Combining MONSSTER and LES/PME to Predict Protein Structure from Amino Acid Sequence: Application to the Small Protein CMTI-1

Carlos Simmerling,[†,§] Matthew R. Lee,[†] Angel. R. Ortiz,[‡] Andrzej Kolinski,[‡,||] Jeffrey Skolnick,[‡,||] and Peter A. Kollman*[,†]

Contribution from the Department of Pharmaceutical Chemistry, University of California, 513 Parnassus, San Francisco, California 94143-0446, Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037

Received August 27, 1999. Revised Manuscript Received March 7, 2000

Abstract: A combined method for the prediction of protein tertiary structures from sequence is presented. This multistep procedure initially uses a simplified approach to protein structure prediction, MONSSTER, that assembles structures from initial extended conformations and scores them. Then, using the lowest-energy low-resolution model as a starting conformation, a detailed atomic model is built and refined using molecular dynamics simulations that employ the locally enhanced sampling (LES) methodology with the particle mesh Ewald (PME) technique for calculation of long-range electrostatic interactions. The combined method is applied to a small disulfide-rich 29-residue protein CMTI-1, a trypsin inhibitor found in squash seeds. Starting with an initial low-resolution model from MONSSTER, which has an rmsd from the native conformation of 3.7 Å (5.0 Å) for C_α atoms (all heavy atoms), LES/PME refinement leads to a structure that is only 2.5 Å (3.3 Å) from native, with a C_α rmsd of only 1.7 Å for residues 5−29. These rmsd values should be compared to C_α rmsd values of 1.2 Å (all residues) or 0.8 Å (residues 5−29) found in PME molecular dynamics simulations that start with the native conformation.

Introduction

Despite many attempts, the inability to predict from first principles the three-dimensional structure of a protein from its amino acid sequence remains a central unsolved problem in contemporary molecular biophysics.[1] The solution to the protein folding problem remains elusive due to the lack of potentials capable of recognizing the native structure from a sea of alternatives and the lack of efficient search protocols to navigate the resulting conformational space.[2,3] Having the possibility of predicting tertiary structure from sequence is not only of fundamental interest, but would also have immediate practical applications in a broad spectrum of fields. This need has become even more urgent as genome sequencing projects provide thousands of protein sequences for which structural information is nonexistent.[4]

However, it is becoming apparent that this overwhelming accumulation of sequence information can be used as a tool in evolutionary-based approaches to the structure prediction problem. Since protein evolution imposes a larger memory for structural features than for sequence patterns, it is possible to group sequences together in structurally conserved families by scoring only their sequence similarity.[5−7] Thus, it can be reliably assumed that all sequences grouped together in an alignment share the same basic fold, even if experimental structural information is absent for all elements in the alignment. It is also now becoming better established that the sequence variability in these alignments has an important nonrandom component that enforces sequence correlations, originating in part from constraints imposed by topological factors derived from the common fold.[5−7] Such sequence correlations can be exploited to predict both secondary structure and tertiary contacts between residues.[8] Penalty functions can then be derived from these target distances, allowing a directed search of conformational space and thereby facilitating the prediction of tertiary structure.[9]

The MONSSTER method[10] is a technique that implements this evolutionary-based approach to structure prediction. It uses multiple sequence alignment derived contacts to predict low-resolution folds of small proteins. Other low-resolution prediction techniques have also been developed by different authors, and the relatively large number of entries in the recent CASP2 and CASP3 protein structure prediction challenges[11] (over 30

* To whom correspondence should be addressed.
[†] University of California.
[‡] The Scripps Research Institute
[§] Current Address: Department of Chemistry, State University of New York, Stony Brook, New York 11794-3400.
[||] Current Address: Laboratory of Computational Genomics, Donald Danforth Plant Science Center, St. Louis, Missouri 63105.

(1) Dobson, C. M. *Nat. Struct. Biol.* **1995**, *2*, 513−517.
(2) Sternberg, M. J.; Thornton, J. M. *Nature* **1978**, *271*, 15−20.
(3) Wolynes, P. G.; Onuchic, J. N.; Thirumalai, D. *Science* **1995**, *268*, 960−961.
(4) Holm, L.; Sander, C. *Science* **1996**, *273*, 595−602.
(5) Sander, C.; Schneider, R. *Proteins: Struct., Funct., Genet.* **1991**, *9*, 56−68.
(6) Ortiz, A. R.; Kolinski, A.; Skolnick, J. *J. Mol. Biol.* **1998**, *277*, 419−448.
(7) Ortiz, A. R.; Skolnick, J. Manuscript in preparation.
(8) Gobel, U.; Sander, C.; Schneider, R.; Valencia, A. *Proteins: Struct., Funct., Genet.* **1994**, *18*, 309−317.
(9) Smith-Brown, M. J.; Kominos, D.; Levy, R. M. *Protein Eng.* **1993**, *6*, 605−614.
(10) Skolnick, J.; Kolinski, A.; Ortiz, A. R. *J. Mol. Biol.* **1997**, *265*, 217−241.
(11) Moult, J.; Hubbard, T.; Bryant, S. H.; Fidelis, K.; Pedersen, J. T. *Proteins: Struct., Funct., Genet.* **1997**, 2−6.

10.1021/ja993119k CCC: $19.00
Published on Web 08/19/2000

groups participated in the ab initio prediction category[12]) demonstrates the variety of approaches to this important problem. All of these approaches have in common that, even when successful, resulting low-resolution structures have root-mean-squared-deviations (rmsd) for the C_α atoms of 3−7 Å compared to the native conformations.[13−16] For many applications, this level of accuracy is inadequate.

Our current understanding of the protein folding problem clearly indicates that the coarse features of the sequences, and not the detailed interactions, are the main determinants of the fold adopted by the protein. Thus, to make the search through the vast conformational space tractable, most (if not all) computational models forego an explicit inclusion of solvent, and many use a reduced model of the protein in which each amino acid is represented by only a few particles. In addition, conformational space is typically not continuous, but a lattice model is used or only a small number of rotamers for each of the many rotatable bonds are available. Even with high-quality energy functions, these approximations may limit the inherent accuracy of the resulting predicted structures.

A reasonable approach to improving the quality of the predictions is to carry out an additional search of conformational space in the region of the low-resolution predicted structure using a more accurate model. This approach is attractive in that the detail is only added to the model when necessary; a simple model is used for the initial screening of topologies, with the most promising (based on energy) selected for refinement using a more accurate (but also more time-consuming) model. This idea is not new, and researchers in the past have attempted to use molecular dynamics (MD) simulations with atomic detail with an explicit solvent shell to refine low-resolution predicted structures.[17,18] Results were mixed, with only one of three proteins (GCN4 leucine zipper) showing improved agreement with the experimental structure. In addition, these simulations included helical backbone dihedral restraints. Other researchers used MD to successfully refine the same protein, starting from an idealized initial geometry, but also included helical and inter-helical restraints.[19,20]

This type of detailed simulation has two inherent limitations: first, the quality of the potential energy function must be high enough so that the native conformation will be the most favorable, and second, the simulation must be able to overcome the barriers to the conformational transitions on the pathway between the initial model and the native conformation. These problems are not independent, since the energy function determines the potential energy surface of the molecule, which in turn influences the characteristic time scales of the conformational transitions. It has been found that increasingly more accurate energy functions often require correspondingly greater computational resources, since the more fine grained topographical description of the potential energy surface either requires more interaction centers for its description or is concomitant with a larger roughness of this surface. Thus, in many cases a compromise had to be found between the accuracy of the energy function (and treatment of solvation effects) and the sampling of conformations. We provide clear evidence, however, that both both the quality of the force field and the extent of conformational sampling are critical to success.

In this article, we show for the first time that the two-step approach to protein structure prediction can be effective for small globular proteins by presenting a test-case application to the small protein CMTI-1, a 29-residue disulfide-rich serine protease inhibitor. We combine the MONSSTER approach, which employs a low-resolution lattice model, with state-of-the-art molecular dynamics (MD) simulations in AMBER,[21] using a force field[22] that is optimized for simulation in explicit aqueous solvation. We evaluate several MD simulation protocols for their ability to improve the similarity between the model and native[23] conformations, and find that only when the locally enhanced sampling[24] (LES) technique is combined with the particle mesh Ewald[25] (PME) treatment of long-range electrostatic interactions do the simulations provide a very significant improvement of the initial model conformation.

Methods

The first step in the combined structure prediction algorithm is the generation of the overall topology using a reduced model of the protein with the MONSSTER protocol. While this procedure has been published elsewhere,[10] the specific application to CMTI-1 will be briefly summarized. The method to predict a protein structure using MONSSTER can be divided into three stages: restraint derivation, structure assembly, and fold selection.

For restraint derivation, a multiple sequence alignment with the sequence of interest is generated. For CMTI-1, the multiple sequence alignment was obtained from the HSSP database, and consisted of 19 homologous sequences with percentage identities ranging from 97 to 64% (Table 1). Predicted secondary structure restraints are then obtained from a standard secondary structure prediction scheme, in this case the PHD method, supplemented by the prediction of loop residues or "U-turns", implemented in the program LINKER. For the CMTI-1 case, PHD predicted only a small helix for residues 5−7 (with a reliability index higher or equal 5), and therefore only this small region was restrained to the helical conformation. We note that this helix assignment is actually incorrect, as this region of the CMTI-1 molecule is not in reality helical. As a matter of fact, this misassignment turned out to be the only uncorrected error in the structure after the MD refinement (vide infra). As for the rest of the chain, predictions from LINKER were used (Table 1). After the overall secondary structure assignment, about half of the chain remained unrestricted in the folding simulations (Table 1). Tertiary restraints are then predicted from multiple sequence alignments using the package DRACULA, which implements a method to predict subsets of the contact map from a multiple sequence alignment using a combination of multivariate statistics and local threading. Details of this method are given in a separate publication. For CMTI-1, three contacts were predicted (see Table 1). Additionally, the three disulfide bridges of the protein were considered to be known and used as additional contacts. In total, only six contacts were used as restraints in the folding runs.

(12) Lesk, A. M. *Proteins: Struct., Funct., Genet.* **1997**, 151−166.
(13) Ortiz, A. R.; Kolinski, A.; Skolnick, J. *Proteins: Struct., Funct., Genet.* **1998**, *30*, 287−294.
(14) Ortiz, A. R.; Kolinski, A.; Skolnick, J. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 1020−1025.
(15) Yue, K.; Dill, K. A. *Protein Sci.* **1996**, *5*, 254−261.
(16) Srinivasan, R.; Rose, G. D. *Proteins: Struct., Funct., Genet.* **1995**, *22*, 81−99.
(17) Skolnick, J.; Kolinski, A.; Brooks, C. L.; Godzik, A.; Rey, A. *Curr. Biol.* **1993**, *3*, 414−423.
(18) Vieth, M.; Kolinski, A.; Brooks, C. L.; Skolnick, J. *J. Mol. Biol.* **1994**, *237*, 361−367.
(19) Brunger, A. T.; Nilges, M. *Q. Rev. Biophys.* **1993**, *26*, 49−125.
(20) Brunger, A. T.; Clore, G. M.; Gronenborn, A. M.; Saffrich, R.; Nilges, M. *Science* **1993**, *261*, 328−331.
(21) Case, D. A.; Pearlman, D. A.; Caldwell, J. A.; Cheatham, T. E.; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crowley, M.; Ferguson, D. M.; Radmer, R. J.; Seibel, G. L.; Singh, U. C.; Weiner, P. K.; Kollman, P. A. *AMBER 5.0*; University of California: San Francisco, 1997.
(22) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.
(23) Nilges, M.; Habazettl, J.; Brunger, A. T.; Holak, T. A. *J. Mol. Biol.* **1991**, *219*, 499−510.
(24) Elber, R.; Karplus, M. *J. Am. Chem. Soc.* **1990**, *112*, 9161−9175.
(25) Essman, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577−8593.

Table 1. Data from the Restraint Derivation, Structure Assembly, and Fold Selection Stages

A. Restraint Derivation Summary Data: Comparison with Experimental Structure Using 3cti[23]

number of residues	29
number of aligned sequences	19
ID of the sequences used in the MSA	itr1_cucma, itr4_cucma, itr3_cucpe, iti1_lagle, itr1_lufcy, itr2_lufcy, itr2_cucsa, itr1_momre, itr1_citvu, itr2_brydi, itr4_cucsa, itr3_lufcy, itr2_ecbel, itr3_cucmc, itr2_momch, iel1_momch, itr1_momch, itra_momch, itr4_lufcy
secondary structure prediction accuracy (Q3)	82.4
secondary structure assignment[10]	11552225555333555555511331111
number of predicted contacts	6
contact prediction accuracy ($\delta = 0$)[a]	83.3
contact prediction accuracy ($\delta = 1$)[a]	100.0
tertiary restraint list**:**	8−17 (0.922)
pair of residues predicted to be in	7−27 (0.809)
contact and used in the folding	14−21 (0.577)
simulations; numbers in parenthesis	3−20 (SS link)
correspond to the Pearson correlation	10−22 (SS link)
coefficient of the mutational behavior.	16−28 (SS link)

B. Energy Properties of the Predicted Folds of CMTI-1 with the MONSSTER Force Field

properties of lowest-average energy topology				properties of first excited-state topology			
$\langle E \rangle$[b]	σ[c]	$\langle rmsd \rangle$[d]	E_{min}[e]	$\langle E \rangle$[b]	σ[c]	$\langle rmsd \rangle$[d]	E_{min}[e]
−107	7.4	3.8	−125	−103	7.8	6.7	−116

C. Structural Comparison of the Predicted Structure (after MD Refinement) of CMTI-1[23]

structural database search results							comparison of predicted and native structure
closest matching structure to predicted structure[f]	comparison of closest match and predicted structure			comparison of closest match and native structure			structural alignment[m]
	Z_{scr}[g]	rmsd[h]	%[i]	Z_{scr}[j]	rmsd[k]	%[l]	
4cpa-1	0.7	1.9	86	1.3	1.5	76	5−29, 5−29

[a] Percentage of all predicted contacts within δ residues of a native contact. [b] $\langle E \rangle$ is the average energy (in kT units) of the isothermal run. [c] σ is the standard deviation of the energy in the isothermal run. [d] $\langle rmsd \rangle$ is the average coordinate root-mean-square deviation (in Å) from the native structure in the isothermal run. [e] Minimum energy (in kT units) found during the isothermal run. [f] First hit using the predicted conformation against the set of DALI representative folds of the protein database. [g] Statistical significance (*Z*-score) of the structural alignment. [h] rmsd between the structure that best matches the structure in database structure and the predicted structure. [i] % of residues aligned between the predicted structure and the first hit in the structural alignment. [j] Statistical significance (*Z*-score) of the structural alignment between the database structure that best matches the predicted structure in the database and the experimental structure. [k] rmsd between the structure chosen by DALI that best matches the predicted structure and the experimental structure. [l] % of residues aligned between DALI's first hit and the native structure. [m] Residues aligned. The first entry corresponds to the predicted structure and the second one to the native structure.

In structure assembly, the set of predicted restraints is used in the MONSSTER program to drive the conformational search in a lattice protein model. The model uses a simplified representation of the protein chain, with two interacting particles per residue, and a statistical potential derived from the protein database. Restraints are implemented with a soft potential to avoid inaccurate contact predictions that compromise correct fold assembly. A series of independent simulated annealing structure assembly runs are carried out (50 folding runs were performed for CMTI-1), and the resulting structures are clustered and fold representatives selected. The lowest-energy representatives of each topology are subjected to isothermal simulations, and the resulting average energy is calculated. The predicted structure with the lowest average energy is then selected for the final stage, *structure refinement*.

Having a predicted low-resolution model in hand, a detailed atomic model is built using the MODELLER program.[26] The initial structure is then subjected to subsequent refinement using the AMBER suite of programs[21] and the Cornell et al. force field[22] to perform MD simulations. Several different simulation protocols (discussed below) were evaluated to determine the most efficient method of refining the predicted structures.

The quality of the energy function used in simulations attempting to refine biomolecular structures is related to the amount of experimental data included; large numbers of restraints may alone be sufficient to force the protein into a nativelike conformation. Refinement of structures based on NMR data are often carried out in the absence of solvent and neglect electrostatic interactions.[27] When limited experimental data is available, results are improved when more accurate representations, including solvent, are included.[28] In the present case, no experimental restraints besides the location of the cross-links are included, and the results obtained will therefore depend entirely on the energy function to guide the structure to the native conformation. We therefore employ the most accurate energy function that we can afford computationally, and explicitly include the effects of solvent molecules.

At room temperatures, normal nanosecond-length molecular dynamics simulations have difficulty overcoming barriers to conformational transitions and only sample conformations in the neighborhood of the initial structure. Among the various techniques to enhance sampling during a simulation, LES stands out as a promising strategy. This mean-field technique allows the selective application of additional computational effort to a portion of the system, increasing the sampling of the region of interest. The enhanced sampling is achieved by replacing the region(s) of interest with multiple copies. These copies do not interact with each other and interact with other LES regions and the rest of the system in an average way. During the simulation, the copies are free to move apart and explore different regions of conformational space, thereby increasing the statistical sampling.

It has been argued on the basis of theoretical grounds[29] and demonstrated in practice using potential of mean force calculations[30] that the barriers to conformational transitions in a LES system are reduced as compared to the original system, resulting in more frequent conformational changes. Moreover, a key feature of the LES system is

(26) Sali, A.; Blundell, T. L. *J. Mol. Biol.* **1993**, *234*, 779−815.
(27) James, T. L. *Curr. Opin. Struct. Biol.* **1994**, *4*, 275−284.
(28) Chiche, L.; Gaboriaud, C.; Heitz, A.; Mornon, J. P.; Castro, B.; Kollman, P. A. *Proteins: Struct., Funct., Genet.* **1989**, *6*, 405−417.
(29) Roitberg, A.; Elber, R. *J. Chem. Phys.* **1991**, *95*, 9277−9287.
(30) Simmerling, C.; Fox, T.; Kollman, P. A. *J. Am. Chem. Soc.* **1998**, *120*, 5771−5782.

that the global energy minimum occurs when all copies occupy the position of the global energy minimum in the original system.[29] This means that optimization of the LES system directly provides information about the original system without complicated mapping procedures. Another major advantage of LES over other methods to reduce barriers or improve sampling is that it is compatible [31,32] with current state-of-the-art simulation techniques such as explicit aqueous solvation and the particle mesh Ewald technique for accurate treatment of long-range electrostatic interactions.

For LES simulations in this study, the ADDLES module of AMBER 5.0 was used to divide the entire protein into regions of 4 consecutive residues (5 in the last region), with the region boundaries placed between the C_α and C atoms,[32] for a total of 7 regions in CMTI-1. Five copies were used in each region, and particle masses were not scaled. MD simulations were carried out without LES for evaluation of the improvement obtained using LES, and both types of simulation were performed employing a cutoff on all nonbonded interactions and repeated using PME.

Several steps were required before MD simulations could be performed. Hydrogen atoms were placed using the Edit module of AMBER. Since the disulfide restraints were employed only during the initial fold prediction, and not during side chain placement, pairs of sulfur atoms corresponding to the three disulfide bonds were initially far apart (6.6−6.8 Å). Minimization and dynamics were carried out in vacuo with a dielectric constant of $4R_{ij}$ and positional restraints on C_α atoms and the harmonic terms for the S−S bonds replaced by flatwell distance restraints with no penalty for the range of 0.5 to 4.5 Å. Over 5 ps of dynamics, this restraint was converted into the appropriate AMBER harmonic bond. The protein was subsequently placed in a periodic box approximately 40 Å in each dimension, along with 1653 TIP3P[33] water molecules. A 20 ps simulation with protein atom positions fixed was carried out to equilibrate the water around the protein. After these steps, the C_α rmsd from the initial structure was only 0.2 Å, the C_α rmsd from the native conformation remained 3.7 Å, and the heavy atom rmsd was 4.7 Å. Simulations starting from the native conformation were heated to 300 K over 20 ps of MD following addition and equilibration of solvent molecules.

MD simulations were performed with the SANDER module in AMBER 5.0 after modification to support LES/PME simulations[31] and the Cornell et al. force field. The time step was 2 fs, and SHAKE[34] was applied to all bonds involving hydrogen. Simulations were carried out in the NPT ensemble at a temperature of 300 K and pressure of 1 atm. The cutoff on all nonbonded interactions for simulations without PME and vdW interactions for PME simulations was 8 Å unless noted otherwise. The nonbonded neighbor pairlist was residue-based and updated every 10 steps. PME simulations used a charge grid spacing of ~1 Å with cubic B-spline interpolation and a direct sum tolerance set to 10^{-5}. Center of mass velocity was removed each 20 ps.

Results

With CMTI-1, two structural families were generated by the MONSSTER program (Figure 1), that differed in the placement of the C-terminal fragment. The lowest-energy representatives of each topology were subjected to isothermal simulations (using a reduced temperature of $T = 1$), and the resulting average energy was calculated in the *structure selection* stage.[14,15] The nativelike topology has slightly lower average and minimum energies (see Figure 1 and Table 1). The predicted conformation of the lattice model is shown in Figure 2 superimposed on the NMR solution structure, having a C_α rmsd of 3.8 Å (Table 1).

Demonstration of the correct fold prediction is done by an automatic and unbiased method. We define a fold prediction as correct when a significant structural similarity can be found between the predicted structure and the experimental fold, with the predicted fold significantly (in the statistical sense) more similar to the target fold than to any other fold in the structure database. We carried out this automatic comparison using DALI,[35,36] matching the predicted fold (after MD refinement, vide infra) against a representative set of the protein database. Results can be found in Table 1. The correct fold is selected as

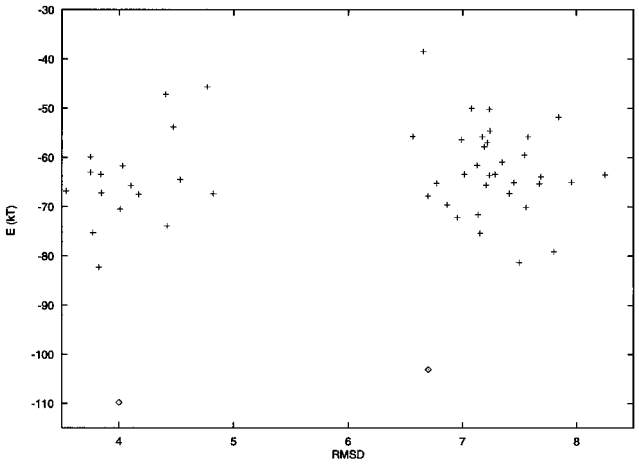

Figure 1. Energy versus rmsd plot for the assembly runs of CMTI-1. Crosses indicate the average energy of the final structures in the topology *assembly* process. The final lowest-energy structures of the two resulting topological families are then subjected to isothermal simulations at a reduced temperature of $T = 1$. The value of the average energy in these simulations is indicated by diamonds.

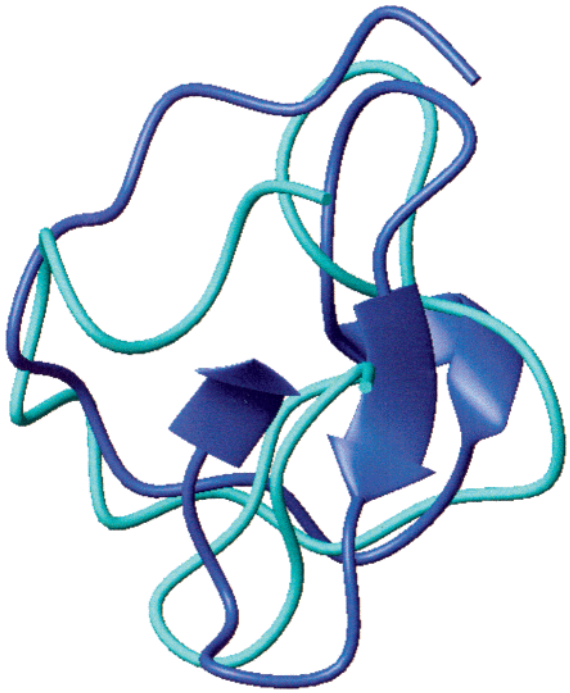

Figure 2. Predicted lattice protein model (cyan) is shown superimposed with the average structure from the experimental NMR family of structures (blue). Deviations can be seen in the binding region (left side), the 3_{10} helix (right) and, most notably, the β-sheet region (lower middle). In the predicted structure, the C-terminal strand is exposed to solvent rather than packed against the protein core. Figure produced with MOLMOL.[43]

(31) Simmerling, C.; Miller, J. L.; Kollman, P. A. *J. Am. Chem. Soc.* **1998**, *120*, 7149−7155.
(32) Simmerling, C.; Elber, R. *J. Am. Chem. Soc.* **1994**, *116*, 2534−2547.
(33) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.
(34) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327−341.
(35) Holm, L.; Sander, C. *Nucleic Acids Res.* **1997**, *25*, 231−234.
(36) Holm, L.; Sander, C. *J. Mol. Biol.* **1993**, *233*, 123−138.

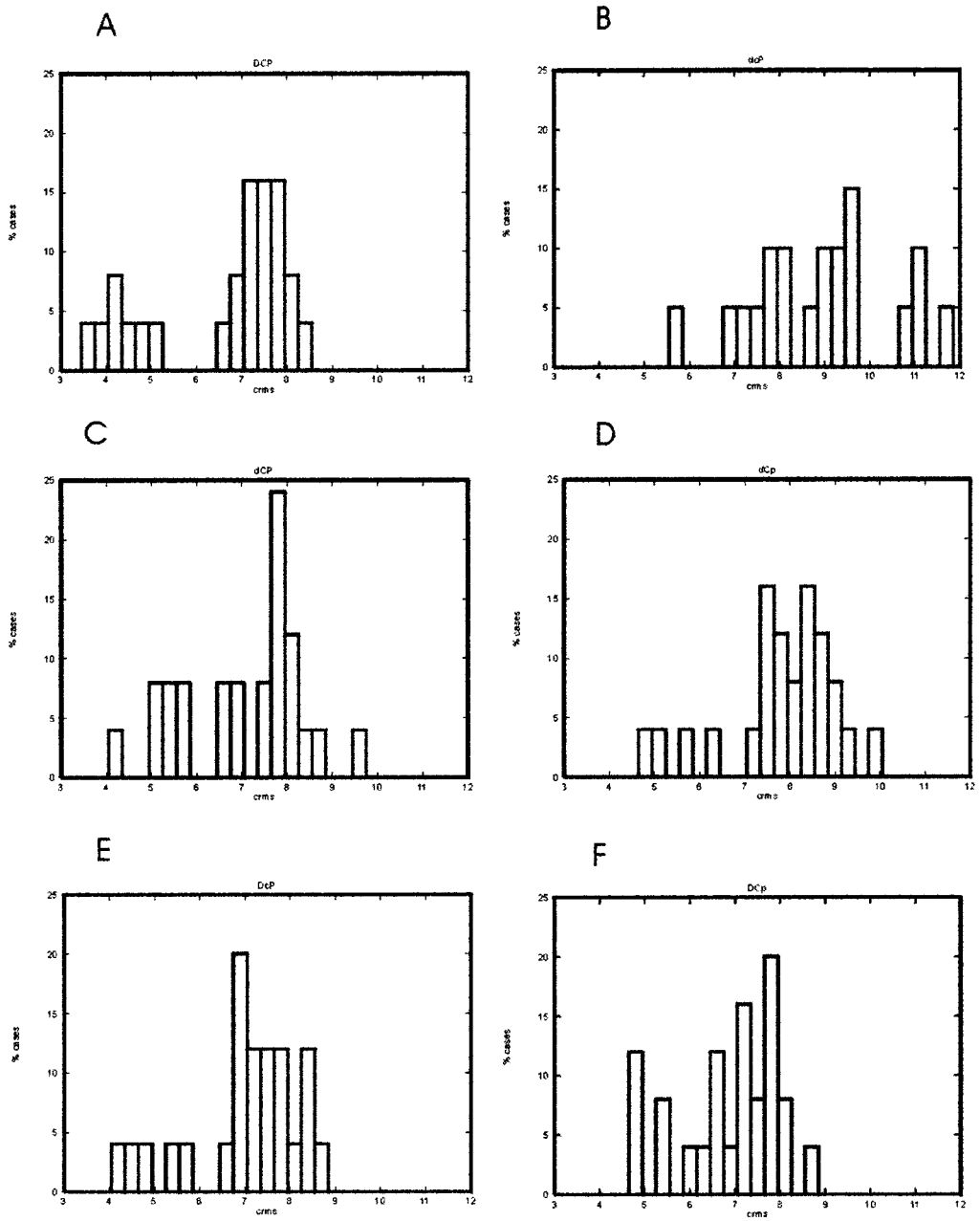

Figure 3. Effect of different protocols (see text for details) on the ability of MONSSTER to assemble the native fold of 3cti. The figure shows histograms indicating the percentage of cases having a given coordinate rmsd from native under different conditions (DR = disulfide restraints; PC = predicted contacts; PP = Pair potential): (A) DR: Yes; PC: Yes; PP: Yes; (B) DR: No; PC: No; PP: Yes; (C) DR: No; PC: Yes; PP: Yes; (D) DR: No; PC: Yes; PP: No; (E) DR: Yes; PC: No; PP: Yes; (F) DR: Yes; PC: Yes; PP: No.

the first hit, aligning 86% of the structure (residues 5−29) with a rmsd of only 1.9 Å for the C_α atoms.

We studied in detail the influence of different protocols on fold assembly. In our calculations on CMTI-1, we have included the three native disulfide bridges as restraints because this situation corresponds to a rather realistic case: On many occasions, it is possible to experimentally determine the disulfide connectivity of small cysteine-rich proteins. Having a method that could exploit this limited structural data together with additional information coming from theoretical approaches to provide near-native conformations would be of great interest. However, it is first necessary to test whether the introduction of this information, in the context of this model, is already enough to determine the fold of the protein by itself. This is an appropriate question to ask, since it would seem that a few core constraints could by themselves determine the topology of a small protein of 29 residues.

We have therefore conducted a series of studies exploring the influence of different factors on the ability to assemble the nativelike fold. In particular, we have studied the influence of the disulfide restraints, the predicted contacts and the pair potential in the likelihood of assembling a nativelike state. All eight possibilities were tested. However, and for the sake of clarity and brevity, only the six more relevant of the eight possibilities of the computational experiment are shown (Figure 3). A nativelike conformation was defined as any conformation having a value of the coordinate rmsd of less than 4.5 Å from native. Fifty simulations with each of the eight protocols were conducted. Here, the probability of obtaining a given rmsd value is shown in the form of histograms. It can be observed that the

Table 2. Best-Fit Root-Mean-Square-Deviations (rmsd), Compared to the Native Conformation, for the Final Structures Obtained from the Four MD Simulations[a]

	initial model	CUT1	PME1	CUT5	PME5
C_α rmsd (Å) 1−29	3.7	2.8	3.1	2.8	2.5
C_α rmsd 5−29	3.3	2.6	2.8	2.5	1.7
bb rmsd 4−9	1.9	1.9	1.6	2.1	1.7
bb rmsd 11−16	2.3	0.9	1.4	1.2	0.7
bb rmsd 21−29	2.8	2.1	2.1	1.6	0.9

[a] None of the simulations is successful at significantly improving the fragment from residues 4−9. LES combined with PME (PME5) is the most successful protocol.

probability of assembly with the original protocol, including pair potential, predicted contacts, and disulfide restraints is ∼25−30%. If the disulfide restraints are not incorporated, assembly is still possible, but now the frequency of successful assembly drops to around ∼5%. A similar situation is observed if only disulfide restraints are used; in this case the assembly frequency is ∼10%. However, when only the pair potential is used, no successful folding runs were observed, and a similar situation was found when both the disulfide restraints and the predicted contacts were used but not the pairwise potential.

Thus, from this computational experiment, we could conclude the following: (1) Disulfide restraints by themselves do not determine the overall fold of small proteins in our low-resolution models. (2) A small number of predicted contacts is enough to assemble small proteins when used in combination with pair potentials; however, the yield is still too small to be considered a robust predictor when the number of restraints is very small. We have previously reported that this number should be above *N*/7, where *N* is the number of residues. (3) Both predicted restraints and statistical potentials are required in order to assemble nativelike folds. (4) Robust and reproducible fold assemblies for small proteins seem to be possible when disulfide connectivity is used together with restraints derived from multiple sequence alignments and statistical potentials are used to evaluate pair interactions.

As seen in Table 2, the structure obtained after the *structural refinement/selection* stage of the MONSSTER protocol still differs from the native conformation, with an rmsd of 3.7 Å for C_α atoms and 5.0 Å for all non-hydrogen atoms. Since these rmsd values are averages over the entire protein, they may not provide insight concerning the ability of MD simulations to improve specific features of the protein topology. Three areas in the model deviate significantly from the native conformation (Figure 2, Table 2). First, residues 2−6 in the binding region (residues 2−9) are in a helical conformation, rather than the more extended conformation found in the native structure. An incorrect hydrogen bond is present between residues 3 and 6 in the model (all hydrogen bonds listed are between backbone atoms, with the convention that the first residue corresponds to the carbonyl oxygen atom and the second residue to the hydrogen). The ring of Pro4 is also on the incorrect side of the chain and should be flipped 180°, although the model does have the correct *trans* amide bond conformation. Second, residues 12−15 form a single hairpin turn rather than the 3_{10} helix at residues 12−16 found in the native conformation, with a best-fit backbone rmsd for the 11−17 segment of 2.3 Å. Third, the native β-sheet formed by residues 21−29 is irregular in that 2 residues (23−24) in one strand are paired with 3 residues (25−27) in the opposite strand. The model has a conventional β-sheet, resulting in a shift of some hydrogen bond pairs by 1 residue. The *best-fit* backbone rmsd for residues 21−29 is 2.8 Å. In addition, this sheet is also incorrectly packed against the protein core, with strand 21−24 contacting the core rather than strand 25−29 as found in the native state (Figure 2). Correction of this packing would require a ∼180° rotation of the β-sheet. We tested the ability of MD simulations to improve each of these inaccuracies, as well as to reduce the overall C_α rmsd value.

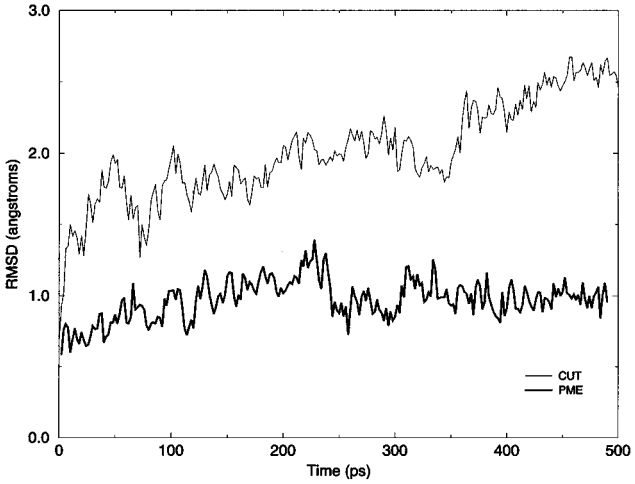

Figure 4. The time course for two simulations, starting from the NMR conformation, showing the C_α rmsd from the initial structure. The structure in the PME simulation (thick line) remains closer to the native conformation than when using an 8 Å cutoff on nonbonded interactions (thin line). The PME simulation was extended to 2 ns, and the rmsd remained ∼1 Å.

As noted above, the ability to successfully improve the model conformation using MD depends on two factors. First, nativelike conformations should be stable in the force field employed. Second, the conformational sampling technique must be able to overcome the barriers to transitions between the model and native conformations. To test the former, we carried out simulations starting from the native conformation for CMTI-1 under a variety of conditions and monitored the C_α rmsd as a function of simulation time (Figure 4). When an 8 Å cutoff is employed on nonbonded interactions, the native conformation is relatively unstable and changes to a new conformation that differs by ∼2.5 Å. Since the native conformation is not stable under these conditions even when the protein is placed there, we expect the cutoff approach to have limited value in refining structures to an accuracy greater than this level. However, when long-range electrostatic interactions are included by using the aforementioned PME method, the protein moves only ∼1 Å from the native conformation. Previous simulations have also shown that the use of PME results in more accurate protein simulations.[37]

Results of the refinement simulations are summarized in Table 3. Simulation CUT1 employed an 8 Å cutoff on all nonbonded interactions and was 4 ns in length. Simulation PME1 used PME for long-range electrostatic interactions and was 2 ns in length. Simulations CUT5 and PME5 used the same parameters as simulations CUT1 and PME1, respectively, and additionally employed LES.

In Figure 5 we show the C_α rmsd as a function of time for the four trajectories as compared to the *initial model* conformation. The protein moves ∼2.5−3.5 Å away from the model in all four simulations, demonstrating that some structural rearrangement takes place during MD. Somewhat larger and more rapid conformational changes are observed in the LES simulations. In Figure 6 we show the rmsd for the same atoms, but

(37) Cheatham, T. E.; Miller, J. L.; Fox, T.; Darden, T. A.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 4193−4194.

Table 3. C_α Root-Mean-Square-Deviation (rmsd) Values for Structure Pairs[a]

	native	native equ	CUT1	PME1	CUT5	PME5
native	0.					
native equ	0.9	0.				
CUT1	3.2	2.9	0.			
PME1	2.9	2.7	2.1	0.		
CUT5	2.8	2.7	2.9	2.8	0.	
PME5	2.4	1.8	3.0	2.7	2.7	0.

[a] The "native equ" structure corresponds to the average MD structure over 500 ps of PME MD starting with the native structure. The structures from the refinement simulations are snapshots after 2 ns MD. Note that PME5 provides a structure much more similar to the equilibrated native structure than the other refinement protocols.

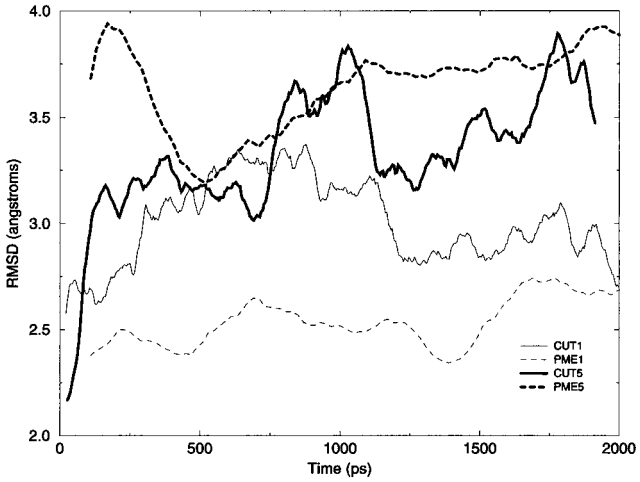

Figure 5. The time course for four simulations, starting from the model conformation, showing the C_α rmsd from the initial structure. Structures in LES simulations move farther from the initial model, with single copy PME showing the least amount of conformational change. CUT1 and PME5 were extended to 4 ns with only minor changes in rmsd values.

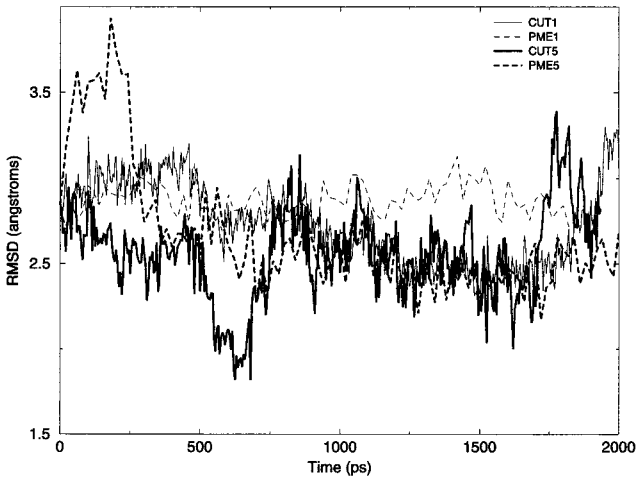

Figure 6. The time course for four simulations, starting from the model conformation, showing the C_α rmsd from the NMR structure. All four simulations result in lower rmsd values than the initial 3.7 Å. CUT1 and PME5 were extended to 4 ns with only minor changes in rmsd values.

compared to the *native* conformation. In all simulations, the final C_α rmsd was lower than that of the initial model (3.7 Å), with final values of 2.5−3.0 Å. The final structures from the two simulations using a cutoff are very similar, with a C_α rmsd of only 1.4 Å to each other, but both have C_α rmsd values of ∼3.5 Å compared to the cutoff-equilibrated native structure.

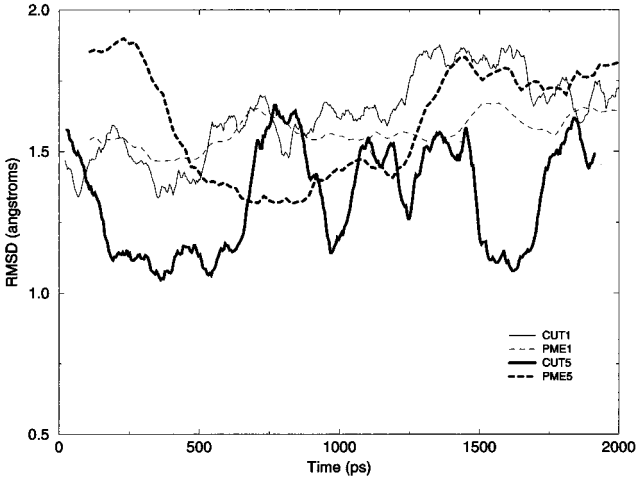

Figure 7. The time course for four simulations, starting from the model conformation, showing the backbone rmsd in residues 4−9 compared to the NMR structure. LES using a cutoff occasionally samples conformations as close as 1.0 Å. None of the simulations are successful at finding and maintaining nativelike conformations for this region. CUT1 and PME5 were extended to 4 ns; the rmsd in PME5 fell to ∼1.5 Å while that in CUT5 rose to 2.0 Å.

Final structures from LES and single-copy PME simulations differed by ∼3.0 Å from each other. In Table 3 we show pairwise C_α rmsd values for the native, PME equilibrated native, and structures after 2 ns refinement using each protocol.

As mentioned above, these rmsd values are averages over the entire protein and may not accurately reflect local deviations or improvements in the structure. We therefore examined each of the three major differences (described above) between the initial model and the native conformations, comparing segments of the protein for these regions. We begin with the incorrect helical conformation in the binding region and the position of the Pro4 ring (Figure 7).

In simulation CUT1, the segment of the backbone between Cys3 and Cys10 adopts an extended conformation, and the Pro4 ring moves to the correct position within 600 ps. However, at 2 ns the formation of a β-turn with a hydrogen bond for residues 6:9 causes this segment to deviate further from the native conformation. In simulation CUT5, no hydrogen bonds are present in this region and structures ∼1 Å from the native conformation were sampled. However, the final structure for this region is similar to that found in simulation CUT1. In simulations PME1 and PME5 an additional hydrogen bond formed between residues 2:5, stabilizing the incorrect helical conformation found in the initial model. The position of the ring for Pro4 was not corrected in any PME simulation. Despite these differences, the backbone rmsd for the binding loop residues outside this helix (residues 6−9) was improved in simulation PME5 from the initial 1.3 Å rmsd to only 0.7 Å (Figure 8).

It should be noted that two buried water molecules are present in this region of the protein in the crystal structure[38] (for CMTI-1 bound to trypsin). These water molecules form hydrogen bonds between this region and the protein core, and their presence may be necessary to reproduce the experimental structure. The lack of a structured solvent that could provide such a stabilizing network during the fold assembly process is a likely reason for the inaccuracy in the initial model of this region. During the simulation, no water molecules entered this pocket since it is

(38) Bode, W.; Greyling, H. J.; Huber, R.; Otlewski, J.; Wilusz, T. *Febs Lett.* **1989**, *242*, 285−92.

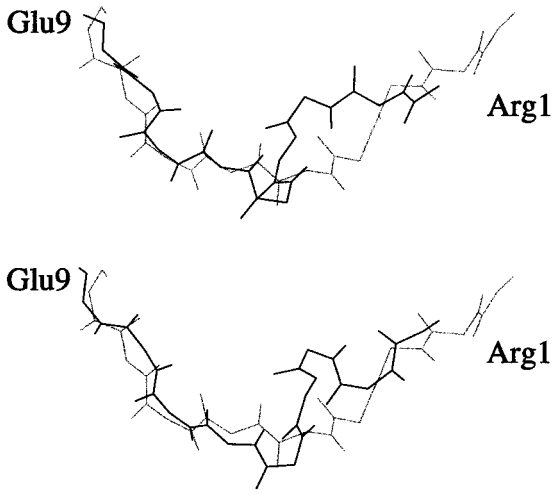

Figure 8. The initial model (upper) and LES PME refined (lower) structures for the segment from residues 1−9 are drawn with dark lines. The NMR structure is shown in gray for comparison.

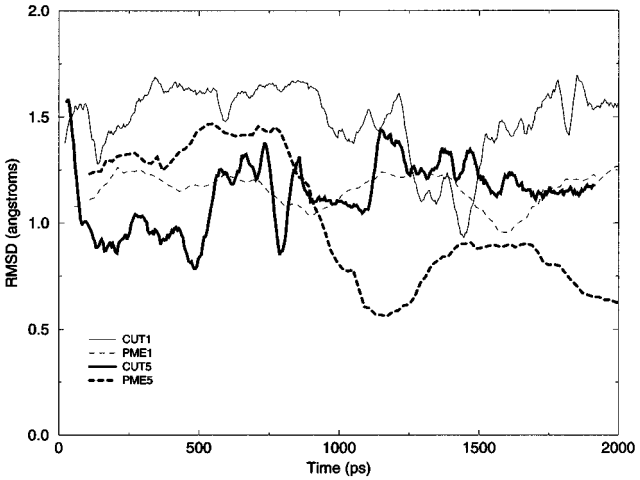

Figure 9. The time course for four simulations, starting from the model conformation, showing the backbone rmsd in residues 11−16 (3_{10} helical region) compared to the NMR structure. Only the LES PME simulation finds the correct conformation for this region. CUT1 and PME 5 were extended to 4 ns and displayed similar behavior for the last 2 ns.

not directly accessible by the bulk solvent and such diffusion is likely slower than the length of the refinement simulations. In previous studies using LES MD with simulated annealing to determine loop conformations for this protein, two water molecules did enter the pocket, and the correct conformation was located for this region. This is likely due to the higher temperatures used during annealing and the larger fluctuations in protein structure as compared to that for the present simulations (at a constant temperature of 298 K). It was also observed[39] that simulations initiated from the NMR structure showed larger deviations from the initial structure in this region if the two solvent molecules were not placed in the locations identified by those annealing simulations and crystallography of the complex. It is therefore not surprising that fold assembly without explicit solvent followed by relatively short LES refinement simulations is unable to reproduce a segment that depends on the presence of water in the protein core. This is a deficiency of the current approach and will be addressed in future refinements to the method.

All four simulations improved the segment from residues 11−16 (Table 2), where a turn of 3_{10} helix exists in the native conformation. Neither of the two hydrogen bonds (residues 12:15 and 13:16) were present in the initial model, although a turn-like conformation was observed for residues 12−15. Simulation CUT1 showed a transient population of the 12:15 hydrogen bond, with a final distance of 3.5 Å. The 13:16 hydrogen bond never formed, and the final backbone rmsd for residues 11−16 compared to the native conformation was 1.2 Å (Figure 9). Simulation PME1 behaved similarly to CUT1, with a final backbone rmsd of 1.4 Å for this segment. Simulation CUT5 again provided results similar to CUT1. The best results were clearly obtained from PME5, in which both hydrogen bonds properly formed, and a final backbone rmsd of only 0.7 Å was obtained for the 11−16 segment (Figure 10).

The largest difference between the model and the native conformations involves the β-sheet in residues 21−29 (Figure 2). In this case, both local and topological changes were present. We monitor the local structural quality using the best-fit backbone rmsd (Figure 11) and the hydrogen bond pattern in the β-sheet (Figure 12). In all four simulations, the β-sheet initially 'unfolded' with a loss of some hydrogen bonds. All simulations except PME1 eventually obtained backbone rmsd

(39) Simmerling, C.; Kollman, P. A. Manuscript in preparation.

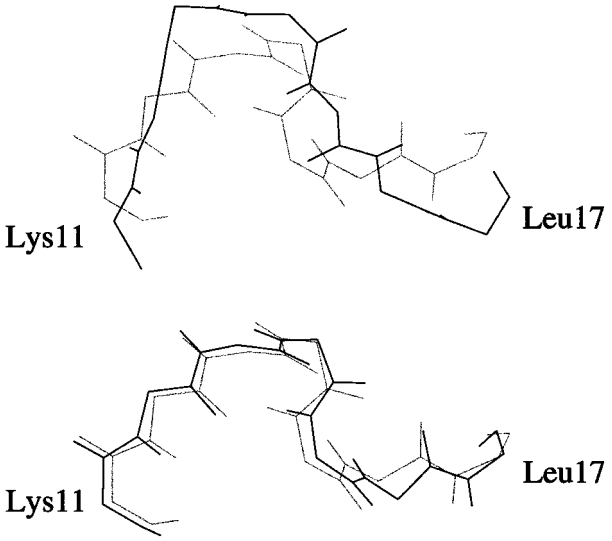

Figure 10. The initial model (upper) and LES PME refined (lower) structures for the segment from residues 11−17 are drawn with dark lines. The NMR structure is shown in gray for comparison. The simulation is successful at refining this segment and allows it to adopt the correct hydrogen bond pattern.

values under ~2 Å (Table 2). The best results were obtained from PME5, converging to a value near 0.7 Å. Only PME5 was able to form the 21:29 hydrogen bond, and none of the correct ones formed in PME1.

The largest deviation between the refined structures is in the C-terminal end of the sheet at residues 28−29, at the location of the final hydrogen bond. If we consider only residues 21−28, all simulations except PME1 result in structures with rmsd values near 0.5 Å (Figure 11). However, the amount of time that is required for this structural change varies; both LES simulations required under 1ns, with PME5 slower than CUT5, possibly related to the increased energy fluctuations resulting from the cutoff. CUT1 was over 1 order of magnitude slower than CUT5, a factor that is in agreement with previous LES PME studies.[31] If this factor is also applicable to the CMTI-1 PME simulations, we would expect that this rearrangement in PME1 would require at least 5 ns.

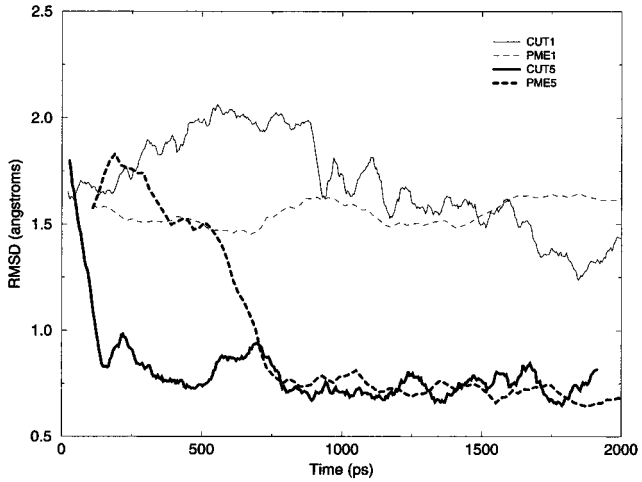

Figure 11. The time course for four simulations, starting from the model conformation, showing the backbone rmsd in residues 21−28 (β−sheet region) compared to the NMR structure. Gly29 is the location of the largest difference between the refined structures and was not included (see text). Note that these data do not evaluate the orientation of the sheet with respect to the rest of the protein. CUT1 and PME5 were extended to 4 ns; the rmsd in PME5 remained ∼0.7 Å and the rmsd in CUT1 fell to ∼0.7 Å at ∼2 ns and remained near that value for the final 2 ns.

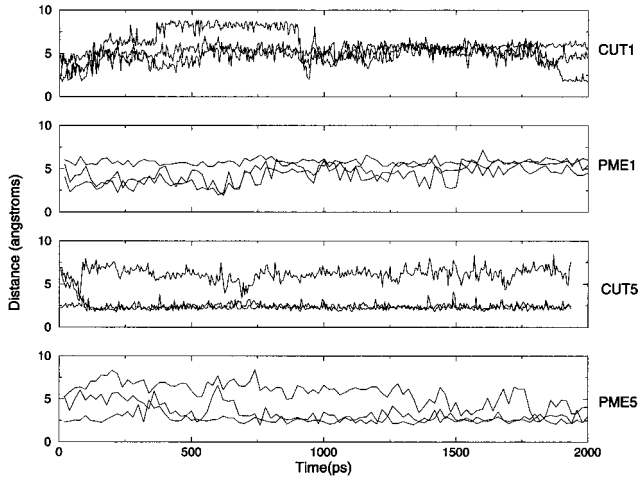

Figure 12. The distances for the atom pairs corresponding to three correct hydrogen bonds in the β-sheet (23:26, 27:23, and 21:29) are shown for each simulation. Despite the low rmsd values shown in Figure 11, only in PME5 are all of the hydrogen bonds formed. In PME1, none of the hydrogen bonds are formed. CUT1 and PME5 were extended to 4 ns; all hydrogen bonds were maintained in PME5, and two hydrogen bonds were formed in CUT5 (see text for details).

As described above, this region is incorrectly packed against the protein core. To estimate the degree to which the reorientation of the β-sheet was successful, we monitored the angle formed by the C_α atoms of residues Leu23 and Gly26 (in the β-sheet) and Glu9, which should contact the 25−29 strand. This angle has a value of 142° in the native conformation and 20° in the initial model, a deviation of 122°. The 4 MD simulations again differed in their efficacy (Figure 13). In simulations CUT1 and CUT5, the final packing angle deviation was ∼50°. In simulation PME1, the sheet angle deviation was 100°. The best results were again obtained in simulation PME5, where helix packing was essentially correct with a packing angle deviation of only 7° within 500 ps. The conformations of this segment for the initial model and PME5 are shown in Figure 14.

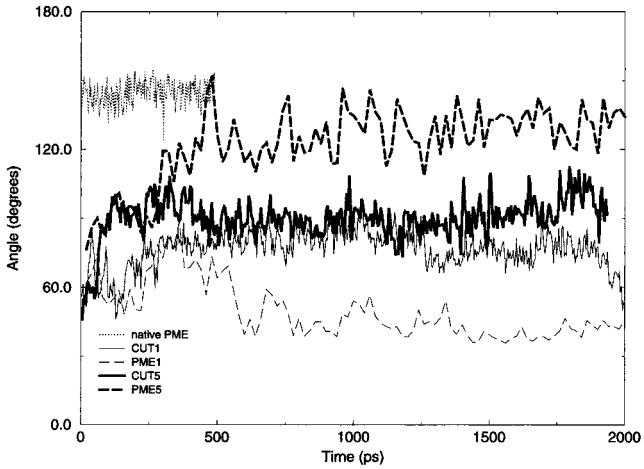

Figure 13. The packing angle of the β-sheet against the protein core for five MD simulations: one single copy PME starting from the native structure, and the four refinement simulations. Only the LES PME simulation is successful at reorientation of the β-sheet. Simulations with cutoff (both LES and single copy) are partially successful, while single copy PME shows slightly worse results than the initial model. CUT1 and PME5 were extended to 4 ns with similar values observed during the final 2 ns.

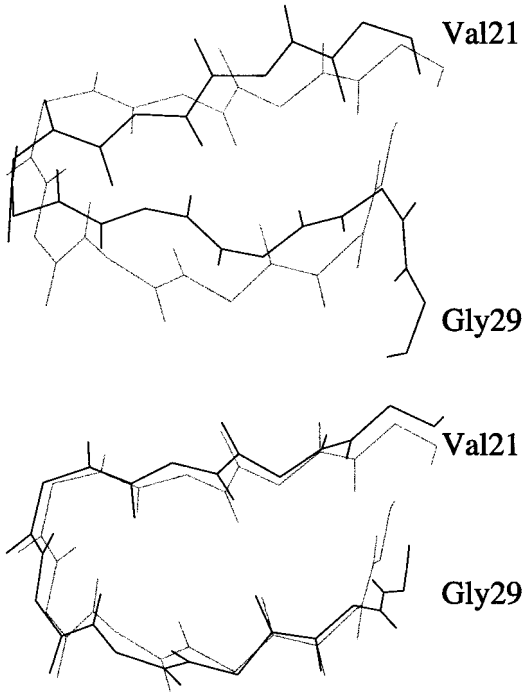

Figure 14. The initial model (upper) and LES PME refined (lower) structures for the segment from residues 21−29 are drawn with dark lines. The NMR structure is shown in gray for comparison. The simulation is successful at refining this segment and adopts the correct structure for the sheet. Note that the CO of Leu23 interacts with NH groups of both Gly26 and Tyr 27 in the correct structure.

In all cases, simulation using LES combined with PME showed improved agreement with experiment over single copy simulations, including the ability to correct significant errors in the model with the exception of the helix in residues 2−6. Overall, the C_α rmsd was reduced from an initial value of 3.7 to 2.5 Å compared to the native structure and 1.8 Å compared to the PME equilibrated native structure. Neglecting residues 1−4, the C_α rmsd value was reduced from 3.3 to 1.7 Å compared to the native structure and only 1.3 Å compared to the equilibrated native structure. The rmsd of all heavy atoms

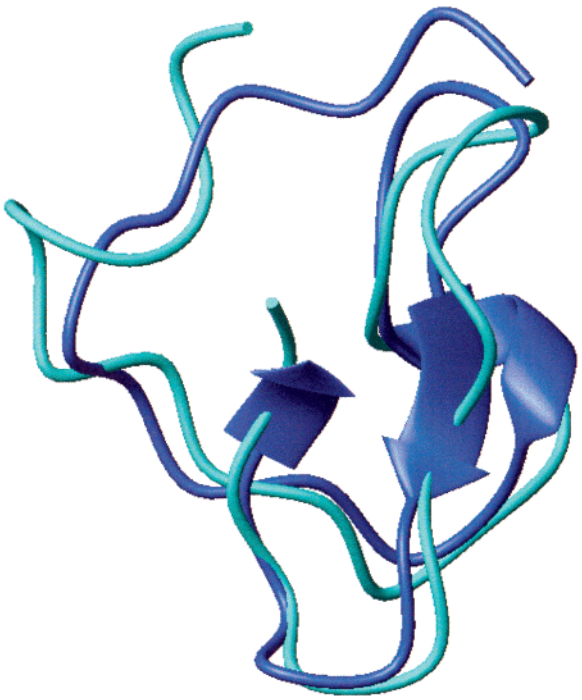

Figure 15. Comparison of the LES PME refined model structure (cyan) and the average experimental structure (blue) (see Figure 2). With the exception of the region at the C-terminus, the simulation corrected both local and topological errors in the protein structure. Most notable is the near 180° rotation of the C-terminal β-sheet to correct the packing of this secondary structure against the remainder of the protein.

that have unique positions (well-defined atoms were defined as those atoms where none of the structures has a deviation from the average position of more than 1 Å; this reduced the number of non-H atoms considered from 222 to 167, mostly neglecting the side chains of solvent-exposed charged residues) in the family of NMR structures was reduced from 4.1 to 2.8 Å for residues 1−29, and from 3.7 to 2.1 Å for residues 5−29. The corresponding rmsd values for comparison to the PME equilibrated native structure are 2.3 and 1.7 Å. The final refined conformation is shown in Figure 15.

As mentioned previously, the conformational variability of this protein is limited by the presence of the three disulfide bonds. This is true for disulfide-rich proteins in general, and the 7 kcal/mol barrier to conversion between $g+$ and $g-$ S−S rotamers can make rearrangement of segments connected by disulfide bonds difficult to observe in nanosecond length MD simulations. However, in our application of LES to CMTI-1 the Cys residues in each pair are in different LES regions. Each of the five copies of a given Cys residue therefore interacts with all five copies of the other, resulting in 25 copies of each disulfide bond and a larger reduction in the corresponding barriers compared to that for interactions inside a given region. This should permit easier adjustment of the disulfide bonds and the segments that they link. In the MONSSTER model of CMTI-1, two of the three disulfide bond rotamers were correct (Table 4). All four simulations corrected the 10:22 disulfide in the first few ps of MD, during initial structural relaxation. After this period, however, both CUT1 and PME1 showed no further disulfide conformational changes, whereas both CUT5 and PME5 showed many such transitions during periods of conformational change in the rest of the protein. In addition to the sampling, however, the energy function appears to play a role; although the model was correct for the 16:28 disulfide, CUT1 and CUT5 both rapidly moved to the incorrect rotamer while

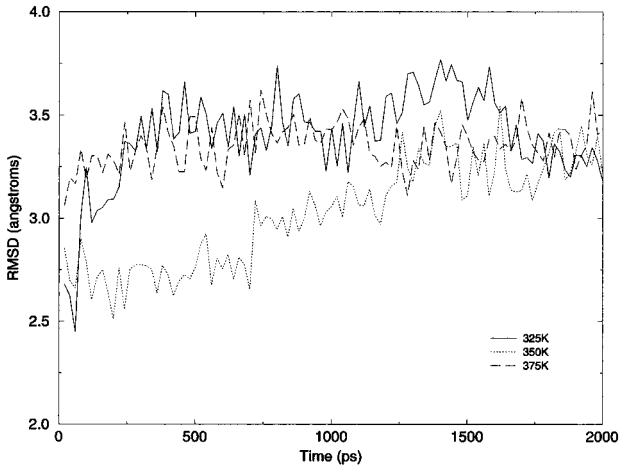

Figure 16. The C_α rmsd (compared to the NMR conformation) as a function of simulation time for three simulations that were initiated with the model conformation using single copy PME with elevated temperature to increase conformational sampling. None of the simulations is as successful as any of the simulations at 300 K.

Table 4. Disulfide Bond Rotamers for the Native, Model, and Refined Structures[a]

	3:20	10:22	16:28
native	$g+$	$g+$	$g-$
model	$g+$	$g-$	$g-$
CUT1	$g+$	$g+$	$g+$
PME1	$g+$	$g+$	$g-$
CUT5	$g+$	$g+, g-$	$g+$
PME5	$g+$	$g+$	$g-$

[a] In CUT5, some copies were in each rotamer.

the correct conformation was maintained in both PME1 and PME5. The refined PME1 and PME5 structures are correct for all three disulfides.

Several control simulations were carried out to test the sensitivity of the results to the simulation parameters. First, the LES simulation using a nonbonded cutoff of 8 Å was repeated with an 11 Å cutoff. Results similar to those for CUT5 were obtained; neither of the hydrogen bonds in the 3_{10} helix properly formed, only two of three hydrogen bonds in the β-sheet formed, and the sheet packing angle error remained above 40°. However, the structure was different from that obtained using an 8 Å cutoff, with a C_α rmsd of 3.0 Å between the two final structures. In addition, this simulation was very computationally expensive, requiring 115% greater effort than that for CUT5 and 80% greater than that for the more successful PME5.

We also tested whether explicit solvent molecules were needed by running a simulation in vacuo, using a distance-dependent dielectric constant to implicitly model solvation effects. This simulation began with the initial model and moved to a structure 3.8 Å away from the native conformation. In addition, none of the three specific areas of deviation improved, and the β-sheet and 3_{10} helix were lost. We therefore conclude that explicit solvation contributes to the success of the refinement protocol.

Next, we tested whether simply raising the temperature of the single copy PME simulations would provide results similar to that obtained using LES. In Figure 16 we show the C_α rmsd compared to the native conformation as a function of time for simulations at 325, 350, and 375 K. In all three cases, the results were worse than at 300 K, with final rmsd values of ~3.5 Å. This demonstrates that higher temperature simulations do not provide all of the benefits of LES despite their ability to cross

energetic barriers at higher rates. A LES PME simulation was carried out at 350 K to determine whether slightly increased temperature would aid in LES simulations, and to provide independent LES PME results. This 1 ns simulation was also initiated with the MONSSTER model and converged to a structure very similar to that obtained at 300 K using LES and PME; the rmsd between the two final structures for C_α (heavy) atoms was only 1.3 Å (2.4 Å). In this simulation, however, the rearrangement of the β-sheet was complete within only 100 ps.

An additional simulation was carried out to test the effect of a multistep refinement protocol in which initial refinement is carried out using LES and an 8 Å cutoff, followed by more accurate simulations using LES with PME. This protocol is consistent with our general approach in which additional detail and accuracy (and therefore computational effort) is added in a stepwise fashion. The structure after 2 ns of CUT5 simulation was used as the initial structure for a PME5 simulation. After only 200 ps with PME, the C_α rmsd (compared to the native structure) for residues 5−29 dropped from 2.8 Å (after CUT5) to 1.3 Å and the overall C_α rmsd went from 2.8 to 2.5 Å. These results are similar to those obtained from the PME5 simulation. This demonstrates that inclusion of long-range electrostatic interactions is at least partially responsible for the success of the refinement protocol.

One final variation tested the sensitivity of the results to the force field used. The four refinement simulations (single copy and LES combined with cutoff and PME) were repeated with a modified version of the AMBER force field[40] in which backbone torsion parameters were adjusted to improve the agreement to the quantum mechanical energy difference between α and β conformations for a tetrapeptide. However, none of the 2−4 ns simulations resulted in significant improvement of either the overall C_α rmsd or the three regions of largest deviation, with final values ranging from 3.2 to 3.9 Å compared to the NMR structure and 2.8 to 3.5 Å compared to the results obtained for the corresponding simulations with the original force field. Moreover, none of the correct hydrogen bonds in either the 3_{10} helix or the β-sheet were formed in any of the simulations. In this case (a protein in aqueous solution) the unmodified force field provides better results despite the improved agreement of the modified force field with quantum mechanical calculations for short peptides in vacuo.

Conclusions

Some improvement of the initial CMTI-1 model structure generated by MONSSTER was observed for all four of the simulation protocols, demonstrating that MD simulation with an accurate force field and explicit solvation can improve the predicted structures. However, the best results were provided by the combination of LES and PME. The use of a cutoff on nonbonded interactions is not desirable, given that the native conformation undergoes significant conformational changes under these conditions. However, cutoff simulations were still more effective than single copy simulations with PME; the fluctuations in energies and forces resulting from atoms crossing the cutoff boundary may assist the protein in crossing barriers to conformational transitions. The LES + PME combination provides effective barrier reduction but under conditions in which the native conformation is more stable.

Although increased temperature was not successful when combined with single copy PME simulations, higher-temperature LES PME simulations improved the protein conformation more rapidly than at 300 K. This result is encouraging and suggests that the combination of LES and simulated annealing may be a promising approach to refinement of low-resolution structures.

Many of the potential uses of predicted protein structures will require accuracy beyond the 3.5−6.5 Å provided by MONSSTER. A stepwise process of applying increasingly detailed models at each stage of refinement may be the most efficient approach to prediction of protein structures with atomic-level accuracy. We therefore investigated whether model conformations were amenable to further refinement using MD simulations employing an all-atom model with explicit solvation. We carried out such simulations for CMTI-1, a protein that is small enough so that several simulations could be carried out, but which has a well-defined NMR solution structure. We found that our best protocol (LES/PME) was able to significantly improve agreement with the native conformation. This method not only reduced the overall C_α rmsd from 3.7 to 2.5 Å (1.7 Å for residues 5−29), but corrected local mistakes such as sheet and helix hydrogen bond patterns as well as a topological error in the packing of secondary structural elements. These improvements demonstrate the quality of both the force field and sampling protocol. Although not all errors were completely corrected, these data are also beneficial and provide critical feedback, demonstrating where MONSSTER may be further improved. However, these results are for a small protein with disulfide bonds, and we intend to test the generality of the approach and investigate whether such refinement procedures can be equally successful when applied to larger proteins.

The significance of the results obtained in this work can be better evaluated by putting them in perspective and by comparing the accuracy of the CMTI-1 model with that of small proteins solved by NMR spectroscopy in the early days, one decade ago. The solution NMR structures of barley serine protease inhibitor 2 (BSPI-2), when superimposed with the corresponding X-ray structure, yielded an average backbone rmsd of 1.9 ± 0.2 Å and an all-atom rmsd of 3.0 ± 0.3 Å.[41] Similarly, model calculations with crambin based on the expected number and distribution of restraints at that time yielded rmsd values of 1.5 to 2.2 Å for the backbone atoms and from 2.0 to 2.8 Å for all atoms.[42] Here, neglecting residues 1 to 4, for CMTI-1, the structures obtained after LES/PME MD refinement have a C_α rmsd of 1.7 Å and an all heavy-atom rmsd of 2.6 Å which, in light of these figures, can be considered a very encouraging result in the structure prediction of small proteins.

Acknowledgment. M.R.L gratefully acknowledges support from NIH Training Grant No. 2-T32-GM07175-21. P.A.K. is grateful to the NIH (GM-29072) for research support. J.S. is grateful to the NIH (GM-37408 and RR-12225) for research support. A.R.O. acknowledges partial support from the Spanish Ministry of Education.

JA993119K

(40) Kollman, P.; Dixon, R.; Cornell, W.; Fox, T.; Chipot, C.; Pohorille, A. *The Development of a "Minimalist" Organic/Biochemical Molecular Mechanics Force Field Using a Combination of Ab Initio Calculations and Experimental Data*; van Gunsteren, W., Weiner, P. and Wilkinson, A., Ed.; ESCOM Press, 1997; Vol. 3.

(41) Clore, G. M.; Gronenborn, A. M.; James, M. N.; Kjaer, M.; McPhalen, C. A.; Poulsen, F. M. *Protein Eng.* **1987**, *1*, 313−8.

(42) Clore, G. M.; Breunger, A. T.; Karplus, M.; Gronenborn, A. M. *J. Mol. Biol.* **1986**, *191*, 523−51.

(43) Koradi, R.; Billeter, M.; Wuthrich, K. *J. Mol. Graphics* **1996**, *14*, 51+.