
Refinement of homology-based protein structures by molecular dynamics simulation techniques

HAO FAN AND ALAN E. MARK

Groningen Biomolecular Sciences and Biotechnology Institute (GBB), Department of Biophysical Chemistry, University of Groningen, 9747 AG Groningen, The Netherlands

(RECEIVED August 20, 2003; FINAL REVISION September 10, 2003; ACCEPTED September 10, 2003)

Abstract

The use of classical molecular dynamics simulations, performed in explicit water, for the refinement of structural models of proteins generated *ab initio* or based on homology has been investigated. The study involved a test set of 15 proteins that were previously used by Baker and coworkers to assess the efficiency of the ROSETTA method for *ab initio* protein structure prediction. For each protein, four models generated using the ROSETTA procedure were simulated for periods of between 5 and 400 nsec in explicit solvent, under identical conditions. In addition, the experimentally determined structure and the experimentally derived structure in which the side chains of all residues had been deleted and then regenerated using the WHATIF program were simulated and used as controls. A significant improvement in the deviation of the model structures from the experimentally determined structures was observed in several cases. In addition, it was found that in certain cases in which the experimental structure deviated rapidly from the initial structure in the simulations, indicating internal strain, the structures were more stable after regenerating the side-chain positions. Overall, the results indicate that molecular dynamics simulations on a tens to hundreds of nanoseconds time scale are useful for the refinement of homology or *ab initio* models of small to medium-size proteins.

Keywords: protein structure prediction; homology modeling; molecular dynamics; structure refinement

It is clear that the discrepancy between the rate at which novel protein sequences are discovered and the rate at which detailed structural information on proteins can be obtained from X-ray diffraction (X-ray) or nuclear magnetic resonance spectroscopy (NMR) will continue for the foreseeable future. For this reason, there is a pressing need for theoretical methods to predict protein structure from sequence (Abagyan and Batalov 1997; Yang and Honig 1999; Standley et al. 2001; Williams et al. 2001). At present, it is not possible to reliably predict protein structure from sequence taking a truly *ab initio* approach. Instead, proteins are generally modeled based on the fact that sequence homology

normally implies structural similarity (Tramontano et al. 2001), and several large-scale efforts to model all proteins of particular genomes have been initiated (Fischer and Eisenberg 1997; Sanchez and Sali 1998; Guex et al. 1999). The question is: How useful are such exercises? Homology models are imprecise by definition. They contain little additional information, compared with the template structure, unless they can be further refined. In cases of low sequence homology, errors in the secondary structure definition and packing of secondary structure elements are common. Even the general fold may not be correct. In cases of high sequence homology, the basic framework of the protein can normally be predicted with high accuracy. Nevertheless, errors still occur in variable loops, the relative orientations of secondary structure elements, and in the details of atomic packing. Even small errors in critical regions, however, are sufficient to prevent the use of models in sensitive applications such as in rational drug design and the prediction of protein-protein interactions.

Reprint requests to: Alan E. Mark, Groningen Biomolecular Sciences and Biotechnology Institute (GBB), Department of Biophysical Chemistry, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands; e-mail: mark@chem.rug.nl; fax: 31-50-3634800.

Article and publication are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.03381404>.

Present attempts to refine homology models and thus correct the errors inherent when using a template approach are normally based either on energy minimization, limited conformational sampling using molecular dynamics in conjunction with a detailed force field, or more extensive sampling using simplified force fields. Generally these approaches have proved ineffective (Schonbrun et al. 2002). This has been one of the findings of the Critical Assessment of Techniques for Protein Structure Prediction (CASP) competitions (Venclovas et al. 2001), and the inability to refine protein models to atomic resolution now stands as a major obstacle to the wider use of models generated ab initio or based on homology in structural genomics.

That refinement schemes based on simplified representations and/or limited sampling fail is not surprising. Proteins are densely packed, which makes the searching of conformational space difficult. In addition, the native conformation is frequently only marginally stable. There is a fine balance of competing interactions between the solvent and the protein as well as between alternate packing arrangements of side chains that cannot easily be captured in simplified representations. To be able to refine protein models to atomic resolution, we must ultimately turn to physically reasonable representations, extensive sampling, and/or advanced search techniques.

In the present work, we investigate the use of molecular dynamics simulations performed using atomic-based empirical force fields in explicit solvent for the refinement of protein structures generated ab initio or based on homology. This study is based on a set of 15 proteins used previously by Baker and coworkers to verify the ROSETTA structure

prediction algorithm (Orengo et al. 1999; Simons et al. 1999, 2001). For each of the 15 proteins, four models, generated using ROSETTA, and two controls, based on the experimentally derived structure (90 structures in total), were simulated under identical conditions. Key structural and dynamic properties of the systems were then analyzed. By keeping all other parameters constant, the aim of the study was to assess objectively the efficiency of classical MD simulation techniques in the refinement of such structural models.

Results and Discussion

Model refinement

As an indication of the degree of refinement during the simulations, the positional root mean square deviation (RMSD) from the respective experimental structure (NMR or X-ray) after a least-squares best fit was calculated for each structure investigated (four models and two controls for each of 15 proteins) as a function of the simulation time for all backbone atoms. The RMSD values calculated after 5 nsec of simulation for all structures are given in Table 1. Also indicated in Table 1 are some structural properties of the proteins including the number of residues involved in particular secondary-structure elements. The values in parentheses indicate the RMSDs of the original ROSETTA models from the experimental structure. Taking a 10% change in RMSD (compared with the RMSD of the original ROSETTA model) as an indication of significance, we find that after 5 nsec of simulation, 11 of the 60 models show

Table 1. Structural properties of the proteins and backbone RMSD after 5 nsec of MD simulation at 300 K

No.	PDBID	N_{res}	Exp. N_{α}	Exp. N_{β}	RMSD exp.	RMSD s-chains	RMSD model 1	RMSD model 2	RMSD model 3	RMSD model 4
1	1aoy	78	32	8	0.45	0.35	0.72 (0.59)	0.92 (0.84)	0.99 (0.95)	1.39 (1.19)
2	1stu	68	26	20	0.29	0.30	1.03 (0.62)	0.95 (0.79)	0.99 (0.80)	0.79 (0.86)
3	1vif	60	0	26	0.29	0.29	0.81 (0.62)	0.95 (0.97)	0.96 (0.98)	1.16 (1.08)
4	1sro	76	4	27	0.30	0.29	0.49 (0.45)	0.67 (0.87)	1.02 (0.96)	1.03 (0.97)
5	1tuc	61	3	25	0.24	0.18	0.71 (0.59)	0.75 (0.82)	0.85 (0.90)	0.88 (1.00)
6	1sap	66	20	27	0.36	0.31	0.50 (0.35)	0.87 (0.72)	1.04 (0.94)	1.09 (1.16)
7	1afi	72	21	21	0.17	0.18	0.21 (0.26)	0.76 (0.87)	0.91 (0.93)	1.03 (1.02)
8	1vcc	77	11	25	0.26	0.16	0.68 (0.59)	0.83 (0.84)	1.11 (1.06)	1.17 (1.17)
9	2bby	69	35	4	0.26	0.31	0.69 (0.53)	0.88 (0.86)	1.02 (1.04)	0.84 (1.05)
10	2fmr	65	18	18	0.52	0.35	0.57 (0.44)	0.80 (0.90)	0.94 (0.96)	0.96 (1.02)
11	1alz	83	62	0	0.27	0.36	0.43 (0.55)	0.83 (0.96)	1.20 (1.21)	1.21 (1.28)
12	1ail	70	60	0	0.27	0.35	0.70 (0.54)	0.92 (0.84)	0.97 (1.10)	1.22 (1.24)
13	1coo	81	35	0	0.23	0.67	0.82 (0.58)	1.22 (1.16)	1.22 (1.21)	1.20 (1.34)
14	1lea	72	39	6	0.23	0.26	0.51 (0.40)	1.12 (0.97)	0.98 (0.99)	1.00 (1.06)
15	2ezh	65	45	0	0.30	0.29	0.43 (0.34)	0.66 (0.74)	0.89 (0.88)	0.95 (0.93)

(PDBID) PDB identifier; (N_{res}) number of residues; (N_{α}) the number of helical residues; (N_{β}) the number of strand residues. The backbone root mean square deviation (RMSD) in nanometers with respect to the experimental structure after 5 nsec of simulation starting from different initial configurations. Starting configurations: (exp.) the experimental structure; (s-chains) the experimental structure after side chains were reconstructed using WHATIF; (model 1–4) models generated using ROSETTA (side chains constructed using WHATIF). The values in parentheses correspond in the RMSD of the starting model with respect to the experimental structure. Values in bold indicate a >10% decrease in RMSD with respect to the corresponding starting model. Values in italics indicate a >10% increase in RMSD with respect to the corresponding starting model.

lower RMSD values with respect to the experimental structure than the original model (indicated in boldface in Table 1). For 18 models, the RMSD had increased after 5 nsec of simulation, and for 31 models, there was no significant change.

As alternative measures of whether simulation in explicit solvent improved the modeled structures, the number of native hydrogen bonds (HB) and native contacts (NC) was also considered. The HB and NC were determined based on geometric criteria. An HB was considered to exist if the donor–hydrogen–acceptor angle was $<60^\circ$ and the distance between the hydrogen and the acceptor was <0.25 nm. A contact between two residues was considered significant if the distance between the two corresponding C_α atoms was <0.6 nm and the residues were separated by at least two positions in the amino acid sequence; that is, they were not first or second neighbors.

The number of native HBs for each model after 5 nsec of simulation is presented in Table 2. Two different criteria for determining which HBs should be considered native in the context of the simulations were considered. First, all HBs that existed in the experimental structure were assumed to be native. These were compared with the HBs in the final configuration after 5 nsec of simulation for each of the models. Again taking a change of 10% (of the total number of native HBs) as an indication of significance, we find that seven of the 60 models improve, seven get worse, and 46 remain largely unchanged. It could be questioned, however, if counting the instantaneous number of HBs is truly appropriate. In the simulations, the structures undergo rapid local

fluctuations, and many HBs are only transient. Also, not all HBs found in the experimental structure are structurally significant. For these reasons, an alternative criterion for identifying native HBs was also considered. The HBs were averaged over the last 1 nsec of the 5-nsec simulation started from the experimental structure. An HB was considered significant only if it occurred with a frequency of >0.9 for this period. An analogous procedure was used to identify HBs in the simulations of the model structures. Again taking a change of 10% as an indication of significance, we find 14 models show an increase in native structure, whereas only six show a significant loss of structure. The reason for the large difference when using the two criteria is twofold. First, averaging over the simulation will help identify the HBs that are structurally most significant. Second, in attempting to determine if MD simulations are useful in refinement, it is important to separate questions of sampling from questions related to the force field. For example, looking at the sixth column in Table 1, it is clear that in some cases, the experimental structure itself deviates significantly during the simulations. By considering only those HBs that are stable in the simulation started from the experimental structure as significant, we are able to better judge whether the model structures are being refined. In this respect, we note that in general the most significant improvements in native HBs were observed for the first model (model 1) of each protein. The numbering of the models was chosen such that model 1 (initially) had the lowest RMSD with respect to the experimental structure among the four models.

Table 2. Native hydrogen bonds (HB) after 5 nsec MD at 300 K

No.	PDBID	HB exp.	HB model 1	HB model 2	HB model 3	HB model 4	HB exp*	HB model 1*	HB model 2*	HB model 3*	HB model 4*
1	1aoy	43	20 (20)	23 (22)	11 (10)	<i>12</i> (20)	26	18 (18)	20 (18)	11 (10)	<i>13</i> (18)
2	1stu	32	9 (14)	12 (13)	16 (14)	15 (10)	35	11 (11)	10 (12)	15 (12)	13 (9)
3	1vif	44	0 (0)	0 (2)	0 (0)	0 (0)	19	0 (0)	0 (0)	0 (0)	0 (0)
4	1sro	28	8 (9)	6 (10)	3 (2)	1 (3)	27	8 (3)	8 (11)	1 (1)	2 (3)
5	1tuc	36	10 (3)	4 (5)	7 (5)	6 (5)	21	8 (2)	3 (4)	7 (4)	5 (4)
6	1sap	47	25 (17)	15 (9)	8 (6)	15 (8)	23	20 (13)	8 (8)	5 (4)	9 (7)
7	1afi	44	27 (17)	11 (11)	15 (19)	16 (14)	27	20 (9)	9 (7)	<i>10</i> (13)	11 (10)
8	1vcc	55	9 (15)	21 (16)	6 (10)	7 (12)	28	5 (8)	13 (9)	5 (2)	4 (6)
9	2bby	42	23 (20)	15 (15)	19 (18)	18 (19)	28	17 (14)	13 (11)	16 (16)	16 (16)
10	2fmr	32	12 (10)	9 (9)	9 (11)	14 (11)	17	7 (7)	7 (7)	7 (8)	8 (9)
11	1alz	58	30 (22)	28 (23)	22 (19)	25 (25)	31	20 (13)	21 (16)	12 (13)	19 (16)
12	1ail	67	24 (25)	23 (29)	20 (14)	23 (17)	27	16 (13)	<i>14</i> (17)	12 (8)	16 (12)
13	1coo	37	8 (16)	<i>14</i> (18)	14 (13)	18 (17)	24	<i>10</i> (14)	15 (17)	10 (12)	15 (13)
14	1lea	55	28 (29)	16 (19)	23 (26)	<i>16</i> (24)	21	17 (17)	8 (10)	15 (14)	11 (12)
15	2ezh	47	25 (27)	25 (25)	22 (26)	18 (20)	22	15 (15)	16 (14)	16 (17)	11 (12)

(HB) The number of native hydrogen bonds in a certain conformation; (exp) the number of hydrogen bonds in the experimental structure; (exp*) the number of hydrogen bonds that occur with a frequency >0.9 during the last 1 nsec of the MD simulations started from experimental structure; (models 1–4) the number of native (exp) hydrogen bonds after 5 nsec of simulation starting from the ROSETTA models; (models 1–4*) the number of native (exp*) hydrogen bonds after 5 nsec of simulation starting from ROSETTA models. The values in parentheses correspond to the corresponding number of native hydrogen bonds in the starting model. In bold are indicated cases in which the number of native hydrogen bonds in the model increases by $>10\%$ (of the total number of native HBs). In italics are indicated cases in which the number of native hydrogen bonds in the model decrease by $>10\%$ (of the total number of native HBs).

In Table 3, the results of the analysis of native contacts are presented. Analogous to the HB analysis, NCs were determined using two alternative criteria, an instantaneous NC determined from the experimental structure and an averaged NC in which only those contacts that were present for >90% of the time during the last 1 nsec of the simulation were considered significant. Again using a change of 10% as a measure of significance, we find that using an instantaneous criterion, only two of the 60 models show improvement, whereas seven get worse; considering the averaged NC, seven improve, whereas 11 are worse. Overall, the number of native contacts was found to be the least appropriate measure by which to judge the degree of refinement as native contacts could be still satisfied even if there were errors in local packing and the number of native contacts was, in comparison to the results of the backbone RMSD and native HB analysis, less sensitive to changes in global structure and/or secondary structure motifs, especially when the starting and final conformations were both well-packed structures.

The effect of side-chain reconstruction on model refinement

The models provided by Baker and coworkers at the time this study was conducted contained only the protein main chain and C_β atoms. The side chains for all amino acids other than glycine and alanine were constructed automatically using the CORALL module of WHATIF (Vriend 1990). The construction of the side chains could be expected

to introduce stress within the proteins and is thus potentially a major source of error in the calculations. To test the effect of side-chain generation on the stability of the protein structures and to provide an additional control for the refinement calculations, the side chains of each of the 15 experimental structures were first removed and then reconstructed using CORALL. Each of these modified native structures was then simulated for 5 nsec. The backbone positional RMSD after 5 nsec from the experimental structures is given in column 7 of Table 1. Again taking a change of 10% as a crude indication of significance, we find after deleting and regenerating the side chains, five of the 15 structures deviate less than the experimental structure, four deviate more, and six show similar behavior to the experimental structure. Although this is a small sample and no particular significance should be attributed to whether the RMSD in a particular case fell slightly above or below the cutoff, the results do demonstrate that in most cases generation of the side-chain conformations by WHATIF does not introduce strain into the structures. In fact, deleting and regenerating side chains may even remove the strain inherent in some NMR derived structures (Fan and Mark 2003).

Effect of increasing the length of the simulation

To investigate how the length of the simulations affects the degree of refinement, the trajectories of selected models for two proteins, lafi and lsro, were extended by up to almost 2 orders of magnitude. Specifically, model 1 and model 2 of lafi were simulated for 100 nsec and 400 nsec, respectively,

Table 3. Native contacts (NC) after 5 nsec of MD at 300 K

No.	PDBID	NC exp	NC model 1	NC model 2	NC model 3	NC model 4	NC exp*	NC model 1*	NC model 2*	NC model 3*	NC model 4*
1	laoy	71	37 (37)	35 (38)	22 (29)	30 (34)	53	32 (32)	34 (33)	20 (27)	26 (32)
2	lstu	106	34 (35)	33 (38)	39 (41)	41 (30)	68	25 (31)	28 (27)	33 (31)	34 (21)
3	lvif	71	13 (12)	15 (16)	10 (7)	10 (12)	47	9 (7)	11 (7)	7 (3)	6 (5)
4	lsro	80	31 (30)	27 (36)	19 (22)	10 (17)	76	34 (32)	23 (26)	14 (13)	8 (14)
5	ltuc	68	18 (20)	12 (13)	22 (22)	14 (18)	51	13 (14)	8 (11)	14 (14)	9 (13)
6	lsap	71	50 (45)	34 (31)	27 (28)	34 (35)	57	48 (38)	33 (25)	23 (22)	31 (30)
7	lafi	111	72 (59)	28 (26)	32 (42)	38 (38)	60	50 (41)	20 (23)	30 (32)	29 (28)
8	lvcc	85	29 (32)	34 (35)	30 (25)	20 (20)	58	18 (20)	23 (24)	19 (15)	12 (12)
9	2bby	64	34 (39)	27 (33)	35 (38)	30 (32)	49	27 (31)	22 (27)	28 (33)	26 (27)
10	2fmr	66	25 (27)	18 (19)	25 (29)	28 (33)	36	15 (15)	11 (11)	18 (22)	17 (21)
11	1alz	86	56 (58)	50 (55)	40 (53)	52 (58)	58	44 (46)	40 (42)	32 (43)	42 (47)
12	1ail	75	42 (40)	37 (43)	32 (33)	34 (37)	56	38 (35)	33 (38)	25 (26)	29 (33)
13	1coo	84	24 (37)	40 (39)	26 (39)	43 (45)	54	22 (31)	30 (29)	21 (29)	31 (33)
14	1lea	68	36 (41)	30 (35)	38 (40)	33 (39)	54	36 (39)	26 (32)	35 (37)	30 (36)
15	2ezh	57	39 (45)	40 (44)	39 (47)	32 (39)	49	36 (35)	34 (34)	36 (37)	28 (31)

(NC) The number of native contacts in a certain conformation; (exp) the number of contacts in the experimental structure; (exp*) the number of native contacts that were found during the last 1 nsec of the MD simulations started from experimental structures at 300 K with appearance frequency >0.9; (models 1–4) the number of native (exp) contacts after 5 nsec of simulation starting from the corresponding ROSETTA models; (models 1–4*) the number of native (exp*) contacts after 5 nsec of simulation starting from the corresponding ROSETTA models. The values in parentheses correspond to the number of native contacts in the starting model. In bold are indicated cases in which the number of native contacts in the model increases by >10% (of the total number of native NCs). In italics are indicated cases in which the number of native contacts in the model decrease by >10% (of the total number of native NCs).

and model 2 of 1sro was simulated for 400 nsec. Figure 1 shows the RMSD as a function of simulation time for model 1 of 1afi. The initial backbone RMSD from the experimental structure was 0.26 nm, and this structure was by far the best prediction made by ROSETTA for any protein in this test set. As can be seen in Figure 1, in the simulation there was an initial rise in RMSD during the first several nano-seconds as strain in the structure is released. Then there is a sharp drop in RMSD to ~ 0.2 nm within 5 nsec. A further systematic decrease in RMSD is observed over the following 50 nsec. After 100 nsec, the RMSD is ~ 0.17 nm, a substantial improvement over the original model. Several points should be noted. First, the initial rise in RMSD is observed in almost all simulations starting from a modeled structure. Small errors in packing normally lead to large interatomic forces as the force field is applied. This, in turn, leads to an initial distortion of the structure, which has led many people to conclude, based on inappropriately short simulation time, that MD simulations are too inaccurate to be used to refine protein structural models unless experimental restraints are also applied. Second, the process of repacking is relatively slow. The minimum difference in RMSD (0.12 nm) occurred after ~ 52 nsec. The RMSD then rises slightly and continues to fluctuate during the rest of the simulation. In the simulations, thermal fluctuations lead to the generation of an ensemble of structures. For this reason, there will always be residual differences between the averaged and/or minimized experimental structure and any given configuration from the trajectory. Thus, although in this work we present the results in terms of the configuration after a particular time, in principle an averaged structure or, more correctly, the most probable structure should in fact be compared with that obtained experimentally.

The fact that the structure has undergone substantial refinement during the simulations is illustrated in Figure 2, which shows the initial ROSETTA model, the final struc-

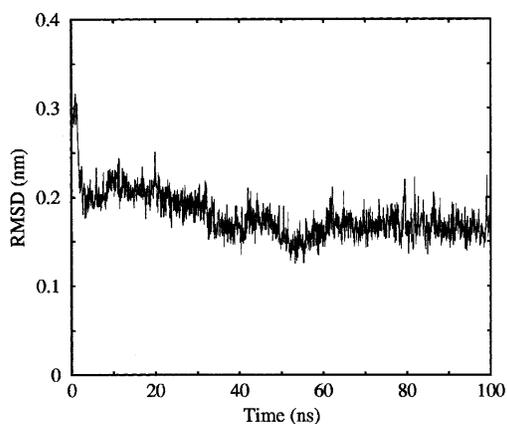


Figure 1. The time evolution of the backbone positional root mean square deviation (RMSD) of the ROSETTA model 1 for the protein 1afi after a least-squares best fit on all backbone atoms of experimental structure.

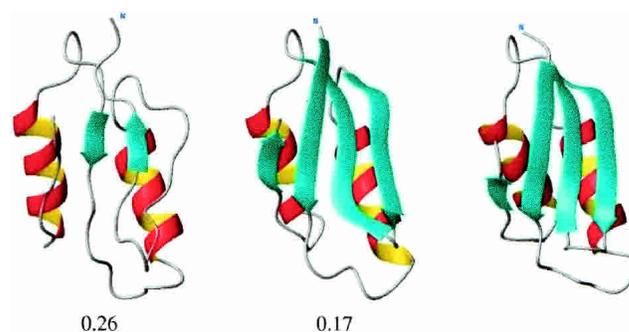


Figure 2. (From left to right) ROSETTA model 1 of the protein 1afi, the final structure after 100 nsec of molecular dynamics (MD) simulation, and the structure of 1afi as determined experimentally using NMR. Below the first and second structures are given the backbone RMSDs (in nanometers) referring to the NMR structure.

ture after 100 nsec, and the experimental structure determined by NMR. As can be seen in Figure 2, the simulations have led to a tighter packing of the helices, the regularization of the elements of the β -sheets, and the improvement in the packing of the structural elements. The final refined structure would be within experimental uncertainty (i.e., the precision to which the structure can be known).

Model 1 of 1afi already lies very close to the experimental structure. To investigate if it is possible to refine more distant models, model 2 of 1afi and model 2 of 1sro were simulated for 400 nsec. The initial deviation from the respective experimental structures was 0.87 nm in both cases. Both examples had shown some improvement in RMSD within 5 nsec. In Figure 3, the RMSD as a function of simulation time is plotted for both cases. For model 2 of 1afi (black curve), the RMSD fluctuates widely over the first 200 nsec, falling to close to 0.6 nm but rising again to above 0.8 nm. After 200 nsec, the system remains relatively stable at an RMSD of 0.68 nm. In the case of model 2 of 1sro (gray curve), there are large fluctuations over the entire 400 nsec. By coincidence, the final RMSD value is also close to 0.7 nm. The striking feature of these comparatively long simulations is the extensive spontaneous rearrangements these proteins undergo. Substantial regions of the proteins unfold then reassemble, as is indicated by the large changes in RMSD.

In Figure 4 is shown the initial ROSETTA model, the final structure after 400 nsec, and the experimental structure for model 2 of 1afi. Although the overall RMSD is still high and the final model is very far from a useful structure, it can be seen there is a marked improvement in the structure, in particular, in the packing of the helices. The two strands that were correctly paired in the model are maintained and move into the correct position relative to the helices. The incorrectly paired strands are lost but do not refold within the 400 nsec. The system instead becomes trapped in a local well toward the end of the simulation, and much longer simula-

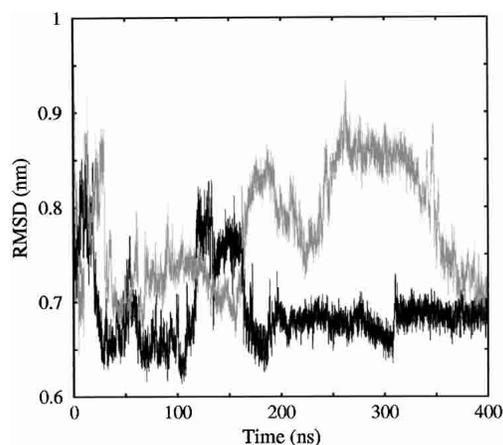


Figure 3. The time evolution of the backbone RMSD of the ROSETTA model 1 for 1afi (black line) and the ROSETTA model 2 for 1sro (gray line) with respect to the corresponding experimental structures.

tion times would be required to obtain further refinement (assuming the force field is appropriate).

In Figure 5 is shown the initial ROSETTA model, the final structure after 400 ns and the experimental structure for model 2 of 1sro. In this case it is interesting to note that the ROSETTA model (Fig. 5, left) is compact, and in fact might be viewed as more structured than the experimental structure (Fig. 5, right). However, the ROSETTA model contains major errors in the arrangements and pairing of the elements of the β -sheet. This can be seen most readily in terms of the location of the N terminus. After 400 nsec of simulation, the final structure is still compact. There has been substantial rearrangement of the sheet regions, relocation of the N-terminal region, and improvement in the overall fold. Nevertheless, the final structure is still far from that obtained experimentally.

Simulation at elevated temperature

It is clear that although substantial refinement of modeled structures is possible given an ability to simulate on a 100-nsec time scale, sampling and the avoidance of trapped states are major issues. One means to help avoid being trapped in local minima is to increase the simulation temperature. To test the influence of the simulation temperature on the process of the refinement, selected models were simulated for 10 nsec at 300 K and 325 K for comparison. The models selected included the model that initially had the lowest deviation (model 1) for 14 of the 15 test proteins. The protein 2fmr was not included in this test as the experimental structure was not stable when simulated at 300 K. In addition to these 14 models, four models that showed large improvements in RMSD after 5 nsec of simulation at 300 K were also included. The backbone RMSD values with respect to the experimental structures are listed in Table 4.

Values are listed for each model simulated at 300 K and at 325 K after 0.1, 1.0, and 10 nsec. For the five models (Table 4, boldface) for which the initial deviation from the experimental structure was the lowest (RMSD ≤ 0.50 nm), the RMSD after 10 nsec of simulation at 325 K was in general lower than the RMSD after 10 nsec of simulation at 300 K. The only exception was the first model of the protein 1afi for which the last conformation from the 325 K simulation was close to that of native structure (RMSD = 0.27 nm). In contrast, for the four models (Table 4, italics) that initially showed the largest deviations (initial RMSD ≥ 0.8 nm), a lower RMSD was generally obtained at 300 K than at 325 K. In the case of model 2 of 1afi, similar performance was observed at both temperatures. Although one must be careful not to attribute too much significance to this result considering the small sample size and the large fluctuations in the data, it does indicate that models closer to their global minimum, within what might be considered to be a potential energy well or the lower regions of a “folding funnel” (Dill and Chan 1997), may respond better to simulation at slightly elevated temperatures. The higher temperature in these cases helped to enhance sampling but did not lead to unfolding. Limited tests with temperatures higher than 325 K (data not shown) did not show further improvement. For models that initially showed very large deviations from the experimental structure, increasing the temperature simply led to unfolding. On this basis alone, it might be possible to discriminate between models that are well folded (close to their native conformation) from models with an incorrect global fold using simulation techniques.

From Table 4 it is also possible to compare the RMSD at three different simulation times for each model. As alluded to earlier, the striking feature is that in the vast majority of cases, there is an initial rise in RMSD after 0.1 nsec, and in many cases, there is a further rise from 0.1 to 1.0 nsec. Similar results have been obtained earlier in studies by Lee et al. (2001a,b), who showed that simulations up to 1 nsec in explicit solvent were insufficient to observe refinement. From the present study, it is clear that only if the simulations

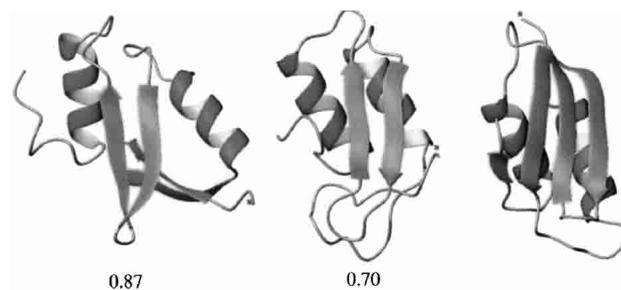


Figure 4. (From left to right) ROSETTA model 2 of 1afi, the final structure after 400 nsec of simulation, and the NMR structure of 1afi. The backbone RMSDs (in nanometers) are shown for each nonnative structure compared with the NMR structure.

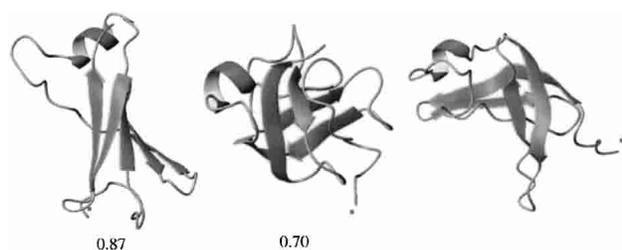


Figure 5. (From left to right) ROSETTA model 2 of 1sro, the final structure after 400 nsec of simulation, and the NMR structure of 1sro. The backbone RMSDs (in nanometers) are shown for each nonnative structure compared with the NMR structure.

are performed on a 10–100-nsec time scale is any systematic refinement observed. It should also be noted that although improvements are observed within 10 nsec, the results of the simulations are chaotic, and large fluctuations in RMSD values are observed. To obtain a refined model, a consensus result from multiple simulations started from different initial conditions (i.e., random velocities) would be desirable.

In Figures 6 and 7, the results for two cases are illustrated. Figure 6 shows the initial ROSETTA model (left), the structure after 10 nsec of simulation at 300 K (top center), the structure after 10 nsec of simulation at 325 K (bottom center), and the experimentally determined structure (right) for model 1 of the protein 1sap. Figure 7 shows the analogous structures for model 1 of the protein 1lea. In both

of these cases, the RMSD after simulating for 10 nsec at 325 K is slightly lower than that of the original model, whereas the RMSD after 10 nsec at 300 K is higher than the original model (Table 4). In the case of 1sap (Fig. 6), the primary differences between the ROSETTA model and the experimental structure lie in the relative positions of the secondary structure elements. In the experimental structure there is a high degree of curvature in the C-terminal helix, and in this projection the N-terminal hairpin lies lower than in the initial model. The global fold is correct. In the simulations the global fold is maintained. There is a shift in the N-terminal hairpin, especially in the simulation at 325 K, and the loss of some of the central sheet region, which is overextended in the initial model. The C-terminal helix also shifts in the simulations. In the final structure from the simulation at 325 K the primary difference with respect to the experimental structure is the orientation of the C-terminal residues. In fact, eliminating the last five residues from the fit causes the RMSD to fall from 0.33 to 0.25 nm.

In the case of 1lea (Fig. 7), the primary difference between the initial model and the experimental structure lies in the C-terminal region, which was modeled as a helix–turn but experimentally was found to be a hairpin. There are also differences in the relative positions of the helices. In the two simulations different regions of the molecules undergo major changes. At 300 K, there is a loss of the C-terminal helix–turn configuration found in the initial model, which results in an intermediate with a large RMSD. At 325 K, there is some rearrangement of the relative positions of the

Table 4. A comparison RMSD at various times in simulations performed at 300 K and 325 K

PDBID	Model	RMSD 0 nsec	RMSD (0.1 nsec)		RMSD (1 nsec)		RMSD (10 nsec)	
			300	325	300	325	300	325
1aoy	1	0.59	0.70	0.66	0.71	0.71	0.69	0.71
1stu	1	0.62	0.70	0.73	0.88	0.84	1.08	0.88
1vif	1	0.62	0.72	0.70	0.74	0.81	0.75	0.78
1sro	1	0.45	0.48	0.44	0.47	0.43	0.53	0.42
1sro	2	<i>0.87</i>	<i>0.90</i>	<i>0.87</i>	<i>0.89</i>	<i>0.68</i>	<i>0.70</i>	<i>0.94</i>
1tuc	1	0.59	0.67	0.64	0.72	0.67	0.66	0.68
1tuc	2	0.82	0.85	0.78	0.84	0.77	0.72	0.91
1sap	1	0.35	0.42	0.36	0.43	0.41	0.64	0.33
1afi	1	0.26	0.29	0.23	0.32	0.26	0.21	0.27
1afi	2	<i>0.87</i>	<i>0.90</i>	<i>0.83</i>	<i>0.76</i>	<i>0.74</i>	<i>0.78</i>	<i>0.78</i>
1vcc	1	0.59	0.58	0.63	0.64	0.73	0.63	0.81
2bby	1	0.53	0.47	0.53	0.54	0.55	0.76	0.62
2bby	4	<i>1.05</i>	<i>1.03</i>	<i>1.05</i>	<i>0.91</i>	<i>0.96</i>	<i>0.74</i>	<i>0.98</i>
1alz	1	0.55	0.43	0.44	0.43	0.45	0.41	0.71
1ail	1	0.54	0.60	0.61	0.59	0.59	0.61	0.94
1coo	1	0.58	0.63	0.68	0.68	0.82	0.95	0.93
1lea	1	0.40	0.44	0.46	0.48	0.45	0.63	0.39
2ezh	1	0.34	0.36	0.39	0.43	0.45	0.47	0.39

(Model) Selected models from those shown in Table 1 for which simulations at 325 K were performed. (RMSD) The backbone RMSD (in nanometers) with respect to the experimental structure at 0.1, 1.0, and 10 nsec. In bold are indicated those cases for which the RMSD of the starting model is ≤ 0.5 nm. In italics are indicated those cases for which the RMSD of the starting models is ≥ 0.8 nm.

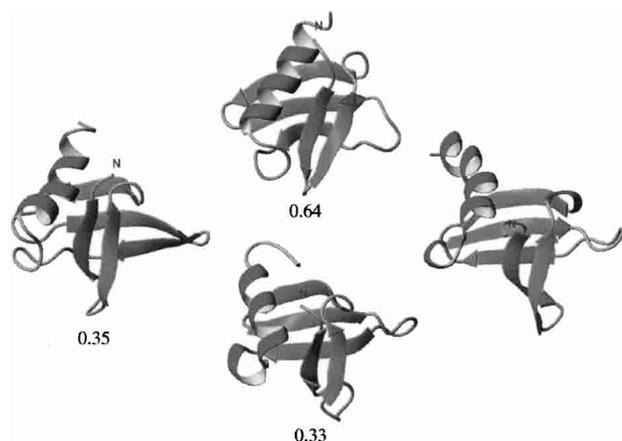


Figure 6. Refinement at different temperatures for model 1 of 1sap. (From left to right) The initial ROSETTA model, the models after 10 nsec of MD simulation at 300 K (top) and at 325 K (bottom), and the NMR structure of 1sap. The backbone RMSDs (in nanometers) are shown for the initial model and two models after simulation with respect to the NMR structure.

helices. There is also a partial loss of the C-terminal helix, but the tight packing (turn) in the C-terminal region is maintained. Together these aspects probably account for the slight reduction in RMSD with respect to the experimental structure. This example also illustrates that global measures such as RMSD are poor indicators of the nature of the changes in the simulations and the degree of refinement in individual cases. When using RMSD in this work, it is primarily for considering trends involving large numbers of structures or for when the structures are very close.

Conclusions

By comparing the results for 60 cases (four models for each of 15 proteins), we have investigated the extent to which classic molecular dynamics simulation techniques performed in explicit solvent are useful for the refinement of the structures of small to medium-size proteins (50–100 amino acids) obtained based on homology or generated *ab initio*. We find that although the structures undergo some initial distortion during the first 1–5 nsec of simulation, significant refinement of the structures is observed at longer time scales in some cases. In the clearest case of refinement, the backbone RMSD was reduced from an initial value of 0.26 nm to 0.17 nm after 100 nsec with values as low as 0.12 nm being sampled. Other cases showed improvements in the number of native hydrogen-bond interactions, and in some cases the loss and refolding of inappropriate elements of secondary structure was observed. The results challenge the widely held belief that molecular dynamics simulations are not useful for the refinement of protein structures unless used in conjunction with experimental restraints (Baker and Sali 2001), a belief based primarily on results from simu-

lations performed on very short times scales or using simplified representations of the protein and/or environment.

The results highlight the fact that to achieve significant local refinement, simulations on an appropriate time scale (>10 nsec) are required. To observe major structural changes, simulations on at least the microsecond time scale will be needed. Increasing the temperature slightly to improve sampling was effective in cases in which the initial model was close to the desired structure but was less effective (in fact, often resulting in major loss of structure) in cases in which the initial model was far from the desired result and not in a local potential energy well. It may be possible to use this as a means of discriminating appropriate from inappropriate models.

Clearly, given the need for extended simulation times and for an accurate representation of the interactions both within the protein and the environment, the use of MD refinement is still limited. We have not demonstrated refinement in all, or even a majority, of the cases investigated. We would also note that in presenting the results of this study, we have been careful not just to illustrate the successful cases. The simulations generate an ensemble of structures and are chaotic in nature. It would, of course, be simple to select from each trajectory the structure that lies closest to the experimental structure in all cases and in this way claim success. However, as there is no reliable means to recognize the “best structure,” an average or consensus result rather than the structure after a given time from a single simulation (as done in this study) should be taken for comparison to experiment. Some comment should be made with respect to the force field. Refinement is only possible if the native structure represents the global minimum for the force field

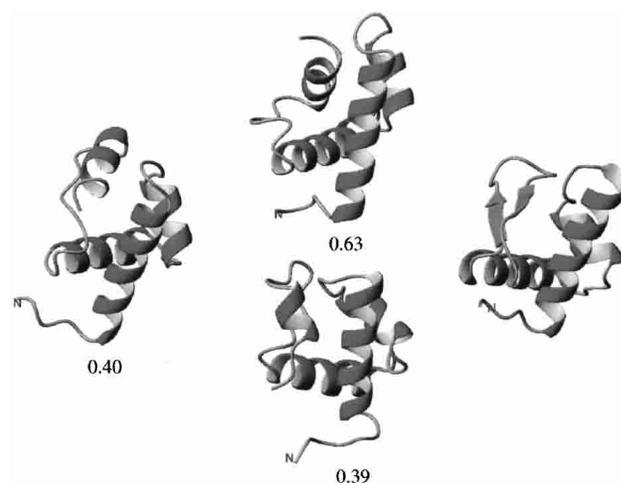


Figure 7. Refinement at different temperatures for model 1 of 1lea. (From left to right) The initial ROSETTA model, the model after 10 nsec of MD simulation at 300 K (top) and at 325 K (bottom), and the NMR structure of 1lea. The backbone RMSDs (in nanometers) are shown for each nonnative structure as in Figure 6.

used in the particular environment simulated. It is important that electrostatic and in particular solvation effects are treated appropriately. This is why the work was performed in explicit solvent. We see from Table 1 that the best results were obtained for 1afi, which also shows the lowest deviation from the starting experimental structure in the control simulations. In this case, the version of the GROMOS96 force field used performs very well. In cases in which little or no refinement was observed, it is difficult to determine if this is a consequence of limited sampling or limitations of the force field.

In conclusion, the results we have presented indicate that MD simulations on a 10–100-nsec time scale performed using an explicit representation of the protein and solvent environment are useful for the refinement of protein models whether based on homology or generated *ab initio*, especially when the initial model has the correct global fold. Clearly, such simulations have an increasingly important role to play in structural genomics.

Materials and methods

Structure selection

The test set used in this study was a selection of 15 proteins ranging in size from 50 to 100 amino acids that had been used previously by Baker and coworkers to verify the structure prediction algorithm ROSETTA (Simons et al. 1999, 2001). The proteins chosen do not contain disulphide bridges, and any bound ions or ligands in the experimental structures were removed before simulation. Four models for each of the 15 proteins were simulated, giving 60 models in total. The models, generated using the ROSETTA procedure, were obtained from the Baker Web site (<http://depts.washington.edu/bakerpg/decoys>). The structures provided contained the peptide backbone and C_β atoms only. The side chains were generated using the program CORALL from the WHATIF package. In addition, the experimentally determined structure of each protein, taken from the Protein Data Bank (PDB; Fogh et al. 1994; Sharma et al. 1994; Bycroft et al. 1995, 1997; Edmondson et al. 1995; Jeon et al. 1995; Narayana et al. 1995; Viguera et al. 1995; Clubb et al. 1997; Liu et al. 1997; Musco et al. 1997; Steele and Opella 1997; Sunnerhagen et al. 1997; Eberstadt et al. 1998; Groft et al. 1998), and the experimentally derived structure in which the side chains of all residues had been deleted and then regenerated using the WHATIF program were simulated and used as controls.

The PDBID and other structural properties of the 15 proteins in the test set are listed in Table 1. Four of the proteins were determined by X-ray crystallography, 11 by NMR techniques. Of the NMR structures, five correspond to energy minimized average structures where only a single structure was given in the PDB. In the remaining cases where multiple structures have been deposited in the PDB, the first structure in each set was chosen to represent the molecule. In total, six structures (four models and two controls) for each of the 15 proteins were simulated, giving 90 systems in all.

All simulations were performed in explicit water using the GROMACS (Groningen Machine for Chemical Simulation) package (Berendsen et al. 1995; Lindahl et al. 2001; van der Spoel et

al. 2001) in conjunction with the GROMOS96 43a1 force field for condensed phases (van Gunsteren et al. 1996). The Simple Point Charge (SPC) model was used to represent the water (Berendsen et al. 1981). The protonation state of ionizable groups in each of the proteins was chosen appropriate for pH 7.0. No counterions were added to neutralize the system. The molecular dynamics simulations were performed at constant temperature and pressure in a periodic truncated octahedral box. The minimum distance between any atom of the protein and the box wall was 1.0 nm. Nonbonded interactions were evaluated using a twin-range method. Coulomb and van der Waals interactions within a shorter-range cutoff of 0.9 nm were evaluated every time step. Longer-range Coulomb and van der Waals interactions between 0.9 and 1.4 nm were updated every five time steps, together with the pairlist. To minimize the effects of truncating the electrostatic interactions beyond the 1.4 nm long-range cutoff, a reaction field correction (Tironi et al. 1995) was applied using a relative dielectric constant of 78. To remove high-frequency degrees of freedom, explicit hydrogen atoms in the force field were replaced by dummy atoms, the positions of which were constructed each step from the coordinates of the heavy atoms to which they are attached. This allows a time step of 4 fsec to be used without affecting the thermodynamic properties of the system significantly (Feenstra et al. 1999). Covalent bonds in the protein were constrained using the LINCS algorithm (Hess et al. 1997). The SETTLE algorithm (Miyamoto and Kollman 1992) was used to constrain the geometry of the water molecules. To generate the starting configuration for each system, the following protocol was used. After energy minimization (EM) using a steepest descent algorithm, 10 psec of molecular dynamics with position restraints on the protein (PRMD) were performed at 250 K to gently relax the system. Unrestrained molecular dynamics (MD) were then performed at 300 K for at least 5 nsec of simulation to assess the stability of the structures. During the simulations the temperature and the pressure were maintained at 300 K and 1 bar by coupling to an external heat and an isotropic pressure bath (Berendsen et al. 1984). The relaxation times were 0.1 psec and 0.5 psec, respectively.

Acknowledgments

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

- Abagyan, R.A. and Batalov, S. 1997. Do aligned sequences share the same fold? *J. Mol. Biol.* **273**: 355–368.
- Baker, D. and Sali, A. 2001. Protein structure prediction and structural genomics. *Science* **294**: 93–96.
- Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., and Hermans, J. 1981. Interaction models for water in relation to protein hydration. In *Intermolecular forces* (ed. B. Pullman), pp. 331–342. Reidel, Dordrecht, The Netherlands.
- Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., Di Nola, A., and Haak, J.R. 1984. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**: 3684–3690.
- Berendsen, H.J.C., van der Spoel, D., and van Drunen, R. 1995. GROMACS: A message-passing parallel molecular dynamics implementation. *Comp. Phys. Comm.* **91**: 43–56.
- Bycroft, M., Grunert, S., Murzin, A.G., Proctor, M., and St. Johnston, D. 1995. NMR solution structure of a dsRNA binding domain from *Drosophila* staufen protein reveals homology to the N-terminal domain of ribosomal protein S5. *EMBO J.* **14**: 3563–3571.
- Bycroft, M., Hubbard, T.J., Proctor, M., Freund, S.M., and Murzin, A.G. 1997.

- The solution structure of the S1 RNA binding domain: A member of an ancient nucleic acid-binding fold. *Cell* **88**: 235–242.
- Clubb, R.T., Schumacher, S., Mizuuchi, K., Gronenborn, A.M., and Clore, G.M. 1997. Solution structure of the I γ subdomain of the Mu end DNA-binding domain of phage Mu transposase. *J. Mol. Biol.* **273**: 19–25.
- Dill, K.A. and Chan, H.S. 1997. From Levinthal to pathways to funnels. *Nat. Struct. Biol.* **4**: 10–19.
- Eberstadt, M., Huang, B., Chen, Z., Meadows, R.P., Ng, S.C., Zheng, L., Leonardo, M.J., and Fesik, S.W. 1998. NMR structure and mutagenesis of the FADD (Mort1) death-effector domain. *Nature* **392**: 941–945.
- Edmondson, S.P., Qiu, L., and Shriver, J.W. 1995. Solution structure of the DNA-binding protein Sac7d from the hyperthermophile *Sulfolobus acidocaldarius*. *Biochemistry* **34**: 13289–13304.
- Fan, H. and Mark, A.E. 2003. Relative stability of protein structures determined by X-ray crystallography or NMR spectroscopy: A molecular dynamics simulation study. *Proteins* **53**: 111–120.
- Feenstra, K.A., Hess, B., and Berendsen, H.J.C. 1999. Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *J. Comp. Chem.* **20**: 786–798.
- Fischer, D. and Eisenberg, D. 1997. Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*. *Proc. Natl. Acad. Sci.* **94**: 11929–11934.
- Fogh, R.H., Otteleben, G., Ruterjans, H., Schnarr, M., Boelens, R., and Kaptein, R. 1994. Solution structure of the LexA repressor DNA binding domain determined by ^1H NMR spectroscopy. *EMBO J.* **13**: 3936–3944.
- Groft, C.M., Uljon, S.N., Wang, R., and Werner, M.H. 1998. Structural homology between the Rap30 DNA-binding domain and linker histone H5: Implications for preinitiation complex assembly. *Proc. Natl. Acad. Sci.* **95**: 9117–9122.
- Guex, N., Diemand, A., and Peitsch, M.C. 1999. Protein modeling for all. *Trends Biochem. Sci.* **24**: 364–367.
- Hess, B., Bekker, H., Berendsen, H.J.C., and Fraaije, J.G.E.M. 1997. LINCS: A linear constraint solver for molecular simulations. *J. Comp. Chem.* **18**: 1463–1472.
- Jeon, Y.H., Negishi, T., Shirakawa, M., Yamazaki, T., Fujita, N., Ishihama, A., and Kyogoku, Y. 1995. Solution structure of the activator contact domain of the RNA polymerase α subunit. *Science* **270**: 1495–1497.
- Lee, M.R., Baker, D., and Kollman, P.A. 2001a. 2.1 and 1.8 Å average C_α RMSD structure predictions on two small proteins, HP-36 and S15. *J. Am. Chem. Soc.* **123**: 1040–1046.
- Lee, M.R., Tsai, J., Baker, D., and Kollman, P.A. 2001b. Molecular dynamics in the endgame of protein structure prediction. *J. Mol. Biol.* **313**: 417–430.
- Lindahl, E., Hess, B., and van der Spoel, D. 2001. GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Modeling* **7**: 306–317.
- Liu, J., Lynch, P.A., Chien, C.Y., Montelione, G.T., Krug, R.M., and Berman, H.M. 1997. Crystal structure of the unique RNA-binding domain of the influenza virus NS1 protein. *Nat. Struct. Biol.* **4**: 896–899.
- Miyamoto, S. and Kollman, P.A. 1992. SETTLE: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comp. Chem.* **13**: 952–962.
- Musco, G., Kharrat, A., Stier, G., Fraternali, F., Gibson, T.J., Nilges, M., and Pastore, A. 1997. The solution structure of the first KH domain of FMR1, the protein responsible for the fragile X syndrome. *Nat. Struct. Biol.* **4**: 712–716.
- Narayana, N., Matthews, D.A., Howell, E.E., and Nguyen-huu, X. 1995. A plasmid-encoded dihydrofolate reductase from trimethoprim-resistant bacteria has a novel D2-symmetric active site. *Nat. Struct. Biol.* **2**: 1018–1025.
- Orengo, C.A., Bray, J.E., Hubbard, T., LoConte, L., and Sillitoe, I. 1999. Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. *Proteins* **37 Suppl 3**: 149–170.
- Sanchez, R. and Sali, A. 1998. Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl. Acad. Sci.* **95**: 13597–13602.
- Schonbrun, J., Wedemeyer, W.J., and Baker, D. 2002. Protein structure prediction in 2002. *Curr. Opin. Struct. Biol.* **12**: 348–354.
- Sharma, A., Hanai, R., and Mondragon, A. 1994. Crystal structure of the amino-terminal fragment of vaccinia virus DNA topoisomerase I at 1.6 Å resolution. *Structure* **2**: 767–777.
- Simons, K.T., Bonneau, R., Ruczinski, I., and Baker, D. 1999. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* **37 Suppl 3**: 171–176.
- Simons, K.T., Strauss, C., and Baker, D. 2001. Prospects for ab initio protein structural genomics. *J. Mol. Biol.* **306**: 1191–1199.
- Standley, D.M., Eyrich, V.A., An, Y., Pincus, D.L., Gunn, J.R., and Friesner, R.A. 2001. Protein structure prediction using a combination of sequence-based alignment, constrained energy minimization, and structural alignment. *Proteins Suppl 5*: 133–139.
- Steele, R.A. and Opella, S.J. 1997. Structures of the reduced and mercury-bound forms of MerP, the periplasmic protein from the bacterial mercury detoxification system. *Biochemistry* **36**: 6885–6895.
- Sunnerhagen, M., Nilges, M., Otting, G., and Carey, J. 1997. Solution structure of the DNA-binding domain and model for the complex of multifunctional hexameric arginine repressor with DNA. *Nat. Struct. Biol.* **4**: 819–825.
- Tironi, I.G., Sperb, R., Smith, P.E., and van Gunsteren, W.F. 1995. A generalized reaction field method for molecular dynamics simulations. *J. Chem. Phys.* **102**: 5451–5459.
- Tramontano, A., Leplae, R., and Morea, V. 2001. Analysis and assessment of comparative modeling predictions in CASP4. *Proteins Suppl 5*: 22–38.
- van der Spoel, D., van Buuren, A.R., Apol, E., Meulenhoff, P.J., Tieleman, D.P., Sijbers, A.L.T.M., Hess, B., Feenstra, K.A., Lindahl, E., van Drunen, R., et al. 2001. *Gromacs User Manual version 3.0*. Nijenborgh 4, 9747 AG Groningen, the Netherlands. <http://www.gromacs.org>.
- van Gunsteren, W.F., Billeter, S.R., Eising, A.A., Hünenberger, P.H., Krüger, P.K.H.C., Mark, A.E., Scott, W.R.P., and Tironi, I.G. 1996. *Biomolecular simulation: The GROMOS96 manual and user guide*. vdf Hochschulverlag AG, Zürich.
- Venclovas, C., Zemla, A., Fidelis, K., and Moulton, J. 2001. Comparison of performance in successive CASP experiments. *Proteins Suppl 5*: 163–170.
- Viguera, A.R., Blanco, F.J., and Serrano, L. 1995. The order of secondary structure elements does not determine the structure of a protein but does affect its folding kinetics. *J. Mol. Biol.* **247**: 670–681.
- Vriend, G. 1990. WHAT IF: A molecular modeling and drug design program. *J. Mol. Graph.* **8**: 52–56.
- Williams, M.G., Shirai, H., Shi, J., Nagendra, H.G., Mueller, J., Mizuguchi, K., Miguel, R.N., Lovell, S.C., Innis, C.A., Deane, C.M., et al. 2001. Sequence-structure homology recognition by iterative alignment refinement and comparative modeling. *Proteins Suppl 5*: 92–97.
- Yang, A.S. and Honig, B. 1999. Sequence to structure alignment in comparative modeling using PrISM. *Proteins Suppl 3*: 66–72.