Protein Science

Modeling mutations in protein structures

Eric Feyfant, Andrej Sali and András Fiser

Protein Sci. 2007 16: 2030-2041
Access the most recent version at doi:10.1110/ps.072855507

Modeling mutations in protein structures

ERIC FEYFANT,[1] ANDREJ SALI,[2] AND ANDRÁS FISER[3,4]

[1]Wyeth Research, Chemical and Screening Sciences, Cambridge, Massachusetts 02421, USA
[2]Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research, University of California at San Francisco, San Francisco, California 94158, USA
[3]Department of Biochemistry, Albert Einstein College of Medicine, Bronx, New York 10461, USA
[4]Institute of Enzymology and Alfred Renyi Institute of Mathematics, Hungarian Academy of Sciences, H-1113 Budapest, Hungary

(RECEIVED March 5, 2007; FINAL REVISION June 19, 2007; ACCEPTED June 19, 2007)

Abstract

We describe an automated method for the modeling of point mutations in protein structures. The protein is represented by all non-hydrogen atoms. The scoring function consists of several types of physical potential energy terms and homology-derived restraints. The optimization method implements a combination of conjugate gradient minimization and molecular dynamics with simulated annealing. The testing set consists of 717 pairs of known protein structures differing by a single mutation. Twelve variations of the scoring function were tested in three different environments of the mutated residue. The best-performing protocol optimizes all the atoms of the mutated residue, with respect to a scoring function that includes molecular mechanics energy terms for bond distances, angles, dihedral angles, peptide bond planarity, and non-bonded atomic contacts represented by Lennard-Jones potential, dihedral angle restraints derived from the aligned homologous structure, and a statistical potential for non-bonded atomic interactions extracted from a large set of known protein structures. The current method compares favorably with other tested approaches, especially when predicting long and flexible side-chains. In addition to the thoroughness of the conformational search, sampled degrees of freedom, and the scoring function type, the accuracy of the method was also evaluated as a function of the flexibility of the mutated side-chain, the relative volume change of the mutated residue, and its residue type. The results suggest that further improvement is likely to be achieved by concentrating on the improvement of the scoring function, in addition to or instead of increasing the variety of sampled conformations.

Keywords: point mutation; protein structure; comparative modeling

Residue type differences at a single position in a protein are important consequences of genetic variation among the individuals (Strausberg et al. 2003). Point mutations in a protein sequence may result in a change or loss of the native structure, which in turn may cause a change or loss of function, and ultimately yields different phenotypes. In addition to the natural variations among the individuals, researchers frequently introduce single amino acid residue replacements by site-directed mutagenesis in the laboratory to explore structural and functional features of proteins. For example, site-directed mutagenesis is often applied to study functional specificity (Wu et al. 1999), structure stability (Matthews 1995), kinetics and mechanism of protein folding (Ladurner and Fersht 1997), oligomerization (Chattopadhyay et al. 2006), and the stability of protein complexes (Otzen and Fersht 1999). Site-directed mutagenesis is also used to introduce

Reprint requests to: Andrej Sali, Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research, Byers Hall at Mission Bay, Suite 503B, University of California at San Francisco, 1700 4th Street, San Francisco, CA 94158, USA; e-mail: sali@salilab.org; fax: (415) 514-4231; or András Fiser, Department of Biochemistry, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY 10461, USA; e-mail: andras@fiserlab.org; fax: (718) 430-8565.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.072855507.

binding sites for heavy atoms in preparation for X-ray crystallographic experiments and other markers, such as spin labels, for various spectroscopic experiments (Perozo et al. 1999).

Experimental exploration of different positions in a protein structure with various residue types is a time-consuming and expensive process. Such an exploration is generally facilitated by three-dimensional (3D) modeling of side-chain mutations (Dunbrack and Karplus 1993; Vasquez 1996; Koehl and Delarue 1997; Levitt et al. 1997; Xiang and Honig 2001). While the modeling of a single side chain in a given atomic environment seems to be one of the easiest of all protein structure prediction problems, it is still not solved (Fiser 2004). Seemingly insignificant change of a side-chain may lead to a significant change or loss of protein function (Wu et al. 1999). This observation implies that side-chain conformation prediction is useful only if it is highly accurate, which makes it a challenging problem.

Two simplifications are frequently applied in the modeling of side-chain conformations. First, amino acid residue replacements often leave the backbone conformation almost unchanged (Chothia and Lesk 1986). As a consequence, many algorithms fix the backbone during the search for the best side-chain conformations. Second, it was observed that most side-chains in high-resolution crystallographic structures can be represented by a limited number of conformers that comply with stereochemical and energetic constraints (Janin and Wodak 1978). This observation motivated Ponder and Richards to develop the first library of side-chain rotamers for the 17 types of residues with dihedral angle degrees of freedom in their side-chains, based on 10 high-resolution protein structures determined by X-ray crystallography (Ponder and Richards 1987). Subsequently, a number of additional libraries have been derived (Tuffery et al. 1991; Dunbrack and Karplus 1993; Dunbrack and Cohen 1997; Mendes et al. 1999; Xiang and Honig 2001).

Rotamers on a fixed backbone are often used when all the side chains need to be modeled on a given backbone. This approach overcomes the combinatorial explosion associated with a full conformational search of many side chains and is applied by some comparative modeling (Blundell et al. 1987) and protein design approaches (Desjarlais and Handel 1999). In addition, it has been shown that the accuracy of side-chain modeling on a fixed backbone decreases rapidly when the backbone errors are >0.5Å (Chung and Subbiah 1996). Fortunately, these two approximations may be unnecessary in the modeling of a single-point mutation that in general does not trigger changes in many dihedral angles (Xiang and Honig 2001).

Earlier methods for side-chain modeling often put less emphasis on the energy or scoring function. The function was usually greatly simplified and consisted of the empirical rotamer preferences and simple repulsion terms for non-bonded contacts (Dunbrack and Karplus 1993). Nevertheless, these approaches have been justified by their performance. For example, a method based on a rotamer library compared favorably with that based on a molecular mechanics force field (Cregut et al. 1994). More recent and efficient methods are also based on rotamer libraries, albeit some of these methods radically expand the rotamer library size, up to as many as approximately 50,000 rotamer states (Xiang and Honig 2001; Canutescu et al. 2003; Peterson et al. 2004). In contrast, much attention has been paid to the optimization procedure. The various approaches include Monte Carlo simulation (Eisenmenger et al. 1993; Jain et al. 2006), simulated annealing (Lee and Levitt 1991), a combination of Monte Carlo and simulated annealing (Holm and Sander 1992), the dead-end elimination theorem (Lasters and Desmet 1993; Looger and Hellinga 2001), genetic algorithms (Tuffery et al. 1991), neural network with simulated annealing (Hwang and Liao 1995), mean field optimization (Koehl and Delarue 1994), and combinatorial searches (Dunbrack and Karplus 1993; Bower et al. 1997; Petrella et al. 1998).

Several recent papers focused on the testing of more sophisticated potential functions for conformational search (Petrella et al. 1998; Jacobson et al. 2002) and development of new scoring functions for side-chain modeling (Liang and Grishin 2002), reporting improved accuracy compared to earlier studies. In retrospect, one reason for relative success of the early simplified energy models may be a surprisingly small role of entropy in determining side-chain conformational preferences (Hu and Kuhlman 2006). Most recent methods rely on all-atom molecular mechanics force fields in addition to rotamer preferences (Jain et al. 2006; Zhang and Duan 2006). Both of these approaches applied a cooperative rearrangement of atoms where groups of side-chain atoms were deleted from a side chain in a particular region and then regrown with the generation of trial positions to achieve a more continuous sampling of rotamer space and a smoother potential surface. Estimates of the accuracy of side-chain modeling techniques vary substantially, partially caused by inaccuracies in the experimentally determined structures (Shapovalov and Dunbrack 2007).

In this paper, we describe a method for the modeling of point mutations in the context of several different types of a fixed environment provided by the rest of the protein structure. We approach this task through its three aspects: a representation of the modeled system, a scoring function that depends on the conformation of the system, and an optimization procedure for finding good scoring conformations of the system. The modeling protocol is tested on a set of 717 side-chain mutations that have

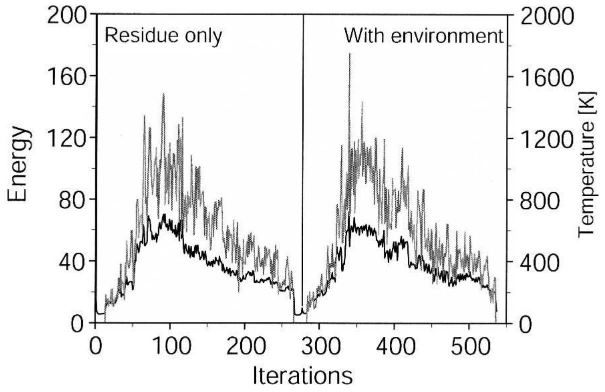

Figure 1. Optimization of a side chain. The scoring function (thick line and *left Y*-axis) is shown as a function of progress during optimization. The optimization starts with a conjugate gradient minimization, followed by a molecular dynamics optimization, using simulated annealing protocol (temperature is shown in the thin line and *right Y*-axis), and finished with a conjugate gradient minimization again. Two cycles of such an optimization are carried out: first, considering only interactions between the atoms of the mutated residue (residue only); and second, also including the interactions with the environment (with environment).

been defined by X-ray crystallography. We show that the best-performing scoring function takes advantage of homology-derived dihedral angle restraints. In contrast, homology-derived distance restraints overconstrain the model and lead to a decrease in the prediction accuracy. Optimally, all atoms of the mutated residue, including its main-chain atoms, need to be optimized in the search for the most accurate prediction. The performance of the prediction method is explored as a function of sampling, residue type, relative change in the occupied volume, side-chain flexibility, and self-consistency. The most influential terms in the scoring function are discussed.

Results and Discussion

We focused on the search for the best-performing scoring function for modeling point mutations in proteins. Even for the modeling of loops, which is usually a larger optimization problem than the modeling of a single residue, it was demonstrated that the hurdle to a more accurate prediction is a more accurate scoring function, not a more thorough sampling (Fiser et al. 2000, 2002). Another advantage of focusing on the scoring function is that it helps toward understanding the importance of the different determinants of protein structure. During the search for a well performing scoring function, we explored three components: (1) internal energy as captured by the terms from the CHARMM-22 molecular mechanics force field (MacKerell et al. 1998), (2) nonbonded energy terms, and (3) statistical preferences that restrain dihedral angles and distances as implied by known protein structures in general (i.e., statistical potentials) or by similar structures only (i.e., homology-derived restraints). The scoring function was optimized using a protocol that combines conjugate gradient minimization and molecular dynamics simulation with simulated annealing (Fig. 1). In this section, we discuss the results, separately for exposed and buried residues.

We also explored the choice of the environment that needs to be fixed for the optimization to achieve the most accurate prediction. First, we assumed that the backbone largely remains unchanged by a mutation and modeled the new side chain on a fixed backbone. However, it has already been shown that the lack of backbone flexibility limits the prediction accuracy (Desjarlais and Handel 1999), so in a second approach we optimized the whole residue. It has also been shown for some mutations that their environment adapts to the volume changes (Liu et al. 2000). Therefore, in a third approach, we also optimized the environment of the mutated side chain.

The side-chain prediction approach is analyzed on a test set of experimentally solved 717 side-chain mutations. The accuracy of the method is compared to that of SCWRL (version 2.9) (Bower et al. 1997). SCRWL is available, widely used, and one of the existing programs for side-chain modeling, the most accurate.

Best-performing scoring functions

Different methods were applied for two subsets of mutations, buried and exposed. This subdivision by solvent accessibility of the mutation type corresponds to different prediction problems. In the case of buried residues, a side chain has to adapt its conformation to fit the local packing, even if that induces a conformational strain in the side chain itself. In contrast, at the exposed position, a side chain can in principle adopt several conformations due to a less restraining environment. Five-hundred and thirty-one of 726 mutations were classified as buried with 30% or less relative solvent accessibility. Thirty-six alternative protocols have been compared by measuring the percentage of correctly predicted $\chi_{1,}$ χ_2, and χ_{1+2} dihedral angles for the mutated residue in the model and in the actual structure. The accuracy of predicted χ_1 dihedral angles of the buried residues is similar for the 12 tested scoring functions and does not depend significantly on the type of the environment (Fig. 2A).The best result, 76.3% (σ = 3.5%) of correct χ_1 angles, was achieved by optimizing only the mutated side chains with respect to the scoring function consisting of the Lennard-Jones potential with the full set of restraints (Full-LJ, protocol 5) including homology-derived and statistically calculated restraints. By evaluating χ_1 only, it is not possible to discriminate among the few best approaches, because the differences between the accuracies are comparable to standard deviations

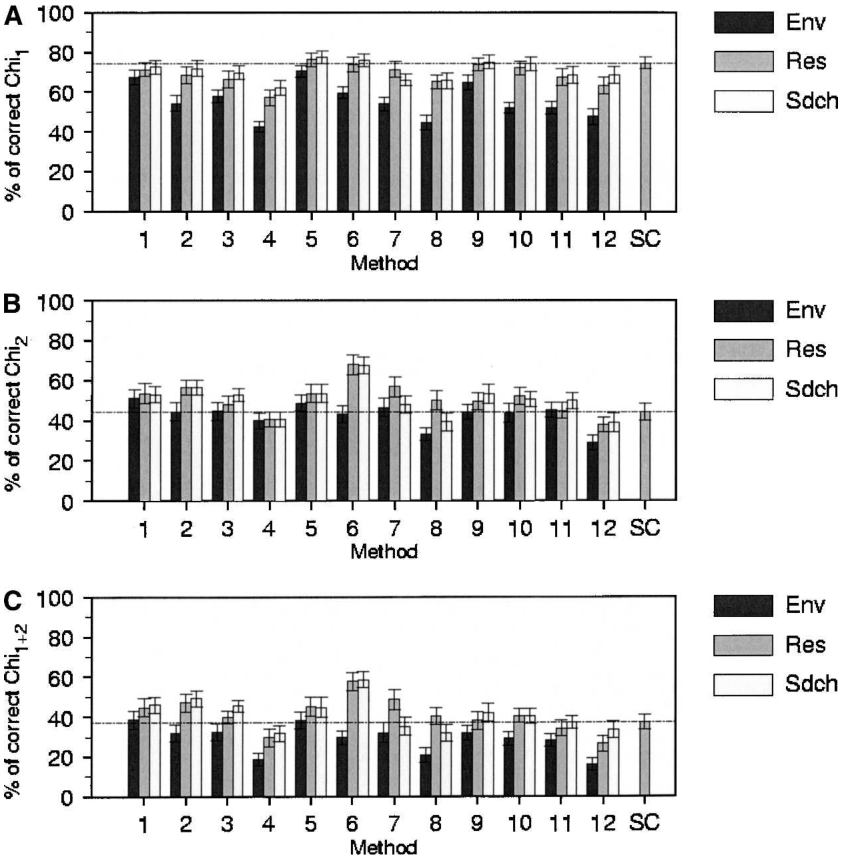

Figure 2. The average accuracy of modeling mutated residues in buried locations is shown as the fraction of correctly predicted χ_1, χ_2, and χ_{1+2} dihedral angles, on panels *A, B,* and *C,* respectively. Standard deviations of average accuracies are indicated. Twelve different scoring functions are shown on each panel for three different environments: (1) dark bars: all atoms of the mutated residue and all the atoms within 4.5 Å of the residue; (2) gray bars: all atoms of the mutated residue; and (3) white bars: only the side-chain atoms of the mutated residue. The components of the 12 scoring functions are detailed in Table 1. The horizontal line corresponds to the average SCRWL prediction accuracy.

(typically, ~3.5%). For χ_2 (Fig. 2B) and χ_{1+2} (Fig. 2C), the differences become more significant, clearly favoring one particular protocol (StlibH-LJ, protocol 6). This protocol relies on a combination of stereochemical restraints with the Lennard-Jones potential as well as the dihedral angle restraints from many known protein structures (phi, psi, and omega) and the template structure (chi_i). This scoring function clearly outperforms others by at least 10% (accuracy is 67.7% ± 4.2% for χ_2 and 58.4% ± 4.0% for χ_{1+2}).

The differences between the accuracies of the best-scoring functions are less accentuated in case of exposed residues, but the most accurate protocol still depends on optimizing the whole residue with respect to the StlibH-LJ scoring function (Fig. 3A–C). The accuracy for χ_1, χ_2, and χ_{1+2} is 74.2%, 63.6%, and 51.4%, respectively. These results are only a few percentage points worse than those for buried residues. It is expected that the exposed residues are more difficult to predict than the buried residues because of their less restraining environment. In the present study, this expectation is confirmed, but the differences are somewhat smaller than reported before (Dunbrack and Karplus 1994): The differences between the accuracies of χ_1, χ_2, and χ_{1+2} predictions for the buried and exposed residues are 1.9%, 4.3%, and 7.0%, respectively.

As described in Materials and Methods, we used non-hydrogen atom representation of our system. However, the Lennard-Jones non-bonded terms were originally parameterized using the all-atom representation that includes hydrogen atoms. Therefore, for the scoring functions that employed the Lennard-Jones terms, we tested the all-atom representation as well. The prediction accuracies did not change significantly (data not shown).

In this work, interactions of the protein with the solvent are not treated explicitly. However, there is an implicit partial consideration of solvation through the statistical pair potential. It is reasonable to expect that additional

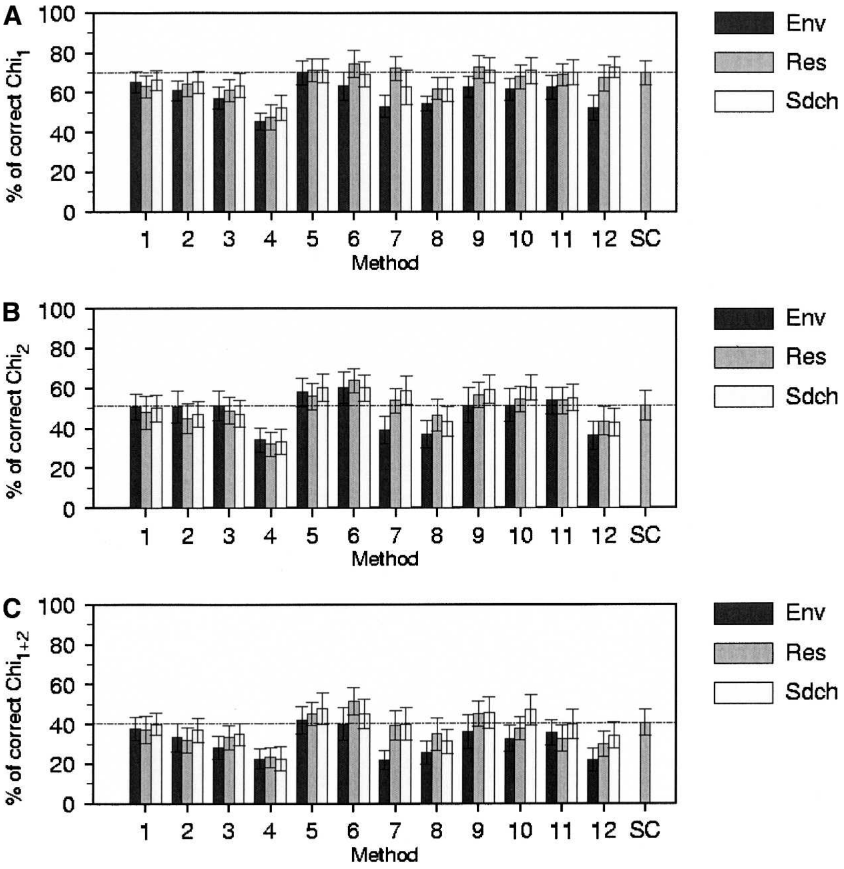

Figure 3. The same as Figure 2 but for mutations of exposed residues.

terms responsible for solvation free energy would improve the prediction accuracy, especially for the exposed side chains. Therefore, we used a generalized Born implicit solvation potential to calculate the electrostatic solvation free energy in combination with a nonpolar hydration free energy estimator (GBSA) (Gallicchio and Levy 2004). The latter model combines an estimator of the solute-solvent van der Waals interaction and a surface area term corresponding to cavity formation. In one approach, our best-performing point mutation modeling protocol first generated 10 conformations that were then ranked by the solvation term. In another approach, the GBSA term was added to the scoring function to obtain the single best-scoring conformation. The solvation term dominates the scoring function. In both approaches, we observed a similar decrease in accuracy, sometimes statistically significant. The accuracy for buried (exposed) χ_1, χ_2, and χ_{1+2} side-chain dihedral angles is 62.9% (61.4%), 47.2% (47.8%), and 32.9% (37.5%), respectively. For χ_4 of a few long side chains, the accuracy was 42.9% (60%), corresponding to a minor, statistically insignificant improvement. We conclude that the current combination of the GBSA term, template-dependent restraints, statistical and van der Waals potentials does not improve the modeling of single-point mutations with our current sampling protocol. However, the modest improvement of the conformations of exposed long side chains indicates that further work in this direction may be warranted.

Accuracy as a function of conformational sampling

Each point mutation modeling corresponds to a set of several independent optimizations. The final model was selected as the one with the lowest scoring function value. We explored the question of how many final conformations needed to be generated to obtain the highest possible accuracy for a given scoring function. One-hundred models were built for each side chain, using our best prediction protocol. Figure 4 shows only the first 20 sampled models for buried and exposed mutations, respectively. The accuracy did not improve with the increase in sampling beyond building 10 models. This suggests that our conformational sampling of the tested scoring functions is essentially exhaustive. Therefore, further improvement can be expected by concentrating

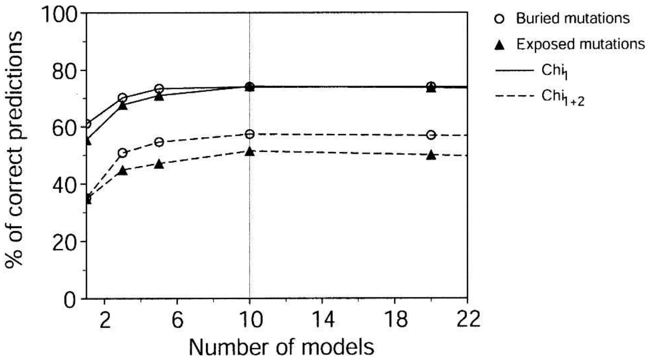

Figure 4. Accuracy as a function of sampling, for the buried and exposed mutations.

on the improvement of the scoring function and not by extending the conformational sampling. However, a further perfection in the scoring scheme would allow one to select and accurately model larger environments of a point mutation with the possibility of a better capturing packing rearrangement and in turn delivering more accurate side-chain conformations.

Most influential terms in the scoring function

As described in Materials and Methods, there are three main components of the scoring function: (1) CHARMM terms responsible for the local stereochemistry, (2) terms describing non-bonded interactions, and (3) spatial restraints on dihedral angles and distances derived from the template (“homology-derived restraints”) or from many known structures (“statistical potentials”). The CHARMM terms were included in all tested versions of the scoring function, while we varied the other two in our testing. It is difficult to draw a sharp conclusion about the accuracy of the applied non-bonded energy terms. The Lennard-Jones and soft sphere repulsive potentials perform quite similarly, the former being slightly more accurate. The combination of a statistical and soft sphere repulsive potential, which proved to be the most accurate choice for loop modeling (Fiser et al. 2000), seems to be the least accurate in the modeling of side chains, especially for χ_1 and χ_{1+2} (Figs. 2A–C, 3A–C).

In contrast to the non-bonded terms, a clear pattern emerges for the spatial restraints. Exclusion of all the spatial restraints (i.e., homology-derived distance and dihedral angle restraints; and dihedral angle restraints from many known structures) leads to a decrease of accuracy irrespectively of the non-bonded terms and the optimized set of atoms. If at least one type of the spatial restraints are included in the scoring function (e.g., restraints on dihedral angles derived from many known structures), approximately the highest achieved level of accuracy is obtained. Further adding the template-derived dihedral angle restraints improves the accuracy by <2%, resulting in the most accurate scoring function (StlibH-LJ) using the full residue representation. Adding homology-derived distance restraints seems to overconstrain the mutated residue and its structural environment, leading to lower prediction accuracies.

Role of the environment

Protein structure can adapt to a mutation by rearranging the spatial environment of the mutated residue; sometimes, only the backbone shifts slightly. For less densely packed neighborhoods, it is also possible that a mutation fits without causing any shifts or distortions. Accordingly, three different environment selections were explored for our optimizations: (1) only mutated side-chain atoms, (2) the whole mutated residue, and (3) the mutated residue and all atoms within 4.5 Å of any of the atoms of the original residue.

The most accurate results were achieved if we considered the full residue in the optimization. Both in case of buried and exposed residues optimizing the environment of the mutation generally results in a less accurate prediction (Figs. 2, 3). An exclusion of homology-derived restraints on dihedral angles and distances results in a significant decrease in prediction accuracy. The more homology-derived restraints are added to the scoring function, the higher is the absolute accuracy and the smaller are the differences between protocols optimizing different environments.

For the exposed residues, the accuracy of the protocol generally does not depend as strongly on the selected environment as it does for the buried residues. One possible explanation is that the environment plays a smaller role in restraining the modeled side chain. Nevertheless, the pattern remains the same as that for the buried residues.

The cutoff distance for defining the environment was increased from 4.5 Å to 6 Å, and 8 Å, but no improvement in the accuracy was observed (data not shown).

Accuracy as a function of volume change and B-factor

It is reasonable to expect the residue mutations that significantly alter the volume will have a more pronounced effect on the rest of the structure. We investigated this hypothesis from the viewpoint of prediction accuracy. Three situations were distinguished: A mutation does not change significantly the volume; a mutation decreases the volume by replacing a given residue with a smaller one; and a mutation increases the volume by replacing a residue with a larger one. We quantified changes as significant if the relative residue volume change is larger or smaller than 10^3 Å. This cutoff classifies the mutations in our benchmark into three equally populated groups. Figure 5 shows the correlation between the volume change and the accuracy of the prediction and confirms that, for both buried and exposed mutations, predictions are more accurate when the volume does not change significantly. This observation may provide a rule of thumb to estimate the accuracy of a given prediction.

We also explored the question of whether or not side-chain flexibility influences the prediction accuracy. Side-chain flexibility was quantified by the crystallographic B-factor. B-factors cannot be compared directly between different PDB files. Therefore, we calculated a B-factor Z-score of a mutated residue as ([B − A]/S), where B is the B-factor of the mutated residue, A is the average B-factor of all residues, and S is the standard deviation of all residue B-factors. No statistically significant dependence of prediction accuracy on the mutated residue B-factor Z-score was observed (data not shown). However, a recent detailed study that analyzed accuracies of side-chain conformations using original diffraction maps argues that accuracy of side-chain resolution is a critical factor in properly assessing prediction results (Shapovalov and Dunbrack 2007). The lack of a similar conclusion in our survey may just underline the unreliability of B-factors in structure files.

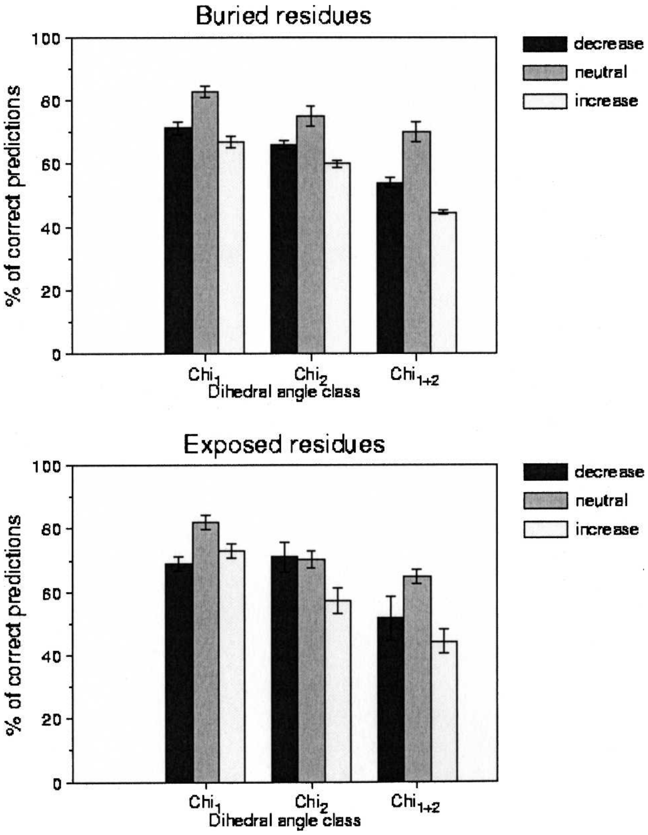

Figure 5. Accuracy as a function of relative volume change of mutations. Relative volume change is classified into three groups: volume decrease, neutral, or volume increase. At least a 1000 Å^3 relative volume change is required upon mutation to qualify for a decreased or increased class, otherwise the mutation is deemed as neutral.

Accuracy of prediction as a function of residue type

The residue type itself is expected to be an important predictor of prediction accuracy, partly because the available conformational space varies with the residue type. For example, proline has a much more restricted conformational flexibility than leucine. It may also be informative to characterize the accuracy for classes of substitutions, such as replacing a polar residue in the core of the protein or an exposed hydrophobic residue.

For a buried mutation, the larger residues, especially the aromatic ones, are predicted accurately whereas other residues, including proline, seem to be less well modeled (Fig. 6). For the exposed residues, the difference in accuracy between the different residue types is smaller, and significant differences can be observed only between the particularly large aromatic residues and the remaining types (not shown). The fact that buried, large residues are predicted more accurately than small ones is consistent with the fact that a buried neighborhood is more restraining than a flexible, exposed neighborhood.

The small number of test cases for most residue types in the benchmark of 717 point mutations precludes a reliable evaluation of the dependence of prediction accuracy on the mutant residue type.

Predicting the error in the model

What is the probability that the best-scoring model is correct? Or, how does one detect an error in the predicted structure? We examined the possibility that the accuracy is correlated with the fraction of independent optimizations resulting in conformations similar to the best-scoring prediction. For each of the 717 mutations in the benchmark, we plotted the frequency of the most frequently occurring conformation among the 10 best-scoring conformations (Fig. 7). In general, the majority of the 10 best-scoring models have the same conformation, whether or not it is correct. When the relative frequency of the most populated conformation is >50%, there is a tendency that this conformation is correct. However, the trend is weak and may not be useful for predicting the accuracy of the prediction.

The high relative frequency of the incorrect predictions among the best-scoring solutions again suggests that the method is limited by the scoring function and not the sampling.

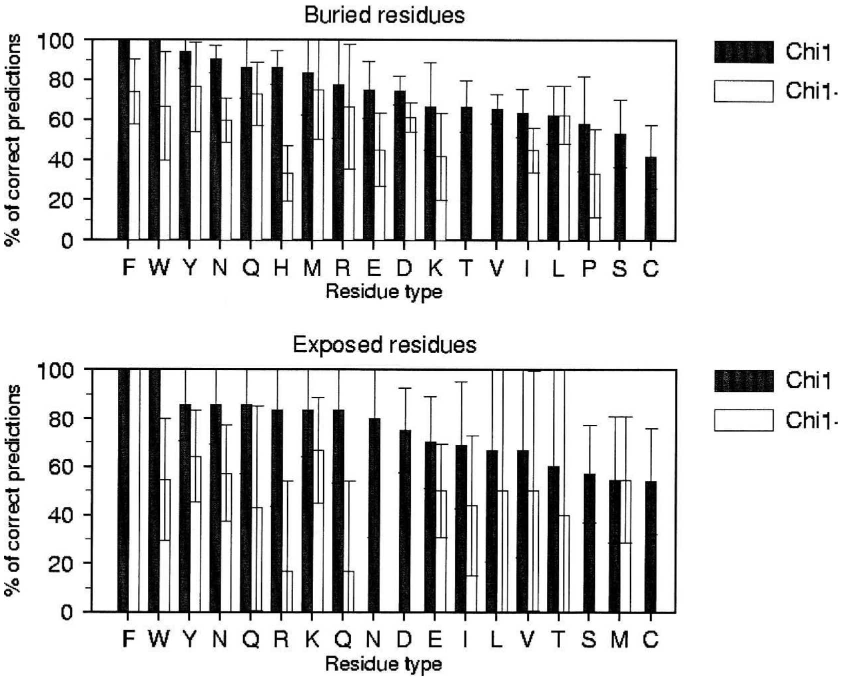

Figure 6. Accuracy of prediction as a function of the residue type. Standard deviations of averages are indicated. For Phe and Trp, only a few mutations occurred and they were all predicted correctly.

Comparing method accuracy to that of SCWRL

The performance of our best side-chain modeling protocol (StlibH-LJ) and using full residues for optimization was compared to that of SCWRL (Bower et al. 1997). SCRWL is widely accepted as one of the most accurate methods. SCWRL relies on a rotamer library and places side chains on a fixed backbone. SCRWL results are shown in Figures 2A–C and 3A–C. For χ_1, the accuracies of SCWRL and our protocol are comparable; the differences between accuracies are approximately as large as the standard deviations. SCWRL achieved the χ_1 accuracy of 74.4% (± 2.92%) and 69.9% (± 6.07%) for buried and exposed mutations, respectively, whereas StlibH-LJ performs at 73.8% and 74.2%. For χ_2, SCRWL's accuracy is 44.6% (± 4.03%) and 51.4% (± 7.31%) for buried and exposed mutations, respectively, whereas our protocol performs somewhat better at 67.9% and 63.8%. For χ_{1+2}, SCRWL's accuracy is 37.5% (± 3.89%) and 40.6% (± 6.43%) for buried and exposed, mutations, respectively, compared to our higher accuracies of 57.6% and 51.4%. We note in fairness that SCRWL was designed for predicting all side chains in a structure at the same time and was presumably not optimized for predicting single-point mutations.

Side-chain prediction accuracy, particularly for the exposed residues, strongly depends on the crystal environment of proteins (Jacobson et al. 2002). One-hundred and forty-nine of the 717 protein pairs in our benchmark were crystallized in different unit cells. The prediction accuracy for χ_1 dihedral angles of buried residues drops to 69% when the unit cells of the native and mutant proteins are different and increases to 75% when they are the same. As expected, this difference becomes more accentuated for the exposed mutations, where the χ_1 prediction accuracy changes to 62% and 79% for different and identical unit cells, respectively. These data confirm the observations of Jacobson and colleagues (2002) and

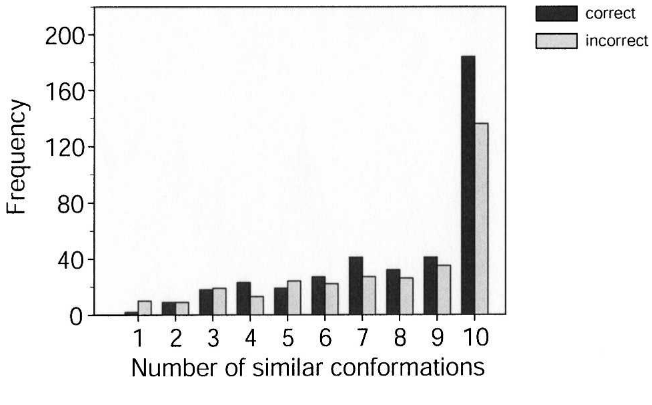

Figure 7. Distribution of the top 10 best-scoring conformations, which are identical to single best-scoring prediction. The plot is shown separately for the correct and incorrect best-scoring predictions.

highlight the need to assess different prediction methods by the same benchmark, as was done here.

Materials and Methods

The method for the modeling of a point mutation in a given environment is described here by specifying its three main components: (1) the representation of a protein, (2) the restraints that define the scoring or "energy" function, and (3) the method for optimizing the energy function. The modeling method is entirely automated and is implemented in the program MODELLER-8 (Sali and Blundell 1993) (http://salilab.org/modeller/).

Representation

Two different protein representations are applied: (1) Only the non-hydrogen atoms are used for scoring functions without the Lennard-Jones terms; and (2) all atoms, including all hydrogen atoms, are used for scoring functions with the Lennard-Jones terms. No explicit solvent molecules or ligands are included in general, although they could be added in special cases. The degrees of freedom in model optimization are the Cartesian coordinates of the atoms to be optimized. The atoms of the optimized residues "feel" the rest of the environment, but this "environment" does not move during the optimization procedure.

Environment of a point mutation

We investigated the optimal environment of a point mutation for modeling by testing three plausible selections of the optimized atoms: (1) only the atoms of the mutated side chain; (2) all atoms of the mutated residue, including its main-chain atoms; and (3) all atoms of the mutated residue and all other atoms within 4.5 Å of any of the atoms of the mutated residue.

Scoring function

The tested scoring functions for the modeling of point mutations are the sum of three major types of terms: (1) potential energy terms enforcing proper stereochemistry, (2) non-bonded energy terms, and (3) spatial restraints that express statistical preferences of distances, angles, and dihedral angles defined by the main-chain and side-chain atoms, depending either on the known protein structures in general or only on the template structure (Sali and Blundell 1993; Sali and Overington 1994; Fiser et al. 2000).

First, the internal energy terms that describe the stereochemical features are captured from the CHARMM molecular mechanics force field, F_{stereo} (MacKerell et al. 1998). Four terms were employed, restraining the length of the covalent bonds, the angles, the dihedral angles, and the planarity of the peptide bond. When statistical preferences of dihedral angles are used, the dihedral angle terms from the force field are omitted.

Second, for the non-bonded terms, $F_{\text{non-bonded}}$, three alternative functions were tested: (1) the Lennard-Jones terms from the CHARMM-22 force field (MacKerell et al. 1998), (2) "soft sphere repulsion" modeled by a harmonic lower bound (Equation 22 in Sali and Blundell 1993), and (3) a combination of the soft sphere repulsion with an atomistic, distance-dependent statistical potential of mean force ("statistical potential"), (Equations 1 and 4 in Fiser et al. 2000), (Sippl 1990; Melo and Feytmans 1998).

Third, main-chain and side-chain dihedral angles were also restrained by statistical preferences extracted from known protein structures in general (F_{stat}). Restraints on the main-chain dihedral angles ϕ and ψ depend on the residue type and correspond to the natural logarithm of the probability density of ϕ and ψ dihedral angles in a set of high-resolution protein structures, $\ln[p^m(\phi,\psi/R)]$ (Fiser et al. 2000). Similar restraints are also applied to the ω main-chain dihedral angle, $\ln[p^s(\omega/R)]$, and all side-chain dihedral angles χ_ι (up to four per residue), $\ln[p^s(\chi_\iota/R)]$ (Sali and Blundell 1993; Sali and Overington 1994; Fiser et al. 2000).

Finally, the main-chain and side-chain dihedral angles (Equations 25 and 26 in Sali and Blundell 1993) as well as distance restraints between C_α–C_α and N–O atoms (Equations 23 and 24 in Sali and Blundell 1993) were restrained by the template structure. These homology-derived restraints bias the target model toward the template structure.

Twelve combinations of the various components of the scoring function were explored in the three environments, resulting in 36 different protocols (Table 1). Twelve of these 36 protocols were tested with both the all-atom and non-hydrogen atom representations, while the remaining protocols used only the non-hydrogen atom representation.

Optimization of the scoring function

We used the same optimization protocol that was applied to the prediction of loops in protein structures (Fiser et al. 2000). One prediction consists of optimizing independently a number of randomized initial structures and picking as the final model the conformation that has the lowest value of the scoring function. A good compromise between efficiency and performance is achieved by 10 independent optimizations (Results and Discussion).

An individual optimization begins by generating starting coordinates for the atoms whose positions need to be optimized: (1) The selected atoms are built in their ideal positions based on the remaining atoms and the internal coordinates in the CHARMM-22 residue topology library; (2) these coordinates with locally ideal geometry

Table 1. *Summary of the different combinations of scoring function terms explored for side-chain modeling*

F_{stereo}	$F_{non\text{-}bonded}$	F_{stat}	Notation	Protocol #
Stereochemical restraints from CHARMM 22 potential force field	Distance-dependent statistical potential (Stat)	Full homology	Full-Stat	1
		StlibH	StlibH-Stat	2
		Stlib	Stlib-Stat	3
		None	Stereo-Stat	4
	Lennard-Jones (LJ)	Full homology	Full-LJ	5
		StlibH	StlibH-LJ	6
		Stlib	Stlib-LJ	7
		None	Stereo-LJ	8
	Lower bound harmonic potential (Soft sphere = SP)	Full homology	Full-Stat	9
		StlibH	StlibH-Stat	10
		Stlib	Stlib-Stat	11
		None	Stereo-Stat	12

Each of the 12 combinations was tested in three possible environments, resulting in 36 tested protocols (Materials and Methods). "Stlib" refers to standard side-chain rotamer library. "StlibH" indicates Stlib as combined with dihedral angle restraints that are derived from the equivalent residue in the template structure. "Full homology" refers to a combination of StlibH and homology-derived distance restraints.

are randomized by adding a random number distributed uniformly from −5 to 5 Å.

The procedure for optimizing a single initial conformation begins with a conjugate gradients minimization, continues with molecular dynamics with simulated annealing, and finishes by conjugate gradients again (Fig. 1). The first conjugate gradients phase is designed to relax the system and consists of five successive minimizations of up to 200 steps each, gradually increasing the scaling factors for the non-bonded restraints from 0, 0.01, 0.1, 0.5, to 1.0, respectively. In this phase, the atoms are allowed to pass near each other without having to surmount large energy barriers. This stage is followed by a relatively rapid heating up of the system consisting of 200 4 fs steps of "molecular dynamics" at 150°K, 250°K, 400°K, 700°K, and 1000°K. The heating stage is followed by the main optimization stage that consists of gradual cooling by molecular dynamics of 600 4 fs steps at 1000°K, 800°K, 600°K, 500°K, 400°K, and 300°K. Finally, the optimization is completed by a conjugate gradients minimization consisting of up to 1000 steps. There are, in fact, two cycles of the conjugate gradients, molecular dynamics with simulated annealing, and conjugate gradients phases: In the first cycle, only those non-bonded atom pairs are considered that contain the set of atoms selected for optimization (i.e., the side chain does not "feel" its environment). In the second cycle, the atom pairs that contain up to one environment atom are also included in the energy function (i.e., the side chain does "feel" its environment).

Test set

Pairs of protein structures that differ by a single residue type and were solved by X-ray crystallography at resolution of 2.0 Å or better were extracted from PDB (Berman et al. 2000). Point mutations were not considered if these happened within two positions of either termini of the sequences. The test set contained 431 pairs of protein structures. Excluding side chains without dihedral angles and considering the modeling of each one of the members in a pair based on the other member, the test set was comprised of 717 test cases for the modeling of point mutations.

Assessment of prediction accuracy

Following the usual convention, the accuracy of the protocols was assessed by the percentage of the correct χ_1 and χ_2 dihedral angles, as well as the correct pairs (χ_{1+2}) of these angles for the side chain of the mutated residue. A dihedral angle was defined to be correct if it was within 40° of the corresponding angle in the crystallographic structure of the modeled protein (Dunbrack and Karplus 1993; Jacobson et al. 2002).

Solvent accessibility

The solvent accessibility of a residue was calculated as implemented in MODELLER (Sali and Blundell 1993). The fractional surface area was obtained by dividing the contact area of a given residue by the standard contact area of the corresponding residue type X in the extended tripeptide Gly–X–Gly. Each residue with a fractional surface area of 30% or less was considered as buried.

Discussion

We described a method to model point mutations in protein structures. Our most accurate scoring function captures the internal energy that describes local stereochemical features through CHARMM force field terms, restraining the length of the covalent bonds, the angles, the dihedral angles, and the planarity of the peptide bond. The scoring function employs Lennard-Jones potential for non-bonded terms and combines homology-derived spatial restraints on dihedral angles from the template with statistical preferences observed in many representative structures. The highest accuracy is obtained if the full residue is considered for optimization. The algorithm consists of 10 independent optimizations. The procedure for optimizing a single initial conformation begins with a conjugate gradients minimization, continues with molecular dynamics with simulated annealing, and finishes by conjugate gradients again. There are two cycles of the conjugate gradients: molecular dynamics with simulated annealing, and conjugate gradients phases. In the first cycle, only those non-bonded atom pairs are considered that contain the set of atoms selected for optimization. In the second cycle, the atom pairs that contain up to one environment atom are also included in the energy function.

The prediction accuracies for the buried (and exposed) mutations are 76.3% (74.2%) for χ_1, 67.7% (63.6%) for χ_2, and 58.4% (51.4%) for χ_{1+2}, respectively.

The main difference between the different tested scoring functions was in the homology-derived terms. We attempted to incorporate terms in the scoring function that depend on the wild-type residue conformation and interactions between the mutated residue and its environment. In contrast to comparative modeling of whole structures, the modeling of a mutation is restrained by the environment of the mutated residue. Consequently if the environment is not modeled appropriately (e.g., when homology-derived distance restraints are strictly enforcing template-based conformations), modeling of the mutation is less accurate (Figs. 2, 3). As a result the highest success was achieved if the template-dependent dihedral angle preferences were used in combination of general statistical preferences while avoiding homology-derived distance restraints between the mutated side chain and its environment.

For the tested scoring functions and the sampling scheme, it is optimal to refine the whole mutated residue, but not any additional atoms in its environment. This result is consistent with the observation that the neighborhoods of most point mutations are essentially not distorted by the mutation (main-chain RMSD <0.4 Å). Therefore, refining the environment merely increases the demands on the scoring function to identify the correct conformation among many more decoys. More accurate modeling of a flexible environment may require both a more accurate scoring function and a more thorough sampling scheme.

Our most accurate protocol produces similar best-scoring conformations, even when they are incorrect. This observation suggests that the sampling is sufficient, but the scoring function cannot always identify the correct structure as the best-scoring one.

The prediction accuracy depends on the relative volume change during mutation. Neutral volume changes can be predicted most accurately, while a significant increase or decrease of the occupied volume upon mutation makes thorough sampling more difficult. Particularly difficult is the prediction of those mutations that increase significantly the volume (by >1000 Å^3).

Acknowledgments

This work has been supported in part by grants from NIH R01 GM54762 (A.S.), NIH U54 GM62529 (A.S., A.F.), The Sandler Family Supporting Foundation (A.S.), and hardware gifts from IBM, HP, Intel, and Netapp (A.S.).

References

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res.* **28:** 235–242. doi: 10.1093/nar/28.1.235.

Blundell, T.L., Sibanda, B.L., Sternberg, M.J., and Thornton, J.M. 1987. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* **326:** 347–352.

Bower, M.J., Cohen, F.E., and Dunbrack Jr., 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.

Canutescu, A.A., Shelenkov, A.A., and Dunbrack Jr., R.L. 2003. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* **12:** 2001–2014.

Chattopadhyay, K., Bhatia, S., Fiser, A., Almo, S.C., and Nathenson, S.G. 2006. Structural basis of inducible costimulator ligand costimulatory function: Determination of the cell surface oligomeric state and functional mapping of the receptor binding site of the protein. *J. Immunol.* **177:** 3920–3929.

Chothia, C. and Lesk, A.M. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5:** 823–826.

Chung, S.Y. and Subbiah, S. 1996. How similar must a template protein be for homology modeling by side-chain packing methods? In *Pacific Symposium Biocomputing* (eds. R.B. Altman et al.), pp. 126–141. Uniformed Services University of the Health Sciences, Bethesda, MD.

Cregut, D., Liautard, J.P., and Chiche, L. 1994. Homology modeling of annexin I: Implicit solvation improves side-chain prediction and combination of evaluation criteria allows recognition of different types of conformational error. *Protein Eng.* **7:** 1333–1344.

Desjarlais, J.R. and Handel, T.M. 1999. Side-chain and backbone flexibility in protein core design. *J. Mol. Biol.* **290:** 305–318.

Dunbrack Jr., R.L. and Cohen, F.E. 1997. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* **6:** 1661–1681.

Dunbrack Jr., R.L. and Karplus, M. 1993. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol.* **230:** 543–574.

Dunbrack Jr., R.L. and Karplus, M. 1994. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. *Nat. Struct. Biol.* **1:** 334–340.

Eisenmenger, F., Argos, P., and Abagyan, R. 1993. A method to configure protein side-chains from the main-chain trace in homology modelling. *J. Mol. Biol.* **231:** 849–860.

Fiser, A. 2004. Protein structure modeling in the proteomics era. *Expert Rev. Proteomics* **1:** 97–110.

Fiser, A., Do, R.K., and Sali, A. 2000. Modeling of loops in protein structures. *Protein Sci.* **9:** 1753–1773.

Fiser, A., Feig, M., Brooks III, C.L., and Sali, A. 2002. Evolution and physics in comparative protein structure modeling. *Acc. Chem. Res.* **35:** 413–421.

Gallicchio, E. and Levy, R.M. 2004. AGBNP: An analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. *J. Comput. Chem.* **25:** 479–499.

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.

Hu, X. and Kuhlman, B. 2006. Protein design simulations suggest that side-chain conformational entropy is not a strong determinant of amino acid environmental preferences. *Proteins* **62:** 739–748.

Hwang, J.K. and Liao, W.F. 1995. Side-chain prediction by neural networks and simulated annealing optimization. *Protein Eng.* **8:** 363–370.

Jacobson, M.P., Friesner, R.A., Xiang, Z., and Honig, B. 2002. On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol.* **320:** 597–608.

Jain, T., Cerutti, D.S., and McCammon, J.A. 2006. Configurational-bias sampling technique for predicting side-chain conformations in proteins. *Protein Sci.* **15:** 2029–2039.

Janin, J. and Wodak, S. 1978. Conformation of amino acid side-chains in proteins. *J. Mol. Biol.* **125:** 357–386.

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. 1997. The native sequence determines side-chain packing in a protein, but does optimal side-chain packing determine the native sequence? In *Proceedings of the Pacific Symposium on Biocomputing*, pp. 198–209. World Scientific, Singapore.

Ladurner, A.G. and Fersht, A.R. 1997. Glutamine, alanine or glycine repeats inserted into the loop of a protein have minimal effects on stability and folding rates. *J. Mol. Biol.* **273:** 330–337.

Lasters, I. and Desmet, J. 1993. The fuzzy-end elimination theorem: Correctly implementing the side-chain placement algorithm based on the dead-end elimination theorem. *Protein Eng.* **6:** 717–722.

Lee, C. and Levitt, M. 1991. Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. *Nature* **352:** 448–451.

Levitt, M., Gerstein, M., Huang, E., Subbiah, S., and Tsai, J. 1997. Protein folding: The endgame. *Annu. Rev. Biochem.* **66:** 549–579.

Liang, S. and Grishin, N.V. 2002. Side-chain modeling with an optimized scoring function. *Protein Sci.* **11:** 322–331.

Liu, R., Baase, W.A., and Matthews, B.W. 2000. The introduction of strain and its effects on the structure and stability of T4 lysozyme. *J. Mol. Biol.* **295:** 127–145.

Looger, L.L. and Hellinga, H.W. 2001. Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: Implications for protein design and structural genomics. *J. Mol. Biol.* **307:** 429–445.

MacKerell Jr., A.D., Bashford, D., Bellott, M., Dunbrack Jr., 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.

Matthews, B.W. 1995. Studies on protein stability with T4 lysozyme. *Adv. Protein Chem.* **46:** 249–278.

Melo, F. and Feytmans, E. 1998. Assessing protein structures with a nonlocal atomic interaction energy. *J. Mol. Biol.* **277:** 1141–1152.

Mendes, J., Soares, C.M., and Carrondo, M.A. 1999. Improvement of side-chain modeling in proteins with the self-consistent mean field theory method based on an analysis of the factors influencing prediction. *Biopolymers* **50:** 111–131.

Otzen, D.E. and Fersht, A.R. 1999. Analysis of protein-protein interactions by mutagenesis: Direct versus indirect effects. *Protein Eng.* **12:** 41–45.

Perozo, E., Cortes, D.M., and Cuello, L.G. 1999. Structural rearrangements underlying K^+-channel activation gating. *Science* **285:** 73–78.

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.

Petrella, R.J., Lazaridis, T., and Karplus, M. 1998. Protein sidechain conformer prediction: A test of the energy function. *Fold. Des.* **3:** 353–377.

Ponder, J.W. and Richards, F.M. 1987. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193:** 775–791.

Sali, A. and Blundell, T.L. 1993. Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* **234:** 779–815.

Sali, A. and Overington, J.P. 1994. Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci.* **3:** 1582–1596.

Shapovalov, M.V. and Dunbrack, R.L. 2007. Statistical and conformational analysis of the electron density of protein side chains. *Proteins* **66:** 279–303.

Sippl, M.J. 1990. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213:** 859–883.

Strausberg, R.L., Simpson, A.J., and Wooster, R. 2003. Sequence-based cancer genomics: Progress, lessons and opportunities. *Nat. Rev. Genet.* **4:** 409–418.

Tuffery, P., Etchebest, C., Hazout, S., and Lavery, R. 1991. A new approach to the rapid determination of protein side chain conformations. *J. Biomol. Struct. Dyn.* **8:** 1267–1289.

Vasquez, M. 1996. Modeling side-chain conformation. *Curr. Opin. Struct. Biol.* **6:** 217–221.

Wu, G., Fiser, A., ter Kuile, B., Sali, A., and Muller, M. 1999. Convergent evolution of *Trichomonas vaginalis* lactate dehydrogenase from malate dehydrogenase. *Proc. Natl. Acad. Sci.* **96:** 6285–6290.

Xiang, Z. and Honig, B. 2001. Extending the accuracy limits of prediction for side-chain conformations. *J. Mol. Biol.* **311:** 421–430.

Zhang, W. and Duan, Y. 2006. Grow to Fit Molecular Dynamics (G2FMD): An ab initio method for protein side-chain assignment and refinement. *Protein Eng. Des. Sel.* **19:** 55–65.