Probing the Free Energy Landscape of the FBP28 WW Domain Using Multiple Techniques

XAVIER PERIOLE,[1] LUCY R. ALLEN,[2] KAMIL TAMIOLA,[1] ALAN E. MARK,[1,3] EMANUELE PACI[2,4]

[1] *Department of Biophysical Chemistry, Groningen Biomolecular Sciences and Biotechnology Institute (GBB), University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands*

[2] *Physics and Astronomy and Astbury Centre for Structural Molecular Biology, University of Leeds, Leeds LS2 9JT, United Kingdom*

[3] *School of Molecular and Microbial Sciences and the Institute of Molecular Biosciences, University of Queensland, St Lucia, QLD 4072, Australia*

[4] *Institute of Molecular and Cellular Biology, University of Leeds, Leeds LS2 9JT, United Kingdom*

Received 14 May 2008; Revised 28 July 2008; Accepted 22 August 2008
DOI 10.1002/jcc.21128
Published online 22 October 2008 in Wiley InterScience (www.interscience.wiley.com).

Abstract: The free-energy landscape of a small protein, the FBP 28 WW domain, has been explored using molecular dynamics (MD) simulations with alternative descriptions of the molecule. The molecular models used range from coarse-grained to all-atom with either an implicit or explicit treatment of the solvent. Sampling of conformation space was performed using both conventional and temperature-replica exchange MD simulations. Experimental chemical shifts and NOEs were used to validate the simulations, and experimental ϕ values both for validation and as restraints. This combination of different approaches has provided insight into the free energy landscape and barriers encountered by the protein during folding and enabled the characterization of native, denatured and transition states which are compatible with the available experimental data. All the molecular models used stabilize well defined native and denatured basins; however, the degree of agreement with the available experimental data varies. While the most detailed, explicit solvent model predicts the data reasonably accurately, it does not fold despite a simulation time 10 times that of the experimental folding time. The less detailed models performed poorly relative to the explicit solvent model: an implicit solvent model stabilizes a ground state which differs from the experimental native state, and a structure-based model underestimates the size of the barrier between the two states. The use of experimental ϕ values both as restraints, and to extract structures from unfolding simulations, result in conformations which, although not necessarily true transition states, appear to share the geometrical characteristics of transition state structures. In addition to characterizing the native, transition and denatured states of this particular system in this work, the advantages and limitations of using varying levels of representation are discussed.

© 2008 Wiley Periodicals, Inc.   J Comput Chem 30: 1059–1068, 2009

Key words: protein folding; free-energy landscape; transition state; molecular dynamics simulation; chemical shifts; NOEs

Introduction

Understanding the interplay between the forces that drive a protein to adopt its native fold is still an open challenge. Although it has been known for more than 40 years[1] that the amino acid sequence is sufficient to determine the native structure of a protein, it is still not possible to predict with certainty if, how fast or to what structure a given sequence will fold. While in principle computer simulations might allow complete analysis of folding and misfolding, in practice, the complexity and ruggedness of the free-energy surface mean that the timescales on which most proteins fold are too great to be accessed computationally. It is also unclear whether current force-fields can describe the relative free energies of the native and non-native states of a protein, and equally importantly the

Additional Supporting Information may be found in the online version of this article.

Correspondence to: X. Periole; e-mail: x.periole@rug.nl or E. Paci; e-mail: e.paci@leeds.ac.uk

Contract/grant sponsor: HPC-Europa translational access program

Contract/grant sponsor: European Community Training and Mobility of research Grant "Protein mis-folding"; contract/grant number: HPRN-CT-2002-00241

energy barriers and pathways between them with sufficient fidelity to reproduce folding appropriately.

The most reliable information concerning the mechanisms of protein folding is derived from experiment. While a number of novel techniques[2,3] are providing new insights into protein folding, the largest and most comprehensive set of data on the importance of particular residues in the thermodynamics of folding are so-called ϕ values.[4] Experimental ϕ values are related to changes in the rates of folding and unfolding associated with single-point mutations; ϕ values are commonly interpreted as the degree of nativeness of the environment of the mutated residue at the rate-limiting step (or transition state [TS]) in the folding pathway of the protein and have been widely used to benchmark (or "validate") unfolding simulations at high temperature.[5] Information contained in sets of experimental ϕ values has also been used to explore the ensemble of structures which could simultaneously satisfy a structural interpretation of these ϕ values.[6,7] Although the assumptions on which this approach is based have been questioned,[8] the method provides a picture of a remarkably heterogeneous transition state ensemble.

In this study, we have focused on the folding of the FBP28 WW domain. WW domains are small (typically 30–40 residues) ubiquitous proteins which are involved in protein recognition.[9,10] They form a slightly curved three-stranded antiparallel β-sheet and are fast two-state folders. The FBP28 WW domain is the fastest folder of the entire family, with a folding time of the order of 10 μs.[11] Both experimental and theoretical studies of this protein suggest that the rate limiting step in folding involves the formation of the loop between the β-strands 1 and 2.[12] The small size of the FBP28 WW domain, the availability of the three-dimensional structure,[13] of experimental ϕ values,[12] and the high degree of conservation of the fold upon mutation all make this protein particularly suitable for computational studies on folding.

Here we use a range of simulation techniques and molecular models to probe the free-energy landscape of this protein. Together the different techniques and models provide a detailed picture of native and denatured basins, and a transition state which are compatible with the available experimental data. The work has also allowed us to further our understanding of the available computational tools. Our results suggest that sampling is still the primary obstacle when using explicit solvent. All atom models simulated in explicit solvent are accurate enough to discriminate the native and non-native states of the protein; however, the barrier between the two is never crossed and therefore can not be characterised from equilibrium simulations. In contrast to the all atom model, the specific implicit solvent model used here appears to predict a native state which differs slightly but significantly from the experimental structure, and to stabilize denatured states more swollen and solvent exposed than the explicit representation of the solvent. Considering the continuous development of implicit solvent models,[14] this behaviour might not be general to implicit solvent models. A simple structure-based model provides a reasonable description of the native and denatured basins (comparable with that obtained using explicit solvent) but underestimates the free-energy barrier for folding. Experimental ϕ values, interpreted in terms of the native structure, suggest that the transition state is very compact, native like and lies between the native and denatured basins (in the RMSD/Q_{SC} projection) described by the explicit solvent model used in this study. In contrast, the transition state ensemble obtained when the experimental ϕ values are used as restraints in simulations is highly heterogeneous, suggesting that the results of such simulations must be interpreted with care.

Models and Methods

System

The solution structure of the FBP 28 WW domain has been determined by NMR spectroscopy[13] (PDB entry 1E0L). The first of the ensemble of NMR models deposited in the PDB was used as the reference structure. Simulations of both the full length and a truncated (residues 6–33) form of the protein were performed. It has been demonstrated experimentally that the stability of the protein is only marginally affected by the truncation of the N- and C-termini and that the truncated protein presents the characteristics of a two-state folder;[11] for this reason we will focus on the truncated form.

Models

Different models were used to represent the protein. The most detailed model is the united-atom GROMOS 43a1[15] force-field for the protein and the SPC[16] water model for the solvent. This will be referred to as the ES (explicit solvent) model. The system was solvated by ~3000 water molecules and equilibrated at 300 K and 1 atm. The box was sufficiently large (edge length $\geq$ 4 nm) to ensure that, even in the denatured state (with a radius of gyration of 1.01 ± 0.23 nm), the peptide was fully solvated and never interacted with its periodic images. Both the temperature and pressure were maintained close to their target values using the weak coupling algorithm[17] ($\tau_{\mathrm{T}} = 0.1$ ps and $\tau_{\mathrm{P}} = 1$ ps). A twin-range cutoff (1.0–1.4 nm) was used for the nonbonded interactions. Interactions within the short-range cutoff were evaluated every time step (2 fs), whereas interactions within the long-range cutoff were evaluated every 10 steps together with the neighbor-list. To correct for the truncation of electrostatic interactions beyond the long-range cutoff a reaction-field correction was applied ($\epsilon = 78$).[18] Bond lengths were constrained using the LINCS algorithm[19] for the protein and the SETTLE algorithm[20] for the water.

Simulations were also performed using the implicit solvent model EEF1, which is based on the united atom force field CHARMM19.[21] The model assumes that the solvation free energy of a protein is a sum of group contributions parameterized depending on their exposure to the solvent against a set of small model compounds. The ionic side-chains are neutralized, and a distance-dependent dielectric constant is used to approximate charge-charge interactions in solution. Simulations were performed with a 2 ps timestep with constraints applied to bonds involving hydrogen atoms. Constant temperature was maintained by using Langevin dynamics with a friction coefficient of 0.1 ps^{-1}. This will be referred to as the IS (implicit solvent) model.

The third model used was a "structure based" (SB) model in which each residue is represented by a single bead (at the Cα position) and the energy is defined by the native contacts present in the experimental structure.[13] Such models are also known as Go models and are computationally very efficient; the model proposed by Karanicolas and Brooks[22] was used. The distinguishing feature of this model compared with other Go models is that the magnitude and range of the interactions depend on the chemical properties of the residues and on the presence of hydrogen bonds in the native structure. A small sequence-dependent effect is also included via the dihedral term of the potential (see ref. 22 for details). Cα-Cα bonds were constrained and the integration timestep was 15 fs. Constant temperature was maintained by using Langevin dynamics with a friction coefficient of 0.1 ps^{-1}. This will be referred to as the SB (structure based) model.

Simulations

MD simulations of the protein, starting from the experimental model of the native structure, were performed at 300 K with the three models described above. T-REMD simulations were performed with the ES and IS models. In a T-REMD simulation,[23] a fixed number of independent simulations are carried out simultaneously at different temperatures ranging from T$_1$ to T$_M$, where M is the number of replicas. At regular time intervals, neighboring replicas i and j ($j = i \pm 1$) are allowed to exchange conformations according to the Metropolis criterion:

$$p = \min\{1, \exp[(\beta_i - \beta_j)(E_i - E_j)]\}, \qquad (1)$$

where p is the probability of replicas i and j exchanging, $\beta_i = 1/k_B T_i$, and E_i is the potential energy of replica i. By allowing conformations to explore temperature space, T-REMD enables the system to use thermal energy to cross energetic barriers and thus explore a larger conformation space than at low temperatures. The temperatures were chosen according to an exponential distribution and range from 282 to 497 K and 270 to 500 K for the ES and the IS models, respectively.

MD simulations were also performed using the IS and SB models with restraints applied so that the experimental ϕ values (ϕ^{exp}) were satisfied according to a structural interpretation of ϕ.[7] Such conformations do not necessarily correspond to the transition-state[8,24] of the underlying force-field; instead, they represent a structural model of the experimental transition-state under the assumptions implicit to the restraints. ϕ values were interpreted as the fraction of native contacts present in the TSE (see ref. 7 for details). For the SB model, a contact was defined as being present if two Cα atoms were within a distance of 0.85 nm and separated by at least two residues in the sequence. For the IS model, a contact was counted when two side-chain heavy atoms were at a distance less than 0.55 nm and separated by at least two residues in the sequence. For both models, contacts are defined as native if they are present in the energy minimized experimental structure. These definitions are consistent with the widely used protocol in which ϕ^{exp} are used as restraints.[7]

For the analysis of the trajectories, two alternative definitions for contacts between two residues were used. In the first definition (similar to when ϕ^{exp} are used as restraints with the SB model) a contact was considered to exist if the distance between two Cα atoms was less than 0.85 nm and the residues were separated by at least two positions in the sequence ($Q_{\mathrm{C}\alpha}$). The second definition was based on the center of mass of two side-chains: a contact was considered to exist if the centers of mass were closer than 0.65 nm and the residues separated by more than one position (Q_{SC}). The set of native contacts was defined using a 300 ns simulation of the native structure with ES model. Cα and side-chain (center of mass) contacts were considered native when present more than 75% and 70% of the time, respectively. A contact map calculated using the 10 NMR model structures was very similar to that defined using the simulation data, indicating that the simulation contact list provides a reliable basis for analysis.

Trajectories

Conventional long MD simulations at 300 K with the three models starting from the native structure have been performed and ran for 300, 400, and 1500 ns for the ES, IS, and SB models, respectively. T-REMD simulations were performed with the ES and IS models. Exchange trials were made every 1 and 2 ps with acceptance ratios of 0.1 and 0.2 in ES and IS REMD simulations, respectively, giving an average of 10 ps interval between temperature exchanges in both cases.

For both the ES and IS models, two sets of T-REMD simulations were performed, one starting from only folded structures and one from only unfolded structures. The NMR model equilibrated at the target temperature for 50 ps was taken as the initial folded conformation, whereas the unfolded conformations were extracted from a simulation at very high temperature (1000 K), with a minimum root mean square deviation (RMSD) of 0.8 nm from the native structure and each conformation separated by more than 1 ns along the trajectory. In this particular simulation, the dihedral angle describing the peptide-bond was modified to maintain the trans conformation. The 38 structures selected (see supplementary material Figure SM3) cover a large variety of highly unfolded, yet low energy conformations.

Simulations incorporating ϕ values as restraints were performed with both the SB and IS models. To maximize sampling, restrained simulations were performed using the T-REMD scheme; ϕ values were satisfied simultaneously for all residues and replicas.

All ES simulations using the GROMOS 43a1 force field (G43a1) were performed with the GROMACS simulation package[25] while the simulations using either the SB or IS models were performed with CHARMM version 34.

Chemical Shifts

Chemical shifts of backbone Hα and HN were predicted using the SHIFTX 1.0 software[26] with standard settings and all optimization routines switched-on. The predictions were made on conformations extracted from the simulations in which the hydrogen atoms were first removed and then regenerated according to the SHIFTX protocol.

Experimental values for the ^{1}H chemical shifts were obtained from Macias et al.[13] (and personal communication). After discarding those cases where the assignment was ambiguous there were 22 Hα and 27 HN shifts. The confidence level of the predictions was estimated from the IUPAC-referenced protein chemical shift database[27] as the standard deviation of the reported HN (42,674 entries) and Hα (37,733 entries) chemical shifts, 0.68 ppm and 0.50 ppm, respectively.

NOEs

NMR Nuclear Overhauser enhancement (NOE) intensities were also used to verify the simulations. The NOE intensity between

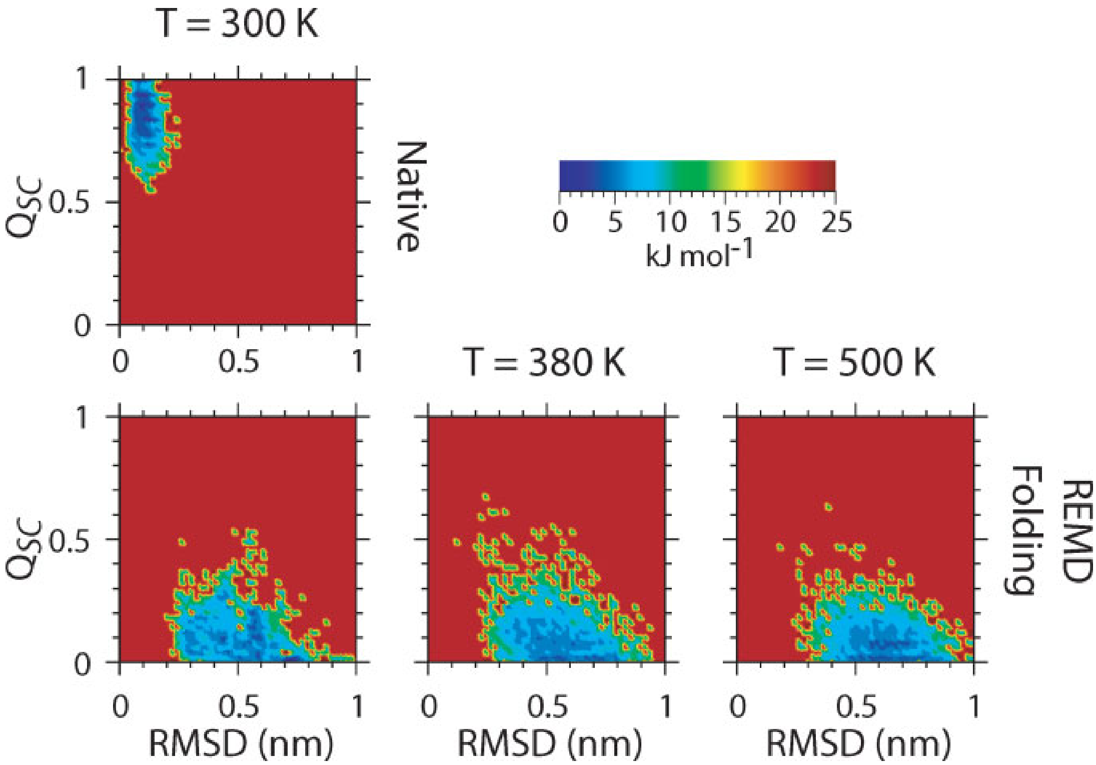

Figure 1. Projection of $-k_BT\ln P$ onto RMSD/Q_{SC} for the FBP28 WW domain simulated with the ES model. The top panel shows the native state surface explored with conventional MD simulation at 300 K. The bottom panels are obtained from T-REMD started from denatured conformations. The bins with a zero probability were assigned the arbitrary value of 25 kJ mol^{-1}.

two protons i and j (I_{ij}^{NOE}) is generally interpreted as being proportional to the inverse sixth power of the distance between them, $r_{ij}^{NOE} = (I_{ij}^{NOE})^{-1/6}$. r_{ij}^{NOE} is normally treated as an upper-bound distance.[28,29] For the truncated protein (residues 6–33), a set of 487 upper-bound distances were determined from the experimental set of I_{ij}^{NOE}. To compare the NOE derived distances to the simulations, distances between protons i and j in the simulation were averaged according to $r_{ij}^{sim} = \langle r_{ij}^6\rangle^{-1/6}$; where the average is over all the conformations present in the simulation.[30,31] Distances above 0.6 nm were ignored.

Violations of experimental upper-bound NOE distance, r_{ij}^{NOE}, were calculated as $v_{ij} = r_{ij}^{sim} - r_{ij}^{NOE}$. Averaged violations were calculated as $1/N\sum v_{ij}$, where the sum is over the N violated r_{ij}^{NOE} in the set. Since r_{ij}^{NOE} represent upper-bound distances, a violation is only counted when $r_{ij}^{sim} > r_{ij}^{NOE}$. The number of NOEs predicted by each simulation and the number of experimental NOEs matched by a simulation are also reported.

Results

Native Basin

Simulations of the full-length peptide (37 residues) performed at 300 K starting from the native structure using all the three force fields (ES, IS, SB) confirmed that truncation does not affect the stability of the peptide. In both the full and truncated simulations the region of the protein between residues 10 and 30 (4–24 in the truncated version) is structured whereas the remaining part of the sequence is flexible (data not shown). For this reason, we focus our study on the truncated version of the protein and the analysis reported relates to residues 4–24 if not otherwise indicated.

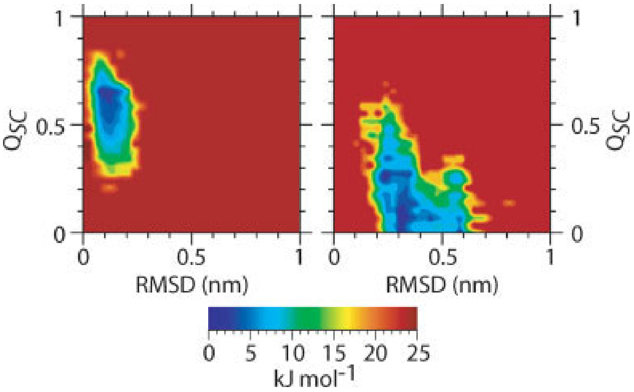

Figure 2. Projections of $-k_BT\ln P$ onto RMSD/Q_{SC} for the IS model. Left: native state from conventional MD at 300 K; right: denatured state from T-REMD started from unfolded structures, data collected at 300 K.

Using each of the three force-fields, conventional MD simulations at 300 K started from the native state explore a well defined region of conformation space. The average backbone RMSD from the experimental reference structure is relatively low for all three force-fields (0.13, 0.17, and 0.09 nm for ES, IS, and SB simulations, respectively), indicating that the main features of the peptide backbone found in the experimental structure are well conserved. Although in all three cases the native basins show approximately the same range of RMSDs values they differ significantly in terms of the fraction of native contacts. In both the ES and IS simulations a high proportion native side-chain contacts, Q_{SC}, are observed; however, the native contacts are much better preserved using the ES model (at least 60% always present) (Fig. 1) than using the IS model (between 30 and 80% present) (Fig. 2). The low Q_{SC} value suggests that, despite the RMSD from the experimental native structure being low, the IS model stabilizes a state in which the local contacts differ significantly from the experimental structure. Close inspection of the trajectory revealed that the twist which is evident in the experimental structure and ES model, is lost in the IS model. As the SB model does not contain side chains the simulations were interpreted in terms of Cα contacts ($Q_{\mathrm{C}\alpha}$). The range of $Q_{\mathrm{C}\alpha}$ values is very narrow, between 0.85 and 1, as expected using a SB model which is parameterized based on native contacts.

The calculated proton chemical shifts deviate from the experimental values for Hα and HN with a RMSD of 0.43 and 0.37 ppm, respectively, in the ES simulation, and 0.56 and 0.51 ppm in the IS simulation. While this suggests that the ES simulation reproduces the conformations sampled under the experimental conditions better than the IS simulation, the proton chemical shifts cover a very narrow range of values (3–6 ppm). The degree of agreement reported here is consistent with the uncertainty in the prediction of proton chemical shifts,[26] and our results are therefore not conclusive.

Violations of the NMR-derived upper-bound proton distances, $r_{\mathrm{ij}}^{\mathrm{NOE}}$, are reported in Table 1; 462, 452, and 456 out of the 487 experimental interproton distances were predicted by the ES, IS, and NMR models, respectively. The average violation of the $r_{\mathrm{ij}}^{\mathrm{NOE}}$ again indicates that the ES model more faithfully reproduces the experimental data than the IS model, with a level of agreement comparable with that of the NMR models.[13] The three sets of structures predict more NOEs than observed in the experiment (Table 1); this is a consequence of the use of r^{-6} averaging which heavily weights transiently sampled short distances and therefore predicts additional potential NOEs. This underlines the difficulty in interpretation of violations of experimental NOEs simply in terms of interproton distances.[32] Interestingly in this context, Zagrovic and van Gunsteren recently showed that even ensembles of unfolded structures could largely satisfy a set of experimental NOEs corresponding to the proposed native state.[31]

The deviation of the IS model from the experimental native state is further illustrated by the solvent accessible surface and the radius of gyration (R_{g}). The average values of the solvent accessible surface and R_{g} of the protein in the IS model simulation of the native state, 27.5 ± 0.8 nm^2 and 0.95 ± 0.01 nm, respectively, are significantly larger than the values obtained using the 10 NMR model structures, 25.9 ± 0.2 nm^2 and 0.91 ± 0.01 nm, respectively. In contrast the simulation with ES model agrees well with the NMR models, with values of 25.7 ± 0.7 nm^2 and 0.92 ± 0.02 nm, respectively.

Denatured Basin

The denatured basin was explored with all three force-fields. For the SB model reversible folding allowed the denatured state to be explored in a conventional simulation at the melting temperature T_{m}, which corresponds to a sharp peak in the specific heat (data not shown). For the ES and IS models T-REMD simulations were carried out starting from a set of thermally denatured conformations (see Models and Methods). Using the ES model, 38 replicas covering the temperature range of 282 to 497 K were employed. It has recently been suggested that in the case of a small β-peptide, as a result of increased conformational sampling and efficient sorting of multiple copies, T-REMD is about one order of magnitude computationally more efficient than a single continuous MD simulation.[33] Assuming that this factor holds for larger systems, the data presented here (38 × 305 ns) would be equivalent to ~100 μs, which is several times the expected folding time of the protein (~10 μs).[11]

The conformations sampled by the T-REMD simulation using the ES model were projected onto the coordinates RMSD/Q_{SC} as

Table 1. Violations of NMR-Derived Upper-Bound Distances for the ES, IS, and NMR Models.

	ES (1174 predicted NOEs)				IS (1118 predicted NOEs)				NMR models (1101 predicted NOEs)				ES/unfolded (8758 predicted NOEs)			
Set	Hits	$v < 0$	$\langle v \rangle$	(N)	Hits	$v < 0$	$\langle v \rangle$	(N)	Hits	$v < 0$	$\langle v \rangle$	(N)	Hits	$v < 0$	$\langle v \rangle$	(N)
all	462	247	0.069	(215)	452	244	0.081	(208)	456	237	0.068	(219)	391	252	0.084	(131)
Hα-Hα	3	3	–	(0)	3	3	–	(0)	3	3	–	(0)	3	1	0.180	(2)
HN-HN	12	4	0.047	(8)	12	1	0.073	(11)	12	9	0.035	(3)	11	4	0.101	(7)
Hα-HN	61	47	0.055	(14)	61	47	0.055	(14)	61	40	0.024	(21)	61	45	0.115	(16)
Hα-SC	92	58	0.073	(34)	89	58	0.095	(31)	90	53	0.082	(37)	77	60	0.078	(17)
HN-SC	113	54	0.053	(59)	113	49	0.063	(64)	112	39	0.060	(73)	113	70	0.050	(43)
SC-SC	181	81	0.080	(100)	174	86	0.094	(88)	178	93	0.081	(85)	126	72	0.097	(54)

The number of NOEs predicting a distance <0.42 nm (maximum reported in the experiment), the number of experimental NOEs predicted by a simulation (hits), the number of NOEs satisfied ($v < 0$), the averaged violation ($\langle v \rangle$), and the number of violations (N) are reported for each set of structures and grouped into different types of proton–proton interactions (set).

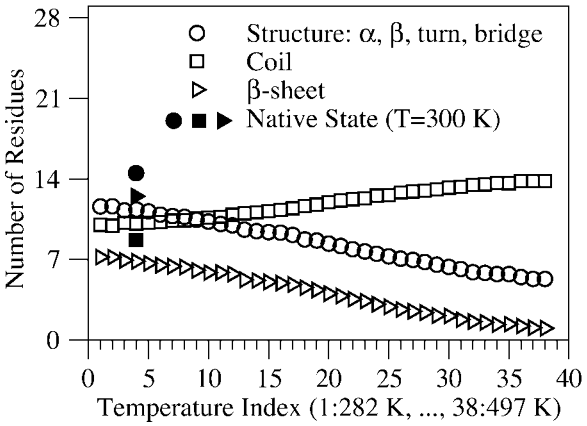

Figure 3. Secondary structure content observed in the denatured state of the FBP28 WW domain. The data was extracted from a REMD simulation (38 × 305 ns) started from denatured conformations.

shown in Figure 1. The native simulation at 300 K is shown for comparison. The projections (see also supplementary material Figure SM1) show that a few conformations with an RMSD as low as 0.14 nm and 65% of the side-chain native contacts are explored by the denatured simulations. Although these conformations mostly found at higher temperatures overlap with the edge of the native basin obtained at 300 K, they are not truly native; they are instead transient structures and do not represent a transition to a long-lived folded state. This highlights the danger of using projections or collective variables as reaction coordinates to identify folded states. Nevertheless, such projections do give a clear indication of the separation between the two basins, and allow us to identify and characterise the two states. The region dividing the two basins corresponds to a free-energy barrier which is not crossed on the timescale of these T-REMD simulations. Note that the use of R_g and the number of residues involved in a β-sheet structure as reaction coordinates does not differentiate the denatured and native states (data not shown). It is also remarkable that the denatured state for the ES model is more compact than the native state with an average R_g of 0.81 nm and solvent accessible surface of 25.01 nm^2.

The presence of configurations with low RMSD values (<0.3 nm) within the denatured basin suggests that native-like backbone conformations are present in the denatured state. As the simulation setup corresponds to conditions under which the FBP28 WW domain folds it is not surprising that the peptide approaches the native structure. However, the simulations suggest that the acquisition of a native-like backbone conformation is not sufficient for the peptide to fold.

To further characterise the denatured state explored in the REMD simulations the secondary structure content of the ES denatured basin was quantified using DSSP.[34] As expected, the secondary structure and particularly the β-sheet content decreases with temperature, whereas the amount of coil increases (Fig. 3). Notably, in the unfolded state at 300 K 25% of the residues form β-sheets compared with about 50% in the native state. On average only 2% of the residues are involved in α-helices in the unfolded state. This is consistent with growing evidence that residual secondary structure is often found in the denatured state of the protein even in denaturing conditions.[35,36]

As an additional analysis of the denatured state described by the ES model, the violations that this ensemble makes of NMR-derived upper-bound proton distances characterising the experimental native state, were calculated (Table 1). Unsurprisingly, the number of violations and their amplitudes are clearly higher in the ES denaturated state than in the experimental native states. However, a number of the NOEs predicted for the denatured ensemble match experimental ones, indicating that native contacts may be present in the denatured ensemble. The large number of predicted NOEs, also not present in the experimental data, strongly indicates the presence of numerous non-native contacts.

The results from analogous REMD simulations performed using the IS force-field are shown in Figure 2. Here, 16 replicas between 270 and 500 K were used, and the system simulated for 16×80 ns. As noted previously, the native basin is shifted to lower values of Q_{SC} compared with the ES model. The denatured basin, in contrast, while covering approximately the same range of Q_{SC} as the ES model, has a significant proportion of structures with Q_{SC} >40% (almost absent in ES) and a much narrower range of RMSD (rarely exceeding 0.6 nm). Conformations sampled within the denatured basin have Q_{SC} values characteristic of the native basin, but higher RMSD values. This result indicates that the IS force-field stabilizes a denatured state that is different from the ES model. The difference is further illustrated by the higher average R_g, 0.958 nm, and higher solvent accessible surface, 26.55 nm^2, than the ES model. Again the two basins approach each other very closely on this projection.

Figure 4 shows the free energy surface given by $-k_BT \ln P(\text{RMSD}, Q_{C\alpha})$ for both the SB model at equilibrium at T_m and for the ES model. Both models describe a narrow native basin and a large denatured ensemble in this projection. The SB model explores structures with RMSD values larger than 1 nm, which are never seen with the ES (and IS) models. Moreover, in the SB model, the native and unfolded states are not clearly separated: this could either be because the free-energy barrier is very small at T_{m}, or because the projection variables are not appropriate to separate the states using the SB model. To distinguish between these two possibilities, the free energy surface was recalculated using a method that does not involve a projection onto a particular set of geometrical criteria.[37] The method aims to group conformations into free energy minima based on equilibrium dynamics rather than geometrical criteria. The barrier measured using this method is very small, around 1 kJ mol^{-1}, and is in reasonable agreement with the free energy barrier in the RMSD/Q_{SC} projection. This indicates that the overlap of the two states is not due to the projection variables but to the small size of the energy barrier.

Transition State Ensembles

Because of the transient nature of folding transition states, detailed and reliable information regarding their structure is difficult to obtain experimentally. Experimental ϕ values provide information related to the effect of mutations on the thermodynamic properties of the transition state (TS) ensemble, but do not provide direct structural information on members of the TS ensemble. Simulations using detailed models could in principle be used to predict TS structures;

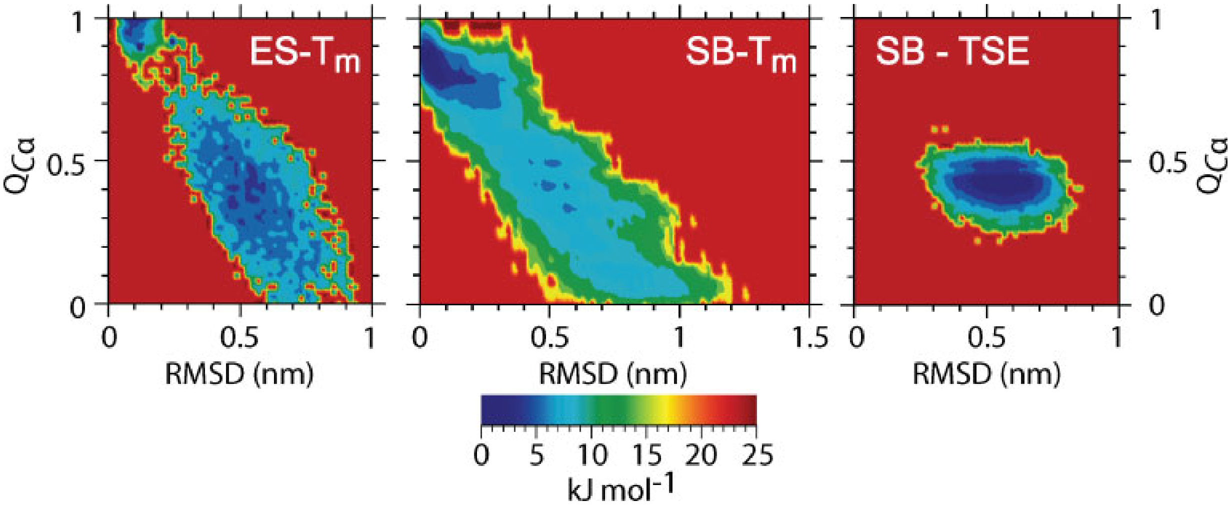

Figure 4. Projections of $-k_BT \ln P$ on RMSD/$Q_{C\alpha}$. The ES (left panel) and the SB results (middle panel) are shown. The ES surface was built from the native simulation at 300 K and the "folding" T-REMD simulation at 380 K, with each being given equal weighting. The SB model surface was obtained from an equilibrium simulation at 380 K. The right panel shows the projection of structures obtained using ϕ value restrained MD with the SB model.

however, in practice this is not possible with current computational resources. To circumvent this, several methods have been developed which aim to provide structural models of the experimental TS. Essentially a structural interpretation is proposed for the experimental ϕ values which are either used as restraints in simulations,[6,7] or to identify possible TS structures from high temperature unfolding simulations.[38] The results from both approaches have to be interpreted with caution.[8,39] For example Wang and Wade[40] have recently shown that the mechanism of unfolding of small peptides varies with the temperature. Here the results of the two approaches are compared and discussed in relation to the RMSD/Q_{SC} surface obtained with the ES model, shown in Figure 1, and the experimental ϕ values.

Figure 5 shows the results of IS simulations with experimental ϕ values as restraints. When projected on the RMSD/Q_{SC} surface, the ensemble of conformations satisfying the experimental ϕ values is broad and includes structures belonging to both the folded and unfolded basins as described by both the ES and IS models. However, the largest density of structures is found in a region which lies between the folded and unfolded basins in the ES model. Representative structures of the four most populated clusters in the TSE are shown in Figure 6. Strands 1 and 2 and loop 1 are native, reflecting the ϕ values which were used as restraints.

Simulations using the SB model with experimental ϕ values as restraints resulted in structures which lay firmly in the denatured basin as defined by both the ES and SB models (Fig. 4). This result is not entirely surprising as experimental ϕ values reflect the change in folding rate upon deletion of side-chain interactions, which are not included in the SB model.

We also used ϕ values to extract putative TS structures from a 100 ns T-REMD simulation using the ES model started from the native state. The conformations sampled in this simulation fall not only in the native and denatured basins but also populate the region between the two as unfolding events are observed (see supplementary material). To test to what degree the conformations sampled at 300 K probe the experimental transition state of the protein, we extracted structures characterised by calculated ϕ values close to the experimental ϕ values; i.e., structures for which $\rho = (1/N_\phi)\sum_i^{N_\phi}(\phi_i^{\text{calc}} - \phi_i^{\text{exp}})^2$ is less than 0.1, where N_ϕ is the number of ϕ values, ϕ_i^{calc} is the calculated value of ϕ for residue i (i.e., the fraction of native contacts formed by residue i) and ϕ_i^{exp} is the experimental value of ϕ for residue i. The structures found at 300 K are projected on the RMSD/Q_{SC} surface of the ES model (white circles in Fig. 5); remarkably, all are located in the region between the native and denatured basins. Representative structures

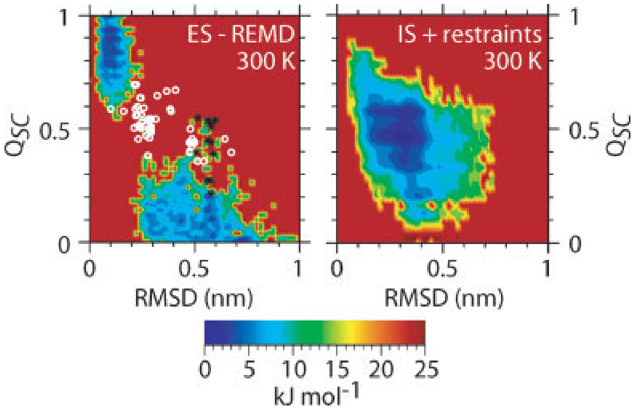

Figure 5. Projections of $-k_BT \ln P$ of TSEs on RMSD/Q_{SC}. Left panel: the native and denatured basins from ES simulations; white crosses and black stars represent those structures which satisfy $\rho < 0.1$ in REMD simulations starting from folded and unfolded conformations, respectively. Right panel: putative transition state structures calculated using MD restrained by ϕ^{exp} with the IS model at 300 K.

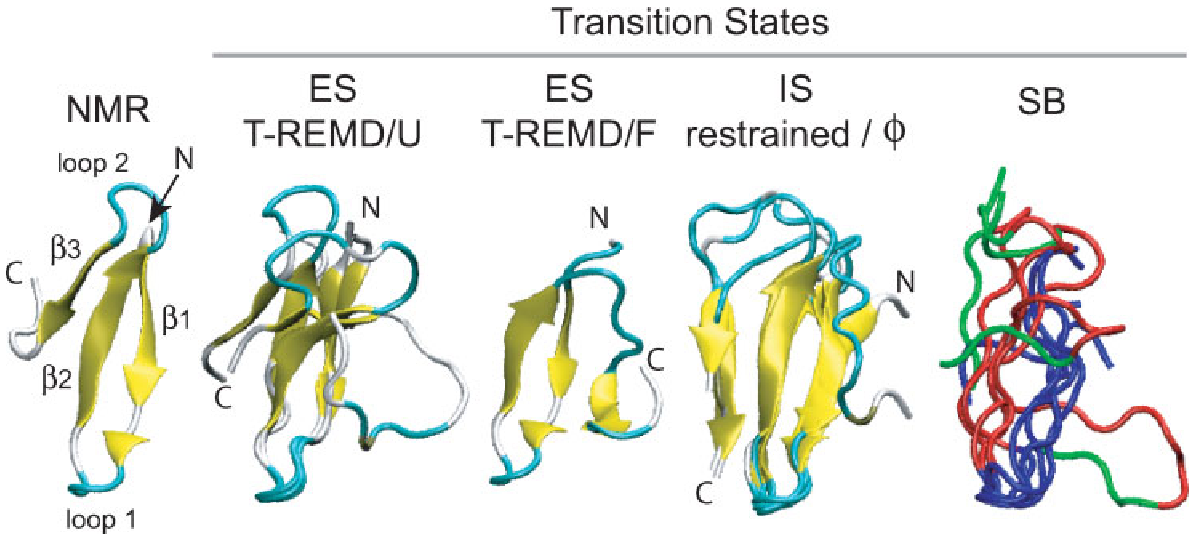

Figure 6. Structure of the transition states (TS). The representative structures of the three most populated clusters (cutoff 0.2 nm) with the lowest deviation from ϕ^{exp} found in T-REMD-unfolding (second from the left) and folding (third from the left) and restrained-IS ensembles (fourth from left) are shown in ribbon representation. Representative structure of the SB model TS are shown on the right. On the left the NMR structure is shown as a reference. [Color figure can be viewed in the online issue which is available at www.interscience.wiley.com.]

of the TSE obtained using this method are shown in Figure 6; again, strands 1 and 2 and loop 1 are native like, while strand 3 and loop 2 are mainly unstructured reflecting the ϕ values used in selecting the structures.

Interestingly, the conformations satisfying $\rho < 0.1$ but this time extracted from the T-REMD simulation started from unfolded conformations (black stars in Figure 5 at 300 K; see supplementary material Figure SM2 for a more complete range of T) overlap with a subset of conformations satisfying the same criterion from the T-REMD started from native structures. These structures (Fig. 6) also have strands 1 and 2, and the loop 1 native-like whereas loop 2 and strand 3 are clearly non-native. This suggests that the formation of strands 1 and 2 and loop 1 occurs prior to complete folding but does not necessarily lead to it.

The criterion used here ($\rho < 0.1$) is not very strict. The ϕ values deviate significantly from the experimental ϕ value by 0.3 on average. However, the individual ϕ values averaged over the ensemble of structures extracted from the unfolding simulations with $\rho < 0.1$ show (Fig. 7) that for most residues the difference between ϕ^{calc} and ϕ^{exp} lies within the uncertainty of ϕ^{calc}.

Finally, as reversible folding was observed using the SB model, in principle it should be possible to determine the TS for this model. However, the low free energy barrier means that the native and denatured basins are poorly separated at T_{m}. Using the RMSD/$Q_{\mathrm{C}\alpha}$ projections of the trajectories at 300 K and 450 K as indications of the positions of the native and denatured basins, structures in region between the two basins from the trajectory at T_{m} were isolated. Representative structures of the most populated clusters are shown in Figure 6. The structures correspond neither to the structures suggested by a qualitative analysis of the ϕ^{exp}, nor with a quantitative analysis based on the fraction of native contacts present in the structures. Although the SB model can represent the position and heterogeneity of the native and denatured basins reasonably well, it does not predict correctly the transition between the two basins.

Discussion

Using three force-fields with differing levels of approximation in conjunction with both conventional MD and T-REMD, we have analysed the energy landscape of the FBP28 WW domain. The use of a range of approaches has allowed the identification of features of the energy landscape that are model-independent. It has also allowed us to highlight the strengths and weaknesses of the alternative approaches. In addition experimental data has been used

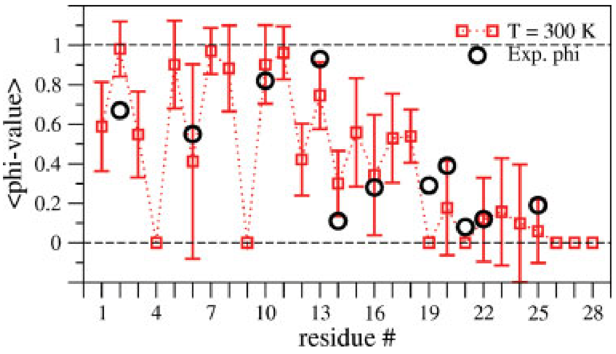

Figure 7. Individual ϕ values averaged over the ensemble of conformations satisfying $\rho < 0.1$ in the T-REMD-unfolding simulations. The ϕ^{exp} used to extract the TS are shown as black circles. [Color figure can be viewed in the online issue which is available at www.interscience.wiley.com.]

as an aid to simulation, both to assess the reliability of the different force-fields, and to extract putative TS structures. Our results suggest that using such a multi-method approach is not only valuable but also necessary; the use of only a single force-field (even the most accurate available) without validation from experimental data may provide accurate data on some areas of the energy landscape, but could be misleading in others.

The most detailed model used here, which combined the GROMOS96 43a1 force field and an explicit treatment of the solvent, appeared to describe the native state of the protein most accurately, with the available NMR data being well reproduced. The native and denatured basins were distinct and well-defined. The denatured state was more compact than the native basin and contained significant amounts of secondary structure, with many structures having native-like backbone conformations. This is in agreement with recent experimental data suggesting that the denatured state of proteins contains significant amounts of residual secondary structure.[35,36] Projection of the T-REMD trajectory onto the Q_{SC} and RMSD coordinates shows that, although the denatured basin closely approaches the native basin, no folding events occur. This is despite the fact that the total simulation time is equivalent to or even greater than the experimental folding time. This result is in line with previous simulation studies of the same protein.[41,42] One possible explanation for the lack of folding events is that the force-field used, which is primarily parametrized to reproduce low energy states of the system, overestimates the free energy barrier between the folded and unfolded states.

The implicit solvent model used in this work (IS), which describes the protein at atomic resolution but accounts for the effect of solvation via an additional term in the Hamiltonian, performed relatively poorly compared with the explicit solvent model. The experimentally determined native state was unstable with simulations started from this state quickly adopting a state which lacked the twist present in the experimental structure. Using the IS model the denatured basin was more extended and solvent exposed than that seen in the explicit solvent simulations. It should be noted that alternative IS models are being continuously developed,[14] so we have also probed more recent forcefields than the one used in this study. The native and denatured states of the protein were both simulated with the models GBSW[43] and ACE[44] and the results compared with those obtained with the EEF1 model. While the native contacts in the native state are slightly better preserved with ACE, the denatured state is in both cases narrow on the RMSD/Q_{SC} projection compared with ES simulations. A recent thorough analysis of the ability of several implicit solvent models to reproduce experimental kinetic and thermodynamic properties of disordered peptide states has shown that all implicit solvent models considered overestimate compactness, and EEF1 less so than any other model.[45]

A simple structure based (SB) model led to native and denatured basins which, at least in terms of the RMSD/$Q_{C\alpha}$ projection, agreed well with the more detailed models. However, the energy barrier between the native and denatured states was very small and the transition state ensemble predicted by the model did not agree with the available experimental data.

Experimental ϕ values were used together with MD simulations to predict possible transition state ensembles using two alternative approaches: (i) by restraining implicit solvent simulations with ϕ^{exp} and (ii) by isolating structures from explicit solvent unfolding simulations which satisfy the ϕ^{exp} at 300 K. In both cases, a structural interpretation of ϕ^{exp} was used. The first method gave a broad ensemble of structures, the majority of which fell in the region between the native and denatured basins as defined by the ES model in the RMSD/Q_{SC} projection. The second method gave TS structures which lay between the native and denatured states in the RMSD/Q_{SC} surface of the ES model. The second approach suggested that the TS lies between the native and denatured states in the RMSD/Q_{SC} projection and that the TSE obtained using ϕ^{exp} as restraints may be too broad. Remarkably, the conformations that satisfied ϕ^{exp} in the denatured state trajectory resulted in a number of structures at the edge of the RMSD/Q_{SC} transition region, overlapping with structures extracted from the unfolding simulation. Since those structures did not lead to the native state, this indicates that structures satisfying the ϕ^{exp} are not necessarily true transition states.

Both methods suggested that in the transition state strands 1 and 2, and the loop between them (loop 1) are native-like, while loop 2 and strand 3 do not show a consistent structure. This is in global agreement with the experimental ϕ values obtained for the full length version of this protein and with putative transition state structures extracted from high temperature unfolding simulations.[12] However, our results suggest the presence of contacts between strands 2 and 3 and that the TS structures have a low RMSD with respect to the native state.

Each of the methods used to determine putative transition state ensembles have potential artefacts. In particular, all rely on a structural interpretation of ϕ^{exp} and there is no objective means to test if the structures suggested actually correspond to true transition states.[8,46] Nevertheless, isolating structures from unfolding T-REMD simulations which satisfy the ϕ^{exp} appears to give the most transition-state-like conformations, in that the structures lie in the same region of the RMSD/Q_{SC} surface as when using the IS model with ϕ values as restraints. However, this method does not give any information on the heterogeneity of the TSE, which might be relevant for some proteins,[47,48] and is dependent on the validity of a structural interpretation of ϕ^{exp}.

In conclusion, the free energy landscape of the FBP28 WW domain was explored using alternative models for the protein and the solvent, and alternative methods to sample the conformation space. The native and denatured basins occupy two distinguishable and well-defined regions of the RMSD/Q_{SC} surface. The native state covers a narrow range of RMSD values (<0.2 nm from the experimental structure) and has more than 90% and 60% of the Cα and side-chain native contacts, respectively. The denatured basin is compact and contains a significant amount of residual secondary structure. Conformations having an RMSD value as low as 0.2–0.3 nm from the native structure and up to 70–80% of the native Cα contacts were found within the denatured basin, however these structures had rarely more than 40% of the side-chain native contacts. The use of experimental ϕ values to extract putative transition states (TS) of the protein from the unfolding simulations, defined a TS ensemble at 300 K which lies between the native and denatured basins on the same RMSD/Q_{SC} surface. The TS structures were very compact with low RMSD values from the native structure. Finally, whereas the use of experimental ϕ as restraints in simulations resulted in a TS ensemble which was possibly too broad, the majority of the structures were also found between the native and denatured basins. Both approaches predicted that strands 1 and 2, and loop 1 are native-like in the TS.

Acknowledgments

We thank Maria Macias for providing the values for the ^{1}H chemical shifts and the NOE intensities. LRA acknowledges EPSRC for a PhD studentship. The use of the supercomputing facilities of the NWO SARA and the University of Groningen is gratefully acknowledged.

References

1. Anfinsen, C. B. Science 1973, 181, 223.
2. Cecconi, C.; Shank, E. A.; Bustamante, C.; Marqusee, S. Science 2005, 309, 2057.
3. Nettels, D.; Gopich, I. V.; Hoffmann, A.; Schuler, B. Proc Natl Acad Sci USA 2007, 104, 2655.
4. Fersht, A. R. Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding; W. H. Freeman & Co.: New York, 1999.
5. Li, A.; Daggett, V. Proc Natl Acad Sci USA 1994, 91, 10430.
6. Vendruscolo, M.; Paci, E.; Dobson, C. M.; Karplus, M. Nature 2001, 409, 641.
7. Paci, E.; Vendruscolo, M.; Dobson, C. M.; Karplus, M. J Mol Biol 2002, 324, 151.
8. Allen, L. R.; Paci, E. J Phys: Condensed Matter 2007, 19, 285211.
9. Bork, P.; Sudol, M. Trends Biochem Sci 1994, 19, 531.
10. Macias, M. J.; Wiesner, S.; Sudol, M. FEBS Lett 2002, 513, 30.
11. Nguyen, H.; Jager, M.; Moretto, A.; Gruebele, M.; Kelly, J. W. Proc Natl Acad Sci USA 2003, 100, 3948.
12. Petrovich, M.; Jonsson, A. L.; Ferguson, N.; Daggett, V.; Fersht, A. R. J Mol Biol 2006, 360, 865.
13. Macias, M. J.; Gervais, V.; Civera, C.; Oschkinat, H. Nature Struct Biol 2000, 7, 375.
14. Chen, J.; Brooks, C. L.; Khandogin, J. Curr Opin Struct Biol 2008, 18, 140.
15. 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 Vdf. Hochschulverlag AG an der ETH Zürich, Zürich, Switzerland, 1996.
16. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In B. Pullman, Ed.; Intermolecular Forces, Reider: Dordrecht, Germany, 1981; pp. 331–342.
17. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. J Chem Phys 1984, 81, 3684.
18. Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. J Comp Chem 1995, 102, 5451.
19. Hess, B.; Bekker, H.; Berendsen, H. J. C. J Comp Chem 1997, 18, 1463.
20. Miyamoto, S.; Kollman, P. A. J Comp Chem 1992, 13, 952.
21. Lazaridis, T.; Karplus, M. Curr Opin Struct Biol 2000, 10, 139.
22. Karanicolas, J.; Brooks, C. L. I. Proc Natl Acad Sci USA 2003, 100, 3954.
23. Sugita, Y.; Okamoto, Y. Chem Phys Lett 1999, 314, 141.
24. Hubner, I. A.; Shimada, J.; Shakhnovich, E. I. J Mol Biol 2004, 336, 745.
25. van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. J Comp Chem 2005, 26, 1701.
26. Neal, S.; Nip, A. M.; Zhang, H.; Wishart, D. S. J Biomol NMR 2003, 26, 215.
27. Neal, S.; Wishart, D. S. J Biomol NMR 2003, 25, 173.
28. Wuthrich, K. NMR of Proteins and Nucleic Acids; Wiley: New York, 1986.
29. Roberts, G. C. NMR of Macromolecules: A Practical Approach; IRL Press: Oxford, UK, 1993.
30. Tropp, J. J Chem Phys 1980, 72, 6035.
31. Zagrovic, B.; van Gunsteren, W. F. Proteins 2006, 63, 210.
32. Feenstra, K. A.; Peter, C.; Scheek, R. M.; van Gunsteren, W. F.; Mark, A. E. J Biomol NMR 2002, 23, 181.
33. Periole, X.; Mark, A. E. J Chem Phys 2007, 126, 014903-1.
34. Kabsch, W.; Sander, C. Biopolymers 1983, 22, 2577.
35. Shortle, D. R.; Ackerman, M. S. Science 2001, 293, 487.
36. Matsuo, K.; Sakurada, Y.; Yonehara, R.; Kataoka, M.; Gekko, K. Biophys J 2007, 92, 4088.
37. Krivov, S. V.; Karplus, M. J Phys Chem B 2006, 110, 12689.
38. Daggett, V.; Li, A.; Itzhaki, L. S.; Otzen, D. E.; Fersht, A. R. J Mol Biol 1996, 257, 430.
39. Taskent, H.; Cho, J.-H.; Raleigh, D. P. J Mol Biol 2008, 378, 699.
40. Wang, T.; Wade, R. C. J Chem Theory and Comp 2007, 3, 1476.
41. Mu, Y.; Nordenskiold, L.; Tam, J. P. Biophys J 2006, 90, 3983.
42. Freddolino, P. L.; Liu, F.; Gruebele, M. H.; Schulten, K. Biophys J 2008, 94, L75.
43. Im, W.; Lee, M. S.; Brooks, C. L. J Comp Chem 2003, 24, 1691.
44. Schaefer, M.; Karplus, M. J Phys Chem 1996, 100, 1578.
45. Feige, M. J.; Paci, E. J Mol Biol 2008, 382, 556.
46. Periole, X.; Vendruscolo, M.; Mark, A. E. Proteins 2007, 69, 536.
47. Paci, E.; Friel, C. T.; Lindorff-Larsen, K.; Radford, S. E.; Karplus, M.; Vendruscolo, M. Proteins 2004, 54, 513.
48. Morton, V. L.; Friel, C. T.; Allen, L. R.; Paci, E.; Radford, S. E. J Mol Biol 2007, 371, 554.