JMB

Refolding of Potato Carboxypeptidase Inhibitor by Molecular Dynamics Simulations with Disulfide Bond Constraints

Marc A. Martí-Renom[1], Roland H. Stote[2*], Enrique Querol[1] Francesc X. Avilés[1] and Martin Karplus[2,3*]

[1]*Institut de Biologia Fonamental and Departament de Bioquímica, Universitat Autònoma de Barcelona 08193 Bellaterra, Barcelona Spain*

[2]*Laboratoire de Chimie Biophysique, ESA 7006 CNRS Institut Le Bel - ISIS 4 rue Blaise Pascal Université Louis Pasteur 67000 Strasbourg, France*

[3]*Department of Chemistry Harvard University, Cambridge MA 02138, USA*

The folding of the potato carboxypeptidase inhibitor (PCI) from partially unfolded conformations by the introduction of native disulfide bond constraints was studied by molecular dynamics simulations in explicit solvent. PCI consists of a globular core (Cys8 to Cys34), two flexible terminal regions (Glu1 to Ile7 and Glu35 to Gly39) and three loop regions characteristic of the family of proteins known as knottins. To generate unfolded conformations, two high temperature (600 K) simulations were performed; one with the native disulfide bonds intact (N600), and one with the disulfide bonds broken (ND600). For comparison purposes, two simulations at 300 K were done; one with the native disulfide bonds (N300), and one with the disulfide bonds broken (ND300). The N300 simulation reached an energetic equilibrium within a few picoseconds and maintained a stable structure during the 500 ps simulation. The three other simulations led to partial unfolding. The largest changes were observed in ND600 simulation with an rms deviation of over 5 Å and radius of gyration 12.5% larger than the crystal structure value. Six structures from the ND600 simulation and one from the N600 simulation were used as starting structures for nine refolding simulations with somewhat different protocols for reforming the native disulfide bonds; in all cases the disulfides were reformed at 600 K and the temperature was decreased to 300 K for equilibration of the folded structures. Except for one structure that was significantly misfolded (final rms of 6.64 Å with respect to N300), the other folding simulations recovered the native simulation structure (N300) to within rms differences ranging from 1.8 to 3.2 Å for the main-chain of the core, relative to the N300, the X-ray and the NMR structures. Of particular interest is the internal and overall refolding behavior of the three loop regions. The more unfolded starting structures led to smaller rms values for the folded structures. Several energetic and solvation models were used to evaluate the X-ray, NMR, N300 and refolded structures. Although most models can distinguish the X-ray, NMR and N300 from the refolded structures, there is no correlation between the rms values of the latter and their estimated stability. Implications of the present results for protein folding by simulations and database search methods are discussed.

© 1998 Academic Press

Keywords: potato carboxypeptidase inhibitor; thermal unfolding; refolding; molecular dynamics

*Corresponding authors

Introduction

The mechanism of protein folding and unfolding is among the major unsolved problems of structural biology. One approach uses molecular dynamics simulations (MD) at high temperatures to characterise the unfolding process (Fan *et al.*, 1991; Daggett & Levitt, 1992a,b, 1993a,b; Mark & van Gunsteren, 1992; Vishveshwara *et al.*, 1993; Tirado-Rives & Jorgensen, 1993; Li & Daggett, 1994; Caflisch & Karplus, 1994, 1995; Alonso & Daggett, 1995; Lazaridis & Karplus, 1997). Corresponding studies of folding (e.g. low temperature dynamics or simulated annealing starting with the denatured state) have not been successful because the denatured polypeptide chain rapidly collapses and is trapped far from the native state (Alonso & Daggett, 1995; Lazaridis & Karplus, 1997). This is not surprising since folding takes microseconds or longer in solution and no simulations of this length have yet been possible.

Because of the difficulty of studying larger systems, it is of interest to use simulations to examine the unfolding and folding of small proteins for which experimental data are available. One such protein is the potato carboxypeptidase inhibitor (PCI), a 39 amino acid protein that plays a role in the defense mechanism of plants (Hass & Ryan, 1981). PCI is a representative of a series of small proteins with very similar topologies called T-knottins (Lin & Nussinov, 1995). They generally have three disulfide bonds which provide significant constraints on their structures. These proteins are further characterized by two short β strands and three loops which together form a topology resembling the letter ''T''. Each of the three loops is delineated by two disulfide bonds and there is a specific pattern to the disulfide bonding. The X-ray crystal structure of PCI complexed with carboxypeptidase A has been determined to a 2.5 Å resolution (Rees & Lipscomb, 1982) and the NMR

structure of PCI in solution has also been determined (Clore *et al.*, 1987); the rms deviation between the two for the core (residues 8 to 34) backbone atoms is 1.2 Å. The overall structure (see Figure 1) consists of N-terminal and C-terminal tails and a central globular core of 27 residues (Cys8 to Cys34) stabilized by three disulfide bonds (Cys8–Cys24, Cys12–Cys27 and Cys18–Cys34). In the crystal structure, there are five hydrogen bonds (9Asn–34CysO, 12CysN–32ArgO, 26AlaN–35GlyO, 34CysN–10LysO and 35GlyN–26AlaO) in the core plus two additional hydrogen bonds (8CysN–5AspO and 10LysN–7IleO) that stabilize two reverse turns outside the core. Only the four latter core hydrogen bonds are identified in the NMR structure. Although, the crystal and solution structures are very similar, there is one region of the protein where significant differences occur. The crystal structure has a 3_{10} helix, Thr14 through Cys18, which is generated by two consecutive type III reverse turns (Rees & Lipscomb, 1982); the solution structure has a small α-helix involving Asp16 through Ala21. No helix (1–4) hydrogen bonds were present in the crystal structure, while in the NMR structure, there are two hydrogen bonds involving residues of the helix (19SerN–16AspO and 21AlaN–18CysO). The three loops are between Cys8 and Cys12, Cys12 and Cys24 and between Cys27 and Cys34. The C-terminal region (Gly35 to Gly39) makes the primary contacts with carboxypeptidase A (CPA) in the inhibited complex (Rees & Lipscomb, 1982); a secondary contact region involves residues from Cys27 to Ser30 and the side-chain of residue His15. In the NMR structure, the two terminal ends are not well defined because of a lack of NOE data, indicating that these regions are flexible in solution. There is no known function associated with the N-terminal region (Glu1 to Ile7). The N and C-terminal tails are attached to the core by proline residues (Pro6 and Pro36).
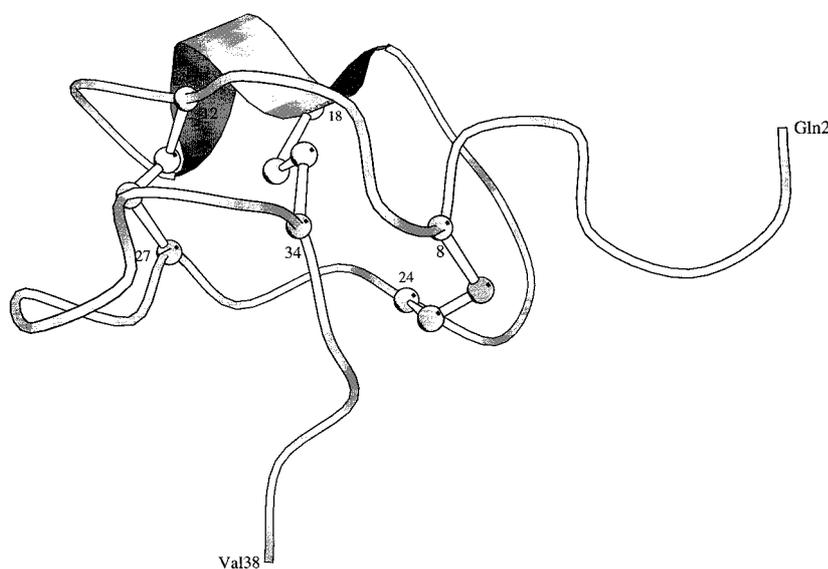

Figure 1. A ribbon representation of the PCI crystal structure (pdb4cpa.ent) (Rees & Lipscomb, 1982) highlighting the helix 3_{10} and the disulfide bonds between cysteine residues. Protein regions: loop 1 from Cys8 to Cys12, helix 3_{10} from Cys12 to Cys18, loop 2 from Cys18 and Cys24 and loop 3 from Cys27 to Cys34. Figure generated using the program MOL-SCRIPT (Kraulis, 1991).

Experimental studies performed on the disulfide folding pathways of PCI by one of our groups (Chang *et al.*, 1994) showed that the reduced and denatured state of this protein refolds spontaneously *in vitro* under oxidizing conditions. Stop/go folding studies and structural characterization by mass spectrometry showed that the disulfide pathway proceeds through an initial stage of non-specific disulfide formation, leading to the formation of scrambled three-disulfide species; a second stage involves disulfide reshuffling to reach the native conformation. Significant populations of five scrambled disulfide species have been found and isolated; three of these are being characterized by NMR solution studies (M. Rico, private communication). Recently, experimental studies on similar proteins have appeared; these include the study of the unfolding and refolding of cardiotoxin III (Chang *et al.*, 1998), a small protein containing four disulfide bonds, and the structure determination of three disulfide isomers of α-conotoxin (Gehrmann *et al.*, 1998), a three-disulfide bond protein. Moreover, proteins with non-native disulfide bonds are of interest because they appear to be involved in the trapping as inclusion bodies in bacterial host systems (Marston, 1986; Speed *et al.*, 1995, 1997).

Here, molecular dynamics simulations in solution with explicit solvent are used to study refolding of PCI with disulfide bond constraints. Molecular dynamics simulations of native PCI at room temperature have been performed previously with the GROMOS program (Oliva *et al.*, 1995; Daura *et al.*, 1996). The study by Oliva and coworkers showed that the core of the proteins was stable and the terminal ends were flexible; the N terminus was more flexible than the C terminus. We first performed a simulation of 500 ps at 300 K of PCI solvated by 1985 water molecules in a stochastic boundary system (Brünger *et al.*, 1984; Brooks *et al.*, 1985; Brooks & Karplus, 1989) with the CHARMM program (Brooks *et al.*, 1983). When compared with the experimental structures and with the simulations of Oliva *et al.* (1995) and Daura *et al.* (1996), the results showed that PCI was stable and well behaved with the all-atom CHARMM22 potential energy function (MacKerell *et al.*, 1998). Unfolded conformers of PCI were generated by a 700 ps simulation at 600 K with broken disulfide bonds (ND600). For comparison, simulations were done at 600 K with the three disulfide bonds present (N600; 610 ps) and at 300 K with broken disulfide bonds (ND300; 625 ps). Starting with different structures obtained from the ND600 simulation, a set of simulations explored the possibility of refolding PCI by introducing disulfide bond constraints. In the refolding simulations, NOE-type restraints between two cysteine sulfur atoms were used to slowly bring together the cysteine residues that form disulfide bonds in the native structure; in most cases the disulfide bonds were introduced when the sulfur atoms of the native pairs were relatively close to each other.

Since there are many proteins where disulfide pairing is known prior to a structure determination, this represents a realistic approach to a limited folding problem. The resulting structures were characterized with primary emphasis on the loop structures, which are an essential element of the folded protein. Most of the refolded structures are relatively close to native (1.45 to 3.13 Å rms) and a variety of approaches were tried to pick out the best structures (in terms of rms) without reference to experiment. No satisfactory criterion was found. The present test of potential functions and statistical methods is much stricter than the original ''threading'' study of misfolded proteins by Novotny *et al*. (1984) and many subsequent applications of threading to pick out the native structures from others that have very different folds (see, for example, Jones, 1996). It supplements recent studies (see e.g. Park & Levitt, 1996) in that it attempts to distinguish structures all of which are close to native in terms of rms.

We first present the methods used in the simulations and then describe the results. The analysis shows that reasonable refolding can be achieved in this small protein with known disulfide bonds. The protocols employed here are being applied to predict the structures of PCI with scrambled disulfide bonds (M.A.M.-R. *et al.*, unpublished).

Methods

Simulation protocols

The X-ray coordinates for PCI were obtained from the Brookhaven Protein Data Bank file 4cpa (Bernstein *et al.*, 1977; Abola *et al.*, 1987; Abola *et al.*, 1997). This file contains the coordinates for PCI complexed with carboxypeptidase A (Rees & Lipscomb, 1982). The coordinates were used as the initial structure for the native protein simulation (N300). Residues Glu1 and Gly39 from PCI were excluded; Glu1 is not well defined in the structure and it is absent in other active isoforms of PCI, while Gly39 is not necessary for the inhibitory activity of the protein (Hass & Ryan, 1980) and appears spatially separated from PCI in the coordinate file as if it had been cleaved by CPA when the complex with PCI was formed.

Hydrogen coordinates were constructed using the HBUILD facility (Brünger & Karplus, 1988) in the CHARMM program (Brooks *et al.*, 1983). The CHARMM22 all-atom potential energy function was used (MacKerell *et al.*, 1998). The simulations were done with the stochastic boundary molecular dynamics (SBMD) method (Brünger *et al.*, 1984; Brooks & Karplus, 1989). The protein (residues Gln2 to Val38) consists of 533 atoms. It was solvated in a 25 Å radius sphere of water by a total of 1985 modified TIP3 water molecules (Jorgensen *et al.*, 1983; Neria *et al.*, 1996). The entire protein was treated by molecular dynamics; the stochastic boundary region was limited to the 3 Å shell of water at the boundary. The boundary potential

was applied only to the water molecules; no constraints were imposed on the protein. Two simulations were done at room temperature (300 K) and two at 600 K. High temperatures were used to permit the protein to pass over the energetic barrier to the unfolded state in a reasonable amount of computer time, the temperature of 600 K is the same as that used in the simulations of Caflisch & Karplus (1994, 1995).

To prepare the crystal structure for the simulations, the isolated protein was energy minimized with 1000 steps of Steepest Descent (SD) minimization (Levitt & Lifson, 1969; Brooks *et al.*, 1983); a distance (*r*)-dependent dielectric function (Gelin & Karplus, 1977) with ε equal to 4 was used; the function 4*r* has been suggested by Teeter (Whitlow & Teeter, 1986). After solvating the proteins, a constant dielectric with ε equal to 1 was employed in all simulations. To equilibrate the water around the fixed protein, 2.5 ps of SBMD at 1000 K was done, followed by 100 steps of SD minimization and 6 ps of SBMD at 300 K. All constraints were then removed from the protein and the simulation was run for 500 ps at 300 K. The structure of PCI at 210 ps from this simulation (N300) was used as the starting structure for the unfolding simulations. For the ND300 simulation, the disulfide bonds were broken by removing all bond, angle and dihedral terms associated with the S–S bonds between two Cys residues. The van der Waals radii and charges of the S atoms were not changed. No hydrogen atoms were introduced so as to minimize the van der Waals repulsion between the two cysteine residues when breaking the disulfide bond. The system was then energy minimized with 300 steps of SD followed by 25 ps of SBMD at 300 K. The simulation was continued for 625 ps.

The N600 and ND600 simulations were started from the same point (i.e. 210 ps) of the N300 simulation as ND300. For the N600 simulation, the system was heated from 300 K to 600 K over 10 ps of SBMD; the simulation was run for 610 ps at 600 K. In the ND600 simulation, the disulfide bonds were broken as described above and the system was heated to 600 K during 10 ps of SBMD. After the 300 ps of simulation at 600 K, some atoms from Glu2 residue extended beyond the surface of the water sphere. A second layer of water was added to increase the size of the water sphere to a radius of 37 Å containing 6556 water molecules. The simulation was restarted and continued for another 400 ps.

During the ND600 simulation, the distances between the C^α atoms and the sulfur atoms of the Cys residues were calculated; six structures were selected, at 319, 358, 474, 540, 725 and 820 ps, where the distances between two of the three native pairs were relatively close to the native values. These structures were used as starting points for the refolding simulations with reformation of the native disulfide bonds. Four different methods were used to construct nine refolded structures. In method A (simulations A319, A358,

A474, A530, A725 and A820) NOE-type restraints (Brooks *et al.*, 1983; B. R. Brooks, unpublished) between native pairs of Cys sulfur atoms were introduced with the equilibrium distances equal to the current distances; in many of the cases studied here the distance between the two sulfur atoms was 10 Å or greater (see Table 2). Restrained molecular dynamics were run at 600 K, in intervals of 10 ps. At the beginning of every 10 ps interval, the equilibrium distance in the NOE-type restraint was reduced, slowly forcing the Cys groups together. During the 10 ps interval, one structure was extracted every 1 ps and the distances between the S–S pairs were checked. The structure with the shortest distances was used as the starting structure in the next 10 ps interval. When the S–S distances were less than 2.9 Å, disulfide bonds were re-introduced. After equilibrating the new systems for 25 ps at 600 K, the systems were cooled from 600 to 300 K over 10 ps and the simulations were continued for 200 ps. Methodology B is essentially the same as methodology A, but the interval for restrained dynamics was 20 ps instead of 10 ps and rather than searching over the interval for structures where the native S–S pairs were closest, the last structure from the interval was used as the starting structure for the next 20 ps interval of restrained dynamics. This methodology was used in simulation B358. Method C was applied in the one simulation (C474). It differs from methods A and B in that the NOE restraint was applied only to the native S–S pair with the shortest distance; the system was then cooled from 600 to 300 K over 10 ps. The simulation was restarted at 300 K and the distances between sulfur and C^α atoms were calculated. During the simulation, if a native cysteine pair was close in distance (less than 2.9 Å for the S–S distance), the disulfide bridge was constructed and the simulation was restarted from this point. The two disulfide bridges were introduced at 878 and 1003 ps respectively; the simulation for PCI with all three S–S bonds was continued for 200 ps. In method D, the simulation (D810) was started from the last structure of the N600 simulation (810 ps); in this case the native disulfide bonds were intact. This structure was cooled from 600 to 300 K as described above and the simulation was continued for another 200 ps.

In all simulations the time step was 2 fs and the temperature was controlled by the Langevin heat bath (Berendsen, 1984). Table 1 summarizes the different protocols (see also Figure 2). Table 2 gives the distances between C^α and SG atoms when the disulfide bridges were introduced. The distances between C^α and SG atoms for the disulfide bonds after refolding are within the expected distances (Thornton, 1981).

Analysis of the simulations

For the control and unfolding simulations, the potential energy time series and the van der Waals and electrostatic energy time series were deter-

Table 1. Principal properties of the simulations

	Initial structure	Temperature (K)	Disulfide bridges	Simulation time
Control at 300 K (N300)	X-ray	300	Yes	500 ps
Reduced PCI at 300 K (ND300)	210 ps from N300	300	No	835 ps
Native PCI at 600 K (N600)	210 ps from N300	600	Yes	820 ps
Reduced PCI at 600 K (ND600)	210 ps from N300	600	No	910 ps
Refolded PCI at 300 K (A319)	319 ps from ND600	300	Yes	200 ps
Refolded PCI at 300 K (A358)	358 ps from ND600	300	Yes	200 ps
Refolded PCI at 300 K (A474)	474 ps from ND600	300	Yes	200 ps
Refolded PCI at 300 K (A540)	540 ps from ND600	300	Yes	200 ps
Refolded PCI at 300 K (A725)	725 ps from ND600	300	Yes	200 ps
Refolded PCI at 300 K (A820)	820 ps from ND600	300	Yes	200 ps
Refolded PCI at 300 K (B358)	358 ps from ND600	300	Yes	200 ps
Refolded PCI at 300 K (C474)	474 ps from ND600	300	Yes	200 ps
Refolded PCI at 300 K (D810)	Cooling from N600	300	Yes	200 ps
Misfolded PCI at 300 K	1175 ps from ND600	300	Yes	200 ps

The simulations are defined by a letter and a number; in the control and unfolding simulations, the letter refers to the state of the protein (with or without disulfide bonds) and the number to the temperature, while in the refolding simulations, the letter refers to the methodology used and the number to the time in picoseconds from the starting structure. See the text.

mined; the average energy and its fluctuations were calculated. The root mean square (rms) deviations between the crystal structure and the average structure from the last 50 ps of each simulation (25 ps for ND600 simulation) were calculated as were the rms fluctuations for the protein core (from Cys8 to Cys34); these were compared to the rms fluctuations determined from the crystallographic B factors. Root mean square deviations averaged over 10 ps intervals were plotted as a function of time. To assess further the global deformations of PCI, the average radius of gyration (R_g) and radius of gyration time series were calculated.

The solvent accessible surface area (SASA) was determined using the algorithm of Lee & Richards (1971) implemented in CHARMM; a water radius of 1.4 Å was used.

For the folding simulations, the overall rms deviations for the protein core from the experimental structures (crystal and NMR) and the average structure from the control simulation (N300) were calculated. The rms deviation on a residue-by-residue basis was calculated between the average refolded structures and the average N300 structure. The radius of gyration and SASA for each refolded structure were calculated and compared with the
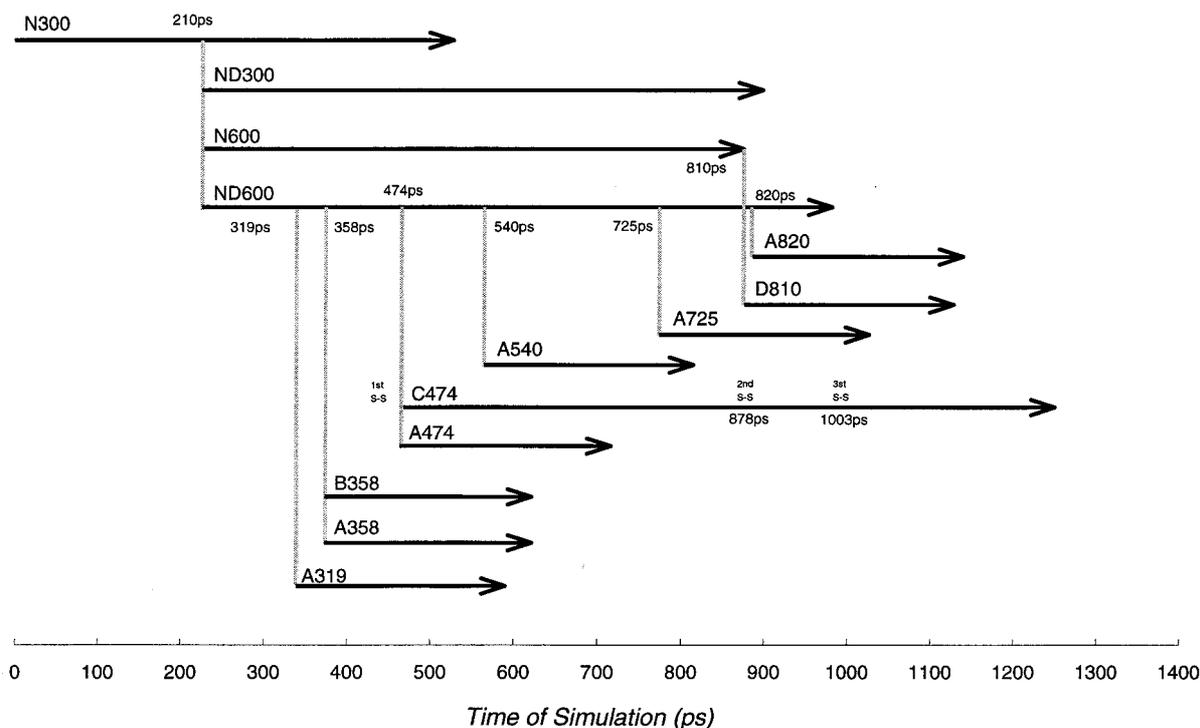

Figure 2. Simulation protocols. Horizontal lines indicate the simulations performed and the time scale. Vertical lines indicate origin of the starting structures for each simulation. The times of disulfide bridge formation have been marked in the C474 simulation.

Table 2. Original and final distances between C$^{\alpha}$ and SG atoms in refolding simulations

Original structure	Disulfide bridge	Before constraints		After constraints[a]		After construction[a]	
		C$^{\alpha}$ atoms	SG atoms	C$^{\alpha}$ atoms	SG atoms	C$^{\alpha}$ atoms[b]	SG atoms
319 ps	Cys8–Cys24	6.40	5.58	6.31	2.85	6.02	2.00
	Cys12–Cys27	5.80	4.88	6.76	2.86	6.34	1.94
	Cys18–Cys34	7.04	4.47	8.06	2.87	6.46	2.00
358 ps	Cys8–Cys24	7.14	4.19	6.29/8.17	2.78/2.87	6.05/5.74	2.00/1.99
	Cys12–Cys27	6.87	7.65	7.93/7.01	2.85/2.83	5.43/5.44	1.99/1.99
	Cys18–Cys34	7.71	3.77	7.24/8.47	2.86/2.85	6.16/6.33	2.00/2.01
474 ps	Cys8–Cys24	14.45	9.19	6.66/6.06	2.76/2.09	6.17/6.77	2.01/2.01
	Cys12–Cys27	8.95	10.98	5.82/7.29	2.81/2.85	6.31/6.34	1.99/1.90
	Cys18–Cys34	8.23	3.40	7.93/6.57	2.74/1.97	6.59/6.62	2.02/2.03
540 ps	Cys8–Cys24	9.70	9.96	8.35	2.86	6.32	2.01
	Cys12–Cys27	6.61	4.48	7.05	2.87	6.80	1.98
	Cys18–Cys34	15.21	13.22	8.06	2.77	6.40	2.01
725 ps	Cys8–Cys24	12.07	11.71	8.08	2.85	6.33	1.99
	Cys12–Cys27	5.17	7.08	7.08	2.85	6.23	2.02
	Cys18–Cys34	14.93	15.01	7.84	2.78	5.92	2.00
820 pa	Cys8–Cys24	12.12	13.27	8.35	2.79	6.38	1.98
	Cys12–Cys27	6.97	6.48	7.44	2.83	6.32	1.99
	Cys18–Cys34	15.13	10.90	8.32	2.83	6.18	2.02

All distances are given in Å.
[a] When two simulations were performed from the same original structure, the equilibrium distances are separated by a slash character; for example, for the original structure at 358 ps, the results for A358 and B358 are given, respectively. For the original structure at 474 ps, A474 and C474 results are given, respectively. The after constraints distances for C474 simulation correspond to the distances after construction of the third disulfide bond at 1003 ps.
[b] Distances between C$^{\alpha}$ range from 4.6 to 7.4 Å (Thornton, 1981).

average N300 structure and the experimental structures. Root mean square deviations between the initial structure from which the folding simulation was started and the average folded structure were also obtained; this assesses the degree of structural change that occurred during the folding process. Since one of the important structural elements of the T-knottin proteins are the loops, their internal and global structures were compared in detail.

An analysis of hydrogen bonds was made for the control, the unfolding and the refolding simulations. For the control and unfolding simulations, hydrogen bonds between the solvent and both the backbone and side-chain atoms of the proteins, as well as the intramolecular hydrogen bonds of the protein, were examined as a function of time. The criterion for a hydrogen bond was that the distance between the donor and the acceptor atom was less than 3.0 Å and that the donor-H-acceptor angle was greater than 90° (Baker & Hubbard, 1984). With this criterion, the percentage of hydrogen bonds maintained during simulations relative to the hydrogen bonds in the X-ray structure were calculated. For the folding simulation, backbone–backbone and backbone–side-chain hydrogen bonds which appeared for 20% or more of the final 200 ps of the simulation were tabulated.

To try and discriminate between the crystal structure, the NMR structures and the different structures obtained from simulations, two empirical solvation models for proteins were used. The model of Lüthy et al. (1992), available through the web site http://www.doe-mbi.ucla.edu/Services/Verify3D.html, assumes that the total solvation free energy is a sum of contributions from the constituent groups of the protein. The model of Koehl and Delarue (KD) generalizes this model by including atom–atom contact distributions (Koehl & Delarue, 1994). Parameters for the models, as given in the references, were determined from experimental measurements of hydration or hydrophobicity of amino acids and/or analogs. Both models were used to generate 3D profiles of the various PCI structures. A different approach to the characterization of proteins uses an effective energy function based on the CHARMM polar hydrogen force field (Neria et al., 1996) combined with a solvent exclusion model for the solvation free energy (Lazaridis & Karplus, 1997). In this model, which has been shown to work well in threading tests (T. Lazaridis & M.K, unpublished), the internal energy components are considered as well as the non-bonded terms. Energies calculated using this model were applied to the different structures.

Results

This section is divided in two parts. We first describe the room temperature control simulation and the unfolding simulations. In the second part, the results from the refolding simulations are presented and analyzed.

Room temperature and high temperature unfolding simulations

Two room temperature simulations were made; the first (N300) had the three disulfide bonds of the native structure and the second (ND300) had the disulfides broken as described in Methods. The

Table 3. Overall properties as a function of simulation conditions

Property	Crystal structure	N300 simulation	ND300 simulation	N600 simulation	ND600 simulation
Potential Energy[a]	–	-1439 ± 39	-1515 ± 56	-697 ± 61	-803 ± 123
PCI Potential E.[a]	–	182 ± 28	220 ± 39	526 ± 42	639 ± 67
PCI-Water Potential E[a]	–	-1621 ± 50	-1735 ± 79	-1223 ± 71	-1451 ± 160
PCI van der Waals E.[a]	–	-85 ± 9	-75 ± 8	-48 ± 14	-22 ± 14
PCI-Water van der Waals E.[a]	–	-135 ± 15	-141 ± 15	-110 ± 17	-116 ± 24
PCI electrostatic E.[a]	–	-372 ± 25	-312 ± 33	-424 ± 34	-303 ± 65
PCI-Water electrostatic E.[a]	–	-1485 ± 55	-1594 ± 80	-1113 ± 70	-1302 ± 178
Backbone r.m.s.d[b]	–	1.686 Å	2.936 Å	3.083 Å	5.847 Å
Backbone r.m.s.f.[c]	–	0.989 Å	0.950 Å	1.761 Å	2.544 Å
Radius of gyration[d]	8.409 Å	8.963 Å	8.840 Å	8.962 Å	9.057 Å
S.A.S.A.[e]	0.000 (1929.39 Å^2)[f]	150.37 Å^2	247.42 Å^2	-0.62 Å^2	472.84 Å^2

All energies are in kcal/mol.

[a] Mean and standard deviation of total and PCI-water interaction potential energy and PCI van der Waals energy.

[b] The backbone rms deviation from the crystal structure for core residues (8Cys to 34Cys) averaged over the last 50 ps; the ND600 simulation average is over the last 25 ps.

[c] The backbone atoms rms fluctuation about the average structure for last 50 ps; for the ND600 simulation it is over the last 25 ps.

[d] The radius of gyration for core residues (8Cys to 34Cys) averaged over the last 50 ps (except for the ND600 simulation over the last 25 ps).

[e] Solvent accessible surface area (SASA) for core residues (8Cys to 34Cys) averaged over the last 50 ps (except ND600 simulation over last 25 ps) relative to the crystal structure value.

[f] Original crystal structure value of SASA.

two corresponding simulations at 600 K are labelled N600 and ND600, respectively. The behavior of the N300 simulation, which serves as a control, is described first and the three perturbed systems (broken S–S bonds, higher temperature or both) are considered subsequently in terms of their differences from N300. The primary purpose of these simulations is to generate a collection of unfolded structures (N600 and ND600) as well as to have control simulations (N300 and ND300) for comparison. However, their analysis is of interest for a comparison with other unfolding simulations (Daggett & Levitt, 1992a,b, 1993a,b; Mark & van Gunsteren, 1992; Vishveshwara *et al.*, 1993; Tirado-Rives & Jorgensen, 1993; Li & Daggett, 1994; Caflisch & Karplus, 1994, 1995; Alonso & Daggett, 1995).

N300 (disulfide bonds intact, 300 K)

After the first 20 ps, which involves reheating of the system to 300 K from the energy minimised structure (see Methods), the total potential energy remains essentially constant, indicating that the simulation is well equilibrated; the PCI-self and PCI-water potential energies are also stable, in accord with the fact that the protein structure is well preserved. The van der Waals and electrostatic energy time series, respectively, show the same behavior as the other parts of the potential energy. In Table 3, the average energies and fluctuations are given for all simulations.

The rms deviations of the protein core as a function of time averaged over 10 ps intervals is shown in Figure 3(a). The N300 simulation is structurally stable during the entire simulation. Table 3 also shows the rms deviations for the core backbone atoms for the average structures from all simulations with respect to the crystal structure. Figure 4(a) shows the rms deviation as a function

of residue number. The N-terminal end of the protein is the most mobile part, having the largest deviations in all simulations. The largest deviations in the core of the protein occur in the regions between cysteine residues 12 and 24 (see Figure 1); this region, comprising the helix 3_{10} and loop 2,
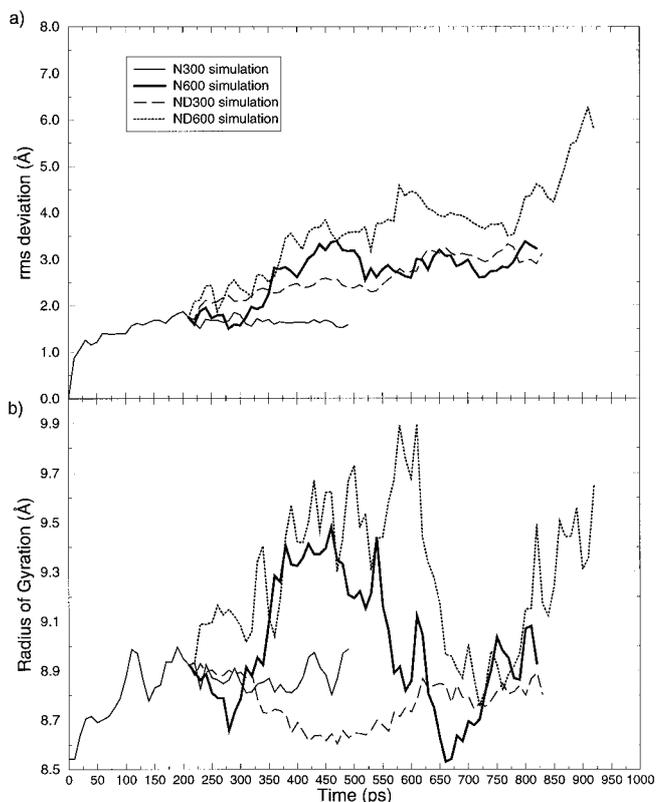

Figure 3. Representation of rms deviation and radius of gyration as a function of time for every simulation. (a) Rms deviation for the backbone atoms from the core of the protein. (b) Radius of gyration for the atoms from the core of the protein.
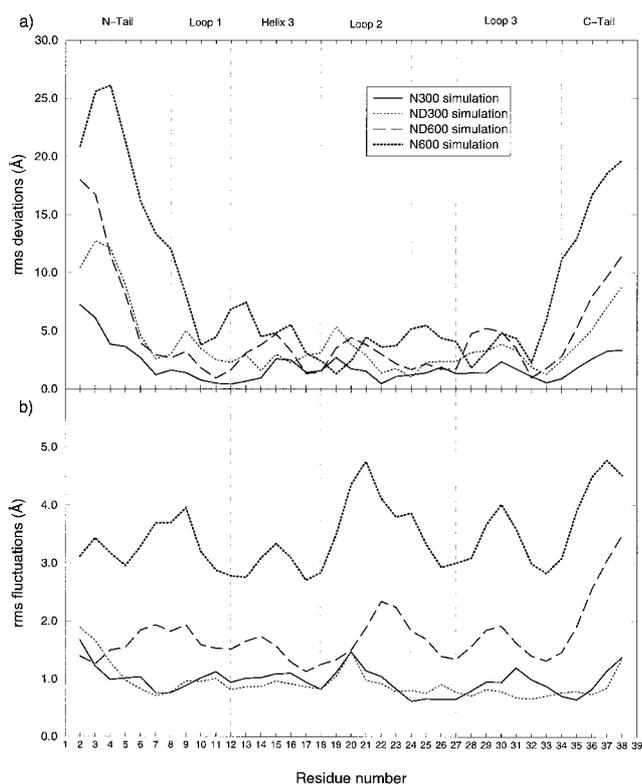
Figure 4. Simulation properties as a function of residue number. (a) Backbone rms deviations of the average structures over the last 50 ps of every simulation (25 ps for ND600 simulation) from crystal structure. (b) Backbone rms fluctuations of the average structures. Cysteine positions are marked by the vertical dotted lines.

exhibits larger rms fluctuations than the other loop regions and region between 24 and 27 in the N300 average structure. These results are in agreement with the simulations by Oliva *et al.* (1995) and Daura *et al.* (1996). Figure 4(b) shows the rms fluctuations as a function of residue number relative to the average structures. The N300 structure is most flexible at the N and C-terminal ends, in the loop 2 region (Cys18 to Cys24) and in the loop 3 region (Cys27 to Cys34). Although the fluctuations in the X-ray structure, calculated from the crystallographic *B* factors, are lower than in N300 average structure, the general pattern is similar.

The radii of gyration of the protein core calculated from the average structures are given in Table 3 and the radii values as a function of time (averaged over 10 ps intervals) are shown in Figure 3(b); the values for the crystal structure and mean NMR structure are 8.41 Å and 8.10 Å, respectively. The time evolution of the radius of gyration in the N300 simulation shows minor changes. During the first 100 ps of the simulation, the radius of gyration increased, in part, as a result of the protein solvation. After this, the radius of gyration stabilized around the 8.96 Å. This is in agreement with the behavior observed in the earlier simulations (Oliva *et al.*, 1995).

The solvent accessible surface area (SASA) for the atoms of the protein core for the N300 simulation with respect to the X-ray structure shows a small increase (8%) relative to the crystal structure. This is in accord with the behavior of the radius of gyration.

To characterize further the structure, a list of hydrogen bonds is presented in Table 4. It shows the hydrogen bonds in the X-ray and the average NMR structure and the percentage of time hydrogen bonds are present in the N300, as well as the refolded structures. Figure 5(a) and (b) shows the percentage of conserved crystal protein–protein hydrogen bonds averaged over 10 ps intervals and the percentage of the hydrogen bonds that are protein–protein hydrogen bonds, relative to the total number of PCI hydrogen bonds (including solvent) averaged over 10 ps intervals, respectively. In the crystal structure, there are nine main-chain hydrogen bonds and five main-chain–side-chain hydrogen bonds, while the NMR structure has eight and four, respectively. Only four main-chain and three main-chain–side-chain hydrogen bonds are present in both the crystal and NMR structures; of these, two main-chain hydrogen bonds (26AlaN–35GlyO and 35GlyN–26AlaO) are involved in the short β-strand interactions. This suggests that most of the protein hydrogen bonds are of marginal stability. This is in accord with the N300 simulation, which shows a sharp drop in the number of crystal hydrogen bonds during the first 50 ps (Figure 5(a)). The N300 simulation maintains the percentage of internal *versus* total hydrogen bonds at around 30% (Figure 5(b)). This indicates that, although the N300 simulation lost certain hydrogen bonds present in the crystal structure, other internal hydrogen bonds are formed in the protein core.

The hydrogen bonds that appear for more than 40% of the time during each 50 ps segment of the N300 simulation are shown in Table 5. The backbone hydrogen bonds that are present during most of the simulation are primarily those that were found in both the crystal and NMR structures; they are 12CysN–32ArgO, 26AlaN–35GlyO and 34CysN–10LysO. The only one that does not appear is 35GlyN–26AlaO; in the average structure the distance between the donor and acceptor increased to 3.47 Å, somewhat larger than the maximum distance used to define a hydrogen bond (see Methods). Of the hydrogen bonds that occur only in the crystal structure (8CysN–5AspO, 9AsnN–34CysO, 10LysN–7IleO, 23PheN–21AlaO and 30SerN–28TrpO), see Figure 6(a), only 9AsnN–34CysO and 10LysN–7IleO were present some of the time; 9AsnN–34CysO is quite stable. None of the hydrogen bonds present only in the NMR structure (19SerN–16AspO, 12AlaN–18CysO, 28TrpN–26AlaO, 31AlaN–28TrpO and 32ArgN–28TrpO) are present for a significant fraction of time in the simulation. There are a number of backbone hydrogen bonds formed in the simulation (19SerN–14AspO, 21AlaN–19SerO, 25GlnN–23PheO and 32ArgN–30SerO) that are

Table 4. Representative hydrogen bonds in PCI; comparison with refolding simulations

Donor	Acceptor	X-ray	NMR	N300	A319	A358	A474	A540	A725	A820	B358	C474	D810
				Backbone hydrogen bonds									
7IleN	5AspO	No	No	–	–	–	32.5	–	–	–	–	–	51.0
8CysN	5AspO	**Yes**	No	–	–	–	–	–	–	–	–	–	–
8CysN	6ProO	No	No	–	23.5	–	–	–	–	–	–	–	–
9AsnN	6ProO	No	No	–	23.0	–	–	–	–	–	–	–	–
9AsnN	7IleO	No	No	–	–	78.5	–	–	–	–	–	–	–
9AsnN	34CysO	**Yes**	No	50.2	–	–	–	–	–	–	–	–	–
10LysN	7IleO	**Yes**	No	24.7	–	–	–	–	–	–	–	–	–
10LysN	8CysO	No	No	–	31.0	–	–	81.5	–	–	69.0	–	–
12CysN	10LysO	No	No	–	33.0	–	–	–	–	–	93.5	–	–
12CysN	32ArgO	**Yes**	**Yes**	43.3	37.0	–	51.0	–	–	30.0	–	70.5	51.0
12CysN	33ThrO	No	No	–	–	–	–	–	31.0	–	–	–	–
16AspN	14ThrO	No	No	–	–	–	–	–	–	–	–	–	28.0
17AspN	13LysO	No	No	–	36.0	–	–	–	–	–	–	–	–
18CysN	11ProO	No	No	–	–	–	–	–	–	88.5	–	–	–
18CysN	13LysO	No	No	–	37.5	–	–	–	–	–	–	–	–
19SerN	16AspO	No	**Yes**	–	–	–	–	–	–	–	–	–	–
19SerN	17AspO	No	No	27.6	–	–	–	–	–	–	–	–	–
20GlyN	18CysO	No	No	–	–	–	–	–	–	–	59.0	–	–
20GlyN	25GlnO	No	No	–	–	–	–	–	–	–	–	–	25.5
21AlaN	18CysO	No	**Yes**	–	–	–	–	–	–	–	–	–	–
21AlaN	19SerO	No	No	23.3	–	–	–	–	–	–	–	–	–
21AlaN	24CysO	No	No	–	–	38.0	28.5	26.5	30.0	81.5	–	–	71.0
22TrpN	20GlyO	No	No	–	–	–	–	–	–	–	27.5	–	–
23PheN	21AlaO	**Yes**	No	–	–	–	–	–	–	–	–	–	–
25GlnN	23PheO	No	No	38.4	–	–	44.5	21.5	–	–	–	–	–
25GlnN	35GlyO	No	No	–	54.0	79.0	–	–	–	–	–	–	–
26AlaN	35GlyO	**Yes**	**Yes**	79.3	–	–	–	–	–	–	93.5	–	–
28TrpN	26AlaO	No	**Yes**	–	–	–	–	–	–	–	–	–	–
28TrpN	33ThrO	No	No	–	–	54.5	–	–	–	–	–	34.0	–
30SerN	28TrpO	**Yes**	No	–	–	49.5	–	–	–	–	–	–	–
31AlaN	28TrpO	No	**Yes**	–	–	–	–	–	–	–	–	–	–
32ArgN	28TrpO	No	**Yes**	–	–	–	–	–	–	–	–	–	–
32ArgN	30SerO	No	No	20.6	–	–	–	–	–	–	–	–	–
33ThrN	31AlaO	No	No	–	43.0	–	61.5	–	–	–	21.5	30.0	–
34CysN	10LysO	**Yes**	**Yes**	82.6	–	–	–	44.5	–	57.0	53.0	–	40.0
35GlyN	26AlaO	**Yes**	**Yes**	–	–	–	39.5	–	–	–	56.0	–	–
				Backbone-side-chain hydrogen bonds									
7IleN	5AspOD1	**Yes**	**Yes**	–	–	–	–	–	–	–	–	–	–
9AsnND2	34CysO	No	No	–	–	–	–	–	–	–	56.5	–	–
10LysNZ1/2/3[a]	17AspO	**Yes**	**Yes**	28.1	28.0	–	–	33.5/23.5	–	–	–	–	32/29.5/27
13LysN	17AspOD1/2[a]	No	**Yes**	84.8	82.0	–	96.5	–	–	91.0	–	44.0	66.0/33.0
14ThrN	16AspOD	No	No	65.4	–	–	–	–	38.5	–	–	–	–
14ThrN	17AspOD	No	No	–	–	–	–	–	–	93.5	–	–	–
15HisN	14ThrOD	No	No	–	–	–	83.5	59.0	78.5	–	70.5	42.5	32.5
17AspN	16AspOD2	No	No	–	–	–	–	–	62.0	–	–	–	–
20GlyN	19SerOG	No	No	59.1	–	–	–	59.5	55.5	45.5	–	–	31.5
22TrpNE1	3HisO	No	No	44.1	–	–	–	–	–	–	–	–	–
22TrpNE1	4AlaO	**Yes**	**Yes**	–	–	–	–	–	–	–	–	–	–
22TrpN	5AspOD2	No	No	20.6	–	–	–	–	–	–	–	–	–
23PheN	5AspOD2	**Yes**	No	25.9	–	–	–	–	–	–	–	–	–
25GlnNE2	21AlaO	No	No	–	–	–	–	22.7	31.0	–	–	–	–
25GlnNE2	22TrpO	**Yes**	No	–	–	–	–	–	–	–	–	–	–
28TrpNE1	36ProO	No	No	–	–	–	–	–	–	–	25.5	–	–
30SerN	29AsnOD	No	No	–	–	–	–	40.0	–	–	29.0	–	–
31AlaN	30SerOG	No	No	72.7	80.0	54.0	–	–	71.5	–	56.0	39.0	49.5
32ArgNH2	11ProO	No	No	–	81.5	89.5	83.5	–	30.5	–	50.0	–	–
32ArgNE	12CysO	No	No	–	55.5	89.5	32.0	90.5	–	–	–	–	–
32ArgNH1/2[a]	12CysO	No	No	37.0	–	–	–	–	–	–	–	–	61.5
32ArgNE	13LysO	No	No	49.5	–	–	–	–	–	–	–	–	29.0
32ArgNH1/2[a]	14ThrO	No	No	–	–	–	–	–	–	–	–	–	94.5/42.1

All bonds with more than 20% of appearance are shown. X-ray or NMR structures are written in the first and second column, respectively.

[a] A switching between the atoms marked was found. The percentage of appearance is separated by slash character.
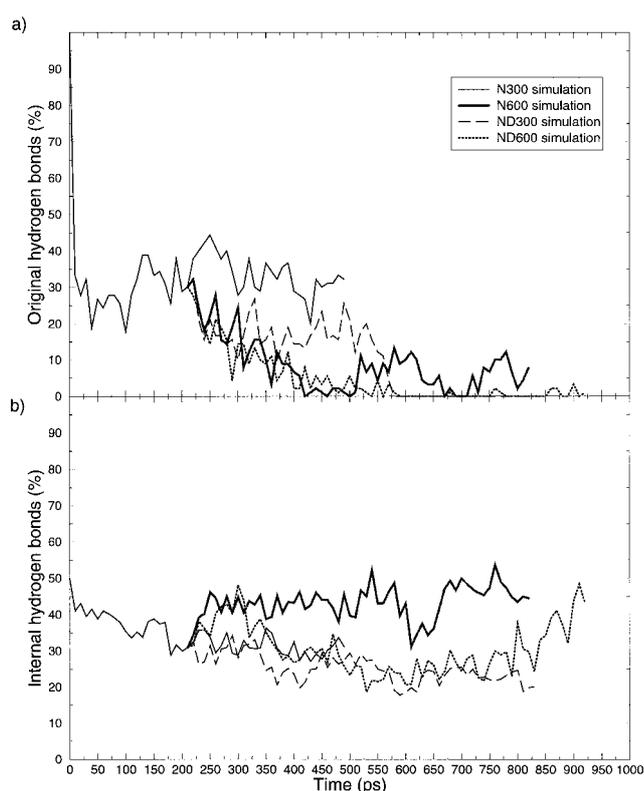
a)

b)

Figure 5. Representation of percentage of hydrogen bonds as a function of time for every simulation. (a) Percentage of maintenance of the original crystal hydrogen bonds. (b) Percentage of internal hydrogen bonds over the total hydrogen bonds.

observed neither in the crystal nor the NMR structure (see Table 4). The most stable hydrogen bonds observed in N300 simulations are those presented in Table 5 and Figure 6(b). Hydrogen bonds formed between side-chain atoms and backbone atoms are: 13LysN–17AspO$^{\delta_2}$, 14ThrN–16AspO$^{\delta_2}$, 20GlyN–19SerO$^{\gamma}$, 22TrpN$^{\epsilon_1}$–3HisO, 31AlaN–30Ser$^{\gamma}$ and 32ArgN$^{\epsilon}$–13LysO. In previous simulations of PCI (Oliva *et al.*, 1995; Daura *et al.*, 1996) stable backbone hydrogen bonds were found between 12CysN–32ArgO, 26AlaN–35GlyO, 34CysN–10LysO and 35GlyN–26AlaO. It is of interest that the hydrogen bonds present in the crystal and NMR structures are sequentially distant for the most part, suggesting that they have a role in maintaining the tertiary structure of the protein. In the simulation, a number of experimental protein–protein hydrogen bonds were replaced by protein–water hydrogen bonds; examples are: 10LysN–7IleO was replaced by 10LysN–782H$_2$O for 44.9% of the time, 28TrpN–26AlaO was replaced by 28TrpN–332H$_2$O for 59.3% and 18CysN–15HisO was replaced by 18CysN–1503H$_2$O for 45.5%.

ND300 (disulfide bonds broken, 300 K)

The total potential energy for the ND300 simulation is essentially the same as that for the N300 simulation. A comparison of the PCI-self energy and the PCI-water interaction energy shows that increases in the PCI-self potential energy are compensated by a decrease of the PCI solvent interaction potential energy. This is mainly an electrostatic effect; there is little change in the van der Waals energy. These changes in energy are correlated with the changes in the radius of gyration discussed below. The mean values of the potential energy (Table 3), show that there are stronger interactions with water molecules and an overall lower potential energy value than when the disulfide bonds were intact. The PCI-self potential energy is higher than for N300 simulation and is compensated by a more favorable PCI-water interaction potential energy.

The ND300 structure has a larger rms deviation from the crystal structure than N300 (see Table 3). The rms deviation of the protein core (residues Cys8 to Cys34) as a function of time is given in Figure 3(a). It shows that upon cleavage of the disulfide bonds, the protein slowly moves away from the crystal structure. The rms deviation from the crystal structure reaches a first plateau (rms of 2.4 Å) at about 350 ps and remains there until 550 ps. When the rms begins to increase again, it reaches a second plateau (rms of 3.1 Å) at 650 ps where it remains until the end of the simulation. The radius of gyration shows more complex behavior (Figure 3(b)). It decreases until 550 ps and then increases until it reaches a plateau at 650 ps; the value of R_g over the last 50 ps is slightly smaller (8.84 Å) than for the N300 simulation (9.96 Å). Figure 4 shows the average rms deviation and fluctuation as a function of residue number for the average structure of the final 50 ps of the simulation. The major structural changes involve the N and C tails where the rms deviation reaches more than 9 Å from the crystal structure. In the protein core, as in the N300 simulation, the three loops between the cysteine residues have large deviations. The cysteine residues remain relatively close to their crystal structure positions. However, the final S–S distances for the three broken disulfides are 4.3 Å (8CysSG–24CysSG), 5.6 Å (12CysSG–27CysSG) and 3.8 Å (18CysSG–34CysSG); significantly larger than for an S–S bond (2 Å).

The SASA of the average structure (Table 3) has increased significantly with respect to the N300 value. This is somewhat surprising since R_g is smaller (see above and Table 3) suggesting that a change in shape is involved. That the protein-solvent interaction energy has decreased (become more negative) is in accord with the surface area result. As for N300, there is a loss of the original hydrogen bonds, and the percentage of internal hydrogen bonds is smaller, see Figure 5(a) and (b); i.e. the number of internal hydrogen bonds

Table 5. Representative hydrogen bonds in N300 simulation

Backbone hydrogen bonds												
Atom type		Simulation time period (ps)										
Donor	Acceptor	0 50	50 100	100 150	150 200	200 250	250 300	300 350	350 400	400 450	450 500	%
9 AsnN	**34CysO**											50.2
10LysN	**7IleO**											24.7
12CysN	**32ArgO**											43.3
19SerN	17AspO											27.6
21AlaN	19SerO											23.3
25GlnN	23PheO											38.4
26AlaN	**35GlyO**											79.3
32ArgN	30SerO											20.6
34CysN	**10LysO**											82.6

Backbone−side-chain hydrogen bonds												
Atom type		Simulation time period (ps)										
Donor	Acceptor	0 50	50 100	100 150	150 200	200 250	250 300	300 350	350 400	400 450	450 500	%
10LysNZ	**17AspO**											28.1
13LysN	**17AspOD2**											84.8
14ThrN	16AspOD2											65.4
20GlyN	19SerOG											59.1
22TrpNE1	3HisO											44.1
22TrpN	5AspOD2											20.6
23PheN	**5AspOD2**											25.9
31AlaN	30SerOG											72.7
32ArgNH1	12CysO											37.0
32ArgNE	13LysO											49.4

All bonds with more than 20% of appearance are shown. The bonds that are present in the X-ray, NMR or in both structures are written in bold letters. The dark cells indicate that the hydrogen bonds appear in more than 40% of the trajectory in this period of time. The last column shows the appearance in percentage of trajectory over the total simulation time.

decreases and the number of protein solvent hydrogen bonds increases. Only two backbone hydrogen bonds exist for more than 20% of the time (10LysN–8CysO and 34CysN–10LysO). As already seen in the N300 simulation, most of the protein hydrogen bonds are only marginally stable. When no disulfide bonds are present to constrain the core structure, they rapidly disappear.
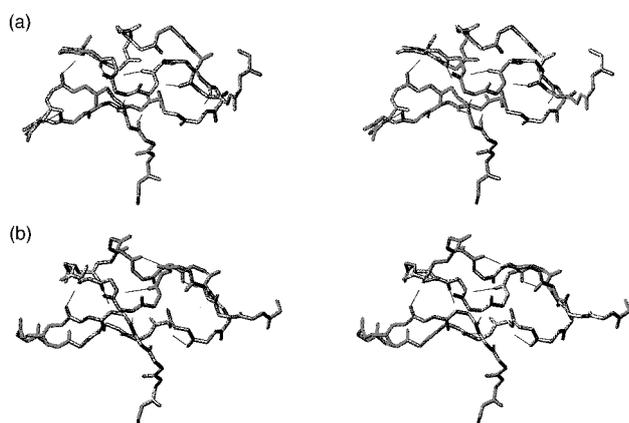
Figure 6. Stereo drawing of the most populated hydrogen bonds. (a) Crystal structure of PCI. (b) N300 simulation average structure.

N600 (disulfide bonds intact, 600 K)

In the N600 simulation, the rms deviation reaches a plateau value in about 350 ps (Figure 3(a)). It is slightly above 3 Å, significantly larger than that for the N300 simulation (1.7 Å). That a limiting value is reached is due to the constraints provided by the disulfide bonds. The largest structural changes occur at the two termini, which are free to move. In the core of the protein, the largest deviations are found in the three loops between cysteine residues, similar to the behavior of the N300 and ND300 simulations (see above). The R_g value fluctuates greatly during the simulation (Figure 3(b)). At 350 ps it reaches a value of 9.4 Å and then the value falls to less than 8.6 Å, below that of the N300 and ND300 simulations, and rises again during the last part of the trajectory. These changes take place without variation of the core rms deviation (see Figure 4(a)); i.e. they involve reorganization of the two termini.

The SASA of the N600 simulation is decreased relative to N300, showing that the structure becomes more compact. The SASA was calculated using the average structure of the last 50 ps of dynamics, where the radius of gyration of the N600 is slightly larger than the values of the N300 and ND300. When the SASA was calculated using a structure earlier in the simulation (with an R_g of

9.3 Å) the value was about 336 Å^3 larger than the value for the N300 average structure.

The protein hydrogen bonds show an initial behavior that is similar to that of the ND300 simulation. There is a rapid decrease in the percentage of crystal hydrogen bonds which drops to near zero in about 200 ps. This is again compensated by an increase in the number of alternative internal hydrogen bonds (see Figure 5). All the original backbone hydrogen bonds are lost (i.e. exist for less than 40% of the simulation time, though 12LysN–32ArgO and 34LysN–10LysO are present for 28.2 and 24.2%, respectively) and two backbone–side-chain hydrogen bonds are maintained over 40% of the time.

ND600 (disulfide bonds broken, 600 K)

In the ND600 simulation, the total potential energy increases as the temperature is raised from 300 to 600 K, as expected. However, as the ND600 results in more unfolding than the N600 simulation, the energies behave differently over the rest of the simulation. After the initial increase, the total potential energy of ND600 decreases relative to the N600 simulation due primarily to stronger PCI-water interactions in the former. In fact, the PCI self energy increases while the PCI-water energy decreases. This is due almost entirely to the electrostatic term. The total energy increases again toward the end of the simulation. During the ND600 simulation, the PCI-self van der Waals energy increased; similar behavior was observed in the unfolding simulations of barnase by Caflisch & Karplus (1995). This is related to the unpacking of the hydrophobic core of the protein and could contribute to the energetic barrier between the native and unfolded structures.

The average structure of the ND600 simulation has a deviation of 5.5 Å from the crystal structure indicating that PCI has undergone significant structural changes. The average structure was calculated from the last 25 ps of the simulation rather than the last 50 ps as in the other simulations because (see Figure 3) the protein underwent large structural changes during the last 50 ps. The three loops again were the most unstable regions of the protein core and exhibit larger fluctuations than in the other simulations.

Figure 3(b), gives the radius of gyration as a function of time and Table 3 the average value for the last 25 ps; the average value from ND600 differs significantly from the crystal structure value (9.06 Å *versus* 8.41 Å for the crystal structure). During most of the simulation, R_g is greater than 9.40 Å; however, between 660 and 790 ps, the value drops to under 9 Å. The SASA of the ND600 average increases by 473 Å^2 relative to the original value of 1929 Å^2 for the crystal structure.

During the simulation, there was a complete loss of original hydrogen bonds. The percentage of the internal hydrogen bonds decreased over 500 ps to 25%; during the last 100 ps there is an increase in

the number of internal hydrogen bonds to 40%, relative to the initial value of 31%. No backbone hydrogen bonds appear for more than 40% of the time and all the hydrogen bonds that appear over 20% of the time are between residues close in sequence. These hydrogen bonds are responsible for maintenance of local structure (secondary structures), some of which still exists for part of the time. The hydrogen bonds separated in sequence are all broken. The backbone–side-chain hydrogen bond, 13LysN–17AspO$^{\delta_2}$, is maintained in all simulations for 34 to 98% of the time. This indicates that a stable salt bridge is present.

Refolding simulations

Since the structural difference between the X-ray structure and the equilibrated N300 structure is small (rms deviation of 1.68 Å) and the latter is the structure that is stable for the choice of potential function, comparisons for refolding simulations are made with respect to the average and minimized average structures from the N300 simulation between 160 ps and 210 ps. The structure at 210 ps was the starting structure for all unfolding simulations. The difference between this structure and the average structure for the last 50 ps of the N300 simulations is 0.45 Å. After the refolding procedures (see Methods) had been applied and the proper three-disulfide species had been generated, the protein was simulated for 200 ps at 300 K. The results described below correspond to averages and minimized averages over the last 50 ps, unless stated otherwise.

Overall structural analysis

Table 6 shows the general properties for the average structures from all the refolding simulations. The conformations are structurally stable as demonstrated by the rms fluctuations which are in the range of 0.9 to 1.5 Å for the protein core. This is to be compared with the value of 0.99 Å for the N300 simulation. There is no correlation between the rms deviation from the N300 structure and the rms fluctuations. There is some indication that the refolding results depend on the method used. For the structures refolded at 358 ps, B358 has a smaller rms deviation from the N300 average structure than A358. Also, for the structures refolded at 474 ps (A474 and C474), the application of the NOE constraints to all three cysteine residues at the same time (A474) rather than one at a time (C474) gives better results for refolding. The rms deviation of the structure from D810, where the disulfide bonds were intact, remains essentially unchanged during refolding and does not recover the original conformation (see Table 7). This is probably due to the reduced conformational flexibility arising from the presence of the disulfide bonds so that the protein was rapidly trapped in a local minimum close to the starting structure. The overall rms values show that the structures fall

Table 6. Properties of N300 simulation and refolded PCI

Property	N300 simulation	A319 simulation	A358 simulation	A474 simulation	A540 simulation	A725 simulation	A820 simulation	B358 simulation	C474 simulation	D810 simulation
Backbone rmsd[a] (Å)	0.000	1.987	2.776	1.807	1.760	1.792	2.571	2.082	2.212	2.847
Backbone rmsd[a1] (Å)	1.686	2.720	3.126	2.705	2.476	2.437	2.998	2.602	2.996	3.198
Backbone rmsd[a2] (Å)	1.450	2.643	3.055	2.616	2.352	2.385	2.942	2.575	2.780	3.131
Backbone rmsf[b] (Å)	0.989	1.020	0.934	0.954	1.360	1.508	1.370	0.956	0.880	0.915
Radius of gyration[c] (Å)	8.963	8.870	9.030	9.196	9.399	8.954	9.133	8.404	9.578	8.930
SASA[d] (Å²)	2079.76[e]	227.65	269.48	360.42	213.46	116.67	258.12	1.65	422.71	35.41

[a] The backbone atom rms deviation from the N300 average structure for core residues (8Cys to 34Cys) averaged over the last 50 ps.
[a1] The backbone atom rms deviation from the crystal structure for core residues (8Cys to 34Cys) averaged over the last 50 ps.
[a2] The backbone atom rms deviation from the crystal structure for core residues (8Cys to 34Cys) of the averaged structure minimized by 300 steps using ABNR.
[b] The backbone atoms rms fluctuation about the average structure for last 50 ps.
[c] The radius gyration for core residues (8Cys to 34 Cys) averaged over the last 50 ps.
[d] Solvent accessible surface area (SASA) for core residues (8Cys to 34Cys) averaged over the last 50 ps relative to the N300 average structure value.
[e] Original N300 average structure values of SASA.

Table 7. Matrix table for backbone rms difference (in Å) between control structures and refolding simulations

Simulations	Crystal	NMR	N300	A319	A358	A474	A540	A725	A820	B358	C474	D810
Crystal	0.00	1.16	1.69	2.72	3.13	2.71	2.48	2.44	3.00	2.60	3.00	3.20
NMR	1.16	0.00	1.76	2.59	3.10	2.71	2.51	2.28	3.10	2.67	2.98	2.93
N300	1.69	1.76	1.69	1.99	2.78	1.81	1.76	1.79	2.57	2.08	2.21	2.85
A319	2.07	2.27	1.50	1.73	1.77	1.21	1.57	1.95	2.52	1.73	1.28	2.91
A358	2.57	2.69	1.90	1.28	2.28	2.14	2.15	2.20	2.16	2.12	1.50	3.56
A474	3.82	4.03	3.58	2.94	2.39	2.98	1.70	2.12	2.27	1.74	1.17	2.93
A540	3.89	4.03	3.27	2.91	2.82	3.53	2.93	1.69	2.21	1.79	1.45	2.40
A725	3.64	3.80	3.53	3.11	3.39	4.48	2.73	3.36	2.14	2.00	2.02	2.82
A820	4.44	4.63	4.03	3.61	3.61	4.41	2.57	2.52	3.28	2.35	1.93	3.43
B358	2.57	2.69	1.90	1.28	0.00	2.39	2.82	3.39	3.61	2.22	1.81	3.33
C474	3.82	4.03	3.58	2.94	2.39	0.00	3.53	4.48	4.41	2.39	2.81	2.91
D810	3.12	2.90	2.88	3.09	3.17	4.13	4.22	4.34	5.05	3.17	4.13	0.88

Values for rms differences between average structures from last 50 ps (upper-right diagonal) and starting structures for the refolding simulations (lower left diagonal). These rms deviation were calculated after re-orienting the structures to best overlap the core backbone. Along the diagonal are the rms differences between the starting and final structures of the respective refolding simulation.

into several groups in terms of their rms from the N300 structure: A474, A540 and A725 have rms values in range 1.79($\pm$0.02) Å, A319 and B358 have a range 2.00($\pm$0.07) Å, C474 is at 2.21 Å, while A358, A820 and D810 are considerably worse (2.73($\pm$0.14) Å). Within group A, the rmsd between structures ranges from 1.21 to 2.52 Å; in some cases, the structures tend to be closer to each other than to N300. Energy minimization slightly improves the agreement with the crystal structure. For the N300 structure, energy minimization (Table 8) improves the results slightly for some structures (e.g. by 0.02 Å for A358) but makes it worse for most others (0.06 Å for A725 and 0.04 for A319).

The reduction in rmsd for the refolded structures is not simply a consequence of the decrease in the radius of gyration on formation of the disulfide bonds. There appears to be no correlation between the rms difference and the radius of gyration. For D810, which misfolds, the rmsd in the ''folded'' structure is slightly larger than in the unfolded structure, while the R_g decreases from 9.7 to 8.9 Å upon introducing the disulfide bonds. Also, the best refolded structure (A540, rmsd = 1.76 Å)

has a larger R_g (R_g = 9.4 Å) than the worst (rms = 2.85 Å; R_g = 8.9 Å). These results show that there is refolding of PCI as measured by an improvement of rmsd that is not correlated with collapse as measured by a decrease in R_g.

The simulations leading to the refolded structures were started from unfolded structures where the rms differences with the native simulation structure ranged for 1.50 to 4.03 Å (see Table 7). After the refolding simulations, features of the native structure are regained and the rms differences with the average structure from the N300 simulation varied between 1.76 and 2.78 Å. The results indicate, interestingly, that the refolded structures which agree less well with N300 were obtained from structures that had smaller rms deviations from it and were more compact, as measured by the radius of gyration (e.g. compare A319 and A358 with A474, A540 and A720). In fact, the A319 and A358 simulations, which had the least unfolded starting structures became less like N300 during ''refolding'' (see Table 7). An inverse relationship between folding ability and the compactness of the starting structures has been observed in some cases using lattice model simu-

Table 8. Matrix table for the backbone rms differences (in Å) between control structures and refolding simulations

Simulations	Crystal	NMR	N300	A319	A358	A474	A540	A725	A820	B358	C474	D810
Crystal	0.55	1.21	1.45	2.64	3.06	2.62	2.35	2.39	2.94	2.58	2.78	3.13
NMR	1.16	0.00	1.72	2.56	3.11	2.71	2.47	2.28	3.08	2.66	2.88	2.88
N300	1.69	1.76	0.28	2.03	2.76	1.85	1.75	1.85	2.63	2.09	2.21	2.94
A319	2.07	2.27	1.50	1.80	1.87	1.32	1.47	2.08	2.49	1.73	1.32	2.95
A358	2.57	2.69	1.90	1.28	2.26	2.16	2.11	2.29	2.09	2.16	1.50	3.61
A474	3.82	4.03	3.58	2.94	2.39	2.97	1.57	2.24	2.27	1.69	1.20	2.98
A540	3.89	4.03	3.27	2.91	2.82	3.53	2.93	1.78	2.19	1.76	1.27	2.52
A725	3.64	3.80	3.53	3.11	3.39	4.48	2.73	3.38	2.23	2.08	2.10	2.87
A820	4.44	4.63	4.03	3.61	3.61	4.41	2.57	2.52	3.33	2.42	1.93	3.40
B358	2.57	2.69	1.90	1.28	0.00	2.39	2.82	3.39	3.61	2.22	1.76	3.35
C474	3.82	4.03	3.58	2.94	2.39	0.00	3.53	4.48	4.41	2.39	2.76	2.90
D810	3.12	2.90	2.88	3.09	3.17	4.13	4.22	4.34	5.05	3.17	4.13	0.82

Crystal, NMR and average structures were minimized by 300 steps of ABNR.

Values for rms differences between average structures form last 50 ps (upper-right diagonal) and starting structures for the refolding simulations (lower-left diagonal). These rms deviations were calculated after re-orienting the structure to best overlap the core backbone. Along the diagonal are the rms differences between the starting and final structures of the respective refolding simulation.

lations (Gutin *et al.*,1995; Camacho & Thirumalai, 1996; Mirny *et al.*, 1996). Experimental studies of PCI folding (Pavia, personal communication) and hirudin (Chang, 1995) concluded that compact scrambled structures were less productive in conversion to the native state.

Here, the refolding of the proteins was driven by slowly forcing the disulfide bonds together using an NOE-type constraint. To assess the effect of this procedure on the final structure, an unfolded conformation was allowed to refold without the NOE-type constraint. This structure was taken from the final steps of the ND600 simulation, where the rms deviation before refolding was 5.9 Å. Refolding the protein by just allowing the system to cool to 300 K gave a structure with an rms deviation of 5.6 Å from the N300 average structure. Such behavior is similar to that observed by Alonso & Daggett (1995) and Lazaridis & Karplus (1997). This is to be compared to A820 (4.03 Å), which folded to an rms of 2.57 Å.

A structure taken from the end of the ND600 simulation with an rms deviation of 6.1 Å and a radius of gyration of 8.87 Å was refolded using methodology A. The final structure from this refolding simulation had a radius of gyration and SASA values close to the original N300 values, but it was significantly misfolded (see Figure 7) with an rms deviation of 6.6 Å. The properties of the average structure from this simulation are presented in Table 9. The relative orientation of the disulfide bonds in the protein core was significantly different from that in the N300 average structure (Figure 7); this resulted in the incorrect orientation of the different segments within the protein core, as well as the terminal ends.

Detailed structural analysis

We focus on the behavior of the loop regions since they are the most important structural elements of this protein. In addition, we consider the hydrogen bonds present in both the X-ray and NMR structure. The 3_{10} helix observed in the X-ray structure but not in the NMR structure is not present in the N300 simulation nor in any of the refolded structures.

The deviations as a function of residue number are shown in Figure 8 for the initial and final structures. The starting structures (Figure 8(a)) had large rms deviations in the core of the protein, with the exception of the structures 319 ps and 358 ps (see also Table 7); the largest deviations occur for structures 540 ps and 820 ps. The major deviations are located in the loops; particularly in loop 1 (Cys8 to Cys12) and loop 3 (Cys27 to Cys34). Figure 8(b) shows the rms deviations for all the refolded structures relative to the N300 average structure, while Figure 8(c) shows the average of the three best A structures and the average of the three worst A structures. As can be seen, the main difference in the two sets occurs in the 3_{10} helix region and in the loop 3 region; the two sets
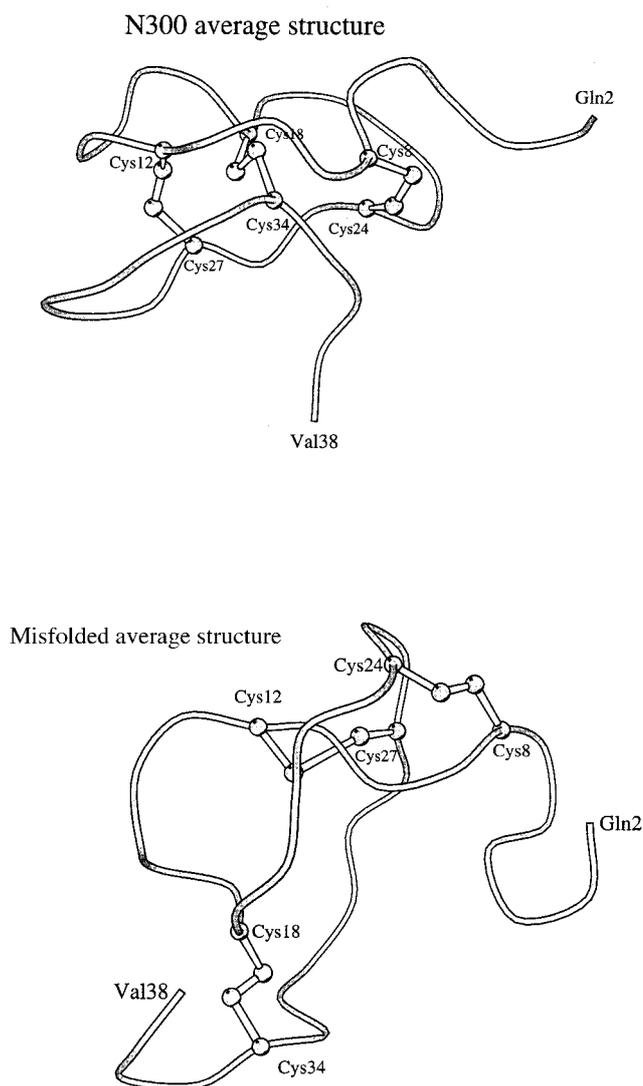

Figure 7. A ribbon representation of the PCI highlighting the disulfide bonds between cysteine residues. (a) N300 average structure. (b) Misfolded average structure. Computer graphic Figure generated using the program MOLSCRIPT (Kraulis, 1991).

have essentially the same deviations in loop 1 and loop 2. Only simulations A540 and A725 recovered completely the orientation of the loop 1 and 3_{10} helix regions. The rms deviation of loop 2 was relatively large from residues Cys18 to Gly20 in all cases (Figure 8(b)). Visual inspection of the average structures (see Figure 9) shows that the 3_{10} helix becomes more turn-like and forces loop 2 to change its orientation in all structures (except A540 and A725). For comparison, Figure 8(d) shows the rms deviation as a function of the residue number between the 11 NMR structures and the N300 average structure as well as between the crystal structure and the N300 average structure. As in the refolding simulations, the major differences occur in the 3_{10} helix and loop regions. Similar behavior is found when comparing the average NMR struc-

Table 9. Overall properties of the misfolded structure

Property	N300 simulation	Average A refolding	Misfolded simulation
Potential PCI E[a]	182 ± 28	280 ± 36	1469 ± 24
van der Waals PCI E[a]	-85 ± 9	-62 ± 9	300 ± 15
Electrostatic PCI E[a]	-371 ± 25	-291 ± 21	-429 ± 25
Backbone rmsd (Å)[b]	–	2.116	6.640
Backbone rmsf (Å)[c]	0.989	1.191	1.395
Radius of gyration (Å)[a]	8.963	9.097	8.770
SASA (Å^2)[e]	0.00 (2079.76)[f]	240.97	179.47

[a] Mean and standard deviation of PCI energy terms (in kcal/mol).
[b] The backbone atoms rms deviation from the N300 average structure for core residues (8Cys to 34Cys) averaged over the last 50 ps.
[c] The backbone atoms rms fluctuation about the average structure for last 50 ps.
[d] The radius of gyration for core residues (8Cys to 34 Cys) averaged over the last 50 ps.
[e] Solvent accessible surface area (SASA) for core residues (8Cys to 34Cys) averaged over the last 50 ps relative to the N300 average structure value.
[f] Original N300 average structure values of SASA.

ture with the X-ray structure. The largest deviations are near residue Gly20 (average rms deviation of 2.22 Å from Ser19 to Ala21). Overall, the distribution of rms deviations in the NMR structure is similar to that obtained using methodology A.

Loop regions can deviate from the experimental (or here, average N300 structure) for two reasons. The loop can have the correct internal structure and differ in orientation and/or the internal structure of the loop can be distorted. These two factors have been examine in loop modelling studies (van Vlijmen & Karplus, 1997). Table 10 shows the deviations obtained for the backbone atoms of the various structures by superimposing the entire protein or each loop/helix region by itself. The result for the four different loop/helix regions of the protein are given, i.e. loop 1 between Cys8 and Cys12, helix 3_{10} between Cys12 and Cys18, loop 2 between Cys18 and Cys24 and loop 3 between Cys27 and Cys34. From the results for N300 it is evident that even in the native state simulation, there are significant deviations both overall and internally; loops 1 and 2 have undergone mainly an overall displacement, while in helix 3_{10} and loop 3 the internal rearrangement is quite large. In the three best refolding simulations (best in terms of rms) and in many of the others, the overall and the internal structure of the loop/helix regions is significantly improved after refolding. For some cases, even though the overall orientation of the loop/helix is not close to the orientation in the N300 average structure (e.g. B358), the internal structure is still improved. Structure A540 has particularly low rms values, while structure A725 has significantly higher values for two of the loops. The A319 simulation is the only simulation to lose completely the native conformation in the helix 3_{10} region; the other simulations all recovered a certain degree of the helix 3_{10} configuration. The results of the loop 2 superposition yield an average internal rms deviation of 1.09 Å after refolding. Simulation B358 lost the native conformation in this region on refolding (0.97 Å before; 1.64 Å after). Of the

different regions, loop 3 showed the largest average internal deviation (1.35 Å); the result is skewed by the results from the D810 simulation; when this is removed, the average rms deviation decreases to 1.23 Å. In most of the simulations, the internal conformation of loop 3 does not improve very much on refolding; the exceptions are simulations A474 and A540, where the values before and after refolding were 1.12 Å/0.85 Å and 1.52 Å/0.96 Å, respectively.

Table 11 shows the structural properties of the disulfide bonds after refolding in comparisons with the N300 simulation. The rms deviations for the atoms of the two cysteine residues involved in the disulfide bonds range from 0.7 to 1.8 Å after refolding (relative to 1.2 to 5.5 Å before refolding). The χ_3 dihedral angles between cysteine residues are also reported in Table 11. In most cases, the values after refolding are in the expected range, around $-90°$ for left-handed disulfide bonds and $+90°$ for right-handed disulfide bonds (Thornton, 1981). However, there are some dihedral angles that are far from the expected values; Cys12–Cys27 with 178.15° in the A319 simulation and 162.76° in the C474 simulation; and Cys8–Cys24 with 133.33° in the B358 simulation. None of the simulations, other than D810 in which the disulfides were not broken, have all of the disulfide bond dihedral angles in the correct region; e.g. the lowest rms deviation structures (A474, A540 and A725) have one (A474 and A540) to two (A725) incorrectly twisted disulfides. This is an important source of error in the final structures.

In Table 4, the appearance of specific hydrogen bonds after refolding are reported. The refolded structure for which the starting structure was closest in terms of rms to the N300 average structure (A319 ps) has more internal hydrogen bonds than those with larger rms. However, only one of the backbone hydrogen bonds (12CysN–32ArgO) and two of the side-chain hydrogen bonds in A319 (10LysN$^{\zeta_1}$–17AspO and 13LysN–17AspO$^{\delta_1}$) are native-like. The D810 simulation, which had the disulfide bonds intact during the entire simu-
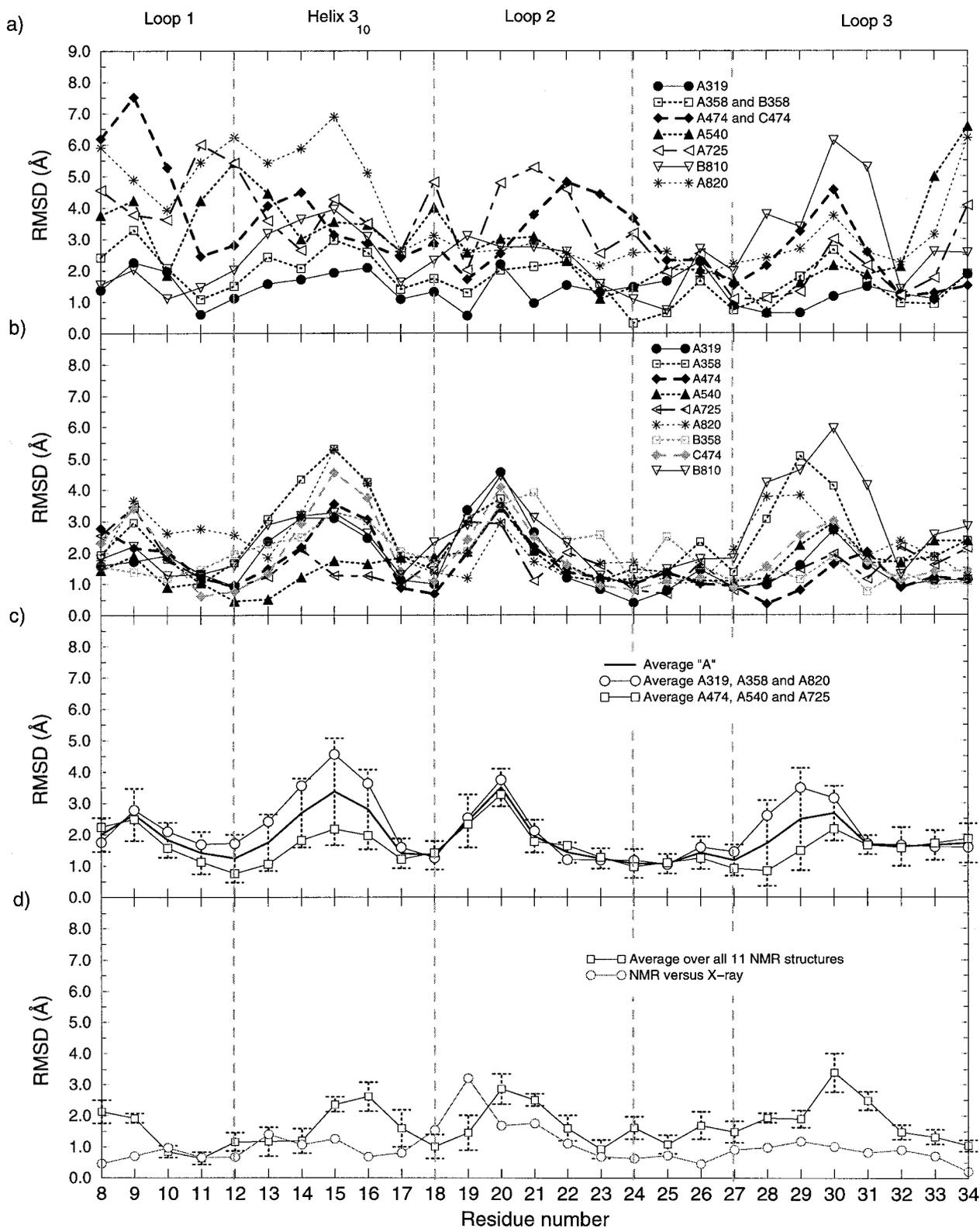
Figure 8. Rms deviation of refolded structures from N300 average structure as a function of residue number. Cysteine positions are marked by the vertical dotted lines. (a) Results for the seven different starting structures. (b) Results for average structures of the last 50 ps of every refolding simulation. (c) Average rmsd and standard deviation for refolding simulations using methodology A. (d) Average rmsd and standard deviation for NMR PCI structures and the rmsd between the mean NMR structure and the crystal structure.
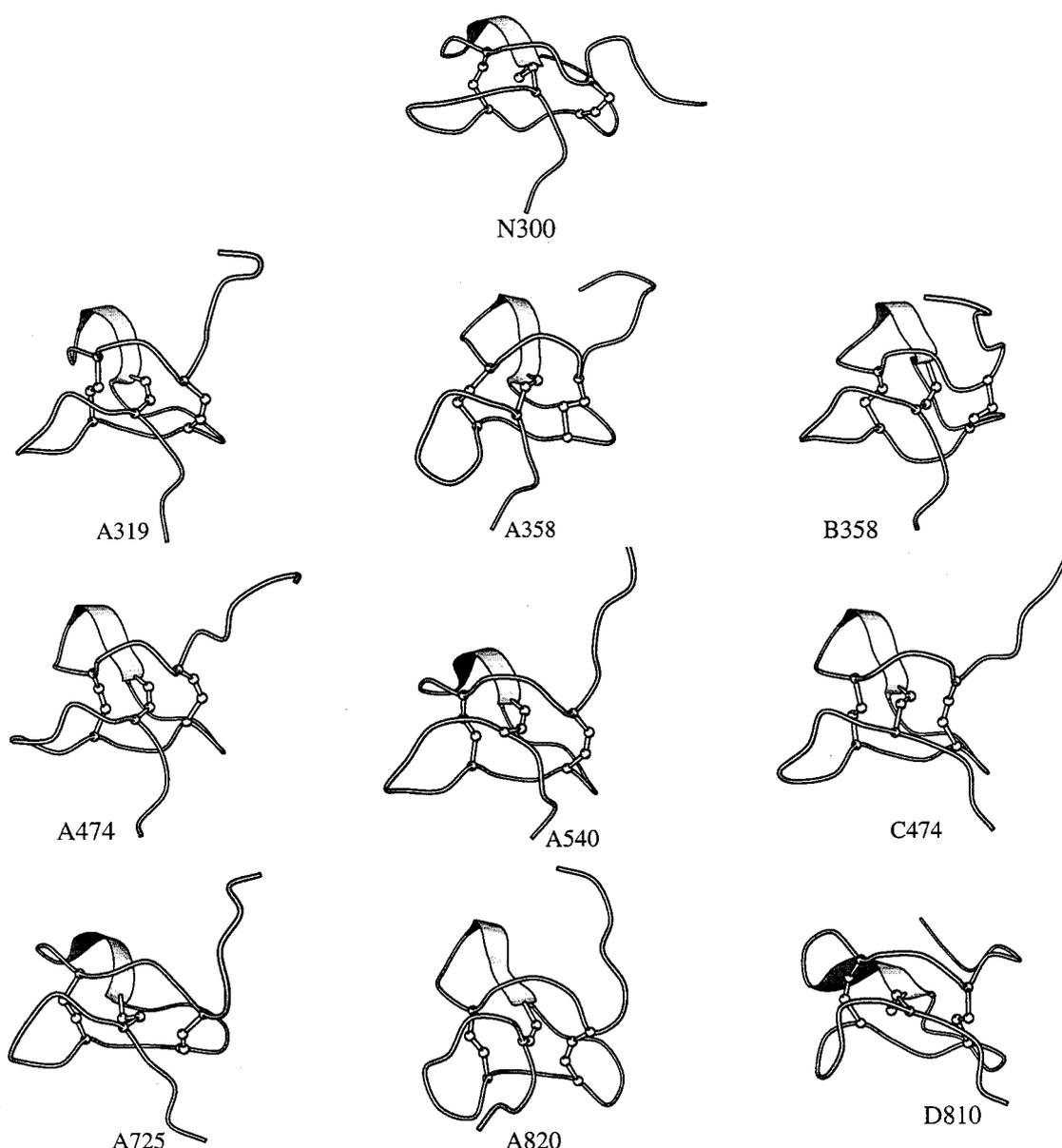
Figure 9. A ribbon representation of the average structures for N300 and refolding simulations. Figure generated using the program MOLSCRIPT (Kraulis, 1991).

lation, also has several internal backbone hydrogen bonds of which two (12CysN–32ArgO and 34CysN–10LysO) are native-like. The other simulations lost most of the original hydrogen bonds during the unfolding and refolding cycle but recovered at least one hydrogen bond; B358 recovered three backbone hydrogen bonds and no backbone–side-chain hydrogen bonds while A474 recovered two backbone hydrogen bonds present in both the X-ray and NMR structures and A820 also recovered two hydrogen bonds. In many cases, there was a switching of the hydrogen bond acceptors. Examples are: 10LysN–7IleO changed to 10LysN–8CysO in simulations A319, A540 and B358 and 21AlaN–18CysO changed to

21AlaN–24CysO in simulations A358, A474, A540 and D810. Often, a water molecule replaced the original hydrogen bond; for example, the hydrogen bond 35GlyN–26AlaO was replaced in simulation B358 by 35GlyN–971H$_2$O (49.5%) and the same hydrogen bond was replaced by 35GlyN–3379H$_2$O (50.0%) in the A540 simulation. Several hydrogen bonds would have been recovered if a more generous criterion were used. This is true for the β-strand hydrogen bonds; for example, the average distance between the donor and the acceptor in the hydrogen bond 26AlaN–35GlyO was 3.3 Å in the A474 and, in the simulations A319 and A358, the hydrogen bond 35GlyN–26AlaO had an average distance

Table 10. Regional rmsd between N300 control structures and refolding simulations

Protein region	N300[a]	A319 Before	A319 After	A358 Before	A358 After	A474 Before	A474 After	A540 Before	A540 After	A725 Before	A725 After	A820 Before	A820 After	B358 Before	B358 After	C474 Before	C474 After	D810 Before	D810 After
Overall Loop 1	0.94	1.47	1.49	2.08	1.88	4.85	1.82	3.91	1.34	4.68	1.98	5.28	2.67	2.08	1.51	4.85	1.83	1.63	1.63
Internal Loop 1	0.31	1.05	0.89	1.09	1.16	1.52	0.94	1.22	0.58	1.00	1.71	1.13	1.32	1.09	0.78	1.52	0.96	0.73	0.73
Overall Helix 3_{10}	1.44	1.55	2.01	2.11	3.14	3.23	1.82	3.79	1.32	3.84	1.35	5.03	2.89	2.11	2.41	3.23	2.25	2.83	2.45
Internal Helix 3_{10}	1.07	0.92	1.33	1.25	1.29	1.39	1.08	1.15	0.57	0.89	0.73	1.56	1.08	1.26	0.74	1.39	1.10	1.52	1.35
Overall Loop 2	1.46	1.34	2.00	1.63	2.01	3.41	1.69	2.50	1.89	3.89	1.87	2.64	1.68	1.63	2.46	3.41	1.91	2.31	2.46
Internal Loop 2	0.73	1.13	0.88	0.97	1.12	1.58	0.80	1.73	0.95	2.48	1.07	1.99	1.24	0.97	1.64	1.58	0.85	1.30	1.30
Overall Loop 3	1.33	1.13	1.37	1.48	2.65	2.26	1.11	2.71	1.96	1.97	1.53	3.17	2.46	1.48	1.21	2.26	1.70	3.40	3.42
Internal Loop 3	1.21	1.00	1.08	1.14	1.72	1.12	0.85	1.52	0.96	1.53	1.31	1.48	1.78	1.14	1.07	1.12	1.05	2.33	2.29

All values are given in Å.
Values of rms difference for starting and average refolding structures with respect to the N300 average structure.
[a] N300 average structure with respect to the X-ray structure.

Table 11. Disulfide bond properties of refolding simulations

Analysis	Disulfide bond	X-ray	NMR	N300	A319	A358	A474	A725	A540	A820	B358	C474	D810
Bond distance[a] (Å)	8–24	–	0.82	0.52	1.55/0.74	1.48/1.42	4.56/1.38	4.21/1.83	2.39/1.14	4.03/1.48	1.48/1.55	4.56/1.52	1.37/1.38
	12–27	–	0.98	0.67	1.15/1.35	1.72/1.49	2.64/1.20	2.66/1.11	1.64/1.15	2.07/1.67	1.72/0.71	2.64/1.39	0.79/0.60
	18–34	–	0.76	0.59	1.45/1.35	1.30/0.98	1.68/1.29	5.18/1.38	5.49/1.32	5.09/1.27	1.30/1.32	1.68/1.30	0.94/0.92
Dihedral angle[b] (degree)	8–24	92.6	–57.2	97.5	92.4	–91.4	97.3	–97.3	95.7	–82.9	133.3	–101.3	86.3
	12–27	–145.1	91.5	–101.5	–178.2	105.3	84.2	–70.9	100.1	–90.5	–102.5	–162.8	–102.1
	18–34	–103.0	–77.1	–94.4	–94.3	85.7	–112.3	76.0	–93.4	–74.0	–78.0	67.7	–78.8

[a] Values of rms deviation for starting and average refolding structures (separated by slash) with respect to the N300 average structure.
[b] Values of dihedral angle $\chi3$ of disulfide bonds for N300 and refolding average structures.

of 3.2 Å. Several hydrogen bonds were never recovered. Table 4 shows that the hydrogen bond 12CysN–32ArgO was most often present, confirming the importance of this bond in the native structure of PCI. The same was found for the side-chain–backbone hydrogen bond 13LysN–17AspO$^{\delta 1,2}$, already mentioned.

Criteria for determining the accuracy of the refolded structures

Given the various refolded structures with a range of rms deviations of 1.76 to 2.78 Å (excluding D810) from the N300 target structure, it would be very useful to be able to have criteria which can pick out those that have the smaller rms. This is an important problem which is essential to predicting the tertiary structure from the sequence (Levitt *et al.*, 1997). Many approaches to folding when applied to a protein of known structure generate a set of structures with a wide variation in their rms from the native structure. Often, the rms range of "folded structures" varies between 1.5 and 10 Å for small proteins (Levitt *et al.*, 1997). A useful "screening" function would be able to pick out the near-native folds, In one survey of six energy functions, Park & Levitt (1996) showed that the best structures in terms of rms were identified rather well, but that there were "always many grossly misfolded decoys with energies more favorable than some of the near-native folds" (Levitt *et al.*, 1997). In the present case we have generated a set of structures, all of which, except for the misfolded structure, are near-native (rmsd of 1.76 to 2.85 Å relative to N300). We apply a number of criteria that have been used for evaluating protein structures to this set. For comparison, we apply the same criteria to the X-ray structure, the set of NMR structures, and the N300 average structure. Table 12 shows the potential energy, the van der Waals and the electrostatic energy (PCI-self, PCI-water and the total corresponding to PCI-self plus PCI-water) of each simulation over the last 200 ps the values from the N300 simulation are listed for ease of comparison. All the refolding simulation structures were stable during that period as is evident from the standard deviations. Examination of the various energies shows that neither the intramolecular energy of PCI, the interaction energy between PCI and water, or the two combined can distinguish between the native-like N300 average structure and the refolded structures or among the various refolded structures. Only for the misfolded structure is the van der Waals energy noticeably higher than that of the other structures (Table 9).

In the simulations where the refolded structures originated from structures with reduced disulfide bonds, the total energies are all more favorable relative to the energy of the N300 average structure (except simulations A725 and A820). For the refolded structure originating from the N600 simulation where the disulfide bonds were kept intact (D810), the energy is less favorable than for the N300 average structure. Greater conformational flexibility in the ND600 simulation apparently allows the refolded structure to find a lower local minimum. There is a good agreement between the potential energy of interaction between PCI and water for all refolding simulations, except for the simulations done using the 474 ps structure. These two simulations (A474 and C474) had residues more exposed to the solvent (see SASA results from Table 6), thus allowing better protein-solvent interactions. The contribution of the van der Waals interactions to the potential energy has a similar behavior. The largest difference from the N300 value was observed for simulation C474. That this average structure had a less negative value of van der Waals energy for PCI alone (−50 *versus* −85 kcal/mol) can also be explained by its relatively greater exposure to solvent. The differences between the refolding simulations on electrostatic interactions showed a similar behavior to the van der Waals energy.

A number of functions that make use of solvation potentials have been proposed in recent years for evaluating protein structures. These empirical solvation models have been used to discriminate between native conformations and compact non-native ones in proteins. The models generally assume that the total solvation free energy is a sum of contributions from the constituent groups of the protein. Parameters for the models are determined from experimental measurements of hydration and hydrophobicity of amino acids and/or analogs. The models appear to discriminate between grossly misfolded proteins generated by threading methods and the native conformation. Here, we use two methods based on this approach, that of Eisenberg and co-workers (Lüthy *et al.*, 1992) and that of Koehl & Delarue (1994), to see to what degree these models can discriminate between the crystal structure, the NMR structures and the different structures obtained from simulations. The model of Lüthy *et al.* (1992) was used to generate 3D profiles of the various structures. In this model, the higher the score, the better is the structure; Figure 10 shows the residue by residue score for the average structures, energy minimized structures, the crystal structure and the NMR structures. The residue by residue scores are determined using a 21-residue sliding window, so the scores for the first several and last several residues have little or no meaning. The crystal structure and the mean NMR structure give the highest scores for most of the residues; the latter has better scores than the former, even though the profiles are based on X-ray structures. The average structure from the N300 simulation gives scores which are better, in general, than those of the refolded structures but worse than the two experimental structures. There is no correlation between the rms values of the refolded structures and the scores. Although the refolded structures do less well than the N300 structure, there are some that are close; for example A725 and B358. Of those structures

Table 12. Average and standard deviation energy values of refolding simulations

Energy		N300 simulation	A319 simulation	A358 simulation	A474 simulation	A540 simulation	A725 simulation	A820 simulation	B358 simulation	C474 simulation	D810 simulation
Potential	PCI	128 ± 28	248 ± 48	280 ± 44	322 ± 23	305 ± 43	245 ± 25	276 ± 32	214 ± 41	383 ± 28	185 ± 34
	PCI-Water	−1621 ± 50	−1727 ± 64	−1792 ± 53	−1951 ± 48	−1769 ± 78	−1600 ± 50	−1617 ± 53	−1680 ± 54	−2052 ± 49	−1574 ± 54
	Total	−1439 ± 39	−1478 ± 83	−1512 ± 79	−1629 ± 45	−1463 ± 83	−1355 ± 42	−1341 ± 47	−1466 ± 74	−1670 ± 47	−1389 ± 35
van der Waals	PCI	−85 ± 9	−71 ± 9	−71 ± 8	−57 ± 8	−61 ± 9	−56 ± 9	−57 ± 11	−83 ± 7	−50 ± 8	−78 ± 10
	PCI-Water	−135 ± 15	−146 ± 12	−140 ± 14	−160 ± 11	−138 ± 13	−107 ± 15	−109 ± 14	−141 ± 11	−153 ± 16	−129 ± 13
	Total	−220 ± 16	−217 ± 17	−212 ± 16	−217 ± 12	−198 ± 14	−163 ± 18	−166 ± 18	−224 ± 13	−203 ± 17	−208 ± 15
Electrostatic	PCI	−372 ± 25	−318 ± 27	−273 ± 16	−237 ± 18	−270 ± 32	−328 ± 22	−320 ± 21	−341 ± 17	−193 ± 21	−386 ± 34
	PCI-Water	−1485 ± 55	−1581 ± 70	−1651 ± 54	−1791 ± 51	−1631 ± 78	−1493 ± 51	−1508 ± 55	−1539 ± 54	−1900 ± 51	−1445 ± 58
	Total	−1857 ± 40	−1898 ± 52	−1924 ± 44	−2028 ± 42	−1901 ± 60	−1821 ± 40	−1828 ± 46	−1880 ± 46	−2093 ± 42	−1832 ± 41

The term Total refers to the sum of the PCI and PCI-water terms. All energies are in kcal/mol.
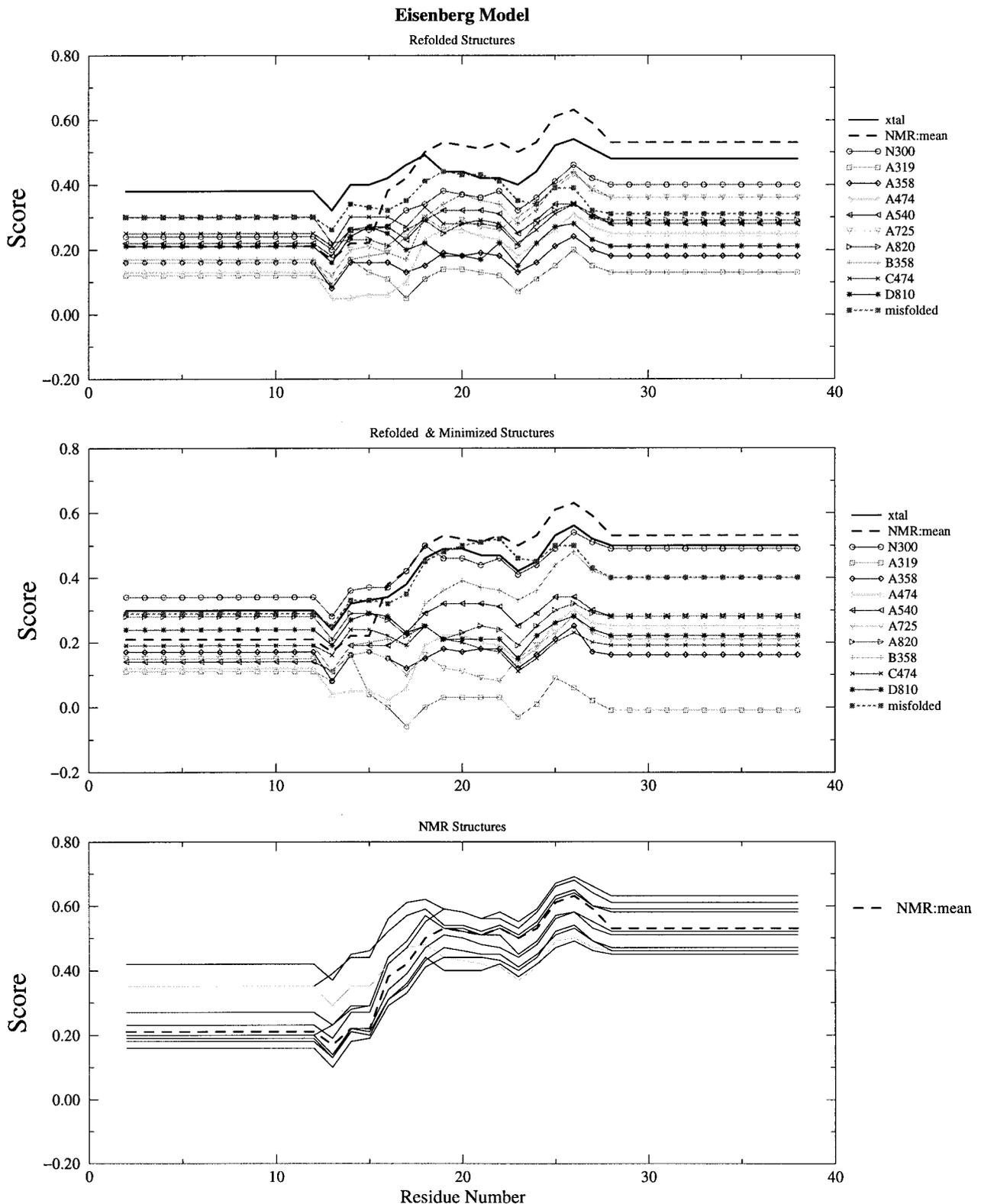
Eisenberg Model

Figure 10. Scores calculated from the model of Lüthy *et al.* (1992) on a residue-by-residue basis. The top Figure is for the NMR, crystal and simulation structures; the middle Figure is for the energy minimized NMR, crystal and simulation structures; and the bottom Figure is for the family of NMR structures.

that are closest in rms to N300 (A474, A540 and A725), only A725 has scores close to N300. The misfolded structure, however, scores well with respect to N300 and for a major part of the protein,

surpasses the corresponding N300 values. After energy minimization (Figure 10(b)), the N300 scores closer to the experimental structures, the refolded structures score less well relative to N300

Table 13. Total scores from the model of Lüthy *et al.* and the total environment free energies from the Koehl and Delarue model (in kcal/mol)

| | Lüthy *et al.* | | Koehl-Delarue | |
	Simulation	Minimized	Simulation	Minimized
Crystal	15.0	15.0	−70.0	−69.0
NMR:mean	16.08	15.27	−69.4	−69.4
N300	12.15	15.64	−62.3	−65.4
A319	4.62	1.59	−54.9	−57.2
A358	6.26	6.08	−56.6	−58.8
A474	6.96	6.45	−51.0	−52.9
A540	9.79	8.61	−53.1	−55.1
A725	9.91	6.29	−49.4	−50.9
A820	10.46	9.97	−53.4	−56.9
B358	10.15	11.22	−56.9	−58.8
C474	10.11	7.13	−46.3	−48.7
D810	7.86	8.56	−55.3	−59.4
Misfolded	12.22	13.97	−53.3	−55.2

Values are given for both the average structures from simulation and the minimized structures.

and the experimental structures while the misfolded structure scores as well or better than N300. The set of NMR structures have a rather wide distribution (see Figure 10(c)). Some of the refolded structures fall inside the envelope of scores from the family of NMR structures. The total scores given by the sum of scores for each amino acid from the model of Lüthy *et al.* for the average and minimized structures are given in Table 13. From this perspective, the N300 structure compares well to the crystal and NMR structures; the refolded structures all score less well and the misfolded structure scores better than the N300 structure.

The second model of this type was that proposed by Koehl & Delarue (KD; 1994). This model extends the model of Luthy *et al.* by including atom−atom contacts. The KD model estimates an environment free energy; the lower the value, the better the structure. In contrast to the model of Lüthy *et al.*, the KD results do not give a clear separation between the experimental structures (X-ray, NMR) the N300 structures, and the refolded set (Figure 11). For many residues, the free energy values are better in the experimental structures than the structures determined from simulation, but the differences are small. Moreover, for some residues, the simulation structures give better free energies (Lys10, for example). As with the Lüthy *et al.* model, the misfolded structure scores well relative to the experimental and simulation structures. From a linear fit of the environment free energy to the length of the protein, using a sdatabase of 82 proteins, Koehl and Delarue found the following relation.

$$\Delta G_e = 51.1 - 2.75\,N \qquad (1)$$

where ΔG_e is the environment free energy and N is the total number of residues. For a 37 residue protein, the value of ΔG_e given by the above expression is −50.65 kcal/mol. The total ΔG_e, taken as the sum of ΔG_e for each amino acid position, for each average and minimized structure in

the present set is given in Table 13. The total environment free energies, ΔG_e, for the crystal and NMR structures are between −68 and −74.8 kcal/mol, the N300 structure gives values of −62.3 and −65.4 kcal/mol for the average and minimized structures, respectively. The refolded structures give values ranging from −46.3 and −58.8 kcal/mol; the misfolded structure gives ΔG_e values of −53.3 and −55.2 kcal/mol for the average and energy minimized structures, respectively. All the refolded structures have higher values than the N300 average structures. As for the model of Lüthy *et al.*, there is no correlation between rms and the ΔG_e score. Neither model is able to clearly distinguish between the misfolded structure and the native and near native structures.

A different approach to the characterization of misfolded proteins uses an effective energy function based on the CHARMM polar hydrogen force field (Neria *et al.*, 1996) combined with a solvent exclusion model for the solvation free energy (Lazaridis & Karplus, 1997). In this model, the internal energy components are considered as well as the non-bonded terms. Results for the entire protein and for the core are given in Table 14. The mean NMR structure, as in the other models, gives the best energy, followed by the crystal structure and the N300 structure. Some of the refolded structures, for example B358 and A725, have energies close to the energy of N300. The behavior of the core energies is similar to that for the entire protein. In contrast to the results with the pure solvation models described above, this model discriminates between the misfolded conformation and the near-native conformations; the energy of the misfolded protein is significantly higher than the others, due to the inclusion of internal energy terms. This suggests that internal energy terms, in addition to solvation energies may need to be considered when characterizing the quality of a protein structure. As found for the protein solvent interaction energy (see Table 12), many of the refolded structures have more negative solvent contributions than the experimental structures.

Concluding Discussion

The problem of protein folding is of inherent interest both from a predictive and kinetic viewpoint. Here we have examined a specific "prediction" problem. Can the introduction of known disulfide bonds for the unfolded polypeptide chain make it possible to determine the native structure by molecular dynamics simulations? We have chosen the three-disulfide bond protein, potato carboxypeptidase inhibitor, because it is small, representative of a sizable class of T-knottins, and it is possible to construct non-native/scrambled three-disulfide species whose structure has not yet been determined. Use of such a small protein with
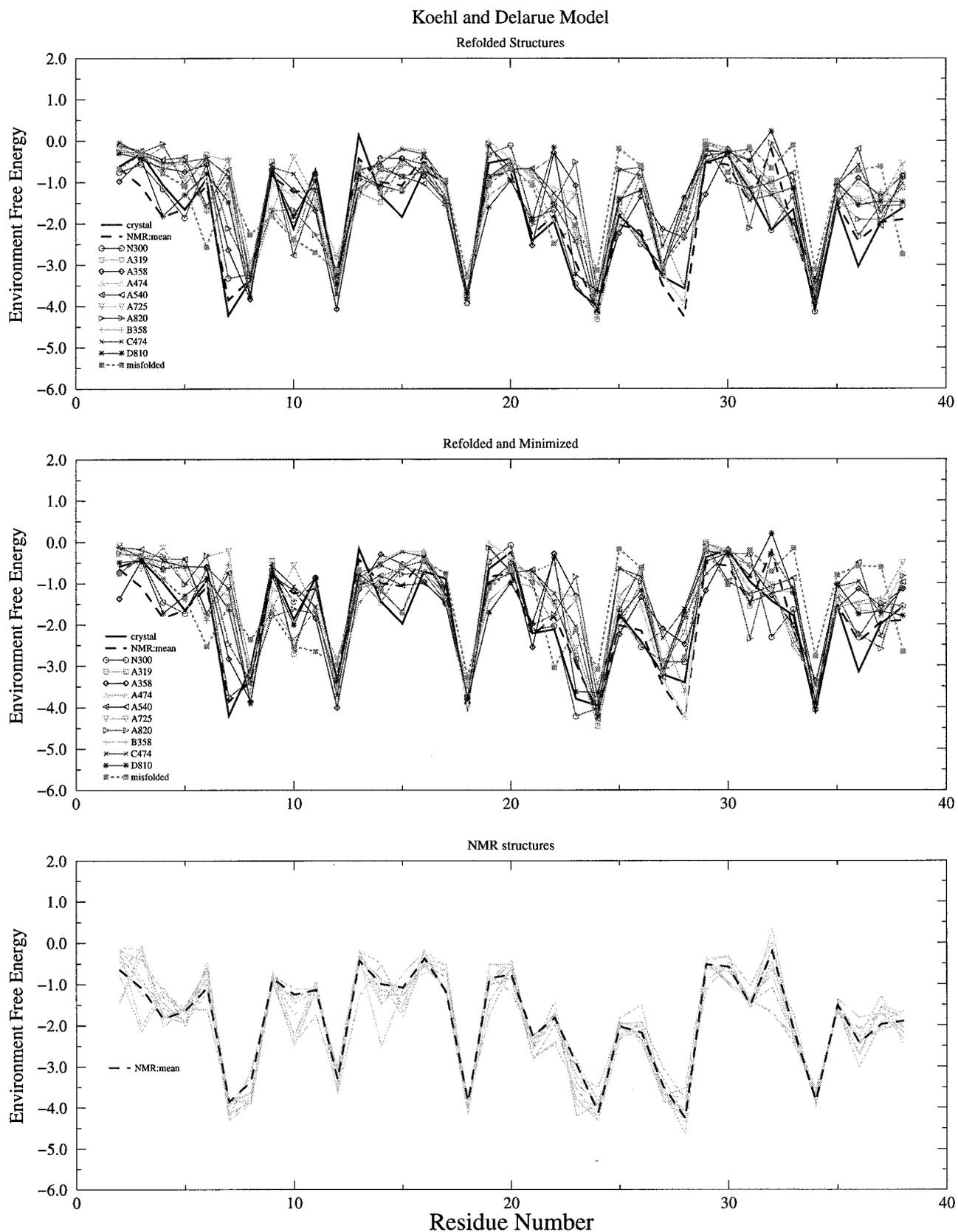
Figure 11. Environment free energies calculated from the model of Koehl and Delarue on a residue-by-residue basis. The top Figure is for the NMR, crystal and simulation structures; the middle Figure is for the energy minimized NMR, crystal and simulation structures; and the bottom Figure is for the family of NMR structures.

Table 14. Total energy and components calculated using the model of Lazaridis and Karplus

	Energy		Internal		van der Waals		Electrostatic		Solvent	
Crystal	−977.6	−709.8	145.3	108.1	−192.4	−123.7	−544.0	−400.0	−386.5	−294.3
NMR[a]	−1023.8	−744.9	151.9	102.3	−198.4	−132.5	−612.1	−439.1	−365.1	−275.6
N300	−959.2	−690.9	133.2	96.9	−199.5	−127.5	−493.2	−351.0	−399.8	−309.3
A319	−941.7	−693.9	135.0	93.3	−181.1	−132.6	−482.5	−352.5	−413.1	−302.0
A358	−944.2	−688.5	131.8	95.6	−176.5	−122.4	−488.9	−347.2	−410.6	−314.6
A474	−927.6	−675.3	129.4	88.0	−163.3	−117.8	−467.5	−328.0	−426.2	−317.5
A540	−917.5	−678.6	135.5	92.7	−157.9	−110.1	−464.9	−336.4	−430.2	−324.8
A725	−950.0	−698.7	138.6	96.5	−171.8	−127.9	−506.7	−367.1	−410.1	−300.3
A820	−924.0	−683.9	144.2	100.9	−161.6	−112.7	−483.5	−360.9	−423.1	−311.2
B358	−958.9	−705.4	139.7	95.1	−184.1	−134.8	−507.0	−368.7	−407.6	−297.0
C474	−891.3	−648.3	147.2	100.3	−149.9	−112.7	−454.3	−314.8	−434.3	−321.1
D810	−948.4	−697.1	153.5	110.9	−172.9	−119.5	−529.3	−391.4	−399.7	−297.1
Misfold	576.2	854.0	1139.3	1107.3	343.5	402.0	−500.7	−337.4	−405.9	−317.8

Energies are given in kcal/mol. Two energies are given for each entry; the first is for the average structure energy minimized for 300 steps using ABNR. The second is for just the protein core, residues 8 to 34 after minimization.

[a] Mean NMR structure.

specific structural elements in a compact core (three loop regions) provides a first test of the approach before applying it to scrambled species or larger systems such as BPTI. As described here, successful refolding is achieved, at least in some trials from an rmsd = 3.53 Å after unfolding to an rmsd = 1.79 Å after folding, although no criterion was found to select the best structures in terms of their rmsd.

Prior to refolding, molecular dynamics simulations of native PCI at room temperature were performed. They showed that the protein is stable in the core region and that the N and C-terminal ends are rather mobile; an average backbone rms from the X-ray structure of 1.69 Å for the core was obtained in a 500 ps simulation. These results agree in general with the earlier simulations made with the GROMOS force-field (Oliva *et al.*, 1995; Daura *et al.*, 1996). Unfolding of PCI was initiated by temperature perturbation at 600 K, disulfide bond reduction, or both. The simulation with reduced disulfides at high temperatures showed the most pronounced unfolding; the final rms after 910 ps was 5.85 Å and the radius of gyration had increased from 8.96 to 9.06 Å. The simulations with reduced disulfides (ND300) or high temperature with disulfide bonds (N600) did not show significant unfolding although the structures deviated from the X-ray value more than the native room temperature simulation (rms of 2.94 Å for ND300 and 3.08 Å for N600); for lysozyme with disulfide bonds present (M. Buck & M. Karplus, personal communication), it was found that only by going to 900 K was there significant unfolding in 100 ps. After initiation of unfolding, the van der Waals energy of the systems increased very quickly; the same energetic behavior was observed in the unfolding simulations of barnase done by Caflisch & Karplus (1995) and is due to the unpacking of the hydrophobic core of the protein during unfolding.

Refolding simulations were started with high temperature structures with reduced disulfide bonds at various time points during the simu-

lation. The native S−S bonds were reformed over different time periods and the resulting structures were found to be stable over 200 ps simulations at room temperature. The rms averages for the core of the refolded structures range from 2.44 to 3.20 Å from the X-ray structure and from 1.76 to 2.85 Å from the N300 average simulation structure. The largest deviations arise from the three loop regions in most cases; the internal structures of the loops tend to be correct but they are spatially misoriented. Few of the native state hydrogen bonds are reformed. This is not surprising because the hydrogen bonds appear to be of marginal stability; e.g. the X-ray and NMR structures have only four backbone and two backbone−side-chain hydrogen bonds in common.

An approach, where specific information concerning S−S bond pairing is introduced to aid in protein structure prediction, has recently been used in a genetic algorithm approach (Dandekar & Argos, 1997). Dandekar and Argos find that when disulfide bond pairing information is included in their simulations, the predicted structures tend to be in better agreement with the experimental structures; in the case of crambin, which has three disulfide bonds, they find that the structure predicted is closer to the experimental structure when using the disulfide bond information (4.1 Å) than when the information is not included (5.4 Å).

It is found that the refolded structures starting from less compact states tend to be closer to the X-ray structure. A similar trend has been observed experimentally in the case of hirudin (Chang, 1995), where experimental studies of folding intermediates showed that about half of the species with non-native disulfide bonds are sensitive to denaturants. This suggests that they are compact relative to the denatured state. These species are also less productive in converting to the native structure and more resistant against reduction. Several species of PCI with non-native disulfide bonds (scrambled species) have been isolated (Chang *et al.*, 1994). Experimental results have demonstrated that the species which are

more sensitive to the denaturants are the ones that have a slow rate of conversion to the native structure. The scrambled species that are less sensitive to the denaturants are likely to be less compact. They have faster rates of conversion to the native structure (S. Pavia *et al.*, personal communication). This suggests that the introduction of the S−S bond constraints in a more open structure permits a broader search that leads to more rapid folding to well-formed structures.

Tests were made with the overall CHARMM energy (protein and protein-solvent), an approximate solvent corrected CHARMM energy function, and two functions that evaluate the residue environment based on empirical data from protein structures. Comparison of the average CHARMM energies of the N300 and refolded simulations did not distinguish the former from the latter. The other functions were able to distinguish the experimental crystal and NMR structures and the N300 average structure from the refolded and misfolded structures. The solvation models of Lüthy *et al.* (1992) and of Koehl & Delarue (1994), were able to distinguish between the native structures and the near-native refolded structures; however, they were unable to differentiate between the refolded structures and the misfolded structure. In contrast, the CHARMM function corrected for solvation distinguished the N300 structure from the refolded structures and was able to differentiate between the near-native structures and the misfolded structure. Although the various functions can distinguish correctly folded proteins from those that are "incorrect" (incorrect as used here refers to an rms difference greater than 2.0 Å from X-ray or NMR), there seems to be no correlation between the rms deviations and the values of the various functions. This is an important point because many tests of empirical and other "energy" functions use such a comparison. Although globally one would hope that such a correlation exists, there is no evidence for it from the present results. It is likely that if a very large sample of structures is generated there will be an overall correlation between rms and energy, but it is also likely that certain structures will violate this correlation.

Acknowledgments

References

Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987). Protein Data Bank, in Crystallographic Databases-Information Content, Software Systems, Scientific Applications, pp. 107–132. (Allen, F. H., Bergerhoff, G. and Sievers, R. eds.), Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester.

Abola, E. E., Sussman, J. L., Prilusky, J. & Manning, N. O. (1987). Protein Data Bank Archives of Three-Dimensional Macromolecular Structures. In *Methods in Enzymology*. Vol. 277, pp. 556–571 (Carter, C. W., Jr and Sweet, R. M. eds), Academic Press, San Diego.

Alonso, D. & Daggett, V. (1995). Molecular dynamics simulations of protein unfolding and limited refolding: characterization of partially unfolded states of ubiquitin in 60% methanol and in water. *J. Mol. Biol.* **247**, 501–520.

Baker, E. N. & Hubbard, R. E. (1984). Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* **44**, 97–179.

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. J., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, S., Swaminathan, S. & Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.* **4**, 187–217.

Brooks, C. L. III & Karplus, M. (1989). Solvent effects on protein motion and protein effects on solvent motion. *J. Mol. Biol.* **208**, 159–181.

Brooks, C. L. III, Brünger, A. T. & Karplus, M. (1985). Active site dynamics in protein molecules: a stochastic boundary molecular-dynamics approach. *Biopolymers*, **24**, 843–865.

Brünger, A., Brooks, C. L., III & Karplus, M. (1984). Stochastic boundary conditions for molecular dynamics simulations of ST2 water. *Chem. Phys. Letters*, **105**, 495–500.

Brünger, A. T. & Karplus, M. (1988). Polar hydrogen positions in proteins: empirical energy placement and neutron diffraction comparison. *Proteins: Struct. Funct. Genet.* **4**, 148–156.

Caflisch, A. & Karplus, M. (1994). Molecular dynamics simulation of protein denaturation: solvation of the hydrophobic cores and secondary structure of barnase. *Proc. Natl Acad. Sci. USA*, **91**, 1746–1750.

Caflisch, A. & Karplus, M. (1995). Acid and thermal denaturation of barnase investigated by molecular dynamics simulations. *J. Mol. Biol.* **252**, 672–708.

Camacho, J. C. & Thirumalai, D. (1996). Denaturants can accelerate folding rates in a class of globular proteins. *Protein Sci.* **5**, 1826–1832.

Chang, J.-Y. (1995). The properties of scrambled hirudins. *J. Biol. Chem.* **270**, 25661–25666.

Chang, J.-Y., Canals, F., Schindler, P., Querol, E. & Avilés, F. X. (1994). The disulfide folding pathway of potato carboxypeptidase inhibitor. *J. Biol. Chem.* **269**, 22087–22094.

Chang, J.-Y., Krisnaswamy, T. & Yu, C. (1998). Unfolding and refolding of cardiotoxin III reversible conversion of the native and scrambled species. *Biochemistry,* **37**, 6745–6751.

Clore, G. M., Gronenborn, A. M., Nilges, M. & Ryan, A. C. (1987). Three-dimensional structure of potato carboxypeptidase inhibitor in solution. A study using nuclear magnetic resonance, distance geometry, and restrained molecular dynamics. *Biochemistry,* **26**, 8012–8023.

Daggett, V. & Levitt, M. (1992a). A model of the molten globule state from molecular dynamics simulations. *Proc. Natl Acad. Sci. USA,* **89**, 5142–5146.

Daggett, V. & Levitt, M. (1992b). Molecular dynamics simulations of helix denaturation. *J. Mol. Biol.* **223**, 1121–1138.

Daggett, V. & Levitt, M. (1993a). Realistic simulations of native-protein dynamics in solution and beyond. *Annu. Rev. Biophys. Biomol. Struct.* **22**, 353–380.

Daggett, V. & Levitt, M. (1993b). Protein unfolding pathways explored through molecular dynamics simulations. *J. Mol. Biol.* **232**, 600–619.

Dandekar, T. & Argos, P. (1997). Applying experimental data to protein fold prediction with the genetic algorithm. *Protein Eng.* **10**, 877–893.

Daura, X., Oliva, B., Querol, E., Avilés, F. X. & Tapia, O. (1996). On the sensitivity of MD trajectories to changes in water-protein interaction parameters, the potato carboxypeptidase inhibitor in water as a test case for the GROMOS force field. *Proteins: Struct. Funct. Genet.* **25**, 89–103.

Fan, P., Kominos, D., Kitchen, D. B., Levy, R. M. & Baum, J. (1991). Stabilization of α-helical secondary structure during high-temperature molecular-dynamics simulations of α-lactalbumin. *Chem. Phys.* **158**, 295–301.

Gelin, B. & Karplus, M. (1977). Mechanism of tertiary structural change in hemoglobin. *Proc. Natl Acad. Sci. USA,* **74**, 801–805.

Gehrmann, J., Alewood, P. F. & Craik, D. K. (1998). Structure determination of the three disulfide bond isomers of α-conotoxin GI: a model for the role of disulfide bonds in structural stability. *J. Mol. Biol.* **278**, 401–415.

Gutin, A. M., Abkevich, V. & Shakhnovich, E. (1995). Is burst hydrophobic collapse necessary for protein folding? *Biochemistry,* **34**, 3066–3076.

Hass, G. M. & Ryan, C. A. (1980). Cleavage of the carboxypeptidase inhibitor from potatoes by carboxypeptidase A. *Biochem. Biophys. Res. Commun.* **97**, 1481–1486.

Hass, G. M. & Ryan, C. A. (1981). Carboxypeptidase inhibitor from potatoes. *Methods Enzymol.* **80**, 778–791.

Jones, D. T. (1996). Potential energy functions for threading. *Curr. Opin. Struct. Biol.* **6**, 212–216.

Jorgensen, W. L., Chandrasekhar, J., Madura, J., Impey, R. W. & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935.

Koehl, P. & Delarue, M. (1994). Polar and nonpolar atomic environments in the proteins core: impli-

cations for folding and binding. *Proteins: Struct. Funct. Genet.* **20**, 264–278.

Kraulis, P. J. (1991). MOLSCRIPT, a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946–950.

Lazaridis, T. & Karplus, M. (1997). ''New view'' of protein folding reconciled with the old through multiple unfolding simulations. *Science,* **278**, 1928–1931.

Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.

Levitt, M. & Lifson, S. (1969). Refinement of protein conformations using a macromolecular energy minimization procedure. *J. Mol. Biol.* **46**, 269–279.

Levitt, M., Gerstein, M., Huang, E., Subbiah, S. & Tsai, J. (1997). PROTEIN FOLDING: the endgame. *Annu. Rev. Biochem.* **66**, 549–579.

Li, A. & Daggett, V. (1994). Characterization of the transition state of protein unfolding by use of molecular dynamics: chymotrypsin inhibitor 2. *Proc. Natl Acad. Sci. USA,* **91**, 10430–10434.

Lin, S. L. & Nussinov, R. (1995). A disulfide-reinforced structural scaffold shared by small proteins with diverse functions. *Nature Struct. Biol.* **2**, 835–837.

Lüthy, R., Bowie, J. U. & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature,* **356**, 83–85.

MacKerell, A. D., Jr, Bashford, D., Bellott, M., Dunbrack, R. L., Jr, Evanseck, J., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo., T., Nguyen, D. T., Prodham, B., Rieher, W. E., III, Roux, B., Schlenkirch, M., Smith, J., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D. & Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of protein. *J. Phys. Chem.* **B102**, 3586–3616.

Mark, A. E. & van Gunsteren, W. F. (1992). Simulation of the thermal denaturation of hen egg white lysozyme: trapping the molten globule state. *Biochemistry,* **31**, 7746–7748.

Marston, F. A. O. (1986). The purification of eukaryotic polypeptides synthesized in *Escherichia coli. Biochem. J.* **240**, 1–14.

Mirny, L. A., Abkevich, V. & Shakhnovich, E. (1996). Universality and diversity of the protein folding scenarios: a comprehensive analysis with the aid of a lattice model. *Folding Design,* **1**, 103–116.

Neria, E., Fischer, S. & Karplus, M. (1996). Simulation of activation free energies in molecular systems. *J. Chem. Phys.* **105**, 1902–1921.

Novotny, J., Bruccoleri, R. E. & Karplus, M. (1984). An analysis of incorrectly folded protein models: implications for structure predictions. *J. Mol. Biol.* **177**, 787–818.

Oliva, B., Daura, X., Querol, E., Avilés, F. X. & Tapia, O. (1995). Structure and atomic fluctuation patterns of potato carboxypeptidase A inhibitor protein. *Eur. Biophys. J.* **24**, 1–11.

Park, B. & Levitt, M. (1996). Energy functions that discriminate X-ray and near native folds from well-structured decoys. *J. Mol. Biol.* **258**, 367–392.

Rees, D. C. & Lipscomb, W. N. (1982). Refined crystal structure of the potato inhibitor complex of carboxypeptidase A at 2.5 Å resolution. *J. Mol. Biol.* **160**, 475–498.

Speed, M. A., Wang, D. I. C. & King, J. (1995). Multimeric intermediates in the pathway to the aggre-

gated inclusion body state for the P22 tailspike polypeptide chains. *Protein Sci.* **4**, 900–908.

Speed, M. A., Wang, D. I. C. & King, J. (1997). Conformation of the P22 tailspike folding and aggregation intermediates probed by monoclonal antibodies. *Protein Sci.* **6**, 99–100.

Thornton, J. M. (1981). Disulphide bridges in globular proteins. *J. Mol. Biol.* **151**, 261–287.

Tirado-Rives, J. & Jorgensen, W. L. (1993). Molecular dynamics simulations of the unfolding of apomyoglobin in water. *Biochemistry*, **32**, 4175–4184.

van Vlijmen, H. & Karplus, M. (1997). PDB-based protein loop prediction: parameters for selection and methods for optimization. *J. Mol. Biol.* **267**, 975–1001.

Vishveshwara, S., Ravishanker, C. & Bevirdge, D. L. (1993). Differential stability of β-sheets and α-helices in β-lactamase: a high temperature molecular dynamics study of unfolding intermediates. *Biophys. J.* **65**, 2304–2312.

Whitlow, M. & Teeter, M. M. (1986). An empirical examination of potential energy minimization using the well-determined structure of the protein crambin. *J. Am. Chem. Soc.* **108**, 7163–7172.