers, T.M., Fridovich-Keil, J.L., and Holden, H.M. 2000. Crystallographic evidence for Tyr 157 functioning as the active site base in human UDP-galactose 4-epimerase. *Biochemistry* **39**: 5691–5701.
- Tosatto, S.C.E. 2005. The Victor/FRST function for model quality estimation. *J. Comput. Biol.* **12**: 1316–1327.
- Wang, G. and Dunbrack, R.L. 2003. PISCES: A protein sequence culling server. *Bioinformatics* **19**: 1589–1591.
- Wang, G. and Dunbrack, R.L. 2005. PISCES: Recent improvements to a PDB sequence culling server. *Nucleic Acids Res.* **33**: W94–W98.
- Xiang, Z. and Honig, B. 2001. Extending the accuracy limits of prediction for side-chain conformations. *J. Mol. Biol.* **311**: 421–430.
- Xie, W. and Sahinidis, N.V. 2006. Residue-rotamer-reduction algorithm for the protein side-chain conformation problem. *Bioinformatics* **22**: 188–194.