ELSEVIER

Biophysical Chemistry 100 (2003) 261–280

Biophysical Chemistry

www.elsevier.com/locate/bpc

Addition of side chains to a known backbone with defined side-chain centroids☆

Rajmund Kaźmierkiewicz[a,b], Adam Liwo[a,b,c], Harold A. Scheraga[a,*]

[a]*Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853-1301, USA*
[b]*Faculty of Chemistry, University of Gdansk, ul. Sobieskiego 18, 80-952 Gdansk, Poland*
[c]*Academic Computer Center in Gdansk TASK, Technical University of Gdansk, ul. Narutowicza 11/12, 80-952 Gdansk, Poland*

Received 4 December 2001; accepted 3 May 2002

Abstract

An automatic procedure is proposed for adding side chains to a protein backbone; it is based on optimization of a simplified energy function for peptide side chains, given its backbone and positions of side-chain centroids. The energy is expressed as a sum of the energies of interaction between side chains, and a harmonic penalty function accounting for the preservation of the positions of the C^{α} atoms and the side-chain centroids. The energy of side-chain interactions is calculated with the soft-sphere ECEPP/3 potential. A Monte Carlo search is carried out to explore all possible side-chain orientations within a fixed backbone and side-chain centroid positions. The initial, usually extended, side-chain conformations are taken directly from the ECEPP/3 database. The procedure was tested on six experimental (X-ray or NMR) structures: immunoglobulin binding protein (PDB code 1IGD, an $\alpha+\beta$-protein); transcription factor PML (PDB code 1BOR, a 49–104 fragment of the ring finger domain, predominantly β-protein); bovine pancreatic trypsin inhibitor (crystal form II) (PDB code 1BPI, an $\alpha+\beta$-protein); the monomer of human deoxyhemoglobin (PDB code 1BZ0, an α-helical structure); chain A of alcohol dehydrogenase from *Drosophila lebanonensis* (PDB code 1A4U); as well as on the 10–55 portion of the B domain of staphylococcal protein A (PDB code 1BDD). In all cases except 1BPI, the data for the algorithm (i.e. the backbone or C^{α} coordinates and the positions of side-chain centroids) were taken from the experimental structures. For protein A, the C^{α} coordinates and positions of side-chain centroids were also taken from the 1.9-Å-resolution model predicted by the UNRES force field. In all comparisons with experimental structures, *complete* side-chain geometry was reconstructed with a root-mean-square (RMS) deviation of approximately 0.6–0.9 Å from the heavy atoms when complete backbone and side-chain-centroid coordinates were used in reconstruction, or approximately 1.0 Å when the C^{α} and centroid coordinates were used.

Keywords: Protein backbone modeling; Side chain addition; United-residue force field

☆ In memory of John T. Edsall.

*Corresponding author. Tel.: +1-607-255-4034; fax: +1-607-254-4700. ACS Postdoctoral Fellow with John T. Edsall, 1946–1947.

E-mail address: has5@cornell.edu (H.A. Scheraga).

0301-4622/03/$ - see front matter
PII: S0301-4622(02)00285-5

1. Introduction

In a hierarchical approach [1–12] to predicting protein structure, the polypeptide chain is initially treated as a united-residue (UNRES) model, and its conformational space is searched with a conformational space annealing method. This approach provides low-resolution protein structures with the backbone expressed in terms of α-carbons, and the side chains treated as ellipsoids with defined centroids rather than at atomic resolution. After the *region* of the global minimum of the UNRES force field is identified, the whole chain is converted to an all-atom model, and the global optimization procedure is continued. We have already developed a method to reconstruct an all-atom backbone from its C^{α}-trace [13,14]; in this paper, we present a fully automatic energy-based procedure to convert the side-chain centroids from the UNRES results to atomic coordinates (with an optimal set of dihedral angles χ) attached to the all-atom backbone.

Many attempts have been made to reconstruct an all-atom polypeptide chain with a focus on adding side chains, either for a known backbone [15–20], or for a model with a known backbone, constructed either from a C^{α}-trace [21,23], the side-chain centroids [24], or by sequence homology [25–30]. The problem in predicting side-chain conformations is a combinatorial one. A systematic search of all possible side-chain orientations for even a small protein is a challenging task for today's computers. Two different approaches have been used to accomplish this computational task. The first strategy was to reduce the dimensions of the problem by incorporating as much empirical information as possible in the computational procedure. For example, in homology modeling, the conformations of the side chains are copied from a template and the full structure is optimized [25,26,31,32]. However, there are some cases in which identical residues in homologous proteins adopt different conformations [32]. Another observation used in adding side chains is that conformations can be grouped into sets having similar values of torsional angles χ. Thus, families of rotamers have been categorized and described. Very complete libraries of naturally occurring families of side-chain conformations have been published [16,18,33,34]. The use of this type of library dramatically decreases the dimensions of the problem. The second strategy was to significantly reduce the computing time by searching for an approximate solution using a simulated annealing protocol [15,22,35,36], a Monte Carlo procedure [15,23,35], a genetic algorithm [16], a clustering approach in which only parts of the proteins are handled systematically [15,22,35,36], a so-called *dead-end elimination* method [17,20,37], or a mean-field theory [19]. Roitberg and Elber [38] used a combination of the locally enhanced sampling (LES) [39–41] and simulated annealing methods in side-chain placement.

The basic idea of the method presented here is to attach a side chain to each C^{α}-atom of the known backbone using a simplified force-field and a Monte Carlo [35] method, with preservation of the positions of the side-chain centroids from the UNRES models. The UNRES representation of polypeptide chains uses united side chains as interacting sites separated from the α-carbons [2,3]. Their orientation with respect to the backbone is a result of structure determination (as is the α-carbon trace). It is therefore justifiable to preserve centroid positions from UNRES simulations when reconstructing a full-atom chain. Most of the existing approaches do not use the positions of the centroids as input data, because they assume that only the backbone coordinates are available (this is the case for the structure produced by homology modeling and for most of the threading approaches). A method that makes use of centroid positions was designed by the Skolnick and Kolińki group [24] to reconstruct all-atom side chains from structures produced using their SICHO model [42–44], in which centroid positions are available. However, this method makes use of rotamer libraries constructed from the PDB database. Because our goal is to predict protein structure without explicit use of information from structural databases, we report a different approach in this paper.

As in our previous work [1,13], we assume that, for a given C^{α}-trace, the problem of reconstruction of an all-atom chain from a united-residue chain, as in the case of UNRES, can be separated into two discrete tasks: (i) reconstruction of the all-

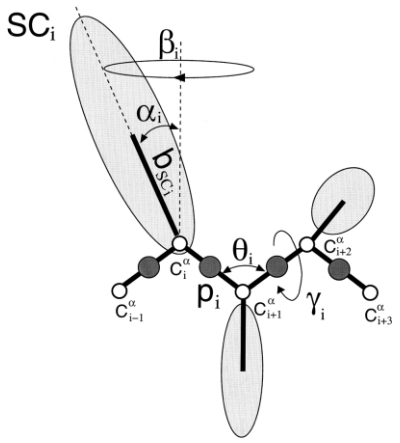

Fig. 1. The UNRES model of polypeptide chains. The interaction sites are side-chain ellipsoids of different sizes (SC) and peptide-bond centers (p) indicated by shaded circles, whereas the α-carbon atoms (*small empty circles*) are introduced to define the backbone-local interaction sites and to assist in defining the geometry. The virtual C^{α}—C^{α} bonds have a fixed length of 3.8 Å, corresponding to a *trans* peptide group; the virtual-bond (θ) and dihedral (γ) angles are variable. Each side chain is attached to the corresponding α-carbon with a different but fixed 'bond length', b_{SC_i}, variable 'bond angle', α_{SC_i}, formed by SC_i and the bisector of the angle defined by C^{α}_{i-1}, C^{α}_{i} and C^{α}_{i+1}, and with a variable 'dihedral angle' β_{SC_i} of counterclockwise rotation about the bisector, starting from the right side of the C^{α}_{i-1}, C^{α}_{i}, C^{α}_{i+1} frame.

atom backbone and (ii) reconstruction of the side-chain geometry for a given backbone. This is justified by the observation that the backbone geometry of well-defined secondary-structure elements (α-helices, β-sheets, β-turns, etc.) does not vary significantly with the type of amino-acid residue.

2. Methods

2.1. The UNRES model of polypeptide chains

The UNRES model of polypeptide chains [1–12] provides information about the geometry of a C^{α}-trace and the positions of side-chain centroids (SC), separately defined for each residue as the averaged position over all side-chain heavy atoms and the C^{α}-atom (see Fig. 1). The distance of a centroid from the α-carbon to which it is attached is fixed and depends on the residue type [3], while the orientation is not fixed and depends on two angles α and β, which are indicated in Fig. 1. A local potential is imposed on these angles, which was determined by fitting the distribution of centroid orientations from the PDB using the maximum-likelihood principle. This potential depends on residue type, and its minima in the (α,β)-space correspond to centroid orientations of well-defined side-chain rotamers [3]. The C^{α}–side-chain distances were determined from the PDB as mean distances between the geometric centers of the heavy atoms of the side chains and the α-carbons to which they are attached [3]; we found that the variation of the C^{α}–SC distance is statistically insignificant for a given residue type [3]. As far as long-range interactions are concerned, the side chains are considered as ellipsoids of revolution, the rotation axis being the C^{α}–SC axis [2]. We use a modified Gay–Berne anisotropic potential to compute the side-chain interaction energy [2]. It should be noted that ellipsoid anisotropies, which are parameters of this potential, are anisotropies of van der Waals-like interactions, and therefore have no direct relation to the components of the moments of inertia of the side chains calculated from the coordinates and masses of the constituent atoms. For a detailed description of the UNRES model and energy function, the reader is referred to the original papers [1–12].

2.2. Formulation of the problem

Having already developed a procedure to compute the all-atom backbone [14] from the UNRES model, we focus here on producing an algorithm that combines this backbone treatment with a method to add side chains to the known backbone. The method is based on two principles: (i) deviation of the added side chains from the positions of the centroids of the parent UNRES model should be as low as possible; and (ii) overlaps between the atoms of side chains, as well as between those of side chains and the backbone are avoided. Because the side chains are ellipsoids of revolution in the UNRES model (i.e. their interactions are averaged over the rotation about the C^{α}–SC axes) and the anisotropies of their interaction potential are fixed, UNRES models provide no additional

information about side-chain geometry, except for the location of their centroids.

The above formulation leads directly to the minimization of a target function comprising the centroid-deviation-penalty and overlap-penalty terms (see Section 2.4). In addition to variations in the χ torsional angles, the side chains are allowed to move with respect to the backbone, with restrictions imposed by an energy penalty function, which is described later in this section. We found that this results in quicker removal of overlaps, compared to allowing only internal side-chain rotations. To be consistent with the UNRES definition of the centroid and to allow freedom in moving the side chain, we introduce a virtual dummy C^{α}-atom, which is attached to the C^{β}-atom of the moving side chain. We calculate the actual position of the side-chain centroid using the defined dummy C^{α}-atom, and treat the real C^{α}-atom only as the point in space to which to attach the side chain. The ECEPP/3 database [45] serves as the source of the starting conformations of the side chains. We adopt the rigid-body ECEPP/3 approximation, and allow only rotations about bonds, as in the ECEPP/3 algorithm. The side-chain models in the ECEPP/3 database are usually in the extended conformation.

2.3. Conservation of the C^{α} configuration

The orientation of the C^{α}–C^{β} bond in the L-configuration can be analytically calculated directly from the geometry of the protein backbone. To accomplish this, we consider two sets of backbone coordinates. The first is directly obtained from the original all-atom backbone derived from UNRES [14]; the second is taken from the ECEPP/3 database. A schematic representation of the protein backbone together with the attached C^{β} atom is presented in Fig. 2.

Although the residue geometry from the ECEPP/3 database includes a full-atom structure of the single residue, at this stage we need information only about the relative orientation of the backbone atoms and the C^{β}-atom. We assume that the dot product of the vectors C^{α}–N and C^{α}–C^{β}, as well as the dot product of the vectors C^{α}–C' and C^{α}–C^{β} calculated from the ECEPP/3 database, should maintain the proper three-dimensional structure of the fragment in Fig. 2 for each added side chain. These dot products, divided by the length of the C^{α}–C^{β} bond (to provide a unit vector parallel to the C^{α}–C^{β} bond), are denoted as c_1 and c_2 in the system of Eq. (1):

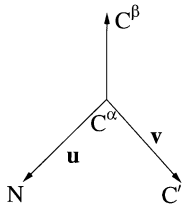

Fig. 2. Schematic representation of the protein backbone. The vectors associated with the C^{α}–N (u) and C^{α}–C (v) bonds are depicted.

$$
\begin{aligned}
&xu_1+yu_2+zu_3=c_1\\
&xv_1+yv_2+zv_3=c_2\\
&x^2+y^2+z^2=1
\end{aligned}
\tag{1}
$$

where x,y,z are the components of the unit vector parallel to the C^{α}–C^{β} bond, and (u_1,u_2,u_3) and (v_1,v_2,v_3) are the components of the vectors **u** and **v**, respectively. The first equation corresponds to the dot product of the vectors C^{α}–N and C^{α}–C^{β}, the second to the dot product of the vectors C^{α}–C' and C^{α}–C^{β}, and the third equation is the square of the norm of the unit vector parallel to the C^{α}–C^{β} bond. From the first two equations of the system Eq. (1) we obtain:

$$z=(c_1-u_2y-u_1x)/u_3 \tag{2}$$

$$y=A+Bx \tag{3}$$

where

$$A=\frac{c_1v_3-c_2u_3}{u_2v_3-v_2u_3} \tag{4}$$

$$B=\frac{v_1u_3-u_1v_3}{u_2v_3-v_2u_3} \tag{5}$$

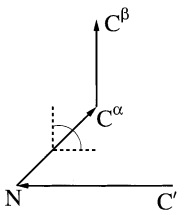

Fig. 3. Illustration of the definition of the torsional angle C–N–C^{α}–C^{β}, which enables us to distinguish between the solution of the system of Eq. (1) corresponding to L- and D-amino acid configurations.

Substituting equations Eqs. (2)–(5) into the third equation of the system Eq. (1), we obtain:

$$\begin{aligned}&(a_1+a_2B+a_4B^2)x^2\\&\quad+(a_2A+a_3+2a_4AB+a_5B)x+a_4A^2+a_5A\\&\quad+a_6\\&\quad=0\end{aligned}\tag{6}$$

where $a_1=1+u_1^2/u_3^2$, $a_2=2u_1u_2/u_3^2$, $a_3=-2c_1u_1/u_3^2$, $a_4=1+u_2^2/u_3^2$, $a_5=-2c_1u_2/u_3^2$ and $a_6=c_1^2/u_3^2-1$. Finally, the quadratic equation Eq. (6) can be solved analytically. There are two solutions of the system of Eq. (1). The first corresponds to the direction of the C^{α}–C^{β} vector in an L-amino acid residue, and the second corresponds to its direction in a D-amino acid residue. To distinguish between the two possible solutions, we examined the torsional angle between the four atoms C′–N–C^{α}–C^{β}, illustrated in Fig. 3.

For all L-amino acid residues in the ECEPP/3 database, this angle is approximately $-120°$; for a D-amino acid residue with an 'ideal' sp^3 C^{α} atom, it would have a value of about $+120°$.

2.4. Target (*pseudo-energy*) *function*

In addition to the correct assignment of the C^{α}-configuration, there are two other problems in adding the side chains properly to the known protein backbone. The first is to avoid the repulsion between closely located atoms. The second is the multitude of possible side-chain orientations. This is the reason why the final structure is not unique. Our goal is not to find the unique solution, but to obtain the lowest possible deviation of the centroids of the added side chains from the given starting set of centroids. The orientation of the side chain is determined by the direction of the C^{α}–C^{β} vector and the set of $\chi_1,\chi_2,\ldots,\chi_n$ torsional angles, where n is the number of bonds about which rotation can take place in the side chain of the single residue. In this work, we use the torsional angle χ numbers, which conform to the IUPAC-IUB convention [46].

The significant force driving the side chains to accommodate to a particular conformation is repulsion. To describe the repulsion between two atoms i and j, we use the soft-sphere potential energy function similar to one from the ECEPP/3 force field [47]:

$$E_{\mathrm{rep};ij}=w_{\mathrm{rep}}\begin{cases}(d_{ij}-d^{\mathrm{o}})^4 & \text{if} \quad d_{ij}<d^{\mathrm{o}}\\ 0 & \text{otherwise}\end{cases}\tag{7}$$

In this work we assumed $w_{\mathrm{rep}}=6.5\times10^3$ and $d^{\mathrm{o}}=3.5$ Å. We found that it is sufficient to choose one arbitrary value of $d^{\mathrm{o}}=3.5$ Å for all interacting atom pairs (using a smaller value of 3.0 Å or altering it depending on the atom type did not change the results). As interacting atoms, we consider only the non-hydrogen atoms separated by at least three bonds. For rotations, we use the bonds defined in the ECEPP/3 database. Because hydrogens are not considered, we ignore the rotations of the methyl, hydroxyl and amino groups. We also omit the rotations inside the guanidino group of arginine residues. Since we do not use torsional potentials, it is more realistic to fix the internal geometry of the guanidino group. As mentioned in Section 2.2, in our model the side chains are allowed to move with some restrictions. The most important is that the distance between the dummy C^{α}-atom and the real C^{α}-atom should be as small as possible, and should be zero at the end of the procedure. This can be expressed in mathematical form as the penalty function defined by:

$$E_{\mathrm{penalty};C^{\alpha};k}=w_{\mathrm{penalty}}\times d^2_{C^{\alpha};\mathrm{dummy}\ C^{\alpha};k}\tag{8}$$

The subscript k denotes the number of the residue and, at the same time, the number of the centroid; $d_{C^{\alpha};\mathrm{dummy}\ C^{\alpha};k}$ is the distance between the

dummy C^{α}-atom and the real C^{α}-atom. Except for the restrictions mentioned above, the side chain has the same number of degrees of freedom as in the ECEPP/3 force field. In order to preserve the positions of the side-chain centroids of UNRES, we added the penalty function defined by:

$$E_{\text{penalty;centroid};k} = w_{\text{penalty}} \times d^2_{\text{centroid;UNRES};k} \tag{9}$$

The value of $d_{\text{centroid;UNRES};k}$ in Eq. (9) denotes the distance between the actual position of the centroid and the position of the centroid defined by UNRES. We have chosen the same arbitrary value of $w_{\text{penalty}} = 2.0$ for both penalty functions, as determined by numerical experiments in which we tried various values of w_{penalty} to determine the optimal one that leads to fastest convergence. The actual position of a side-chain centroid changes whenever any part of the all-atom side chain is rotated or the side chain is moved. Since the repulsion energy $E_{\text{rep};ij}$ has a non-zero value only for distances smaller than d°, we used an arbitrary 'centroid' based cut-off value of 7.0 Å for all interactions. This means that we consider two residues as non-interacting if their UNRES centroids are outside the cut-off range. We found the cut-off of 7.0 Å to be sufficient. Clearly, there might be cases of prolate side chains (e.g. two lysine side chains) contacting head-to-head, where use of this cut-off value between centroids would result in apparently overlapping atoms. However, such prolate side chains are hydrophilic and their head-to-head contacts are rare, because they usually occur on the surface of the globule with side-to-side contacts. At least in the examples considered, we did not find the 7.0-Å cut-off to be too low. The matrix of distances between the centroids of the side chains is calculated only once at the beginning of the program, because the centroids are restricted to the positions in the parent structure. As a consequence of this approximation, the execution time of the program is reduced. Finally, the total energy of interaction is expressed by:

$$E = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E_{\text{rep};ij} + \sum_{k=1}^{\text{nres}} E_{\text{penalty;C}^{\alpha};k} + \sum_{k=1}^{\text{nres}} E_{\text{penalty;centroid};k} \tag{10}$$

In Eq. (10), the first summation extends over all interacting atom pairs (including the atom pairs within the same side chain), and the second and the third sums are defined over the number of residues.

2.5. Rotations

We use quaternions for representation of rotations about arbitrary axes in three-dimensional space. They are an ordered set of four numbers: one real and three imaginary components. Quaternion operators can rotate vectors about a given axis [48]. The rotation matrix **R** is defined in terms of the quaternion parameters by the following expression:

$$\mathbf{R} = \begin{pmatrix} q_0^2+q_1^2-q_2^2-q_3^2 & 2(q_1q_2+q_0q_3) & 2(q_1q_3-q_0q_2) \\ 2(q_1q_2-q_0q_3) & {}^{2}_{0}+q_2^2-q_1^2-q_3^2 & 2(q_2q_3+q_0q_1) \\ 2(q_1q_3+q_0q_2) & 2(q_2q_3-q_0q_1) & q_0^2+q_3^2-q_1^2-q_2^2 \end{pmatrix} \tag{11}$$

where q_0, q_1, q_2 and q_3 are the components of a unit quaternion:

$$q_0 = \cos(\theta/2)$$
$$q_i = p_i \sin(\theta/2) \qquad \text{for} \quad i = 1,2,3 \tag{12}$$

The p_i values for $i = 1,2,3$ are the components of the unit vector parallel to the bond chosen for the rotation, and θ is the rotation angle. The quaternion components q_0, q_1, q_2 and q_3 fulfill the condition:

$$\sum_{i=0}^{3} q_i^2 = 1 \tag{13}$$

The use of quaternions simplified the description of the rotation and assured relatively fast calculations.

2.6. Procedure

We use Monte Carlo optimization to minimize the target function given by Eq. (10). The procedure is equivalent to the Metropolis method [35] with zero temperature, i.e. we accept a new conformation only if the target function is lowered. We split the optimization into two subtasks: (i) generating a conformation compatible with centroid positions in the parent structure, which is equivalent to considering only the geometric penalties [i.e. E_{rep} is omitted from Eq. (10)]; and (ii) full minimization of both the geometric and overlap penalties. In task (i), each side chain is independently optimized, because changing coordinates of a given side chain affects only its geometry, leaving the geometry of the other side chains unchanged.

The procedures to carry out tasks (i) and (ii) are summarized below.

2.6.1. Generating a configuration compatible with centroid positions in the parent structure

1. Read the coordinates of the backbone, and add side chains in the extended conformation from the ECEPP/3 database. Compute the ECEPP/3 side-chain centroid from the extended conformation. Define the translation vector **t** from the ECEPP/3 centroid to the UNRES centroid. Move the side chain to the position of the UNRES centroid by adding the coordinates of the translation vector **t** to the coordinates of each atom of the side chain.
2. Orient all side chains in the L-configuration using the procedure described in Section 2.3.
3. For each consecutive side chain perform steps 4–8.
4. For a given side chain k, calculate the initial geometric penalty function (pseudo-energy), E_o, (i.e. the sum $E_{penalty;C^{\alpha};k} + \sum_{k=1}^{nres} E_{penalty;centroid;k}$) corresponding to this side chain.
5. Perturb either a selected torsional angle χ_i by adding a random increment $\Delta\chi$, where $\Delta\chi_{max} = \pi/3$, or move the side chain by adding the coordinates of a random translation vector **s** to the coordinates of each atom of the selected side chain. Either choice is randomly chosen.
6. Calculate the pseudo-energy (E_1) of the perturbed configuration.
7. If $E_1 < E_o$, accept the new configuration and replace E_o with E_1, otherwise reject the new configuration.
8. Iterate steps 5–7 until the energy no longer decreases or a pre-defined maximum number of steps (usually 10 000) has been tried.

This procedure converges very quickly (in 1000–3000 MC steps per side chain) and takes a few seconds, even for large proteins. It usually produces zero differences from target centroids; exceptions are residues with valence geometry significantly different from the ideal ECEPP geometry.

2.6.2. Minimization of overlaps and geometric penalties

1. Calculate the total energy of the structure generated by the procedure for generating a conformation compatible with the centroid positions in the parent structure (see above).
2. Calculate the energy difference between the new and the old structure using Eq. (10).
3. Choose the side chain with the highest energy, and select one of its torsional angles χ_i at random.
4. Define a random translation vector **s** of length 0.05 Å. Either change the selected torsional angle χ_i by adding a random increment $\Delta\chi$, where $\Delta\chi_{max} = \pi/3$, or move the side chain by adding the coordinates of vector **s** to the coordinates of each atom of the selected side chain. Either choice is randomly chosen.
5. Calculate the difference between the energies of the unperturbed and perturbed configurations, ΔE.
6. Accept the new value of χ_i or the new coordinates of the moved side chain if $\Delta E < 0$. Otherwise restore the old values of the side-chain coordinates.
7. Iterate steps 2–6 until the energy no longer decreases, or a pre-defined maximum number of steps has been tried. The maximum number of steps depends on the protein size and equals

30 000 for proteins with size less than 100 amino acid residues, and 100 000 for larger proteins.

3. Results and discussion

As tests of this procedure, side-chain positions were predicted, based on the correct backbone atoms, for which coordinates were taken from the PDB files: 1IGD [49], 1BPI [50], 1BOR [51], 1BZ0 [52], 1A4U [53]. We also calculated the orientations of the side chains in the model of protein A for which the C^{α}-trace and set of centroids were predicted by the UNRES force field. We applied the method to this set of six proteins ranging in size from 46 to 254 residues. These proteins were chosen for their different numbers of residues; three of them (1IGD, 1BPI, 1BOR) have similar size, but adopt diverse folds.

In the following subsections, we present six test cases: the 61-residue immunoglobulin binding protein (an $\alpha+\beta$-protein; PDB code 1IGD [49]); the 56-residue transcription factor PML, (a 49–104 fragment of the acute promyelocytic leukemia proto-oncoprotein PML, which binds two Zn ions; a predominantly β-protein; PDB code 1BOR [51]); the 58-residue bovine pancreatic trypsin inhibitor (crystal form II) (a protein which has almost the same amount of α-helix and β-sheet, also with some loop regions; PDB code 1BPI [50]); the 141-residue monomer (chain A) taken from the structure of human deoxyhemoglobin (with an α-helical structure; PDB code 1BZ0 [52]); and the 254-residue alcohol dehydrogenase from *Drosophilia lebanonensis* (chain A) (an α/β-protein; PDB code 1A4U [53]). We also applied the method to the model of protein A for which the C^{α}-trace was predicted by the UNRES force field. Protein A is a three-α-helix bundle for which the NMR structure was determined by Gouda et al. [54].

We used the experimental (X-ray or NMR) structures of the backbones and side chains of five proteins. The positions of the side-chain centroids were directly calculated by averaging the positions of all side-chain heavy atoms and the position of the C^{α}-atom in the X-ray structures. For each protein, we performed two runs of the reconstruction procedure. In the first run, complete backbone coordinates from the experimental structure were used, while in the second run, C^{α} coordinates were used and, consequently, the backbone was reconstructed using our dipole-path method [13,14]. The backbone of the UNRES model of protein A was reconstructed from the C^{α}-trace by our dipole-path method [13,14]. To add side chains to the model of the backbone of protein A, we used the positions of the centroids provided by UNRES. For comparison, the backbone and side-chain centroids from the NMR structure were also used.

3.1. Immunoglobin-binding protein (an α/β protein)

The structure of 1IGD contains a four-stranded β-sheet (formed from the N- and the C-terminal parts of the protein) packed against an α-helix formed by residues Ala^{28}–Asp^{41}. The superposition of the side chains from the crystal structure and the side chains resulting from the reconstruction are presented in Fig. 4.

The RMS deviation over all side-chains heavy atoms is 0.62 Å. Such RMSD values are typically obtained when fitting crystal structures to rigid ECEPP/3 geometry [13]. The RMSD values of the heavy atoms of reconstructed side chains from the crystal structure of 1IGD are presented in Fig. 5. Most of them, except for that of Lys^{24} are of the order of 1.0 Å or lower. Larger deviation for Lys^{24} is understandable, because the corresponding side chain is large, and fixing its side-chain centroid does not provide sufficient constraints to position the side chain. It should also be noted that, apart from side chains with many χ angles, given the positions of side-chain centroids it is also impossible to fix the coordinates of such side chains as Phe, Tyr, His, Asp, Asn, Glu and Gln, in which cases the rotation about the C^{β}–C^{γ} or C^{γ}–C^{δ} bond virtually does not change the position of a centroid. In fact, after the first stage of our procedure, which optimizes only the geometric-penalty function, the phenyl rings and other symmetric or nearly symmetric groups were usually rotated differently than in the crystal structure. In Fig. 4, which shows the result of the application of the complete procedure, these groups, however, are nearly superimposed on their experimental counterparts; this is a result of optimizing the

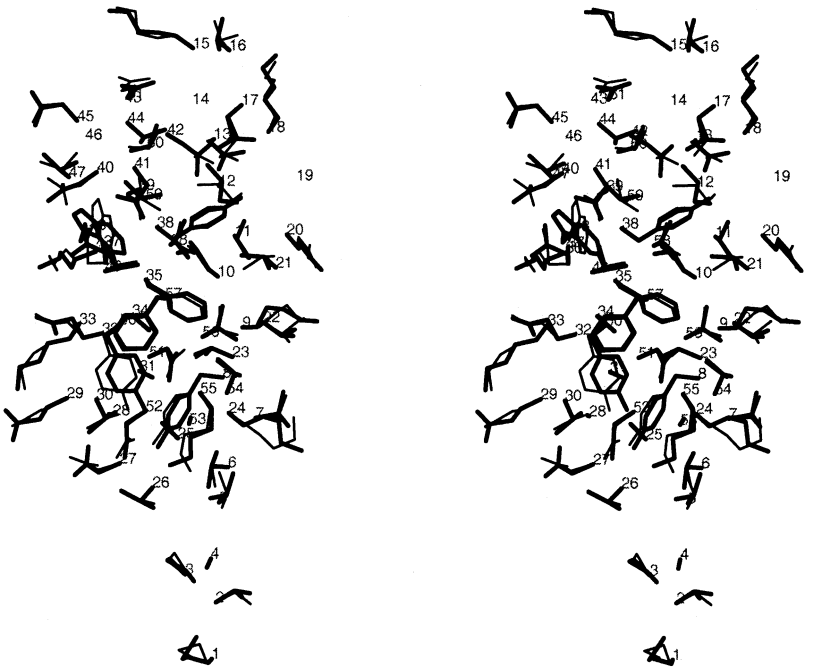

Fig. 4. Stereoview of the superposition of the side chains of the X-ray structure of 1IGD [49] (*thick lines*) on the structure resulting from the application of our procedure to the experimental backbone and centroid coordinates (*thin lines*). The RMSD for the side-chain heavy atoms is 0.62 Å. Amino acid residue numbering is provided.

overlaps. The values of zero on the RMSD plot are usually an indication of the positions of Gly residues in the protein sequence.

The RMSD of the positions of the centroids of the reconstructed structure from those of the experimental structure is 0.18 Å (Table 1). This indicates that the original positions of the centroids are well conserved. We calculated the coordinates of the side-chain centroids separately for each residue as the averaged positions over all side-chain heavy atoms and the C^{α}-atom. The deviations result from the fact that the valence geometry of the experimental polypeptide chain is different from the ECEPP/3 ideal valence geometry.

We also calculated the differences between the experimental and predicted χ torsional angles. The RMSD over all χ angles is 46.8° (Table 1). The percentages of correctly predicted χ_1 and $\chi_1+\chi_2$ angles (these are commonly used measures when assessing the quality of methods for side-chain rebuilding) are 84.0 and 76.9%, respectively.

We also tested the capability of the method developed in this work and the method developed earlier for all-atom-backbone reconstruction [14] to reconstruct the all-atom chain from C^{α} and centroid coordinates. The results are summarized in Table 1. As shown, in this case the RMSD over side-chain heavy atoms is 0.95 Å and the percentage of correctly predicted χ angles is 52.5%. The percentages of correctly predicted χ_1 or $\chi_1+\chi_2$ angles are 64.0 and 59.3%, respectively. These values are lower than the values obtained by Feig et al. [24] using their conversion method to reconstruct side-chain coordinates from centroid and C^{α} positions; it should be noted, however, that those authors reported results averaged over structures of various size and, moreover, in contrast to their work, we do not use any rotamer libraries in reconstruction.

3.2. Transcription factor PML

This is a β-protein with two short helical fragments: Glu^{34}–Met^{38}, a right-handed α-helix and Ala^{45}–Pro^{48}, a right-handed 3_{10} helix; PDB code 1BOR [51]. The superposition of the side chains

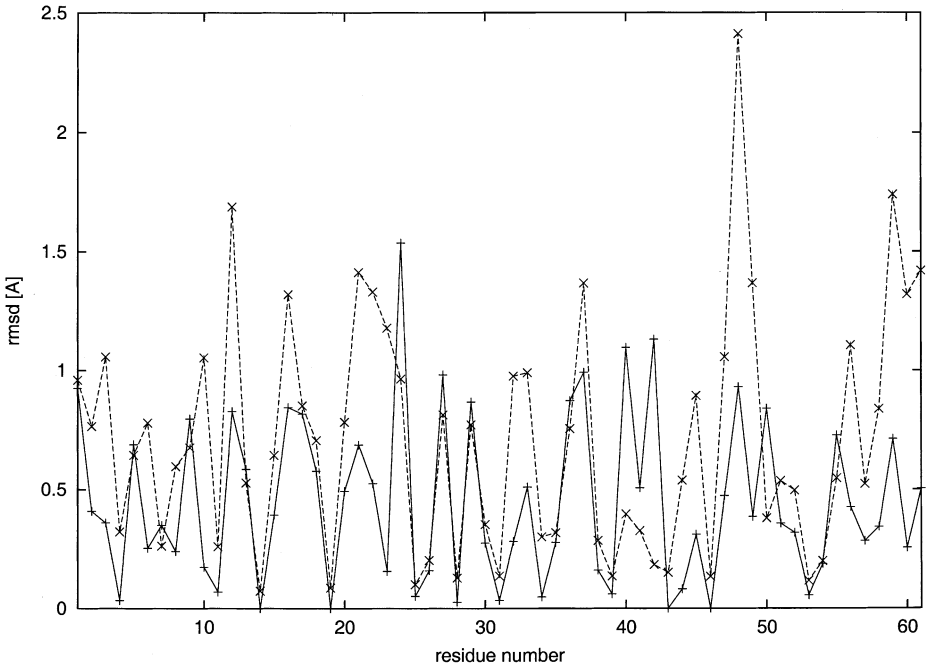

Fig. 5. Plot of the RMSD values between the heavy atoms of the experimental side chains of 1IGD [49] and those reconstructed by our procedure from experimental backbone and centroid coordinates (*solid line*), and C^{α} and centroid coordinates (*dashed line*) against the residue number. The overall RMSD values are 0.62 and 0.95 Å, respectively.

from the NMR structure and the side chains resulting from the reconstruction are presented in Fig. 6.

The RMS deviation over all side-chains heavy atoms is 0.81 Å (Table 1); it is worth noting that this value is in the same range as that of 1IGD. The RMSD values between the predicted and experimental positions of side-chain heavy atoms for individual side chains are shown in Fig. 7. The percentages of correctly predicted χ angles are 74.0% for all angles, 87.2% for the χ_1 angles and 82.9% for $\chi_1+\chi_2$ angles. The highest RMSD values occur for residues Gln^5, Leu^7, Gln^{11}, Leu^{21}, Leu^{22}, His^{26} and Leu^{49}. It can be observed (Fig. 6) that, in all these cases, the branched groups of these side chains are rotated by 180° with respect to the experimental structure, which leaves the positions of the centroids virtually unchanged and also results in the same overlaps, because the branched groups either contain different atoms at their ends (O and N for Gln, or C and N for His) or are non-planar [the terminal $-CH(CH_3)_2$ group of Leu]; therefore, their rotation by 180° results in a distinguishable configuration. It should be noted that 1BOR is an NMR structure in which many leucine residues have side chains in unlikely conformations, which might contribute to the difference in the reconstructed leucine side chains from the experimental counterparts.

As in the case of 1IGD, worse results are obtained when reconstructing the all-atom chain from C^{α} and side-chain-centroid coordinates (Table 1).

3.3. Bovine pancreatic trypsin inhibitor

This protein has almost the same amount of α-helix and β-sheet, as well as some loop regions. The superposition of the side chains from the crystal structure and the side chains resulting from the reconstruction is presented in Fig. 8.

The RMS deviation over all side-chain heavy atoms is 0.67 Å, while that over all centroid positions is 0.14 Å. The RMSD values of the

Table 1
RMS deviations for all side-chain heavy atoms, the χ torsional angles and side-chain centroids between the reconstructed and experimental structures, as well as the percentages of correctly predicted χ torsional angles of the six proteins tested

Protein PDB code	nres	RMS (Å)	RMS_c (Å)	RMS_χ (°)	RMS_{χ_1} (°)	$RMS_{\chi_{12}}$ (°)	$\%\chi$ (%)	$\%\chi_1$ (%)	$\%\chi_{12}$ (%)	Time (min)
1BDD(bb,cen)[a]		0.90	0.16	63.6	27.7	52.9	60.6	85.0	68.8	37
1BDD(C^α,cen)[b]	46	0.94	0.34	61.3	30.0	49.4	59.6	85.0	68.8	39
1BDD(UNRES)[c]		4.09	3.71	78.6	76.2	70.6	37.2	42.5	42.8	39
1IGD(bb,cen)[a]	61	0.62	0.18	46.8	28.5	39.5	73.3	84.0	76.9	39
1IGD(C^α,cen)[b]		0.95	0.25	64.1	49.3	54.5	52.5	64.0	59.3	42
1BPI(bb,cen)[a]	58	0.67	0.22	52.8	32.1	46.5	72.4	84.8	77.8	39
1BPI(C^α,cen)[b]		1.00	0.50	59.9	47.2	58.1	62.9	71.7	66.7	41
1BOR(bb,cen)[a]	56	0.81	0.14	57.5	25.5	50.0	74.0	87.2	82.9	32
1BOR(C^α,cen)[b]		1.07	0.48	63.5	50.2	58.7	58.0	74.5	64.6	35
1BZ0(bb,cen)[a]	141	0.69	0.14	45.0	20.3	36.9	77.5	93.8	82.8	43
1BZ0(C^α,cen)[b]		0.88	0.33	54.0	36.4	45.4	67.0	80.5	71.8	59
1A4U(bb,cen)[a]	254	0.69	0.13	48.9	22.1	42.3	75.3	93.0	80.1	63
1A4U(C^α,cen)[b]		0.90	0.34	56.8	36.9	51.0	64.0	81.0	69.5	120

nres, the number of residues; RMS, the overall side-chain RMSD between the reconstructed and experimental structure; RMS_c, the RMSD between side-chain centroids of the reconstructed and experimental structure; RMS_χ, the RMSD between the experimental and reconstructed χ torsional angles; RMS_{χ_1}, computed only over the χ_1 torsional angles; $RMS_{\chi_{12}}$, computed only over the χ_1 and χ_2 torsional angles; $\%\chi$, the percentage of correctly predicted χ torsional angles, defined as the percentage of χ torsional angles that deviate by less than 40° from those of the experimental structure; $\%\chi_1$, computed only over the χ_1 torsional angles; $\%\chi_{12}$, computed only over the χ_1 and χ_2 torsional angles; time, time necessary to perform the conversion, which is equal to the time for performing side-chain reconstruction for the (bb,cen) entries, the sum of the times for backbone reconstruction using the dipole-path method [13,14] and side-chain reconstruction for the (C^α,cen) and the (UNRES) entries, respectively.

[a] (bb,cen): reconstructed from experimental backbone and centroids.

[b] (C^α,cen): reconstructed from experimental C^α coordinates and centroids.

[c] UNRES: reconstructed from the UNRES model.

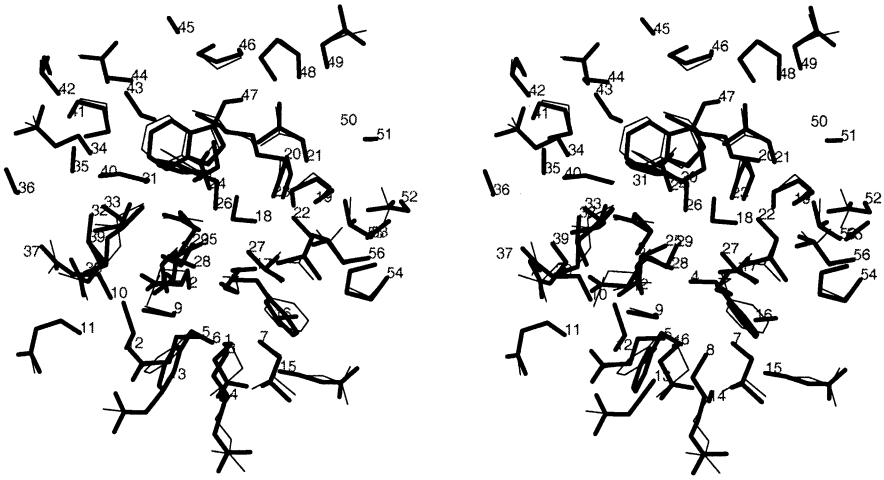

Fig. 6. Stereoview of the superposition of the side chains of the NMR structure of 1BOR [51] (*thick lines*) on the structure resulting from the application of our procedure (*thin lines*). The RMSD for the side-chain heavy atoms is 0.81 Å. Amino acid residue numbering is provided.

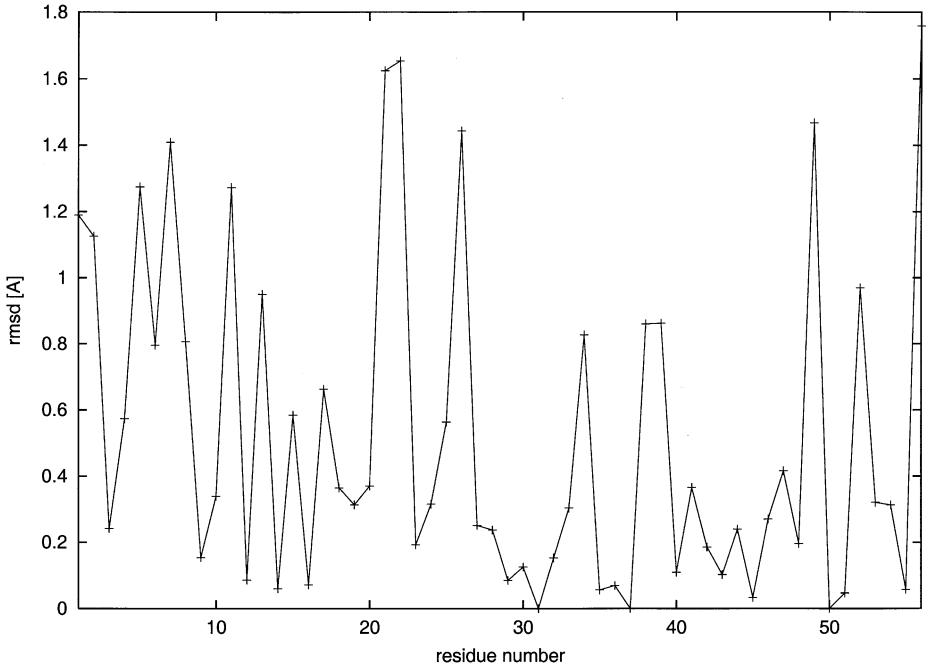

Fig. 7. Plot of the RMSD values between the side-chain heavy atoms predicted by our procedure and those taken from the NMR structure of 1BOR [51] against the residue number.

positions of side-chain heavy atoms of the reconstructed structure from those of the crystal structure are presented in Fig. 9. The highest deviations occur for Arg^{42} and Arg^{53}, for which the χ_1 angles differ by approximately 180° from the experimental values, while the other χ angles have opposite sign. This results in a nearly mirror-image configuration of a side chain with the same centroid position and comparable overlaps with the neighboring atoms.

3.4. Larger proteins

We also tested the ability of our method to add side chains to longer backbones within a reasonable time. The selected proteins were: chain A of human deoxyhemoglobin, PDB code 1BZ0 [52]; and chain A of the alcohol dehydrogenase from *Drosophilia lebanonensis*, PDB code 1A4U [53]. The RMSD values over side-chain heavy atoms, centroid positions and χ angles, as well as the percentage of correctly predicted χ angles are shown in Table 1. As can be observed, for reconstruction from complete backbone coordinates, these measures are comparable with values obtained for smaller proteins, which shows that the method is convergent. It is also interesting to note that, for reconstruction from the C^{α} and centroid coordinates, the percentages of predicted χ angles have appreciably increased compared to those obtained for smaller proteins (Table 1), and approach the values reported in the work of Feig et al. [24] using their method of reconstructing all-atom chains from C^{α} and centroid coordinates. The superpositions of the side chains from the crystal structure and the side chains resulting from the reconstruction, along with the RMSD values for side-chain heavy atoms from the experimental structures, are presented in Figs. 10 and 12 for 1BZ0 and in Figs. 11 and 13 for 1A4U, respectively.

3.5. Protein A

Protein A [54] is a three-α-helix bundle. It served as a convenient test case protein for UNRES

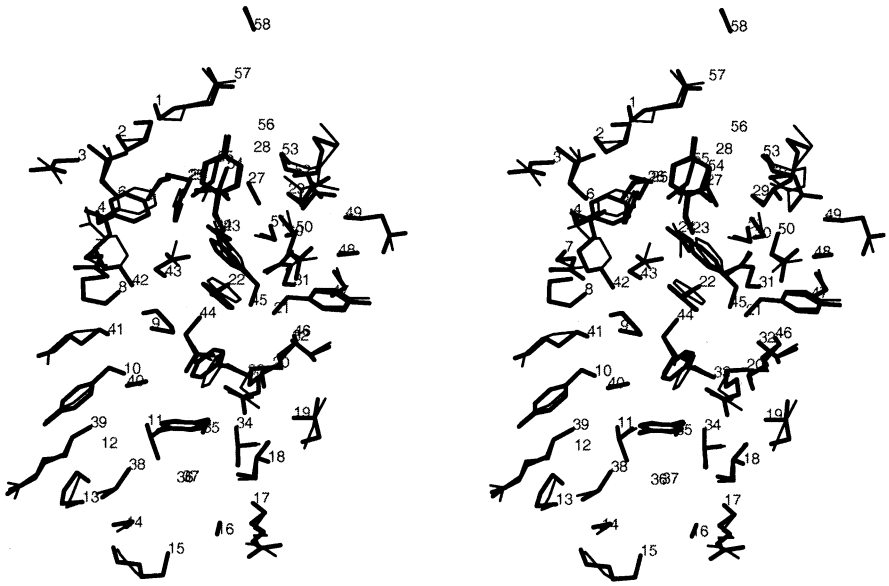

Fig. 8. Stereoview of the superposition of the side chains of the X-ray structure of 1BPI [50] (*thick lines*) on the structure resulting from the application of our procedure (*thin lines*). The RMSD for the side-chain heavy atoms is 0.67 Å. Amino acid residue numbering is provided.

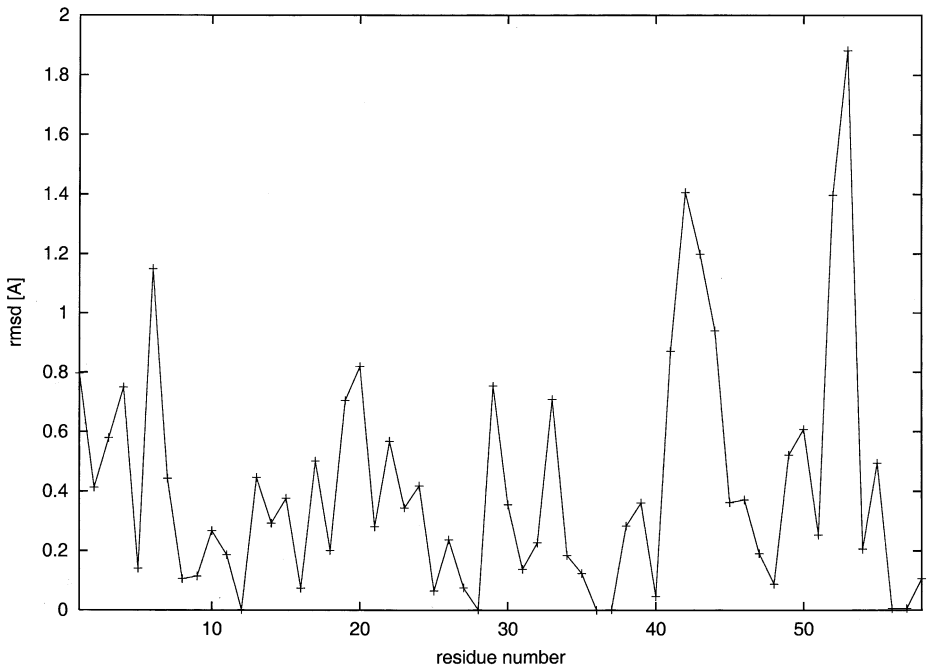

Fig. 9. Plot of the RMSD values between the side-chain heavy atoms reconstructed by our procedure from experimental backbone and centroid coordinates and those taken from the X-ray structure of 1BPI [50] against the residue number.

Fig. 10. Stereoview of the superposition of the side chains of the X-ray structure of 1BZ0 [52] (*thick lines*) on the structure resulting from the application of our procedure (*thin lines*). The RMSD for the side-chain heavy atoms is 0.69 Å. Amino acid residue numbering is provided.

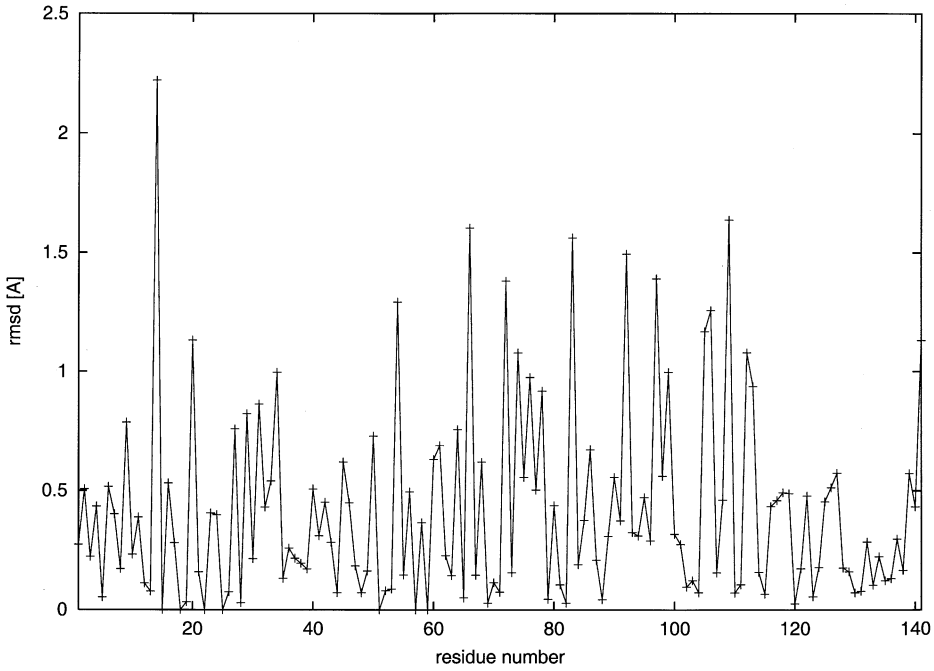

Fig. 11. Plot of the RMSD values between the side-chain heavy atoms reconstructed by our procedure from experimental backbone and centroid coordinates and those taken from the X-ray structure of 1BZ0 [52] against the residue number.

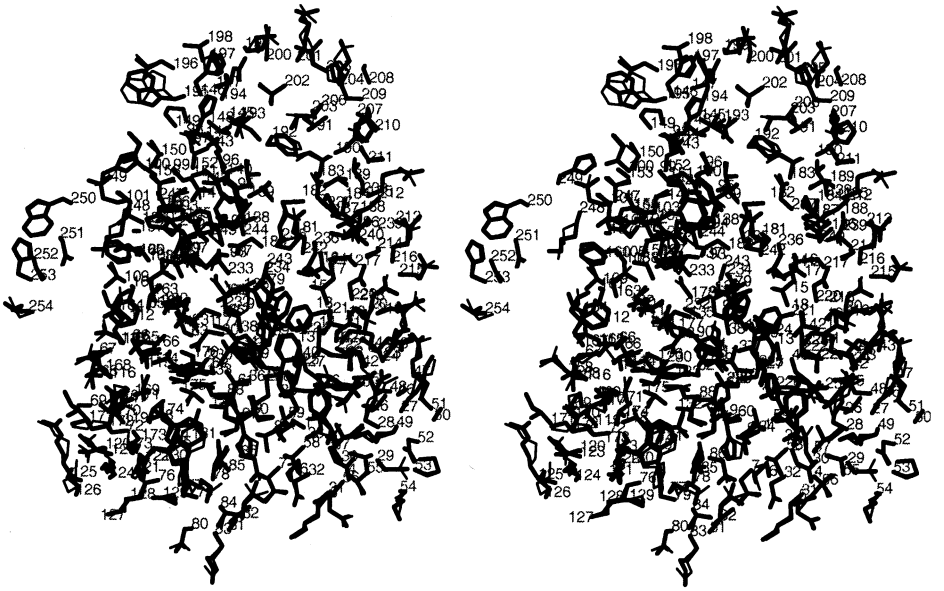

Fig. 12. Stereoview of the superposition of the side chains of the X-ray structure of 1A4U [53] (*thick lines*) on the structure resulting from the application of our procedure (*thin lines*). The RMSD for the side-chain heavy atoms is 0.69 Å. Amino acid residue numbering is provided.

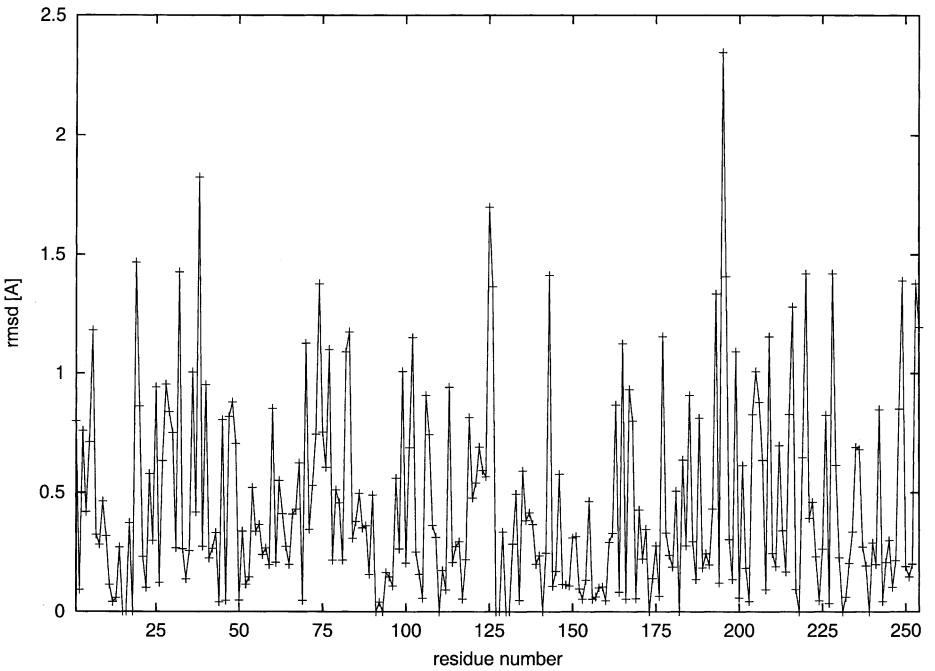

Fig. 13. Plot of the RMSD values between the side-chain heavy atoms reconstructed by our procedure from experimental backbone and centroid coordinates and those taken from the X-ray structure of 1A4U [53] against the residue number.

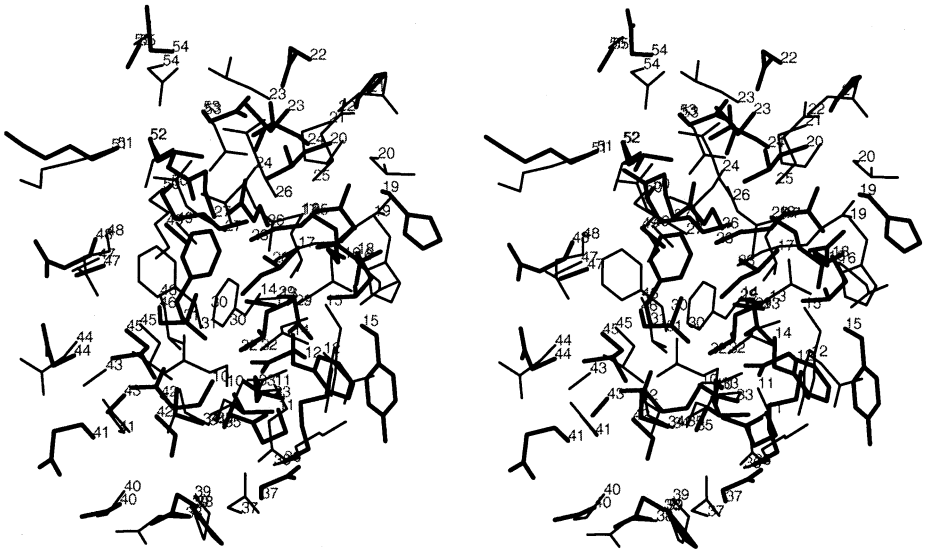

Fig. 14. Stereoview of the superposition of the side chains of the NMR structure of protein A [54] (*thick lines*) on the structure resulting from the application of our procedure to the C^{α} and centroid coordinates of a 1.96-Å-resolution UNRES model (*thin lines*). The RMSD for the side-chain heavy atoms is 4.09 Å. Amino acid residue numbering is provided for both the experimental structure and the UNRES model.

force-field simulations [11], and the RMS deviation of the C^{α} trace predicted by the UNRES force field from the NMR structure was 1.96 Å. For the backbone reconstruction, we used a recently improved version [14] of the dipole-path method [13,14]. The overall RMSD of the side-chain heavy atoms from the NMR structure is 4.09 Å and the percentage of correctly predicted χ torsional angles is 37.2% (Table 1). This unfavorable result arises from the fact that centroid positions are poorly reproduced in the UNRES model; the RMSD from the centroids calculated from the NMR structure is 3.62 Å, which is nearly two-fold worse than the C^{α} RMSD. Such strong dependence of the side-chain orientation on the backbone conformation was also observed in earlier works [18,55]. It should be noted that the RMSD of the centroids of the reconstructed side chains from those of the parent UNRES model is only 0.10 Å, which shows that the procedure also performed well in this case. The superposition of the predicted side chains and the NMR structure is presented in Fig. 14, and side-chain RMSD values are plotted as a function of residue number in Fig. 15.

We also tried to reconstruct the side chains of protein A using the NMR backbone and side-chain centroids and NMR C^{α} coordinates and side-chain centroids. The results are summarized in Table 1, and the side-chain RMSD is plotted in Fig. 15. It can be observed that the results are of the same quality as for other proteins.

We also examined the RMS deviations over all C^{α}–C^{β} pairs for each protein tested. Their values vary between 0.04 and 0.13 Å, which indirectly confirms that all added side chains are in the L-configuration. This relatively small value of the RMS deviation is a consequence of replacing real residues by the ECEPP/3 rigid-body models. All values reported for the RMS deviations, along with the protein sizes and times of program execution are summarized in Table 1.

3.6. Timing

The computation time ranged from 900 s (30 000 MC steps performed) for the 1BPI system to 3600 s (100 000 MC steps performed) for 1A4U. We combined our current program for side-

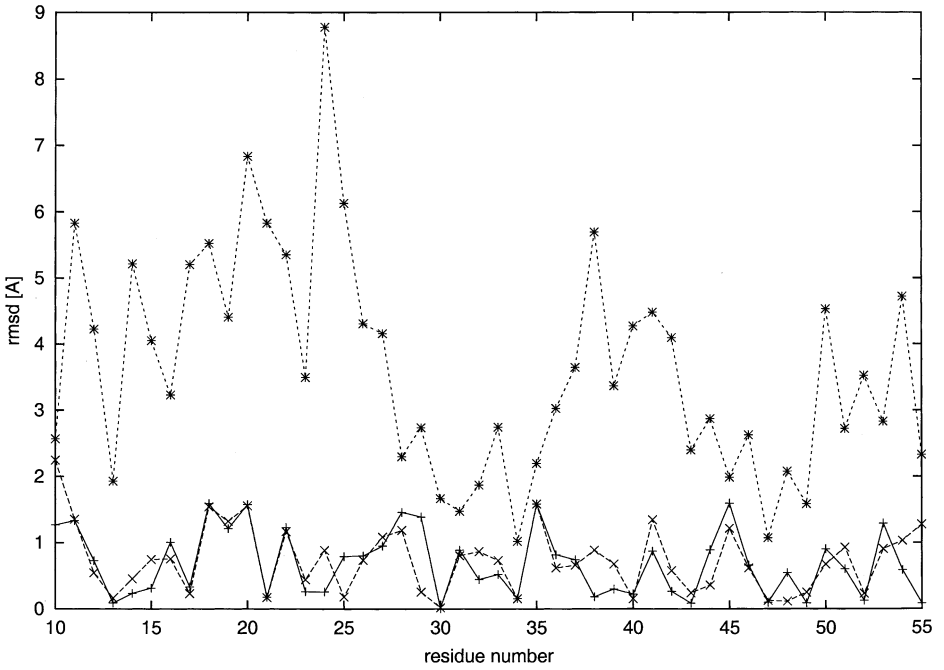

Fig. 15. Plots of the RMSD values between the side-chain heavy atoms reconstructed by our procedure from experimental backbone and centroid coordinates (*solid line*), from experimental C^{α} and centroid coordinates (*dashed lines*), and from the 1.96-Å-resolution UNRES model [11] (*dotted lines*) and those taken from the NMR structure of protein A [54] against the residue number. The RMSD over all centroids is 0.10Å.

chain prediction with the dipole-path method [13,14], which, besides its ability to reconstruct the backbone from the C^{α}-trace, can now add side chains to the reconstructed backbone, while preserving the side-chain centroids. At the same time, we improved the speed of part of the program (reconstructing the backbone from the C^{α}-trace) by introducing a cut-off of 7.0 Å for the interaction energy defined in our previous work [14]. The time for reconstructing backbones now ranged from 108 s for protein A to 3600 s for 1A4U. All these data refer to a single processor of a 500-MHz Pentium III computer running LINUX.

4. Conclusions

The algorithm for adding side chains to a fixed backbone with preservation of the side-chain centroids, described in this work, has been shown to be capable of restoring the correct side-chain geometry within 0.6–0.9 Å RMSD when complete backbone coordinates and side-chain-centroid coordinates are used as input data, or approximately 1.0 Å when C^{α} and centroid coordinates are used as input data, which is of comparable quality to that of other methods [15–20,24,30,55]. Moreover, the predicted side-chain centroid positions are within 0.1–0.2 Å RMSD from the original experimental structures when complete backbone coordinates and side-chain-centroid coordinates are used as input data, or 0.2–0.3 Å when C^{α} and centroid coordinates are used as input data. The example of the UNRES model of protein A shows that the quality of the restored side chains strongly depends on how well the centroid positions are predicted by the coarse-grain model. Most of the coarse-grained energy functions, especially those which use C^{α} coordinates as the only interacting sites, are optimized to reproduce a C^{α}-trace, while reproduction of the side-chain-centroid positions is given less attention. The method presented here is fully based on energy optimization, and imple-

ments a simplified energy function to facilitate the search for optimal orientations of the side chains. This method is designed to reconstruct an all-atom protein structure from a coarse-grain model, predicted with the UNRES force field, in reasonable time. It should be noted that the energy function used here to compute the interactions involving side chains is very simple. It can easily be enhanced to include explicit torsional potentials and a simplified potential accounting for hydrogen-bond formation between side-chains bearing proton-donor and/or proton-acceptor groups. This is currently under investigation in our laboratory.

In this work, we also extended our backbone reconstruction method [14] to make possible computation of the complete all-atom geometry from its C^{α}-trace and the positions of the side-chain centroids; this information is contained in coarse-grain structures predicted with the UNRES force field. By applying a distance cut-off in the evaluation of the dipole-interaction energy, we reduced the computational time of the previous procedure [14].

5. Note added in proof

Some technical details about changes in the program that have been implemented after submitting the manuscript are presented below.

We focused our efforts on improving the speed of the program and the accuracy of the calculation of the torsional angles χ of the added side chains. To achieve better agreement with the values of the torsional angles χ observed in crystal structures of proteins, we modified the energy function. Since our method for adding side chains to the backbone does not use rotamer libraries, we included explicit torsional potentials from the ECEPP/3 force field [45] in our simplified energy function. The ability do differentiate between up and down conformations of proline residues was also added. In addition, we included the deterministic SUMSL algorithm [56] for energy minimization.

The algorithm is now being extended to coarse-grain structures of multichain proteins.

Acknowledgments

This work was supported by grants from the National Institutes of Health (GM-14312), the National Science Foundation (MCB00-03722), the Fogarty Foundation (R03 TW1064), the NIH National Center for Research Resources (P41RR-04293), and the Polish State Committee for Scientific Research, KBN (3 T09A 111 17). Support was also received from the National Foundation for Cancer Research. The computations were carried out at the Cornell Theory Center in Ithaca, NY, the Informatics Center of the Metropolitan Academic Network (IC MAN) in Gdańsk, and the Interdisciplinary Center for Mathematical Modeling (ICM) in Warsaw, Poland.

References

[1] A. Liwo, M.R. Pincus, R.J. Wawak, S. Rackovsky, H.A. Scheraga, Prediction of protein conformation on the basis of a search for compact structures; test on avian pancreatic polypeptide, Protein Sci. 2 (1993) 1715–1731.

[2] A. Liwo, S. Ołdziej, M.R. Pincus, R.J. Wawak, S. Rackovsky, H.A. Scheraga, 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. 18 (1997) 849–873.

[3] A. Liwo, M.R. Pincus, R.J. Wawak, S. Rackovsky, S. Ołdziej, H.A. Scheraga, A united-residue force field for off-lattice protein-structure simulations. II. Parameterization of local interactions and determination of weights of energy terms by *Z*-score optimization, J. Comput. Chem. 18 (1997) 874–887.

[4] A. Liwo, R. Kaźmierkiewicz, C. Czaplewski, et al., United-residue force field for off-lattice protein-structure simulations. III. Origin of backbone hydrogen-bonding cooperativity in united-residue potentials, J. Comput. Chem. 19 (1998) 259–276.

[5] J. Lee, A. Liwo, H.A. Scheraga, 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 96 (1999) 2025–2030.

[6] A. Liwo, J. Lee, D.R. Ripoll, J. Pillardy, H.A. Scheraga, Protein structure prediction by global optimization of a potential energy function, Proc. Natl. Acad. Sci. USA 96 (1999) 5482–5485.

[7] J. Lee, A. Liwo, D.R. Ripoll, et al., Hierarchical energy-based approach to protein-structure prediction. Blind-test evaluation with CASP3 targets, Int. J. Quant. Chem. 77 (2000) 90–117.

[8] A. Liwo, J. Pillardy, C. Czaplewski, et al., UNRES—a united-residue force field for energy-based prediction of protein structure-origin and significance of multibody

terms, in: R. Shamir, S. Miyano, S. Istrail, P. Pevzner, M. Waterman (Eds.), RECOMB 2000, Proceedings of the Fourth Annual International Conference on Computational Molecular Biology, ACM, New York, 2000.

[9] J. Pillardy, C. Czaplewski, A. Liwo, et al., Recent improvements in prediction of protein structure by global optimization of a potential energy function, Proc. Natl. Acad. Sci. USA 98 (2001) 2329–2333.

[10] A. Liwo, C. Czaplewski, J. Pillardy, H.A. Scheraga, Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field, J. Chem. Phys. 115 (2001) 2323–2347.

[11] J. Lee, D.R. Ripoll, C. Czaplewski, J. Pillardy, W.J. Wedemeyer, H.A. Scheraga, Optimization of parameters in macromolecular potential energy functions by conformational space annealing, J. Phys. Chem. B 105 (2001) 7291–7298.

[12] J. Pillardy, C. Czaplewski, A. Liwo, et al., Development of physics-based energy functions that predict medium-resolution structures for proteins of the α, β, and α/β structural class, J. Phys. Chem. B 105 (2001) 7299–7311.

[13] A. Liwo, M.R. Pincus, R.J. Wawak, S. Rackovsky, H.A. Scheraga, Calculation of protein backbone geometry from α-carbon coordinates based on peptide-group dipole alignment, Protein Sci. 2 (1993) 1697–1714.

[14] R. Kaźmierkiewicz, A. Liwo, H.A. Scheraga, Energy-based reconstruction of a protein backbone from its α-carbon trace by a Monte Carlo method, J. Comput. Chem. 23 (2002) 715–723.

[15] C. Lee, S. Subbiah, Prediction of protein side-chain conformation by packing optimization, J. Mol. Biol. 217 (1991) 373–388.

[16] P. Tuffery, C. Etchebest, S. Hazout, R. Lavery, A new approach to the rapid determination of protein side-chain conformations, J. Biomol. Struct. Dyn. 8 (1991) 1267–1289.

[17] J. Desmet, M. De Maeyer, B. Hazes, I. Lasters, The dead-end elimination theorem and its use in protein side-chain positioning, Nature 356 (1992) 539–542.

[18] R.L. Dunbrack Jr, M. Karplus, Backbone-dependent rotamer library for proteins. Application to side-chain prediction, J. Mol. Biol. 230 (1993) 543–574.

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

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

[21] L.S. Reid, J.M. Thornton, Rebuilding flavodoxin from C^{α} coordinates: a test study, Proteins: Struct. Funct. Genet. 5 (1989) 170–182.

[22] P.E. Correa, The building of protein structures from α-carbon coordinates, Proteins: Struct. Funct. Genet. 7 (1990) 366–377.

[23] L. Holm, C. Sander, Database algorithm for generating protein backbone and side-chain co-ordinates from a C^{α} trace. Application to model building and detection of co-ordinate errors, J. Mol. Biol. 218 (1991) 183–194.

[24] M. Feig, P. Rotkiewicz, A. Koliński, J. Skolnick, C.L. Brooks III, Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models, Proteins: Struct. Funct. Genet. 41 (2000) 86–97.

[25] N.L. Summers, M. Karplus, Construction of side-chains in homology modeling. Application to the C-terminal lobe of rhizopuspepsin, J. Mol. Biol. 210 (1989) 785–811.

[26] M. Levitt, Accurate modeling of protein conformation by automatic segment matching, J. Mol. Biol. 226 (1992) 507–533.

[27] C. Wilson, L.M. Gregoret, D.A. Agard, Modeling side-chain conformation for homologous proteins using an energy-based rotamer search, J. Mol. Biol. 229 (1993) 996–1006.

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

[29] M.J. Bower, F.E. Cohen, R.L. Dunbrack Jr, Prediction of protein side-chain rotamers from a backbone-dependent rotamer library, J. Mol. Biol. 267 (1997) 1268–1282.

[30] R.L. Dunbrack Jr, Comparative modeling of CASP3 targets using PSI-BLAST and SCWRL, Proteins: Struct. Funct. Genet. Suppl. 3 (1999) 81–87.

[31] T.L. Blundell, B.L. Sibanda, M.J.E. Sternberg, J.M. Thornton, Knowledge-based prediction of protein structures and the design of novel molecules, Nature 326 (1987) 347–352.

[32] N.L. Summers, W.D. Carlson, M. Karplus, An analysis of side-chain orientations in homologous proteins, J. Mol. Biol. 196 (1987) 175–198.

[33] J.W. Ponder, F.M. Richards, Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes, J. Mol. Biol. 193 (1987) 775–791.

[34] S.C. Lovell, J.M. Word, J.S. Richardson, D.C. Richardson, The penultimate rotamer library, Proteins: Struct. Funct. Genet. 40 (2000) 389–408.

[35] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equation of state calculations by fast computing machines, J. Chem. Phys. 21 (1953) 1087–1092.

[36] M. Nilges, A.T. Brünger, Automated modeling of coiled coils: application to the GCN4 dimerization region, Protein Eng. 4 (1991) 649–659.

[37] I. Lasters, J. Desmet, The fuzzy-end elimination theorem: correctly implementing the side-chain placement

algorithm based on the dead-end elimination theorem, Prot. Eng. 6 (1993) 717–722.

[38] A. Roitberg, R. Elber, Modeling side chains in peptides and proteins: application of the locally enhanced sampling and the simulated annealing methods to find minimum energy conformations, J. Chem. Phys. 95 (1991) 9277–9287.

[39] R.B. Gerber, V. Buch, M.A. Ratner, Time-dependent self-consistent field approximation for intramolecular energy transfer. I. Formulation and application to dissociation of van der Waals molecules, J. Chem. Phys. 77 (1982) 3022–3030.

[40] R. Elber, M. Karplus, Enhanced sampling in molecular dynamics: use of the time-dependent Hartree approximation for a simulation of carbon monoxide diffusion through myoglobin, J. Am. Chem. Soc. 112 (1990) 9161–9175.

[41] R. Czerminski, R. Elber, Computational studies of ligand diffusion in globins. I. Leghemoglobin, Proteins: Struct. Funct. Genet. 10 (1991) 70–80.

[42] A. Koliński, P. Rotkiewicz, J. Skolnick, Application of high-coordination lattice model in protein structure prediction, in: P. Grassberger, G.T. Barkema, W. Nadler (Eds.), Monte Carlo Approach to Biopolymers and Protein Folding, World Scientific, Singapore/London, 1998.

[43] A. Koliński, P. Rotkiewicz, B. Ilkowski, J. Skolnick, Protein folding: flexible lattice models, Prog. Theor. Phys. Suppl. 138 (2000) 292–300.

[44] A. Koliński, M.R. Betancourt, D. Kihara, P. Rotkiewicz, J. Skolnick, Generalized comparative modeling (GENECOMP): a combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement, Proteins: Struct. Funct. Genet. 44 (2001) 133–149.

[45] G. Némethy, K.D. Gibson, K.A. Palmer, et al., Energy parameters in polypeptides. 10. Improved geometrical parameters and non-bonded interactions for use in the ECEPP/3 algorithm with application to proline-containing peptides, J. Phys. Chem. 96 (1992) 6472–6484.

[46] IUPAC-IUB Commission On Biochemical Nomenclature, Abbreviations and symbols for the description of the conformation of polypeptide chains, Biochemistry 9 (1970) 3471–3479.

[47] M. Vásquez, H.A. Scheraga, Calculation of protein conformation by the build-up procedure. Application to bovine pancreatic trypsin inhibitor using limited simulated nuclear magnetic resonance data, J. Biomol. Struct. Dyn. 5 (1988) 705–755.

[48] S.K. Kearsley, On the orthogonal transformation used for structural comparisons, Acta Cryst. A 45 (1989) 208–210.

[49] J.P. Derrick, D.B. Wigley, The third IgG-binding domain from streptococcal protein G. An analysis by X-ray crystallography of the structure alone and in a complex with Fab, J. Mol. Biol. 243 (1994) 906–918.

[50] S. Parkin, B. Rupp, H. Hope, Structure of bovine pancreatic trypsin inhibitor at 125 K: Definition of carboxyl-terminal residues Gly57 and Ala58, Acta Cryst. D 52 (1996) 18–29.

[51] K.L.B. Borden, M.N. Boddy, J. Lally, et al., The solution structure of the RING finger domain from the acute promyelocytic leukaemia proto-oncoprotein PML, EMBO J. 14 (1995) 1532–1541.

[52] J.S. Kavanaugh, W.F. Moo-Penn, A. Arnone, Accommodation of insertions in helices: the mutation in hemoglobin catonsville (Pro 37^{α}–Glu–Thr 38^{α}) generates a $3_{10} \rightarrow \alpha$ bulge, Biochemistry 32 (1993) 2509–2513.

[53] J. Benach, S. Atrian, R Gonzalez-Duarte, R. Ladenstein, The refined crystal structure of *Drosophila lebanonensis* alcohol dehydrogenase at 1.9 Å resolution, J. Mol. Biol. 282 (1998) 383–399.

[54] H. Gouda, H. Torigoe, A. Saito, M. Sato, Y. Arata, I. Shimada, Three-dimensional solution structure of the B domain of staphylococcal protein A: comparisons of the solution and crystal structures, Biochemistry 31 (1992) 9665–9672.

[55] E.S. Huang, P. Koehl, M. Levitt, R.V. Pappu, J.W. Ponder, Accuracy of side-chain prediction upon near-native protein backbones generated by ab initio folding methods, Proteins: Struct. Funct. Genet. 33 (1998) 204–217.

[56] D.M. Gay, Subroutines for unconstrained minimization using a model trust-region approach, ACM Trans. Math. Software 9 (1983) 503–524.