Brute-Force Molecular Dynamics Simulations of Villin Headpiece: Comparison with NMR Parameters

David van der Spoel*

Department of Cell and Molecular Biology, Uppsala University, Husargatan 3, Box 596, SE-751 24 Uppsala, Sweden

Erik Lindahl

Department of Structural Biology, Stanford University School of Medicine, Stanford, California 94305

Received: January 15, 2003; In Final Form: July 9, 2003

Most biomolecular force fields are well-suited for studying static properties of equilibrium structures; however, they have hardly been tested on very-long-time dynamics. This is an important area of research, because computers shortly will be fast enough for quantitative computational studies of folding and stability of small protein domains. In this work, we present an extensive series of 28 50-ns particle-mesh Ewald simulations (a total of 1.4 μs of data) of the Villin headpiece in water, to compare the results of the OPLS−AA/L and GROMOS96 force fields, the influence of water models, protonation states, and the effect of using virtual "dummy" particles for H atoms to increase the simulation time step. In addition to normal properties such as the Cα root-mean-square deviation, hydrogen bonding, and secondary structure, we also calculate the NMR distance restraints, *J*-couplings, and chemical shifts of Villin from the simulation and compare directly with the experimental data. We argue that this comparison of simulated and experimental *ensembles of structures* is a much more accurate tool than the classical comparison of average or momentaneous protein structures. For the Villin system simulated, we find that distance restraint violations in the simulations are comparable to, or even lower than, those from the ensemble obtained by distance geometry. Furthermore, we find that the OPLS−AA/L force field works considerably better with the TIP4P water model than with the simple point charge model. In further simulations, we have tested the effect of the protonation state of the two Glutamate residues, because the structure was determined at pH 3.7. We do find slightly better agreement of the simulations with experimental data when the Glutamate residues are protonated. Finally, we have used multiple different starting structures for most simulations and find that the results are very similar.

1. Introduction

Biomolecular simulation is slowly coming of age.[1] In the past few years, several papers have been published that describe long (>10 ns) dynamics simulations[2,3] or simulations of large systems such as micelles[4] and lipid bilayers.[5,6] Simulations of membrane proteins in a realistic environment are also starting to appear,[7−10] and distributed computing projects for protein folding have been initiated.[11,12] Many of these works have been facilitated by the very-well-optimized GROMACS simulation software,[13,14] which is considerably faster than other packages, because of the implementation of various algorithmic optimizations, e.g., the use of a single sum for virial calculations[15] (which yields a factor of ∼2, because the virial calculation is taken from the inner loops), optimized noniterative constraint algorithms,[16−18] the implementation of methods allowing for long time steps[19] and the use of assembly code for the most time-consuming portions of the program[14] (yielding another factor of 2). The use of single precision further increases the performance, although the effect varies for different processor architectures. The GROMACS software also supports the $\mathcal{O}(N \log N)$ particle-mesh Ewald (PME) algorithm,[20,21] which means that accurate treatment of electrostatics can be used in an efficient simulation package.[22] Thus, the time seems to be ripe for simulations of long-time-scale $\mathcal{O}(\mu s)$ simulations of rare events such as protein folding.

Although most force fields for protein simulation have a long history behind them,[23−26] they have hardly been tested on very-long-time dynamics. Instead, the available force fields were based on models for amino acids and small molecules[27,28] and optimized, at least in part, to maintain the structure of a protein close to the experimental structure from X-ray crystallography.[24] This is not necessarily correct: the environment of a protein in a crystal can be quite different from that of a protein in solution at 300 K; however, it was a necessary tradeoff, because long biomolecular simulations were not feasible a decade ago. Obviously, we want the results of a computer simulation to reproduce all experimental properties of a protein in vivo as closely as possible, but which properties are these?

It is important that the average structure of our simulated protein resemble the experimental structure. "Average" is a key word here, because the experimental structure also represents a time and ensemble average. The positional fluctuations of atoms can, in principle, be compared to the crystallographic B-factors, although these are partially determined by crystal artifacts such as mosaicity.[29] An average structure can also be calculated from experimental NMR data.

A statistical survey comparing the quality of NMR structures to those from X-ray crystallography indicates that NMR

* Author to whom correspondence should be addressed. E-mail: spoel@xray.bmc.uu.se.

10.1021/jp034108n CCC: $25.00
Published on Web 09/06/2003

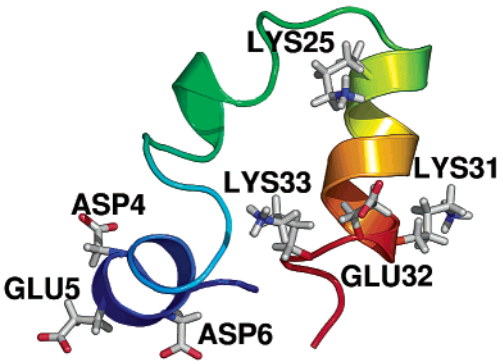

Figure 1. Schematic representation of the Villin headgroup protein domain. Note that Glu32 is surrounded by Lys residues (Lys30 is behind the backbone and not shown).

structures are not as accurate.[30] The main reason for this difference is that NMR, until recently, lacked methods to derive long-range distance information within a protein. The recent advent of methods based on dipolar couplings,[31] in combination with alignment techniques,[32] gives hope for better NMR structures.[33] However, the accuracy of the derived structure is not a problem when comparing simulations to experimental NMR data, as long as the data are accurate and complete (i.e., cover all regions of the protein). It is not very complicated to calculate the average distance restraints from the simulated coordinates and, thus, bypass a large portion of the experimental modeling.

Under physiological conditions, proteins (except, of course, membrane proteins) are usually completely solvated. Therefore, in this work, we will focus on the properties of a protein in solution. More specifically, we have studied the 35-residue Villin headpiece protein subdomain (Figure 1) in water. The NMR structure of this small protein[34] was determined using nuclear Overhauser effect (NOE) distance restraints and dihedral restraints based on *J*-coupling measurements. We compare the simulations using different force fields (GROMOS96,[35] with united-atom aliphatic C atoms, and the all-atom OPLS−AA/L simulation[26,36,37]) to the NMR data and discuss the properties of both force fields for long-time-scale simulations. Furthermore, we test the possibility of replacing H atoms by dummy particles, which was recently shown to allow for larger time steps.[19] In all simulations, we use the PME algorithm[20,21] for treatment of the Coulomb interactions, because long-range electrostatic effects may be important for protein folding.[38]

It is "common knowledge" among computational chemists that each protein force field has a corresponding water model: for example, the Amber[25,39] and CHARMM[23,40] force fields used the TIP3P model,[41] whereas the latest incarnation of the OPLS−AA force field[26,37] was optimized for the TIP4P model.[41] The GROMOS87,[42] as well as the GROMOS96 force field,[35] were optimized for the single point charge (SPC) water model.[43] These simple water models have been compared to test how well they reproduce properties of pure water.[41,44,45] However, tests of force fields with different water models are scarce[46,47] and limited to short peptides. Here, we do compare the performance of the OPLS−AA/L force field for the long-time dynamics simulation of Villin with both SPC water[43] and TIP4P water.[41] Finally, we have performed simulations with different starting conditions to test the reproducibility and statistical sampling.[48]

2. Methods

Protein coordinates were taken from the Protein Databank (PDB entry 1VII) for the NMR structure of the 35-residue Villin headpiece subdomain.[34] The amino acid sequence used for structure determination contains a N-terminal Met residue (for expression) and, therefore, is 36 residues long. The protein consists of a single domain only and has no disulfides; therefore, it can readily unfold under appropriate conditions (even though it is known to have high thermostability[49]) and, thus, represents a good test for a force field. Two additional different starting structures were created by downloading Villin conformations from McKnight's website[50] and selecting two structures—"K" and "P", which have the largest Cα root-mean-square deviation (RMSD) difference (0.26 nm)—giving three reasonably different structures of the same protein. In addition, the PDB structure was used for simulations with three different sets of initial velocities, yielding trajectories labeled A, B, and C.

The protein was dissolved in a truncated octahedron-shaped box that contained 2992 SPC water molecules[43] and ∼70 mM NaCl, to neutralize the protein and provide some electrostatic screening for the relatively large number of charged residues (10 out of 36, not including the termini). In some cases, the SPC water was replaced by TIP4P molecules. Each of the simulation setups described below was applied to 2−5 different combinations of structures and velocities, giving a total of 28 simulations of 50 ns, i.e., almost 1.5 μs of total simulation data. All simulations were used to calculate numerical properties, and nine for each force field (three per setup) are presented graphically in figures in this paper.

Because the NMR structure for Villin was determined at pH 3.7, and the protein contains two glutamates that have a pK_a of 4.3,[51] two extra simulation setups were created with both glutamates protonated. Two different force fields were used: GROMOS96[35] (with united alkane atoms, explicit aromatic H atoms) and the all-atom OPLS force field.[26,36,37] The OPLS−AA/L force field was optimized for use with the TIP4P water model,[41] although it is usually used with the computationally simpler TIP3P or SPC models. To test the importance of the water model, we have performed OPLS simulations with both SPC and TIP4P. Furthermore, in two of the GROMOS96 simulation setups and one OPLS setup, the H atoms were replaced by virtual (dummy) particles. It was recently shown that this makes it possible to use much-longer time steps;[19] in this case, we used 4 fs, which provides an almost twofold acceleration of the simulation. The temperature and pressure were controlled using Berendsen weak coupling[52] at 300 K and 1 bar, with coupling constants of 100 fs for temperature and 20 ps for pressure. A cutoff of 0.8 nm for van der Waals interactions was used, and the smooth PME algorithm[20,21] was used for Coulomb interactions. Finally, neighbor lists were used and updated every fifth integration step. Before each production simulation, the protein structures were energy-minimized and a 200-ps simulation with position restraints on the protein (force constant of 1000 kJ mol^{-1} nm^{-2}) was performed to allow the solvent to relax around the protein. An overview of the simulation setups is given in Table 1. Finally, the 29 structures derived from the experimental data by distance geometry calculations[34] from McKnight's website[50] were used for analysis and comparison. All simulations and analyses were performed using the GROMACS software,[14,53] except as otherwise stated.

3. Results

3.1. Comparison to NMR Distances. The Villin structure was determined by NMR experiments,[34] using distance information from NOEs and dihedral information from *J*-couplings. In the experimental data, 532 distances are defined, which involve 1073 atom pairs. We have measured the distance restraint

TABLE 1: Overview of Simulation Setups[a]

simulation[b]	starting structures[c]	force field	water	dummy	GLUH	Na^+	Cl^-	Δt (fs)
GS	B,C,K,P	GROMOS96	SPC	no	no	3	5	2
GSD	A,B,C,K,P	GROMOS96	SPC	yes	no	3	5	4
GSDH	A,B,C,K,P	GROMOS96	SPC	yes	yes	2	6	4
OS	B,C,K,P	OPLS-AA/L	SPC	no	no	3	5	2
OSH	A,B,C,K	OPLS-AA/L	SPC	no	yes	2	6	2
OT	B,C,K,P	OPLS-AA/L	TIP4P	no	no	3	5	2
OTD	K,P	OPLS-AA/L	TIP4P	yes	no	3	5	4

[a] Each setup was used in 2−5 50-ns simulations, each with different starting structures. [b] Letters in the simulation name designate force field (G/O), water model (S/T), and the use of dummy particles (D) and protonated glutamates (H). [c] Letters A, B, and C started from the PDB structure (with different initial velocities) and the letters K and P refer to structures from the set distributed by McKnight.[50]

TABLE 2: Distance Restraint Violations Averaged over Different Trajectories[a]

		All 532 distances						325 core distances only					
		Linear Averaging			Third-Power Averaging			Linear Averaging			Third-Power Averaging		
simulation[b]	starting structure	#V	Vm (nm)	ΣV (nm)	#V	Vm (nm)	ΣV (nm)	#V	Vm (nm)	ΣV (nm)	#V	Vm (nm)	ΣV (nm)
NMR		68	0.30	3.3	67	0.30	3.1	42	0.17	1.7	42	0.17	1.7
GS	B	99	0.54	8.7	73	0.51	5.8	39	0.20	1.8	32	0.14	1.2
GS	C	118	0.87	12.5	80	0.75	6.7	49	0.18	2.0	31	0.15	1.2
GS	K	105	0.54	9.3	71	0.50	5.8	41	0.19	1.8	29	0.13	1.2
GS	P	135	0.70	16.9	88	0.63	9.2	68	0.55	6.9	45	0.27	3.6
GSD	A	102	0.55	9.1	71	0.51	6.0	41	0.21	1.8	31	0.15	1.2
GSD	B	139	0.83	16.2	96	0.69	8.2	54	0.26	2.8	40	0.23	1.8
GSD	C	136	0.74	13.0	69	0.61	6.2	48	0.23	2.0	30	0.16	1.3
GSD	K	109	0.68	11.1	79	0.57	6.9	42	0.20	2.0	34	0.15	1.3
GSD	P	100	0.65	8.9	76	0.59	5.6	40	0.20	1.9	31	0.14	1.3
GSDH	A	122	0.59	12.6	78	0.56	6.9	46	0.19	2.3	33	0.14	1.4
GSDH	B	136	0.73	18.7	89	0.55	9.9	59	0.64	7.4	48	0.44	4.2
GSDH	C	107	0.64	10.7	73	0.56	6.1	42	0.20	1.9	31	0.14	1.2
GSDH	K	106	0.65	9.6	77	0.58	5.3	42	0.18	1.7	31	0.13	1.1
GSDH	P	112	0.64	13.3	86	0.58	8.1	45	0.20	2.2	36	0.14	1.5
OS	B	128	0.57	15.2	100	0.43	9.3	60	0.22	3.0	47	0.18	2.2
OS	C	130	0.61	14.5	110	0.59	10.6	66	0.25	4.6	58	0.19	3.2
OS	K	119	0.60	12.7	94	0.58	7.7	57	0.22	3.1	44	0.20	2.2
OS	P	124	0.57	14.6	94	0.41	8.7	58	0.23	2.9	44	0.15	2.0
OSH	A	135	0.64	14.4	101	0.61	9.8	66	0.31	3.9	52	0.21	2.4
OSH	B	125	0.58	12.0	90	0.50	7.1	60	0.26	3.1	41	0.19	2.0
OSH	C	134	0.55	13.7	104	0.53	8.6	65	0.22	3.3	47	0.18	2.2
OSH	K	130	0.59	12.5	101	0.49	7.4	56	0.25	2.8	45	0.22	2.0
OT	B	118	0.59	12.2	93	0.49	8.0	55	0.29	2.7	40	0.23	1.9
OT	C	135	0.65	17.9	103	0.61	10.0	56	0.27	3.0	42	0.22	2.0
OT	K	131	0.53	15.5	95	0.50	9.8	61	0.28	3.0	39	0.20	2.0
OT	P	129	0.58	11.8	94	0.47	7.3	65	0.27	3.9	49	0.16	2.5
OTD	K	105	0.45	10.7	93	0.40	7.4	51	0.34	2.6	45	0.31	2.0
OTD	P	108	0.44	11.0	91	0.41	7.4	48	0.32	2.3	41	0.28	1.9

[a] Legend for column headers: #V, number of violations; Vm, largest violation; and ΣV, sum of violations. [b] NMR indicates 29 structures derived from distance geometry; see Table 1 for simulation names.

violations that result from simulations using the different force fields. Some of the experimental restraints represent interactions between, for example, methyl groups, which lead to a single NOE. In these cases, the distances between all the protons were summed by the sixth power:

$$d = (\sum_i \sum_j d_{ij}^6)^{1/6}$$

where i and j designate the elements in the individual methyl groups (or other groups with more than one proton) and d is the effective distance used for computing the distance violations. The NOE signal, as measured, is an average over time. However, it is not a linear average, but rather a third-power average:[54]

$$\langle d \rangle = \langle d(t)^3 \rangle^{1/3}$$

In Table 2, we have summarized the number of violations, the largest violation, and the sum of violations as the average over all the simulation trajectories. We have performed the distance restraint analysis on two groups of distances: all restraints and core restraints, the latter indicating restraints with only backbone atoms or side chain atoms up until Hγ. If we select a simple linear average, there are no large differences between simulations; however, when the more-accurate third-power averaging is applied, the differences are more pronounced, apparently in favor of the GROMOS96 force field. The ratio of linear and third-power averaging for the sum of violations of core distances was computed for all the GROMOS96 simulations and all the OPLS simulations. For GROMOS96, the ratio is 1.58, and for OPLS, it is 1.44, which indicates that the protein in the GROMOS96 simulation samples a larger portion of conformation space and can, hence, satisfy the different distance restraints more easily. Also, note that, in 11 out of 14 simulations, the sum of violations for the GROMOS96 simulations is actually lower than that computed based on the 29 experimental structures. Finally, the OPLS force field performs better in combination with the TIP4P water than with SPC, and in both OPLS- and GROMOS96-based simula-

TABLE 3: Correlation between Measured $^3J_{NH}$ Coupling Constants and Simulated Constants[a]

residue[b]	experimental data[c]	NMR	GS	GSD	GSDH	OS	OSH	OT
L2	6.1	3.5 ± 1.5	5.9 ± 2.5	7.1 ± 2.5	6.3 ± 2.7	8.1 ± 1.6	8.0 ± 1.6	7.9 ± 1.7
S3	5.2	5.6 ± 2.1	7.2 ± 2.2	7.8 ± 2.1	7.5 ± 1.9	8.7 ± 1.2	8.4 ± 1.3	8.3 ± 1.5
D4	3.0	3.1 ± 1.9	2.9 ± 1.3	3.1 ± 1.1	3.2 ± 1.1	3.5 ± 1.5	3.9 ± 1.5	4.4 ± 1.9
E5	4.7	2.3 ± 0.3	3.8 ± 1.6	4.2 ± 1.3	4.6 ± 1.2	5.6 ± 1.6	5.3 ± 1.7	5.9 ± 1.7
D6	6.3	2.7 ± 0.1	3.7 ± 1.5	4.0 ± 1.2	4.0 ± 1.1	6.4 ± 1.7	6.9 ± 1.7	7.4 ± 1.8
F7	3.0	2.7 ± 0.3	3.7 ± 1.3	4.3 ± 1.4	3.8 ± 1.0	3.9 ± 1.6	3.5 ± 1.5	4.0 ± 2.2
K8	4.1	1.9 ± 0.2	3.3 ± 1.3	3.8 ± 1.2	3.8 ± 1.1	5.3 ± 1.6	5.1 ± 1.6	3.8 ± 1.7
A9	4.8	9.5 ± 0.4	4.2 ± 1.8	4.6 ± 1.6	5.0 ± 1.5	8.0 ± 1.7	8.2 ± 1.6	8.2 ± 1.6
V10	7.3	7.5 ± 1.0	6.0 ± 2.1	6.7 ± 2.1	6.4 ± 1.9	6.0 ± 2.0	5.9 ± 2.0	6.2 ± 2.1
F11	9.0	7.0 ± 0.7	8.3 ± 1.7	8.1 ± 1.9	8.4 ± 1.9	8.2 ± 1.5	8.2 ± 1.4	7.5 ± 1.9
M13	3.3	7.9 ± 0.6	6.8 ± 2.7	8.6 ± 1.8	7.1 ± 2.4	7.9 ± 1.5	7.9 ± 1.5	7.9 ± 1.6
T14	6.9	9.4 ± 0.7	5.7 ± 2.0	5.1 ± 1.9	5.1 ± 1.6	6.1 ± 1.6	5.8 ± 1.6	6.6 ± 1.7
R15	3.0	5.5 ± 1.9	4.7 ± 1.8	3.7 ± 1.2	3.9 ± 1.2	4.0 ± 1.5	3.6 ± 1.4	4.2 ± 1.4
S16	4.5	3.7 ± 0.7	3.6 ± 1.4	4.2 ± 1.2	4.0 ± 1.1	5.1 ± 1.5	5.1 ± 1.5	5.0 ± 1.5
A17	5.7	5.4 ± 1.0	4.1 ± 1.6	4.4 ± 1.2	4.4 ± 1.2	7.1 ± 1.7	7.1 ± 1.7	7.1 ± 1.7
F18	3.3	3.2 ± 0.9	3.7 ± 1.4	4.1 ± 1.1	4.1 ± 1.1	5.1 ± 1.6	5.0 ± 1.5	5.6 ± 1.7
A19	3.8	7.1 ± 0.9	3.9 ± 1.5	4.4 ± 1.3	4.2 ± 1.1	5.3 ± 1.6	5.2 ± 1.6	5.2 ± 1.5
N20	9.1	8.6 ± 1.3	4.8 ± 2.2	6.1 ± 2.3	5.3 ± 1.9	7.7 ± 1.6	7.6 ± 1.6	7.5 ± 1.7
L21	6.9	6.9 ± 1.9	5.3 ± 2.5	5.7 ± 2.3	5.8 ± 2.2	7.8 ± 1.5	7.9 ± 1.5	7.9 ± 1.5
L23	3.0	3.5 ± 0.3	3.2 ± 1.5	3.3 ± 1.2	3.3 ± 1.1	3.1 ± 1.2	3.1 ± 1.2	3.0 ± 1.2
W24	3.0	5.5 ± 0.5	3.3 ± 1.5	3.8 ± 1.1	3.8 ± 1.1	4.2 ± 1.4	4.5 ± 1.6	4.5 ± 1.5
K25	8.0	6.2 ± 0.3	4.2 ± 1.8	4.4 ± 1.2	4.4 ± 1.2	7.7 ± 1.5	7.7 ± 1.5	7.4 ± 1.5
Q26	2.5	5.9 ± 0.4	3.7 ± 1.4	3.8 ± 1.0	3.7 ± 1.0	5.5 ± 1.7	5.1 ± 1.7	5.2 ± 1.6
Q27	4.0	2.4 ± 0.3	3.9 ± 1.4	4.2 ± 1.1	4.2 ± 1.1	5.4 ± 1.6	5.0 ± 1.5	5.0 ± 1.6
N28	4.3	5.7 ± 0.7	3.6 ± 1.4	3.8 ± 1.1	3.9 ± 1.1	6.6 ± 2.1	6.1 ± 2.0	5.6 ± 1.9
L29	4.6	3.0 ± 0.6	4.2 ± 1.6	4.2 ± 1.2	4.3 ± 1.1	7.2 ± 1.8	6.1 ± 2.0	6.0 ± 1.9
K30	4.2	6.0 ± 1.0	3.8 ± 1.6	4.0 ± 1.2	4.1 ± 1.1	7.2 ± 2.5	5.1 ± 2.1	5.3 ± 1.7
K31	4.8	3.5 ± 0.8	4.7 ± 2.2	4.6 ± 1.7	4.7 ± 1.4	7.3 ± 2.2	7.0 ± 1.9	7.0 ± 2.1
E32	4.5	7.3 ± 1.0	5.2 ± 2.4	5.4 ± 1.9	4.8 ± 1.7	8.5 ± 1.4	7.7 ± 1.9	7.0 ± 2.1
K33	8.9	8.9 ± 1.0	8.6 ± 1.6	8.1 ± 2.0	8.5 ± 1.7	7.5 ± 2.0	7.5 ± 1.9	8.2 ± 1.6
L35	9.7	9.1 ± 0.9	8.3 ± 2.0	7.8 ± 1.9	8.7 ± 1.6	8.3 ± 1.6	8.1 ± 1.6	8.1 ± 1.6
F36	8.5	8.5 ± 1.1	9.0 ± 1.2	8.9 ± 1.2	8.9 ± 1.4	6.8 ± 1.7	5.7 ± 1.3	6.8 ± 1.6
CC		0.57	0.71	0.67	0.74	0.62	0.65	0.69
a		0.62 ± 0.2	0.58 ± 0.1	0.54 ± 0.1	0.58 ± 0.1	0.46 ± 0.1	0.48 ± 0.1	0.49 ± 0.1
b		2.2 ± 0.1	1.8 ± 0.6	2.3 ± 0.6	2.0 ± 0.6	4.0 ± 0.6	3.6 ± 0.6	3.7 ± 0.5

[a] A linear fit through a plot with the correlation between experiment and simulation was made, and the correlation coefficient (CC), the slope of the fit (*a*), and the intercept (*b*) are given in the last three rows. NMR indicates 29 structures derived from distance geometry; see Table 1 for simulation names. [b] Residues M1, G12, P22, and G34 were missing from the experimental data. [c] Experimental data from ref 49.

tions, we see that, at least in some cases, the peptide with protonated Glutamate side chains gives slightly lower violations.

3.2. Comparison to NMR *J*-Couplings. $^3J_{NH}$ coupling constants were estimated from simulation C (starting from the PDB structure) using a Karplus relation:[55]

$$J(\theta) = A\cos^2\theta + B\cos\theta + C$$

with parameters $A = 6.51$, $B = -1.76$, and $C = 1.6$.[56] Although several other parametrizations for the $^3J_{NH}$ Karplus relation have been suggested,[57−59] we have used the parameters of Vuister and Bax,[56] because these were derived from a specifically designed experiment. A comparison of experimental data and simulation results is given in Table 3. The correspondence between simulation and experiment is not perfect, although the experimental values fall within the standard deviations of the simulation in most cases, with the exceptions being residues S3, M13, N20, and Q26. The correlation between simulations and data is quantified in Table 3. The fact that the slope in the regression analysis (variable *a* in Table 3) is <1 highlights the problem that the very high and very low values are not well reproduced. The simulations using the GROMOS96 force field perform slightly better than the OPLS−AA force fields; however, once again, we see that OPLS performs considerably better in combination with TIP4P than with SPC water. Furthermore, the simulations with the protonated Glu residues (GSDH and OSH) have higher correlation constants than the corresponding simulations using unprotonated Glutamates (GS and OS). The average $^3J_{NH}$ coupling constants from the experimental distance geometry ensemble are actually slightly *worse* than those from the simulations, stressing the necessity of sampling many structures.

3.3. Comparison to Chemical Shifts. Chemical shifts were calculated for the Hα protons, using the "shifts" program from the Case group,[60−62] and are compared to the experimental values from ref 49. Simulation C (starting from the PDB structure) was used for this analysis. To compare fairly, we computed the deviation of chemical shifts from the random coil values,[63] and the results are listed in Table 4. In most cases, the experimental value is within the error bars of the simulated values, except for residues F7 and K8. The correlation between experimental data and simulated results is also given in Table 4. As in the case of the $^3J_{NH}$ coupling constants, the slope of the regression analysis is <1, which indicates that, in particular, large deviations from the random coil shifts are reproduced poorly. The correlation coefficients are comparable to those for the $^3J_{NH}$ coupling constants. Here, too, we see that the chemical shifts computed from the distance geometry ensemble do not correlate with the measured chemical shifts as well as the simulation averages do.

3.4. Comparison to Average NMR Structure. The positional RMSD from the experimental structure is the simplest measure to compare structures. The Cα RMSD values are plotted in Figure 2. There are significant fluctuations that

TABLE 4: Correlation between Measured Hα Chemical Shifts and Simulated Chemical Shifts[a]

residue[b]	experimental data[c]	NMR	GS	GSD	GSDH	OS	OSH	OT
S3	0.03	−0.1 ± 0.2	0.19 ± 0.3	0.11 ± 0.3	0.19 ± 0.3	0.18 ± 0.2	0.15 ± 0.2	−0.02 ± 0.3
D4	−0.26	−0.3 ± 0.2	−0.25 ± 0.3	−0.41 ± 0.4	−0.36 ± 0.3	−0.21 ± 0.2	−0.22 ± 0.2	−0.19 ± 0.2
E5	−0.19	−0.3 ± 0.0	−0.19 ± 0.2	−0.23 ± 0.2	−0.17 ± 0.2	−0.21 ± 0.2	−0.22 ± 0.2	−0.17 ± 0.2
D6	−0.12	−0.4 ± 0.0	−0.23 ± 0.2	−0.26 ± 0.2	−0.20 ± 0.2	−0.18 ± 0.2	−0.07 ± 0.2	−0.08 ± 0.2
F7	−0.76	−0.6 ± 0.2	−0.43 ± 0.3	−0.44 ± 0.4	−0.35 ± 0.3	−0.40 ± 0.2	−0.32 ± 0.2	−0.34 ± 0.2
K8	0.04	−0.2 ± 0.1	−0.39 ± 0.4	−0.52 ± 0.4	−0.59 ± 0.4	−0.51 ± 0.2	−0.60 ± 0.2	−0.46 ± 0.2
A9	−0.24	−0.0 ± 0.1	−0.25 ± 0.3	−0.26 ± 0.3	−0.23 ± 0.3	−0.07 ± 0.2	−0.03 ± 0.2	−0.09 ± 0.2
V10	−0.58	−0.2 ± 0.1	−0.17 ± 0.4	−0.23 ± 0.5	−0.26 ± 0.4	−0.37 ± 0.3	−0.36 ± 0.3	−0.28 ± 0.2
F11	−0.39	−0.3 ± 0.2	0.07 ± 0.4	−0.07 ± 0.4	−0.13 ± 0.4	−0.04 ± 0.3	−0.22 ± 0.3	−0.28 ± 0.2
G12	−0.07	0.0 ± 0.1	0.04 ± 0.3	0.14 ± 0.3	0.18 ± 0.3	0.14 ± 0.2	0.07 ± 0.2	−0.01 ± 0.2
G12	0.07	0.2 ± 0.1	−0.01 ± 0.3	0.09 ± 0.3	0.12 ± 0.3	0.22 ± 0.2	0.15 ± 0.3	0.08 ± 0.2
M13	0.32	0.2 ± 0.1	−0.01 ± 0.3	0.09 ± 0.3	0.11 ± 0.3	0.01 ± 0.2	0.00 ± 0.2	−0.04 ± 0.2
T14	0.16	−0.3 ± 0.6	0.19 ± 0.3	0.10 ± 0.3	0.10 ± 0.3	−0.07 ± 0.2	−0.09 ± 0.2	0.02 ± 0.2
R15	−1.08	−0.3 ± 0.5	−0.71 ± 0.6	−0.76 ± 0.7	−1.15 ± 0.5	−0.85 ± 0.4	−0.91 ± 0.4	−0.31 ± 0.3
S16	−0.41	−0.3 ± 0.1	−0.17 ± 0.2	−0.23 ± 0.2	−0.24 ± 0.2	−0.28 ± 0.2	−0.30 ± 0.2	−0.22 ± 0.2
A17	−0.23	−0.2 ± 0.1	−0.30 ± 0.3	−0.32 ± 0.3	−0.34 ± 0.2	−0.14 ± 0.2	−0.14 ± 0.2	−0.10 ± 0.2
F18	−0.51	−0.4 ± 0.1	−0.37 ± 0.3	−0.46 ± 0.2	−0.43 ± 0.2	−0.30 ± 0.2	−0.32 ± 0.2	−0.26 ± 0.2
A19	−0.37	−0.6 ± 0.1	−0.24 ± 0.3	−0.28 ± 0.3	−0.34 ± 0.3	−0.41 ± 0.2	−0.44 ± 0.2	−0.46 ± 0.2
N20	−0.10	−0.0 ± 0.2	−0.23 ± 0.3	−0.19 ± 0.3	−0.24 ± 0.3	−0.19 ± 0.2	−0.20 ± 0.2	−0.20 ± 0.2
L21	−0.02	−0.3 ± 0.2	−0.34 ± 0.3	−0.34 ± 0.3	−0.30 ± 0.3	−0.19 ± 0.2	−0.21 ± 0.2	−0.19 ± 0.2
W24	−0.22	−0.2 ± 0.0	−0.28 ± 0.2	−0.27 ± 0.2	−0.25 ± 0.2	−0.31 ± 0.2	−0.27 ± 0.2	−0.28 ± 0.2
K25	−0.66	−0.3 ± 0.1	−0.63 ± 0.4	−0.75 ± 0.4	−0.71 ± 0.4	−0.35 ± 0.3	−0.40 ± 0.4	−0.44 ± 0.4
Q26	−0.92	−0.5 ± 0.2	−0.61 ± 0.4	−0.72 ± 0.4	−0.65 ± 0.4	−0.68 ± 0.3	−0.58 ± 0.3	−0.54 ± 0.3
Q27	−0.35	−0.3 ± 0.1	−0.41 ± 0.2	−0.40 ± 0.2	−0.39 ± 0.3	−0.35 ± 0.2	−0.32 ± 0.2	−0.38 ± 0.2
N28	−0.24	−0.2 ± 0.1	−0.31 ± 0.3	−0.36 ± 0.3	−0.41 ± 0.3	−0.42 ± 0.3	−0.36 ± 0.3	−0.43 ± 0.3
L29	−0.07	−0.3 ± 0.1	−0.23 ± 0.3	−0.34 ± 0.3	−0.33 ± 0.3	−0.04 ± 0.2	−0.17 ± 0.2	−0.13 ± 0.2
K30	−0.25	−0.2 ± 0.4	−0.13 ± 0.4	−0.10 ± 0.4	−0.15 ± 0.4	0.10 ± 0.3	−0.21 ± 0.3	−0.10 ± 0.3
K31	−0.19	−0.5 ± 0.1	−0.15 ± 0.3	−0.37 ± 0.4	−0.37 ± 0.4	−0.20 ± 0.3	−0.27 ± 0.2	−0.24 ± 0.2
E32	−0.21	−0.2 ± 0.1	−0.12 ± 0.3	−0.16 ± 0.3	−0.23 ± 0.3	−0.01 ± 0.2	−0.10 ± 0.2	−0.15 ± 0.2
K33	−0.01	−0.0 ± 0.2	0.07 ± 0.3	0.03 ± 0.4	−0.03 ± 0.3	−0.18 ± 0.3	−0.28 ± 0.3	−0.15 ± 0.2
G34	−0.16	−0.3 ± 0.3	−0.18 ± 0.3	−0.34 ± 0.6	−0.38 ± 0.6	0.03 ± 0.5	−0.59 ± 0.7	−0.17 ± 0.3
G34	0.02	−0.0 ± 0.4	−0.22 ± 0.4	−0.40 ± 0.7	−0.44 ± 0.7	0.13 ± 0.5	−0.54 ± 0.7	−0.16 ± 0.4
L35	0.15	0.2 ± 0.2	−0.24 ± 0.4	−0.12 ± 0.4	−0.14 ± 0.4	−0.16 ± 0.2	−0.19 ± 0.2	−0.03 ± 0.2
F36	−0.21	0.0 ± 0.2	0.01 ± 0.3	−0.04 ± 0.3	−0.03 ± 0.3	−0.38 ± 0.2	−0.50 ± 0.2	−0.31 ± 0.2
CC		0.62	0.69	0.71	0.71	0.71	0.59	0.65
a		0.42 ± 0.1	0.48 ± 0.09	0.55 ± 0.10	0.63 ± 0.11	0.56 ± 0.10	0.44 ± 0.11	0.33 ± 0.07
b		−0.11 ± 0.04	−0.10 ± 0.03	−0.13 ± 0.04	−0.12 ± 0.04	−0.06 ± 0.04	−0.16 ± 0.04	−0.13 ± 0.03

[a] A linear fit through a plot with the correlation between experiment and simulation was made, and the correlation coefficient (CC), the slope (*a*), and the intercept (*b*) are tabulated below. For parameters *a* and *b*, the standard deviations are given in parentheses. NMR indicates 29 structures derived from distance geometry; see Table 1 for simulation names. [b] Residues M1, L2, P22, and L23 were missing from the experimental data, whereas for G12 and G34, chemical shifts of both Hα protons were determined. [c] Experimental data from ref 49.

correlate well with the radius of gyration; however, this is not surprising, considering that the total amount of simulation covers 0.9 μs. Setup GSDH for the "P" structure could be partially unfolding; however, because there is no corresponding increase in the radius of gyration, it might only be a slow change of conformation.

Several of the OPLS simulations show long periods with relatively small RMSD values (<0.2 nm), with a few large fluctuations and corresponding increases in the radius of gyration (Figure 3). In contrast, in the GROMOS96 simulation, the protein seems to undergo larger and more-frequent conformational changes. These are also correlated with changes in the radius of gyration, but less so than the OPLS simulations. For setup GSDH, both the run starting from the PDB structure and conformation P show partial unfolding, with RMSD values up to 0.45 nm. However, the run that starts from the PDB structure converges down to 0.25 nm again at the end of the simulation, which indicates that this setup is not inherently unstable. The same setup without protonated Glu residues (GSD) shows nice results; therefore, we have no reason to believe that this fluctuation was caused by the use of dummy particles. However, the RMSD, with respect to the experimental structure, as a function of time, is not the best way to compare a molecular dynamics (MD) trajectory to an experimental structure. What we really want to quantify is how well the average from the

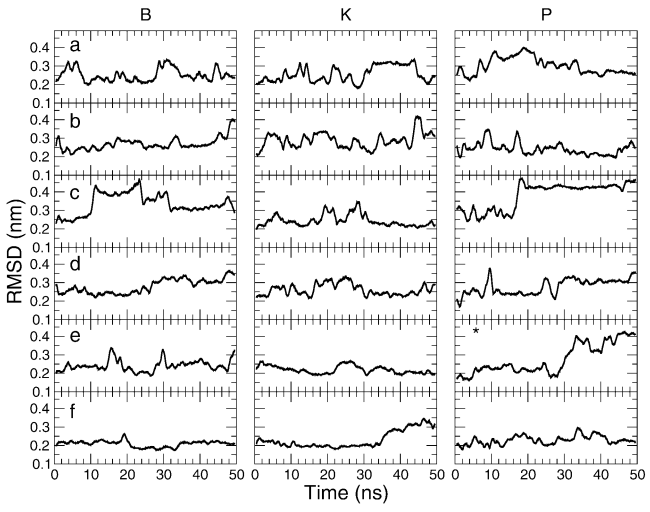

Figure 2. Cα root-mean-square deviation (RMSD) from the NMR structure of Villin, as a function of time. (a) GS simulations, (b) GSD simulations, (c) GSDH simulations, (d) OS simulations, (e) OSH simulations (asterisk (*) indicates that starting structure A rather than P was used here), and (f) OT simulations. For clarity, a running average over 1 ns is plotted. The difference in the initial RMSD between setups is due to changes during energy minimization and equilibration. See Table 1 for an explanation of the simulation names.

TABLE 5: Average Properties from Simulations[a]

simulation[b]	starting structure	SASA (nm^2)	R_g (nm)	Cα RMSD (nm)	RMST	SS	$\#HB_{PP}$	$\#HB_{PS}$	$\#HB_{tot}$
NMR		16.7	0.94	0.19	0.27	19.6	20.4		
GS	B	15.6	0.92	0.22	0.49	22.8	26.6	88.2	141.1
GS	C	14.8	0.91	0.19	0.59	21.9	26.0	90.4	141.4
GS	K	15.8	0.92	0.22	0.46	22.8	26.7	88.2	141.6
GS	P	15.2	0.91	0.25	0.53	19.8	25.4	88.7	139.5
GSD	A	15.7	0.92	0.22	0.39	22.6	26.6	87.8	141.0
GSD	B	15.3	0.93	0.21	0.69	20.6	25.9	88.3	140.1
GSD	C	16.0	0.93	0.20	0.57	22.3	25.4	90.5	140.9
GSD	K	15.7	0.92	0.24	0.52	22.4	26.1	88.7	140.9
GSD	P	15.4	0.91	0.20	0.56	22.6	26.8	87.0	140.6
GSDH	A	16.2	0.91	0.31	0.46	19.7	24.8	81.2	130.8
GSDH	B	16.3	0.92	0.26	0.67	19.5	25.0	80.7	130.7
GSDH	C	16.1	0.92	0.21	0.59	22.8	26.3	79.0	131.6
GSDH	K	15.7	0.91	0.21	0.51	22.8	26.2	80.0	132.4
GSDH	P	16.4	0.92	0.30	0.54	22.4	26.7	78.4	131.8
OS	B	17.3	0.97	0.25	0.45	20.0	23.6	100.9	148.1
OS	C	16.8	0.95	0.17	0.39	17.0	21.5	104.5	147.5
OS	K	16.8	0.96	0.23	0.40	19.8	25.0	98.7	148.7
OS	P	17.3	0.97	0.24	0.52	19.7	23.5	101.1	148.1
OSH	A	18.2	0.97	0.21	0.57	17.7	20.9	98.1	139.9
OSH	B	17.4	0.97	0.18	0.66	19.6	22.9	94.4	140.2
OSH	C	17.8	0.96	0.21	0.45	19.5	23.0	93.2	139.2
OSH	K	16.8	0.96	0.17	0.42	18.9	22.7	95.0	140.4
OT	B	16.8	0.96	0.18	0.37	18.9	24.0	99.8	147.8
OT	C	17.7	0.98	0.26	0.55	17.8	23.9	99.4	147.2
OT	K	16.3	0.96	0.19	0.43	17.8	22.9	101.9	147.7
OT	P	17.1	0.98	0.18	0.52	17.3	23.1	100.4	146.6
OTD	K	16.8	0.95	0.16	0.39	19.5	26.3	97.4	150.0
OTD	P	16.6	0.96	0.18	0.41	19.3	25.8	99.1	150.7

[a] Table legend is as follows: SASA, hydrophobic solvent-accessible surface area; R_g, radius of gyration; Cα RMSD, root-mean-square deviation of average structure from the experimental structure; RMST, maximum Cα RMSD between any two structures in the trajectory; SS, number of residues in α-helix; $\#HB_{PP}$, number of intramolecular hydrogen bonds; $\#HB_{PS}$, number of hydrogen bonds between protein and solvent; and $\#HB_{tot}$, total number of hydrogen bonds ($\#HB_{tot} = 2(\#HB_{PP}) + \#HB_{PS}$). There are three lines for each setup, corresponding to simulations with different starting structures. [b] NMR indicates 29 structures derived from distance geometry; see Table 1 for simulation names.

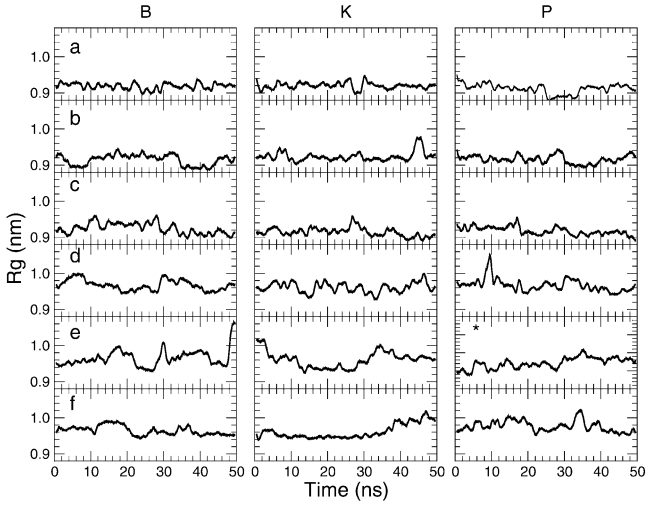

Figure 3. Radius of gyration, as a function of time in the simulations. (a) GS simulations, (b) GSD simulations, (c) GSDH simulations, (d) OS simulations, (e) OSH simulations, and (f) OT simulations, For clarity, a running average over 1 ns is plotted. See Table 1 for simulation names.

MD trajectory compares to the experimental data. Therefore, we have computed the average coordinates from all the trajectories and computed the RMSD values of these (unphysical) structures to the experimental PDB structure. In Table 5, we see that this RMSD is 0.19 nm for the NMR distance geometry ensemble. For the simulations, the results are comparable, with the best results achieved for OPLS and TIP4P water (0.16 nm—lower than the internal spread in the NMR ensemble). The worst case is that for GSDH simulation P, where we suspected partial unfolding: an average RMSD of 0.30 nm seems to support this observation. Recall that the experimental structure is also the (energy-minimized) average over 29 structures and, therefore, is not necessarily physically realistic. The average radii of gyration in the simulations are listed in Table 5 as well; in the GROMOS96 simulations, the peptide is somewhat more compact than the experimental structure, whereas in the OPLS simulations, the simulation average is consistently slightly larger than the experimental value. The extent of the conformational space sampled in the simulations can (very crudely) be characterized by the largest RMSD between any pair of structures in the trajectory. We have tabulated this measure (RMST) in Table 5. The distance geometry ensemble samples a relatively small space, because of the fact that they are all derived from the same set of distances. It is interesting that the OPLS simulations, on average, seem to sample a smaller space than the simulations using the GROMOS force field, which could, at least in part, explain why the distance restraints were not reproduced as well.

3.5. Solvent-Accessible Surface Area. The solvent-accessible surface area (SASA) was computed using the double cube lattice method[64] with a probe radius of 0.14 nm. All nonpolar atoms in the protein were used to compute the hydrophobic SASA. In Figure 4, the results are plotted as a function of time. The proteins in the OPLS simulations behave in a more hydrophilic manner, because more hydrophobic surface area is exposed here than in the GROMOS simulations. The simulated proteins have a slightly larger SASA than that of the distance geometry

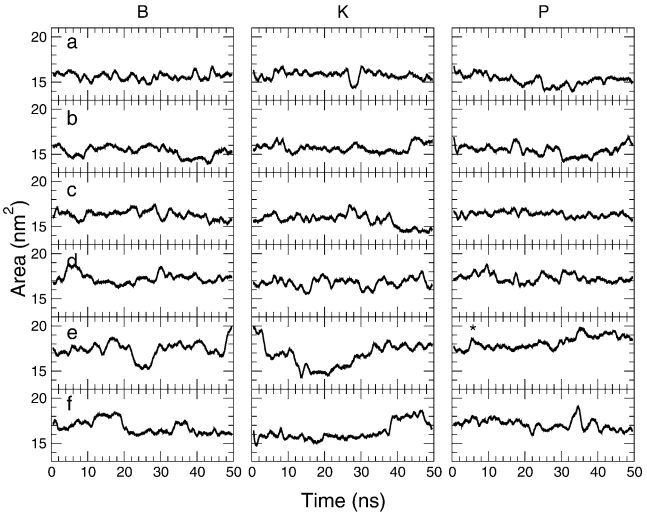

Figure 4. Hydrophobic solvent-accessible surface area, as a function of time in the simulations. (a) GS simulation, (b) GSD simulation, (c) GSDH simulation, (d) OS simulation, (e) OSH simulation (asterisk (*) indicates that starting structure A rather than P was used here), and (f) OT simulation. For clarity, a running average over 1 ns is plotted. See Table 1 for simulation names.

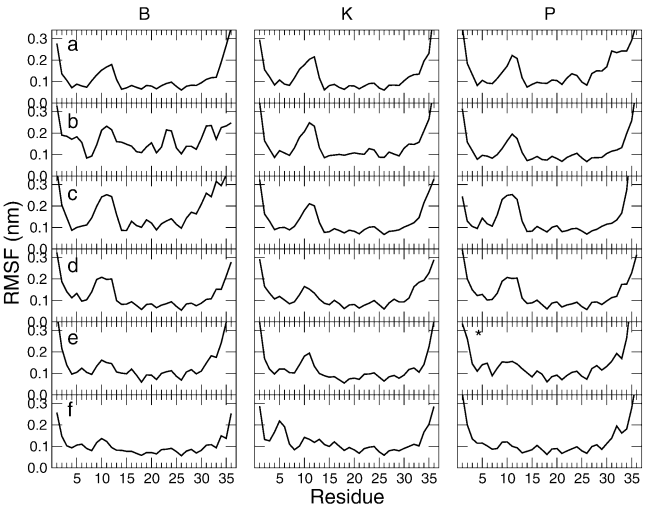

Figure 5. Root-mean-square fluctuations (RMSFs), as a function of residue, averaged over the trajectories, and over the NMR ensemble. (a) GS simulation, (b) GSD simulation, (c) GSDH simulation, (d) OS simulation, (e) OSH simulation (asterisk (*) indicates that starting structure A rather than P was used here), and (f) OT simulation. See Table 1 for simulation names.

ensemble (except simulation OS, which contracts noticably during the simulation). Average SASA values are given in Table 5.

3.6. Fluctations. The positional root-mean-square fluctuations (RMSFs) are computed as

$$\mathrm{RMSF}_i = \left(\frac{1}{N}\sum_i^N (\mathrm{x}_i - \langle \mathrm{x} \rangle)^2\right)^{1/2}$$

The Cα RMSF values are plotted as a function of residue in Figure 5. In almost all the simulations except OT (and partly OSH), there is a peak around Phe11, which is the center of a short loop between helix 1 and helix 2. As expected, the termini—in particular, the C-terminal residues—are very flexible, whereas the α-helical regions are more rigid. Large fluctuations in the GSDH simulations exist around this loop, although simulation GSDH/B (which refolds according to the RMSD plot) also exhibits high fluctuations at the C-terminus.

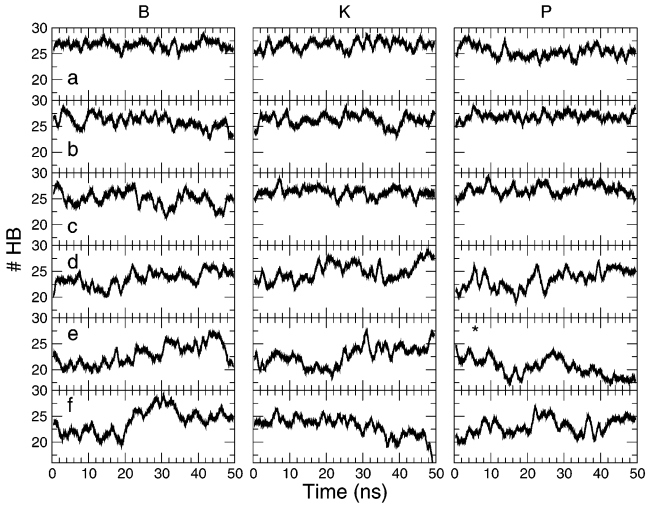

Figure 6. Protein−protein hydrogen bonds, as a function of time in the simulations. (a) GS simulation, (b) GSD simulation, (c) GSDH simulation, (d) OS simulation, (e) OSH simulation (asterisk (*) indicates that starting structure A rather than P was used here), and (f) OT simulation. See Table 1 for simulation names. For clarity, a running average over 1 ns is plotted.

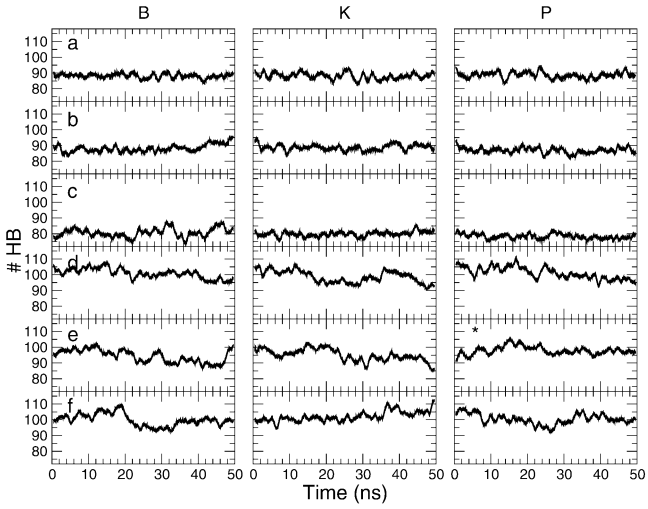

Figure 7. Protein−solvent hydrogen bonds, as a function of time in the simulations. (a) GS simulation, (b) GSD simulation, (c) GSDH simulation, (d) OS simulation, (e) OSH simulation (asterisk (*) indicates that starting structure A rather than P was used here), and (f) OT simulation. See Table 1 for simulation names. For clarity, a running average over 1 ns is plotted.

There is also a peak in one of the OT simulations near the start of the first α-helix, and quite large fluctuations are visible in simulation GSD/B, including the second α-helix and further, until the start of the third α-helix.

3.7. Hydrogen Bonding. Hydrogen bonds were analyzed using a geometrical criterium: a hydrogen bond was counted when the distance between a H atom and an acceptor was 0.25 nm or less and when the donor−hydrogen−acceptor angle was >120°. In Figure 6, the intramolecular hydrogen bonds in the simulations are plotted, whereas in Figure 7, we have plotted the protein−solvent hydrogen bonds. In all simulations (except OS), there are more intramolecular hydrogen bonds than are present in the distance geometry ensemble. It seems, however, that, in the OS simulation, the combination of the OPLS force field (which has greater partial charges than GROMOS96) with the SPC water yields too many protein−solvent hydrogen bonds. The balance is better in the OT simulation; in fact, the number of intramolecular hydrogen bonds in the OT simulation is very

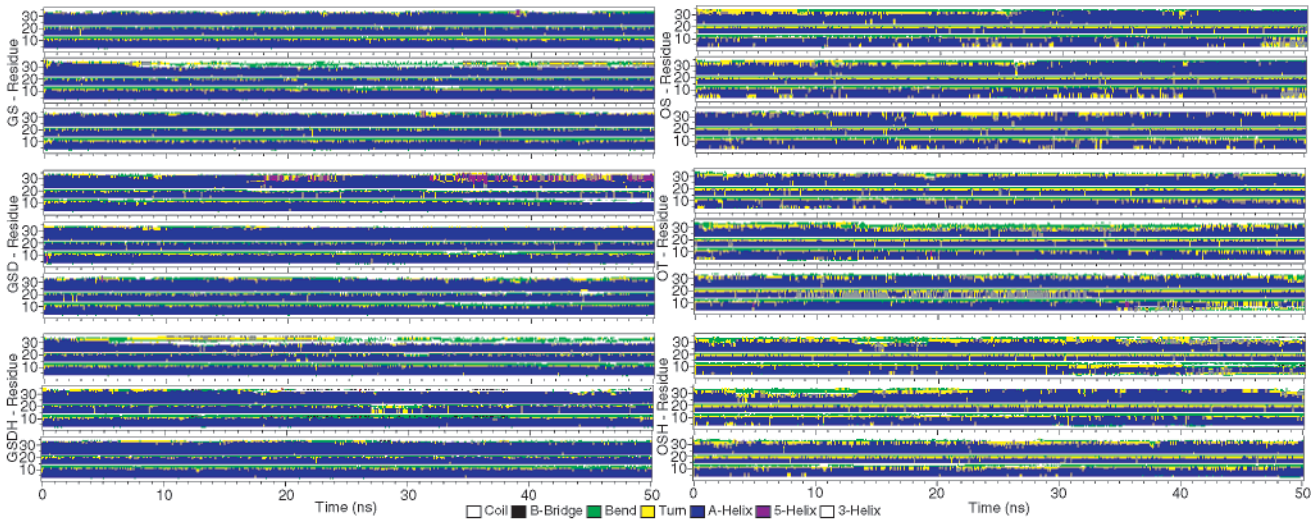

Figure 8. Secondary structure, as a function of time, in the different simulations. On the *Y*-axis, the simulation setup and residue number is indicated. For each setup, the top panel is the simulation starting from the PDB structure, the middle starting from conformation "P", and the bottom panel corresponds to conformation "K". See Table 1 for setup names. The difference in secondary structure confirms the differences in radius of gyration: the GROMOS96 structures are slightly more compact than the NMR ensemble for this protein, whereas OPLS is slightly looser.

similar to that in the GROMOS-based simulations. Interestingly, we see that the total number of hydrogen bonds with which the protein is engaged is almost the same for the GS and GSD simulations, and also for OS and OT (see Table 5). Therefore, we conclude that the protein force field is the most important determinant of hydrogen bonding. In both GSDH and OSH, we see that the number of hydrogen bonds is reduced when compared to GS(D) and OS, respectively. The neutralization of two charged groups clearly has significant impact on the hydrogen bonding properties of Villin.

3.8. Secondary Structure. The secondary structure of the protein conformations in the trajectories was computed using the DSSP program.[65] In Figure 8, we have plotted the secondary structure as a function of time for nine different simulations for each force field. In setup GSDH, the simulation starting from conformation P exhibit complete but reversible unrolling of the second helix (~30 ns). It is interesting that, despite the large increase in the RMSD for simulation GSDH/P, the secondary structure remains largely intact. The reversible nature of the secondary structure fluctuations in all the simulations indicate that the main-chain hydrogen bonding pattern is rather stable for both force fields. However, it is interesting that the larger radius of gyration for OPLS is clearly reflected in a more flexible secondary structure and slightly shorter helixes (not necessarily negative, because the GROMOS96 structures are more compact than the experimental NMR ensemble!).

4. Discussion

Proteins in solution are not static entities. Loops and solvent-exposed side chains fluctuate more or less freely. "Real proteins" (as opposed to peptides) are characterized by having a neatly packed hydrophobic core, i.e., an internal part that is not exposed to the solvent and, consequently, are not as flexible. A means to determine the stability of the core is to test which parts of the molecule are protected from solvent using H/D exchange.[66] This has been explicitly tested for the Villin headpiece. At least five amide protection factors were found to be $\gg$2000 in H/D exchange experiments,[49] which indicates that the protein adopts a fixed, unique conformation in solution. Furthermore, mutation studies have been performed on three phenylalanine (F) residues that form a cluster in the core of the protein to leucine (L).[67] The assumption in that work was that a real protein can withstand a mutation in the core without unfolding, and it was indeed found that all three single F $\rightarrow$ L mutants are stable and even the double mutant F47L, F51L is. Thus, it can be concluded that Villin can be regarded as a protein.[49,67] The minimum size for a peptide to become a protein is yet to be determined. Simulation work based on fragments from lactate dehydrogenase[68] showed that something as small as an α-helix, with an isolated β-strand folding back on the α-helix (23 residues), was reasonably stable. However, these simulations were rather short (2 ns), and no experimental evidence is available to prove the stability of that peptide.

There is considerable interest in mini proteins, among others, in the context of protein design but also for protein folding simulations and structure prediction.[69] The record for the smallest-size protein without disulfides is currently held by Trp, which is a stable 20-residue protein.[70] Brute-force folding studies[71] do suffer to some extent from the Levinthal paradox;[72] thus, it is good to have small targets in trying to optimize force fields and algorithms. However, it is not trivial to determine whether a computer simulation of a small protein is a good model of reality. Because of the inherent flexibility of small proteins, "reality" is a fuzzy target. Here, we have demonstrated that quantitative comparisons between experimental NMR data and simulation trajectories are possible, and that they provide a better comparison than RMSD for NMR structures.

We find that the structure of the hydrophobic core of Villin is reproduced well in some of the simulations. The sum of violations of 1.2 nm (GS, GSDH) is, in fact, even lower than the average over the distance geometry structures in the PDB file (1.7 nm). Although RMSD from a reference structure may be a good measure of simulation quality in the case of structures that have been determined by X-ray crystallography (and when the simulation is performed in the crystal[73,74]), it is probably not a good measure for solution structures of small proteins. Here, we find RMSD of the average structures to lie between 0.17 (OT/K) and 0.30 nm (GSDH/P), with no obvious correlation to the quality of the simulation (see Table 5). We therefore strongly prefer to make direct comparisons to experimental data such as distance restraints.[3,75]

We have also examined experimental observations other than NOE restraints. $^3J_{NH}$ coupling constants were computed through a Karplus relation.[55] Here, it should be noted that the accuracy

of the $^3J_{NH}$ calculation is limited, because the available parametrizations[56] are based only on the backbone angle ϕ. This may even adversely affect experimental structure determination, because there, dihedral angles are derived from coupling constants. In earlier simulations of random coil (GXG) peptides, experimental $^3J_{NH}$ coupling constants were reproduced to within 0.4 Hz for seven out of eight residues tested,[76] which indicated that it is possible to achieve good agreement for such rather-sensitive observables. For the Villin headpiece, we also obtain reasonable correlations between experiment and simulation (see Table 3), and once more, we find that the distance geometry ensemble, on average, has a lower correlation to the measured $^3J_{NH}$ coupling constant than the simulated ones. However, the problem remains that it is not possible to determine whether the deviation from experimental values arises from an incorrect trajectory due to the force field or from errors in the calculation of the $^3J_{NH}$ coupling constant.

Chemical shifts can be computed from structures in various ways, and this procedure has been used before as an aid in structure determination.[77−80] We have earlier used[46,76] the "total" program,[81] which is a variant of the shifts program[60] that has been used in this work. We also tried a statistical method that was developed by Wishart and Nip[80] to predict Hα (and other) shifts based on backbone torsion angles; however,[76] our results indicated that the total program, which (similar to the shifts program) is based on a physical model (both include, for instance, ring-current effects that are due to aromatic groups[61]) has a greater predictive power. In the comparison between simulation and data, we find here that the ^{1}Hα shift correlates well between simulation and experiment. Once again, we find that the simulation results are in better agreement with the data than the distance geometry results (see Table 4).

An important question follows from our comparison with NMR data: do the distance geometry structures form a good ensemble? It seems that we can conclude that the MD trajectories of Villin form a better ensemble than the distance geometry ensemble, based on the results given in Tables 2−4. For the distance restraint, we find that 11 out of 14 simulations that use the GROMOS96 force field[35] actually yield a lower sum of violations for the core restraints than the NMR ensemble. Although it might seem likely a priori that the use of an explicit solvent should help to improve the model, compared with the distance geometry ensemble, we cannot directly conclude that from our results. For the distance restraints in particular, we see that the restraints in the core of the protein are well reproduced but those which involve side chains (i.e., those in contact with the solvent) are not as good. This highlights the problem of water models[41,43] that are optimized for simulations of bulk water but may not be as suitable for the simulation of protein solvation.[46] Finally, we can conclude that the OPLS force field performs considerably better with the more-detailed four-site TIP4P model than with the SPC model.

Earlier simulations of Villin have focused on the folding pathway that starts from a completely unfolded peptide[2,82] and, similarly, that which starts from a structure prediction.[83] These simulations were performed using the AMBER force field[25,39] and were very successful in the sense that the native conformation was reproduced to within 0.3 nm RMSD[2] or even 0.21 nm.[83] A drawback[84] of the first folding simulation of Duan and Kollman[2] was that the Coulomb interactions were truncated at 0.8 nm. This was necessary to be able to perform the calculations on massively parallel computers and, hence, reach a trajectory length of 1 μs.[82] Interestingly, Lee et al. used the AMBER force field in combination with the PME algorithm,[20] rather than a simple truncation, and obtained lower RMSD values. Zagrovic et al. studied the folding of Villin using many short simulations with an implicit solvent model[85,86] and did also get low average RMSD values (0.17−0.19 nm). In this work, we have used explicit solvent molecules with the PME algorithm in all our simulations, even though neither GROMOS96 nor OPLS were optimized for use with PME (and not AMBER either, for that matter). In most simulations, we do find a RMSD that is comparable to that of Lee et al.[83] but slightly lower than that of Duan and Kollman.[2] It seems likely that this observation can, at least in part, be attributed to the electrostatics treatment, as suggested earlier by Åqvist.[38] Other simulations of Villin that should be mentioned are those by Sullivan and Kuntz,[87,88] which use Villin to study the free-energy landscape of a small protein, in the spirit of earlier work from the Berendsen group.[89,90]

In some of the simulations (GSD, GSDH, OTD), we used virtual particles to evaluate whether these produce results that are comparable to normal simulations. In the original implementation,[19] only short, 1-ns simulations were used for testing. Later work by the same authors[91] indicates that for very long (>100 ns) simulations of a nonapeptide, the trajectory averages of conformation and energies are similar to simulations without virtual particles. Table 2 shows that the simulations that use dummy particles yield results that are very similar to the normal simulation, which indicates that the use of dummy particles does not affect the quality of the simulation.

A final comment must be made about the protonation state of Villin in solution. At pH 3.7, solvent-exposed Glu residues will be protonated, unless there is a pK_a shift because of surrounding residues. In Villin, Glu-5 is embedded between two Asp residues (see Figure 1), so this is most likely protonated. Glu-32, on the other hand, is situated in a sequence Lys−Lys−Glu−Lys, and therefore it might well be charged even at pH 3.7. We have simulated Villin with either both residues charged, or with both residues neutral. Further simulations in which only Glu-5 was protonated are now underway. In previous simulations, it has been observed that pH (or, rather, protonation state) can influence the conformation.[92] This observation is consistent with our findings. Obviously, in the NMR experiments, there will also be a mixture of states; thus, it is important that pH be considered in structure determination and refinement that is based on NMR data.

Acknowledgment. The implementations of GROMOS96 and OPLS−AA/L force fields in GROMACS would not have been possible without the helpful assistance of Alan Mark, William Jorgensen, and Julian Tirado-Rives, who made their parameters freely available and patiently answered our numerous questions. Any remaining mistakes should be blamed on us.

References and Notes

(1) Berendsen, H. J. C. *Science* **1996**, *271*, 954−955.
(2) Duan, Y.; Kollman, P. A. *Science* **1998**, *282*, 740−744.
(3) Daura, X.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. *J. Mol. Biol.* **1998**, *280*, 925−932.
(4) Marrink, S. J.; Tieleman, D. P.; Mark, A. E. *J. Phys. Chem. B* **2000**, *104*, 12165−12173.
(5) Lindahl, E.; Edholm, O. *Biophys. J.* **2000**, *79*, 426−433.
(6) Marrink, S. J.; Lindahl, E.; Edholm, O.; Mark, A. E. *J. Am. Chem. Soc.* **2001**, *123*, 8638−8639.
(7) Tieleman, D.; Berendsen, H. *Biophys. J.* **1998**, *74*, 2786−2801.
(8) Izrailev, S.; Crofts, A. R.; Berry, E. A.; Schulten, K. *Biophys. J.* **1999**, *77*, 1753−1768.
(9) de Groot, B. L.; Grubmüller, H. *Science* **2001**, *294*, 2353−2357.
(10) Tajkhorshid, E.; Nollert, P.; Jensen, M. O.; Miercke, J. L. W.; O'Connell, J.; Stroud, R. M.; Schulten, K. *Science* **2002**, *296*, 525−530.
(11) Shirts, M. R.; Pande, V. S. *Phys. Rev. Lett.* **2001**, *86*, 4983−4986.
(12) See http://folding.stanford.edu for more information.

(13) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43−56.
(14) Lindahl, E.; Hess, B. A.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306−317.
(15) Bekker, H.; Berendsen, H. J. C.; Dijkstra, E. J.; Achterop, S.; van Drunen, R.; van der Spoel, D.; Sijbers, A.; Keegstra, H.; Reitsma, B.; Renardus, M. K. R. In *Physics Computing 92*; de Groot, R. A., Nadrchal, J., Eds.; World Scientific: Singapore, 1993; pp 257−261.
(16) Miyamoto, S.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 952−962.
(17) Barth, E.; Kuczera, K.; Leimkuhler, B.; Skeel, R. D. *J. Comput. Chem.* **1995**, *16*, 1192−1209.
(18) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463−1472.
(19) Feenstra, K. A.; Hess, B.; Berendsen, H. J. C. *J. Comput. Chem.* **1999**, *20*, 786−798.
(20) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(21) Essman, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577−8592.
(22) The GROMACS program is available free of charge under the GNU general public license from http://www.gromacs.org.
(23) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187−217.
(24) Hermans, J.; Berendsen, H. J. C.; van Gunsteren, W. F.; Postma, J. P. M. *Biopolymers* **1984**, *23*, 1513−1518.
(25) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. *J. Comput. Chem.* **1986**, *7*, 230−252.
(26) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657−1666.
(27) Hagler, A.; Huler, E.; Lifson, S. *J. Am. Chem. Soc.* **1974**, *96*, 5319−5327.
(28) Lifson, S.; Hagler, A.; Dauber, P. *J. Am. Chem. Soc.* **1974**, *101*, 5111−5121.
(29) Drenth, J. *Principles of Protein X-ray Crystallography*; Springer: New York, 1994.
(30) Doreleijers, J. F.; Rullmann, J. A. C.; Kaptein, R. *J. Mol. Biol.* **1998**, *281*, 149−164.
(31) Tolman, J. R. *Curr. Opin. Struct. Biol.* **2001**, *11*, 532−539.
(32) Sass, H. J.; Musco, G.; Stahl, S. J.; Wingfield, P. T.; Grzesiek, S. *J. Biomol. NMR* **2001**, *21*, 275−280.
(33) Spronk, C. A. E. M.; Linge, J. P.; Hilbers, C. W.; Vuister, G. W. *J. Biomol. NMR* **2002**, *22*, 281−289.
(34) McKnight, C. J.; Matsudaira, P. T.; Kim, P. S. *Nature Struct. Biol.* **1997**, *4*, 180−184.
(35) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular Simulation: The GROMOS96 Manual and User Guide*; Hochschuleverlag AG an der ETH Zürich: Zürich, Switzerland, 1996.
(36) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474−6487.
(37) Jorgensen, W. L. *Encyclopedia of Computational Chemistry*; Wiley: New York, 1998; Vol. 3, Chapter OPLS, Force Fields, pp 1986−1989.
(38) Åqvist, J. *FEBS Lett.* **1999**, *457*, 414−418.
(39) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.
(40) MacKerell, A. D.; Wiorkiewicz-Kuczera, J.; Karplus, M. *J. Am. Chem. Soc.* **1995**, *117*, 11946−11975.
(41) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.
(42) van Gunsteren, W. F.; Berendsen, H. J. C. *Gromos-87 Manual*; Biomos BV: Groningen, The Netherlands, 1987.
(43) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In *Intermolecular Forces*; Pullman, B., Ed.; D. Reidel Publishing Company: Dordrecht, The Netherlands, 1981; pp 331−342.
(44) van der Spoel, D.; van Maaren, P. J.; Berendsen, H. J. C. *J. Chem. Phys.* **1998**, *108*, 10220−10230.
(45) Mark, P.; Nilsson, L. *J. Phys. Chem. A* **2001**, *105*, 9954−9960.
(46) van der Spoel, D.; van Buuren, A. R.; Tieleman, D. P.; Berendsen, H. J. C. *J. Biomol. NMR* **1996**, *8*, 229−238.
(47) Mark, P.; Nilsson, L. *J. Phys. Chem. B* **2001**, *105*, 8028−8035.
(48) Elofsson, A.; Nilsson, L. *J. Mol. Biol.* **1993**, *233*, 766−780.
(49) McKnight, C. J.; Doering, D. S.; Matsudaira, P. T.; Kim, P. S. *J. Mol. Biol.* **1996**, *260*, 126−134.
(50) See http://people.bu.edu/cjmck/index.htm.
(51) Schulz, G. E.; Schirmer, R. H. *Principles of Protein Structure*; Springer−Verlag: New York, 1978.
(52) Berendsen, H. J. C.; Postma, J. P. M.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(53) Berendsen, H. J. C.; Mavri, J. *Int. J. Quantum Chem.* **1996**, *57*, 975−984.
(54) Torda, A. E.; Scheek, R. M.; van Gunsteren, W. F. *Chem. Phys. Lett.* **1989**, *157*, 289−294.
(55) Karplus, M. *J. Chem. Phys.* **1959**, *30*, 11−15.
(56) Vuister, G. W.; Bax, A. *J. Am. Chem. Soc.* **1993**, *115*, 7772−7777.
(57) Pardi, A.; Billeter, M.; Wüthrich, K. *J. Mol. Biol.* **1984**, *180*, 741−751.
(58) Smith, L. J.; Sutcliffe, M. J.; Redfield, C.; Dobson, C. M. *Biochemistry* **1991**, *30*, 986−996.
(59) Ludvigsen, S.; Andersen, K. V.; Poulsen, F. M. *J. Mol. Biol.* **1991**, *217*, 731−736.
(60) Ösapay, K.; Case, D. A. *J. Am. Chem. Soc.* **1991**, *113*, 9436−9444.
(61) Case, D. A. *J. Biomol. NMR* **1995**, *6*, 341−346.
(62) Dejaegere, A. P.; Bryce, R. A.; Case, D. A. In *Modeling NMR Chemical Shifts*; Facelli, J. A., de Dios, A. C., Eds.; American Chemical Society: Washington, DC, 1999; pp 194−206.
(63) Merutka, G.; Dyson, H. J.; Wright, P. E. *J. Biomol. NMR* **1995**, *5*, 14−24.
(64) Eisenhaber, F.; Lijnzaad, P.; Argos, P.; Sander, C.; Scharf, M. *J. Comput. Chem.* **1995**, *16*, 273−284.
(65) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577−2637.
(66) Baldwin, R. L. *Curr. Opin. Struct. Biol.* **1993**, *3*, 84−91.
(67) Frank, B. S.; Vardar, D.; Buckley, D. A.; McKnight, C. J. *Protein Sci.* **2002**, *11*, 680−687.
(68) van der Spoel, D.; Vogel, H. J.; Berendsen, H. J. C. *Proteins: Struct. Funct. Genet.* **1996**, *24*, 450−466.
(69) Gellman, S. H.; Woolfson, D. N. *Nature Struct. Biol.* **2002**, *9*, 408−410.
(70) Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H. *Nature Struct. Biol.* **2002**, *9*, 425−430.
(71) Feldman, H. J.; Hogue, C. W. V. *Proteins: Struct. Funct. Genet.* **2002**, *46*, 8−23.
(72) Levinthal, C. *J. Chim. Phys.* **1968**, *65*, 44−45.
(73) van Gunsteren, W. F.; Karplus, M. *Nature* **1981**, *293*, 677−678.
(74) Walser, R.; Hünenberger, P. H.; van Gunsteren, W. F. *Proteins* **2002**, *48*, 327−340.
(75) Colombo, G.; Roccatano, D.; Mark, A. E. *Proteins: Struct. Funct. Genet.* **2002**, *46*, 380−392.
(76) van der Spoel, D. *Biochem. Cell Biol.* **1998**, *76*, 164−170.
(77) Wishart, D. S.; Sykes, B. D. *Methods Enyzmol.* **1994**, *239*, 363−391.
(78) Case, D. A.; Dyson, H. J.; Wright, P. E. *Methods Enyzmol.* **1994**, *239*, 392−415.
(79) Ösapay, K.; Theriault, Y.; Wright, P. E.; Case, D. A. *J. Mol. Biol.* **1994**, *244*, 183−197.
(80) Wishart, D. S.; Nip, A. M. *Biochem. Cell Biol.* **1998**, *76*, 153−163.
(81) Williamson, M. P.; Asakura, T. *J. Magn. Reson. Ser. B* **1993**, *101*, 63−71.
(82) Duan, Y.; Wang, L.; Kollman, P. A. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 9897−9902.
(83) Lee, M. R.; Baker, D.; Kollman, P. A. *J. Am. Chem. Soc.* **2001**, *123*, 1040−1046.
(84) Berendsen, H. J. C. *Science* **1998**, *282*, 642−643.
(85) Zagrovic, B.; Snow, C. D.; Shirts, M. R.; Pande, V. S. *J. Mol. Biol.* **2002**, *323*, 927−937.
(86) Zagrovic, B.; Snow, C. D.; Khaliq, S.; Shirts, M. R.; Pande, V. S. *J. Mol. Biol.* **2002**, *323*, 153−164.
(87) Sullivan, D. C.; Kuntz, I. D. *Proteins: Struct. Funct. Genet.* **2001**, *42*, 495−511.
(88) Sullivan, D. C.; Kuntz, I. D. *J. Phys. Chem. B* **2002**, *106*, 3255−3262.
(89) de Groot, B. L.; Amadei, A.; Scheek, R. M.; van Nuland, N. A. J.; Berendsen, H. J. C. *Proteins: Struct. Funct. Genet.* **1996**, *26*, 314−322.
(90) Amadei, A.; de Groot, B. L.; Ceruso, M. A.; Paci, M.; Di Nola, A.; Berendsen, H. J. C. *Proteins: Struct. Funct. Genet.* **1999**, *35*, 283−292.
(91) Feenstra, K. A.; Peter, C.; Scheek, R. M.; van Gunsteren, W. F.; Mark, A. *J. Biomol. NMR* **2002**, *23*, 181−194.
(92) Smith, L. J.; Dobson, C. M.; van Gunsteren, W. F. *Proteins: Struct. Funct. Genet.* **1999**, *36*, 77−86.