PROTEINS: Structure, Function, and Genetics 6:405–417 (1989)

Use of Restrained Molecular Dynamics in Water to Determine Three-Dimensional Protein Structure: Prediction of the Three-Dimensional structure of *Ecballium elaterium* Trypsin Inhibitor II

Laurent Chiche,[1] Christine Gaboriaud,[2] Annie Heitz,[3] Jean-Paul Mornon,[2] Bertrand Castro,[3] and Peter A. Kollman[1]

[1]Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, California 94143-0446; [2]Laboratoire de Mineralogie-Cristallographie, CNRS UA09, Universites P6 et P7, T16, 75252 Paris, cedex 05, France; [3]Centre CNRS-INSERM de Pharmacologie-Endocrinologie, 34094 Montpellier, cedex 2, France

ABSTRACT **Refinement of distance geometry (DG) structures of EETI-II (Heitz et al.: *Biochemistry* 28:2392–2398, 1989), a member of the squash family trypsin inhibitor, have been carried out by restrained molecular dynamics (RMD) in water. The resulting models show better side chain apolar/polar surface ratio and estimated solvation free energy than structures refined "in vacuo." The consistent lower values of residual NMR constraint violations, apolar/polar surface ratio, and solvation free energy for one of these refined structures allowed prediction of the 3D folding and disulfide connectivity of EETI-II. Except for the few first residues for which no NMR constraints were available, this computer model fully agreed with X-ray structures of CMTI-I (Bode et al.: *FEBS Lett.* 242:285–292, 1989) and EETI-II complexed with trypsin that appeared after the RMD simulation was completed. Restrained molecular dynamics in water is thus proved to be highly valuable for refinement of DG structures. Also, the successful use of apolar/polar surface ratio and of solvation free energy reinforce the analysis of Novotny et al. (*Proteins* 4: 19–30, 1988) and shows that these criteria are useful indicators of correct versus misfolded models.**

Key words: protein tertiary structure, incorrect and correct folding, molecular surfaces, solvation free energy, solution and crystal structure, disulfide connectivity determination, squash inhibitor family

INTRODUCTION

A new family of trypsin inhibitors, known as the squash family, has been recently isolated from various cucurbitacae plants.[1–4] All the members are highly homologous and, despite their very small size—about 30 residues, three disulfide bridges—their association constant with trypsin is among the largest reported for serine proteases ($5*10^{10}$–$5*10^{11}$).[2] These inhibitors are thought to be the smallest serine protease inhibitors and, possibly the smallest rigid proteins.[2,3]

Although knowledge of three-dimensional structures for these new inhibitors would be of considerable interest, very few structural studies have been reported.

A secondary structure with 40–45% of helix has been proposed based on circular dichroism studies and Chou and Fasman analysis[5]; modeling on the basis led to an assignment for the three disulfide bridges quite different from a former one based on comparison with wheat germ agglutinin.[6]

However, we recently reported two-dimensional NMR studies and distance geometry (DG) calculations conducted on a 28 residues synthetic peptide bearing the sequence of EETI-II, a member of the squash family extracted from *Ecballium elaterium*.[7] We showed that EETI-II lacks extensive classic secondary structure, alpha-helix, or beta-sheet, and is mainly composed of loops and turns arranged around a core of disulfide bridges that maintain the molecular conformation; also, a new assignment for the three disulfide bridges could be proposed as the most likely based on the number of violations of interproton distances deduced from the NMR spectra.

However, DG calculations do not incorporate detailed energy terms. The resulting structures often display significant deviations from ideal covalent geometry and may lack favorable electrostatic interactions if they are not introduced as constraints.[8,9]

Received June 7, 1989; revision accepted August 31, 1989.

Address reprint requests to Laurent Chiche, Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA 94143-0446.

Abbreviations used: CMTI-I: cucurbita maxima trypsin inhibitor I; DG: distance geometry; EETI-II: *ecballium elaterium* trypsin inhibitor II; NMR: nuclear magnetic resonance; REM: restrained energy minimization; RMD: restrained molecular dynamics; RMS: root mean square.

Thus the quality of DG structures can be greatly improved by molecular mechanics or molecular dynamics refinement using a well-tested empirical potential energy function with the NMR constraints included as additional pseudoenergy terms[9–14]; these refinements have also been shown to improve the agreement with the experimental NMR data.[10,13,14] Hence, in order to improve the accuracy of the solution structure of EETI-II and to insure the determination of the disulfide connectivity, restrained energy minimization (REM) and restrained molecular dynamics (RMD), in vacuo and in water surrounding, were performed on DG structures of EETI-II[7] and are described here; molecular surfaces and estimated free energies of solvation are reported and compared to those of other proteinase inhibitors and small proteins.

We will show that the use of explicit solvent in the refinement is important because it helps to make the exposed surface more like others proteins. Interestingly it also lead to lower violations of the NMR data and permits a rational assignment of the disulfide connectivity which confirm our previous one. This disulfide connectivity is consistent with the recent X-ray structure0 of CMTI-I1,[5] a member of the squash family, that appeared when this work was just completed. The crystallographical model of the complex between porcine trypsin and EETI-II, almost completed, will be presented here for comparative purposes.

METHODS

Energy Calculations

Energy minimizations and molecular dynamics were carried out on either a VAX 8650 or a FPS 264 computer using the program AMBER.[16] We used an all atom force field[17] to compute the intrinsic strain energy, and added potential energy terms representing interproton distance restraints and dihedral angle constraints arising from the NMR study.[7] The additional energy terms used here are described below.

Interproton distance potential

The potential energy term Ed used for interproton distances constraints is a simple half-parabolic function successfully applied by others,[10,11,13] although more sophisticated functions have been proposed.[12,14]

$$Ed = 0 \text{ if } 0 < r < ro$$
$$Ed = Kd^*(r - ro)^2 \text{ if } ro < r < ro + 1$$

where r is the actual distance between protons, and ro is the upper bound on the distance as determined by NMR. During the RMD runs, the term was taken as linear for large discrepancies in order to avoid too large constraining forces that could reduce the conformational search[13]:

$$Ed = Kd^*[2(r - ro) - 1] \text{ if } ro + 1 < r$$

Force constants of 10–40 kcal·mol^{-1}·Å^{-2} have often been used for interproton distance constraints,[11–14] but values as low as 0.6 and as high as 240 kcal·mol^{-1}·Å^{-2} have also been reported.[9,20] We choosed to use here values of either 4 or 40 kcal·mol^{-1}·Å^{-2} for the REM calculations.

For constraints for which proton stereospecific assignment was not possible from the NMR study, three different methods have been used:

Heavy atom constraints. Since the AMBER program does not handle the pseudoatoms used in the DG calculations, we apply the constraints to the heavy atoms that bear the equivalent protons; this is close to applying the constraints through the united atom representation as described by others.[12] The same corrections on the distance constraints previously used in the DG calculation were applied here since the distances between heavy atoms and protons are close to the ones between psuedoatoms and protons. A listing of the NMR constraints along with the corrections due to pseudoatoms are available as supplementary material of reference 7.

Stereospecific constraints. To do this, the two closest protons among all possible couples of equivalent atoms are selected to apply the constraint. No corrections were applied to the distances in this case.

R-6 weighting. A single mean actual distance was computed and used for all the equivalent protons: $r = (\langle r^6 \rangle)^{1/6}$

This quantity is useful as it is heavily weighted toward the shorter distance[18] but still increases significantly with the number of equivalent protons. To avoid larges constraining forces in these special cases where NMR data are the least accurate, the corresponding ro values were increased with the number of equivalent protons averaged:

ro*1.1	when	3 or 4	atoms	are averaged
ro*1.2	when	6	atoms	are averaged
ro*1.3	when	9	atoms	are averaged

Dihedral angle potential

The potential energy term Et for those dihedral angles that could be evaluated by NMR[7] was taken identical to the torsional energy strain, except that a range of ±50° around the predicted value is allowed:

$$Et = Kt^*[1 + \cos(t - t1)] \text{ if } t >= t1 = to + 50$$
$$Et = 0 \text{ if } t2 < t < t1$$
$$Et = Kt^*[1 + \cos(t - t2)] \text{ if } t <= t2 = to - 50$$

were t is the actual dihedral angle, and to is the predicted value.

For the REM calculations, we used values of 4 or 40 kcal·mol^{-1} for Kt, leading to a similar effect of a 90° dihedral deviation and 1 Å distance deviation.

Computational Procedures:

"In vacuo" simulation. We used a 8 Å residue based cut-off for computing the nonbonded pair list; the dielectric constant was distance dependant to mimic the screening effect of the solvation shell.[19]

The DG structures were first minimized (300 cycles, steepest descent) with the main chain constrained in its position to remove bad contact arising from the introduction of hydrogen atoms instead of the pseudoatoms used in the DG calculations. Then 500 cycles steepest descent REM (restraint energy minimization) without main chain constraint were applied followed by full conjugate gradient refinement until the RMS gradient fell below 0.01 (maximum 10,000 cycles).

Simulations in water surrounding. For simulations in water, the distance geometry structures were immersed in a large box of Monte Carlo TIP3P water molecules[21]; then all water molecules farther than (maxdist + 6) Å from the protein center, or nearer than 1.5 Å from any protein atom were removed (with maxdist the larger distance between the protein center and any other protein atom). The water molecules were aligned along the electric field before simulations, and a spherical half-parabola potential with weak force constant (0.5 kcal · mol^{-1} · $Å^{-2}$) was applied to prevent evaporation during heating phases. The covalent bond lengths were kept constant by applying the SHAKE algorithm.[22]

The newly introduced water molecules were allowed to move, with the protein frozen to remove bad contacts between protein and water (Belly option in AMBER) for 200 steepest descent cycles. Then, the whole system was minimized with explicit treatment of bond vibrations not containing protons to reduce the high strain usually present in DG structures (200 cycles steepest descent). Then the SHAKE option was turned on for all bonds and 200 conjugate gradient cycles were applied.

The system was then submitted to RMD (restrained molecular dynamics) simulations with initial velocities taken from Maxwellian distribution at 10°K, and the translational and rotational motion was removed. This motion was removed again every 600 steps.

The MD time step was Dt = 0.0015 ps, the nonbonded list was updated every 20 steps, and the temperature was regulated by coupling the system to a heat bath with a coupling constant t = 0.1 ps.

The temperature was quickly increased from 10°K to 300°K, then the system was equilibrated for 5 ps at 300°K, with force constants for NMR constraints Kd = 1 kcal·mol^{-1}·$Å^{-2}$ and Kt = 1 kcal·mol^{-1}. During the following 5 ps, both Kd and Kt were gradually increased from 1 to the maximum value of 10, at 300°K. Then the temperature was quickly increased and kept at 600°K for 5 ps, after which the system was allowed to evolve for 5 ps at 300°K.

Final REM cycles were then applied with stereospecific constraints and the same high values of the force constants to lead to the refined structures.

Residual Violations of the NMR Constraints

The residual violation of the NMR constraints (RVIOL) was computed for each structure using either the heavy atoms constraints or the stereospecific constraints described above:

RVIOL = sum over all the constraints of delta with

$$\text{delta} = r - ro \quad \text{if } r > ro$$
$$\text{delta} = 0 \quad \text{if } r <= ro$$

The residual violation computed on the stereospecific constraints is significantly larger than the one computed on heavy atoms or psuedoatoms constraints; this occurs because the corrections made on upper bound distances due to the use of pseudo or heavy atoms suppose a good alignment between these and the real atoms: misalignment results in shortening the distance between pseudo or heavy atoms more than between real atoms and makes the constraints on psuedo or heavy atoms easier to satisfy.

Molecular Surfaces

Molecular surfaces were computed with the program MS[23], using a sphere probe radius of 1.4 Å and a density of 0.5 (approximately 2.5 points/$Å^{12}$). The surface dots were grouped following two distinct classifications:

By residue type: All the dots belonging to side chains containing only C and S atoms form the apolar side chain surface; those belonging to other side chains form the polar side chain surface.

By atom type: All dots belonging to side chain C or S atoms form the apolar side chain surface; those belonging to side chain N or O atoms form the polar side chain surface.

The sum of surface area associated with each dot over all the dots of a class furnishes the surface area of the class. The ratios of apolar/polar surface areas were computed for the two classifications. In order to take in account the different intrinsic polarities of the compounds, we computed the same apolar/polar ratios for theoretical extended structures: this was done by calculating the apolar and polar surface areas for all amino acids in tripeptides GLY-X-GLY in extended conformations (Phi = −120, Psi = 140, extended side chains). Then summation of the surface areas for the residues in the sequence allowed estimation of apolar/polar surface ratios for extended, unfolded proteins.

We designate as the "folding effect" on the apolar/polar surface ratio the ratio of this value for the folded protein versus the estimated value for the theoretical unfolded protein.

TABLE I. Restrained Energy Minimization of the S1-DG Structure With Different NMR Constraints

	Stereospecific constraints				Heavy atom constraints			
	K* = 4		K = 40		K = 4		K = 40	
	C†	R†	C	R	C	R	C	R
Potential energy*	−243	−276	−102	−276	−231	−285	−199	−279
Residual violation‡	6.0	12.0	1.9	12.7	8.6	26.0	3.1	17.2
Deviation from the DG structure**	1.1	1.2	1.4	1.4	0.9	1.2	1.0	1.1

*Force constants K and potential energies are in kcal/mol.
†C = structures minimized with the NMR constraints; R = released structures minimized further without constraints.
‡The values are the sum of violations of all heavy atom constraints.
**RMS deviation between minimized and DG structures fitted for main chain heavy atoms N, CA, C, O of residues 1–28.

Solvation Free Energy

The free energy of solvation was evaluated with the program of Eisenberg and McLachlan.[24] It should be noticed that this program uses the solvent-accessible area defined by Lee and Richards.[25] This surface is not the molecular surface described above, but is 1.4 Å away from the contact surface; it is larger and contains only convex parts, although the MS surface contains both convex (contact surface) and concave (reentrant surface) parts.

The computed solvation contribution to the free energy of protein folding (SFE) is given by the solvation free energy difference between folded proteins and theoretical extended, unfolded proteins. This energy has an absolute value that increase with the size of the compound, probably because larger volumes make easier to bury large hydrophobic groups during the folding process, thus lowering the solvation free energy of the folded protein. We found that a useful way to make comparisons is to compute the solvation free energy per residue as reported in Tables II, III, and V, although these values still increase with the size of the protein (Table III).

RESULTS AND DISCUSSIONS

In Vacuo Simulations

In order to choose both the method to apply the NMR constraints and the force constant for these constraints, we carried out eight REM starting with our best DG structure, with heavy atom and stereospecific constraints and with force constant of 4 kcal·mol^{-1}·Å^{-2} and 40 kcal·mol^{-1}·Å^{-2}. Each structure was further refined without constraints to see how stable the constrained structures are.

The potential energies and the residual violations are reported in Table I for each structure along with the RMS deviation from the starting DG structure: The refined structures stay close to the DG structure, with a maximum deviation of 1.4 Å, and the stereospecific constraints afford roughly similar energies but clearly lower residual violations and more stable structures than heavy atom constraints. For stereospecific constraints, the 4 kcal/mol value seems reasonable with rather low energy and mean residual violation value; it also give the most stable structure with the smaller deviation between constrained and released structure (0.34 Å). The situation here is quite different than that of the DG calculations where nothing is known and pseudoatoms have to be used. Here we already have an approximate 3D model and the REM calculations are not expected to produce large deviations that could reverse the stereospecific assignment, but rather afford a local minimum energy conformation close to the initial one.[9,13]

On this basis, we decided to run REM calculations for the six best DG structures using stereospecific constraints and a force constant of 4 kcal·mol^{-1}·Å^{-2}.

The results are reported in Table II.

In every case except S6 the residual violation has been lowered by the REM; however the refined structures diverge more than the DG structures (ref. 7 and Table VI) and the best potential energy (S3) does not correspond to the best residual violation (S1). Although the best DG structure still has the lowest residual violation after minimization, the fact that the energy does not follow the same way precludes any conclusion on the disulfide bridging at this point.

Since they were recently proposed as a good tool to distinguish between correct and incorrect folding,[26,27] we computed for each structure the side chain apolar/polar surface area and the solvation free energy.[24] The values are reported in Table II. We also computed the same values for some other proteinase inhibitors and proteins for which X-ray structures are available from the Protein Data Bank[28] for comparison (see Table III). The backbone surface area seems less representative of the correctness of the folding[26,27] and have not been used here.

Since we are dealing with very small proteins, we felt that comparison with larger ones could only be achieved if we take in account the intrinsic polarity of the molecules; hence we divided the ratios of apo-

TABLE II. Energies, Residual Violations, Surface Properties, and Solvation Free Energies for the Six Best DG Structure Refined "In Vacuo" With Stereospecific NMR Constraints

Structure*	Potential energy† Total	−Electro** −h bond**	Residual violation‡	Folding effect on apolar/polar surface ratios: Residue based	Atom based	Solvation free energy†	Solvation free energy per residue
S1-DG	4,041		17.9	0.85	0.87	−12.5	
REM	−259	112	11.6	0.95	1.03	−10.8	−0.39
S2-DG	8,253		22.0	0.78	0.77	−16.6	
REM	−237	101	16.3	0.80	0.91	−15.1	−0.54
S3-DG	4,341		27.9	0.90	0.86	−12.8	
REM	−292	116	20.4	1.02	1.42	−8.8	−0.31
S4-DG	5,975		23.1	0.75	0.73	−15.3	
REM	−228	122	18.5	0.80	0.91	−11.7	−0.42
S5-DG	5,037		24.9	0.69	0.77	−14.2	
REM	−257	113	17.9	0.73	0.96	−14.2	−0.51
S6-DG	6,331		20.7	0.76	0.70	−14.9	
REM	−227	128	21.9	0.72	0.86	−16.1	−0.57

*Disulfide connectivity: S1(2–19,9–21,15–27); S2(2–19,9–15,21–27); S3(2–9,15–19,21–27); S4(2–15,9–19,21–27); S5(2–19,9–27,21–15); S6(2–7,9–15,19–21).
†Potential and solvation energies are in kcal/mol.
‡Residual violations are for stereospecific constraints.
**All the electrostatic and H-bond terms have been subtracted.

TABLE III. Surface Properties and Solvation Free Energies for Some Proteins and Serine Protease Inhibitors

PDB file name*	Folding effect on apolar/polar surface ratios: Residue based	Atom based	Solvation free energy†	Solvation free energy per residue†
4CPA(I)	0.66	0.81	−24.6	−0.66
1CRN	0.85	0.75	−31.5	−0.68
2OVO	0.75	0.85	−45.2	−0.81
5PTI	0.70	0.72	−56.5	−0.97
2SSI	0.70	0.80	−99.6	−0.93
1HMQA	0.68	0.67	−114.2	−1.01
2PTN	0.45	0.57	−235.1	−1.04

*4CPA(I) = potato carboxipeptidase inhibitor; 5PTI = basic pancreatic trypsine inhibitor; 1CRN = crambine; 2OVO = ovomucoid inhibitor; 2SSI = streptomyces subtilisin inhibitor; 1HMQA = hemerythrin; 2PTN-trypsin.
†Energies are in kcal/mol.

lar/polar surface for the folded structures by the values for theoretical extended conformations of the same protein. The values for the extended proteins were computed by a similar method as that described by Miller et al.,[29] except that here we use the molecular surface instead of the accessible surface.[30] There are rather large differences between the values reported by Miller et al.[29] for the apolar/polar surface ratios of extended amino acid residues and the values computed here (Table IV); from the values in Table IV and the corresponding values in reference 29, it is clear that the accessible surface emphasize the surface of the polar side chain end atoms versus the apolar part in the polar and charged residues more than the molecular surface. In any case, as long as one consistently uses a given surface in both known and unknown structures, it shouldn't matter whether one uses molecular or accessible surfaces.

What we see is that the apolar/polar ratios as well as the solvation free energies per residue are globally too high for our computer models of EETI-II, especially for the refined structures, when compared with the values of other proteins, even the small potato carboxipeptidase inhibitor (4 cpa, 37 residues). On the other hand, these values are again consistent with neither the potential energies nor the residual violation, and the structure S1 which clearly has the lowest residual violation has an apolar/polar ratio too high and a solvation energy too positive.

The increase of the apolar/polar ratios and of the

TABLE IV. Molecular Surface Areas (A^2) for Amino Acid Residues in Tripeptides Gly-X-Gly in an Extended Conformation

Residue	Total surface	Backbone surface	Side chain surface	
			Apolar	Polar
Ala	66	36	30	0
Arg	134	37	58	39
Asn	88	36	26	26
Asp	85	36	23	26
Cys	84	37	47	0
Gln	103	37	38	28
Glu	104	37	39	28
His	113	36	62	15
Ile	111	37	74	0
Leu	105	35	70	0
Lys	114	37	63	14
Met	115	37	78	0
Phe	124	34	90	0
Pro	83	25	58	0
Ser	69	35	20	14
Thr	84	35	35	14
Trp	142	34	148	0
Tyr	126	34	77	15
Val	93	37	56	0

solvation free energies during the REM could be ascribed to the lack of polar environment in the simulation. And indeed, examination of the refined structures on a graphics display device shows that the charged amino and carboxi termini and some polar and charged side chains tend to fold back toward the protein to find hydrogen bonding partners. We thus decided to solvate the DG structures with a simple shell of water and run RMD calculations, to see if we can improve the poor surface characteristics and solvation energy of the S1 structure; the RMD method was indeed already proved to be an efficient technique to generate low-energy conformations.[9,13]

Simulations in Water Surrounding

As indicated in methods, the protein was solvated by a 6 Å shell of water prevented from evaporation by a weak half-parabola potential. The RMD computation procedure was mainly taken from De Vlieg et al.[13] and slightly modified: the same force constant for the NMR constraints were used, and a procedure intermediate between the simple (10 ps) and the elaborate (50 ps) procedure of reference 13 was employed and is described above. A total RMD time of 20 ps was used: 5 ps equilibration at 300°K with Kd = 1 kcal·mol^{-1}·Å^{-2} and Kt = 1 kcal·mol^{-1} were followed by 5 ps during which Kd and Kt are raised to the maximum value of 10 kcal·mol^{-1}·Å^{-2}; then the system was heated at 600°K for 5 ps then cooled at 300°K for 5 ps.

Since the RMD technique is expected to yield larger deviations than REM, we do not want to restrain the conformational search by use of stereospecific NMR constraints; however, our previous experiments showed that heavy atom constraints do not yield good results; we thus decided to implement the $(<r^{-6}>)^{1/6}$ weighting scheme described by Clore et al.[18] for 3D structure determination with RMD.

At the end of the RMD runs, 2,000 conjugate gradient REM was performed to remove the kinetic energy from the system. We noticed that none of these final minimizations converged, probably due to a too smooth energy surface of the water molecules. The final RMS energy gradients were rather high (~14 kcal/(mol·Å)).

In order to limit the cpu-time use, the above method was applied only to three structures: S1 and S2, which have the best residual violations, and S3, which has the best potential energy. The energies, apolar/polar surface ratios, and the solvation energies are reported in Table V. In order to make residual violation comparisons availables, we also report

TABLE V. Energies, Residual Violation, Surface Properties, and Solvation Free Energies for Three Structures Refined in a Water Surrounding

	Potential energy*				Folding effect on apolar/polar surface ratios		Solvation* free energy	
	Prot-Prot							
	Total	Elect −h bond	Prot-Wat	Residual violation†	Residue based	Atom based	Total	Per residue
S1	−107	147	−867	5.2	0.67	0.70	−19.2	−0.69
S1‡				6.5 (7.9)	0.89	1.08	−9.6	−0.34
S2	−65	154	−952	9.6 (10.8)	0.77	0.72	−17.7	−0.63
S3	−166	144	−737	7.1 (14.0)	0.97	1.14	−13.5	−0.48

*Potential and solvation energies are in kcal/mol.

†The residual violation is the sum of violations of the stereospecific constraints; values in parentheses are for restraint energy minimizations with a force constant of 10 kcal/mol. The residual violation in parentheses are for restrained energy minimized structures "in vacuo" with force constants for the NMR restraints of K = 10 kcal/mol.

‡Values here are for the S1-DG structure refined with the same restrained molecular dynamic procedure conducted "in vacuo."

TABLE VI. Some RMS Deviations Between Computed Structures With Different Disulfide Bridges Assignment*

	s1	s2	s3	s4	s5	s6
S1	—	1.76	2.60	1.68	2.07	2.01
S2	1.66	—				
S3	3.21	3.93	—			

*The values reported are for fit of C-alpha of residues 5–28. Above diagonal entries are between structures refined "in vacuo"; below diagonal are between structures refined in water.

the values for structures S1, S2, and S3 of Table II further minimized with the same force constant used here, i.e., Kd = 10 kcal·mol^{-1}·Å^{-2} and Kt = 10 kcal·mol^{-1}; furthermore, an RMD calculation was run on S1 "in vacuo" with the same procedure used for the solvated proteins and is also reported.

Although we report the energy terms for solvated molecules, these cannot be compared with each other since the number of water molecules vary and protein-protein interactions can be modified by corresponding variations of water-water or protein-water interactions.

We can see that all the values of Table V are improved when compared with the Table II values: the solvation energy and the apolar/polar ratio are lowered, as could be expected by the use of explicit polar solvent, but also the residual violation is significantly reduced in all three cases; the RMD run "in vacuo" applied to S1 also give rather low residual violations, but the apolar/polar surface ratio is much larger.

It is interesting to notice that the values for the apolar/polar surface ratios and the solvation energy compare now quite well with the values in Table III for X-ray determined inhibitors and enzymes.

The structure S1 shows now the best values for the residual violation, the apolar/polar surface ratio, and the solvation free energy; on this basis, we consider the structure S1, shown in Figure 1, as very likely to be the one that bears the actual disulfide bridging, and the refined structure S1 is referred to below as the computer model of EETI-II.

Comparison and Analysis of the EETI-II Models

The RMS deviations for fit between various structures are given in Tables VI and VII. The largest deviations in Table VI are for the S3 structure: this could arise from the very special disulfide bridge arrangement with no overlapping. This structure is therefore less constrained than others, and this could also be responsible for the lowest energy in Table II. Others deviations are around 2 Å, meaning that the global folding is rather similar for the 5–28 segment of all the structures and thus rather well defined by the NMR data.

If we compare the different models for the S1 structure (Table VII), the RMS deviations are also close to 2 Å, but the deviations are mainly located in three segments: the loop 1–6 and the two external turns 22–25 and, to a lesser extent, 16–19. Superposition of structures S1-DG and S1-RMD is depicted in Figure 2.

As shown on Figure 3, this can be related to the number of NMR constraints and is likely arising from enhanced flexibility for these segments in solution. Hence the loop 1–6 that bears the reactive bond ARG4-ILE5 is indeed the least well-defined part of the molecule; this is not very surprising since the reactive loop of uncomplexed inhibitors were shown to have the largest thermal motion of the molecules by X-ray structure determination.[31,32] This flexibility will probably allow the loop to adapt its conformation to slighly different receptors.

Simulation for Potato Inhibitor in Water

To evaluate the deviation from the X-ray structure that could be induced by the simulation procedure we used, we carried out the same RMD calculations in water on the potato inhibitor of carboxipeptidase A[33] from the protein databank (4 cpa),[28] but without any constraints. However, we removed the three first residues since they have very few contacts with the rest of the protein and they are very poorly resolved in the X-ray structure.[33] The first residue is never seen; this probably arise from high mobility of the 1–7 segment. This deletion rendered the molecule more spherical and allowed us to solvate the molecule more easily.

At the end of the RMD calculation, the initial and the final structures are fitted and the RMS deviation obtained is 1.67 Å, a rather reasonable value. The two structures deviate mainly in external loops; however, there are no restraints here, and we can expect smaller deviations when restraints based on NMR data are applied.

Comparison of the EETI Model With Recent X-Ray Structures

While we were writing up our results, the X-ray crystal structure of the complex between bovine B-trypsin and CMTI-I, another member of the squash inhibitor family, appeared in the literature.[15] The disulfide connectivity thus determined for CMTI-I is in complete agreement with the one we proposed for EETI-II after our DG calculations[7]; in this study we strongly support that assignment on different grounds.

At the same time the crystallographical model of the complex between porcine trypsin and EETI-II, almost completed, became available and will be fully described elsewhere (Fig. 4). The present R value (defined as $\Sigma\|Fo| - |Fc\| / \Sigma|Fo|$) for 6,501 reflections between 5.0 and 1.9 Å is 0.180. One par-

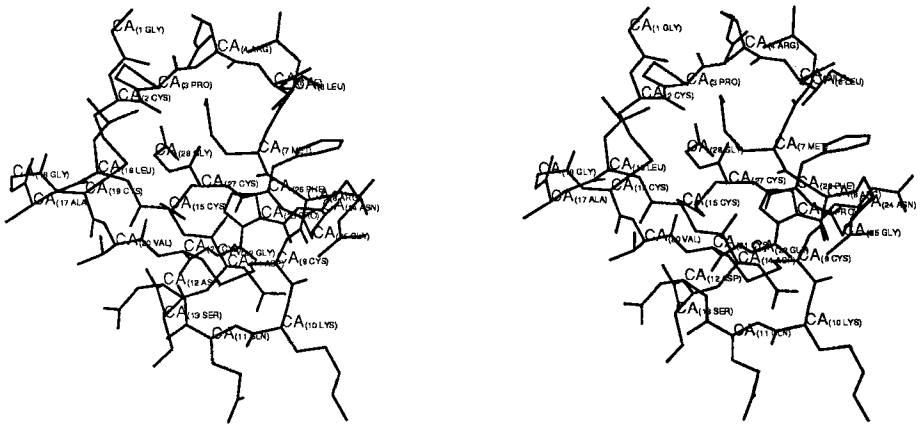

Fig. 1. Final refined structure S1-RMD.

TABLE VII. RMS Deviations Between DG Structures, Structures Refined In Vacuo, and Structures Refined in Water*

	Vacuo-DG	Water-DG	Vacuo-water
S1	0.94	1.99	2.05
S2	1.11	1.50	1.49
S3	1.33	1.17	1.46

*RMS deviations are for C-alpha atoms of residues 1–28.

ticular feature of this crystallographical data set is the very large amount of solvent content in the crystals (60% according to Matthews[34]) which should prevent important modification of the inhibitor structure due to crystal packing constraints; only one intermolecular contact is observed with ALA17. Refinement of the description of solvent structure and the LYS10 and GLN11 side chains that do not properly fit in electron density is in progress.

Though we do not have the coordinates of CMTI-I, a rough comparison can be made between our computer model and these two X-ray structures:

The 1–6 loop displays the same conformation in the two X-ray structures but is clearly different in our model; however, we already noticed that this is the least defined part; then, from residue 7 to 28, our model is very close to both X-ray structures with the short, irregular 3–10 helix looking very similar, although we lack the hydrogen bond between GLN11 O and ASP14 N (pitches GLN11 Ca . . . ASP12 CA and ASP12 CA . . . CYS15 CA and 5.9 and 5.1 Å in the model instead of 6.0 Å for a regular 3–10 helix). The only significant deviation that appears between CMTI-I and EETI-II is for residues PRO23 and ASN24, but these residues belong to an external turn, probably rather flexible, for which we have only few NMR constraints and for which the electron density is very weak in the X-ray study. Moreover, differences in sequences (22–25: LEHG in CMTI and GPNG in EETI) are likely to produce some local deviations; however, in both CMTI and EETI, GLY25 is located near CYS9, with the hydrogen bond between GLY25 O and CYS9 N that was induced from the 2D NMR experiments[7] (Fig. 5).

The rms deviation between the model and X-ray structure of EETI for backbone fit of residues 7 through 28 (atoms N, CA, C, and O) is 1.23 Å; the same value for the DG, REM, and in vacuo RMD structures are, respectively, 1.70, 1.78, and 2.04 Å. The improvement afforded by the molecular dynamic run in water is thus obvious for this sufficiently constrained part of the protein. The effect of the molecular dynamics on the 1–6 loop is much less clear, but a qualitative evaluation of the improvement of the exposed surface that was discussed in above can be made. In the DG structure the large apolar residues ILE5, LEU6, and PHE26 point outside; in the computer model, as in the X-ray structures, ILE5 and LEU6 fall more flat on the surface, although in the model ILE5 fills approximately the space of LEU6 in the X-ray structures and LEU6 fills the space of ILE5. This move between the DG and the refined models thus brings an apolar group close to the PHE26 therefore lowering its exposed surface (the accessibility of PHE26 is lowered from 92% in the DG structure to 66% for the refined model), although the apolar side chain is ILE5 instead of LEU6.

As noted above, the water content of the inhibitor structure is very important, mainly because the active loop is linked to the central part of the molecule through a water molecule; this may possibly be part of the explanation for the importance of explicit solvent use during the refinement of the DG structures. Indeed, a water molecule has been trapped between the active loop and the core of the molecule in the RMD model, although the location is different from the X-ray structure. The water molecule in the X-ray structure lies along the SS bond between CYS2 and CYS19, not too far from the main chain carbon-

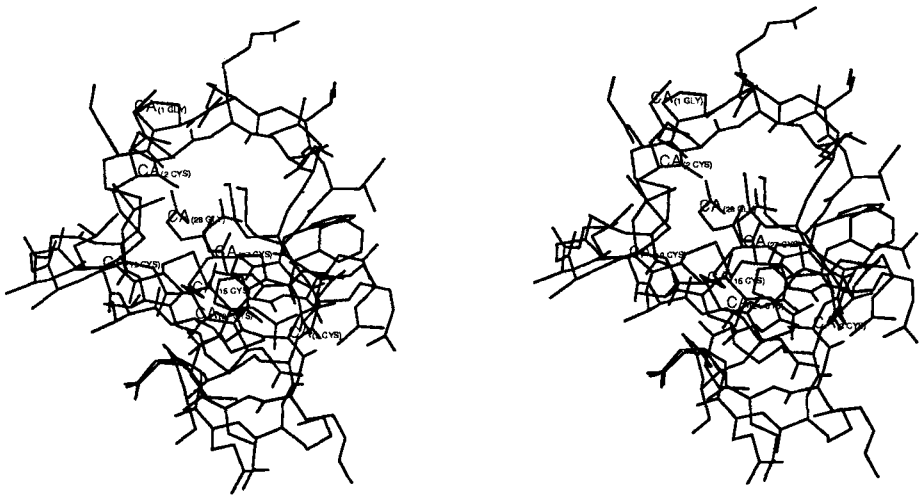

Fig. 2. Superposition of structure S1-RMD (labeled) and the starting S1-DG structure.

yls of residues 3 and 5 (distances: WAT O . . . PRO3 O = 3.50 A, WAT O . . . ILE5 O = 3.98 Å) (Fig. 6).

On the other hand, the water molecule in the RMD structure is located in a more polar area, surrounded by main chain atoms LEU6 N, MET7 N and O, and CYS27 N and O at respective distances of 3.64, 3.45, 2.96, 3.04, and 2.89 Å from the oxygen atom of the water molecule (Fig. 7).

Quite interestingly, the water molecules in the X-ray and RMD structures appear to occupy two distinct sites which correspond roughly to two of the three water molecules included in the CMTI-I X-ray structure.[15] In the RMD structure, the methyl group of MET7 is more buried than in the X-ray structure, filling approximately the space of the X-ray water molecule. Figure 8 shows superposition of S1-RMD and X-ray structures, with the water molecule of each model depicted as Van Der Waals spheres around the water oxygen atom.

We want to point out here the usefulness of joint NMR and X-ray structural studies: First the approximate DG structure[7] served as a starting geometry for the X-ray study. Later when the RMD solution structure described above was already completed, a partial X-ray model was made available; in this model, the turn 22–25 was basically undefined due to a very low electron density in this region. The coordinates of the S1-RMD structure for these few residues were then successfully used to fit the electron density, affording, after refinement, a significant improvement of the R factor. On the other hand, the X-ray structure allows confirmation of the validity of the refinement method and parameters used for the solution structure.

CONCLUSIONS

The computer simulation described here was completed before we learned about the two X-ray structures of the complex of CMTI-I and EETI-II with trypsin, and was conducted in a fully predictive way. It is therefore noteworthy that on one hand, the RMD simulation in water allowed a significant improvement of the initial DG structure, and on the other hand, that we have been able to predict the correct disulfide connectivity on rational grounds.

We thus think that, despite the much larger cputime, restrained molecular dynamics in water surrounding can be very useful and should be more extensively used. This refinement reduced the violations of the NMR constraints and the deviation from the X-ray structure, but also led to more realistic exposed surfaces. It is, however, of interest to develop other ways to take into account the solvation and hydrophobic effects while avoiding the time consuming use of explicit solvent. The introduction of solvation or hydrophobic interaction free energy term in the energy function during molecular dynamics calculations have been suggested[24,35] and are presently under investigation.

Also, the apolar/polar side chain surface area corrected for the intrinsic polarity of the protein as well as the solvation contribution to the free energy of folding per residue have been shown to be significant enough to allow correct prediction of the disulfide connectivity of EETI-II, although different connectivity can be accommodated without large changes in the folding. This is a successful application of the criteria proposed by Novotny et al.[27] for the distinction between incorrect and correct folding in a different case than the one they studied. These criteria thus seem reliable enough to be applied in different applications of prediction of protein folding.

Considering the improved knowledge of the 3D structure for the squash inhibitor family arising from the recent X-ray structure of complexed CMTI-

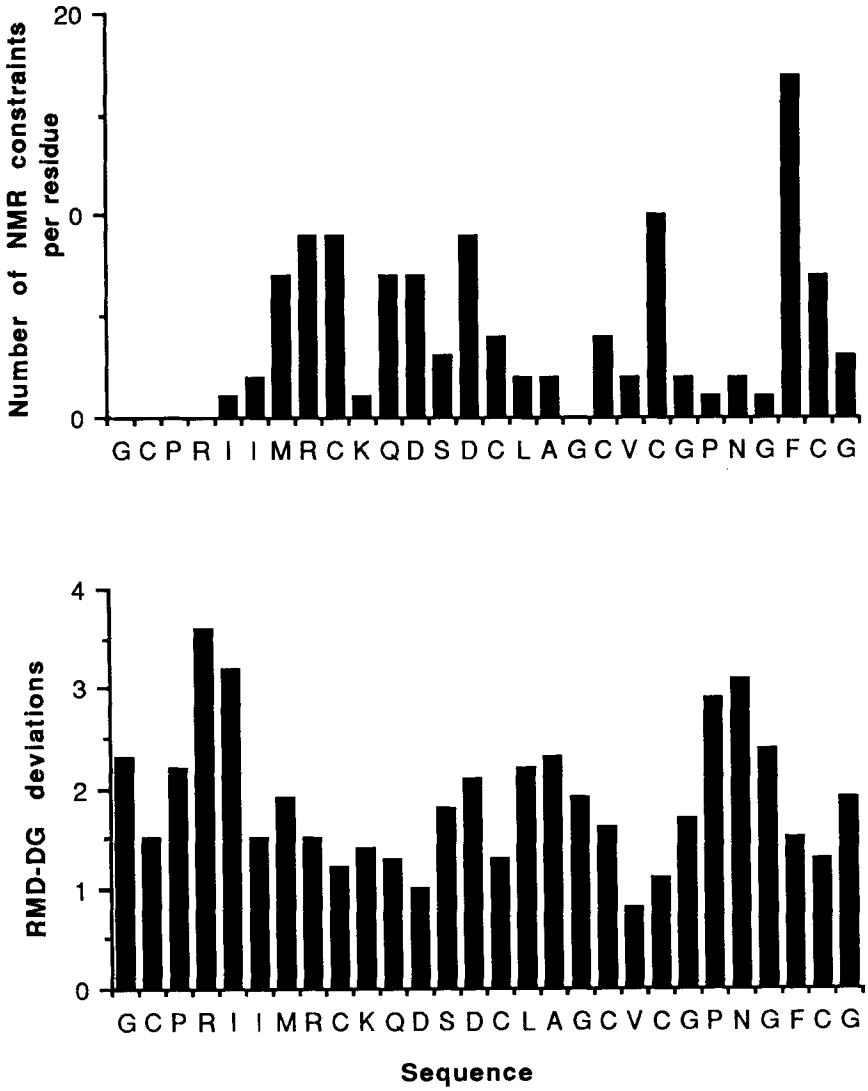

Fig. 3. Deviation between C-alpha atoms of structures S1-RMD and S1-DG and number of NMR constraints per residue reported along the sequence of EETI-II.

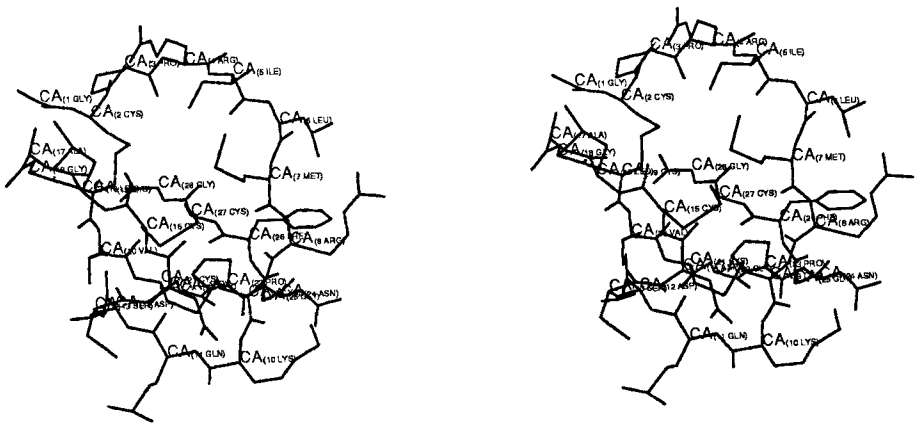

Fig. 4. X-ray structure of EETI- II complexed with trypsin.

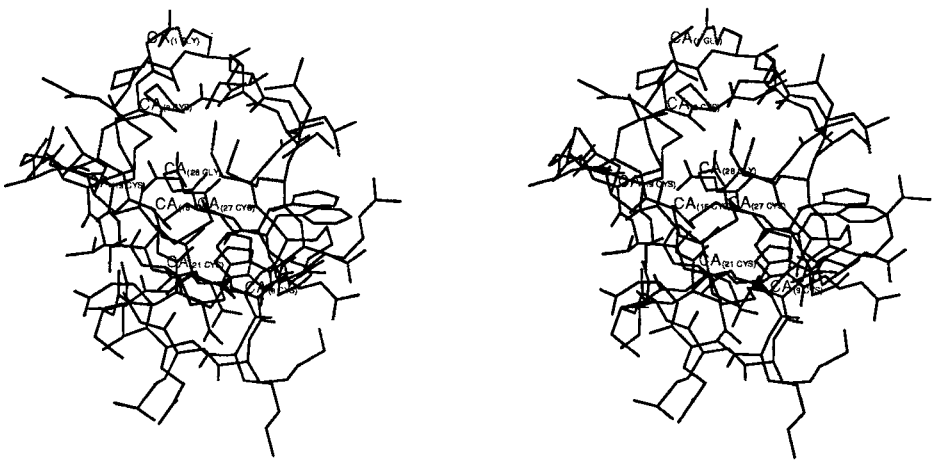

Fig. 5. Superposition of structure S1-RMD (labeled) and the X-ray structure of EETI-II.

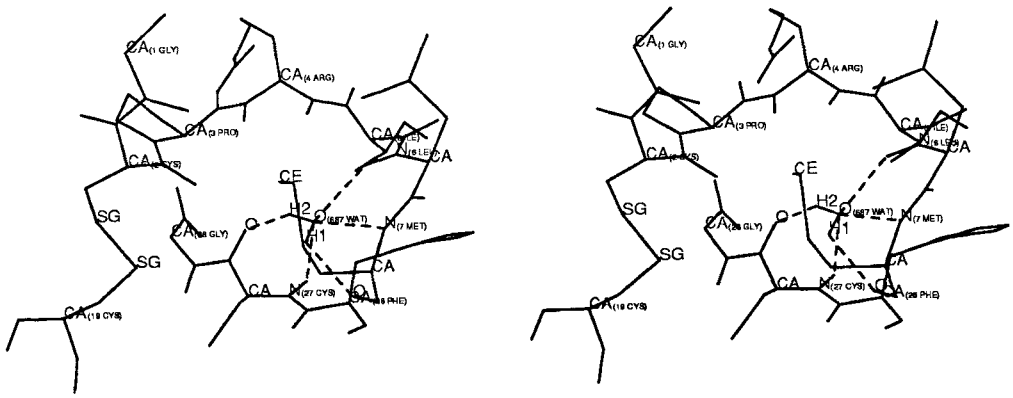

Fig. 6. Location of the buried water molecule in the X-ray structure of EETI-II.

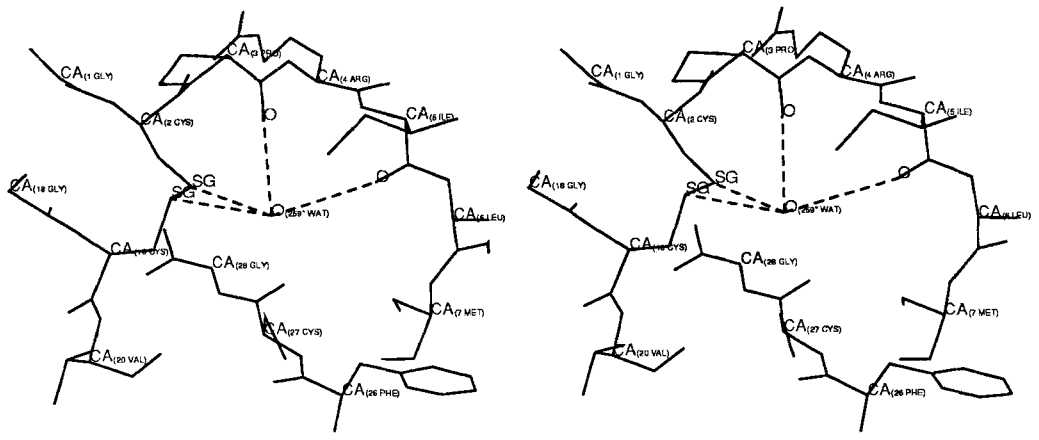

Fig. 7. Location of the buried water molecule in the S1-RMD structure.

I[15] and from the solution structure of the free inhibitor and the almost completed X-ray structure of the complex with porcine trypsin for EETI-II described here, and given their small size that allows total synthesis, these inhibitors are exciting substrates for modeling studies. Indeed, we are currently un-

Fig. 8. Partial view of superposed X-ray and S1-RMD structures with the buried water molecule of each structure.

dertaking computer simulations in order to predict and understand the effect of simple changes in the sequence of EETI-II on the affinity and selectivity toward serine proteases.

ACKNOWLEDGMENTS

This work was supported by a grant (to L.C.) from the National Science Fondation and the Centre National de la Recherche Scientifique (International exchange program) and by research support to P.A.K. through GM-29072 and DARPA grant ONR-34 (R. Langridge, P.I.). The use of the UCSF Computer Graphics Laboratory (NIH-RR-1081 to R. Langridge) is gratefully acknowledged.

REFERENCES

1. Favel, A., Mattras, H., Coletti-Previero, M.-A., Zwilling, R., Castro, B. Protease inhibitors from *Ecballium elaterium* seeds. Int. J. Pept. Protein Res. 33:202–208, 1989.
2. Wieczorec, M., Otlewski, J., Cook, J., Parks, K., Leluk, J., Wiliowska, A., Polonavski, A., Wilusz, T., Laskowski, M. The squash family of serine protease inhibitors. Aminoacid sequences and association equilibrium constants of inhibitors from squash, summer squash, zucchini, and cucumber seeds. Biochem. Biophys. Res. Commum. 126:646–652, 1985.
3. Joubert, F. Trypsin isoinhibitors from momordica repens seeds. Phytochemistry 23:1401–1406, 1984.
4. Hojima, Y., Pierce, J.V., Pisano, J.J. Pumpkin seeds inhibitor of human factor XIIa (activated Hageman factor) and bovine trypsin. Biochemistry 21:3741–3750, 1982.
5. Hider, R.C., Drake, A.F., Morrison, I.E.G., Kupryszewski, G., Wilusz, T. Structure analysis of trypsin inhibitors isolated from cucurbitacae seeds. Int. J. Pept. Protein Res. 30:397–403, 1987.
6. Siemon, I.Z., Wilusz, T., Polanowski, A. On the genetic and structural similarities between the squash seeds polypeptide trypsin inhibitor and wheat germ agglutinin. Mol. Cell. Biochem. 60:159–161, 1984.
7. Heitz, A., Chiche, L., Le-Nguyen, D., Castro, B. 1H 2D NMR and distance geometry study of the folding of ecballium elaterium trypsin inhibitor, a member of the squash inhibitors family. Biochemistry 28:2392–2398, 1989.
8. Wuthrich, K. Protein structure determination in solution by nuclear magnetic resonance spectroscopy. Science 243: 45–50, 1989.
9. Kaptein, R., Boelens, R., Schleek, R.M., Van Gunsteren, W.F. Protein structures from NMR. Biochemistry 27: 5389–5395, 1988.
10. Kaptein, R., Zuiderverg, E.R.P., Scheek, R.M., Boelens, R., Van Gunsteren, W.F. A protein structure from nuclear magnetic resonance data. Lac repressor headpiece. J. Mol. Biol. 182:179–182, 1985.
11. Moore, J.M., Case, D.A., Chazin, W.J., Gippert, G.P., Havel, T.F., Powls, R., Wright, P.E. Three-dimensional solution structure of plastocyanin from the green algua scenedesmus obliquus. Science 240:314–317, 1988.
12. Holak, T.A., Kearsley, S.K., Kim, Y., Prestegard, J.H. Three-dimensional structure of acyl carrier protein determined by pseudoenergy and distance geometry calculations. Biochemistry 27:6135–6142, 1988.
13. DeVlieg, J., Scheek, R.M., Van Gunsteren, W.F., Berendsen, H.J.C., Kaptein, R., Thomason, J. Combined procedure of distance geometry and restrained molecular dynamics techniques for protein structure determination from nuclear magnetic resonance data: Application to the DNA binding domain of Lac repressor from Escherichia coli. Proteins 3:209–218, 1988.
14. Nilges, M., Gronenborn, A.M., Brunger, A.T., Clore, G.M. Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints. Aplication to crambin, potato carboxypeptidase inhibitor and barley serine protease inhibitor 2. Protein Eng. 2:27–38, 1988.
15. Bode, W., Greyling, H.J., Hubert, R., Otlewski, J., Wilusz, T. The refined 2.0 A X-ray crystal structure of the complex between bovine B-trypsine and CMTI-I, a trypsin inhibitor from squash seeds (Cucurbita Maxima). FEBS Lett. 242: 285–292, 1989.
16. Singh, U.C., Weiner, P.K., Caldwell, J.W., and Kollman, P.K. AMBER (UCSF), version 3.0. Department of Pharmaceutical Chemistry, University of California San Francisco, 1986.
17. Weiner, S.J., Kollman, P.A., Nguyen, D.T., Case, D.A. An all atom force field for simulation of proteins and nucleic acids. J. Comput. Chem. 7:230–252, 1986.
18. Clore, G.M., Brunger, A.T., Karplus, M., Gronenborn, A.M. Application of molecular dynamics with interproton distance restraints to three-dimensional protein structure determination. A model study of crambin. J. Mol. Biol. 191:523–551, 1986.
19. Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S., Jr., Weiner, P. A new force field for molecular mechanical simulation of nucleic acids and proteins. J. Am. Chem. Soc. 106:765–784, 1984.
20. Tapin, M.J., Pastore, A., Norton, R.S., Freer, J.H., Campbell, I.D. High-resolution 1H NMR study of the solution structure of d-hemolysin. Biochemistry 27:1643–1647, 1988.
21. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., Klein, M.L. Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. 79:926–935, 1983.
22. Van Gunsteren, W.F., Berendsen, H.J.C. Algorithms for macromolecular dynamics and constraints dynamics. Mol. Phys. 34:1311–1327, 1977.
23. Connolly, M.L. Surfaces of proteins and nucleic acids. Science 221:709–713, 1983.

24. Eisenberg, D., McLachlan, A.D. Solvation energy in protein folding and binding. Nature 319:199–203, 1986.
25. Lee, B., Richards, F.M. The interpretation of protein structures: Estimation of static accessibility. J. Mol. Biol. 55: 379–400, 1971.
26. Novotny, J., Bruccoleri, R., Karplus, M. An analysis of incorrectly folded protein models: Implications for structure predictions. J. Mol. Biol. 177:787–818, 1984.
27. Novotny, J., Rashin, A.A., Bruccoleri, R.E. Criteria that discriminate between native proteins and incorrectly folded models. Proteins 4:19–30, 1988.
28. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M.J. The protein databank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
29. Miller, S., Janin, J., Lesk, A.M., Chothia, C. Interior and surface of monomeric proteins. J. Mol. Biol. 196:641–656, 1987.
30. Richard, F.M. Areas, volumes, packing, and protein structure. Annu. Rev. Biophys. Bioeng., 6:151–176, 1977.
31. McPhalen, C.A., James, M.N.G. Crystal and molecular structure of the serine protease inhibitor CI-2 from barley seeds. Biochemistry 26:261–269, 1987.
32. Bode, W., Epp, O., Huber, R., Laskowski, M., Jr., Ardelt, W. The crystal and molecular structure of the third domain of silver pheasant ovomucoid (OMSVP3). Eur. J. Biochem., 147:387–395, 1985.
33. Rees, D.C., Lipscomb, W.N. Refined crystal structure of the potato inhibitor complex of carboxipeptidase A at 2.5 Å resolution. J. Mol. Biol. 160:475–498, 1982.
34. Matthews, B.W. Solvent content of protein crystals. J. Mol. Biol. 33:491–495, 1968.
35. Richmond, T.J. Solvent accessible surface area and excluded volume in proteins. J. Mol. Biol. 178:63–89, 1984.