NIH Public Access
Author Manuscript
Proteins. Author manuscript; available in PMC 2009 September 14.

Published in final edited form as:
Proteins. 2007 March 1; 66(4): 814–823. doi:10.1002/prot.21099.

Prediction of Side-Chain Conformations on Protein Surfaces

Zhexin Xiang[1,*], **Peter J. Steinbach**[1], **Matthew P. Jacobson**[2], **Richard A. Friesner**[3], and **Barry Honig**[4]

[1] Center for Molecular Modeling, Center for Information Technology, National Institutes of Health, Bethesda, Maryland 20892-5624

[2] Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, California 94143-2240

[3] Department of Chemistry, Columbia University, New York, New York 10027

[4] Howard Hughes Medical Institute, Center for Computational Biology and Bioinformatics and Department of Biochemistry and Molecular Biophysics, Columbia University, 1130 St. Nicholas Avenue, Room 815, New York, New York 10032

Abstract

An approach is described that improves the prediction of the conformations of surface side chains in crystal structures, given the main-chain conformation of a protein. A key element of the methodology involves the use of the colony energy. This phenomenological term favors conformations found in frequently sampled regions, thereby approximating entropic effects and serving to smooth the potential energy surface. Use of the colony energy significantly improves prediction accuracy for surface side chains with little additional computational cost. Prediction accuracy was quantified as the percentage of side-chain dihedral angles predicted to be within 40° of the angles measured by X-ray diffraction. Use of the colony energy in predictions for single side chains improved the prediction accuracy for χ_1 and χ_{1+2} from 65 and 40% to 74 and 59%, respectively. Several other factors that affect prediction of surface side-chain conformations were also analyzed, including the extent of conformational sampling, details of the rotamer library employed, and accounting for the crystallographic environment. The prediction of conformations for polar residues on the surface was generally found to be more difficult than those for hydrophobic residues, except for polar residues participating in hydrogen bonds with other protein groups. For surface residues with hydrogen-bonded side chains, the prediction accuracy of χ_1 and χ_{1+2} was 79 and 63%, respectively. For surface polar residues, in general (all side-chain prediction), the accuracy of χ_1 and χ_{1+2} was only 73 and 56%, respectively. The most accurate results were obtained using the colony energy and an all-atom description that includes neighboring molecules in the crystal (protein chains and hetero atoms). Here, the accuracy of χ_1 and χ_{1+2} predictions for surface side chains was 82 and 73%, respectively. The root mean square deviations obtained for hydrogen-bonding surface side chains were 1.64 and 1.81 Å, with and without consideration of crystal packing effects, respectively.

Keywords

side-chain prediction; entropy; conformational mobility; colony energy; protein structure

[*] Correspondence to: Zhexin Xiang, Center for Molecular Modeling, Center for Information Technology, National Institutes of Health, Building 12A Room 2051, 12 South Drive, Bethesda, MD 20892-5624. E-mail: xiangz@mail.nih.gov.

[*] This article is a US Government work and, as such, is in the public domain in the United States of America.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

INTRODUCTION

The prediction of side-chain conformations, given the protein backbone conformation, is an important problem in areas such as protein design,[1] protein–protein association,[2] protein–DNA interactions,[3] homology modeling,[4] and flexible ligand docking. The problem has been studied for many years and has witnessed steady technological advances and the development of widely used programs such as SCWRL.[5] The prediction of the conformations of buried side chains is simplified by packing restraints, and recent papers have reported accuracies of about 0.7 Å root mean-square deviation (RMSD),[6–9] which is close to the experimental accuracy for an X-ray structure at 2.0-Å resolution.[10–12] Success has been possible because the combinatorial problem for buried side chains is not severe,[13] given excluded-volume effects, and thus large rotamer libraries can be used to sample conformational space.[6] However, the prediction of conformations for side chains on protein surfaces (side chains more than 50% exposed), especially for polar residues, remains largely unsolved, and has been essentially ignored in the literature. Difficulties in the prediction of surface side-chain conformations reflect the need to estimate side-chain mobility and to account for interactions with the environment.

Surface side chains generally have fewer geometrical restrictions than buried side chains and are much more mobile, complicating attempts at prediction. Although entropic effects have been of wide interest in the literature,[14–17] few attempts have been made to account for side chain mobility in prediction algorithms, which are mostly based on the frequency of side chain rotamers in the Protein Data Bank (PDB)[8,18] or mean-field methods.[19] Our treatment of mobility makes use of the so-called colony energy, which has been applied previously to the problem of loop prediction.[20] The sampled conformations are evaluated according to their structural similarity and energetic distribution, and a phenomenological term is used to energetically reward structures with a large number of neighbors. Thus, entropic effects are approximated by favoring conformations in frequently sampled regions of conformational space.

Although conformations of surface side chains are often difficult to resolve, a large number of them are seen by experiment to adopt unambiguous conformations, especially those involved in salt bridges or hydrogen bonds. The superposition of an ensemble of NMR models can indicate the existence of unambiguous conformations of surface side chains.[21,22] Such conformations generally have relatively low temperature factors in X-ray structures.[23] Furthermore, since we define surface side chains here as those more than 50% exposed, most surface side chains experience some motional hindrance, and their conformations are predictable to some extent. One example is in the vicinity of protein binding sites where many residues have intermediate solvent accessibility but also experience some steric restraints.[24]

As mentioned earlier, there appears to be little room for improvement in predicting side-chain conformations for core residues given the protein backbone, and even the most recent programs[9] yield results for buried residues that are essentially identical to those obtained previously.[6] However, the prediction of the conformations of surface side chains has not been studied extensively to date. Here, we analyze various factors related to the accuracy of such predictions and develop a method to estimate side-chain mobility. Our results suggest that the predictions made for surface side chains are in good agreement with experiment if conformational mobility and crystal packing effects are reasonably accounted for.

METHODS

Conformational Energy

Using the program SCAP,[6] the conformational energy of a side chain was approximated simply as:

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$E = E_{\mathrm{vdw}} + E_{\mathrm{torsion}} + E_{\mathrm{hbond}} \quad (1)$$

$$\begin{aligned} &E_{\mathrm{hbond}} = \min(0, \mathrm{S}[-16+12\Omega]\cos(\theta_{\mathrm{DHA}}) \\ &\cos(1.5\,\theta_{\mathrm{HAC}})/d_{\mathrm{HA}}^{3}), \\ &\text{if } 2\text{Å} < \mathrm{d}_{\mathrm{HA}} < 3\text{Å}, \theta_{\mathrm{DHA}} > 90°, \text{and } \theta_{\mathrm{HAC}} > 60°; \\ &\text{else } E_{\mathrm{hbond}} = 0. \end{aligned} \quad (2)$$

Torsional energies, E_{torsion}, and van der Waals energies, $E_{\mathrm{vdw,}}$ are calculated using either CHARMM22 or CHARMM19 parameters and expressions. Instead of evaluating hydrogen bond energies using statistical methods,[25] we parameterized the empirical expression in Eq. (2), obtained for the training set of proteins, to account for both hydrogen bond geometry and solvent accessibility. Ω is the fractional solvent accessible surface area (SASA) of the side chain, i.e., the ratio of the SASA in the protein to the SASA of the same conformation for a lone side chain in solution. D is the hydrogen donor, H is the polar hydrogen, A is the hydrogen acceptor, and C is the carbon atom bonded to A. θ_{DHA} and θ_{HAC} are the angles defined by the coordinates of the respective atoms, and d_{HA} is the distance between atoms H and A. The scale factor S was set to 1.5 for hydrogen bonds between oppositely charged residues[26] and 1.0 for all other hydrogen bonds. Although the value of θ_{HAC} depends on whether the atomic orbital of the acceptor is sp^2 or sp^3, the θ_{HAC} angle is nevertheless close to 120°. As the rotamer library is discretized, we relaxed the standard requirement that θ_{HAC} should be larger than 90°.[27] E_{hbond} is defined to assume its minimum value when d_{HA} is 2 Å, θ_{DHA} is 180° and θ_{HAC} is 120°. Constants in Eq. (2) were chosen so that the minimum E_{hbond} values for completely buried and completely exposed side chains are −2 and −0.5 kcal/mol, respectively, representative of experimental data for hydrogen bonds.[28]

We did not include covalent bond energy, since our coordinate rotamer libraries were compiled from high-resolution structures that have already been subjected to energy refinement. Finite-difference Possion–Boltzmann (FDPB) calculations were performed with the Delphi program[29] to compare the effects of the colony energy with solvation effects. In these FDPB calculations, a 1-Å lattice was used, the hydrogen-bonding term was turned off, and the protein interior and exterior were assigned dielectric constants of 4 and 80, respectively. In all other calculations, atomic partial charges were set to zero, and the contribution of electrostatics to hydrogen bonding was included implicitly in Eq. (2). For the 20 proteins in the training set, 165 of the 198 salt bridges conform to standard hydrogen-bond geometry.

Colony Energy

Here we adapt the colony energy[20] for application to side-chain prediction. For a given residue of N rotamers, the colony energy of rotamer i, G_i, is calculated with the following equation:

$$G_i = -\mathrm{RT}^{*}\ln\left[\sum_j \exp(-E_j/(\mathrm{RT}) - \beta(\mathrm{RMSD}_{ij}/\mathrm{RMSD}_{\mathrm{avg}})^{\gamma})\right] \quad (3)$$

where R is the gas constant, T is absolute temperature and E_i is the conformational energy [Eqs. (1) and (2)] of the rotamer conformation i of a given residue. The sum is for all rotamers of a given residue, i.e., j ranges from 1 to N, including i. RMSD_{ij} is the heavy-atom root mean square distance between rotamers i and j. $\mathrm{RMSD}_{\mathrm{avg}}$ is the average of RMSD_{ij} between any two rotamers for a given residue. The symmetry in Asp, Glu, Phe, Tyr, and Arg was taken into account when calculating RMSD values. Results obtained with the training data set for single side-chain prediction were not sensitive to the values of the parameters β and γ, which were

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

set to −ln(1/2) and 1, respectively. However, when predicting loop conformations,[20] a value of 3 was used for γ. The colony-energy parameters β and γ are determined empirically. Here, they were chosen based on single side-chain prediction. The ranges of conformational energies and three-dimensional structures sampled in a particular application call for the use of a γ value that balances the conformational-energy and RMSD-based factors appropriately.

Equation (3) is of the familiar form $G_i = -\mathrm{RT} \ln Z_i$, but Z_i differs from a true partition function in important ways. Z_i is a function of the single microstate *i*; it does not describe a thermodynamic state or the entire system. The Z_i sum is only over the conformations actually sampled (e.g., all rotamers in the current application), and the Boltzmann-like factor associated with a given conformation is weighted by a factor that increases with the similarity of that conformation to conformation *i*. This weight ranges from near zero for structures distant from conformation *i* to unity for conformation *i* itself. Structures of low conformational energy reduce the colony energy of neighboring structures of higher conformational energy, and thus, the colony-energy surface is smoother than the conformational-energy surface.[20] Moreover, conformations sampled along with many similar conformations have more terms in the Z_i sum with large weights, reducing their colony energy. In this way, Eq. (3) approximates entropic effects by favoring those conformations found in regions of configuration space that are most frequently visited.

Conformational Sampling

We have adopted a simple conformational sampling strategy for side-chain prediction similar to that used previously.[6] In brief, side-chain predictions were carried out on fixed polypeptide backbones. For a given protein, a number of calculations were performed starting from different initial conformations. The first calculation was begun with each side chain in its lowest-energy rotamer. The next calculations used initial conformations obtained from the previous runs in which side chains with negative or positive conformational energies were repositioned in randomly chosen rotamers with a probability of 30 or 70%, respectively. Given an initial conformation, energy minimization was performed one residue at a time, while all other residues were kept fixed. The minimization procedure simply tests all rotamers for a given residue and generally picks the one lowest in colony energy. However, if the rotamer of lowest conformational energy participates in a hydrogen bond, the conformational energy is used instead of the colony energy because entropic effects generally do not favor hydrogen bonding, and an accurate balance between hydrogen-bond energy and entropy is difficult to achieve in a simplified force field. The minimization proceeded from the first residue to the last and was repeated until all the side-chain conformations retained the same rotamer upon further iteration. The final prediction was the conformation of lowest colony energy found in any of the calculations. Previously,[6] the final results were obtained by iterating independently on 120 initial conformations. Here, prediction was typically based on a single starting conformation, because the conformations of neighboring surface side chains are not highly correlated.

Rotamer Library

The rotamer library used in this article was built from a large data set, i.e., 648 protein chains (297 chains used in the previous prediction) taken from a 2001 culled PDB list (http://dunbrack.fccc.edu/). Each of the 648 proteins has resolution better than 1.6 Å with no sequence identity higher than 25% between any two chains. For each of the proteins, the hydrogen atoms were positioned using the WHATIF package. The general approach is similar to that used previously,[6] except that here we always start from the most often observed rotamers to trim off redundant side-chain conformations, which produces a more representative rotamer library. Two rotamer libraries were created, with 40° and 10° resolution. The 40° rotamer library was created using a cutoff of 40° for all torsion angles of a side chain. In the 10° library, a 10° cutoff was used for the first two side-chain torsion angles, 15° for the third, and 20° cutoff

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

for all other side-chain torsions, except that a 15° cutoff was used for the second side-chain torsion angle of Ser and Thr side chains that determines the position of the polar hydrogen. Two side chains are assigned to one rotamer if all the torsion angles are within the cutoff range of each other. Based on the 648 protein chains, the total number of rotamers for the 20 amino acids is 831 and 6737, respectively, for the 40° and 10° rotamer libraries. However, only 623 and 4447 rotamers are required to represent 96% of side chain conformations in the two libraries, half of which come from two residues, Arg and Lys. Here, the rotamer library of 96% representation was used. In the following sections, the 10° library of 4447 rotamers was used unless otherwise specified.

Protein Sets and Accuracy Analysis

Three sets of proteins were used: a training set and two test sets that facilitate comparison of our work with SMOL,[8] SCCOMP,[9] and SCWRL.[18] We use a backbone-independent rotamer library, whereas SMOL, SCCOMP, and SCWRL use a backbone-dependent rotamer library. The rotamer libraries used in our work were compiled based on proteins released prior to 2002. Members of both the training and testing sets are single-chain proteins with sequence similarity cutoff of 25% and resolution better than 1.6 Å (except 1.8 Å for test set 2). The training set includes 20 proteins randomly selected from the culled PDB list (http://dunbrack.fccc.edu/): 1ra9, 1tca, 1hfc, 1cvl, 3nul, 1cbn, 1cex, 5pti, 1ixh, 2pth, 5p21, 1aho, 3lzt, 1ctj, 1igd, 7rsa, 1aac, 1eca, 1plc, 1rcf. To remove the influence of the native rotamer conformation on the prediction accuracy, in test set 1, we restricted our selection to protein structures released in 2002, with less than 25% sequence similarity to any proteins released before 2002. The influence of rotamer library on prediction accuracy has been discussed.[6] Generally speaking, prediction accuracy increases if the native side-chain conformation is included in the rotamer library, particularly when only a few rotamers exist for a residue. Twenty proteins were randomly chosen for test set 1 according to the above criteria: 1eaz, 1gv9, 1iqz, 1is3, 1jm1, 1k4n, 1kng, 1kr4, 1ky3, 1lmi, 1mb3, 1mvo, 1n0r, 1gxu, 1jni, 1jo8, 1k7j, 1kqr, 1kr7, 1lf7. Test set 2 includes all proteins that are single chain, better than 1.8-Å resolution, have no sequence similarity higher than 25%, and were released between January 2003 and April 2006. There are 162 proteins in test set 2. Hydrogen atoms were placed on the proteins of the training and test sets with the WHATIF package.[30] Side chains were subsequently removed, leaving only backbone heavy atoms and protons. The structures are described in Table I.

In analyzing the results, a dihedral angle was considered correctly predicted if it was within 40° from that of the native structure. χ_1 indicates the percentage of side chains for which the first torsion angle was correctly predicted, and χ_{1+2} represents the percentage of side chains for which both the first and the second torsion angles were correctly predicted. We also use RMSD to indicate the heavy-atom root mean square deviation of the predicted conformation from that determined by X-ray diffraction. Protein residues with fractional SASA of 0.5 or more were classified as surface residues. Core residues were defined as those with fractional SASA less than 0.1, and the remaining residues were defined as partially buried. With these definitions, roughly one third of the residues can be classified in each category, as shown in Table I. All comparisons were made to crystallographic coordinates.

RESULTS AND DISCUSSION

Single Side-Chain Prediction

All-atom model—The prediction of the conformation of individual side chains while all others are held fixed at their native conformations provides a test of the energy function used by greatly reducing the challenge of conformational sampling. Results obtained for the training set are shown in Table II. The first two rows characterize the use of the conformational energy [Eqs. (1) and (2)] in the case of the united-atom and the all-atom model, respectively. As shown

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

in the table, the all-atom model is more accurate than the united-atom model, especially for surface side chains. When side chains are partially buried (50–90%), the all-atom model is 4% more accurate than the united-atom model in terms of χ_{1+2}; for surface side chains (more than 50% exposed), the difference between the all-atom and the united-atom model increases to 2% and 9% for χ_1 and χ_{1+2}, respectively. The additional hydrogen atoms present in the all-atom model provide a more detailed structural description but can result in a doubling of the computational cost (Table II).

Colony energy and solvation effects—The effect of the colony energy on the prediction accuracy is evident by comparing the first two rows with the second two rows in Table II. There is a significant improvement in prediction accuracy for surface side chains but not for side chains in the core. The latter observation is not unexpected, as the conformations of core residues are determined primarily by packing effects. In the case of the united-atom model (rows 1 and 3), the colony energy increased χ_1 and χ_{1+2} for surface side chains from 65 and 40% to 74 and 59%, respectively. For the all-atom model (rows 2 and 4), use of the colony energy had a smaller, but still significant impact on the prediction accuracy, improving χ_1 and χ_{1+2} for surface residues from 67 and 49% to 73 and 60%. For partially buried side chains, the effect of the colony energy was not as significant as for surface residues. In the united atom model, use of the colony energy improved the RMSD of the prediction slightly, from 1.27 to 1.13 Å, while in the all-atom model, the colony energy did not result in noticeable improvement.

The results in Table II suggest that the use of the colony energy and solvation energy (FDPB) does not seem to constitute significant improvement over the use of colony energy alone. The colony energy may capture some characteristics of solvation for surface residues, since the motion of polar residues toward the solvent favored by a good solvation model is also likely to be favored by a term approximating entropic effects.

As most protein active sites involve polar residues with significant solvent exposure, the correct placement of these residues is especially important. Table III shows the prediction accuracy for polar residues that are exposed to the solvent. The colony energy is particularly effective in improving prediction accuracy for this group of side chains. As shown in the first row (united-atom model) in which the colony energy was not applied, the prediction accuracy for polar residues on the surface is only 60% and 38% for χ_1 and χ_{1+2}, respectively. However, when the colony energy is applied (row 3), the accuracy increases to 71% and 59%, respectively (0.7-Å improvement in RMSD). The improvement obtained with the colony energy is a little less for the all-atom than for the united-atom model, but it is still significant, increasing the accuracy about 7% and 11% for χ_1 and χ_{1+2}, respectively. Table III also indicates that the prediction accuracy for residues involved in hydrogen bonds is higher than that for other residues. For example, as shown in the last row, the average accuracy for polar residues on the surface is only 2.17 Å, while for hydrogen-bonded residues the accuracy is 1.72 Å. Comparing Tables II and III, it is clear that the conformations of polar residues are, in general, relatively difficult to predict.

As demonstrated in Figure 1, the colony energy [Eq. (3)] often discriminates native from nonnative conformations more effectively than does the simplified force field alone [Eqs. (1) and (2)]. Figure 1 plots energy vs. RMSD for all side-chain rotamers available for Arg96 of 1cex. Over 70% of the side chains of Arg96 are exposed to solvent. The conformation of lowest energy has an RMSD of 4.62 Å [Fig. 1(a)]; the conformation of lowest colony energy (−18.4 kcal/mol) has an RMSD of 1.86 Å. In addition, the rotamer in best agreement with experiment (RMSD of 0.5 Å) has colony energy of −18.2 kcal/mol. Note also that the colony-energy surface is much smoother than the conformational-energy surface. Figure 2 depicts another example showing the enhanced smoothness of the colony-energy surface and an improved prediction obtained using the colony energy. Lys29 is a residue on the surface of protein 1aac. The rotamer

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

of lowest conformation energy has an RMSD of 4.69 Å, while the rotamer with the lowest RMSD ranked 254th [Fig. 2(a)]. By contrast, the conformation of lowest colony energy has an RMSD of only 0.93 Å [Fig. 2(b)]. Figures 1(b) and 2(b) are both characterized by a smooth envelope with a single minimum, suggesting that use of the colony energy effectively smoothes the potential energy over all conformations accessible to the given side chain.

Complete Side-Chain Prediction

Rotamer library and conformation sampling—In the previous section, results were presented for single side-chain prediction in which there is no combinatorial problem. Of course, prediction quality will degrade when the conformations of all side chains are predicted simultaneously. Table IV summarizes these results obtained for the training set for both the united and all-atom force fields. Comparing the second rows of Tables II and IV, which report the use of the same potential, there is essentially no change in RMSD for surface side chains (0.01 Å) simply because, without accounting for solvation and entropic effects, predictions for surface side chains are little better than random. For core residues, the difference between single side-chain and all side-chain prediction does not depend sensitively on the energy expression because the result depends primarily on packing effects. However, it does depend on sampling. This is evident by noting the improvement for core residues in the fifth row, where 120 initial conformations were used, relative to the fourth row where only one starting conformation was used. In our previous work, we emphasized the importance of using a detailed rotamer library to obtain an accurate prediction of core residues.[6] This requirement is evident in the fourth and last two rows in Table IV, where there is clearly a significant degradation in accuracy (from 0.84 to 1.43 Å) when the 40° rotamer library is used (623 rotamers in total). In contrast, the size of the rotamer library has almost no influence on surface side-chain prediction. The last two rows also indicate that the effect of the colony energy is less significant when the 40° rotamer library is used. In addition, use of multiple starting conformations does not seem to be necessary for prediction for surface side chains, which are not very restricted by packing effects. As shown in the fourth and fifth rows, the use of 120 starting conformations improved prediction accuracy only about 0.05 Å relative to use of a single starting conformation.

Colony energy—As can be seen from Table IV and Figure 3, use of the colony energy has a significant effect on prediction accuracy. The colony energy increases χ_{1+2} for surface side chains by 15% for the united-atom model (rows 1 and 3) and 7% for the all-atom model (rows 2 and 4). As is the case for single side-chain prediction, the colony energy has essentially no effect on core residues (Table IV).

Table V summarizes the predictions made for polar residues of the training set. As in the case of single side-chain prediction, side chains involved in hydrogen bonding on the protein surface are predicted more accurately than polar residues on average. In the all-atom model (row 2), the prediction accuracy for polar residues on the surface is 69% and 56% for χ_1 and χ_{1+2}, while it is 73% and 59% for residues participating in hydrogen bonds. As shown in the last entry of rows 1 and 2 in Table V, the colony energy improved the prediction for hydrogen-bonded residues about 6% in terms of χ_{1+2}, while for single side-chain prediction (last column of rows 3 and 4 in Table III) the corresponding number is only 2%. If the colony energy can increase the prediction accuracy for polar residues on average, as is shown in Table V, it increases the likelihood that correct hydrogen bonds are formed.

Temperature factor and crystal packing—Figure 4 shows the prediction accuracy for surface residues in different temperature-factor ranges. The average temperature factor is about 25.0 for surface side chains, 16.6 for partially buried side chains, and 11.5 for core side chains (data not shown). Generally, prediction accuracy, as it is commonly measured, decreases with

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

increasing temperature factor. Large temperature factors are assigned to atoms that are not well localized by the diffraction experiments, including protein atoms that are particularly mobile, e.g., surface residues. For side chains in these mobile regions, the native state may involve several well-populated rotamers. Thus, it is not surprising that when the temperature factor is between 0 and 10, the RMSD of the predicted surface side-chain conformations is 1.50 Å; when the temperature factor exceeds 30, the RMSD is about 2.56 Å (Fig. 3).

It is essentially impossible to predict surface side chains as accurately as core side chains without proper consideration of crystal packing (if the prediction is compared with an X-ray structure). Surface side chains may also interact in the crystal lattice with water molecules, ions, small ligands, etc. Neglect of these small molecules will introduce additional errors in the prediction. The third last row of Table IV shows results which take small molecules into account. All atoms belonging to these molecules are referred to as hetero atoms to indicate that they do not belong to any standard amino acids. The hetero atoms were assigned their three-dimensional coordinates directly from the PDB file and treated as hard balls in the prediction with radii taken as those of similar CHARMM atom types (only van der Waals interactions with the protein side chains were considered). When hetero atoms are accounted for, prediction accuracy for surface side chains improves by about 2–3%. As expected, hetero atoms have no significant effect on prediction accuracy for core residues.

As we have reported previously,[7] significant improvement in side-chain prediction can be achieved if crystal packing effects are taken into account. Here, the crystal lattice was generated based on structural information stored in the PDB file. Any crystal atoms more than 6 Å away from the protein under consideration were ignored. The crystal atoms were simply treated as hetero atoms that restrict the conformational freedom of side chains. As shown in the last row of Table V, the inclusion of crystal packing improves the prediction accuracy for surface residues from 71% and 57% for χ_1 and χ_{1+2} to 83% and 72%, respectively, but has no effect on core residues. With the inclusion of hetero atoms and crystal packing, surface side chains are treated more like partially buried side chains, where the colony energy is expected to have a less significant effect on improving the accuracy (Table V).

Predictions on the Test Sets

Tables VI and VII show prediction results for the first test set. It is clear that the colony energy significantly improves side-chain prediction for this set as well. For surface residues, using the united-atom potentials, prediction accuracy is improved for χ_1 and χ_{1+2} respectively from 65 and 41% to 69 and 52%, while in the all-atom model, the prediction accuracy increases from 66% and 47% to 72% and 55%. When hetero atoms are included, prediction accuracy for χ_1 and χ_{1+2} further improves to 78% and 65%. The most accurate results are obtained when crystal packing effects are taken into account and the colony energy is used. In this case, χ_1 and χ_{1+2} predictions are improved to 82% and 73%, respectively, for all surface side chains and 87% and 75% for side chains that participate in hydrogen bonds (Tables VI and VII).

Comparison With Other Programs

Tables VI and VIII allow comparison of the results obtained here with the program SCAP with those obtained using other programs. Comparing the fourth row of Table VI (hetero atoms and crystal effects not accounted for) with SMOL,[8] SCCOMP,[9] and SCWRL,[18] SCAP produces the most accurate results, especially as measured in terms of RMSD, for core and surface side chains. The very detailed rotamer library used in SCAP likely accounts for the quality of the results obtained for core residues. Although SMOL, SCWRL, and SCCOMP each use the same backbone-dependent rotamer library with half a million records, fewer rotamers are considered for a given residue in a given backbone conformation than are considered for the same residue using SCAP, which considers all rotamers for that residue regardless of its backbone

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

conformation. For surface residues, which have relatively few packing interactions to restrain their motion, the use of a detailed rotamer library is less important, but the treatment of entropic effects, e.g., with the colony energy, becomes critical. For partially buried residues, SMOL performed the best, perhaps because of its more extensive conformation sampling and optimized scoring function.

On a 300-MHz SGI workstation, the average CPU time required per protein was 3, 11,700, 960, and 361 seconds for SCWRL, SMOL, SCCOMP, and SCAP with the colony energy, respectively. SCWRL clearly offers an excellent combination of speed and accuracy. SCAP is slower than SCWRL, but when used with a large rotamer library and all-atom models, it offers a somewhat improved level of prediction accuracy.

DISCUSSION

In this study, we analyzed the factors that affect the prediction of the conformations of surface side chains. This prediction is more complicated than the prediction for buried side chains, because surface side chains are less restricted by steric effects. Indeed, accounting for the effects of hetero atoms and crystal packing effects adds restraints that lead to a significant improvement in prediction accuracy. Similarly, results are better for side chains that participate in intraprotein hydrogen bonds than for those that do not, since hydrogen bond formation provides an additional structural restraint. The colony energy is shown to improve prediction accuracy, especially for side chains with relatively few conformational restraints. Typically, for core side chains, the conformational energy, as calculated with Eqs. (1) and (2), is sufficient to favor the rotamer that optimally packs with neighboring residues. Apparently, for surface side chains, incorrect conformations are not as readily discounted by the simplistic conformational energy alone, and the effective averaging over alternative rotamers inherent in the colony energy proves advantageous. Specifically, by favoring conformations found in the most sampled regions of conformational space, the colony-energy term introduces an entropy-like bias into the ranking of structures that is computed economically. When applied to loop modeling[20] and to the prediction of side-chain conformations, the colony energy results in a smooth correlation between energy and deviation from the experimentally determined structure.

Recently, the colony energy was used to refine homology models submitted to the CASP5[31] and CASP6 blind comparative-modeling tests. Its application elsewhere, e.g., to protein folding or protein–ligand docking, may thus prove advantageous. The utility of the colony energy depends on the empirical choice of the parameters β and γ in Eq. (3). These parameters determine the relative importance ascribed to a structure’s conformational energy and its similarity to all the other structures that are sampled. Thus, for each application of the colony energy, it is necessary to find a test set that can be used to tune the values of β and γ.

As surface side chains interact directly with water molecules, ions, and neighboring molecules, future improvements will likely result from explicit incorporation of these molecules. One important factor that has not been considered here involves the protonation state of different side chains. For example, the conformations of His side chains were poorly predicted (Fig. 3), presumably because a single protonation state (both nitrogen atoms protonated) was used in this work. Programs such as MCCE[32] can be particularly effective in the prediction of protonation state,[33] and it is likely that their use will also affect prediction accuracy. Although prediction accuracy for surface residues relative to that obtained for core residues will ultimately be limited due to their intrinsic mobility, we have shown here that it is possible to obtain good results if the various factors treated in this work are taken into account. On the basis of our results, there is clearly room for continued methodological development.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Acknowledgments

No endorsement by the U.S. Government should be inferred from the mention of trade names, software packages, commercial products, or organizations.

Grant sponsor: NIH; Grant number: GM-30518.

References

1. Dahiyat BI, Mayo SL. Probing the role of packing specificity in protein design. Proc Natl Acad Sci USA 1997;94:10172–10177. [PubMed: 9294182]
2. Ma XH, Wang CX, Li CH, Chen WZ. A fast empirical approach to binding free energy calculations based on protein interface information. Protein Eng 2002;15:677–681. [PubMed: 12364582]
3. Anderson EM, Halsey WA, Wuttke DS. Site-directed mutagenesis reveals the thermodynamic requirements for single-stranded DNA recognition by the telomere-binding protein Cdc13. Biochemistry 2003;42:3751–3758. [PubMed: 12667066]
4. Fiser A, Feig M, Brooks CL, Sali A. Evolution and physics in comparative protein structure modeling. Acc Chem Res 2002;35:413–421. [PubMed: 12069626]
5. Dunbrack RL Jr, Karplus M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J Mol Biol 1993;230:543–574. [PubMed: 8464064]
6. Xiang Z, Honig B. Extending the accuracy limits of prediction for side-chain conformations. J Mol Biol 2001;311:421–430. [PubMed: 11478870]
7. Jacobson MP, Friesner RA, Xiang Z, Honig B. On the role of the crystal environment in determining protein side-chain conformations. J Mol Biol 2002;320:597–608. [PubMed: 12096912]
8. Liang S, Grishin NV. Side-chain modeling with an optimized scoring function. Protein Sci 2002;11:322–331. [PubMed: 11790842]
9. Eyal E, Najmanovich R, McConkey BJ, Edelman M, Sobolev V. Importance of solvent accessibility and contact surfaces in modeling side-chain conformations in proteins. J Comput Chem 2004;25:712–724. [PubMed: 14978714]
10. Janin J. Errors in three dimensions. Biochimie 1990;72:705–709. [PubMed: 2078587]
11. Janin J. Radiocrystallographic analysis of protein structures. Biochimie 1975;57:505–514. [PubMed: 1148336](in French)
12. DePristo MA, de Bakker PI, Blundell TL. Heterogeneity and inaccuracy in protein structures solved by x-ray crystallography. Structure 2004;12:831–838. [PubMed: 15130475]
13. Desmet J, Spriet J, Lasters I. Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. Proteins 2002;48:31–43. [PubMed: 12012335]
14. Lee KH, Xie D, Freire E, Amzel LM. Estimation of changes in side chain configurational entropy in binding and folding: General methods and application to helix formation. Proteins 1994;20:68–84. [PubMed: 7824524]
15. Bromberg S, Dill KA. Side-chain entropy and packing in proteins. Protein Sci 1994;3:997–1009. [PubMed: 7920265]
16. Schafer H, Smith LJ, Mark AE, van Gunsteren WF. Entropy calculations on the molten globule state of a protein: Side-chain entropies of agr;-lactalbumin. Proteins 2002;46:215–224. [PubMed: 11807950]
17. Creamer TP. Side-chain conformational entropy in protein unfolded states. Proteins 2000;40:443–450. [PubMed: 10861935]
18. Canutescu AA, Shelenkov AA, Dunbrack RL Jr. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 2003;12:2001–2014. [PubMed: 12930999]
19. Koehl P, Delarue M. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. J Mol Biol 1994;239:249–275. [PubMed: 8196057]
20. Xiang Z, Soto CS, Honig B. Evaluating conformational free energies: The colony energy and its application to the problem of loop prediction. Proc Natl Acad Sci USA 2002;99:7432–7437. [PubMed: 12032300]

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

21. Kuser PR, Franzoni L, Ferrari E, Spisni A, Polikarpov I. The X-ray structure of a recombinant major urinary protein at 1.75 angstrom resolution. A comparative study of X-ray and NMR-derived structures. Acta Crystallogr D Biol Crystallogr 2001;57:1863–1869. [PubMed: 11717500]

22. Jaroniec CP, MacPhee CE, Bajaj VS, McMahon MT, Dobson CM, Griffin RG. High-resolution molecular structure of a peptide in an amyloid fibril determined by magic angle spinning NMR spectroscopy. Proc Natl Acad Sci USA 2004;101:711–716. [PubMed: 14715898]

23. Ramya Bhargavi G, Sheik SS, Velmurugan D, Sekar K. Side-chain conformation angles of amino acids: Effect of temperature factor cut-off. J Struct Biol 2003;143:181–184. [PubMed: 14572473]

24. Wang C, Schueler-Furman O, Baker D. Improved side-chain modeling for protein-protein docking. Protein Sci 2005;14:1328–1339. [PubMed: 15802647]

25. Kortemme T, Morozov AV, Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. J Mol Biol 2003;326:1239–1259. [PubMed: 12589766]

26. Kumar S, Nussinov R. Salt bridge stability in monomeric proteins. J Mol Biol 1999;293:1241–1255. [PubMed: 10547298]

27. McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. J Mol Biol 1994;238:777–793. [PubMed: 8182748]

28. Efimov AV, Brazhnikov EV. Relationship between intramolecular hydrogen bonding and solvent accessibility of side-chain donors and acceptors in proteins. FEBS Lett 2003;554:389–393. [PubMed: 14623099]

29. Nicholls A, Honig B. A rapid finite-difference algorithm, utilizing successive over-relaxation to solve the Poisson-Boltzmann equation. J Comput Chem 1991;12:435–445.

30. Vriend G. WHAT IF: A molecular modeling and drug design program. J Mol Graph 1990;8:52–56. [PubMed: 2268628]

31. Petrey D, Xiang Z, Tang CL, Xie L, Gimpelev M, Mitros T, Soto CS, Goldsmith-Fischman S, Kernytsky A, Schlessinger A, Koh IY, Alexov E, Honig B. Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. Proteins 2003;53 (Suppl 6):430–435. [PubMed: 14579332]

32. Alexov EG, Gunner MR. Calculated protein and proton motions coupled to electron transfer: Electron transfer from QA to QB in bacterial photosynthetic reaction centers. Biochemistry 1999;38:8253–8270. [PubMed: 10387071]

33. Forrest LR, Honig B. An assessment of the accuracy of the methods for predicting hydrogen positions in protein structures. Proteins 2005;61:296–309. [PubMed: 16114036]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

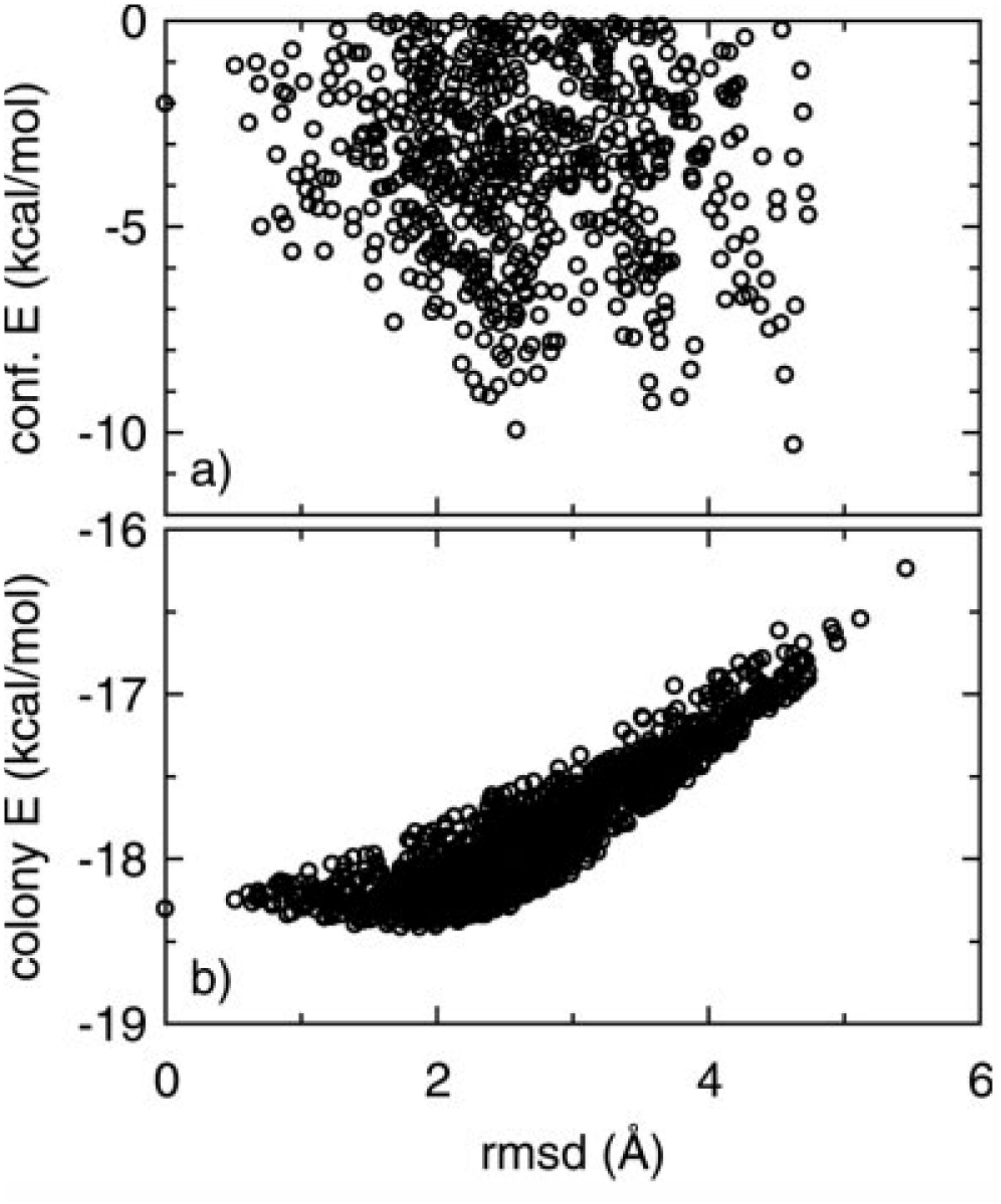

Fig. 1.

Effect of colony energy for residue Arg96 in serine esterase protein 1cex. (**a**) Plot of conformational energy vs. RMSD for the rotamers sampled. Energies greater than the *y*-axis maximum are not plotted. (**b**) Plot of colony energy vs. RMSD for the rotamers shown in (a).

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

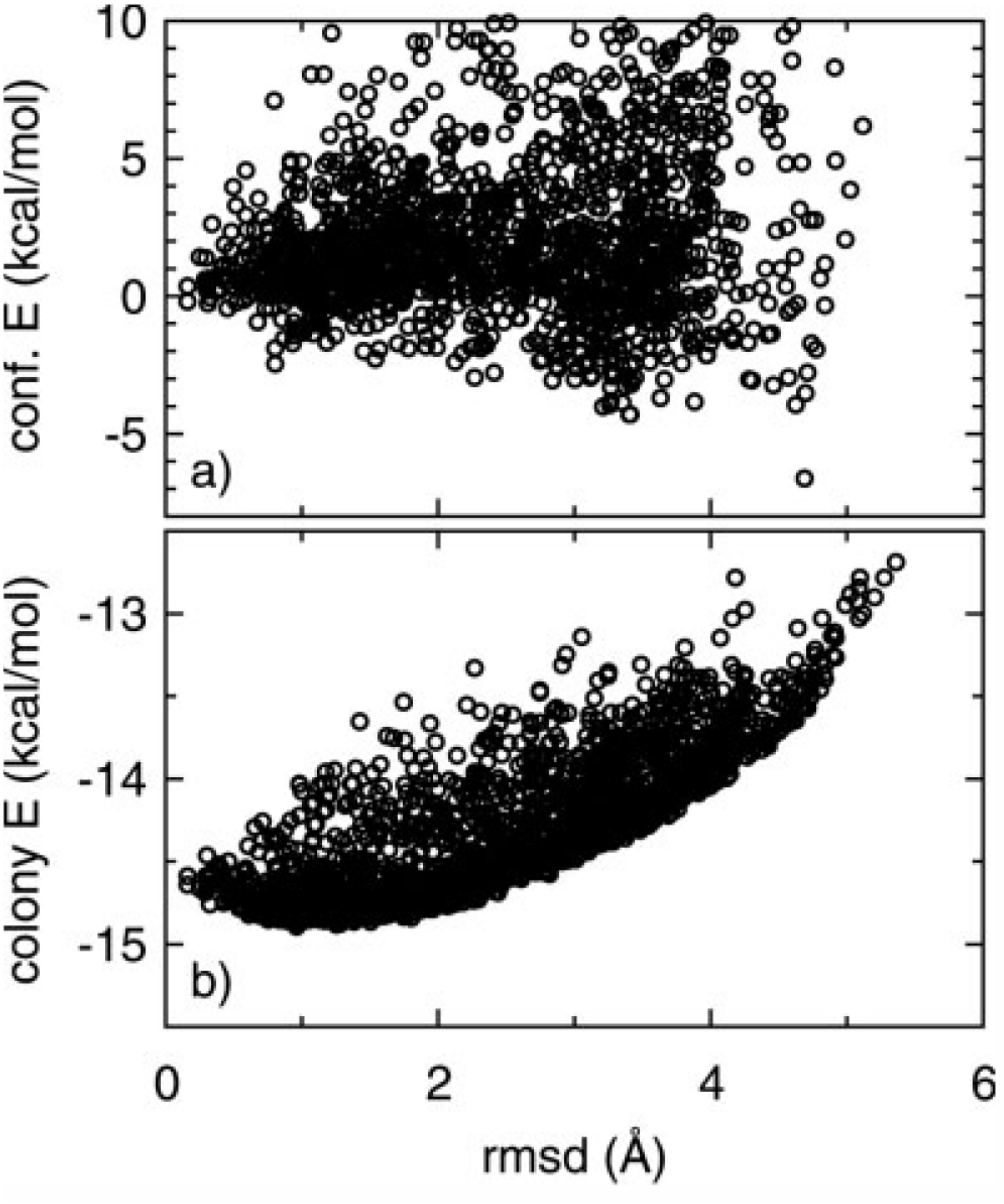

Fig. 2.
Effect of colony energy for residue Lys29 in electron transport protein 1aac, plotted as in Figure 1.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

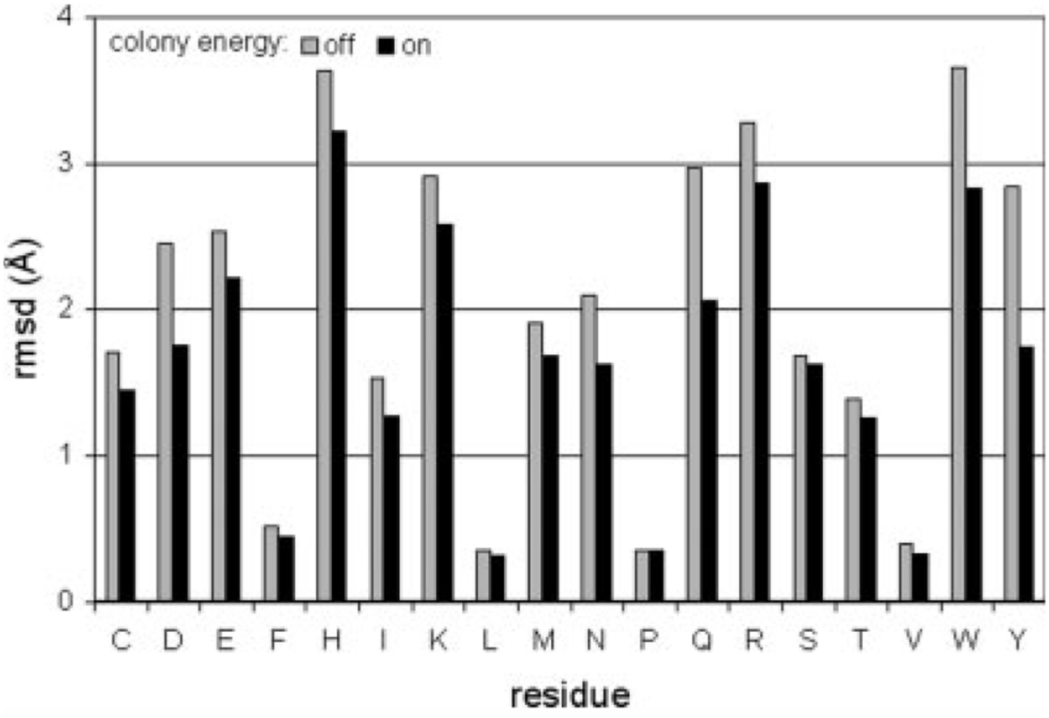

Fig. 3.
Effect of colony energy for surface residues.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

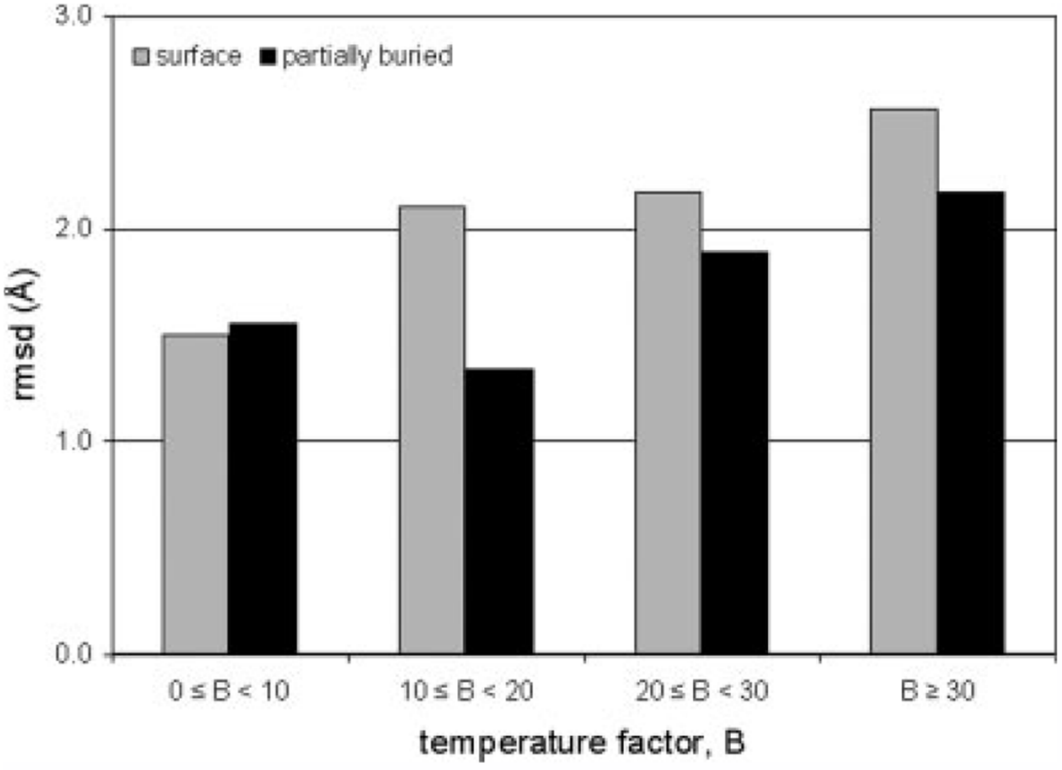

Fig. 4.
Comparison of predictions for surface and partially buried residues, at different values of the experimentally measured temperature factor.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

TABLE I

Number of Residues in Protein Sets

	Residues			Polar residues			Hydrogen-bonded residues		
	S	B	P	S	B	P	S	B	P
Training set	925	926	1182	586	136	604	130	109	319
Test set 1	1003	704	967	617	83	487	119	52	222
Test set 2	5209	13792	9491	4414	3864	6573	539	2414	2523

S, B, and P denote surface, buried, and partially buried residues, respectively.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

TABLE II

Single Side-Chain Prediction Accuracy for Training Set

			Core		Partially buried		Surface		All	
FF	**CE**	**FDPB**	**rmsd (Å)**	**χ_1/χ_{1+2} (%)**	**rmsd (Å)**	**χ_1/χ_{1+2} (%)**	**rmsd (Å)**	**χ_1/χ_{1+2} χ_1/χ_{1+2} (%)**	**rmsd (Å)**	**χ_1/χ_{1+2} (%)**
UA	No	No	0.42	98/97	1.27	91/78	2.81	65/40	1.45	90/78
AA	No	No	0.40	98/98	1.07	91/82	2.52	67/49	1.29	90/82
UA	Yes	No	0.39	98/97	1.13	90/82	2.16	74/59	1.18	91/84
AA	Yes	No	0.38	98/98	1.10	90/83	2.13	73/60	1.15	91/85
AA	No	Yes	0.39	99/98	1.05	91/83	2.26	69/55	1.20	91/83
AA	Yes	Yes	0.40	98/99	1.07	91/83	2.07	73/61	1.12	91/86

FF, CE, UA, and AA denote force field, colony energy, united-, and all-atom models, respectively. FDPB denotes Finite-Difference Poisson–Boltzmann results.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

TABLE III

Single Side-Chain Prediction Accuracy for Polar Residues of the Training Set

		Partially buried		Surface		Partially buried[h]		Surface[h]	
FF	**CE**	**rmsd (Å)**	**χ_1/χ_{1+2} (%)**	**rmsd (Å)**	**χ_1/χ_{1+2} (%)**	**rmsd (Å)**	**χ_1/χ_{1+2} (%)**	**rmsd (Å)**	**χ_1/χ_{1+2} (%)**
UA	No	1.43	89/73	2.92	60/38	1.31	92/85	2.29	78/56
AA	No	1.20	90/80	2.59	63/48	1.12	94/86	1.83	85/63
UA	Yes	1.27	87/79	2.21	71/59	1.17	93/85	1.80	86/64
AA	Yes	1.23	87/80	2.17	70/59	0.80	94/87	1.72	85/65

[h] Side chains participating in hydrogen bonds.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

TABLE IV

All Side-Chain Prediction Accuracy for Training Set

					Core		Partially buried		Surface		All	
FF	CE	Ns	Het	Rot	rmsd (Å)	χ_1/χ_{1+2} (%)	rmsd (Å)	χ_1/χ_{1+2} (%)	rmsd (Å)	χ_1/χ_{1+2} (%)	rmsd (Å)	χ_1/χ_{1+2} (%)
UA	No	1	No	10	1.04	92/89	2.00	79/65	2.81	63/40	1.83	82/70
AA	No	1	No	10	0.81	94/92	1.60	83/72	2.53	68/48	1.54	85/76
UA	Yes	1	No	10	0.93	91/89	1.70	81/70	2.31	66/55	1.55	83/75
AA	Yes	1	No	10	0.84	94/92	1.62	81/74	2.30	68/55	1.49	85/78
AA	Yes	120	No	10	0.65	96/95	1.45	83/77	2.25	69/56	1.38	87/80
AA	Yes	1	Yes	10	0.69	96/95	1.28	87/79	2.11	71/58	1.27	88/81
AA	No	1	No	40	1.36	88/79	1.73	82/69	2.40	66/55	1.73	81/69
AA	Yes	1	No	40	1.43	87/77	1.76	82/69	2.35	68/56	1.75	82/70

Ns is the number of starting conformations used in prediction; Het denotes consideration of hetero atoms (including waters, ions, and ligands) in the PDB file; Rot denotes the resolution of rotamer library, i.e., 10° or 40°.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

TABLE V

All Side-Chain Prediction Accuracy for Polar Residues of Training Set

						Partially buried		Surface		Partially buried[h]		Surface[h]	
FF	CE	Ns	Het	CP	Rot	rmsd (Å)	χ_1/χ_{1+2} (%)	Rmsd (Å)	χ_1/χ_{1+2} (%)	rmsd (Å)	χ_1/χ_{1+2} (%)	rmsd (Å)	χ_1/χ_{1+2} (%)
AA	No	1	No	No	10	1.64	83/70	2.57	68/48	1.70	81/69	2.14	72/53
AA	Yes	1	No	No	10	1.62	81/73	2.29	69/56	1.75	81/76	2.00	73/59
AA	Yes	1	Yes	No	10	1.32	88/78	2.14	71/57	1.50	89/73	1.91	77/60
AA	Yes	1	Yes	Yes	10	1.25	88/80	1.86	83/72	1.36	90/76	1.60	86/74

CP denotes crystal packing effects.

[h] Side chains participating in hydrogen bonds.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

TABLE VI

Comparison of All Side-Chain Prediction Accuracy for Test Set 1

					Core		Partially buried		Surface		All		
FF	**CN**	**Het**	**CP**	**CPU(s)**	**rmsd**	**χ_1/χ_{1+2}**	**rmsd**	**χ_1/χ_{1+2}**	**rmsd**	**χ_1/χ_{1+2}**	**rmsd**	**χ_1/χ_{1+2}**	**Het**
SCAP	UA	No	No	No	134	1.12	89/83	2.04	81/63	2.81	65/41	1.90	82/67
	AA	No	No	No	298	0.91	92/90	1.70	83/70	2.53	66/47	1.66	84/73
	UA	Yes	No	No	180	1.08	93/83	1.89	80/67	2.41	69/52	1.72	83/71
	AA	Yes	No	No	361	0.88	93/89	1.65	83/72	2.31	72/55	1.54	85/75
	AA	Yes	Yes	No	380	0.82	94/88	1.47	86/77	2.12	78/65	1.39	88/79
	AA	Yes	Yes	Yes	406	0.77	93/88	1.25	89/82	1.73	82/73	1.18	90/83
SMOL	AA		No	No	11700	0.97	94/87	1.65	87/74	2.49	70/47	1.62	87/74
SCCOMP	UA		No	No	960	0.89	94/87	1.66	87/73	2.44	68/50	1.59	86/74
SCWRL	UA		No	No	3	1.15	90/82	1.89	84/68	2.40	71/52	1.74	84/70

Rows 1–6 summarize SCAP results using a single starting conformation.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

TABLE VII

Prediction Accuracy for Polar Side Chains of Test Set 1

		Core		Partially buried		Surface		Core[a]		Partially buried[a]		Surface[a]	
Het	**CP**	**Rmsd**	χ_1/χ_{1+2}	**rmsd**	χ_1/χ_{1+2}	**rmsd**	χ_1/χ_{1+2}	**rmsd**	χ_1/χ_{1+2}	**rmsd**	χ_1/χ_{1+2}	**rmsd**	χ_1/χ_{1+2}
No	No	1.07	90/84	1.83	82/68	2.39	73/56	1.30	84/88	1.97	84/66	2.26	79/63
Yes	No	0.83	93/87	1.64	86/74	2.20	79/64	1.06	87/87	1.60	89/77	1.81	83/67
Yes	Yes	0.81	93/87	1.42	88/79	1.82	82/73	1.00	90/89	1.26	91/82	1.64	87/75

Colony energy and all-atom model were used throughout.

[a] Side chains participating in hydrogen bonds.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

TABLE VIII

Comparison of All Side-Chain Prediction Accuracy for Test Set 2

	Core		Partially buried		Surface		Surface[a]		Surface[b]	
	rmsd(Å)	χ_1/χ_{1+2}(%)	rmsd	χ_1/χ_{1+2}	rmsd	χ_1/χ_{1+2}	rmsd	χ_1/χ_{1+2}	rmsd	χ_1/χ_{1+2}
SCAP	0.91	93/89	1.85	80/71	2.31	71/52	2.38	70/53	2.33	74/57
SCCOMP	0.92	93/88	1.78	82/69	2.42	69/50	2.50	67/49	2.50	72/53
SCWRL	1.23	90/79	2.08	79/62	2.36	70/50	2.41	68/49	2.53	70/51
SMOL	1.01	92/86	1.75	83/70	2.46	69/48	2.55	68/46	2.51	73/52

[a] Polar residues.

[b] Side chains participating in hydrogen bonds.