Accurate Reconstruction of All-Atom Protein Representations From Side-Chain-Based Low-Resolution Models

Michael Feig,[1] Piotr Rotkiewicz,[2] Andrzej Kolinski,[2] Jeffrey Skolnick,[2] and Charles L. Brooks III[1]*

[1]*Department of Molecular Biology, The Scripps Research Institute, La Jolla, California*

[2]*Laboratory of Computational Genomics and Bioinformatics,*
Danforth Plant Science Center, St. Louis, Missouri

ABSTRACT **A procedure for the reconstruction of all-atom protein structures from side-chain center-based low-resolution models is introduced and applied to a set of test proteins with high-resolution X-ray structures. The accuracy of the rebuilt all-atom models is measured by root mean square deviations to the corresponding X-ray structures and percentages of correct χ_1 and χ_2 side-chain dihedrals. The benefit of including C_α positions in the low-resolution model is examined, and the effect of lattice-based models on the reconstruction accuracy is discussed. Programs and scripts implementing the reconstruction procedure are made available through the NIH research resource for Multiscale Modeling Tools in Structural Biology (http://mmtsb.scripps.edu). Proteins 2000;41:86–97.**
© 2000 Wiley-Liss, Inc.

Key words: protein backbone; minimization; implicit solvent

INTRODUCTION

Theoretical approaches to the investigation of protein structures and refinement procedures for experimental data involve modeling of proteins at different resolutions. Experimental resolution determines the initial level of model detail, whereas reduced protein models are used to address otherwise computationally challenging tasks such as ab initio protein structure prediction. In both cases, the final aim of experimental or theoretical protein structure determination is a representation in full atomic detail derived from lower resolution models through a reconstruction procedure. Accurate all-atom models can also add more detailed information during the low-resolution modeling process.

So far, most of the attention has been given to reconstruction of all-atom protein models from protein backbones. The backbone chain is most readily available from crystallographic experiments and various models based on backbone centered united residue models have been used to describe proteins at low resolution.[1–4] Protein structure prediction by comparison with homologous proteins also results mainly in a structure for the protein backbone if the sequence identity between the template and the modeled protein is low.

In reconstructing an all-atom protein model from the backbone, the generation of a complete backbone from C_α positions, if the rest of the backbone is unknown, and the placement of side-chains onto a given backbone represent different problems. A variety of methods are available for the calculation of complete backbone geometries from C_α coordinates. Analytical methods have been proposed that find optimal backbone geometries for the valence bond and angle constraints along the peptide linkage with or without using additional empirical potentials[5–7] or by aligning peptide-group dipoles.[8] In another approach the backbone is reconstructed piecewise from known conformations, e.g., from X-ray structures or a library of backbone fragments that are matched against the C_α positions and then mended together to form the complete backbone.[9–14] Recently, new methods have become possible that take advantage of statistical information about preferred conformations from analysis of structures in the Protein Data Bank.[15,16] An algorithm that places backbone atoms at the most likely positions within a local coordinate system spanned by consecutive C_α positions allows much faster reconstructions of peptide chains[15] with similar prediction accuracies as in previous techniques.

Compared with backbone reconstructions, the determination of side-chain orientations for a given backbone presents a much more difficult problem due to the combinatorial nature of searching the large conformational space accessible to most amino acid residues. This problem can be reduced to a manageable size by restricting the search of side-chain conformations to a subset of the complete conformational space. Although different strategies have been explored,[17] most useful has been the observation that side-chains are usually found in one of only a small number rotameric states.[18,19] By using a library of rotameric states[19–21] side-chain conformations can then be sampled efficiently even for larger proteins and optimized according to energetic criteria that avoid steric clashes and may reward favorable packing interactions.[17] Side-chain reconstructions based on such methods are usually fast

Grant sponsor: National Institutes of Health Multiscale Modeling Tools in Structural Biology; Grant number: RR12255.

*Correspondence to: Charles L. Brooks III, Department of Molecular Biology, TPC6, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037. E-mail: brooks@scripps.edu

Received 17 March 2000; Accepted 2 June 2000

and fairly accurate with average RMSD values around 1.5 Å for heavy side-chain atoms.

Although the protein backbone defines secondary structure, the packing interactions of side-chains determine tertiary structure and are most relevant for protein stability and function.[22,23] To reflect the importance of side-chain interactions, recent low-resolution protein models consist of interaction sites located at the side-chain centers with[24,25] or without[26] additional interaction sites at the backbone. The availability of side-chain centers should also improve the accuracy of all-atom reconstructions over reconstructions from backbone-only models by significantly limiting the search space for possible side-chain conformations. However, despite the increased use of side-chain-based low-resolution models, reconstruction procedures that use side-chain centers alone or in conjunction with backbone representations have not been reported to our knowledge.

In this article we present and evaluate such a procedure for the rapid reconstruction of accurate all-atom models from reduced models based on virtual particles at side-chain centers. Emphasis is given to the robustness of the reconstruction procedure with respect to errors in the virtual particle positions. Deviations from the "correct" model particles at the side-chain centers may arise when fitting experimental data or by projection onto lattice grids that are commonly used to increase computational speed in simulations of low-resolution models for protein structure prediction.[2,26] It is also interesting how additional information can improve the quality of reconstructed structures, and we investigate the effect of including C_α positions with the side-chain centers.

In the following we first describe the reconstruction procedure in detail and then present and discuss results from applying the method to side-chain-based low-resolution models of 13 proteins with available high-resolution crystal structures. The quality of the rebuilt structures is compared after different stages of the reconstruction procedure and the influence of lattice grid projections and the addition of C_α positions is discussed.

MATERIALS AND METHODS

Low-Resolution Models

The low-resolution protein model used in this article is based on the SICHO model by Kolinski and Skolnick.[26] It consists of a chain of virtual particles located at the (geometric) side-chain center of each amino acid residue. The side-chain centers are calculated by averaging the positions of all heavy side-chain atoms including the C_α atom. For glycine the virtual particle is placed at the position of the C_α atom. An extended model, called SICHO/C_α, that contains the C_α positions in addition to the side-chain centers also is used to investigate the benefit of an explicit backbone representation for the reconstruction procedure.

For the application in Monte Carlo simulations Kolinski and Skolnick use a projection of the SICHO model onto a cubic lattice grid with a mesh size of 1.45 Å. To understand the influence of model approximations introduced by lattice-based representations we also apply the reconstruction procedure to models where the side-chain centers, and C_α coordinates in the extended model, are projected onto cubic grids with mesh sizes from 0.2 to 2.0 Å.

Reconstruction Procedure

The reconstruction procedure for generating all-atom structures from these side-chain-based low-resolution models consists of three stages as depicted in Figure 1. First, the backbone is rebuilt by constructing a C_α scaffold from the side-chain centers and then adding the remaining C, O, and N atoms based on the C_α positions. In the second stage, side-chain conformations are chosen from a rotamer library according to the configuration of the rebuilt backbone to form a complete all-atom model. The reconstructed structure is refined further by minimization with the CHARMM force field under harmonic restraints for the side-chain centers to ensure that they remain close to the sites in the low-resolution model. If sites at the C_α atoms are included in the low-resolution model, their positions are restrained in a similar way. The different stages are explained in more detail in the following.

Backbone Reconstruction

Most important for the quality of the final all-atom model is an accurate initial representation of C_α atoms. The reconstruction of the complete backbone depends strongly on their positions, whereas the backbone configuration, in turn, determines the subsequent placement of side-chains.

If C_α positions are not available explicitly as part of the low-resolution model, they are rebuilt from the side-chain centers as follows: A local coordinate system (a_x, a_y, a_z) centered at the i-th side-chain center is constructed by

$$a_x = \nu_{i-1,i+1}$$

$$a_z = \nu_{i,i+1} \times \nu_{i-1,i}$$

$$a_y = a_z \times a_x$$

where $v_{i,j}$ denotes the vector between side-chain centers of residues i and j. As illustrated in Figure 2, this results in a coordinate system with the x-axis oriented along the vector between side-chain centers i-1 and i+1 and the z-axis perpendicular to the plane spanned by the side-chain centers at i-1, i, and i+1. The y-axis is chosen such that an orthogonal coordinate system is formed.

Initial estimates for C_α positions from side-chain centers are made by using previously determined average relative C_α positions expressed in the local coordinate system described above. At each residue these average relative C_α positions are applied within the local coordinate system that is defined by the side-chain centers from the low-resolution model and then transformed back into the original frame of reference to provide an estimated C_α position. The use of a side-chain center-based coordinate system in the reconstruction takes side-chain packing at residues i-1, i, and i+1 into account. To include longer range effects that involve residues i-2 to $i+2$, the average C_α position at residue i relative to residue i is combined with average C_α positions at residue i relative to residues i-1 and $i+1$. Weighting factors of 0.5 for the C_α position relative to i and 0.25 for the positions relative to i-1 and $i+1$ has provided the best results and is used here.

Average relative C_α positions were obtained by analyzing ≈200,000 residues in 824 chains of selected nonhomologous, high-resolution PDB structures (Dunbrack, R.L.:http://www.fccc.edu/research/labs/dunbrack/culledpdb.html). For each residue a local coordinate system was calculated from the side-chain centers and used to transform C_α sites accordingly. The relative C_α positions were then accumulated separately for each amino acid type to arrive at the average values used during reconstruction.

Further improvement of the estimated C_α positions is achieved by conjugate gradient minimization under a number of harmonic restraint terms that bias toward regular peptide backbone configurations. The minimization terms are illustrated in Figure 3 and listed in more detail in Table I. First, the C_α positions are restrained weakly to the vicinity of the initial estimate based on PDB statistics. The distance between C_α positions of neighboring residues i and $i+1$ are held at 3.808 Å, the characteristic value for peptide backbones in the usual trans conformation. A special case is the reduction of the distance between C_α positions i and $i+1$ when residue $i+1$ is a cis-proline to ≈3 Å. To allow both cis and trans conformations for proline, the restraint potential is zero for distances between 2.95 Å and 3.808 Å but follows half-sided harmonic potentials outside this interval with zero points at the upper and lower limits. Furthermore, the distance between C_α and the side-chain center is restrained for most residues at a single value, but different distances depending on the rotameric state of the side-chain are possible in arginine, glutamine, glutamic acid, isoleucine, leucine, lysine, methionine, and tryptophan.

As for C_α distances involving cis-proline, the restraint potential is set to zero between the minimum and maximum distances given in Table II to avoid a conformational bias while half-sided harmonic potentials are applied outside the interval limits. Half-sided harmonic potentials are also used to restrain the distance between C_α positions at residues i and $i+2$ to values between 5.4 and 7.3 Å equivalent to virtual angles of 90° to 147° described by the C_α positions at i, $i+1$, and $i+2$. Finally, the statistical distribution of virtual dihedral angles for C_α positions at i, $i+1$, $i+2$, and $i+3$ for protein backbones is taken into account. Simplified potential of mean force (PMF) maps are calculated from the distribution of C_α distances between residues i-1 and $i+1$ (d_1 in Fig. 3) and between i and $i+2$ (d_2 in Fig. 3) with respect to the C_α distance between residues i-1 and $i+2$ (d_0 in Fig. 3). The two maps, shown in Figure 4, are obtained from an analysis of the same set of PDB structures that were used for the averaging of C_α positions. The force constants scaling the individual contributions to the C_α minimization potential were tuned empirically to provide the best overall reconstruction results.

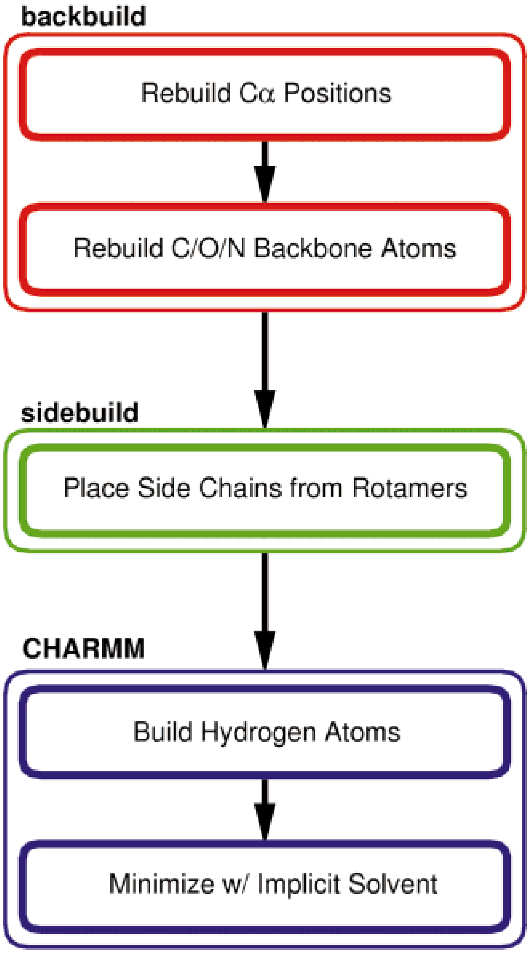
backbuild
Rebuild Cα Positions
Rebuild C/O/N Backbone Atoms
sidebuild
Place Side Chains from Rotamers
CHARMM
Build Hydrogen Atoms
Minimize w/ Implicit Solvent

Fig. 1. Flow diagram of reconstruction procedure.

From the C_α skeleton the rest of the backbone is reconstructed according to the method of Milik et al.[15] Similar to the reconstruction of C_α positions from the side-chain centers, it uses average C, O, and N positions relative to a local coordinate system. This time the coordinate system is derived from the previously determined C_α positions as illustrated in Figure 5. The base vectors (a_x, a_y, a_z) centered at C_α of residue i are defined by using

$$v_1 = r_{i+1}(C_\alpha) - r_i(C_\alpha)$$

$$v_2 = r_{i+2}(C_\alpha) - r_i(C_\alpha)$$

as follows:

$$a_x = \frac{v_1 \times v_2}{|v_1 \times v_2|}$$

$$a_y = \frac{v_2 \times a_x}{|v_2 \times a_x|}$$

$$a_z = a_x \times a_y.$$

Here and in the following $r_i(C_\alpha)$ is used to denote the vector of atomic coordinates for the C_α atom at residue i.

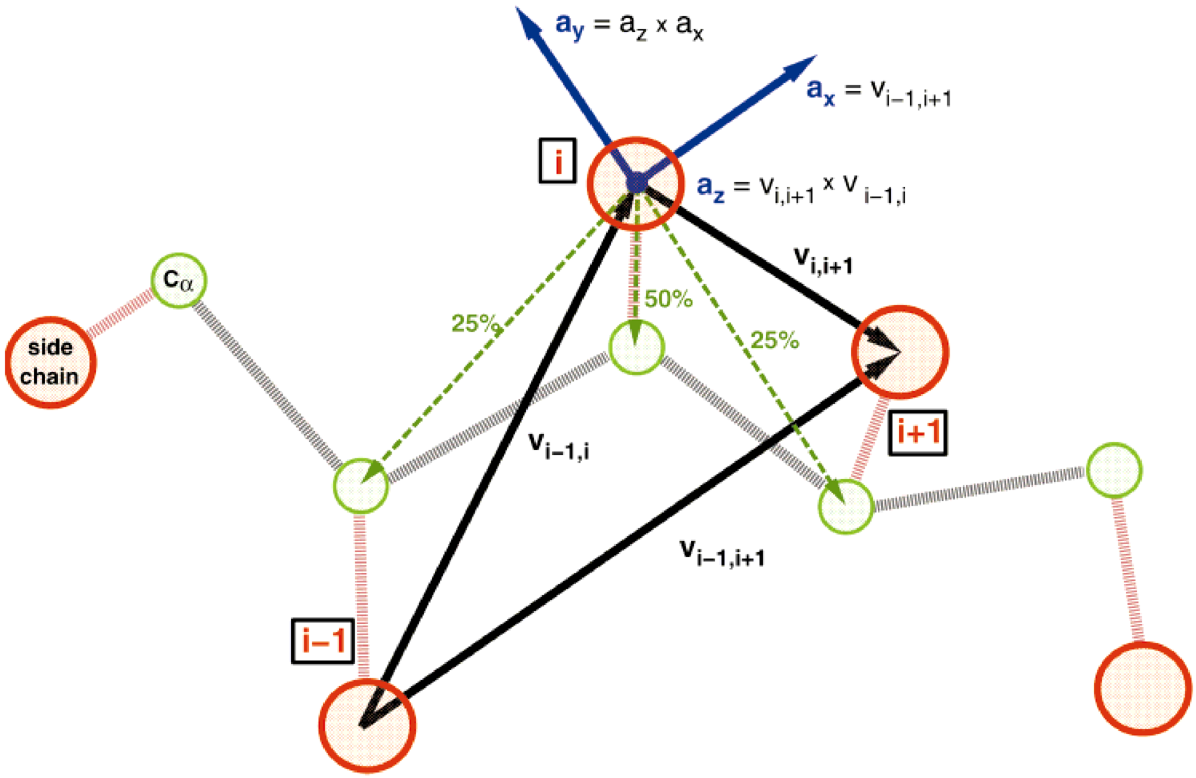

Fig. 2. Initial estimate of C_α position at the *i*-th residue from side-chain centers of mass at *i*-1, *i*, and *i*+1 (see text).

For the non-C_α backbone atoms average relative positions were derived in this local coordinate system by statistical analysis of PDB structures like the average C_α positions relative to the side-chain centers. However, although only one common average C_α position was calculated for each residue type, averages for C, O, and N were determined, depending on the backbone configuration as defined by the C_α coordinates. Following the method by

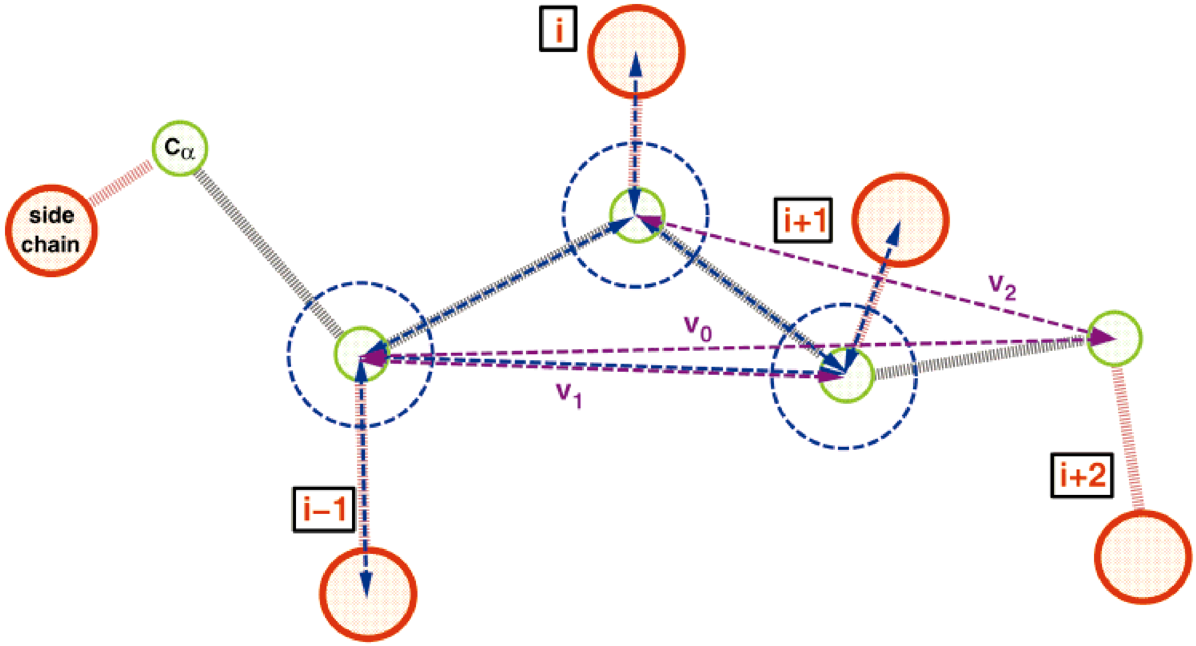

Fig. 3. Potential terms in minimization of estimated C_α positions (see text).

Milik et al. the backbone configuration at residue i is described in a simplified form by the three C_α distances

$$d_0 = |r_{i+2}(C_\alpha) - r_{i-1}(C_\alpha)|$$

$$d_1 = |r_{i+1}(C_\alpha) - r_{i-1}(C_\alpha)|$$

$$d_2 = |r_{i+2}(C_\alpha) - r_i(C_\alpha)|.$$

The chirality of the backbone is taken into account by applying a sign to d_2. Negative values represent left-handed, and positive values represent right-handed conformations as determined from the sign of

$$[(r_i(C_\alpha) - r_{i-1}(C_\alpha)) \times (r_{i+1}(C_\alpha) - r_i(C_\alpha))] \cdot (r_{i+2}(C_\alpha) - r_{i+1}(C_\alpha)).$$

These three distances, together with the residue type, form a four-dimensional grid in which average positions of C, O, and N atoms are accumulated from the PDB structures according to the local backbone configuration measured by d_0, d_1, and d_2 after transformation into the local coordinate system (a_x, a_y, a_z). Values for d_0 and d_1 lie between 4 and 8 Å and for d_2 between 4 and 12 Å. The grid spacing was chosen as 0.25 Å, closer to the typical 0.2 Å atomic resolution of PDB structures than the 0.3 Å used by Milik et al.[15] Sufficient statistics at the higher resolution have become possible because of the rapidly increased number of available PDB structures.

The reconstruction of C, O, and N backbone atoms at a residue i along the peptide chain then involves the following steps. First, the basis vectors a_x, a_y, a_z of the local reference coordinate system and the quantities d_0, d_1, and d_2 are calculated from the C_α coordinates. Average positions for C, O, and N are then taken from the grid described above according to the amino acid type and d_0, d_1, and d_2 and transformed back into the original coordinate system by multiplying with the matrix (a_x a_y a_z).

This procedure provides good initial estimates for the remaining backbone atoms but does not ensure correct bond distances within the peptide backbone. This is corrected by a second conjugate gradient minimization of the now complete backbone with the same potential for the C_α

TABLE I. Terms in Backbone Minimization Potential Used During Initial Backbone Rebuilding Procedure With Corresponding Force Constants k and Minima r_0 of Harmonic Potentials (in Arbitrary Units)

	K	r_0
C_α estimate	0.11	initial estimate
C_α—side chain	1.0	[min, max][a]
C_α—C_α	1.1	3.808[b] Å
C_α—C_α—C_α	10.0	[5.4 Å, 7.3 Å]
C_α—C_α—C_α—C_α	0.25	PMF map
C—O	5.0	1.233 Å
C_α—N	5.0	1.460 Å
C_α—C	5.0	1.525 Å
N—C_{-1}	5.0	1.330 Å

[a] See Table II.

[b] 2.95 Å for cis-proline.

TABLE II. Minimum and Maximum Distances Between C_α and Side Chain Center of Mass in Å

Residue	Minimum	Maximum
ALA	0.763	0.763
ARG	3.250	4.100
ASN	1.987	1.987
ASP	1.984	1.984
GLN	2.210	2.785
GLU	2.240	2.785
GLY	0.000	0.000
CYS	1.378	1.378
HIS	2.699	2.699
ILE	1.660	1.920
LEU	1.930	2.097
LYS	2.600	3.175
MET	2.120	2.640
PHE	2.984	2.984
PRO	1.405	1.405
SER	1.265	1.265
THR	1.455	1.455
TRP	3.450	3.450
TYR	3.350	3.350
VAL	1.471	1.471

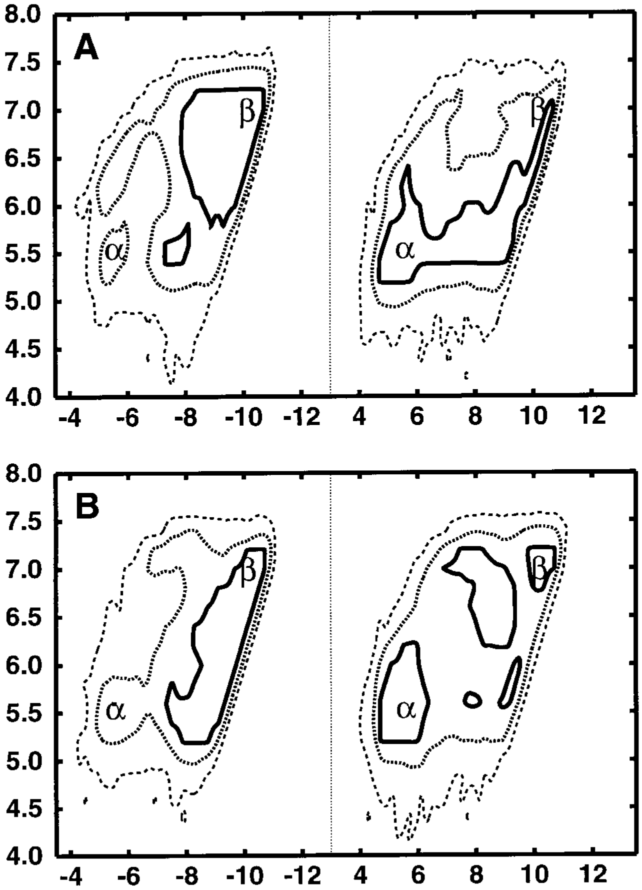

Fig. 4. Potential of mean force maps for distribution of C_α distances d_1 (**A**) and d_2 (**B**) over d_0 (see Fig. 3). All values are given in Å. Negative d_0 values correspond to left-handed, and positive values to right-handed chiralities. Standard α and β configurations are marked for comparison.

positions as in the first minimization and additional terms that restrain C-O, C_α-N, C_α-C, and N-C bond distances to the values given in Table I.

Side-Chain Placement

Once the protein backbone is completely reconstructed, side-chains are placed from a rotamer library. Because ϕ and Ψ angles are available in the reconstructed backbone a backbone-dependent library can be used to provide possible rotameric side-chain conformations at each residue. Here, we are using the latest version (August 1999) of the library by Dunbrack and Karplus.[21] In a typical side-chain rebuilding procedure, e.g., used in homology modeling, the most probable side-chain conformations are chosen first from the rotamer library and then refined to avoid steric clashes and find optimal side-chain packing according to a given intermolecular potential function.[27,28] The latter part presents a challenging problem if no further information on the orientation or position of the side-chain can be taken into account. However, because the side-chain center is available from the low-resolution model, a simplified approach can be used here for the side-chain reconstruction. Instead of starting with the most probable backbone-dependent conformer from the rotamer library for a given side-chain, the conformer with the side-chain center closest to the low-resolution model is selected and then minimized with respect to the distance of the side-chain center to the model position. The refinement of the initial rotamer conformation is performed by searching χ dihedral angles from first to last with all other angles kept fixed within 20° from the initial rotamer value in 2° increments in both directions until a minimum is found. This results in a complete protein model where the side-chains are packed according to the low-resolution model while severe steric clashes are avoided automatically because of fitting of the side-chains to the model side-chain centers.

All-Atom Minimization

After adding hydrogen atoms with a hydrogen-building routine, such as the one found in CHARMM, an all-atom reconstruction from the low-resolution model is available at this point. This structure may be good enough for some purposes because it already provides a good description of the essential features of the protein structure. Further improvements are possible, though, by minimizing intermolecular interactions with a more sophisticated energy function from a standard molecular mechanics force field. However, this requires significant additional computational expense. An all-atom minimization takes minutes or tens of minutes on a typical workstation, depending on the size of the system, compared with seconds for the whole reconstruction procedure described above.

Here, we show results from minimizing rebuilt structures with the CHARMM[29] program using either the PARAM19 united residue force field[29] or the PARAM22 force field[30] with explicit non-polar hydrogen atoms. In addition to minimizations in vacuum we also used a generalized Born implicit aqueous solvent approximation[31,32] in conjunction with the PARAM19 force field. In addition to both force fields, harmonic restraints with a force constant of 50 kcal/mol were applied to keep the side-chain centers (and C_α positions if they are part of the low-resolution model) near the low-resolution model positions. We used a minimization protocol consisting of two steps. First, a 100-step steepest descent minimization is performed to relieve initial energetic stress before the more aggressive adopted basis Newton-Raphson scheme is applied for 1,000 steps or until the energy between subsequent conformations changes by $<10^{-5}$ kcal/mol. For minimizations with a generalized Born solvent approximation the steps above are performed first in a vacuum environment before the generalized Born potential is included in a second run of 1,000 steps of minimization using the adopted basis Newton-Raphson technique.

Test Structures

We tested the reconstruction procedure on 13 structures for which high-resolution data at 1.0 Å or better is available from X-ray crystallography. We do not include lower resolution structures where the quality of reconstructed structures may appear worse than in actuality because of uncertainties in the experimental data.[27] The PDB accession codes of the structures used here and their maximum resolutions and sizes are as follows: 1AB1 (0.89 Å, 46 residues), 1BRF (0.95 Å, 53 residues), 1BYI (0.97 Å, 224 residues), 1CEX (1.00 Å, 197 residues), 1GCI (0.78 Å, 113 residues), 1IXH (0.98 Å, 321 residues), 1NLS (0.94 Å, 237 residues), 1RB9 (0.92 Å, 52 residues), 2ERL (1.00 Å, 40 residues), 2FDN (0.94 Å, 55 residues), 2PVB (0.91 Å, 103 residues), 3LZT (0.92 Å, 129 residues), and 7A3H (0.95 Å, 300 residues).

RESULTS

For all of the test proteins low-resolution side-chain center-based models were generated from the X-ray structures with and without C_α positions. These models were subsequently rebuilt to an all-atom structure by using the reconstruction procedure described above. The reconstructed structures were then compared with the X-ray structures by determining root mean square deviations (RMSD) of the atomic positions and the percentage of first and second side-chain dihedral angles within 40° of the value in the X-ray structure. Because the first and last residues are prone to large fluctuations, they are not included in the calculation of RMSD values and correct percentage of χ dihedrals. We also did not include residues for which alternate conformations are indicated in the X-ray structures. Average values over all test proteins for the reconstruction from SICHO and SICHO/C_α models are shown in Table III. These averages are calculated by accumulating overall residues from the set of test proteins instead of accumulating separate averages for each protein to avoid size effects. Results are shown for both models after the reconstruction procedure without further minimization and after an all-atom minimization with CHARMM using the PARAM19 force field and a generalized Born implicit solvent approximation. The third col-

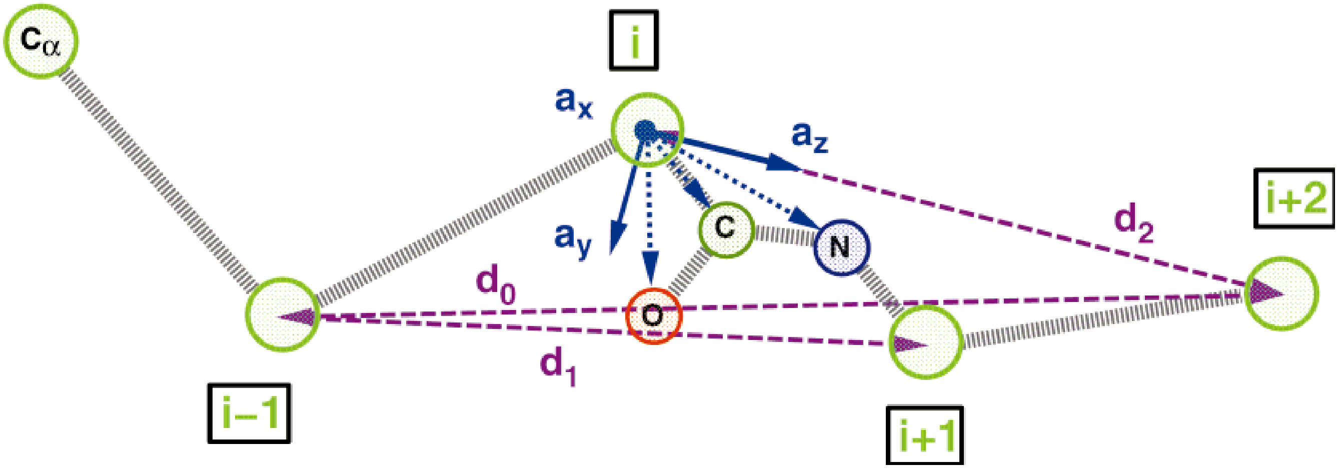

Fig. 5. Initial estimate of backbone atoms C/O/N at the *i*-th residue from C_α positions according to Milik et al.[15]

umn shows how much the crystallographic protein structures deviate on average after minimization with the same protocol that is used after the reconstruction, including the same restraints. Average total RMSDs of 0.2–0.3 Å were found after minimization of the X-ray structures that serve as lower bounds for the possible accuracy of reconstructed structures with the method described here.

On average, side-chain center-based models are reconstructed to an all-atom representation within ≈1 Å from the corresponding X-ray reference. Minimization with CHARMM, while restraining the side-chain centers to the model positions, reduces the total RMS deviation to 0.75 Å. As one might expect, C_α coordinates are represented most accurately, whereas the largest deviations occur for side-chain atoms. The percentage of correctly predicted side-chain dihedral angles also improves after minimization from 73 to 80% for χ_1 and more significantly from 40 to 57% for both χ_1 and χ_2.

The inclusion of C_α positions in the low-resolution model improves the success of the reconstruction procedure

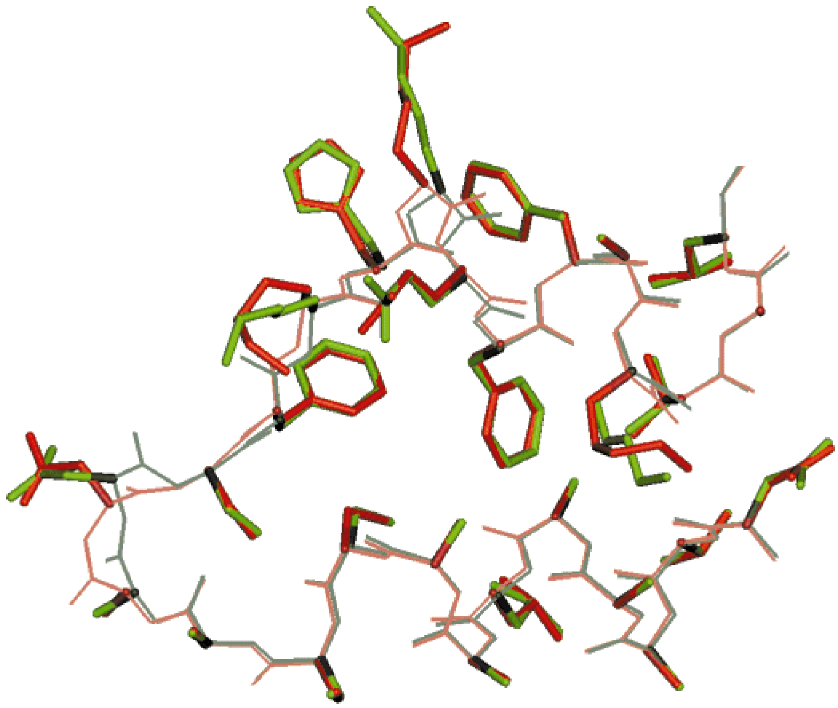

Fig. 6. Detailed view of reconstructed helix-loop-helix region in pike parvalbumin (2PVB). Reconstructed and 0.9-Å crystal structures are shown in red and green, respectively.

TABLE III. Average RMSD Values in Å for X-ray Structures and Percentage of Correct (Within 40°) First (χ_1) and First and Second (χ_2) Side Chain Dihedrals for Rebuilt Structures[†]

	SICHO			SICHO/C_α		
	Rebuilt	Rebuilt/Minimum	Minimum	Rebuilt	Rebuilt/Minimum	Minimum
RMSD						
Total	1.088	0.747	0.258	0.695	0.464	0.219
C_α	0.599	0.308	0.121	0.109	0.104	0.028
Backbone	0.778	0.562	0.197	0.341	0.275	0.134
Side chains	1.367	0.916	0.316	0.958	0.614	0.289
Dihedral						
χ_1	73.3	79.7	99.7	91.8	92.1	99.9
χ_{1+2}	39.5	56.9	90.9	62.5	74.5	91.6

[†] Results are given for side chain only (SICHO) and side chain plus C_α (SICHO/C_α) models after reconstruction procedure without minimization (first column), with minimization (second column), and after minimization of the X-ray structure under the same constraints as during the reconstruction procedure (third column).

significantly. Without CHARMM minimization the average total deviation of reconstructed structures is ≈0.7 Å, which can be reduced to <0.5 Å by minimizing with CHARMM. For the SICHO/C_α model, side-chain centers and C_α positions are restrained to the positions from the low-resolution model. By comparing the RMS deviations of backbone and side-chains, it can be seen that the availability of C_α positions from the low-resolution model improves not just the quality of the reconstructed backbone but the side-chain reconstruction to a similar extent as a consequence of a better backbone representation. This is most apparent in the correct reproduction of >90% of the χ_1 side-chain dihedrals even without CHARMM minimization and 63% and 75% correct χ_1 and χ_2 angles before and after minimization, respectively, if C_α coordinates are included in the low-resolution model.

Table IV provides a more detailed view of the side-chain reconstruction accuracy, depending on the amino acid type. It is not surprising that smaller amino acids are generally rebuilt better than the larger ones. Most problematic are arginine, glutamic, and aspartic acid, histidine, isoleucine, lysine, threonine, and tryptophan because they exhibit the largest degree of conformational freedom. It is interesting that the reconstruction of isoleucine, threonine, and, in particular, tryptophane is improved significantly by a more accurate backbone description when C_α positions are included in the low-resolution model. Another notable aspect is the dramatic improvement of the aromatic residues phenylalanine and tyrosine after CHARMM minimization. In general, the relative accuracy of side-chain reconstructions agree with previous results for backbone-based side-chain predictions.[27]

To illustrate the quality of the reconstructed structures further, Figure 6 shows a typical reconstructed fragment from one of the test structures after CHARMM minimization in comparison with the X-ray structure. For the most part, both structures overlap closely in the backbone and the side-chains with small deviations only in the backbone of the loop region and along the extended lysine residues. The reconstructed structure clearly gives an accurate representation of the backbone conformation and most of the more intricate intermolecular interactions between side-chains found in the X-ray structure that are usually not available in that quality from side-chain predictions based only on the backbone.

Depending on the application, it may be desirable to trade a faster reconstruction procedure for reduced accuracy. Table V shows how much each step in the reconstruction procedure improves the reconstruction from a SICHO model without C_α positions, how much time it requires, and how alternate CHARMM minimization protocols affect the results. C_α coordinates are improved significantly from a relatively crude initial estimate by the subsequent C_α minimization. The more sophisticated initial estimate of C, O, and N backbone atoms, on the other hand, is not improved further in terms of RMSD by the following minimization of all heavy backbone atoms. In fact, the carbonyl group actually becomes slightly worse. Only the C_α positions are moved closer to the reference positions by

TABLE IV. Average RMSD Values in Å for X-ray Structures for Different Amino Acid Types[†]

	SICHO		SICHO/C_α	
	Rebuilt	Rebuilt/Minimum	Rebuilt	Rebuilt/Minimum
ALA	0.642	0.139	0.305	0.126
ARG	1.689	1.147	1.574	0.981
ASN	1.386	1.027	1.184	0.846
ASP	1.088	0.667	0.844	0.536
CYS	0.864	0.543	0.455	0.339
GLN	1.465	1.153	1.067	0.883
GLU	1.236	0.793	0.806	0.556
HIS	1.625	1.173	1.314	1.028
ILE	1.533	1.362	0.784	0.568
LEU	1.044	0.630	0.765	0.528
LYS	1.473	1.104	1.119	0.820
MET	1.240	0.899	0.884	0.616
PHE	1.542	0.379	0.929	0.173
PRO	0.838	0.416	0.393	0.280
SER	0.922	0.643	0.465	0.326
THR	1.438	1.083	0.842	0.741
TRP	1.870	1.357	0.867	0.356
TYR	1.374	0.392	1.127	0.234
VAL	0.979	0.605	0.668	0.413

[†] Results are given for side chain only (SICHO) and side chain plus C_α (SICHO/C_α) models after reconstruction procedure without minimization (first column) and with minimization (second column).

TABLE V. Average RMSD Values in Å for X-ray Structures After Different Stages During the Reconstruction Process for Side Chain Only Model (SICHO)[†]

	Ca	N	C	O	Side chains	Total	CPU time (s)
Ca est.	1.04						0.01
Ca min.	0.62						0.74
C/O/N est.		0.63	0.63	1.08			1.02
Ca/C/O/N min.	0.60	0.63	0.64	1.09			0.37
Side chain reconstr.	0.60	0.63	0.64	1.09	1.36	1.09	0.02
Min. 19, GB, restr.	0.31	0.36	0.41	0.93	0.92	0.75	2056
Min. 19, $e = 1$, restr.	0.35	0.41	0.47	1.06	0.99	0.82	407
Min. 22, $e = 1$, restr.	0.37	0.42	0.47	1.01	1.02	0.83	1164
Min. 19, GB	1.23	1.17	1.20	1.58	1.93	1.63	2085
Min. 19, $e = 1$	1.15	1.08	1.12	1.58	1.89	1.58	403
Min. 22, $e = 1$	1.03	0.99	1.03	1.46	1.73	1.44	1127

[†] Results from different minimization protocols are reported for CHARMM19 (19) and CHARMM22 (22) force fields, with generalized Born solvent approximation (GB) or in vacuum, and with or without restrained side chain centers (restr.). Execution times for 1CEX (197 residues) on SGI R10K (180 Mhz) are shown in the last column. RMSD values for structures following side chain reconstruction and CHARMM minimization are calculated after adjusting to the best fit with the experimental reference structure.

a small amount. For time-sensitive applications with many iterations, the small overall gain in accuracy from the minimization of backbone atoms may not be justified by the time spent, and this step may be skipped.

For the example used in the timing results in Table V, a 197-residue protein, the total time for reconstructing an all-atom structure with 1.1 Å (Table III) from the reference takes slightly more than 2 s. Subsequent CHARMM minimization changes the timescale from seconds to minutes. For the given example, minimization of the reconstructed structure takes between 6.6 and 20 min in vacuum and 35 min with a generalized Born implicit solvent approximation. With this extra computational effort the structure can be improved to between 0.82/0.83 Å total RMSD in vacuum and 0.75 Å with generalized Born as long as side-chain centers are restrained to the positions from the model chain. The more recent PARAM22 force field with explicit aliphatic hydrogens consumes more than twice the time required with the PARAM19 force field because of the significantly larger number of total atoms. The results shown here do not suggest, though, that the extra computational cost of including all hydrogen atoms explicitly improves the quality of reconstructed structures. In fact, the RMS deviations are actually slightly larger with the PARAM22 force field. Using the PARAM19 force field for the kind of structure minimizations described here clearly seems to be the better choice.

CHARMM minimizations without using restraints for the side-chain centers move the rebuilt structures significantly away to between 1.44 Å and 1.63 Å total RMSD from the X-ray structure. This effect is most pronounced for C_α atoms that deviate by 1 Å from the X-ray after CHARMM minimization, as much as after the first estimate from the side-chain centers using only PDB statistics without further optimizations. But the side-chains also move considerably away from their correct positions to 1.7–1.8 Å from the reference structure. This finding suggests that although the reconstructed structures are already quite close to the X-ray structure, they are not in the same local minimum on the energy surface and move further away from the reference structure in a search for the closest local minimum. Hence, the use of side-chain center restraints is essential for further refinement by enforcing the correct side-chain packing from the low-resolution model.

The results presented so far refer to continuous-space, or off-lattice, low-resolution models. Because modeling approaches often use lattice projections of low-resolution models to reduce computational expenses further, we also examined how the approximation introduced by such lattice projections influence the accuracy of all-atom reconstructions. Low-resolution models of the same set of test proteins with and without C_α atoms were projected onto cubic lattices with grid spacings from 0.2 to 2.0 Å. The lattice-based models were then reconstructed in the same way as the continuous-space models above. Average RMS deviations and percentages of correct χ_1 side-chain dihedrals in the rebuilt structures before and after CHARMM minimization are shown in Figure 7 in relation with the grid spacing. The diagram also contains the results from the off-lattice models (at a grid spacing of 0.0 Å). The accuracy of reconstructions from the SICHO model appears to be mostly unaffected up to a grid spacing of 0.6 Å but decrease for larger grid spacings. Only C_α coordinates start to become worse after CHARMM minimization at 0.6 Å. The RMSD of the reconstructed structures is increased by up to ≈30% at a grid spacing of 2.0 Å. At the same time, the benefit of a CHARMM minimization worsens progressively with increasing grid spacing. At 2.0 Å grid spacing CHARMM minimization decreases the RMSD only by a relatively small amount, whereas the percentage of correct χ_1 dihedrals remains essentially the same. If the SICHO/C_α model is projected onto a lattice, the quality of all-atom reconstructions already begins to deteriorate significantly at 0.4 Å at a similar or higher rate than for reconstructions from projected SICHO models. As a consequence, reconstructed structures from SICHO and SICHO/C_α models differ much less at larger grid spacings than in the unprojected case.

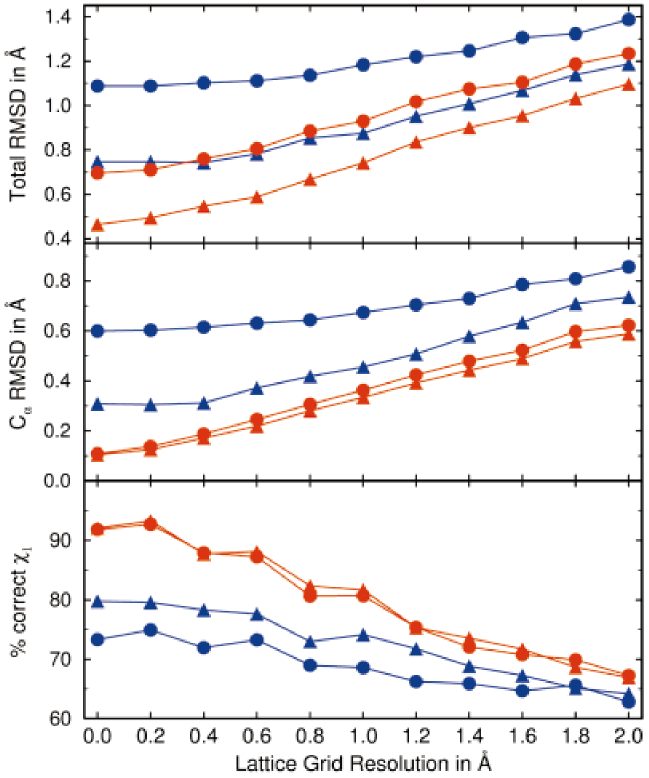

Fig. 7. Quality of reconstructed structures from low-resolution models projected onto lattice grids depending on grid size. RMSD values for all heavy atoms and only C_α and the percentage of correct χ_1 side-chain dihedral angles (within 40°) are compared for SICHO (blue) and SICHO/C_α models (red) with (triangles) and without (circles) CHARMM minimization.

The results for reconstructions from lattice-based low-resolution models are also applicable to the low-resolution modeling of data with a given error. Approximations introduced by a lattice projection can be translated into uncertainties in the side-chain centers or C_α positions if the SICHO/C_α model is used. For a cubic grid with spacing g, the value g/2, the maximum error along the line between two adjacent grid points, can be used as a reasonable estimate of a corresponding error in continuous space positions. The results above then suggest that errors of 0.3 Å in the SICHO model and 0.1 Å in the SICHO/C_α models can be tolerated without significant consequences for the reconstruction of all-atom structures, whereas larger errors begin to affect the accuracy of all-atom reconstructions according to Figure 7.

DISCUSSION

The results presented above can be put into perspective by comparing with other backbone-based reconstruction algorithms.

From C_α coordinates complete backbones can be reconstructed with the fast method by Milik et al.[15] based on PDB statistics at ≈0.4 Å RMSD and with more computational effort by aligning peptide group fragments at ≈0.3 Å RMSD[14] from the reference structure. The reconstruction from side-chain centers is expected to be less accurate. This is reflected in our results with average backbone deviations of 0.77 Å RMSD before and 0.55 Å RMSD after CHARMM minimization. The latter number appears to be quite good, though, considering the lack of an explicit backbone representation in the low-resolution model.

The accuracy of side-chain predictions onto self-backbones can be as good as 1.5 Å RMSD for all heavy side-chain atoms, including C_β, by using a fixed rotamer library[27,33] or principal component analysis.[34] With a flexible rotamer model it is possible to come as close as 1.35 Å RMSD but at considerable computational expense on the order of hours.[33] The same accuracy as with flexible rotamers, 1.36 Å RMSD, is achieved in our reconstruction from side-chain centers already without minimization. CHARMM minimization improves the RMSD to 0.9 Å, well beyond the accuracy obtainable with side-chain predictions that depend only on the protein backbone. As another measure for the accuracy of side-chain conformations the percentage of correct side-chain dihedrals with backbone-based methods is ≈80% for χ_1 and ≈70% for χ_1 and χ_2 in the best cases.[27,33] A similar level of accuracy is reached in the reconstruction from side-chain centers only after CHARMM minimization that improves the correct prediction of χ_1 from 73 to 80% and for χ_1 and χ_2 from 40 to 56%, which is only slightly below the accuracy that can be obtained from backbone-based methods. For the most part, this reflects the dependence of the first side-chain dihedrals on the backbone configuration that is represented only approximately in the reconstruction from side-chain centers.

It should have become obvious at this point that the obtainable accuracy in reconstructed all-atom structures from side-chain centers is limited most severely by the backbone approximation. Apart from improvements in the reconstruction of a backbone from side-chain centers, which may be possible, this can be addressed by including C_α positions in the low-resolution model. By using such an extended low-resolution model, the reconstructed structures improve dramatically. Without CHARMM minimization they are already as good or better than reconstructions from only side-chain centers with CHARMM minimization. The availability of the correct C_α chain allows the accurate prediction of the complete backbone at 0.3 Å RMSD. This in turn improves the backbone-dependent reconstruction of side-chains with almost all (92%) of the χ_1 dihedrals predicted correctly. After CHARMM minimization the reconstructed structures are improved even further. They begin to approach the range of inherent uncertainty expected from the CHARMM force field and after refinement of high-resolution X-ray data. These results suggest that a combined low-resolution representation with model sites at the side-chain center and on the backbone, as for example in the model by Liwo et al.,[24,25] provide an optimal base for accurate all-atom reconstructions.

The situation becomes somewhat different in lattice models. Projecting low-resolution models onto a cubic lattice affects the accuracy of reconstructed structures noticeably. However, even at a grid resolution of 1.4 Å that is used in the lattice-based model by Kolinski et al.,[26] all-atom structures can be reconstructed from side-chain centers reasonably well with a total RMSD of 1.0 Å after CHARMM minimization and 69% correct χ_1 dihedrals. The inclusion of C_α positions in lattice-based models does not improve the structure of rebuilt structures nearly as much as for off-lattice models. With C_α positions, the total RMSD improves only to 0.9 Å and the percentage of correct χ_1 dihedrals becomes 73.5%. This relatively small gain of accuracy may not justify the decreased performance associated with an explicit inclusion of a backbone site in lattice-based models for most applications.

With the availability of an accurate all-atom reconstruction procedure, a description of proteins by very simple models based on virtual particles at the side-chain centers becomes more powerful. This opens the possibility for addressing challenging problems in protein structure modeling, the study of folding and unfolding processes, structure prediction from sequence, large-scale protein dynamics, and refinement of low-resolution structural information from experiment for large macromolecular assemblies like viruses or ribosomal units with multiresolution modeling approaches. Beyond overall structural features and information on side-chain packing available from the low-resolution model, a corresponding all-atom representation can provide further detail about intermolecular interactions because accurate protein structures at atomic level of detail are necessary to ultimately understand protein function.[35] Following earlier ideas of multiresolution modeling,[36] one would expect for such techniques a higher degree of accuracy than for low resolution modeling at moderate computational cost compared to prohibitively expensive complete all-atom modeling that may be reserved for the final stage in a multiresolution modeling approach. Such approaches using side-chain center-based models have already been used successfully for ab initio prediction of protein structures with or without known secondary structures,[26,37] for analysis of protein dynamics,[38] and for improvement of homology-based protein structures predicted by threading algorithms.[39] In all cases, the accurate reconstruction of all-atom models as the final results represents an essential component.

CONCLUSIONS

We have shown that it is possible to rebuild accurate all-atom representations of protein structures from side-chain centers with the reconstruction procedure introduced here. A total RMSD of 1.0 Å is achieved on average in reconstructed all-atom models based only on side-chain centers. With CHARMM minimization the structures are improved to 0.75 Å. By using C_α positions along with the side-chain centers in the low-resolution model, the explicit representation of the backbone all-atom reconstructions can be significantly improved further. A projection of side-chain centers onto lattice grids reduces the accuracy of all-atom reconstructions to an acceptable extent for typical grid spacings in lattice-based protein models. By using this quick reconstruction procedure, multiresolution modeling of protein structures is becoming possible.

REFERENCES

1. Crippen GM, Snow ME. A 1.8 Å resolution potential function for protein folding. Biopolymers 1990;29:1479–1489.
2. Hinds DA, Levitt M. A lattice model for protein structure prediction at low resolution. Proc Natl Acad Sci USA 1992;89:2536–2540.
3. Covell DG. Folding protein α-carbon chains into compact forms by Monte Carlo methods. Proteins 1992;14:409–420.
4. Kolinski A, Skolnick J. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. Proteins 1994;18:338–352.
5. Purisima EO, Scheraga HA. Conversion from a virtual-bond chain to a complete polypeptide backbone chain. Biopolymers 1984;23:1207–1224.
6. Rey A, Skolnick J. Efficient algorithm for the reconstruction of a protein backbone from the α-carbon coordinates. J Comput Chem 1992;13:443–456.
7. Bassolino-Klimas D, Bruccoleri RE. Application of a directed conformational search for generating 3-D coordinates for protein structures from α-carbon coordinates. Proteins 1992;14:465–474.
8. Liwo A, Pincus MR, Rackovsky S, Scheraga HA. Calculation of protein backbone geometry from α-carbon coordinates based on peptide-group dipole alignment. Protein Sci 1993;2:1697–1714.
9. Jones TA, Thirup S. Using known substructures in protein model building and crystallography. EMBO J 1986;5:819–822.
10. Classens M, Van-Cutsem E, Lasters I, Wodak S. Modelling the polypeptide backbone with space parts from known protein structures. Protein Eng 1989;2:335–345.
11. Reid LS, Thornton JM. Rebuilding flavodoxin from Cα coordinates: a test study. Proteins 1989;5:170–182.
12. Correa P. The building of protein structures from α-carbon coordinates. Proteins 1990;7:366–377.
13. Holm L, Sander C. Database algorithm for generating protein backbone and sidechain coordinates from a Cα trace. J Mol Biol 1991;218:183–194.
14. Payne PW. Reconstruction of protein conformations from estimated positions of the Cα coordinates. Protein Sci 1993;2:315–324.
15. Milik M, Kolinski A, Skolnick J. Algorithm for rapid reconstruction of protein backbone from alpha carbon coordinates. J Comput Chem 1997;18:80–85.
16. Wang Y, Huq HI, de la Cruz XF, Lee B. A new procedure for constructing peptides into a given Cα chain. Folding Design 1997;3:1–10.
17. Vasquez M. Modeling side-chain conformation. Curr Opin Struct Biol 1996;6:217–221.
18. Chandrasekaran R, Ramachandran GN. Studies on the conformation of amino acids. XI. Analysis of the observed side group conformations in proteins. Int J Prot Res 1970;2:223–233.
19. Ponder JW, Richards FM. Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. J Mol Biol 1987;193:775–791.
20. Tuffery P, Etchebest C, Hazout S, Lavery R. A new approach to the rapid determination of protein side-chain conformations. J Biomol Struct Dyn 1991;8:1267–1289.
21. Dunbrack RL Jr, Karplus M. Backbone-dependent rotamer library for proteins: application to side-chain prediction. J Mol Biol 1993;230:543–574.
22. Levitt M, Gerstein M, Huang E, Subbiah S, Tsai J. Protein folding: the endgame. Annu Rev Biochem 1997;66:549–579.
23. Bromberg S, Dill KA. Side-chain entropy and packing in proteins. Protein Sci 1994;3:997–1009.
24. Liwo A, Oldziej S, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA. A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. J Comput Chem 1997;18:849–873.
25. Liwo A, Pincus MR, Wawak RJ, Rackovsky S, Oldziej S, Scheraga HA. A united-residue force field for off-lattice protein-structure simulations. II. Parameterization of short-range interactions and determination of weights of energy terms by z-score optimization. J Comput Chem 1997;18:874–887.
26. Kolinski A, Skolnick J. Assembly of protein structure from sparse

experimental data: an efficient Monte Carlo model. Proteins 1998;32:475–494.

27. Bower MJ, Cohen FE, Dunbrack RL. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. J Mol Biol 1997;267:1268–1282.
28. Koehl P, Delarue M. Application of a self-constistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. J Mol Biol 1994;239:249–275.
29. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 1983;4: 187–217.
30. MacKerell AD, Bashford D, Bellott M, Dunbrack JD, Evanseck MJ, Field MJ, Fischer S, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 1998;102:3586–3616.
31. Still WC, Tempczyk A, Hawley RC, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 1990;112:6127–6129.
32. Dominy BN, Brooks CL III. Development of a generalized Born model parametrization for proteins and nucleic acids. J Phys Chem B 1999;103:3765–3773.
33. Mendes J, Baptista AM, Carrondo MA, Soares CM. Improved modeling of side-chains in proteins with rotamer-based methods: a flexible rotamer model. Proteins 1999;37:530–543.
34. Ogata K, Umeyama H. Prediction of protein side-chain conformations by principal component analysis for fixed main-chain atoms. Protein Eng 1997;10:353–359.
35. Wei L, Huang ES, Altman RB. Are predicted structures good enough to preserve functional sites? Structure 1999;7:643–650.
36. Monge A, Lathrop EJP, Gunn JR, Shenkin PS, Friesner RA. Computer modeling of protein folding: conformational and energetic analysis of reduced and detailed protein models. J Mol Biol 1995;247:995–1012.
37. Lee J, Liwo A, Scheraga HA. Energy-based de novo protein folding by conformational space annealing and an off-lattice united-residue force field: application to the 10-55 fragment of staphylococcal protein A and to apo calbindin D9K. Proc Natl Acad Sci USA 1999;96:2055–2030.
38. Kolinski A, Ilkowski B, Skolnick J. Dynamics and thermodynamics of β-hairpin assembly: insight from various simulation techniques. Biophys J 1999;77:2942–2952.
39. Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J. A method for the improvement of threading-based protein models. Proteins 1999;37: 592–610.