Importance of Solvent Accessibility and Contact Surfaces in Modeling Side-Chain Conformations in Proteins

ERAN EYAL, RAFAEL NAJMANOVICH,* BRENDAN J. MCCONKEY,† MARVIN EDELMAN, VLADIMIR SOBOLEV

Department of Plant Sciences, Weizmann Institute of Science, 76100, Rehovot, Israel

Received 10 July 2003; Accepted 3 November 2003

Abstract: Contact surface area and chemical properties of atoms are used to concurrently predict conformations of multiple amino acid side chains on a fixed protein backbone. The combination of surface complementarity and solvent-accessible surface accounts for van der Waals forces and solvation free energy. The scoring function is particularly suitable for modeling partially buried side chains. Both iterative and stochastic searching approaches are used. Our programs (Sccomp-I and Sccomp-S), with relatively fast execution times, correctly predict χ_1 angles for 92–93% of buried residues and 82–84% for all residues, with an RMSD of ~1.7 Å for side chain heavy atoms. We find that the differential between the atomic solvation parameters and the contact surface parameters (including those between noncomplementary atoms) is positive; i.e., most protein atoms prefer surface contact with other protein atoms rather than with the solvent. This might correspond to the driving force for maximizing packing of the protein. The influence of the crystal packing, completeness of rotamer library and precise positioning of C_β atoms on the accuracy of side-chain prediction are examined. The Sccomp-S and Sccomp-I programs can be accessed through the Web (http://sgedg.weizmann.ac.il/sccomp.html) and are available for several platforms.

 J Comput Chem 25: 712–724, 2004

Key words: solvent-accessible surface; surface complementarity; atomic solvation parameter; rotamer library; modeling

Introduction

Side-chain modeling is a necessary and important step in constructing complete structural models of proteins. Various approaches to construct structural models, such as comparative modeling, threading, and *ab initio* prediction usually assign side-chain location artificially as a subtask of the entire modeling procedure.[1] Modeling of side chains is also required for flexible molecular docking, for predicting local structural, and stability changes upon mutations, and as a tool to fill in missing information from experimental data (such as X-ray crystallography).

During the previous decade, the main focus of research in the field was on search procedures, following the common belief that the combinatorial nature of the problem is the main obstacle. A number of searching techniques were applied to this task, including dead end elimination,[2–4] self-consistent–based methods,[5–7] and Monte Carlo approaches.[8–11] Most of these used very simple energy functions, often just van der Waals and torsion terms. It has since been suggested that the combinatorial problem is not that severe in practice,[12] and the focus of research shifted towards the scoring function. More attention is now given to the solvation terms[13,14] and detailed rotamer libraries allow knowledge-based derivation of pseudoenergy values for intraresidue energy and local backbone interactions.[15–18] An optimized scoring function that includes geometric characteristics of interactions gives impressive results and was demonstrated to perform significantly better than standard force fields.[17]

Likewise, surface complementarity can be used to evaluate intermolecular interactions based on the geometry and chemical properties of individual atoms. The method was developed originally for molecular docking,[19] and has been applied to various predictions, such as modeling the quinone binding site in the D1 protein of the photosystem-2 reaction center[20] and the tentoxin binding sites of chloroplast F1-ATPase.[21] In the original implementation, each atom was assigned one of eight chemical types.[19] Weights for interactions between atoms were defined as favorable (complementary chemical contact) or unfavorable (noncomplementary chemical contact) according to the properties of the two

*Present address: European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, United Kingdom

†Permanent address: Department of Biology, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada

Correspondence to: V. Sobolev; e-mail: vladimir.sobolev@weizmann.ac.il

atoms involved. The complementarity score (CS) is the sum of the contacts surfaces S_{ij} multiplied by the appropriate weight W_{ij}:

$$CS = \sum_i \sum_j W_{ij} S_{ij}. \quad (1)$$

For binary scores, CS is therefore simply the sum of contact surface areas of noncomplementary atoms minus the sum of surface areas of complementary atoms. Other weight schemes for W_{ij} are also possible between pairs of atom classes. One such contact-based scoring scheme has recently been used by us to distinguish native proteins from mis-folded decoy structures.[22] The solvent-accessible area correlates well with free energy of solvation.[23,24] Therefore, surfaces in general might be used in scoring functions instead of pseudoenergy-based terms.

However, there are several energy components that clearly do not correlate with surfaces; for example, the torsion energy of a molecule. Such a term is very important for modeling conformations of flexible moieties such as ligands and side chains. Therefore, attempts to model such moieties using surface area require incorporation of other terms to approximate the torsion energy. Accepted procedures[25] include addition of a standard torsion energy term, or using a term based on probabilities derived from rotamer libraries.

Steric overlaps (repulsive van der Waals forces) are also not represented by surface areas. In these cases, a steep term, like the first component of the Lennard–Jones equation, is usually added. Another possibility is to add a term proportional to the excluded volume of the two spheres representing the atoms. Recently, it was shown that such a term gives higher accuracy in side chain modeling.[17]

It is known that optimum side chain conformation is very sensitive to the exact backbone structure.[26,27] This is especially noticeable for *ab initio* protein structure predictions where the main chains generated are often fairly distant (=2.0 Å C_α RMSD) from the native conformation.[28] However, inaccuracy in side-chain placement due to uncertainty in backbone structure can be less problematic in cases where backbone structure is found to be conserved, as in ligand docking[29] and point mutations.[30]

In this report, we develop a new scoring function based on surface complementarity that considers geometric and chemical compatibility and solvent-accessible surface. A function of this sort might be appropriate with only minor modification for both predicting structural changes and estimating stability of point mutations.

Methods

Protein Sets

Three sets of proteins were used (Table 1). Two sets were taken from the work of Liang and Grishim.[17] They are composed of proteins crystallized as monomers without ligands that have less than 50% pair-wise identity and a resolution equal or better than 1.8 Å. One was used as the training set (Set1) and the other as the test set (Set2). The training set consists of about 3000 residues undergoing conformational changes, and more than 10,000 atoms

Table 1. Protein Sets.[a]

PDB ID	Resolution (Å)	Residues		
		Total no.	Flexible[b]	Buried flexible
		Set 1		
1a8q	1.75	274	225	94
1amm	1.2	174	158	46
1bd8	1.8	156	121	36
1cem	1.75	363	292	135
1chd	1.75	199	154	57
1edg	1.6	380	329	143
1ifc	1.19	131	113	27
1mla	1.5	289	227	92
1nar	1.8	150	262	102
1npk	1.8	150	122	33
1thv	1.75	207	167	63
1vjs	1.7	480	391	170
2baa	1.8	243	178	67
2end	1.45	137	118	33
2pth	1.2	193	151	50
		Set 2		
153l	1.6	185	149	50
1ako	1.7	268	234	89
1arb	1.2	263	202	91
1bj7	1.8	150	135	44
1cex	1.0	197	146	53
1dhn	1.65	121	105	27
1hcl	1.8	298	259	90
1koe	1.5	172	144	50
1mml	1.8	251	221	66
1noa	1.5	113	80	22
1thx	1.7	198	97	33
1whi	1.5	122	101	29
2cpl	1.63	164	132	50
2hvm	1.8	273	221	92
2rn2	1.48	155	127	37
		Set 3		
1bgf	1.45	124	112	26
1bkrA	1.1	108	97	29
1byi	0.97	224	177	63
1ep0A	1.5	183	165	54
1es9A	1.3	212	183	66
1ey4A	1.6	136	114	33
1e5mA	1.54	411	315	140
1fcqA	1.6	321	276	107
1fo9	1.5	344	286	120
1gsoA	1.6	420	324	122
1ii5A	1.6	226	173	67
1jb3A	1.6	127	115	30
1ln4A	1.50	98	86	21
1l3kA	1.1	172	140	39
1qgvA	1.4	137	115	37
1t1dA	1.51	100	90	27
1wer	1.6	324	297	98
1bgf	1.45	124	112	26
2bce	1.6	532	436	198
2lisA	1.35	131	115	29

[a]Set1 and Set2 are from Liang and Grishin.[17] Set3 was compiled as described in the text.
[b]All residues except Gly and Ala.

that change position upon side-chain rotation. These atoms form about 50,000 atom–atom contacts, permitting statistical derivation of several parameters. In addition, we created another test set (Set3) based on the October 2002 version of the PISCES site.[31] Set3 has the following characteristics: the proteins share less than 20% identity, have an *R*-factor smaller than 0.25 and a resolution equal or better than 1.6 Å. In addition, any protein that shared more than 20% identity with a protein in Set1 and Set2, contained a ligand, multiple chains or less than 100 residues was removed. Following this, Set3 yielded a final group of 20 proteins.

Rotamer Libraries

The backbone-dependent rotamer library of Dunbrack and Cohen[32] was modified as suggested by Mendes et al.[5] such that each rotamer of His, Gln, and Asn is split to two new rotamers, one with the original torsion values and the other with the terminal χ flipped by 180°. A backbone independent library was used for rotamers of the first and last residue in the polypeptide chain, or for other locations where the ϕ- or ϕ-angles are not defined. Following analysis, a threshold of 0.003 was chosen as the lowest probability for a rotamer to be included in the search, yielding about 200 rotamers for each backbone conformation (without the threshold the number was 338). Set2 was used to examine the suitability of the modified library. It was found that 98% of the χ_{1+2} values are included. The modified library was refined for aromatic residues (His, Phe, Trp, and Tyr) and residues with only one side-chain dihedral angle (Ser, Thr, and Val) by adding conformations $\pm 15°$ off of the rotamer library values. For the small amino acids this expansion does not cost much in computational time, while for the aromatic residues it might be important for prediction accuracy (due to the rigid planar rings, which cause large spatial displacements from small angle differences). If, during the search, side-chain clashing was detected for any type of amino acid, this refinement was included for the χ_1 and χ_2 dihedral angles.

Bond distances and angles used to build rotamers were taken from CHARMM.[33]

Calculation of Surfaces

Our definition of contacting atoms is illustrated in Figure 1. Atom *a* in the protein matrix can have contact with atom *b* if $d_{ab} < R_a + R_b + 2R_w$; i.e., the distance separating them is less than the sum of their van der Waals radii (R_a and R_b) plus two solvent (probe) atom radii (R_w). Traditionally, R_w is equal to a water molecule radius of 1.4 Å. A modified version of a Voronoi tessellation was used to calculate contact surface areas between atoms.[34] This method analytically calculates the contact surfaces following projection of the polyhedra edges on the extended sphere radius (van der Waals plus solvent atom radii). The solvent-accessible area[35] of each atom is what is left after subtraction of the atomic contacts from the sphere surface. This procedure was also used to calculate solvent-accessible surfaces of the side chains in the native structures.

Atom Types

The protein atom types were divided into eight groups as in Sobolev et al.[19] A pairwise interaction between atoms of different

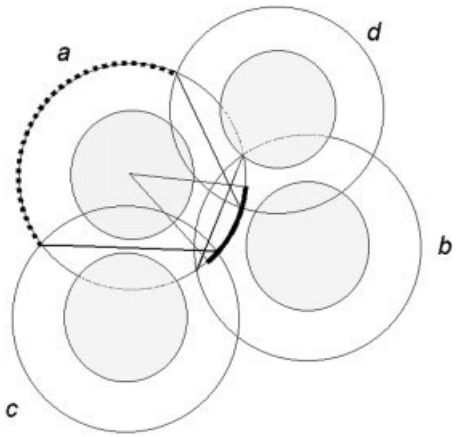

Figure 1. Definition of atomic contacts. Atom ***a*** is in surface contact with atom ***b*** if the distance between the two is less than the sum of their van der Waals radii (gray areas) and two solvent-atom radii (white areas). The contact surface area of atoms ***a*** and ***b*** (represented by the thick solid line) depends on the spatial positions of the additional atoms in contact with both (i.e., atoms ***c*** and ***d***). It can be determined by bisecting the overlap of atoms ***a*** and ***b,*** ***a*** and ***c,*** and ***a*** and ***d*** and projecting lines from the center of atom ***a***, through the intercepts of bisects ***ab*** with ***ac*** and ***ad,*** to the edge of the extended sphere of ***a***. The total surface area of an atom is divided among the contacting atoms and the solvent-accessible surface (SAS, represented for atom ***a*** by the thick dotted line). SAS is determined by implementing a constrained Voronoi procedure.[34]

classes is considered as either favorable (weight of -1) or unfavorable (weight of $+1$). A description of the classes and the weights is given in Table 2. Interaction with the solvent was optimized using a genetic algorithm and a weight of 2 was obtained, as will be described.

Scoring Function

The scoring function for a given side chain conformation takes the general form of:

$$E_{\text{score}} = E_{\text{comp}} + K_{\text{vol}} \cdot E_{\text{vol}} + K_{\text{prob}} \cdot E_{\text{prob}} + K_{\text{sol}} \cdot E_{\text{sol}} \qquad (2)$$

where

$$E_{\text{comp}} = \sum_{a}^{sc} \sum_{b}^{all} W_{\text{ab}} S_{\text{ab}}, \qquad (3)$$

$$E_{\text{vol}} = \sum_{a}^{sc} \sum_{b}^{all} V_{\text{ab}}, \qquad (4)$$

$$E_{\text{sol}} = \sum_{a}^{sc} ASP_a SAS_a, \qquad (5)$$

Table 2. Atom Types and Contact Weights of the Binary-Surface Complementarity Function.[a]

Atom types	I	II	III	IV	V	VI	VII	VIII
Hydrophilic (I)	−	−	−	+	−	−	−	−
Acceptor (II)	−	+	−	+	−	−	−	+
Donor (III)	−	−	+	+	−	−	+	−
Hydrophobic (IV)	+	+	+	−	−	−	−	−
Aromatic (V)	−	−	−	−	−	−	−	−
Neutral (VI)	−	−	−	−	−	−	−	−
Neutral donor (VII)	−	−	+	−	−	−	+	−
Neutral acceptor (VIII)	−	+	−	−	−	−	−	+

[a]Adapted from Sobolev et al.[19] All contact weights are ±1. A designation of "+" represents an unfavorable interaction and "−," a favorable one.

and W_{ab} are the binary weights for surface contacts, S_{ab} is the contact surface of atoms a and b, V_{ab} is the overlapping volume of atoms a and b, ASP_a is an atomic solvation parameter for atom a, and SAS_a is the solvent-accessible surface. E_{prob} is the intraresidue energy, while the Ks are the coefficients calibrating the relative contribution of each term. The numeration of a is over the side-chain heavy atoms excluding C_α, while the numeration of b is over all atoms having contact with side chain atoms. Our scoring function [eq. (2)] contains three parts corresponding to atom–atom interactions (first two terms), a torsion term that depends on dihedral angles (the third term), and a term corresponding to the hydrophobic (or solvation) effect (forth term). Thus, our function is structurally similar to potential energy functions for systems with constant covalent bonds and angles.

Surface Complementarity (E_{comp}) and Solvation (E_{sol}) Terms

For each residue, the complementarity part of the scoring function is calculated according to eq. (3). Two contacting atoms need to be separated by at least four covalent bonds as in standard force fields. It should be noted that the complementarity score for two given atoms may also be dependent on the position of other atoms in space, because it depends on the contact surface allocated to all atoms in contact with a given atom and the solvent. For that reason, it cannot be calculated by analyzing the coordinates of the two atoms alone. As a result, we cannot use a sophisticated search algorithm such as the dead-end elimination[36] in combination with surface complementarity.

A simple solvation term was used based on the SAS [eq. (5)] and *ASP*s were optimized, as will be described in following sections.

Excluded Volume (E_{vol}) Term

This repulsion term is calculated as the volume of overlap of spheres for two atoms having van der Waals radii R_a and R_b. In our study, we use an analytical solution to compute the volume overlap between two atoms a and b as follows:

$$V_{ab} = \tfrac{1}{3}\pi h_a^2(3R_a - h_a) + \tfrac{1}{3}\pi h_b^2(3R_b - h_b) \tag{6}$$

where

$$h_a = \frac{R_b^2 - (d - R_a)^2}{2d}, \quad h_b = \frac{R_a^2 - (d - R_b)^2}{2d} \quad \text{if } (d < R_a + R_b),$$

$$h_a = 0, \quad h_b = 0 \quad \text{if } (d \geq R_a + R_b). \tag{7}$$

where d is the distance between atoms a and b.

Intraresidue Energy (E_{prob}) Term

The intramolecular energy of the residue is an important determinant of side-chain conformations in proteins.[25,37] It could be considered simply by a standard torsion term. However, with recently improved and more detailed rotamer libraries, data regarding commonality and probabilities of rotamers in native proteins have come into use.[15–18] The intraresidue energy term assigns a more favorable (smaller) score for rotamers present more frequently in proteins. We applied here a probabilistic term of the form

$$E_{\text{prob}} = Na_{\text{res}} \cdot \ln(P_{\text{rot}} \cdot Nr_{\text{res}}) \tag{8}$$

where P_{rot} is the probability of the rotamer taken from the backbone dependent rotamer library of Dunbrack and Cohen,[32] Nr_{res} is the number of entries in the rotamer library for residue type res, and Na_{res} is the number of flexible bonds of residue type res. This number might be important in order to fit E_{prob} with other terms in eq. (2). The latter act on the atom level and their contribution is roughly proportional to residue size. Multiplication by Na_{res} makes E_{prob} dependent on residue size as well.

Optimization of Parameters Using a Genetic Algorithm

The relative contribution of various terms in the scoring function was calibrated by implementing a genetic algorithm. The root-mean-square deviation of side chain atoms between the model and native structures was minimized. Each fitness evaluation included modeling one of 2983 side chains in the training set (Set1), while all others were held fixed in the native conformation. The population size was set to 100 individuals, and 150 generations were simulated. The bounds on the parameter values were: excluded volume coefficient (K_{vol}, 0 . . . 1000); solvation coefficient (K_{sol},

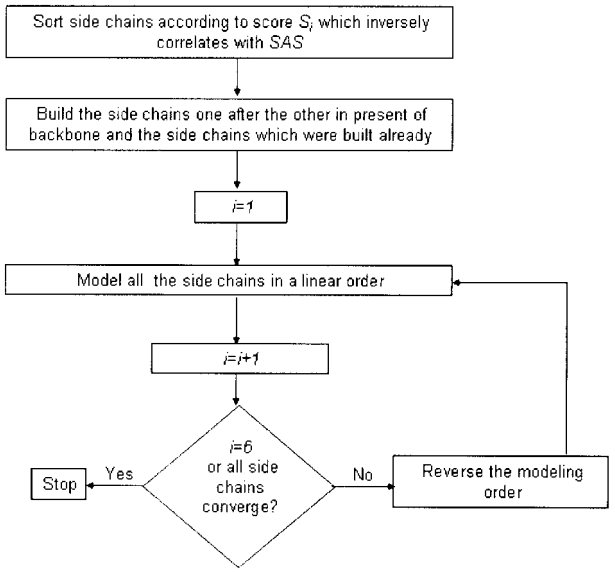

Figure 2. Flow chart of the iterative algorithm for concurrent modeling of all side chains. In the first iteration, side chains are modeled successively in decreasing order of S_i score [eq. (9)], in the presence of the backbone and all previously built side chains. Further iterations are performed similarly except that the modeling order of side chains is reversed in each round. Reiteration is continued until each side chain takes the same conformation in two successive iterations, or until a predetermined number of iterations is reached.

$-10 \ldots 10$); intraresidue energy coefficient (K_{prob}, $-100 \ldots .0$), and atomic solvation parameters (*ASPs*, $-10 \ldots 10$).

Iterative Modeling of Multiple Side Chains

In the iterative self-consistent procedure (Fig. 2), the side chains are modeled in a predetermined order. When a given side chain is modeled, the remaining side chains are held fixed. All side chains are positioned once during an iteration of the algorithm. The procedure is repeated until all side chains converge to the same conformation or until a predetermined maximum number of iterations are reached.

Such an algorithm can obviously be trapped in local minima. Two crucial factors will contribute to its performance. The first is the initial conformations of the side chains, and the second is the order in which they are modeled. Xiang and Honig[12] implemented their procedure with many different random starting conformations and chose the protein model with the lowest energy. In their implementation, the order of modeling was not considered. Here, we run our procedure once and the initial conformations are built one after the other using the scoring function. Every side chain is constructed in the presence of the backbone atoms (including C_β), and all other side chains already built. The order by which the side chains are initially built is a function of their estimated solvent exposure. Side chains were sorted by decreasing number of neighboring residues normalized to side chain size, as follows:

$$S_i = \frac{N_i}{L_i} \tag{9}$$

where S_i is the normalized score of residue i, N_i is the number of neighbors of residue i, and L_i is a characteristic property of each amino acid type roughly equal to the side chain length in extended conformation (from the C_α to the farthest side-chain atom). Two residues, i and j, are considered neighbors if the distance between their C_α atoms is smaller than the sum of their lengths L_i and L_j plus two solvent atom radii. The index S_i was found to correlate with the solvent-accessible area ($r = -0.68$).

After all side chains are initially modeled, the iterative procedure is applied. In each successive iteration, the order in which the side chains are considered is reversed to reduce the possibility of being stuck in a local minimum. This was empirically found to give better results than using an invariant order. The procedure continues until all side chains have approximately the same conformation in two successive iterations, or until a predefined number of iterations is reached. In the latter case, if the total score is higher than the penultimate iteration, the algorithm proceeds until the total energy at the end of an iteration is lower than the total energy at the end of the previous one.

Stochastic Modeling of Multiple Side Chains

The stochastic method used is schematically described in Figure 3. This type of algorithm has been referred to as Gibbs sampling or Heat–Bath algorithm for side chain modeling.[7] In the basic repeated step, a rotamer of a given side chain is chosen at random based on a Boltzmann distribution of the energies of all the

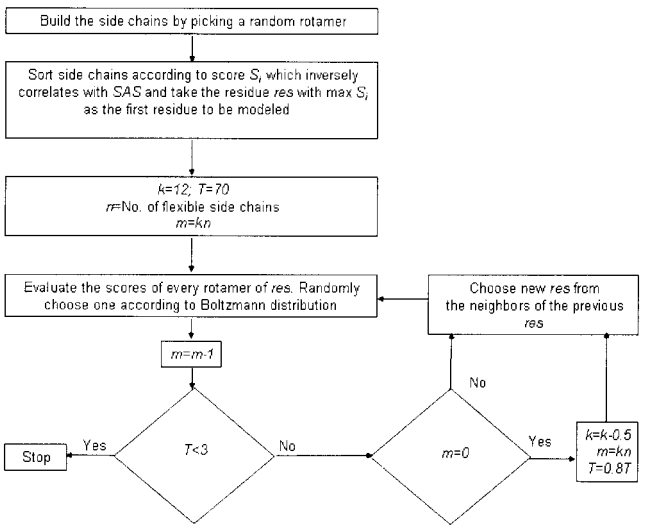

Figure 3. Flow chart of the stochastic algorithm for modeling all side chains concurrently. Side chains are initially built by randomly selecting conformations from the rotamer library. Side chains are then modeled in random order (starting from the one with the highest S_i score), the only restriction being that a given side chain is a neighbor of the previously modeled one. The rotamer chosen at each modeling step is also selected randomly, but with a probability that is exponentially related to its score [eq. (2)] according to the Boltzman distribution [eq. (10)]. The number of modeling steps is proportional to the number of flexible side chains (N). Temperature (T) decreases during the run. More modeling steps take place at higher temperatures, as determined by the parameter k.

rotamers of that side chain, and a time-dependent temperature parameter T. Initially, the side chains have a random conformation. A starting residue (in our implementation, that with the highest S_i score) is chosen and modeled, following which the next residue is chosen at random from its neighbors. Choosing a neighboring residue of the one modeled in the previous step, results in buried residues being sampled more often by the algorithm.

During the modeling of a given side chain, each rotamer j has probability P_j to be accepted, as given by:

$$P_j = \frac{e^{-E_j/T}}{Z} \tag{10}$$

where

$$Z = \sum_j e^{-E_j/T} \tag{11}$$

and E_j is the total score of rotamer j. The initial temperature T was set to 70 K and was scaled by 0.8 every $k \cdot N$ steps, where N is the total number of flexible residues. The algorithm stops when $T = 3$. The value of k decreases gradually (from 12 to 5), such that more modeling steps take place at higher temperatures.

Evaluation Methods

The percentage of correctly predicted χ_1 and χ_{1+2} angles are presented. We consider a correctly predicted angle to be one differing by less than 40° from the angle in the native structure. When considering residues with symmetric terminal groups (Asp, Glu, Phe, and Tyr), or with possibly flipped terminal groups (Asn, Gln, His), the torsion angle that is closer to the native value is taken for the angle comparison and the smaller value for RMSD calculations. The RMSD is calculated for all the protein side chain heavy atoms. C_β atoms were excluded from the calculation. Note that an overall RMSD of 1.5 Å with inclusion of C_β atoms is equal to about 1.7 Å without C_β. The C_β atoms were rebuilt by the programs evaluated and were not considered as part of the backbone.

Buried side chains are defined as those with less than 10% solvent accessibility (solvent radius taken as 1.4 Å), partially buried side chains as those with 10–50% solvent accessibility and exposed side chains as those with more than 50% accessibility.

The evaluation of all programs (ours and others) mentioned in this study was performed using identical evaluation criteria, on the same protein sets and on the same platform (Pentium III, 850 MHz, 512 Mb).

Results and Discussion

Incorporation of Surface Complementarity and Accessible Surface in a Scoring Function

We developed a scoring function in the general form of eq. (2) to predict side-chain conformations. The surface complementarity term [eq. (3)] is the core of the function. A solvation term [eq. (5)] was incorporated based on the solvent contact surface area. Solvation is usually not included in side-chain prediction programs due to the relatively extensive computational time required for its calculation. In our case, however, it is obtained as a side product from calculation of the contact surfaces between atoms. An excluded volume term [eqs. (4) and (6)] for clashing atoms and an intraresidue energy term [eq. (8)] were also included.

The parameters of the combined function were calibrated beginning with the coefficients of the different terms of the scoring function [eq. (2)]. The binary pair-wise weight parameters were applied (Table 2) and the ASPs at this stage were fixed at +1. A genetic algorithm program was used then to calibrate the coefficients K_{vol}, K_{prob}, and K_{sol}. The overall RMSD between predicted and native conformations was optimized for the residues of the proteins in Set1. During analysis of a given side chain, the other side chains were held fixed in the native conformation. This method of evaluation[38–40] avoids the combinatorial problem and the influence of the search procedure. We used the backbone dependent library of Dunbrack and Cohen[32] for the search, with modifications as described in Methods. The top solutions of the genetic algorithm all converged to about the same values, allowing

Table 3. Optimization of Atomic Solvation Parameter Values Using a Genetic Algorithm.[a]

Hydrophilic (I, II, III)	Hydrophobic (IV)	Aromatic (V)	Neutral (VI, VII, VIII)	RMSD
2.91	1.34	2.60	3.39	1.307
2.91	1.02	2.44	3.39	1.307
2.91	1.34	2.44	3.39	1.308
1.65	2.44	2.76	2.44	1.308
1.65	2.44	2.60	2.44	1.308
1.65	2.44	2.44	2.44	1.308
3.23	0.71	2.44	3.23	1.308
1.65	1.97	2.76	2.44	1.309
1.65	1.97	2.60	2.44	1.309
3.23	0.55	2.44	3.23	1.309

[a]The eight atom types of Table 2 were clustered into four groups as shown. The 10 most-fit ASP sets that were generated are presented. Fitness is a function of the root mean square deviation (RMSD) of the model from the native structure (average RMSD for random ASP values is 1.726 Å).

Table 4. Scoring Function Performance.[a]

	No. of residues		χ_1 (%)		χ_{1+2} (%)		RMSD (Å)	
Amino acid	Buried	All	Buried	All	Buried	All	Buried	All
Arg	17	153	94.1	86.3	88.2	73.2	1.99	2.55
Asn	26	134	100.0	89.6	92.3	72.4	0.52	1.14
Asp	28	157	100.0	90.4	67.9	66.2	0.62	1.09
Cys	36	52	100.0	98.1	100.0	98.1	0.32	0.43
Gln	17	111	100.0	85.6	100.0	73.0	0.54	1.62
Glu	8	142	87.5	78.9	62.5	59.9	1.81	1.85
His	15	57	100.0	96.5	100.0	94.7	0.46	0.96
Ile	108	161	100.0	96.3	94.4	88.2	0.38	0.72
Leu	159	242	96.8	95.5	84.9	82.6	0.86	0.95
Lys	4	156	100.0	84.6	100.0	67.9	0.81	2.03
Met	27	39	96.3	92.3	88.9	82.1	0.80	1.22
Phe	69	106	97.1	98.1	95.7	93.4	0.79	0.71
Ser	56	185	78.6	65.4	78.6	65.4	1.11	1.38
Thr	42	172	97.6	89.0	97.6	89.0	0.43	0.83
Trp	23	49	100.0	100.0	100.0	100.0	0.40	0.40
Tyr	47	96	100.0	96.9	97.9	94.8	0.88	1.12
Val	120	184	95.0	94.0	95.0	94.0	0.63	0.68
Pro	20	142	85.0	80.3	80.0	71.8	0.38	0.46
Total	*822*	*2338*						
Average			*96.1*	*88.5*	*90.8*	*79.2*	*0.81*	*1.42*

[a]Set2 was used. The side chains of all residues apart from the one being modeled were held fixed in the native conformation. buried = buried side chains, all = all side chains.

us to fix the values of the parameters as $K_{vol} = 70$, $K_{prob} = -10$ and $K_{sol} = 2$.

The solvation term was refined next. Using atomic solvation parameters for this purpose is well accepted in side chain modeling.[13,40–42] In some cases it was found that the ASP sets of Eisenberg and McLachlan[24] improved prediction, but this was demonstrated only on very small samples,[41,42] or on well-defined, exposed residues taken from NMR data.[13] Others did not show the contribution of the ASP sets to the overall performance[40] or could not improve their method using this ASP model.[6]

Although still popular, a growing body of evidence suggests that ASP sets derived from physicochemical experiments (e.g., refs. 23, 24, and 43) are problematic for molecular modeling of proteins. Such sets contain both negative and positive ASPs. Usually the C and S parameters are the only positive ones, while all the O and N atoms have negative weights. However, uniform positive values of ASP give better results for protein structure refinement than physicochemical ASP sets.[44] Indeed, when Schiffer et al.[41] combined a solvation model with AMBER force field and optimized the ASPs to obtain the minimum driving force for the native structures, they obtained only positive values of ASPs for N and O (but did not utilize this result). Also, it was recently shown that positive values for *all* ASPs are required to interpret binding free energy of quaternary complexes.[45] Likewise, variable physicochemical ASPs were found to be problematic for molecular docking[46] and for accurate calculation of folding free energy.[47] More recently, it was shown that a "stability scale" for both hydrophilic and hydrophobic residues correlates well with average "buried accessible surface,"[48] leading to near uniform ASP values for all atom types. Indeed, using a knowledge-based distribution of atom–atom and atom–solvent contacts, McConkey et al.[22] has now shown that the interaction with solvent is statistically unfavorable for 165 of 167 atom types in proteins (the exceptions are the terminal N of lysine and the terminal O of glutamate). Uniformly positive ASPs are, in fact, consistent with early studies that found a linear correlation between accessible

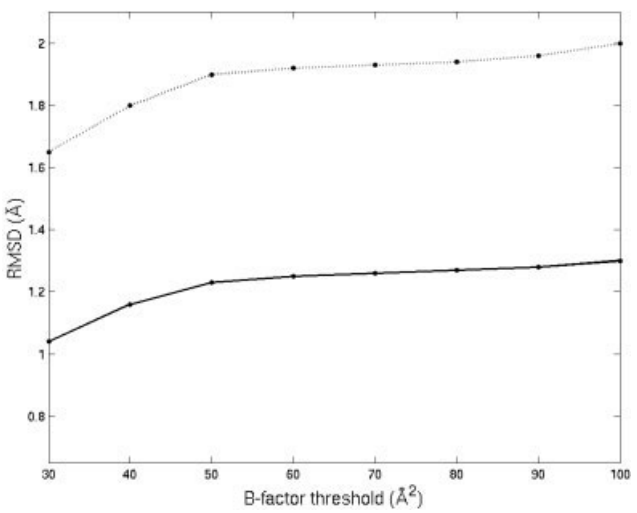

Figure 4. The performance of the scoring function (RMSD) vs. the upper limit of B-factor values. The numbers on the x axis represent the threshold values for a residue to be included in the evaluation. Continuous line: all residues. Dotted line: exposed residues only.

Table 5. Contribution of the Scoring Function Terms to the Prediction Accuracy.[a]

Terms					χ_1 (%)				χ_{1+2} (%)				RMSD (Å)			
Vol	Com	Sol	Int	6–12	Bur	Part bur	Exp	All	Bur	Part bur	Exp	All	Bur	Part bur	Exp	All
+	−	−	−	−	93.4	77.3	66.4	80.5	88.2	63.3	49.7	69.1	0.81	2.03	2.34	1.78
+	+	−	−	−	94.0	84.3	67.2	83.3	89.2	73.8	50.2	73.2	0.77	1.61	2.34	1.61
+	−	+	−	−	93.5	89.5	71.3	86.0	88.6	77.5	53.3	75.1	0.74	1.24	2.26	1.45
+	−	−	+	−	95.0	85.0	74.3	85.9	91.2	73.5	59.2	76.3	0.81	1.75	2.24	1.63
+	+	+	−	−	94.1	91.1	71.8	86.9	89.5	80.9	53.8	76.7	0.71	1.15	2.27	1.42
+	−	+	+	−	95.5	90.7	78.2	89.1	91.8	80.7	64.4	80.4	0.72	1.20	2.09	1.36
+	+	−	+	−	95.5	87.8	76.2	87.5	91.6	77.8	61.6	78.6	0.75	1.49	2.17	1.50
+	+	+	+	−	95.2	92.2	78.7	89.6	91.6	82.5	64.7	81.1	0.71	1.12	2.00	1.30
−	−	−	−	+	91.3	85.7	68.5	83.0	84.9	72.4	46.4	70.0	0.84	1.60	2.46	1.65
−	−	+	+	+	93.2	90.3	76.0	87.5	89.0	79.1	59.0	77.3	0.76	1.17	2.10	1.37

[a]The scoring function was tested using different combinations of its components. The volume term of eq. (2) was always included to avoid atom clashing. The function with traditional 6–12 Lennard–Jones terms was also tested.
+ indicates inclusion of the parameter, − indicates exclusion. Abbreviations: Vol = excluded volume, Com = complementarity term, Sol = solvent accessible surface term, Int = intraresidue energy, bur = buried side chains, part bur = partially buried side chains, exp = exposed side chains, all = all side chains.

surface area and hydrophobicity,[49] while reduction of accessible surface area during folding was almost equal for polar and nonpolar atoms.[35]

In light of all this, we decided not to adopt an experimental ASP set from the literature but, instead, to optimize an ASP set for our specific purpose. Because the coefficient of the complementarity term [E_{comp} in eq. (2)] is 1, we set K_{sol} at this stage also to 1. This allows direct comparison between the optimized ASPs and the parameters for the contact surface area [W_{ab} in eq. (3)]. The number of ASPs was reduced from 8 to 4 by separately grouping the three hydrophilic and three neutral classes in Table 2. Optimization of the resulting ASPs produced a small reduction (1.321 to 1.307 Å) in the RMSD of the top solutions. However, there was no convergence of the ASP values to a single solution (Table 3). We noticed that all the ASPs had positive values, almost always larger than 1 (the weight of contacts between chemically noncomplementary atoms). In terms of contact surface area, intermolecular inter-

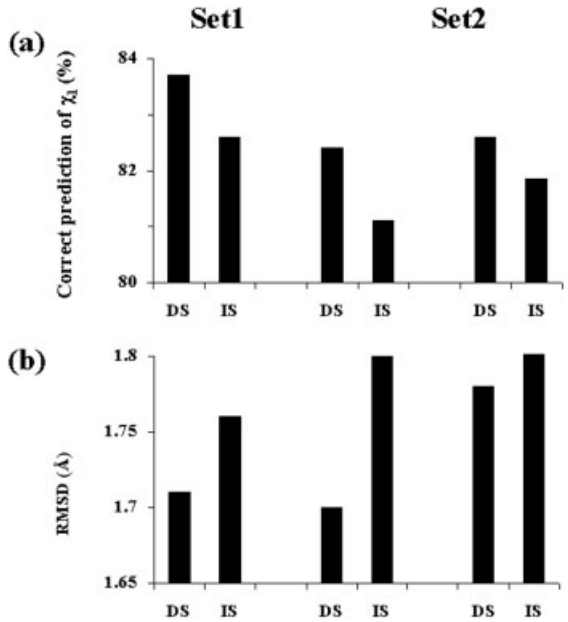

Figure 5. Prediction accuracy as a function of the order of modeling side chains. Side chains were modeled in decreasing (DS) or increasing (IS) order of S scores. (a) χ_1 prediction accuracy; (b) RMSD of side chain atoms.

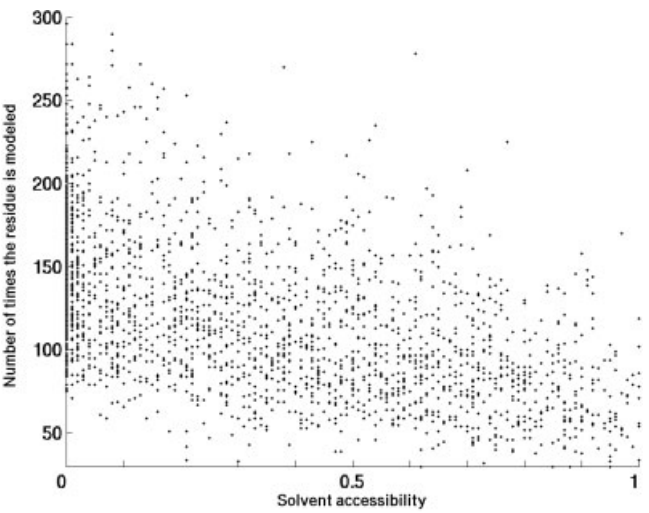

Figure 6. The number of times each side chain is modeled in the stochastic procedure as a function of its solvent accessibility. At any given time, the residue to be modeled is randomly chosen from the neighbors of the previously modeled one. Buried residues (which have more neighbors) are modeled more often. The number of times each side chain is modeled is plotted as a function of its solvent accessibility (correlation coefficient $r = -0.57$).

Table 6. Performance of Modeling Programs.[a]

	Iterative						Stochastic					
	χ_1 (%)		χ_{1+2} (%)		RMSD (Å)		χ_1 (%)		χ_{1+2} (%)		RMSD (Å)	
PDB ID	Bur	All	Bur	All	Bur	All	Bur	All	Bur	All	Bur	All
						Set 2						
153l	94.0	82.6	86.0	71.8	0.94	1.91	94.0	83.2	88.0	74.5	0.81	1.51
1ako	89.9	84.6	86.5	73.1	1.34	1.87	94.4	85.0	87.6	71.4	1.00	1.82
1arb	86.8	86.6	85.7	82.7	1.06	1.45	91.2	90.6	89.0	84.2	1.06	1.35
1bj7	93.2	81.5	86.4	71.1	1.48	1.78	93.2	83.0	81.8	71.9	1.55	1.72
1cex	98.1	86.3	94.3	74.0	0.52	1.67	98.1	90.4	98.1	80.8	0.39	1.49
1dhn	96.3	75.2	85.2	63.8	1.08	2.13	92.6	81.0	88.9	68.6	0.86	1.82
1hcl	86.7	77.6	71.1	59.5	1.32	1.90	83.3	75.7	66.6	57.5	1.44	1.98
1koe	92.0	84.0	90.0	76.4	1.01	1.65	92.0	84.0	90.0	77.1	0.70	1.59
1mml	87.9	80.5	81.8	70.6	0.90	1.57	89.4	82.4	81.8	69.7	0.90	1.45
1noa	95.5	78.8	90.9	73.8	0.69	1.10	95.5	80.0	90.9	72.5	0.71	1.18
1thx	97.0	79.4	93.9	70.1	1.11	1.49	100	82.5	97.0	75.3	0.52	1.25
1whi	93.1	82.2	93.1	73.3	0.64	2.05	93.1	80.2	93.1	73.3	0.64	2.15
2cpl	94.0	83.3	92.0	75.8	0.68	1.63	96.0	87.1	90.0	78.0	0.62	1.59
2hvm	93.5	84.2	88.0	76.0	0.77	1.22	93.5	83.7	89.1	76.5	0.75	1.31
2rn2	100	86.6	97.3	76.4	0.95	1.61	97.3	87.4	91.9	74.8	0.78	1.76
ave	*93.2*	*82.2*	*88.1*	*72.6*	*0.97*	*1.67*	*93.6*	*83.7*	*88.3*	*73.7*	*0.85*	*1.60*
						Set 3						
1e5mA	92.1	85.1	85.7	75.6	1.04	1.67	93.6	84.4	86.4	74.9	0.80	1.66
1fo9	98.3	84.3	87.5	72.4	1.45	1.87	98.3	86.4	90.8	76.9	0.59	1.63
1l3kA	92.3	85.7	92.3	78.6	0.82	1.77	97.4	87.1	97.4	80.0	0.82	1.70
1wer	93.9	78.8	90.8	71.7	1.05	1.86	98.0	80.8	96.9	73.4	0.61	1.69
1ep0A	90.7	80.6	85.2	72.1	1.18	1.76	98.1	81.8	88.9	71.5	0.74	1.78
1ey4A	93.9	79.8	87.9	64.9	0.78	2.01	90.9	80.7	81.8	65.8	0.83	2.01
1fcqA	93.5	81.9	85.0	69.6	0.95	1.91	93.5	83.3	86.0	72.8	0.95	1.76
1tldA	92.6	85.6	88.9	75.6	0.72	1.46	92.6	84.4	92.6	74.4	0.65	1.53
1bkrA	96.6	85.6	89.7	76.3	0.60	1.49	96.6	83.5	89.7	74.2	0.61	1.30
1es9A	92.4	78.7	83.3	69.4	1.90	2.11	93.9	83.1	84.8	72.1	0.91	1.70
1byi	93.7	83.6	90.5	72.9	0.66	1.56	93.7	83.6	90.5	73.4	0.66	1.52
2lisA	100	85.2	86.2	73.9	0.56	2.10	100	84.3	86.2	73.9	0.56	2.09
2bce0	89.4	80.3	83.8	71.8	1.40	1.66	89.9	82.1	83.3	72.7	1.20	1.48
1qgvA	83.8	76.5	78.4	64.3	1.08	2.09	83.8	79.1	78.4	67.8	1.08	1.93
1gsoA	91.8	84.6	83.6	74.4	0.79	1.39	90.2	82.4	83.6	72.8	0.83	1.45
1jb3A	90.0	78.3	86.7	67.8	1.25	2.14	90.0	80.0	90.0	69.6	1.21	2.10
1hztA	85.4	84.0	85.4	72.5	1.47	1.73	90.2	85.5	85.4	73.3	1.07	1.59
1ii5A	92.5	85.0	86.6	78.0	0.81	1.57	92.5	85.0	91.0	80.3	0.69	1.39
1bgf	88.5	77.7	76.9	67.0	2.82	2.03	92.3	75.9	84.6	67.0	1.69	1.94
1ln4A	95.2	90.7	90.5	74.4	0.79	1.71	95.2	90.7	95.2	75.6	0.70	1.70
ave	*92.3*	*82.6*	*86.2*	*72.2*	*1.11*	*1.79*	*93.5*	*83.2*	*88.2*	*73.1*	*0.86*	*1.70*

[a]Side chains of proteins were modeled concurrently. C_β was not included in the RMSD calculation. Abbreviations: bur = buried side chains, all = all side chains, ave = average.

action is therefore always favorable over interaction with the solvent. We checked if particular solutions in Table 3 are more suitable for modeling a specific group of residues (aromatic, polar, exposed, etc.). We found this not to be the case. Thus, the percentage of correctly predicted side chains of different residue types does not depend on the solution chosen. In summary, we did not find preferences for solvent interaction for certain atoms types over others by this procedure. The only general observation is that strong positive (almost uniform) values are favored for all atom types for side-chain modeling. We decided to stay with the initial value of 2 for all the ASPs (while $K_{sol} = 1$). Together, the weights for atomic and solvent contacts favor maximum packing of the protein. Our scoring function therefore takes the final form of:

$$E_{res} = E_{comp} + 70 \cdot E_{vol} - 10 \cdot E_{prob} + 1 \cdot E_{sol} \qquad (12)$$

where E_{sol} is equal to two times the solvent-accessible surface [i.e., ASPs in eq. (12) equal 2]. This function includes only three optimized parameters; attempts at further refinement did not improve the results.

An implicit parameter for both the surface complementarity and solvation terms is the radius of the solvent (probe) atom (R_w).

During our efforts to improve performance we tested different values of R_w in the range 0.0–1.8 Å. We found no real difference in the quality of the prediction when changing this parameter over entire range. Together with Table 5, these results demonstrate that surface complementarity does not contribute much to the prediction of buried side chains. This is not surprising, because the main determinant of core side chains is simply the allowed space, which is determined by the volume term. For *partially buried* and *exposed* side chains, however, we surprisingly found advantage for R_w in the range of 0.7–0.9 Å. The physical meaning of this is not clear.

To study the behavior of our function, we calculated the scores for binary systems of two atoms as a function of interatomic distance using the parameters obtained for the first two terms in eq. (12) (corresponding to atom–atom interactions). The minimum scores are at a distance slightly less than the sum of the van der Waals radii, similar to that in the Lennard–Jones potential. However, at short distances, our function is less steep (and therefore more permissive).

Our scoring function is not presently geared to a direct estimation of free energy. Several physical forces, such as electrostatics, are not represented accurately enough, and entropy is not considered at all. Our function is currently applicable for structural modeling tasks. If desired, another scoring function should be used for free energy estimation, as in the FlexX docking program.[50]

The combined scoring function was tested on a set of 15 protein structures (Set2) by modeling single side chains one at a time, in the same way as was done with the training set (Set1) during optimization. For Set2, our prediction was 96.1% accurate for χ_1 of buried residues and 88.5% for all residues. For χ_{1+2}, the accuracy was 90.8 and 79.5%, respectively, and the RMSDs were 0.80 and 1.42 Å, respectively (Table 4). Similar results were obtained for the proteins of Set3 (for χ_1, a prediction accuracy of 95.9% for buried residues and 87.6% for all residues; for χ_{1+2}, an accuracy of 91.9 and 79.0%, respectively, and RMSDs of 0.69 and 1.45 Å, respectively).

The performance of the scoring function was tested separately for well-defined side chains by considering B-factor values. Prediction accuracy increased when the evaluation was restricted to residues with lower B-factors (Fig. 4). An improved accuracy is also observed when checking exposed residues separately (Fig. 4, dotted line). Therefore, improved accuracy is not a consequence of an increased fraction of buried residues (which are more correctly predicted) in the lower B-factor population. This experiment suggests that some of the "error" in side chain placement reflects flexibility in the positions of the involved residues or uncertainty in their precise location.

We further checked the contribution of every term in the scoring function. From Table 5 it is clear that the intraresidue energy significantly contributes to the prediction of exposed residues. However, for the partially buried side chains (35% of residues) the important terms (apart from the excluded volume) are intraresidue energy, surface complementarity, and interaction with the solvent, especially the latter. Correct prediction of partially exposed residues can be seen as the real test of a scoring function for side chain modeling. For buried residues very simple scoring functions yield good results,[18] while for exposed residues the prediction is complicated by experimental conditions such as crystal packing and ions concentrations,[14] and might at times be meaningless due to the inherent flexibility of these residues.[30,51] Table 5 also compare the results obtained using our surface and volume terms with the traditional Lennard–Jones potentials for reflecting van der Walls forces. We constructed this potential to have its minimum at a distance equal to the sum of the van der Waals radii. The exact relation between the attractive and the repulsive terms were optimized, as well as the relative contribution of the potential to the score. From Table 5 it is apparent that the surface complementarity and volume terms in the scoring function [eq. (2)] compensate well for the absence of a distance dependent term for van der Waals forces. We did not evaluate a classical torsion term because we used a discrete search procedure in rotamer space rather than a continuous one in torsion space. Other traditional energy terms were not evaluated because they have no simple analogs in our scoring function.

Concurrent Modeling of Multiple Side Chains

We implemented our scoring function in a program that predicts conformations of several side chains concurrently. Two methods were applied for this task: a simple iterative self-consistent one (Sccomp-I), and a stochastic method (Sccomp-S) based on the Boltzmann distribution (application of the Gibbs sampling algorithm for side-chain modeling).

In the iterative procedure (Fig. 2), residues are individually modeled one after the other until all χ angles of the side chains being modeled are within 40° of their conformations in the previous iteration, or until a predetermined maximum number of iterations is reached (see Methods). The initial order in which the side chains are considered is of importance for the prediction accuracy. Better results were obtained when side chains with many neighbors (usually buried ones) were modeled first, i.e., in decreasing order of S scores [eq. (9)]. Figure 5 illustrates the correlation between prediction accuracy and order in the initial iteration. The algorithm is reasonably fast, with an average running time of 50 s for a protein of 200 residues.

In the second procedure (Fig. 3) also, every residue is modeled while the rest of the protein is held fixed. However, the rotamer chosen at any given point is according to probabilities derived from the Boltzmann distribution. The order of the modeling is stochastic, such that a modeled residue is randomly chosen from the neighbors of the previous modeled residue. It was expected that buried residues, with higher density of neighbors in the protein

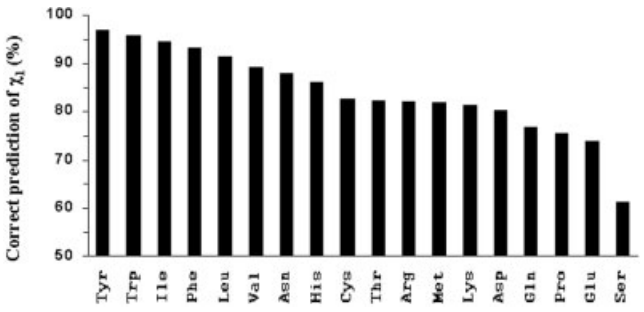

Figure 7. Accuracy of amino acid prediction. The percent accuracy of χ_1 prediction for the different amino acids is shown. All side chains were modeled concurrently.

Table 7. Comparison with Other Side-Chain Modeling Programs.[a]

Method	Execution time	χ_1 (%)		χ_{1+2} (%)		RMSD (Å)	
		Bur	All	Bur	All	Bur	All
Scwrl[18]	8 s	*92.0*	*81.8*	*82.9*	*69.8*	*1.10*	*1.86*
Sccomp–iterative[(This work)]	50 s	*93.2*	*82.2*	*88.1*	*72.6*	*0.97*	*1.67*
Sccomp–stochastic[(This work)]	12 min	*93.6*	*83.7*	*88.3*	*73.7*	*0.85*	*1.60*
Scap[12]	12 min	*94.9*	*81.0*	*89.5*	*71.3*	*0.86*	*1.66*
Smol[17]	45 min	*95.6*	*87.7*	*89.7*	*77.6*	*0.73*	*1.52*

[a]All programs were evaluated locally and identically, using Set2, default parameters, and a PentiumIII, 850 MHz, 512 Mb computer. C_β was not included in the RMSD calculation. Abbreviations: bur = buried side chains, all = all side chains.

matrix, would be modeled more often, and this is indeed what was observed (Fig. 6). The conformation of a buried side chain is strongly dependent on the conformations of its contacting side chains, and, therefore, a more frequent sampling of these during the search is desired. The average running time of this algorithm for a protein of 200 residues is about 12 min.

Performance

The two protocols for side chain modeling were analyzed first with our own test set. The results are shown in Table 6, Set 3. The differences in prediction accuracy between the iterative and stochastic protocols are 1–2% for χ_1 and χ_{1+2} predictions and about 0.1 Å in RMSD. The predictive accuracy for the various amino acids is shown in Figure 7. Prediction of polar and charged side chains is clearly inferior to the prediction of hydrophobic and aromatic ones. This emphasizes the limitations of the scoring function; specifically, the lack of an electrostatic term as well as a more accurate solvation model.

To further evaluate performance, we checked our programs with an additional test set. The results (Table 6, Set2[17]) are very similar to those obtained with Set3 but with slightly smaller RMSD. We also used Set2 to compare our programs to the other recent methods in the field (Table 7). The program of Liang and Grishim[17] (Smol) gave the highest percentage accuracy. Their scoring function includes electrostatic and desolvation terms that require explicit positioning of hydrogens and clearly improve the prediction for polar residues. Smol, however, uses a Monte Carlo search procedure that takes about 45 min for the average sized protein of Set2, the longest time of all the procedures investigated.

The program Scap[12] employs terms from the CHARMM force field and an iterative procedure for the search. This program performs well for buried residues but less so for χ_1 of all side chains. The running time was about 12 min, similar to Sccomp-S. Scap is more advantageous for buried side chains, in part due to its larger rotamer library. However, for partially buried or exposed residues the weight of the scoring function is dominant (Table 5) and the completeness of the library less important (see Fig. 8). Thus, for these residues Sccomp-S is more advantageous.

Scwrl,[18] which employs a simple scoring function using a backbone-dependent rotamer library, is probably the most popular program in the field, due to its simplicity, speed, and availability on the Web. Scwrl often serves as a reference for evaluating side chain modeling programs.[12,16,17,52] The running time on an average size protein is only few seconds. In our test for Scwrl, we obtained a predictive accuracy of 92% for χ_1 of buried residues and 82% for χ_1 of all residues. Thus, Sccomp-I is slightly more accurate than Scwrl for buried (as well as all) residues. This is especially apparent in the mean RMSD for all residues, which is about 0.2 Å lower for Sccomp-I than for Scwrl.

To investigate the effect of crystal packing on accuracy of prediction, the full crystal environment for the proteins of Set3 was built using the program pdbset from the CCP4 suite.[53] Side-chain conformations of the central protein within the fixed framework of the other copies with their native structures were then constructed using Sccomp-I in this environment. The predictive accuracy for χ_1 of all residues improved by 1.8%, for χ_{1+2} by 1.9%, and for RMSD from 1.79 to 1.67 Å. A similar level of improvement was also found for the truly exposed side chains (those with an accessibility >0.5 in the crystal environment). However, as the predictive accuracy still remains quite low (73% for χ_1, 57% for χ_{1+2}

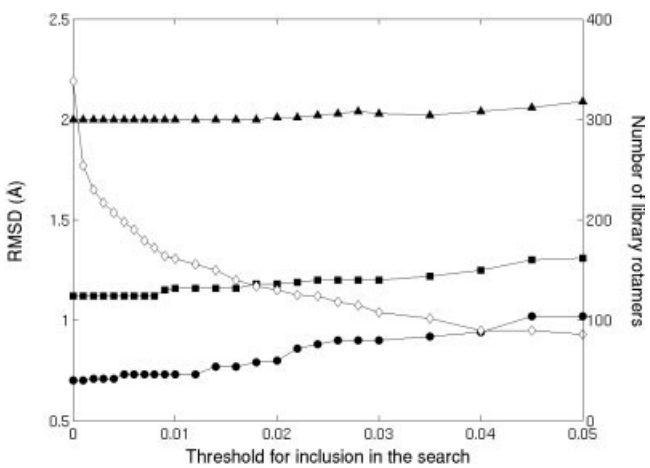

Figure 8. Dependency of prediction accuracy on rotamer library size. The prediction accuracy for rotamers of individual residues in Set1 is shown. The scale of the *x*-axis is the threshold probability for inclusion of a rotamer in the search. It is taken from the rotamer probability appearing in the backbone-independent rotamer library.[32] ●, buried residues; ■, partially buried residues; ▲, exposed residues; ◇, number of rotamers remaining as a function of threshold probability.

Table 8. Influence of the Exact C_β Position on the Prediction.[a]

	C_β from native						C_β built					
	χ_1 (%)		χ_{1+2} (%)		RMSD (Å)		χ_1 (%)		χ_{1+2} (%)		RMSD (Å)	
Modeling program	Bur	All	Bur	All	Bur	All	Bur	All	Bur	All	Bur	All
Scap[12]	95.5	82.4	90.4	73.6	0.78	1.58	94.9	81.0	89.5	71.3	0.86	1.66
Sccomp-S	95.6	85.0	91.0	75.4	0.74	1.55	93.6	83.7	88.3	73.7	0.85	1.60

[a]Abbreviations: bur = buried side chains, all = all side chains. Sccomp-S is from this work.

and 2.2 Å for RMSD), crystal packing is clearly not the major obstacle to accurate prediction of truly exposed side chains.

Some Observations Regarding Side-Chain Modeling

It was recently shown that the use of exact native distances and angles improves modeling.[12] We specifically noticed that although the position of the C_β atom is essentially determined by the backbone, the small deviations between the standard parameters and the native parameters are often sufficient to cause a non-negligible shift in the entire side-chain position. This is demonstrated in Table 8 for two different modeling methods, and emphasizes again the sensitivity of side-chain placement on precise backbone structure. One obvious consequence is that whenever the exact coordinates of C_β atoms are known they should be used and should not be rebuilt.

A threshold of 0.003 for the probability of a rotamer to be included in the search was used throughout this study. At this level, about 60% of the rotamers of the full library are retained. Figure 8 describes the general influence of library size on overall modeling accuracy. For the 0.003 threshold, there is hardly any loss of accuracy. In general, accuracy of prediction (RMSD) declined slowly with size reduction. The decline was the steepest for buried residues and the shallowest for exposed ones. Thus, buried residues adopt uncommon conformations more often, probably because they are more frequently subjected to interactions with other residues.

Conclusions

We presented here a side-chain modeling method that focuses on the scoring function rather than the searching procedure. The combination of weighted contact surface areas and solvent-accessible surface account for van der Waals forces and solvation free energy. The scoring function contributes especially to the modeling of side chains that are not totally buried. For buried residues, simple scoring functions can do well.[9,18] However, buried residues are more sensitive to the completeness of the rotamer library used for the search. Regarding atomic solvation parameters, large, positive values yield the most accurate predictions in terms of RMSD. The differential between the atomic solvation parameters and the contact surface parameters between all atom types should be positive. This differential might reflect the driving force for maximizing packing of the protein.

Software Availability

The Sccomp-S and Sccomp-I programs are available for Unix/Linux and Mac OX platforms. The programs can be accessed through the Web at (http://sgedg.weizmann.ac.il/sccomp.html). They can be used to model all, or a defined set of, the side chains in a protein as well as any number of user-generated changes (mutations). Sccomp can also employ a template PDB to place side-chain conformations at conserved positions and model the remaining ones.

References

1. Levitt, M.; Gerstein, M.; Huang, E.; Subbiah, S.; Tsai, J. Annu Rev Biochem 1997, 66, 549.
2. Desmet, J.; De Maeyer, M.; Hazes, B.; Lasters, I. Nature 1992, 356, 539.
3. Lasters, I.; Desmet, J. Protein Eng 1993, 6, 717.
4. Goldstein, R. F. Biophys J 1994, 66, 1335.
5. Mendes, J.; Soares, C. M.; Carrondo, M. A. Biopolymers 1999, 50, 111.
6. Koehl, P.; Delarue, M. J Mol Biol 1994, 239, 249.
7. Vasquez, M. Biopolymers 1995, 36, 53.
8. Lee, C.; Subbiah, S. J Mol Biol 1991, 217, 373.
9. Holm, L.; Sander, C. Proteins 1992, 14, 213.
10. Kussell, E.; Shimada, J.; Shakhnovich, E. I. J Mol Biol 2001, 311, 183.
11. Shenkin, P. S.; Farid, H.; Fetrow, J. S. Proteins 1996, 26, 323.
12. Xiang, Z.; Honig, B. J Mol Biol 2001, 311, 421.
13. Mendes, J.; Baptista, A. M.; Carrondo, M. A.; Soares, C. M. J Comput Aid Mol Des 2001, 15, 721.
14. Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B. J Mol Biol 2002, 320, 597.
15. Glick, M.; Rayan, A.; Goldblum, A. Proc Natl Acad Sci USA 2002, 99, 703.
16. Mendes, J.; Nagarajaram, H. A.; Soares, C. M.; Blundell, T. L.; Carrondo, M. A. Biopolymers 2001, 59, 72.
17. Liang, S.; Grishin, N. V. Protein Sci 2002, 11, 322.
18. Bower, M. J.; Cohen, F. E.; Dunbrack, R. L. J Mol Biol 1997, 267, 1268.
19. Sobolev, V.; Wade, R. C.; Vriend, G.; Edelman, M. Proteins 1996, 25, 120.
20. Sobolev, V.; Edelman, M. Proteins 1995, 21, 214.
21. Sobolev, V.; Niztaev, A.; Pick, U.; Avni, A.; Edelman, M. Curr Sci 2002, 83, 857.
22. McConkey, B. J.; Sobolev, V.; Edelman, M. Proc Natl Acad Sci USA 2003, 100, 3215.

23. Ooi, T.; Oobatake, M.; Nemethy, G.; Scheraga, H. A. Proc Natl Acad Sci USA 1987, 84, 3086.
24. Eisenberg, D.; McLachlan, A. D. Nature 1986, 319, 199.
25. Dunbrack, R. L. Curr Opin Struct Biol 2002, 12, 431.
26. Tuffery, P.; Etchebest, C.; Hazout, S. Protein Eng 1997, 10, 361.
27. Huang, E. S.; Koehl, P.; Levitt, M.; Pappu, R. V.; Ponder, J. W. Proteins 1998, 33, 204.
28. Samudrala, R.; Huang, E. S.; Koehl, P.; Levitt, M. Protein Eng 2000, 13, 453.
29. Najmanovitch, R.; Kuttner, J.; Sobolev, V.; Edelman, M. Proteins 2000, 39, 261.
30. Eyal, E.; Najmanovich, R.; Edelman, M.; Sobolev, V. Proteins 2003, 50, 272.
31. Wang, G.; Dunbrack, R. L. Bioinformatics 2002, 19, 1589.
32. Dunbrack, R. L.; Cohen, F. E. Protein Sci 1997, 6, 1661.
33. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. J Comput Chem 1983, 4, 187.
34. McConkey, B. J.; Sobolev, V.; Edelman, M. Bioinformatics 2002, 18, 1365.
35. Lee, B.; Richards, F. M. J Mol Biol 1971, 55, 379.
36. De Maeyer, M.; Desmet, J.; Lasters, I. Methods Mol Biol 2000, 143, 265.
37. Dunbrack, R. L.; Karplus, M. Nat Struct Biol 1994, 1, 334.
38. Gelin, B. R.; Karplus, M. Biochemistry 1979, 18, 1256.
39. Petrella, R. J.; Lazaridis, T.; Karplus, M. Fold Des 1998, 3, 353.
40. Wilson, C.; Gregoret, L. M.; Agard, D. A. J Mol Biol 1993, 229, 996.
41. Schiffer, C. A.; Caldwell, J. W.; Kollman, P. A.; Stroud, R. M. Mol Simulat 1993, 10, 121.
42. Cregut, D.; Liautard, J. P.; Chiche, L. Protein Eng 1994, 7, 1333.
43. Wesson, L.; Eisenberg, D. Protein Sci 1992, 1, 227.
44. von Freyberg, B.; Richmond, T. J.; Braun, W. J Mol Biol 1993, 233, 275.
45. Horton, N.; Lewis, M. Protein Sci 1992, 1, 169.
46. Cummings, M. D.; Hart, T. N.; Read, R. J. Protein Sci 1995, 4, 2087.
47. Juffer, A. H.; Eisenhaber, F.; Hubbard, S. J.; Walther, D.; Argos, P. Protein Sci 1995, 4, 2499.
48. Zhou, H.; Zhou, Y. Proteins 2002, 49, 483.
49. Chothia, C. Nature 1974, 248, 338.
50. Kramer, B.; Rarey, M.; Lengauer, T. Proteins 1999, 37, 228.
51. Zhao, S.; Goodsell, D. S.; Olson, A. J. Proteins 2001, 43, 271.
52. Liu, Z.; Jiang, L.; Gao, Y.; Liang, S.; Chen, H.; Han, Y.; Lai, L. Proteins 2003, 50, 49.
53. Collaborative Computational Project, Number 4. Acta Crystallogr 1994, D50, 760.