Stan Galaktionov
Gregory V. Nikiforovich
Garland R. Marshall
Department of Biochemistry and Molecular Biophysics, Washington University, Campus Box 8036, St. Louis, MO 63110

Ab Initio Modeling of Small, Medium, and Large Loops in Proteins

Abstract: *This study presents different procedures for ab initio modeling of peptide loops of different sizes in proteins. Small loops (up to 8–12 residues) were generated by a straightforward procedure with subsequent "averaging" over all the low-energy conformers obtained. The averaged conformer fairly represents the entire set of low-energy conformers, root mean square deviation (RMSD) values being from 1.01 Å for a 4-residue loop to 1.94 Å for an 8-residue loop. Three-dimensional (3D) structures for several medium loops (20–30 residues) and for two large loops (54 and 61 residues) were predicted using residue–residue contact matrices divided into variable parts corresponding to the loops, and into a constant part corresponding to the known core of the protein. For each medium loop, a very limited number of sterically reasonable C^{α} traces (from 1 to 3) was found; RMSD values ranged from 2.4 to 5.9 Å. Single C^{α} traces predicted for each of the large loops possessed RMSD values of 4.5 Å. Generally, ab initio loop modeling presented in this work combines elements of computational procedures developed both for protein folding and for peptide conformational analysis.* © 2001 John Wiley & Sons, Inc. Biopolymers (Pept Sci) 60: 153–168, 2001

Keywords: *ab initio modeling; protein loops; protein folding*

INTRODUCTION

Determining three-dimensional (3D) structures of the loops in proteins is one of the central tasks facing biomolecular modeling. The importance of this task is underlined by the fact that conformational transitions in proteins involving loops are very difficult to observe experimentally: even if the x-ray structure of an entire protein is available, it will provide just one snapshot out of many conformational possibilities existing for its loops. A good example is the ternary complex of the HIV glycoprotein gp120, the cell membrane-bound protein CD4, and an antibody. To crystallize the complex, the authors had to use an engineered "core" gp120 that lacked the highly mobile V1/V2 and V3 loops crucial for almost all biologically relevant interactions.[1]

Current efforts in modeling loops in proteins fall into two main categories. One is ab initio modeling, restoring feasible loop structures based on some general physical principles. The other is a knowledge-based approach, which basically substitutes the 3D structure of a given loop by 3D structure(s) of another loop(s) selected among those with the experimentally known structures based on some criteria of similarity. Both approaches have rapidly developed in the last

Correspondence to: Gregory V. Nikiforovich
Contract grant sponsor: Monsanto Company and National Institutes of Health (NIH)
Contract grant number: EY12113, GM48184, and HL54085
Biopolymers (Peptide Science), Vol. 60, 153–168 (2001)

decade; for the latest review see, e.g., Ref. 2, and the references therein.

The general outline of the most ab initio procedures includes (a) generating the initial set of loop conformers, (b) insertion of this set into the existing structure of the protein, (c) refinement of these structures by energy calculations, and (d) selection of the most "appropriate" final structure(s) of the loop (cf. Ref. 3). Various computational algorithms have been developed to implement these procedures; some examples are described below. One of the earliest was Monte Carlo sampling starting from an extended loop structure constrained by the requirements to keep the endpoints of the loop at the chosen distances [4]. This procedure was applied to eight loops ranging from 7 to 9 residues in size. The best results yielded a root mean square deviation (RMSD) value of 1.53 Å for the heavy atoms of the backbone in the 9-residue loop (the T2 loop[4]), which includes contribution from the constrained end-point residues (The RMSD values mentioned here and further in the text relate to "loop-to-loop" RMSD values (see below), if otherwise not specified). Other authors suggested a Monte Carlo simulated annealing procedure to generate energetically reasonable structures within the existing protein structure starting from a random conformation.[5] In this case, the best result (obtained for the longest loop) was a RMSD value of 1.87 Å for the 9-residue loop bovine pancreatic trypsin inhibitor (BPTI) 10–18. The same procedure with modification of the Monte Carlo procedure resulted in a RMSD = 1.19 Å or the 9-residue loop 2rhe 24–32, and with some modification of the force field used, in a RMSD = 0.93 Å for the same 9-residue loop BPTI 10–18.[7] Combination of the Monte Carlo and molecular dynamics simulations with employment of a solvation model obtained a low energy structure of the 12-residue loop (ribonuclease A 13–24) with the RMSD = 0.80 Å; in this case, however, the RMSD value has been calculated for the backbone atoms of the entire protein.[3] An elegant algorithm, the valence geometry scaling-relaxation, "to fill" the absent parts of the 3D protein structure starting from random segments, has been applied to the 7-residue loop 7rxn 16–22, and yielded the RMSD value of 0.70 Å[8]; the same algorithm enhanced with the multiple copy sampling lowered the RMSD value for the same loop to 0.54 Å[9]. A very recent paper used generation of random conformations for the loop in an already existing protein environment.[2] Subsequent energy minimization of the loop conformers was performed employing a "pseudo-energy" scoring function deduced mainly from distributions for the dihedral angles values in the backbone and side chains that are experimentally observed in the Protein Data Bank (PDB). It appeared that 90% of the low-energy conformers generated for loops up to 8 residues possess RMSD values less than 2 Å; for 12-residue loops, however, this number decreases to 30%, with further reduction to only 5% for the largest 14-residue loops.[2] These results are typical for current ab initio procedures, as well as typical in the problems they are confronting. First, they handle only relatively small loops, since the number of conformers to consider increases exponentially with the size of the loop. Second, the choice of the "right" conformer among those obtained by the procedure is often not straightforward; the lowest-energy conformer does not necessarily correspond to the lowest RMSD value.

The knowledge-based approach depends on a representative database composed of loops of the appropriate size from known 3D structures, which can be used as a "training set" for the initial selection of possible conformers of the loops with an unknown 3D structure. Some sort of energy minimization regularly follows the initial selection. Different aspects of building loop databases are discussed in several recent publications (e.g., Refs. 10–12). The loop databases may be built analyzing either the values of the dihedral angles for protein backbone, as in Ref. 13, or selected interatomic distances (e.g., between C^{α} and C^{β} atoms,[14] or both[12]). One of the first databases was proposed in Ref. 13, where the key angles were the ϕ_{i+1} and ψ_i along the peptide chain. The best results obtained using this database were prediction of ten plausible structures for the 11-residue disulfide-bound crambin fragment 16–26 with RMSD values less than 2.0 Å (out of 250,000 simulations).[13] A loop database constructed from the PDB data was employed to predict low-energy structures of loops up to 9 residues with RMSD values less than 1.79 Å.[14] Recently, a rather sophisticated search procedure over an exhaustive loop database yielded predictions of 8-residue loops with an average RMSD values of 3.8 Å.[12] Obviously, the knowledge-based approach is limited by the average size of loops included in the database; it is hardly feasible to build a representative database of loops larger than 20 residues (less than 100 loops of the size over 19 residues have been found in the PDB recently[2]). Therefore, the only option for dealing with loops of larger size within this approach is to employ homology modeling. The best results obtained by homology modeling at the CASP-1998 event were predictions of several loops of 10–12 residues with a RMSD of 2.5–3.5 Å, and one 20-residue loop with a RMSD of 5.0 Å (reviewed in Ref. 15).

Both approaches have their own obvious limitations, which are, in part, discussed above. However, there is

one more limitation that is not so obvious. Namely, both approaches attempt to predict all loops in proteins with the same procedure, disregarding such important factors as the size of the loop and specific goals of this particular prediction. For instance, if the loop is a short one (3–5 residues) connecting highly structured protein fragments, as transmembrane (TM) helices (see below), it will be easy to generate all possible conformers of such loop with any of several available computational procedures. At the same time, one may expect that all conformers of the short loop will be geometrically similar, so almost any of the energetically reasonable conformers will represent a fairly good prediction. On the other hand, if the loop is more than 30–40 residues long, even estimations of the conformational energy will be less reliable due to inevitable uncertainty in the force field employed. Also, the large loop is expected to be much more flexible than a short one: should the prediction be limited to only one "best" structure, or it would be more meaningful to consider several structures of the loop as equally possible? Those and similar problems, in our view, may be avoided by considering ab initio prediction of different loops in proteins by separate procedures.

This paradigm seems even more reasonable, viewing the problem of ab initio modeling of loops from two different methodological approaches. On the one hand, loop modeling is part of the general problem of ab initio protein folding. Indeed, one plausible scenarios to address the problem of protein folding would be to predict sequence fragments with regular 3D structure, as α-helices and β-strands, pack them together, and then restore the "unstructured" fragments, which often are loops. On the other hand, each individual loop, even in large proteins, is a peptide of the size up to 60–70 residues (as the 70-membered V1/V2 loop in the glycoprotein gp120 (more than 490 residues) in the HIV envelope[1] that retains significant inherent mobility within the framework of the remaining part of the 3D structure of a given protein. Therefore, ab initio loop modeling should, in our view, combine elements of computational procedures developed both for protein folding and for peptide conformational analysis. This study provides examples of applications of the procedures, which we have developed in the past several years to model protein loops varying from 4 to 60 residues.

METHODS

Generating Conformations of Small Loops

Since the small loops we have considered in this study were those connecting two TM helices of bacteriorhodopsin (BR, see below), each loop was represented by the loop itself and by the two flanking N- and C-terminal helical fragments of five residues each. The backbone dihedral angles of the flanking fragments were "frozen" in the values corresponding to the x-ray structure of BR.[16] All possible combinations of local minima of *E*, *F*, *C*, *A*, and *A** types (according to the Zimmerman's notation[17]) for the peptide backbone of each amino acid residue in the loop were considered (minima of *F*, *C*, and *A* types were considered for Pro residues, and of *E*, *F*, *C*, *A*, *A**, *C**, *F**, and *E** types for Gly residues). Rigid valence geometry with the planar *trans*-peptide bonds was assumed. Several filters were used to eliminate conformers from further considerations. First, the backbone structures were constrained by satisfying the requirements $D_{ij} = D^0_{ij} \pm 2$ Å, where D_{ij} are distances between the ith and jth C^α atoms of the N- and C-terminal flanking fragments, and D^0_{ij} are these distances in the x-ray structure. Then, the selected backbone structures were subjected to energy minimization employing the ECEPP/2 force field[18,19]; all dihedral angles in the loop, including the ω angles of the Pro residues, were allowed to rotate. The total energy included also the sum of parabolic potentials $E_{ij} = E_0 (D_{ij} - D^0_{ij})^2$, where $E_0 = 1000$ kcal/mol. The low-energy backbone structures ($\Delta E - E_{\min} < 15$ kcal/mol) were selected. Finally, only structures differing by more than 40° in at least one value of any backbone dihedral angle were selected among low-energy conformers.

Buildup Procedure for Medium Loops

The procedure of a stepwise elongation of peptide backbone to build a loop has been applied for the 18-residue outside loop BR 62–79 (L^{62}GYGLTMVPFGGEQNPIY^{79}) connecting BR2 and BR3 helices. The procedure started from restoring the x-ray structure of the TM helical bundle with the "averaged" conformers representing the entire sets of low-energy conformers of the small outside loops BR 127–133 and BR 190–201 (see the Small Loops subsection) to create a united "framework" for the further loop building. The backbone of the first stem residue of the loop, Leu^{61}, was overlapped with the corresponding residue of the framework. Then, the buildup procedure considered all possible backbone conformations for the newly added residue (one residue at a step, as opposed to random generation of the loop conformers within the existing 3D structure of the protein[2]). The conformations were selected from the set of the preferred conformers for amino acid residues found in proteins,[20] and from the local minima of the Ramachandran map,[17] i.e., of the following ϕ,ψ

points: −140°, 140°; −75°, 140°; −87°, −4°; −65°, −42°; and 77°, 20°. There was an additional ϕ,ψ point for Gly: 107°, −174°; for Pro, the ϕ,ψ points were −75°, 140°; −75°, −4°; and −75°, −42°. At each elongation step, the system of distance limitations was imposed on the growing peptide chain. First, the growing chain was required to be self-avoided, as well as to avoid sterical clashes with residues comprising the framework (i.e., the corresponding C^{α}–C^{α} distances should not be less than 5 Å). Second, the chain could not reach "the point of no return," i.e., the distance between the last C^{α} atom of the growing chain and the C^{α} atom of the "target" residue Trp^{80} should not be larger than $3N$ Å, where N is the number of peptide groups between the last residue of the growing chain and Trp^{80}. At the first step of the buildup procedure, all combinations of local minima for the peptide backbone of the BR 62–72 fragment were considered. Then, in nine steps, the entire set of geometrically possible conformers of the loop backbone was constructed (see also Ref. 21). Energy calculations were then performed for all backbone conformers in the same way as for the small loops (i.e., with the 5-residue flanking helical fragments).

Restoring Loops by Residue–Residue Contact Matrices

General Considerations. This approach is based on residue-residue contact (0,1)-matrices describing the system of contacts in a protein.[22] (A contact is defined as the C^{α}–C^{α} distance less than 8 Å between a pair of residues in the 3D structure of the protein.) Our general procedure for ab initio prediction of the residue–residue contact matrices is as follows. The starting contact matrices are obtained from the protein sequence based on prediction of the residue coordination numbers (number of contacts) for each amino acid residue in the sequence, and on prediction of segments of regular structure (α-helices and β-strands) by a consensus of readily available statistical prediction methods (see, e.g., the list in the URL address[23]). Then, an iterative procedure refines the starting matrices according to (a) the values of known probabilities of contacts between residues with given coordination numbers, (b) the value of the average density of packing for any 3D structure that could be restored from the obtained matrix, and (c) the requirement of geometrical self-consistence for resulting 3D structures. This procedure was employed in this study for restoring the medium loops. For restoring the large loops, the approach additionally exploits certain specific properties of the first few eigenvectors $\mathbf{Y^i}$ of residue–residue contact matrices associated with the largest eigenvalues $\lambda^{\mathbf{i}}$.

Generally, to predict 3D structures of loops in proteins, we divide the matrices into two distinct parts, the invariable part corresponding to the protein itself without loops (this part is known), and the variable part corresponding to loops only (unknown, to be determined). More exactly, we consider three types of residue–residue matrices for each given protein. The first matrix A_c (the "core" or "constant" part) includes the submatrix of the known contacts in the protein without loops, as well as the known contacts between the loop residues and the residues belonging to the invariable part of the protein (for instance, the contacts between positions $a_{i,i+1}$, $a_{i,i+2}$, the standard contacts inside α-helices, disulfide bonds, etc.). The second one A_n is one of "noncontacts." It includes the known "noncontact" positions, or known absence of contacts (i.e., matrix elements corresponding to the C^{α}–C^{α} distances greater than 8 Å) in both the invariable and variable parts of the protein (e.g., the absence of contacts between the ends of longer elements of α-helices or β-strands). The third matrix A_x corresponds to unknown contacts that need to be predicted within each loop, between the loops, and between the loops and the invariable part of the protein. Elements of each of the three matrices are exemplified in Figure 1.

Vector of Coordination Numbers. The algorithm for prediction of the vector of coordination numbers starting with the amino acid sequence and the predicted segments of the regular structure has been described earlier.[24] To describe briefly: the coordination number n_i for each type of amino acid residues is considered a sum of the average value for this type of residue $\langle n_i \rangle$ and some positional increment Δn_i. The average values depend also on the type of the regular (or irregular) structure that contains the residue in question. The same is true for the positional increment values, which depend also on the coordination numbers of the neighboring residues in the amino acid sequence. The values of the average coordination numbers for different types of residues in different regular/irregular structures, as well as the values of coefficients needed to calculate various types of positional increments, have been obtained by processing of high resolution x-ray data of proteins from the PDB (total number of 65 nonhomologous proteins ranging from 52 to 450 residues) and reported previously.[24]

When regular segments are known (in our case, they are predicted by the consensus of several statistical methods), the average values of coordination numbers contributing to the vector of the coordination

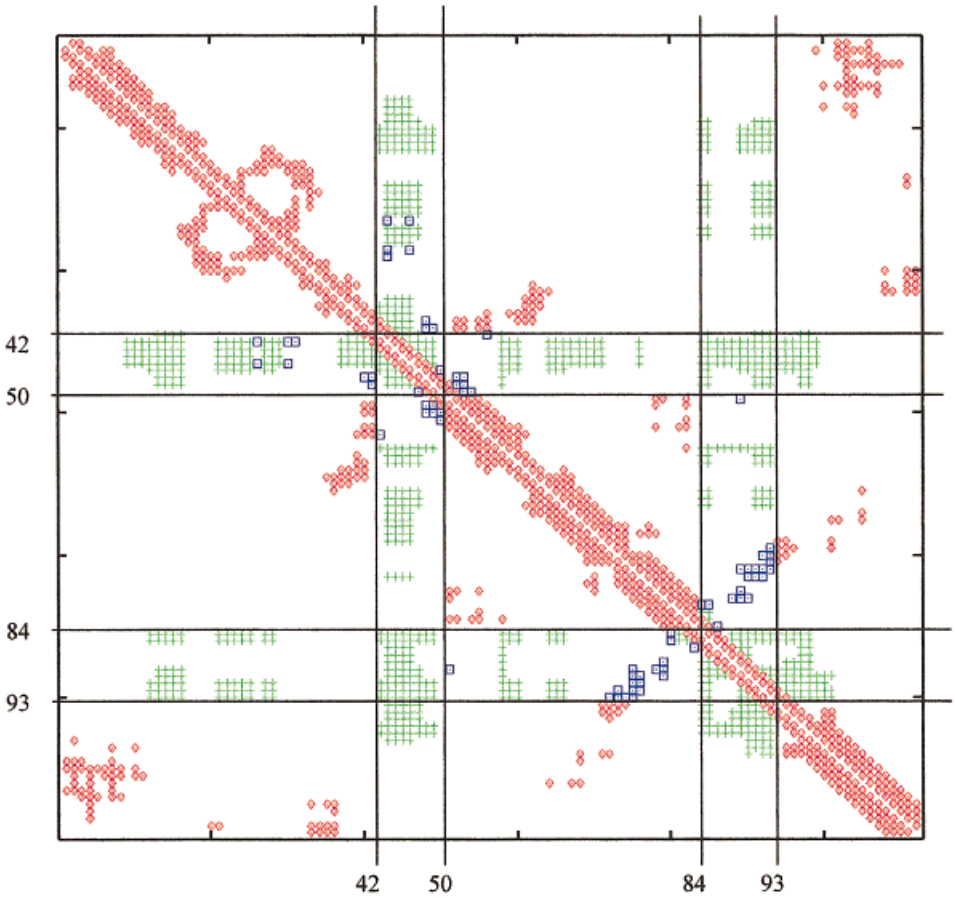

FIGURE 1 Residue–residue contact matrix for predicted 3D structure of 3c2c (blue and green lines in Figure 4b). The constant part A_c is shown in red, the "noncontact" matrix A_n, is shown in green, and predicted variable contacts A_x, are shown in blue. Numbers correspond to the predicted loops.

numbers **N** can be taken directly from those previously reported.[24] The vector of the positional increments **ΔN** can be calculated as a solution of the following linear system:

$$(E - B)\Delta\mathbf{N} = \mathbf{G}$$

where E is a unity matrix, and the values of the coefficients forming the matrix B and the vector **G** also can be taken from the previous work.[24]

For loop prediction, this algorithm was modified by adding the requirement that the total number of contacts predicted for a given residue in the contact matrix consisting of the constant and variable parts should not be smaller than the number of contacts for the same residue in the constant part alone n_{ci}. This was achieved by resolving the following minimization problem:

$$[(E - B)\mathbf{\Delta N} - \mathrm{G}]^2 \rightarrow \min_{\mathbf{\Delta N}}$$

under the conditions $\Delta n_i > (n_{ci} - \langle n_i \rangle)$, where the i index relates to all residues not belonging to loops. This requirement constrains possible solutions for the vector of the coordination numbers to physically reasonable values. In another slight modification of the algorithm, the elements of the vectors **N** were smoothed using a window-like technique (the 5-residue windows) for predictions of the large loops.

Prediction of Matrices. As a first approximation, the contact matrix $A^0{}_x$ was built on the basic of the probability matrix P each element of which, p_{ij}, is the product of probabilities of the contact between residues i and j, which are q_i and q_j, respectively. In turn, $q_i = (n_i - c_i)/v_i$, where n_i is the coordination number of the ith residue, c_i is the number of the known contacts in the ith row of the entire matrix, and v_i is the number of "free vacancies" in the ith row, i.e., the number of residues in a given protein N minus the number of all known contacts and "noncontacts" in

the *i*th row. Accordingly, $v_i = N - c_i - b_i$, where b_i is the number of "noncontacts" that are already present in the *i*th row.

At the first step, the noncontact areas were inserted in the matrices A^0_n. It was observed that the matrix elements for which the p_{ij} values ranks in the lower 20–30% often correspond to actual noncontacts in experimentally determined structures (more than 20 nonhomologous proteins from the PDB ranging from 46 to 123 residues). This observation allows creation of the noncontact areas in the matrices A^0_n; in this study, we have used a threshold of 29%. Similarly, the matrix elements with the highest p_{ij} values may be used for creating the "contact" areas in the matrices A^0_x. In this case, however, we know the exact expected number of contacts in A_x, which is $M_x = \mathbf{IN} - \mathbf{I}A_c\mathbf{I}$, where **I** is a unity vector. Accordingly, we can insert M_x contacts into the matrix A^0_x. (In our experience, unlike prediction of the noncontacts, this procedure predicts correctly only about 50% contacts; obviously, it requires further refinement.)

The second approximation of the matrix A_x was obtained differently for the medium and large loops. For *medium loops*, we used the routine outlined earlier.[22] Namely, we minimize the following penalty function:

$$F = \sum a_{ij} * (1 - p_{ij}) * f1_{ij} * f2_{ij},$$

where summation is over all contacts in the matrix $A = A_c + A_x$. The minimization is performed under condition $a_{ij} < a^2_{ij}$, where a^2_{ij} are elements of the squared matrix A^2. (This condition provides spatial consistence of the 3D structure, as shown earlier.[25]) The normalized squared deviation $f1$ is designed to preserve the observed dependence between the elements of the matrices A^2 and A^3 as follows:

$$f1_{ij} = (\alpha a^2_{ij} - a^3_{ij} + \beta)^2 * [1/(f1^{max}_{ij} - f1^{min}_{ij})],$$

where $\alpha = 2.1$, $\beta = 11.6$. This provides a constraint on the protein density. The normalized squared deviation $f2$ also preserves the dependence observed between the following function of the pair of coordination numbers:

$$\phi(n_i, n_j) = \gamma(n_i - 1)(n_j - 1)/(n_i + n_j - 2) - \omega,$$

where $\gamma = 2.91$ and $\omega = -7.33$, and a^2_{ij} as follows:

$$f2_{ij} = [\phi(n_i, n_j) - a^2_{ij}]^2 * [1/(f2^{max}_{ij} - f2^{min}_{ij})].$$

Both dependencies have been observed earlier [22] analyzing the x-ray protein structures from the PDB (65 nonhomologous proteins, all smaller than 130 residues); the numerical values of the coefficients were also calculated previously.[22]

The function F has been minimized by redistributing the M_x contacts in the matrix A^0_x in the following way. First, 30% of the contacts that correspond to the elements of penalty matrix F_{ij} with the largest values were removed. Then, the 30% of the contacts were distributed over all vacant positions in the matrix A^0_x, which correspond to the smallest elements of penalty matrix F_{ij}. Then, the excessive contacts in those rows and columns of the matrix A_x, where the number of contacts exceeded the predicted coordination numbers were removed; the contacts corresponding to the positions with the largest values of F_{ij} were removed first. The removed contacts were again redistributed over all vacant positions in the matrix A^0_x except the positions from which they have been removed; this cycle continued until less than 5% of all contacts in the matrix A^0_x were allowed to move. Then, the entire procedure was applied repeatedly to the obtained matrices A^i until the difference between A^{i+1} and A^i was less than 5%, or the difference in the successive values of the penalty function, $\Delta F = F^i - F^{i+1}$, was less than 0.01 F^i.

For *large* loops, we obtained the first five terms of the following eigenvector decomposition for the matrix $A = A_c + A^0_x$ as the sum of the direct products of the eigenvectors:

$$\Theta = \sum_{i=1}^{5} \lambda^i \mathbf{Y}^i \mathrm{o} \mathbf{Y}^i$$

The resulting matrix Θ is an approximation of the matrix A, and at the same time, can be regarded as an approximation of the probability matrix P. Therefore, we can use the matrix Θ for construction of the matrices A^0_x and A_n, as we used the matrix P, i.e., to select the highest and lowest values of Θ_{ij} for insertion of the contacts and noncontacts into the matrices A^0_x and A_n, respectively. Note that the important property $\theta_{ij} < \theta^2_{ij}$, analogous to the property $a_{ij} < a^2_{ij}$ shown for the contacts,[25] is fulfilled for the highest values of Θ_{ij} automatically, since the elements of the first eigenvector $\mathbf{Y^1}$ are always positive, and the largest eigenvalue λ^1 is also positive (see the Peron theorem[26]). The obtained matrix $A = A_n + A^1_x$ was again decomposed as above, and the entire cycle was repeated until convergence is achieved, i.e., until the difference between A^{i+1} and A^i was less than 5%. In

Table I Regression Coefficients and Standard Deviations for Expected Distances $\langle d_{ij} \rangle$ Depending on *k*

k	α_k	β_k	$\sigma^k\langle d \rangle$, Å
1[a]	−0.33	7.75	0.85
2	−0.92	12.72	1.16
3	−0.27	16.49	1.94
4	−0.12	20.95	2.52
5	−0.05	25.56	3.02
6	−0.02	29.96	3.4

[a] For the pairs in contact, the regression of $\langle d_{ij} \rangle$ on the elements of squared contact matrix was calculated.

fact, one cycle of this procedure was sufficient for the large loops in question.

Restoring C^α Traces. The dependence of the expected distance between a given pair of residues $\langle d_{ij} \rangle$ on the elements of powers of the contact matrices has been shown earlier.[22] The residue–residue contacts form a network, so one can find several pathways of the shortest length k between each pair of residues. We have observed that the expected distances depend on the values of the elements of kth power of the contact matrix a_{ij}^k at positions where $a^{k-1}{}_{ij} = 0$ and $a_{ij}^k > 0$, as follows:

$$\langle d_{ij} \rangle = \alpha_k a^k \mathrm{ij} + \beta_k,$$

where the regression coefficients α_k and β_k were calculated for the set of the proteins with the known x-ray structures (33 proteins of the size of less than 130 residues). Table I contains the values of the coefficients as well as the values of the standard deviations for the residue–residue distances $\sigma^k\langle d \rangle$. The dependence in question became less pronounced with the increase of the k value. For k = 7 and k = 8 it is not observed anymore; the average distances are 31.0 ±5.4 Å and 32.6 ± 6.5 Å, respectively. On the contrary, the $\sigma\langle d \rangle$ values increase significantly with an increase of k.

One more observation on the same set of proteins relates to the average residue–residue intraglobular distance $\langle d_p \rangle$. It correlates tightly with the cubic root of the number of residues in the protein *N*. The corresponding equation is as follows:

$$\langle d_p \rangle = 4.65N^{1/3} - 4.42,$$

and describes the dependence with good accuracy (the standard deviation $\sigma\langle d_p \rangle = 0.37$ Å).

In the actual protocol, the expected residue–residue distances $\langle d_{ij} \rangle$ were obtained using the above regression equation with coefficients corresponding to the lowest value of k, which met the requirements $a_{ij}^{k-1} = 0$ and $a_{ij}^k < 0$. The distances $\langle d_{ij} \rangle$ were then corrected proportionally to the corresponding standard deviations σ_{ij}^k to satisfy the expected value for the average residue–residue intraglobular distance $\langle d_p \rangle$. The initial C^α trace for the entire protein was then restored employing the standard distance geometry algorithm.[27] After that, the restored C^α trace for the constant part of the protein was replaced with the same C^α trace taken from the x-ray structure (the best fit for all C^α atom coordinates); all subsequent refinement procedures involve only the variable part of the protein.

Refinement of the initial C^α trace began with the insertion of the standard segments of regular structures, i.e., α-helices and β-strands, in their proper places. This was achieved by least square fitting of the ends of the segment and its geometrical center. Then, the refinement procedure was performed for correction of the chosen residue–residue distances and for removal of the possible sterical clashes. For this objective, the penalty function $F(\boldsymbol{R})$ consisting of two relevant terms, was minimized with respect to the set of the C^α coordinates of the loop $\boldsymbol{R}$. The function was as follows:

$$F(\boldsymbol{R}) = \sum w_{ij}(\langle d_{ij} \rangle - |\mathbf{R}_i - \mathbf{R}_j|)^2 + A/(\mathbf{R}_i - \mathbf{R}_j)^6$$

where $w_{ij} = 10.0$ for the elements with $i - j = 1$, $i - j = 2$ and for the distances inside the elements of regular structure, and $w_{ij} = 1/\sigma_{ij}^k$ elsewhere; $A = 500{,}000$. For the next-to-neighboring elements ($i, j = i + 2$), the values of 5.4 or 6.6 Å, whichever was closer to the $\langle d_{ij} \rangle$ value, have been inserted instead of the $\langle d_{ij} \rangle$ values, in accordance with findings previously reported.[28] This substitution allows avoiding unrealistic values of the angles C_i^α—C_{i-1}^α–C_{i+2}^α. The summation included the residue–residue distances within the loops, as well as the residue–residue distances between the loop and the constant part of the protein. The conjugate gradient procedure used for minimization was based on a routine previously described.[29]

An important novel element in refinement of the C^α traces in this study was the systematic variation of orientation and "configuration" of the loop segments. In many loop structures, one can observe the quasiautonomous fragments possessing contacts with the rest of the protein only in the narrow area along the line connecting the beginning and the end of the fragment.

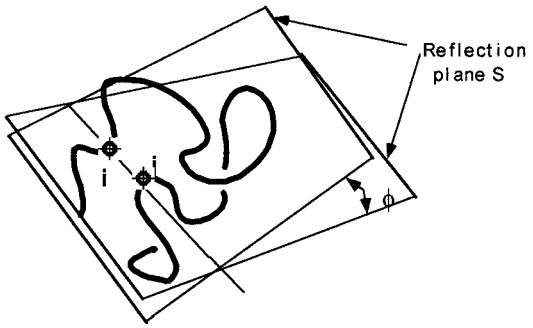

FIGURE 2 Inversion of the configuration of the fragment *i–j*.

In the mirror images of such fragments, the set of contacts/distances between the fragment and the rest of the globule may change only insignificantly; the contacts within the fragment will not change at all. However, the overall structure of the loop comprising this fragment and its relation to the fixed part of the protein may change dramatically.

To examine all of such possible "diastereomers," we developed a procedure that systematically changes orientation and "configuration" of the fragments within a given loop. Namely, for a given original structure of the loop, the mirror image of the each fragment within the loop from the *i*th to the *j*th residues ($i - j > 6$) was calculated using the reflection plane S running through the first and the last residues of the fragment as well as through its geometrical center (see Figure 2). Then, the values of the function $F(\boldsymbol{R})$ corresponding to the various values of the angle ϕ were calculated, the angle ϕ determining rotation of the reflection plane around the vector $\mathbf{R}_i - \mathbf{R}_j$. The minimal value $F(\boldsymbol{R}_{\min})$ was compared with the original value $F(\boldsymbol{R}_{\text{old}})$ to calculate the difference $\Delta F = F(\boldsymbol{R}_{\min}) - F(\boldsymbol{R}_{\text{old}})$. When the procedure has shown that orientation of some "mirrored" fragment can be changed without significant worsening of the function $F(\boldsymbol{R})$, the corresponding new C^{α} trace was submitted to a new cycle of refinement, i.e., a new minimization of $F(\boldsymbol{R})$ with respect to all atomic coordinates is performed. Sometimes the new minimization yielded the new C^{α} trace that possessed the $F(\boldsymbol{R})$ value comparable to that of the original one (difference less than 10%). In these cases, both (or more) C^{α} traces originated from the same contact matrix were regarded as plausible results.

The above procedure for restoring the C^{α} traces from distance matrices has been applied for all medium and large loops. In all cases, except those for 1crn 4–32, 1alc 61–91, 2lzt 64–94, and 1sn3 16–48, the C^{α} traces have been restored from the distance matrices for the entire protein; in the listed cases, only the parts of the distance matrices that correspond to the actual loops were used.

Calculating RMSD Values

To evaluate the quality of our predictions, we have used two different RMSD criteria. Besides the values obtained by the routine procedure overlapping all residues in the isolated predicted loop with the experimental one (referred below as a "loop-to-loop" RMSD), we have calculated also the RMSD differences between the residues of the both loops being fixed in the coordinate system connected with the constant part of the protein (a "loop-in-the-structure" RMSD value). The latter values reflect not only similarity in the internal structures of the both loops, but also similarity in their orientations with respect to the 3D structure of the entire protein (see also definitions of the "local" and "global" RMSD values[2]). In all cases, the RMSD values have been calculated for the coordinates of the C^{α} atoms only.

RESULTS AND DISCUSSION

Small Loops (From 4 to 12 Residues)

We have applied a straightforward procedure for generating low-energy conformations of small loops (see the Methods section) to modeling the interhelical loops in the x-ray structure of bacteriorhodopsin.[16] Summary of the results is given in Table II (see also Ref. 21). For instance, for the 7-residue loop 31–37 in BR (the $G^{31}MGVSDP^{37}$ loop between helices BR1 and BR2), 1594 conformers of the peptide backbone were found to satisfy the distance constraints $D_{ij} = D^0_{ij} \pm 2$ Å. Out of these, 110 low-energy conformers ($\Delta E < 15$ kcal/mol) were obtained. The range of the "loop-to-loop" RMSD values (C^{α} atoms) relative to the x-ray structure for these conformers was 0.87–2.24 Å, quite comparable to the best results of other authors cited above. (Note that the RMSD values for the same loops described in Ref. 21 included 10 flanking residues as well.) At the same time, the "averaged" low-energy conformer for this loop (the conformer obtained by averaging the spatial position of each C^{α} atom of the loop over all 110 low-energy conformers) possesses a RMSD value of 1.04 Å, whereas other low-energy conformers differ from the "averaged" one by a RMSD range of 0.89–2.60 Å. Figure 3 illustrates the spatial difference among the low-energy conformers relative to the "averaged" one; it is evident that the "averaged" conformer fairly

Table II Restored Small Loops in BR

Loop Between Helices	Loop Involves Residues	Size, Residues	Number of Low-Energy Conformers	"Loop-to-Loop" RMSD Values, Å		
				a	*b*	*c*
BR3–BR4	101–104	4	3	1.01	0.54–0.85	0.65–1.01
BR4–BR5	127–133	7	66	2.11	1.20–3.36	1.49–3.15
BR1–BR2	31–37	7	110	1.04	0.89–2.60	0.87–2.24
BR5–BR6	158–165	8	147	1.94	1.00–3.57	1.23–3.22
BR6–BR7	190–201	12	131	4.26	1.67–7.85	2.90–5.74

[a] For the average conformer compared to the x-ray structure.
[b] For all low-energy conformers compared to the average conformer.
[c] For all low-energy conformers compared to the x-ray structure.

represents the entire set of low-energy conformers, at least at the level of their C^{α} traces.

The results listed in Table II include those for the smallest loop (BR 101–104, 4 residues), and for the largest one (BR 190–201, 12 residues). One can see that the "averaged" conformer indeed represents the available set of the loop conformers for loops up to 8 residues. However, geometrical variations between low-energy conformers of the 12-residue loop are significantly larger than those for the 8-residue one. Also, estimations of conformational energies for the larger loops obtained in the same way as for the small loops may be misleading, since they do not consider possible limitations imposed on the loops by the rest of the protein, which become more significant with increasing mobility of the larger loops.

Medium Loops (From 9 to 33 Residues)

Table III contains a general summary of the results obtained for medium loops in ten proteins. In all cases, we used the residue-residue contact matrices divided into the constant "core" part, and the variable parts corresponding to the loops, as described above in the Methods section. The loops listed in Table III have been selected for calculations primarily by two reasons: (a) they belong to relatively small proteins, from the 46-residue crambin, 1crn, to the 129-residue lyzosyme, 2lzt, and (b) conformational mobility for almost all of them is limited by various constraints. For instance, the end residues of the loop 1crn 4–32, Cys^{4} and Cys^{32}, are connected by disulfide bridge; the loop contains also one more disulfide bridge, Cys^{16}–Cys^{26}. The loop 1sn3 16–48 contains three disulfide bridges, namely Cys^{16}–Cys^{41}, Cys^{25}–Cys^{46}, and Cys^{29}–Cys^{48}. Two disulfide bridges are in each of the loops 1alc 61–91 and 2lzt 64–94, which are Cys^{61}–Cys^{77} and Cys^{73}–Cys^{91}, and Cys^{64}–Cys^{80} and Cys^{76}–Cys^{94}, respectively. Two loops in 1bp2, 13–40 and 107–123 (the latter is, actually, the C-terminal tail of the protein, which is why it was not included in Table III), are connected by the disulfide bridge Cys^{27}–

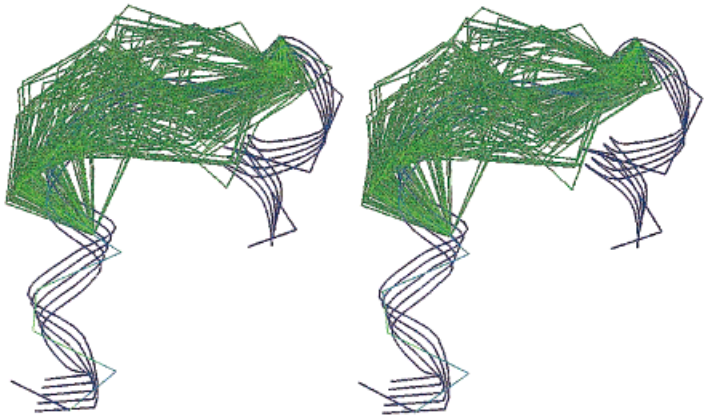

FIGURE 3 Stereoview of all predicted structures of the interhelical loop BR 31–37 (green) and the averaged structure (blue, shown as ribbon).

Table III C^{α} Traces for Medium Loops Obtained from Residue–Residue Contact Matrices

Loop: PDB Entry, End Residues	Size, Residues	No. C^{α} Traces	"Loop-to-Loop" RMSD, Å	"Loop-in-the-Structure" RMSD, Å
3icb 55–63	9	1	3.3	4.4
3c2c 42–50	9	2	2.3; 2.6	2.6; 5.7
3c2c 84–93	10	2	2.3; 2.8	8.3; 6.3
3icb 16–25	10	2	3.1; 3.3	5.0; 4.3
351c 16–26	11	2	3.3; 3.9	4.3; 4.9
4cpv 89–102	14	1	3.5	5.4
4cpv 50–64	15	2	4.0; 4.2	4.5; 4.4
351c 50–67	18	3	3.6; 4.1; 4.3	4.7; 5.0; 4.9
1bp2 13–40	28	2	3.5; 3.8	4.0; 6.1
1crn 4–32	29	1	2.4	n/a[a]
1alc 61–91	31	2	3.9; 6.1	n/a
2lzt 64–94	31	1	5.2	n/a
1sn3 16–48	33	3	5.9; 4.9; 5.3	n/a

[a] As was mentioned, the C^{α} traces for these loops have been restored using the distance matrices for the loops only, which precluded calculation of the "loop-in-the-structure" RMSD values.

Cys^{123}; besides, some residues in the loop 1bp2 13–40 interact with a calcium ion. Each of the two loops of 4cpv, 50–64 and 89–102, interacts with a calcium ion, so the two calcium ions may contribute to stabilization of the loop structures. The same is true for loops 3icb 16–25 and 55–63. The loop 351c 50–67 (but not 16–26) may interact with the porphyrin moiety in this cytochrome protein; in another cytochrome, 3c2c, the porphyrin moiety does not interact with any of selected loops.

The above constraints were used to limit the predicted residue–residue contact matrices by insertion the invariable contacts corresponding to the disulfide bridges in the otherwise variable parts of the matrices. Interactions with the calcium ions as well as with the porphyrin moieties were disregarded. Another limitation was to insert into the contact matrices the invariable contacts, which correspond to segments of α-helices and β-strands in the loops, predicted by a consensus of available statistical methods. Such predicted α-helical segments were 4cpv 98–102; 1crn 6–16; 1alc 85–91; 1sn3 22–30; and 2lzt 90–94. The predicted β-strand fragments were 1alc 61–64, 68–71, 73–76, and 78–92 (the latter in one of the two matrices); 1sn3 37–42; and 2lzt 70–75, 77–80, and 82–86.

The selected loops represent a wide variety of conformational elements in proteins. The two loops in the calcium-binding proteins, 4cpv and 3icb, are interacting with each other (see Figure 4a), whereas the loops 3c2c 42–50 and 84–93 do not interact, since they are separated by helix 51–59 (Figure 4b). Also, the relatively long loop, 1bp2 13–40, that connects two helical fragments interacts with the C-terminal tail 107–123. (The results obtained for the 3icb, 3c2c, 4cpv, 351c, and 1bp2 proteins, where two loops were restored simultaneously, have been briefly discussed earlier.[30]) The loop 1crn 4–32 contains a α-hairpin-like structure, involving 29 out of 46 residues of crambin (Figure 5a). A more complex hairpin-like structure, where the one leg contains the α-helical fragment, and the other leg is a somewhat extended structure, is represented by the loop 1sn3 16–48 (Figure 5b). Both 1alc 61–91 and 2lzt 64–94 loops are of the similar loose "double ring" structures (Figure 5c).

Several conclusions can be drawn for the results of predictions listed in Table III. First, it is interesting to note that the obtained "loop-to-loop" RMSD values do not really depend on the size of the loop. Indeed, the best RMSD values for predicted structures range from 2.3 Å for the 9-residue loop (3c2c 42–50) to 4.9 Å for the 33-residue loop (1sn3 16–48) including a RMSD = 4.0 Å for the 15-residue loop (4cpv 50–64), and a RMSD = 2.4 Å for the 29-residue loop (1crn 4–32). It suggests that the procedure of restoring the variable parts of the contact matrices is not really sensitive to dimension of the variable part, at least in the case of these selected loops. One can see that the results for the small loops (up to 10–12 residues) obtained by the straightforward generation of loop conformers followed by energy minimization are significantly better (Table I). However, restoring C^{α} traces to the level of peptide backbones in all-atomic resolution by the Monte Carlo with Minimization algorithm (MCM[31]), with subsequent energy calculations, which has been performed for the 9- and 10-

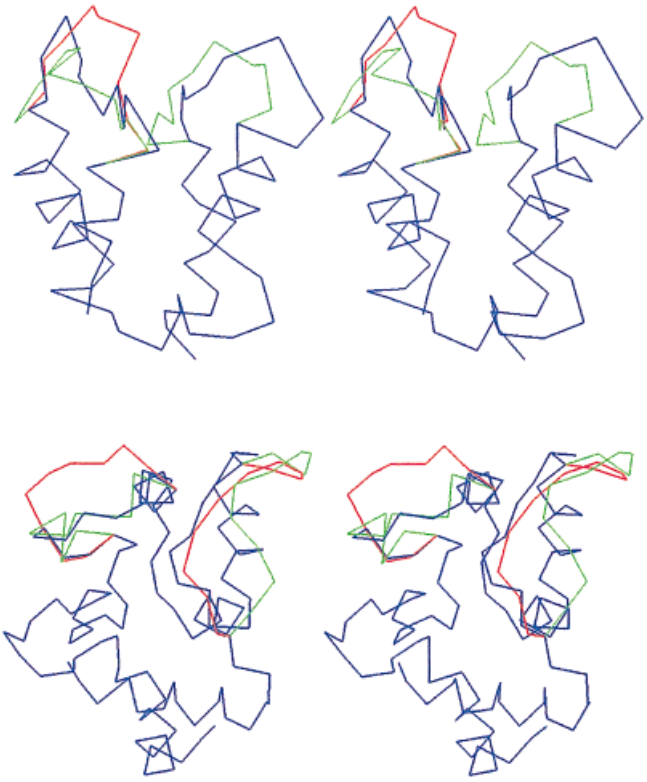

FIGURE 4 (a) Stereoview of the x-ray structure of 3icb (blue) overlapped with predicted loops (red and green). The loop 16–25 is at the left, and the loop 55–63 is at the right. (b) Stereoview of the x-ray structure of 3c2c (blue) overlapped with predicted loops (red and green). The loop 42–50 is at the left, and the loop 84–93 is at the right.

residue loops 3icb 55–63 and 16–25 earlier,[30] changed the resulting RMSD values for 3icb 55–63 from 3.3 to 1.7–5.0 Å (22 low-energy conformers were found), and for 3icb 16–25 from 3.1–3.3 to 1.6–4.2 Å (three low-energy conformers were found). These results are quite comparable to those in Table I for the 8- and 12-residue loops BR 158–165 and BR 190–201 producing, at the same time, lesser numbers of possible backbone conformers. The same trend has been observed in the results obtained by the same calculations for the 14-and 15-residue loops 4cpv 89–102 and 50–64, where the resulting RMSD values changed for 4cpv 89–102 from 3.5 to 1.6–3.4 Å (three low-energy conformers were found), and for 4cpv 50–64 from 4.0 to 2.9–3.5 Å (two low-energy conformers were found).[30] Therefore, it is reasonable to expect that for small loops up to 7–8 residues, the straightforward generation of conformers is still robust and affordable from the point of view of computational resources; for larger loops, the procedure discussed in this subsection followed by restoring of C^{α} traces to all-atomic resolution is probably a better option. At least, the alternative buildup procedure, described in the details in the Methods section, when applied to the 18-residue loop BR 62–79, yielded 56 low-energy structures of the peptide backbone with the RMSD values from 3.4 to 8.9 Å (see also Ref. 21). On the other hand, we achieved only a slight improvement of the results for 1crn 4–32, 1alc 61–91, 2lzt 64–94, and 1sn3 16–48, when all-atomic resolution was restored by fitting the C^{α}–C^{α} distance matrices to the overlapping fragments of the peptide backbone undergoing energy minimization during fitting. In this case, we have obtained 3 low-energy conformers with RMSD = 2.5–4.7 Å for 1crn 4–32; 84 low-energy conformers with RMSD = 3.2–5.7 Å for 1alc 61–91, 32 low-energy conformers with RMSD = 4.9–5.9 Å for 2lzt 64–94; and 2 low-energy conformers both with RMSD = 4.6 Å for 1sn3 16–48.

The successful outcome of the discussed procedure depends mainly on the quality of prediction of contact matrices. For instance, the single matrix predicted for 2lzt 64–94 contained the wrong contact between residues 77 and 86. The contact was predicted as part of a favorable short antiparallel β-sheet involving two β-strands, 77–80, and 82–86, which, in turn, were

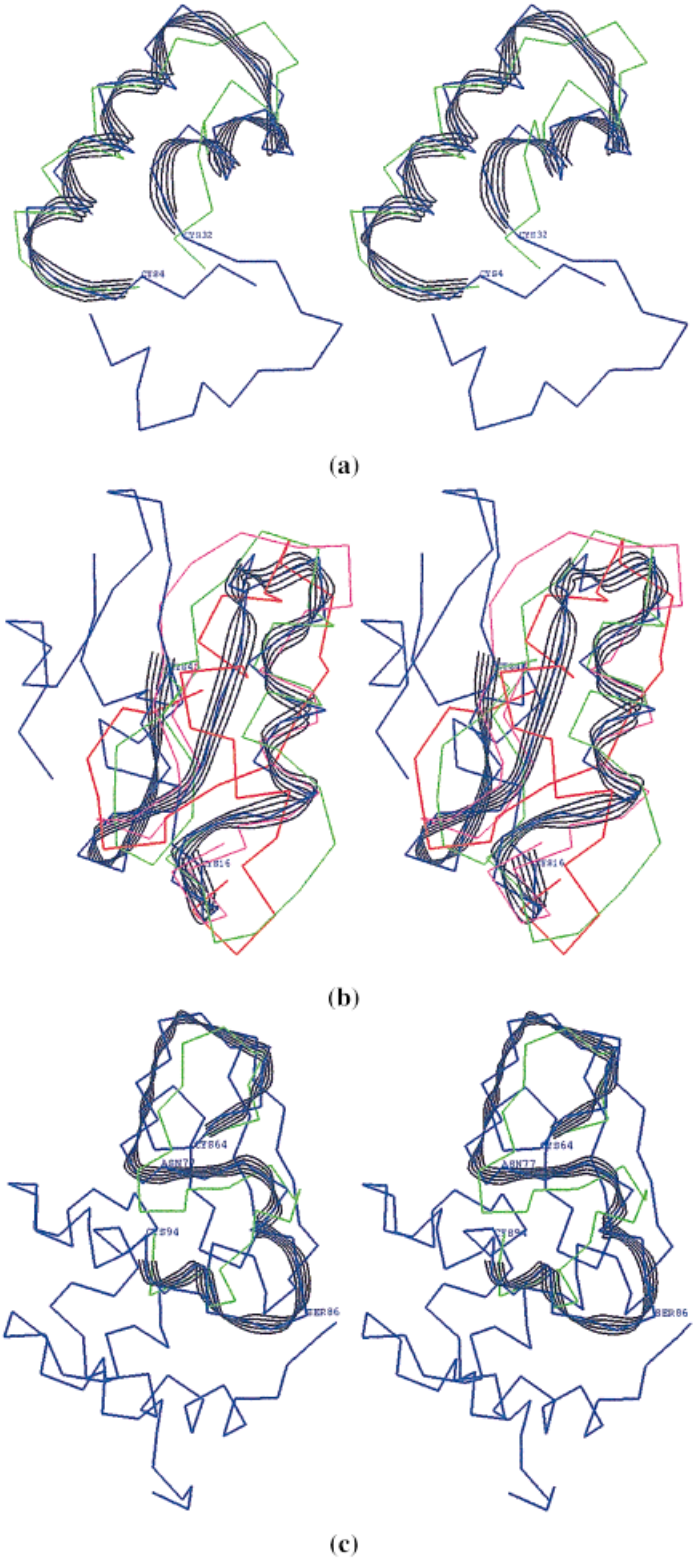

FIGURE 5 (a) Stereoview of the x-ray structure of 1crn (blue) overlapped with the predicted loop 4–32 (green). The loop in the x-ray is shown as ribbon. (b) Stereoview of the x-ray structure of 1snf (blue) overlapped with the predicted loops 16–48 (green, red, and magenta). The loop in the x-ray structure is shown as ribbon. (c) Stereoview of the x-ray structure of 2lzt (blue) overlapped with the predicted loop 64–94 (green). The loop in the x-ray structure is shown as ribbon. Residues 77 and 86, which are mentioned in the text, are labeled.

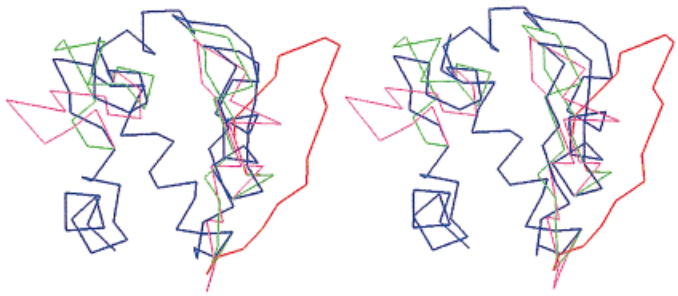

FIGURE 6 Stereoview of the x-ray structure of 351c (blue) overlapped with the predicted loops (red, green, and magenta). The loop 16–26 is at the left, and the loop 50–67 is the right. The mirror images of the fragment 16–22 (at the left, magenta), and of the fragment 53–65 (at the right, red) would be closer to the x-ray structure.

predicted by the consensus of statistical methods (see above). In the x-ray structure of 2lzt, fragment 79–85 is a α-helix, so residues 77 and 86 cannot contact each other. As a result, one part of the "double-ring" structure has been predicted very successfully (the "loop-to-loop" RMSD value at the fragment 64–81 is 2.4 Å), whereas the total RMSD value was only 5.2 Å (see Figure 5c). In a very similar situation, due to the wrong prediction of the β-strand fragments, one of the two matrices predicted for 1alc 61–91 contained the wrong contact between residues 73 and 81, which are separated by the α-helical fragment 76–81; the resulting RMSD in this case was 6.1 Å. However, each particular case of the wrong prediction is different. For instance, the wrong contact 24–32 in one of the two matrices predicted for 1bp2 13–40 resulted only in slight distortion of the corresponding C^{α} trace in the region 21–31; the RMSD value was 3.8 Å compared to 3.5 Å obtained for the C^{α} trace originated from the another matrix.

Even if the contact matrices are predicted correctly, the corresponding C^{α} traces may not reproduce the target 3D structures. As it is shown above, restoring the C^{α}–C^{α} distance matrices from the contact ones depends on the accuracy of estimation of some numerical parameters, which are taken from experimental data. More important is the problem of the mirror images within the C^{α} trace, which was discussed in details in the Methods section. It is still possible to select the wrong "mirror image" of the segment within the C^{α} trace by any distance geometry embedding procedure. For instance, one of the C^{α} traces restored for the loop 351c 16–26 could be much closer to the x-ray structure, if it is replaced by the mirror image of the fragment 16–22 (see the magenta line in Figure 6). In the same protein, the mirror image of the fragment 53–65 in the loop 50–67 also could be closer to the x-ray structure (the red line in Figure 6). One more example is depicted in Figure 4b. In this case, both C^{α} traces for the loop 3c2c 84–93 were originated from the same contact matrix; one of them would be clearly closer to the x-ray structure, if one of the fragments inside the loop will be replaced by its mirror image (the green line in Figure 4b).

The "loop-in-the-structure" RMSD values listed in Table III are usually larger than their "loop-to-loop" counterparts, reflecting the errors in determining the proper orientation of the loops. These errors are mainly due to the fact that the variations in the expected distances $\langle d_{ij} \rangle$, are larger for the more distant residues (see Table I), which are the residues of the loops relative to the most of the residues of the constant part of the protein.

Large Loops (Up to 60 Residues)

Two large loops, the 54-residue loop 1pca 328–381 and the 61-residue loop 1cd5 129–189, which were restored in this study by the residue–residue contact matrix approach are not, in fact, exactly the "loops" often defined as fragments at the surface of the proteins. These subdomains are more or less locally independent parts of the protein 3D structure connecting two α-helical segments, the fragments 311–328 and 381–402 in 1pca, and the fragments 114–129 and 189–195 in 1cd5. For instance, fragment 292–303 intertwines with the loop 328–381 in 1pca (see Figure 7), so this loop contains actually two smaller surface loops, 328–346 and 367–381, as well as the small surface segment 355–358. In 1cd5, definition of the surface loop may be attributed to the 51-residue fragment 139–189, whereas fragment 129–140 is completely buried inside the 1cd5 molecule (see Figure 8).

Our procedures yielded only one type of the C^{α} traces for the large loops in question for both proteins. They have been restored with remarkable accuracy considering their size. For the loop 1pca 328–381, the "loop-to-loop" RMSD value was 4.6 Å, whereas the

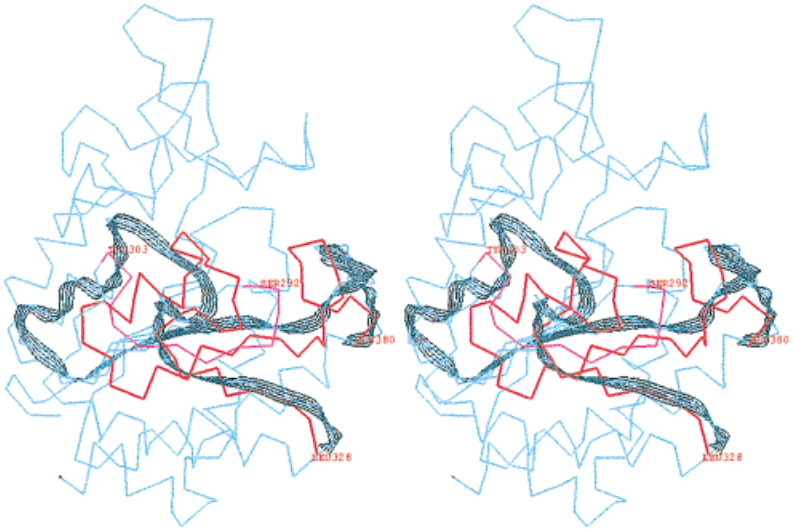

FIGURE 7 The x-ray structure of the fragment 1pca 150–402 (cyan) The loop 328–381 is shown as ribbon, and the predicted loop 328–381 is shown in red. Fragment 292–303 intertwining with the loop 328–381 is shown in magenta. The rest of the 1pca molecule does not interact with the loop 328–381.

"loop-in-the-structure" RMSD value was 6.2 Å. One can see in Figure 7 that the restored C^{α} trace correctly reproduced all main features found for the loop in the x-ray structure; i.e., the extended fragment 328–336, the "loop" 337–346, the α-helical fragment 349–355, the reversal of the peptide backbone at fragment 356–359, the extended fragment 360–367, and the loop 368–381. Note that only one of these fragments (360–366) has been incorporated into the variable part of the contact matrix for 1pca as a β-strand according to consensus of secondary structure predictions. Figure 7 shows also that the loop segments, which are more involved in interactions with the constant part of the protein molecule, were reproduced with the higher accuracy; the less accurate predictions were made for the less constrained segments 337–346 and 368–381.

Figure 8 depicts the restored loop 1cd5 129–189 on the background of the x-ray structure of 1cd5. The "loop-in-the-structure" RMSD value in this case is 8.3 Å, which is higher than for the loop 1pca 328–381, but the "loop-to-loop" RMSD value is 4.5 Å. The loop 1cd5 129–189 is less constrained by the constant part of the protein than the loop 1pca 328–381; that is one of the reasons why our procedures may predict the 3D loop structures, which are somewhat "shifted" relative

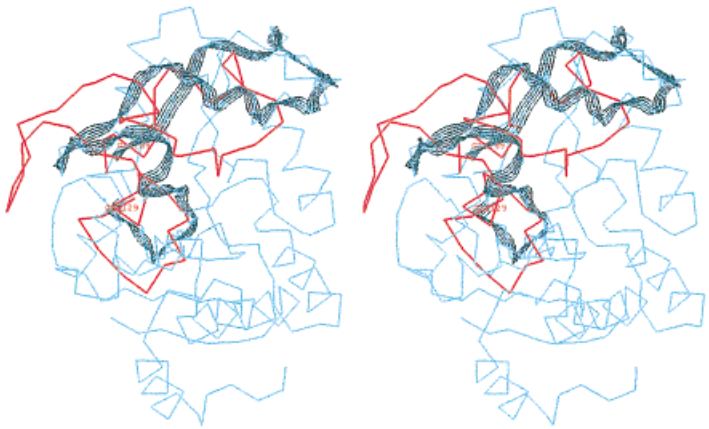

FIGURE 8 The x-ray structure of the fragment 1cd5 (cyan). The loop 129–189 is shown as ribbon, and the predicted loop 129–189 is shown in red.

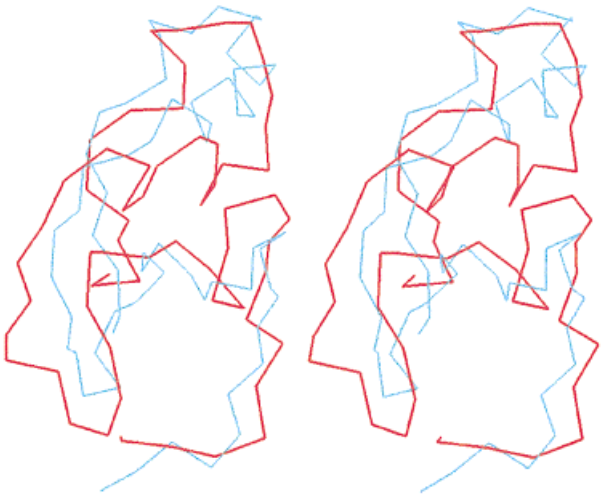

FIGURE 9 Overlapped C^{α} traces for the x-ray structure (cyan) and predicted structure (red) of the loop 1cd5 129–189. Note the α-helical region 164–171 in the upper right part of the x-ray structure.

to the correct spatial position of the loop (see Figure 8). The internal structure of the loop was, however, reproduced much more accurately (see Figure 9). Two fragments, the β-strand 131–136 and the α-helical fragment 155–160, have been incorporated into the variable part of the contact matrix according to consensus of secondary structure predictions. The former roughly corresponded to the actual β-strand 133–136, and has been preserved as such by the procedure of restoring C^{α} traces. The latter one, however, did not correspond to the actual α-helical fragment 164–171, and has been significantly distorted to accommodate the general shape of the fragment 161–185 (Figure 9).

CONCLUSIONS

This study outlines our experience in restoring 3D structures of loops in proteins by ab initio modeling. We have applied different approaches to the loops of different sizes. The small loops (up to 12 residues) have been restored by the direct generation of all reasonable conformations of the peptide backbone of the loop, the approach most widely used in modeling of peptides. On the other hand, the C^{α} traces for the medium and large loops (from 9 to 61 residues) have been restored by predictions of the residue–residue contact matrices, an approach derived from the field of protein structure prediction.

The results of the two approaches do not contradict, but rather complement each other. For instance, the accuracy of reproducing the x-ray structure of the 12-residue loop by the averaged conformer (the "peptide" approach, see Table II) is about the same as the accuracy obtained by restoring C^{α} traces for the loops of the similar size (the "protein" approach, see Table III). It is reasonable to assume that the C^{α} traces obtained by the "protein" approach may stand for the averaged conformers of the loops of larger size, too. In this case, subsequent generation of energetically possible conformations in the vicinity of the C^{α} trace by the "peptide" approach may be the way to describe the entire conformational ensemble of the loop, which is the ultimate goal of loop modeling.

Our results on restoring C^{α} traces for the medium and large loops should be regarded as quite satisfactory. Indeed, we were able to predict 3D structures of the 20–30-residue loops with the accuracy of the "loop-by-loop" RMSD of 3.0–4.0 Å, and of the 50–60-residue loops with the accuracy of RMSD = 4.5 Å. It is comparable with the best ab initio predictions at the CASP-1998 event (the continuous fragments of the size of 60–75 residues have been predicted with the RMSD values of 3.8–4.7 Å[32,33]. However, in our view, these results may be further improved by the procedure similar to that we have applied earlier to the C^{α} traces of the loops in 4cpv and 3icb,[30], i.e., by restoring the all-atomic representation of the loops including the side chains with subsequent energy minimization.

Finally, it is noteworthy that the C^{α} traces for some of the 30-residue loops (1crn 4–32, 1alc 61–91, 2lzt 64–94 and 1sn3 16–48) have been restored from the distance matrices involving the loops only, without the rest of the protein. Nevertheless, the accuracy of prediction remains basically the same as that for the C^{α} traces restored as a part of the entire protein molecule (see, e.g., the 28–residue loop 1bp2 13–40

in Table III). It confirms that those highly constrained loops may be as well regarded as the autonomous molecules, so our procedures may be applicable to small protein molecules as well (see also Ref. 34). Also, our procedures would probably suggest several possible C^{α} traces for the surface loops 1pca 328–346 and 1pca 367–381, if they were considered as individual entities, and not as parts of the united intraglobular segment 1pca 328–381. Perhaps it would lead to the more adequate description as to flexibility of the surface loops; on the other hand, since our procedures are validated for the parts of proteins containing internal fragments, they may be useful for the more general problems of protein structure prediction.

The authors wish to thank the Monsanto Company and the National Institutes of Health for grant support (EY12113, GM48184 and HL54085).

REFERENCES

1. Kwong, P. D.; Wyatt, R.; Robinson, J.; Sweet, R. W.; Sodroski, R.; Hendrickson, W. A. Nature 1998, 393, 648–659.
2. Fiser, A.; Do, R. K. G.; Sali, A. Protein Sci 2000, 9, 1753–1773.
3. Rapp, C. S.; Friesner, R. A. Proteins 1999, 35, 173–183.
4. Collura, V.; Higo, J.; Garnier, J. Protein Sci 1993, 2, 1502–1510.
5. Carlacci, L.; Englander, S. W. Biopolymers 1993, 33, 1271–1286.
6. Vasmatzis, G.; Brower, R.; DeLisi, C. Biopolymers 1994, 34, 1669–1680.
7. Zhang, H.; Lai, L.; Wang, L.; Han, Y.; Tang, Y. Biopolymers 1997, 41, 61–72.
8. Zheng, Q.; Rosenfeld, R.; Vajda, S.; Delisi, C. Protein Sci 1993, 2, 1242–1248.
9. Zheng, Q.; Rosenfeld, R.; DeLisi, C.; Kyle, D. J. Protein Sci 1994, 3, 493–506.
10. Li, W.; Liu, Z.; Lai, L. Biopolymers 1999, 49, 481–495.
11. Lessel, U.; Schomburg, D. Proteins 1999, 37, 56–64.
12. Wojcik, J.; Mornon, J.-P.; Chomilier, J. J Mol Biol 1999, 289, 1469–1490.
13. Sudarsanam, S.; DuBose, R. F.; March, C. J.; Srinivasan, S. Protein Sci 1995, 4, 1412–1420.
14. van Vlijmen, H. W. T.; Karplus, M. J Mol Biol 1997, 267, 975–1001.
15. Bates, P. A.; Sternberg, M. J. E. Proteins Suppl 1999, 3, 47–54.
16. Pebay-Peyroula, E.; Rummel, G.; Rosenbush, J. P.; Landau, E. M. Science 1997, 277, 1676–1681.
17. Zimmerman, S. S.; Scheraga, H. A. Biopolymers 1977, 16, 811–843.
18. Dunfield, L. G.; Burgess, A. W.; Scheraga, H. A. J Phys Chem 1978, 82, 2609–2616.
19. Nemethy, G.; Pottle, M. S.; Scheraga, H. A. J Phys Chem 1983, 87, 1883–1887.
20. Rootman, M. J.; Kocher, J.-P. A.; Wodak, S. J. Biochemistry 1992, 31, 10226–10238.
21. Nikiforovich, G. V.; Galaktionov, S.; Balodis, J.; Marshall, G. R. Acta Biochim Polonica 2001, 48, 53–64.
22. Galaktionov, S. G.; Marshall, G. R. In Proceedings of the 27th Hawaii International Conference on System Sciences, IEEE Computer Society: Washington–Los Alamitos–Brussels–Tokyo, 1994; pp 326–335.
23. http://www.hgmp.mrc.ac.uk/GenomeWeb/prot-2-struct.htlm.
24. Rodionov, M. A.; Galaktionov, S. G. Mol Biol 1992, 26, 777–783.
25. Galaktionov, S. G.; Rodionov, M. A. Biophysics 1980, 25, 395–403.
26. Bellman, R. Introduction to Matrix Analysis; McGraw-Hill: New York–Toronto– London, 1960.
27. Crippen, G. M.; Havel, T. F. Distance Geometry and Molecular Conformation; Research Studies Press: Taunton, England, 1988.
28. Bahar, I.; Kaplan, M.; Jernigan, R. L. Proteins 1997, 29, 292–308.
29. Press, W. H.; Flannery, B. P.; Teukolski, S. A.; Vetterling, W. T. Numerical Recipes in Cambridge University Press: Cambridge–New York–Melbourne–Sydney, 1988.
30. Galaktionov, S.; Shenderovich, M. D.; Nikiforovich, G. V.; Marshall, G. R. In Peptides 1996. Proceedings of the 24th European Peptide Symposium; Ramage, R., Epton, R., Eds.; Mayflower Scientific: Kingswinford, England, 1998; pp 399–400.
31. Li, Z.; Scheraga, H. A. Proc Natl Acad Sci USA 1987, 84, 6611–6615.
32. Simons, K. T.; Bonneau, R.; Ruczinski, I.; Baker, D. Proteins Suppl 1999, 3, 171–176.
33. Lee, J.; Liwo, A.; Ripoll, D. R.; Pillardy, J.; Scheraga, H. A. Proteins Suppl 1999, 3, 204–208.
34. Galaktionov, S. G.; Nikiforovich, G. V.; Marshall, G. R. In Peptides: Frontiers of Peptide Science. Proceedings of the Fifteenth American Peptide Symposium; Tam, J., Kayuma, P. T. P., Eds.; Kluwer Academic Publishers: Dordrecht, 1999; pp 477–478.