Simultaneous modeling of multiple loops in proteins

DROR ROSENBACH AND RAKEFET ROSENFELD

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

(RECEIVED August 18, 1994; ACCEPTED January 10, 1995)

Abstract

The most reliable methods for predicting protein structure are by way of homologous extension, using structural information from a closely related protein, or by “threading” through a set of predefined protein folds (“inverse folding”). Both sets of methods provide a model for the core of the protein – the structurally conserved secondary structures. Due to the large variability both in sequence and size of the loops that connect these secondary structures, they generally cannot be modeled using these techniques. Loop-closure algorithms are aimed at predicting loop structures, given their end-to-end distance. Various such algorithms have been described, and all have been tested by predicting the structure of a single loop in a known protein. In this paper we propose a method, which is based on the bond-scaling-relaxation loop-closure algorithm, for simultaneously predicting the structures of multiple loops, and demonstrate that, for two spatially close loops, simultaneous closure invariably leads to more accurate predictions than sequential closure. The accuracy of the predictions obtained for pairs of loops in the size range of 5–7 residues each is comparable to that obtained by other methods, when predicting the structures of single loops: the RMS deviations from the native conformations of various test cases modeled are approximately 0.6–1.7 Å for backbone atoms and 1.1–3.3 Å for all-atoms.

Keywords: bond-scaling-relaxation loop-closure algorithm; modeling; multiple loops; structure prediction

To date, the most reliable method for predicting the structure of a protein is by way of homologous extension (Greer, 1990; Mas et al., 1992; Boscott et al., 1993; Eisenmenger et al., 1993; Hilbert et al., 1993), i.e., the structure of a protein is modeled from the known structure of a closely related, homologous protein. An additional set of methods that are gaining wide use (Sippl, 1990; Jones et al., 1992; Maiorov & Crippen, 1992; Bowie & Eisenberg, 1993; Bryant & Lawrence, 1993; Lathrop & Smith, 1994) employs the “inverse approach,” in which the most “suitable” structure is chosen from among a collection of predefined folding motifs. Both of these sets of methods provide a model for the core, or structurally conserved regions of the protein. To produce an atomic resolution model for the entire protein, construction of the loops, which connect the secondary structure elements comprising the core, is necessary. These loops are commonly located on the surface of the protein and may therefore be involved in binding or recognition of other molecules. Thus, their correct prediction may be crucial for providing insight into the function of the protein.

Protein loops can vary widely both in sequence and in size, even between two closely related, homologous proteins. Various algorithms have been developed (Go & Scheraga, 1970; Fine

et al., 1986; Jones & Thirup, 1986; Moulton & James, 1986; Brucoleri & Karplus, 1987; Shenkin et al., 1987; Palmer & Scheraga, 1992; Zheng et al., 1993a) for predicting the structures of loops, given their end-to-end distance (loop closure algorithms), and all have been tested by deleting a single loop from a protein whose structure is known and rebuilding its conformation. But when modeling a protein whose conformation is unknown, several loops may need to be modeled, some of which may be spatially close enough to interact with each other. It is clear that any attempt to predict the structures of these close, interacting loops should “close” all loops simultaneously – the results of closing spatially close loops sequentially will, almost invariably, be order dependent.

The conformational space available to even short loops is very large, and hence most loop closure algorithms are limited to predicting the structures of relatively small (7–9 residues) loops. Simultaneous predictions of two or more loops of this size require searching a much larger space and may be further complicated by the strong interactions between the loops: for example, in one 14-residue loop, the major interactions between most residues may be nearest-neighbor interactions, whereas in two proximate 7-residue loops, additional interactions from the second loop will play an important role in determining the conformation of each loop. In this paper we present a strategy for closing two proximate loops – each in the above size range – simultaneously and demonstrate that simultaneous closing of the loops consistently

Reprint requests to: Rakefet Rosenfeld, 44 Cummington Street, Boston University College of Engineering, Boston, Massachusetts 02215; e-mail: kapi@darwin.bu.edu.

results in more accurate predictions than sequential closing. We use the bond-scaling-relaxation algorithm (Zheng et al., 1993a, 1993b), which has previously been shown to give satisfactory results when modeling single loops in a wide range of proteins. Previous work using this algorithm has determined that 100 initial random conformations generally suffice to find loop conformations close to the native structures for loops of seven residues. Scaling this result to a two-loop calculation, $100 \times 100 = 10,000$ initial conformations would have to be tested to give a high probability of finding accurate structures. The strategy suggested herein overcomes this combinatorial explosion, and gives satisfactory results from an order of magnitude of 10^3 starting conformations.

Numerous studies have clearly demonstrated that available molecular-mechanics potential functions cannot distinguish between native and other low-energy conformations (Novotny et al., 1988; Shoichet & Kuntz, 1991; Smith & Honig, 1994). This problem is generally circumvented by a two-step procedure: an initial conformational search using a standard molecular-mechanics potential, followed by evaluation of a manageable number of putative structures using a more rigorous free-energy function, which usually includes hydration and entropic terms (Wilson et al., 1991; Horton & Lewis, 1992; Krystek et al., 1993; Smith & Honig, 1994; Vajda et al., 1994). This work is aimed at the first of these steps, i.e., generating a manageable number of putative, low energy conformations for pairs of proximate loops, among which close-to-native structures are found.

Methodology

The bond-scaling-relaxation algorithm is used in all calculations (Zheng et al., 1993a, 1993b). Briefly, a random conformation is generated and subsequently scaled to meet the end-to-end distance constraint. The equilibrium bond lengths in the empirical energy function are also scaled accordingly, and the conformation is minimized. The equilibrium bond lengths are then gradually restored, concurrent with additional cycles of energy minimization, until correct physical bond lengths are obtained. This procedure works best when the initial conformations have a larger than desired end-to-end distance and are consequently scaled down. We therefore filter the randomly generated conformations and use only those with a larger than desired end-to-end distance. The entire procedure is performed within the rest of the protein (standard equilibrium bond lengths are used for all calculations on the unscaled portions of the protein), thereby allowing the loops to experience all forces from the rest of the protein as they "grow" back to their physical size. To conform with our previous findings, each prediction is obtained by applying the algorithm to 100 random starting conformations.

Initially, we close each loop separately (i.e., with the rest of the protein, including the second loop, intact). The accuracy of the prediction thus obtained provides an upper bound on the expected accuracy of predicting both loops simultaneously. A lower bound is obtained for each loop by predicting its conformation in the absence of the interacting loop. We then close the loops sequentially and finally test two strategies for simultaneous modeling of both loops.

Modeling single loops

For the sake of clarity, we refer to all protein atoms that are not in either of the loops being modeled as the environment of the

loops. Initially, each loop is predicted while the environment and the second loop are constrained to their known conformations. The same 100 random starting conformations used in these calculations are then used to model each of the loops in the absence of the second (i.e., the second loop is deleted, and the environment constrained to its known conformation).

Sequential modeling of two proximate loops

Both loops are deleted from the protein and one is reconstructed. If the loops are of unequal size, the shorter loop is reconstructed first; otherwise the choice is arbitrary. (Because the entire procedure is fast—only slightly longer than performing an energy minimization of the protein—in the case of equal size loops, it is worthwhile to increase the sampling by performing the procedure twice, starting with a different loop each time.) In this calculation, the second loop is missing while the environment is constrained to its known conformation. We then choose the minimum energy conformation obtained and attempt to predict the structure of the second loop. Additional evaluation of the limit of accuracy of the procedure is obtained by testing how well the second loop can be predicted when we choose the most accurate structure (lowest RMS displacement [RMSD] from the crystal conformation) found for the first.

Simultaneous modeling of two proximate loops

In order to keep the number of calculations feasible, predictions were obtained in two steps: (1) calculation of a small number of putative starting conformations for the first loop, followed by (2) simultaneous closure of each of these with 100 random starting conformations for the second loop. Because it is unclear if closing one loop with the second deleted is preferable to closing it with a random conformation of the second loop (i.e., if missing information is preferable to incorrect information), we tested two strategies for step 1: (A) We close one of the loops, while the other is deleted, and the environment constrained to its known conformation. Here too, if the loops are of unequal size, the shorter loop is closed first; otherwise the choice is arbitrary. (B) We close both loops together 100 times, starting from 100 random conformations for each.

For each of these strategies, we choose the five lowest energy conformations obtained (for strategy A the total energy is considered; for strategy B only the internal energy of the loop + the interaction energy of the loop with the environment are considered) and reclose each of these, while simultaneously closing the second loop—i.e., each of the initial conformations that resulted in the five lowest energy structures for the first loop (after relaxation) are placed back into the position of the first loop, and the second is modeled starting from 100 random conformations. Thus, a total of $100 + (5 \times 100) = 600$ initial conformations are modeled in each case.

Implementation

The bond-scaling-relaxation loop closure algorithm has been incorporated into the molecular modeling package CHARMM version 22.g2 (Brooks et al., 1983). All calculations were carried out with the all-hydrogen parameter files using a dielectric constant of 1 and a nonbonded cutoff distance of 10 Å. A force constant of 5 kcal/mol·Å² was used in all constraints. Minimi-

Table 1. *Loops modeled*

Protein	Protein code herein (PDB code)	Loop 1		Loop 2	
		Residues ^a	Sequence	Residues ^a	Sequence
Phosphotransferase	Hpr-a (2hpr)	28-32 (5)	Y-D-A-D-V	66-72 (7)	G-A-D-E-N-D-A
Phosphotransferase	Hpr-b (2hpr)	27-33 (7)	K-Y-D-A-D-V-N	66-72 (7)	G-A-D-E-N-D-A
Ubiquitin	Ubq (1ubq)	19-24 (6)	P-S-D-T-I-E	51-57 (7)	E-D-G-R-T-L-S
Bovine pancreatic trypsin inhibitor	BPTI (8pti)	13-19 (7)	P-C-K-A-R-I-I	33-39 (7)	F-V-G-G-G-C-R

^a Residue range and (size of loop).

zations were performed using the adopted basis Newton Raphson procedure in CHARMM. A minimization procedure was terminated when the energy decrease in 50 steps was less than 5 kcal/mol. All crystal conformations used in this work were taken from the Brookhaven Protein Data Bank and refined by this minimization procedure. The RMSD between two loop conformations was calculated for all backbone heavy atoms *without* rotation and translation. All calculations were carried out on Silicon Graphics workstations.

Results

Calculations were performed on four pairs of interacting loops (Table 1) from three high-resolution (≤ 2.0 Å) protein structures. These were chosen randomly from the Brookhaven Protein Data Bank, the only requirements being that they be small (<200 residues) and that they contain a pair of spatially close, interacting loops. The two loops in the histidine-containing phosphotransferase (Hpr) are 5 and 7 residues long. In order to test the effect of loop length on our predictions, we performed an additional set of predictions for these loops with the shorter loop extended by one residue at each end. All loop pairs modeled are spatially close and strongly interact (their interaction energy is at least 10% of the interaction energy of each of the loops with the environment).

Modeling single loops

To obtain an upper bound on the expected accuracy of modeling each of the seven loops tested (loop 2 of Hpr-a and Hpr-b are the same) and to assess the effects of missing information (deletion of an interacting loop) on the predictions, each was modeled both in the presence (procedure a) and absence (procedure b) of the second loop, using the same 100 random initial starting conformations in both sets of calculations. In general, deletion of the second loop resulted in poorer predictions (Table 2, columns I); in six of the eight cases tested, more than 50% of the initial conformations resulted in more accurate loop conformations in the presence of the second loop than in its absence. Although this result is statistically significant, it is uninteresting because many of these conformations are inaccurate (backbone RMSDs of 3–4 Å). If we look only at the conformations whose backbones are within 2 Å (Table 2, columns II) or 1.5 Å (Table 2, columns III) of the native ones, it is clear that deletion of a spatially close loop strongly hinders our ability to predict loop conformations: for the first loop of Hpr-a, looking at all initial conformations that resulted in predictions within 2 Å

of the native (by at least one of the procedures), procedure a outperformed procedure b 52 times, whereas procedure b outperformed procedure a only 32 times. The corresponding numbers for predictions within 1.5 Å of the native are 26 and 15, respectively. For the longer loops, the difference is even more dramatic: if one looks only at the most accurate conformations (within 1.5 Å of the native), procedure a invariably outperformed procedure b (Table 2, columns III).

Figures 1, 2, 3, and 4 depict the energy-RMSD correlation for each of the eight loops modeled in the presence (A, B) and absence (C, D) of the second loop. Although we do not expect a *strong* correlation between the in-vacuum molecular mechanics potential energy we use and the accuracy of the resulting conformations (RMSD from native), an effective conformational search procedure should find low RMSD conformations among a reasonable number of lowest energy ones. Because the final evaluation of these structures will require a computationally more demanding energy function, it should only be performed on a small subset of the putative structures generated by the conformational search step. Figures 1, 2, 3, and 4 demonstrate that not only are the conformations generated in the absence of the second loop less accurate, but also that the energy-RMSD correlation observed for the low-energy conformations of five of the seven loops (Figs. 1C, 1D, 2D, 3C, 4C, and 4D) is much

Table 2. *Comparison of single loops generated in the presence of a second loop (procedure a) with those generated in its absence (procedure b)*

Protein	Loop 1			Loop 2		
	I ^a	II ^b	III ^c	I ^a	II ^b	III ^c
Hpr-a	61	52/32	26/15	65	4/0	3/0
Hpr-b	56	13/3	5/2	71	4/0	3/0
Ubq	54	24/9	17/1	44	3/4	2/1
BPTI	43	9/0	2/0	56	5/1	1/0

^a Percentage of times that procedure a generated loops with RMSDs lower than those of the corresponding loop generated with procedure b.

^b For generated loops with backbone RMSDs <2.0 Å: (number of loops generated by procedure a that are better predictions than the corresponding prediction by procedure b)/(number of loops generated by procedure b that are better predictions than the corresponding prediction by procedure a).

^c Same as II, but for all generated loops with backbone RMSDs <1.5 Å.

Fig. 1. Energy versus backbone RMSD for loops of Hpr-a. **A:** Loop 1 modeled in the presence of loop 2. **B:** Loop 2 modeled in the presence of loop 1. **C:** Loop 1 modeled when loop 2 is deleted. **D:** Loop 2 modeled when loop 1 is deleted.

weaker in this case. For Hpr-a, a very good energy-RMSD correlation can be seen, and very accurate conformations are observed when the loop is modeled with no missing information (Fig. 1A), whereas a much weaker correlation and less accurate structures are observed when the calculations are performed in the absence of the second loop (Fig. 1C). A similar result can be seen for the second loop of this protein by comparing Figure 1B and D. When the short, five-residue loop of this protein

is extended to seven residues (loop 1 of Hpr-b), our results are poorer (Fig. 2). Although a low RMSD (0.68 Å) structure is found in the presence of loop 2, it is only the 25th lowest energy conformation and, among the 10 lowest energy conformations, the most accurate has a backbone RMSD of 1.05 Å. Also, the energy-RMSD correlation is weak both in the presence and absence of the second loop. For ubiquitin (Ubq), a good energy-RMSD correlation is observed for the low energy conformations

Fig. 2. Energy versus backbone RMSD for loops of Hpr-b. **A, B, C, D:** As in Figure 1.

Fig. 3. Energy versus backbone RMSD for loops of Ubq. **A, B, C, D:** As in Figure 1.

of loop 1 when loop 2 is present in the calculation, and this correlation diminishes when loop 2 is deleted (Fig. 3A,C). The energy–RMSD correlation for loop 2 is very weak in both calculations, but low RMSD conformations are found only in the presence of the first loop (Fig. 3B,D). For bovine pancreatic trypsin inhibitor (BPTI), the conformations generated are generally less accurate than for the other loops, yet a satisfactory energy–RMSD correlation for the low-energy conformations of both loops can be seen when modeled in the presence of the

other (Fig. 4A,B). When the loops are modeled in the absence of the other, not only does the correlation diminish, but the accuracy of the structures found decreases (Fig. 4C,D). It is thus apparent that if the structures of both loops were modeled sequentially, the resulting predictions would be very inaccurate.

Table 3 shows the RMSD of the most accurate structures found for each loop in the presence and absence of the second, as well as the RMSD of the lowest energy structures obtained. In two of the seven loops, the minimum energy conformation

Fig. 4. Energy versus backbone RMSD for loops of BPTI. **A, B, C, D:** As in Figure 1.

Table 3. Upper and lower bounds on expected accuracy of simultaneous modeling of both loops

Protein	Loop	Procedure a ^a						Procedure b ^a					
		Lowest RMSD			Minimum energy			Lowest RMSD			Minimum energy		
		RMSD ^b	Energy	Rank ^c	Energy	RMSD ^b	Rank ^d	RMSD ^b	Energy	Rank ^c	Energy	RMSD ^b	Rank ^d
Hpr-a	1	0.06	-1,393	1	— ^c	—	—	0.68	-1,026	23	-1,101	0.94	6
	2	0.19	-1,394	1	—	—	—	1.71	-1,084	52	-1,175	5.17	77
Hpr-b	1	0.68	-1,219	25	-1,325	1.85	13	1.25	-1,010	5	-1,038	3.64	56
	2	0.19	-1,394	1	—	—	—	1.68	-943	69	-1,044	2.97	7
Ubq	1	0.16	-1,944	5	-1,965	0.33	2	0.61	-1,471	15	-1,507	1.99	22
	2	1.41	-1,799	58	-1,915	3.72	49	1.35	-1,654	54	-1,739	3.18	43
BPTI	1	1.38	-1,532	3	-1,540	1.45	2	2.15	-1,324	34	-1,355	3.73	33
	2	1.09	-1,521	3	-1,529	2.64	24	1.61	-1,332	37	-1,399	2.62	20

^a Loop modeled in the presence (procedure a) or absence (procedure b) of a second loop.^b Backbone RMSD of loop.^c Ranking of energy among all 100 conformations.^d Ranking of RMSD among all 100 conformations.^e Lowest RMSD structure is the minimum energy structure.

obtained is also the most accurate structure, when the calculations are performed in the presence of the second loop, and for three other loops, the lowest RMSD conformation is found among the five minimum energy structures. When the calculations are performed in the absence of the second loop, only in one out of the eight loops (the second loop of Hpr is now modeled twice, because the amount of missing information varies by two residues between Hpr-a and Hpr-b) is the lowest RMSD conformation found among the five minimum energy structures. The result obtained for the second loop of Ubq (the RMSD of the best structure obtained by procedure b is better than that obtained by procedure a) is unexpected. Further examination of the initial structure that resulted in this low RMSD conformation revealed that its initial position clashes with the first loop and, in the presence of the latter, cannot overcome this clash. In its absence, this initial loop results in a more accurate conformation than the others tested. The lowest RMSD structures obtained in the presence and absence of the second loop provide the upper and lower bounds, respectively, on the expected results of modeling both loops simultaneously (with the exception of loop 2 of Ubq). But it is clear that even if a more sophisticated target function were used to choose the most accurate structures for the first loop in a sequential closing procedure, it would be impossible to choose the closest to native structures because the target function would likely include solvation and entropic terms that are proportional to the exposed surface area of the protein, and whose evaluation would be meaningless when the second loop is missing from the calculation. Therefore, these bounds are underestimates (i.e., the actual bounds are larger). Nevertheless, as we show below, simultaneous closure of both loops invariably results in predictions that are between these two bounds.

Modeling two loops

The results of sequentially closing each loop pair are shown in the center column of Table 4. The first loop was modeled in the absence of the second, and the lowest energy structure obtained

was chosen as the prediction. The second loop was then modeled with the first constrained to its predicted conformation. Because the energy-RMSD correlations obtained when modeling the loops with missing information (second loop deleted) was very weak (Figs. 1C,D, 2C,D, 3C,D, 4C,D), predicting the conformation of the second loop with each of the five minimum energy structures obtained for the first did not result in significantly better results (data not shown). Although, as noted above, it is unlikely that any target function would be able to pick out the most accurate structures when the second loop is deleted, we nevertheless attempted to predict the structure of the second loop when the lowest RMSD structure was chosen for the first (Table 4, bracketed numbers). It is interesting that, in the case of Hpr-a, choosing a more accurate structure for loop 1 resulted in a poorer prediction for loop 2. This result emphasizes the difficulty in modeling two loops simultaneously: when loop 1 is in its native conformation, loop 2 is predicted to within 0.19 Å backbone RMSD (Table 3). When the predicted conformation of loop 1 is 0.68 Å from the native, the prediction of loop 2 falls to 2.32 Å, whereas when a 0.94-Å backbone RMSD structure is chosen for loop 1, the resulting prediction of loop 2 has a slightly better RMSD—1.93 Å (Table 4).

For the simultaneous closure of each loop pair, an initial screening of possible starting structures for the first loop was performed. Two strategies were tested for this screening: the second loop was deleted from the calculation (see Table 5, Missing information); and a random structure was placed in the position of the second loop (see Table 5, Incorrect information). On average, structures obtained with incorrect information at the position of the second loop are slightly more accurate than those obtained with the second loop deleted (Table 5). With the exception of the first loop of Hpr-a, the average RMSD of the 10 most accurate structures obtained is either lower when a random structure is placed at the position of the second loop than when this loop is deleted, or comparable in both cases. But the improvement, when found, is small, and if one looks at each of the 10 most accurate structures, one intermittently finds the most accurate structures among the calculations performed in the ab-

Table 4. *RMSDs of sequential versus simultaneous closing of loop pairs*

Protein	Sequential closing ^a	Simultaneous closing ^b
Hpr-a	0.94/1.93 (1.59)	1.02/0.64 (0.82)
		0.61/0.95 (0.83)
	[0.68/2.32 (1.83)]	1.07/0.71 (0.88)
		1.03/0.96 (0.99)
		1.30/0.72 (1.00)
Hpr-b	2.29/3.09 (2.69)	1.09/0.75 (0.94)
		1.24/0.76 (1.04)
	[1.25/1.46 (1.36)]	1.47/0.92 (1.23)
		1.45/1.00 (1.25)
		1.60/1.12 (1.38)
Ubq	1.99/2.46 (2.25)	0.78/1.15 (1.00)
		0.40/1.45 (1.10)
	[0.61/1.73 (1.33)]	0.69/2.03 (1.56)
		1.25/1.81 (1.57)
		1.24/1.85 (1.60)
BPTI	1.52/2.62 (2.07)	1.62/1.11 (1.38)
		1.71/1.39 (1.56)
	[1.27/1.62 (1.45)]	1.49/1.64 (1.57)
		2.01/1.23 (1.66)
		1.72/1.65 (1.68)

^a First line: First loop is predicted in the absence of second, and minimum energy structure is chosen as prediction. Second loop is then predicted. Shown are RMSD of loop 1/RMSD of loop 2 and (in parentheses) RMSD of both loops. Second [bracketed] line: Same as above, but lowest RMSD structure of first loop is chosen as prediction.

^b First loop is predicted 100 times in the absence of second and 100 times in the presence of a random conformation for second loop. Five lowest energy structures are chosen, and reclosed (see Methodology section) with 100 random conformations for second loop. Shown are the five lowest RMSD results obtained.

sence of the second loop (data not shown). Because each of the calculations takes only 15 minutes on a Silicon Graphics R4400 workstation, we decided that both strategies should be performed for each loop pair.

Table 5. *Comparison of the average backbone RMSD of the 10 most accurate structures obtained with missing or incorrect information at the position of the second loop*

Protein	Loop	Average backbone RMSD of 10 most accurate structures	
		Missing information ^a	Incorrect information ^b
Hpr-a	1	0.91	1.14
	2	2.41	2.23
Hpr-b	1	1.73	1.58
	2	2.67	2.12
Ubq	1	1.20	0.90
	2	2.05	2.16
BPTI	1	2.68	1.91
	2	2.06	2.18

^a Second loop is deleted from the calculation.

^b A random initial structure is placed in the position of the second loop.

After the prescreening, the five random structures that resulted in the five lowest energy conformations for each strategy, after relaxation, were chosen as the starting conformations for the first loop in a simultaneous closing procedure. The results using both prescreening strategies were pooled together, and the overall five most accurate structures (ranked by the overall RMSD of both loops) are shown in Table 4. Two conclusions can immediately be drawn from these results: (1) simultaneous closure of both loops results in significantly more accurate predictions; (2) for all four pairs of loops, at least one of the predictions is between the upper and lower expected bounds shown in Table 3, despite the fact that these bounds are underestimated. For three of the eight loops (loop 2 of Hpr-a, loop 2 of Hpr-b, and loop 1 of BPTI) all five lowest RMSD structures are more accurate than the lower bounds estimated (Table 3). For the other loops, at least one of the structures obtained is lower than this bound. The success of the simultaneous closure in predicting loop 2 of Ubq (one of the structures has a lower RMSD than the surprisingly low RMSD obtained in the absence of loop 1) again demonstrates the advantage of this method. A comparison of the most accurate structure obtained in the sequential calculation (bracketed numbers in Table 4) with the five best structures obtained in the simultaneous closure is encouraging; despite the stringency of this comparison (in the sequential closure the lowest RMSD conformation for the first loop was chosen as the prediction, whereas in the simultaneous closure the five minimum energy conformations were chosen), simultaneous prediction consistently resulted in better predictions: if we look at the overall RMSD of both loops, for Hpr-a, all five best simultaneous closure results are more accurate than the most accurate sequential closure result; for Hpr-b, four of the five best simultaneous predictions are better than the most accurate sequential prediction; for Ubq, two of the simultaneous predictions are more accurate; and for BPTI, one of the simultaneous predictions is more accurate than the best possible sequential closing result. If we compare the results of sequential closing when the lowest energy conformation is chosen for the first loop with those of simultaneous closure, the latter invariably outperforms the former.

The higher accuracy achieved by simultaneous closure can also be seen in the C_{α} -traces of the most accurate prediction obtained by sequential and simultaneous closures (Figs. 5, 6, 7, and 8). For Hpr-a (Fig. 5), a good prediction is found for loop 1 (residues 28–32) in both cases. In contrast, the results for loop 2 are significantly more accurate when both loops are predicted simultaneously (Fig. 5B). In particular, residues 66 and 69–72 are close to their native positions in Figure 5B (simultaneous closure), whereas in Figure 5A (sequential closure) these residues deviate from their native conformation substantially. When the short loop of this protein is extended by an additional residue at each end (Hpr-b), the differences between the two procedures becomes more dramatic (Fig. 6): the higher accuracy seen at the ends of loop 2 with the simultaneous closure in Hpr-a is also observed in Hpr-b, but in addition the overall configuration of the first loop is predicted well only with the simultaneous closure; the configuration of the best prediction obtained with the sequential closure is very different from the native (Fig. 6A). The predictions of Ubq shown in Figure 7 show similar results: loop 1 is predicted well in both cases, whereas the prediction of loop 2 is substantially more accurate when both loops are predicted simultaneously. For BPTI (Fig. 8), the improvement is smaller

Fig. 5. Comparison of sequential and simultaneous predictions of Hpr-a loops. C_{α} -traces for the best structures generated (overall minimum backbone RMSD) sequentially (**A**) and simultaneously (**B**). Note that for the sequential closure, the best structure obtained for Hpr-a was that in which the minimum energy structure of loop 1 was chosen (see Table 4), as opposed to all other cases, in which the most accurate structure generated was that in which the minimum RMSD conformation of loop 1 was chosen. Bold line shows the nonloop regions, light-dashed lines show the actual loops, and the bold-dashed lines show the predicted loops.

Fig. 6. Comparison of sequential and simultaneous predictions of Hpr-b loops. **A, B:** As in Figure 5.

because the predictions for these loops are, in general, less accurate than for the other loops tested.

An all-atom representation of the most accurate structures found with the simultaneous closure procedure is shown in Figure 9A for Hpr-a, and in Figure 9B for Ubq. For Hpr-a, the RMSD from the native structure is 1.02/1.24 Å (backbone/all-atom) for the short (five-residue) loop, and 0.65/1.17 Å for the longer (seven-residue) loop. For Ubq, the corresponding numbers are 0.78/1.14 Å and 1.15/2.59 Å for the six- and seven-residue loops, respectively. Although choosing these structures from among the low-energy conformations depends on an adequate function for evaluating the free energy of the different loop conformations (Smith & Honig, 1994), in all proteins modeled, low RMSD conformations were always found among the 10 lowest energy conformations. For Hpr-a, the most accurate structure found (Fig. 9A) has the overall third lowest energy (data not shown), whereas the most accurate structure found for Ubq (Fig. 9B) is also the minimum energy conformation. The

best structures obtained for Hpr-b and Bpti have RMSDs of 1.09/2.36 Å, 0.75/1.38 Å and 1.62/3.30 Å, 1.11/1.79 Å, respectively. These values are well within the ranges obtained when predicting structures of single loops using either the bond-scaling-relaxation algorithm used herein (Zheng et al., 1993b), or various other methods (Jones & Thirup, 1986; Moult & James, 1986; Brucoleri & Novotny, 1992).

Discussion

Modeling proteins inevitably requires loop-closure algorithms, because loop conformations can vary widely even in two closely related homologous proteins. Often, two or three loops are spatially close enough to interact and thus should be modeled simultaneously. This was clearly evident in the work of Brucoleri and Novotny (1992), who sequentially modeled the six hypervariable loops in immunoglobulins: in the HyHEL-5 immunoglobulin, the predicted H1 loop protruded into the space occupied in the native structure by the H2 loop, making an accurate prediction of the latter impossible. Although simultaneously modeling six loops with the method proposed herein would not be feasible, utilizing the canonical structures of the hypervariable loops (Chothia & Lesk, 1987) in conjunction with this method could reduce the conformational space enough to make such calculations feasible.

Fig. 7. Comparison of sequential and simultaneous predictions of Ubq loops. **A, B:** As in Figure 5.

Fig. 8. Comparison of sequential and simultaneous predictions of BPTI loops. **A, B:** As in Figure 5.

The method we have proposed can model multiple loops in proteins simultaneously in such a manner that each loop “feels” the forces from all other loops, as well as from the framework of the protein, as it “grows” to its physical size after being scaled

down to meet the end-to-end distance constraint. Although the computational time does not increase exponentially with loop size, enabling either longer or more loops to be modeled simultaneously, the accuracy obtained decreases. This is clearly seen when comparing the predictions obtained for Hpr-a and Hpr-b: not only does the accuracy of prediction of the first loop decrease when loop size is increased from five to seven residues, but the prediction of the second loop (unchanged in size) is less accurate also. However, because the procedure is fast, this drawback can, in part, be overcome by additional sampling. Additional difficulties that arise in loop modeling are exemplified by the unexpected results obtained for Ubq: the second loop was predicted more accurately in the absence of the first than in its presence (1.35 Å as compared to 1.41 Å), and even more accurate conformations for this loop were found when both loops were modeled simultaneously (1.15 Å).

Modeling protein loops requires both efficient methods to search the conformational space and scoring functions for choosing correct conformations from the various structures generated by the search. We have proposed a method that addresses the first of these tasks. Smith and Honig (1994) have recently addressed the second, and various other empirical free energy functions have recently been proposed (Krystek et al., 1993; Vajda et al., 1994) that could, in principle, be modified for scor-

Fig. 9. Final predictions of the two loops of Hpr-a (**A**) and Ubq (**B**). Stereo images of all heavy atoms. Native conformations have orange backbones and blue side chains; predicted conformations have green backbones and pink side chains.

ing loop conformations. The major problem still remaining in cases where multiple loops are modeled is choosing the correct conformation when only part of the structure is present. When modeling n loops sequentially, one or a few conformations for each loop has to be chosen for the first $n - 1$ loops modeled, in the absence of the $n - 1$ to 1 other loops. Because the free energy functions used for scoring the putative conformations contain entropic and/or solvation terms proportional to the exposed surface area of the molecule, they would not be valid for structures with missing information. This problem also arises in our method because we initially pre-screen putative structures for the first loop and then reclose each of these with the second. In this work, we overcame this problem by extensive sampling: several possible starting conformations for the first loop were tested when modeling the second. Because the method is fast, this was affordable. When extending the calculations to three or more loops, this will not be feasible, and we are currently looking at various scoring schemes that may enable choosing the correct structures when other loops are absent.

Acknowledgments

We are grateful for the generous help of Prof. Charles DeLisi and Prof. Temple Smith, whose encouragement and support throughout this project were imperative. We also thank Dr. Sandor Vajda for his suggestions. R.R. was supported by grant AI30535 awarded to C. DeLisi by NIAID, NIH. D.R. was supported by grant LN05205 awarded to T. Smith by NLM.

References

- Boscott PE, Barton GJ, Richards WG. 1993. Secondary structure prediction for modeling by homology. *Protein Eng* 6:261–266.
- Bowie JU, Eisenberg D. 1993. Inverted protein structure prediction. *Curr Opin Struct Biol* 3:437–444.
- Brooks BR, Brucoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem* 4:187–217.
- Brucoleri RE, Karplus M. 1987. Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers* 26:137–168.
- Brucoleri RE, Novotny J. 1992. Antibody modeling using the conformational search program CONGEN. *Immunomethods* 1:96–106.
- Bryant SH, Lawrence CE. 1993. An empirical energy function for threading protein sequence through a folding motif. *Proteins Struct Funct Genet* 16:92–112.
- Chothia C, Lesk AM. 1987. Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol* 196:901–917.
- Eisenmenger F, Argos P, Abagyan R. 1993. A method to configure protein side-chains from main chain trace in homology modeling. *J Mol Biol* 231:849–860.
- Fine RM, Wang H, Shenkin PS, Yarmush DL, Levinthal C. 1986. Predicting antibody hypervariable loop conformations. II: Minimization and molecular dynamics studies of MCPC603 from many randomly generated loop conformations. *Proteins Struct Funct Genet* 1:342–362.
- Go N, Scheraga HA. 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 mammalian serine proteases. *Proteins Struct Funct Genet* 7:317–334.
- Hilbert M, Bohm G, Jaenicke R. 1993. Structural relationships of homologous proteins as a fundamental principle in homology modeling. *Proteins Struct Funct Genet* 17:138–151.
- Horton N, Lewis M. 1992. Calculation of the free energy of association for protein complexes. *Protein Sci* 1:169–181.
- Jones DT, Taylor WR, Thornton JM. 1992. A new approach to protein fold recognition. *Nature* 358:86–89.
- Jones TA, Thirup S. 1986. Using known substructures in protein model building and crystallography. *EMBO J* 5:819–822.
- Krystek S, Stouch T, Novotny J. 1993. Affinity and specificity of serine endopeptidase–protein inhibitor interactions. Empirical free energy calculations based on X-ray crystallographic structures. *J Mol Biol* 234:661–679.
- Lathrop R, Smith T. 1994. A branch-and-bound algorithm for optimal protein threading with pairwise (contact potential) amino acid interactions. *Proceedings of the 27th Annual Hawaii International Conference on System Sciences*:365–374.
- Maierov VN, Crippen GM. 1992. Contact potential that recognizes the correct folding of globular proteins. *J Mol Biol* 227:876–888.
- Mas MT, Smith KC, Yarmush DL, Aisaka K, Fine RM. 1992. Modeling the anti-CEA antibody combining site by homology and conformational search. *Proteins Struct Funct Genet* 14:483–498.
- Moult J, James MN. 1986. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins Struct Funct Genet* 1:146–163.
- Novotny J, Rashin AA, Brucoleri RE. 1988. Criteria that discriminate between native proteins and incorrectly folded models. *Proteins Struct Funct Genet* 4:19–30.
- Palmer KA, Scheraga H. 1992. Standard-geometry chains fitted to X-ray derived structures: Validation of the rigid body approximation. II systematic searches for short loops in proteins: Applications to bovine pancreatic ribonuclease A and lysozyme. *J Comput Chem* 13:329–350.
- Shenkin PS, Yarmush DL, Fine RM, Wang HJ, Levinthal C. 1987. Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. *Biopolymers* 26:2053–2085.
- Shoichet BK, Kuntz ID. 1991. Protein docking and complementarity. *J Mol Biol* 221:327–346.
- Sippl MJ. 1990. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol* 213:859–883.
- Smith KC, Honig B. 1994. Evaluation of the conformational free energies of loops in proteins. *Proteins Struct Funct Genet* 18:119–132.
- Vajda S, Weng Z, Rosenfeld R, DeLisi C. 1994. The effect of conformational flexibility and solvation on receptor–ligand binding free energies. *Biochemistry* 33:13977–13988.
- Wilson C, Mace JE, Agard DA. 1991. Computational method for the design of enzymes with altered substrate specificity. *J Mol Biol* 220:495–506.
- Zheng Q, Rosenfeld R, Vajda S, DeLisi C. 1993a. Loop closure via bond scaling and relaxation. *J Comput Chem* 14:556–565.
- Zheng Q, Rosenfeld R, Vajda S, DeLisi C. 1993b. Determining protein loop conformation using scaling-relaxation techniques. *Protein Sci* 2:1242–1248.