Determining protein loop conformation using scaling-relaxation techniques

QIANG ZHENG, RAKEFET ROSENFELD, SANDOR VAJDA, AND CHARLES DELISI

Department of Biomedical Engineering and Biomolecular Engineering Research Center, College of Engineering, Boston University, Boston, Massachusetts 02215

(RECEIVED March 8, 1993; REVISED MANUSCRIPT RECEIVED May 4, 1993)

Abstract

We recently developed a rapid loop closure algorithm in which bond lengths are scaled to constrain the ends of a segment to match a known distance and then gradually relaxed to their standard values, with boundary constraints maintained. Although the algorithm predicted the Zif286 zinc-finger loop to within approximately 2 Å, it had a serious limitation that made its more general use tentative: it omitted the atomic environment of the loop. Here we report an extension of the algorithm to take into account the protein environment surrounding a given loop from the outset of the conformational search and show that it predicts structure with an efficiency and accuracy that could not be achieved without continuous environmental inclusion. The algorithm should be widely applicable to structure determination when complete experimental information is unavailable.

Keywords: environment effect; loop closure algorithm; protein modeling; scaling and relaxation

The development of computational methods for the accurate determination of protein structure is of both fundamental and applied importance. Although the three-dimensional structure of a known amino acid sequence can in principle be determined by minimizing the system (protein plus solvent) free energy, in practice such an approach can be implemented to a useful degree of approximation in only a very limited number of cases (Skolnick & Kolinski, 1990). A more immediately useful goal is to determine structure by minimizing potential energy subject to experimental constraints. Although this is a restricted objective, a rapid and accurate procedure can substantially accelerate three-dimensional structure determinations by starting with distance constraints obtained from NMR, or from crystallographically determined coordinates of homologous proteins (Alwyn & Thirup, 1986; Braun, 1987; Greer, 1988). These and other problems are characterized by incomplete information in which the structures of segments are unknown, but information at the boundaries, and to varying degrees beyond, is available. The restricted problem then is to determine the structure of the intervening loop subject to the distance constraint (Go & Scheraga, 1970; Weiner et al., 1983; Bruccoleri & Karplus, 1985; Moult & James, 1986;

Shenkin et al., 1987; Bruccoleri et al., 1988; Summers & Karplus, 1990; Palmer & Scheraga, 1991). This distance can be determined by the positions of the two peptide bonds connecting the loop segment to the rest of the protein (Go & Scheraga, 1970) or by side-chain atoms such as in a disulfide bond or a metal-coordination site (Zheng et al., 1993).

Modeling of loop segments usually consists of two steps. In the first step, loop conformations that meet the imposed distance constraint are generated. This step is usually carried out for the isolated loop segment. In the second step the loops are evaluated, usually by minimization of an energy function. Because the number of possible conformations may (depending on the distance) increase exponentially with the number of amino acids in the loop, a reasonable search of the conformational space is very time consuming. Moreover, because most of the loops generated in the first step tend to have high energies, they will be rejected in the subsequent step, resulting in a relatively small acceptance ratio. The key to accelerating these calculations is therefore to include, from the outset of the conformational search procedure, the interaction of the loop with the surrounding atomic environment in which the loop is embedded.

The bond-scaling relaxation algorithm has been described previously (Zheng et al., 1993) and applied to isolated protein segments. Briefly, a random configuration

Reprint requests to Qiang Zheng at his present address: Scios Nova, Inc., 6200 Freeport Center, Baltimore, Maryland 21224.

is generated and subsequently scaled to meet the distance constraint. The equilibrium bond lengths in the empirical energy parameters are also scaled accordingly, and the random configuration is minimized. The equilibrium bond lengths are then gradually restored, concurrent with additional cycles of energy minimization. The number of intermediate scaling-minimization cycles depends on the size of the initial scaling factors. Here we extend this algorithm to take into account, from the outset of the conformational search, the atomic environment surrounding a given loop and compare the results to those obtained with the more common two-step procedure, in which loop generation is followed by a separate energy evaluation.

Computational method

The procedure is shown schematically in Figure 1. It begins with a randomly generated conformation. The end-to-end distance of the starting conformations is generally larger than the required distance, which for most relevant problems is 2–10 Å.¹ By scaling the initial bond lengths of the segment and the equilibrium bond lengths used for calculating bond energies, we are able to fit a larger loop into its appropriate position. Thus the distribution of energetically possible bond lengths is shifted toward smaller values, and as the equilibrium bond lengths are gradually increased, this distribution shifts back to its physical position. This solves the geometric problem of loop closure. Concomitantly, the scaled segment responds to the force field of the environment (the rest of the protein in the examples herein). The interaction energy surface has many barriers, caused to a large extent by the repulsive term of the van der Waals potential. In order to lower these barriers, this term is also scaled down (and in the same manner as above, is gradually increased during the energy minimization, so that at the end of the procedure all parameters are at their physical values). The release of fixed bond length constraints and the scaling of the repulsive van der Waals interactions creates a smaller and softer segment that finds its proper position more readily.

We focus on loops constrained by two peptide bonds, where the end-to-end distance is measured by the distance between the first N and the last C in the loop. Initial segments are generated by assigning random values to all dihedrals: all side-chain dihedrals and backbone dihedrals of the surface loops are assigned random values between $\pm 180^\circ$. For the extended structures it is necessary to restrict the backbone dihedrals to yield large end-end distances. Consequently, the backbone dihedrals are assigned random values between $-180^\circ \pm 120^\circ$. Similarly, the backbone dihedrals of helical segments are assigned random values between $\phi: -100^\circ \pm 70^\circ$ and $\psi: 60^\circ \pm 120^\circ$. The

Fig. 1. The bond scaling-relaxation algorithm. **A:** Initial random conformation; the distance between the two terminal peptide bonds is d_0 . **B:** Bonds in the loop segment are scaled by a factor of d/d_0 to meet the distance constraint d . **C:** Relaxation of the scaled structure via energy minimization; the scaled bond lengths gradually increase to their standard values during the relaxation process, while the two terminal peptide bonds are fixed to preserve the distance constraint. The van der Waals repulsive term is initially scaled by a factor of $(d/d_0)^2$ for atom pairs within the loop and by d/d_0 for interactions between the loop and the environment atoms. **D:** The final structure with the desired end-to-end distance constraint and all parameters have been restored to their physical values.

random segments are then scaled and subsequently translated and rotated to fit the geometries of the two terminal peptide bonds. The geometric constraint is enforced by applying a harmonic force of $250 \text{ kcal/mol} \cdot \text{\AA}^2$ to the first NH and the last CO groups of the loop. Atoms in the protein environment are restrained to their crystal coordinates by a harmonic force of $5 \text{ kcal/mol} \cdot \text{\AA}^2$. The scaling-relaxation algorithm has been implemented in CHARMM (Brooks et al., 1983) and CONGEN (Brucoleri & Karplus, 1987). The major technical challenge involved is the use of two sets of equilibrium bond lengths and van der Waals parameters, one for the loop segment and one for the protein environment, since most existing modeling packages permit the use of only one parameter set. In order to maintain the integrity of CHARMM and CONGEN the energy and force calculations were modified instead. All energy calculations are done with the polar hydrogen parameter file, param4.mod in CONGEN2.0.

The number of intermediate scaling-relaxation cycles, N_{cycle} , depends on the initial scaling factor d/d_0 as listed in Table 1. The scaling factors of the equilibrium bond lengths in the intermediate cycles are exponentially increased from d/d_0 to 1, in N_{cycle} cycles (Zheng et al., 1993). All minimizations are performed with the adopted basis Newton-Raphson (ABNR) algorithm in CONGEN, with a distance-dependent dielectric $\epsilon = 4R$ and a nonbonded cutoff distance of $9 \text{ \AA}$, where R is the distance between

¹ When extended segments are modeled, the randomization procedure for generating the initial conformations can be tuned to generate more extended structures.

Table 1. *Scaling and relaxation parameters*

Initial scaling factor (d/d_0)	N_{cycle}
<0.2	6
0.2–0.4	5
0.4–0.6	4
0.6–0.8	3
0.8–0.9	2
>0.9	1

two interacting atoms. Energy minimization is terminated when the energy drop in 50 minimization steps is less than 2 kcal/mol. All calculations were done on Silicon Graphics Iris (R3000 series) machines. The time required to generate, for example, a 7-residue loop in the Zif268 zinc-finger domain (26 residues) is approximately 3 min.

Results

We tested the algorithm on eight heptapeptide segments from six proteins with known X-ray coordinates (resolution 1.5–2.7 Å), chosen to represent a variety of protein conformations (Fig. 2; Table 2). Four of the segments are surface loops that lack well-defined secondary structures: rubredoxin residues 16–22; myoglobin residues 119–125 (loop connecting the seventh and eighth helices); immunoglobulins MCPC603 residues 56–62, and Bence–Jones residues 51–57 (both part of the L2 hypervariable loops). The remaining four loops have well-defined secondary structures and are partially or fully internal: myoglobin residues 91–97 are part of an α -helix; Zif268 residues 16–22 begin with a turn, 16–17, followed by an α -helix, 18–22; Zif268 residues 11–17 contain a β -strand, 12–14 followed by a turn, 15–16; bovine pancreatic trypsin inhibitor residues 18–24 are part of a long β -strand buried in the core of the protein.

Fig. 2. Schematic representation of the modeled loops (segments in bold). Zif268:16–22 and 1mba:91–97 are helical segments; Zif268:11–17 and 8pti:18–24 are parts of β -strands; 7rxn:16–22, 1mcp:56–62, 2rhe:51–57, and 1mba:119–125 are surface loops lacking well-defined secondary structures. The proteins are labeled with their Brookhaven Protein Data Bank entry names.

Table 2. Loop segments modeled

Segment class	Protein name	Protein Data Bank name	Resolution (Å)	Residue range	Sequence (one-letter code)
α -Helix	Zif268 zinc finger domain	— ^a	2.1	16–22	R S D E L T R
	Myoglobin	1mba	1.6	91–97	Q F A K E H V
β -Strand	Zif268 zinc finger domain	— ^a	2.1	11–17	D R R F S R S
	Bovine pancreatic trypsin inhibitor (BPTI)	8pti	1.8	18–24	I I R Y F Y N
Surface loops	Rubredoxin	7rxn	1.5	16–22	A E G D P D N
	Immunoglobulin (MCPC603)	1mcp	2.7	56–62	G A S T R E S
	Immunoglobulin (Bence-Jones)	2rhe	1.6	51–57	Y N D L L P S
	Myoglobin	1mba	1.6	119–125	S V A A P P A

^a Zif268 coordinates were kindly provided by N.P. Pavletich and C.O. Pabo.

Two alternative procedures are tested: (1) The bond-scaling relaxation algorithm is applied to the isolated heptapeptide, followed by insertion of the energy-minimized segment into its proper position in the protein, with a final energy minimization of the entire structure. (2) The bond-scaling relaxation algorithm is applied to the segment within the protein environment. Most current loop closure algorithms are similar to procedure 1 in that they first generate a constrained segment in isolation (Go & Scheraga, 1970; Bruccoleri & Karplus, 1985; Moulton & James, 1986; Shenkin et al., 1987), in contrast to procedure 2, in which the environment is included continuously from the start.

For each segment we generate 100 different conformations, and calculate their root-mean-square displacement (RMSD) from the crystal coordinates. Because movement of the constrained environment during minimization is less than 0.3 Å, the RMSD is calculated without rotation or translation of the generated loops. The lowest energy

and lowest RMSD conformations are presented in Table 2. For surface loops the structures predicted by both procedures have similar energy and RMSD values (Table 3). The RMSD per residue is also very similar (Fig. 3).

In contrast, for the internal structures procedure 2 consistently generated structures that are closer to the native conformations (Table 3; Fig. 3). If we use energy as a target function, the predicted structures (those of lowest energy), with the exception of 1mba:91–97 (discussed below), have all-atom RMSD values of 1.4–2.3 Å and backbone RMSD values in the range of 0.8–1.2 Å. In contrast, the structures predicted by procedure 1 have all-atom RMSD values of 2.7–5.4 Å and backbone RMSD values of 1.2–4.0 Å. Moreover, for the four examples tested, the lowest energy structures generated with continuous inclusion of environment coincide with the lowest RMSD structures (in the case of the myoglobin helix, this structure is not unique). This latter result is important, since energy is widely used as a preliminary target function in

Table 3. Energy (kcal/mol) and RMSD (Å) values of predicted loops

Class	Protein	Procedure 2				Procedure 1			
		Lowest energy conformation		Lowest RMSD conformation		Lowest energy conformation		Lowest RMSD conformation	
		Energy	RMSD ^a	Energy	RMSD ^a	Energy	RMSD ^a	Energy	RMSD ^a
α -Helix	Zif268 (16–22)	–242	0.83/1.43 (1)	b	b	–230	4.00/5.35 (5)	–204	2.59/5.00
	1mba (91–97)	–1,263	1.17/2.04 (1) ^c	b	b	–1,257	1.98/3.46 (3)	–1,245	1.91/3.40
β -Strand	Zif268 (11–17)	–250	0.95/3.47 (1)	b	b	–249	1.38/3.70 (2)	–243	0.98/3.60
	8pti (18–24)	–334	0.84/2.33 (1)	b	b	–343	1.24/2.67 (7)	–250	1.08/2.00
Surface loops	7rxn (16–22)	–359	0.70/1.33 (1)	b	b	–351	0.67/1.03 (1)	b	b
	2rhe (51–57)	–844	1.42/2.71 (1)	b	b	–845	1.39/2.69 (1)	b	b
	1mcp (56–62)	–959	1.14/2.31 (1)	b	b	–956	1.14/2.31 (1)	b	b
	1mba (119–125)	–1,305	1.44/2.10 (30)	–1,277	1.02/1.25	–1,307	1.12/1.35 (6)	–1,300	0.95/1.07

^a Backbone RMSD/all-atom RMSD.

^b Indicates correspondence between lowest energy and lowest RMSD conformations.

^c Three conformations within energy difference of 0.07% were observed. The RMSDs not shown are 2.35/3.69 and 4.31/5.50.

Fig. 3. Comparison of the backbone RMSDs per residue for the conformations predicted with procedure 1 (open circles) and those predicted with procedure 2 (closed circles). The ordinate depicts the amino acid sequence (one-letter code).

loop modeling, despite our knowledge that it is insufficient (Novotny et al., 1988, 1989; Shoichet & Kuntz, 1991).

The advantage of continuous inclusion of environment is further revealed by examining the predicted positions of the side chains (Fig. 4). When environmental information is utilized, the side chains can be placed fairly accurately. Examples are the histidine and phenylalanine residues on the right side of the 1mba:91-97 helix, which are part of a hydrophobic core. This is also the case for the side chains of 8pti, with the exception of Tyr²³. Further examination of this structure reveals that in our prediction Tyr²³ points to a cavity that is occupied by three water molecules in the crystal (lack of water in the modeling can be viewed, in this case, as failure to include full environmental information). The phenylalanine in the Zif268:11-17 loop is part of the conserved hydrophobic core of the zinc-finger motif and is also predicted accurately. The predicted backbone of this residue also forms two hydrogen bonds to a distant residue Tyr³, as seen in

the crystal of Zif268. These results clearly demonstrate that the algorithm can utilize environmental information. Furthermore, in cases where environmental information is lacking, side chains cannot be accurately positioned. This is evident from the poor positioning of Arg¹⁶, Asp¹⁸, and Arg²² in the Zif268:16-22 loop. The native conformation is obtained from a protein-DNA complex, in which these side chains contact the DNA (which is environmental information not included in the modeling). Similarly, the misprediction of Arg¹² and Arg¹³ in Zif268:11-17 can also be attributed to the lack of the DNA environment. In the crystal complexes Arg¹² contacts the DNA, and this interaction is stabilized by a charge-charge interaction with the side chain of Asp¹¹. In the predicted conformation Arg¹³ replaces Arg¹² to form a similar charge-charge interaction with Asp¹¹, resulting in a twist of the backbone conformation.

For the correct prediction of surface side chains, the solvent environment should be included in the calculations (Eisenberg & McLachlan, 1986). Because solvent has not been considered here, solvent-exposed side chains are mispredicted. This is the case for the left side of the 1mba:91-97 helix and the side chains of the surface loops of 7rxn, 1mcp, 2rhe, and 1mba:119-125. In addition, the use of side-chain rotamers (Dunbrack & Karplus, 1993; Schrauber et al., 1993) in generating initial random segments and/or a more exhaustive conformational search on the side-chain orientations after loop closure may further improve the prediction.

Although one might expect that after bond scaling the standard geometry would not be satisfactorily restored, our results indicate that this is not the case. Comparison of the predicted and the crystal structures show that the bond lengths of the predicted structures are always within 0.01 Å of the crystal values, the bond angles within 0.1°, and the improper dihedral angles within 0.03°. Furthermore, the overall energy profiles of the predicted loops are very similar to those of the crystal structures (data not shown), suggesting that the algorithm generates conformations that are close conformationally, as well as energetically to the native structures.

Discussion and conclusion

We have demonstrated that a practical method for predicting protein loop conformations, while continually utilizing environmental information, has been developed, and that the environmental information is necessary for accurate prediction. Although surface loops can be generated with equal accuracy in isolation, loops that are fully or partially buried cannot. In addition, a stronger energy-RMSD correlation is obtained with continuous environmental inclusion. But conformational energy is not always sufficient for selecting one or even a few loops from the generated ensemble (Shoichet & Kuntz, 1991; Ring et al., 1992). This is clearly seen in the example of

Fig. 4. Superposition (without rotation or translation) of the conformations predicted with procedure 2 (brown) and the corresponding crystal conformations (green). The backbones are represented by ribbons; the side chains are drawn as stick models.

Imba:91–97. The three lowest energy conformations are energetically indistinguishable (0.07% differences in energies), but their RMSDs from the crystal conformation vary from 1.17 to 4.31 Å. Finding better criteria for discrimination between correct and incorrectly folded protein segments is a long-standing problem in molecular modeling (Novotny et al., 1988, 1989).

An additional advantage of the scaling-relaxation method is its adaptability and applicability: it can easily be adapted to improved energy or other target functions, and additional information such as secondary structure (Zheng et al., 1993) or loop class assignment (Ring et al., 1992) can easily be considered when generating the initial conformations. The method can also be applied with ease to various additional modeling techniques. For example, entropy and temperature effects can be considered by using molecular dynamics instead of energy minimization for relaxation. We have recently applied the scaling-relaxation method in conjunction with a multiple copy simultaneous search (MCSS) method to study peptide docking to class-I major-histocompatibility receptors (R. Rosenfeld, Q. Zheng, S. Vajda, & C. DeLisi, in prep.).

The accuracy and speed of the method stems partially from the fact that we have transformed a constrained minimization problem into a nonconstrained one, without introducing additional energy terms. The energy cost of closing a loop is thus uniformly distributed along the entire loop and its surrounding environment. In contrast, adding penalty terms to the energy function in order to enforce the geometric constraints (Brunger et al., 1986) may result in a nonuniform distribution of tension, especially for long loops, whereas ideally the tension should be spread uniformly along the protein.

Acknowledgments

We thank Robert Bruccoleri for the CONGEN program and helpful discussions. We thank T. Smith for database support. This work was supported by NIAID grant AI30535 to Charles DeLisi.

References

- Alwyn, T. & Thirup, S. (1986). Using known substructures in protein model building and crystallography. *EMBO J.* 5, 819–822.
- Braun, W. (1987). Distance geometry and related methods for protein structure determination from NMR data. *Q. Rev. Biophys.* 19, 115–157.
- Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., & Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* 4, 187–217.
- Bruccoleri, R.E., Haber, E., & Novotny, J. (1988). Structure of antibody hypervariable loops reproduced by a conformational search algorithm. *Nature* 335, 564–568.
- Bruccoleri, R.E. & Karplus, M. (1985). Chain closure with bond angle variations. *Macromolecules* 18, 2767–2773.
- Bruccoleri, R. & Karplus, M. (1987). Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers* 26, 137–168.
- Brunger, A.M., Clore, G.M., Gronenborn, A.M., & Karplus, M. (1986). Three-dimensional structure of proteins determined by molecular dynamics with interproton distance restraints. *Proc. Natl. Acad. Sci. USA* 83, 3801–3805.
- Dunbrack, R.L., Jr. & Karplus, M. (1993). Backbone-dependent rotamer library for proteins. *J. Mol. Biol.* 230, 543–574.
- Eisenberg, D. & McLachlan, A.D. (1986). Solvation energy in protein folding and binding. *Nature* 319, 199–203.
- Go, N. & Scheraga, H.A. (1970). Ring closure and local conformational deformations of chain molecules. *Macromolecules* 3, 178–187.
- Greer, J. (1990). Comparative modeling methods: Application to the family of the mammalian serine proteases. *Proteins Struct. Funct. Genet.* 7, 317–334.
- Moult, J. & James, M.N.G. (1986). An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins Struct. Funct. Genet.* 1, 146–163.
- Novotny, J., Bruccoleri, R.E., & Saul, F.A. (1989). On the attribution of binding energy in antigen–antibody complexes MCPC603, D1.3, and HyHEL-5. *Biochemistry* 28, 4735–4749.
- Novotny, J., Rashin, A.A., & Bruccoleri, R.E. (1988). Criteria that discriminate between native proteins and incorrectly folded models. *Proteins Struct. Funct. Genet.* 4, 19–30.
- Palmer, K.A. & Scheraga, H.A. (1991). Standard-geometry chains fitted to X-ray derived structures: Validation of the rigid-geometry approximation. 1. Chain closure through a limited search of “loop” conformations. *J. Comput. Chem.* 12, 505–526.
- Ring, C.S., Kneller, D.G., Langridge, R., & Cohen, F.E. (1992). Taxonomy and conformational analysis of loops in proteins. *J. Mol. Biol.* 224, 685–699.
- Schrauber, H., Eisenhaber, F., & Argos, P. (1993). Rotamers: To be or not to be? *J. Mol. Biol.* 230, 592–612.
- Shenkin, P.S., Yarmush, D.L., Fine, R.M., Wang, H., & Levinthal, C. (1987). Predicting antibody hypervariable loop conformation. I: Ensembles of random conformations for ringlike structures. *Biopolymers* 26, 2053–2085.
- Shoichet, B.K. & Kuntz, I.D. (1991). Protein docking and complementarity. *J. Mol. Biol.* 221, 327–346.
- Skolnick, L. & Kolinski, A. (1990). Simulations of the folding of a globular protein. *Science* 250, 1121–1124.
- Summers, N. & Karplus, M. (1990). Modeling of globular proteins. A distance-based data search procedure for the construction of insertion/deletion regions and Pro ↔ non Pro mutations. *J. Mol. Biol.* 216, 991–1016.
- Weiner, P.K., Profeta, S., Jr., Wipff, G., Havel, T., Kuntz, I.D., Langridge, R., & Kollman, P.A. (1983). A distance geometry study of ring systems. *Tetrahedron* 39, 1113–1121.
- Zheng, Q., Rosenfeld, R., Vajda, S., & DeLisi, C. (1993). Loop closure via bond scaling and relaxation. *J. Comput. Chem.* 14, 556–565.