32
Molecular Dynamics Simulations to Study Protein Folding and Unfolding

Amedeo Caflisch and Emanuele Paci

32.1
Introduction

Proteins in solution fold in time scales ranging from microseconds to seconds. A computational approach to folding that should work, in principle, is to use an atom-based model for the potential energy (force field) and to solve the time-discretized Newton equation of motion (molecular dynamics, MD [1]) from a denatured conformer to the native state in the presence of the appropriate solvent. With the available simulation protocols and computing power, such a trajectory would require approximately 10–100 years for a 100-residue protein where the experimental transition to the folded state takes place in about 1 ms. Hence, there is a clear problem related to time scales and sampling (statistical error). On the other hand, we think that current force fields, even in their most detailed and sophisticated versions, i.e., explicit water and accurate treatment of long-range electrostatic effects, are not accurate enough (systematic error) to be able to fold a protein on a computer. In other words, even if one could use a computer 100 times faster than the currently fastest processor to eliminate the time scale problem, most proteins would not fold to the native structure because of the large systematic error and the marginal stability of the folded state typically ranging from 5 to 15 kcal mol^{-1}. Interestingly, only designed peptides of about 20 residues have been folded by MD simulations (see Section 32.2.1) using mainly approximative models of the solvent (see Section 32.3.4). Alternatively, protein unfolding which is a simpler process than folding (e.g., the unfolding rate shows Arrhenius-like temperature dependence whereas folding does not because of the importance of entropy, see Section 32.2.1.2) can be simulated on shorter time scales (1–100 ns) at high temperature or by using a suitable perturbation.

MD simulations can provide the ultimate detail concerning individual atom motion as a function of time. Hence, future improvements in force fields and simulation protocols will allow specific questions about the folding of proteins to be addressed. The understanding at the atomic level of detail is important for a complicated reaction like protein folding and cannot easily be obtained by experiments. Yet, experimental approaches and results are essential in validating the force fields

Protein Folding Handbook. Part I. Edited by J. Buchner and T. Kiefhaber

ISBN: 3-527-30784-2

and simulation methods: comparison between simulation and experimental data is *conditio sine qua non* to validate the simulation results and very helpful for improving force fields.

This chapter cannot be comprehensive. Results obtained by using atom-based force fields and MD are presented whereas lattice models [2] as well as off-lattice coarse-grained models (e.g., one interaction center per residue) [3] are not mentioned because of size limitations. It is important to note that the impact of MD simulations of folding and unfolding is increasing thanks to faster computers, more efficient sampling techniques, and more accurate force fields as witnessed by several review articles [1, 4] and books [5–7].

32.2 Molecular Dynamics Simulations of Peptides and Proteins

32.2.1 Folding of Structured Peptides

Several comprehensive review articles on MD simulations of structured peptides have appeared recently [8–10]. Here, we first focus on simulation results obtained in our research group and then discuss the Trp-cage, a model system that has been investigated by others.

32.2.1.1 Reversible Folding and Free Energy Surfaces

***β*-Sheets** The reversible folding of two designed 20-residue sequences, beta3s and ^{D}PG, having the same three-stranded antiparallel β-sheet topology was simulated [11, 12] with an implicit model of the solvent based on the accessible surface area [13]. The solution conformation of beta3s (TWIQNGSTKWYQNGSTKIYT) has been studied by NMR [14]. Nuclear Overhauser enhancement spectroscopy (NOE) and chemical shift data indicate that at 10 °C beta3s populates a single structured form, the expected three-stranded antiparallel β-sheet conformation with turns at Gly6-Ser7 and Gly14-Ser15, (Figure 32.1) in equilibrium with the denatured state. The β-sheet population is 13–31% based on NOE intensities and 30–55% based on the chemical shift data [14]. Furthermore, beta3s was shown to be monomeric in aqueous solution by equilibrium sedimentation and NMR dilution experiments [14].

^{D}PG is a designed amino acid sequence (Ace-VFITS^{D}PGKTYTEV^{D}PG-Orn-KILQ-NH), where ^{D}P are D-prolines and Orn stands for ornithine. Circular dichroism and chemical shift data have provided evidence that ^{D}PG adopts the expected three-stranded antiparallel β-sheet conformation at 24 °C in aqueous solution [15]. Moreover, ^{D}PG was shown to be monomeric by equilibrium sedimentation. Although the percentage of β-sheet population was not estimated, NOE distance restraints indicate that both hairpins are highly populated at 24 °C.

In the MD simulations at 300 K (started from conformations obtained by spon-

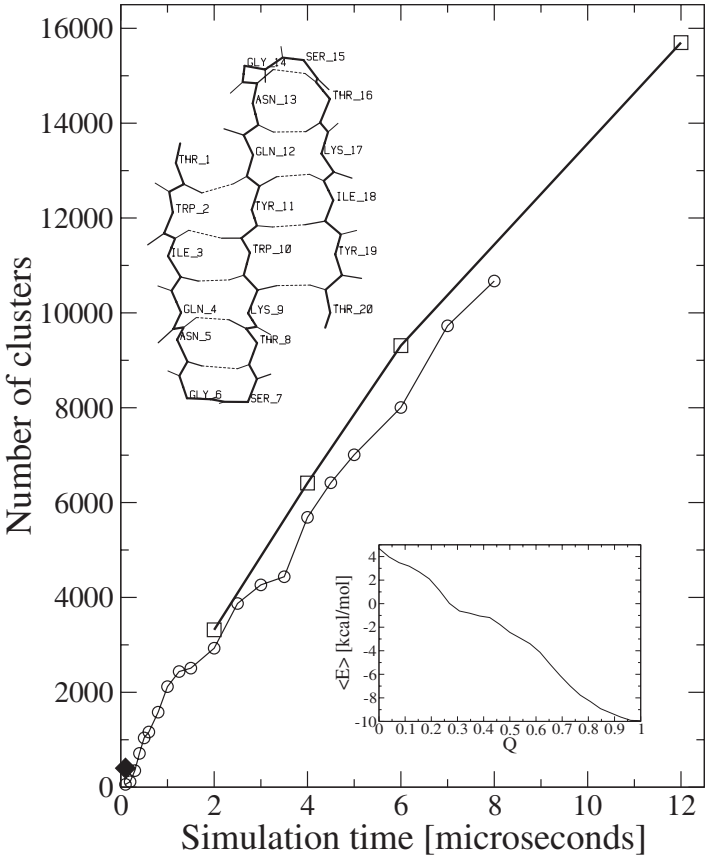

Fig. 32.1. Number of clusters as a function of time. The ''leader'' clustering procedure was used with a total of 120 000 snapshots saved every 0.1 ns (thick line and square symbols). The clustering algorithm which uses the Cα RMSD values between all pairs of structures was used only for the first 8 μs (80 000 snapshots) because of the computational requirements (thin line and circles). The diamond in the bottom left corner shows the average number of conformers sampled during the folding time which is defined as the average time interval between successive unfolding and refolding events. The insets show a backbone representation of the folded state of beta3s with main chain hydrogen bonds in dashed lines, and the average effective energy as a function of the fraction of native contacts Q which are defined in [11]. Figure from Ref. [33].

taneous folding at 360 K) both peptides satisfy most of the NOE distance restraints (3/26 and 4/44 upper distance violations for beta3s and ^{D}PG, respectively). At a temperature value of 360 K which is above the melting temperature of the model (330 K), a statistically significant sampling of the conformational space was obtained by means of around 50 folding and unfolding events for each peptide [11, 12]. Average effective energy and free energy landscape are similar for both peptides, despite the sequence dissimilarity. Since the average effective energy has a downhill profile at the melting temperature and above it, the free energy barriers

are a consequence of the entropic loss involved in the formation of a β-hairpin which represents two-thirds of the chain. The free energy surface of the β-sheet peptides is completely different from the one of a helical peptide of 31 residues, $Y(MEARA)_6$ (see below). For the helical peptide, the folding free energy barrier corresponds to the helix nucleation step, and is much closer to the fully unfolded state than for the β-sheet peptides. This indicates that the native topology determines to a large extent the free energy surface and folding mechanism. On the other hand, the ^{D}PG peptide has a statistically predominant folding pathway with a sequence of events which is the inverse of the one of the most frequent pathway for the beta3s peptide. Hence, the amino acid sequence and specific interactions between different side chains determine the most probable folding route [12].

It is interesting to compare with experimental results on two-state proteins. Despite a sequence identity of only 15%, the 57-residue IgG-binding domains of protein G and protein L have the same native topology. Their folded state is symmetric and consists mainly of two β-hairpins connected such that the resulting four-stranded β-sheet is antiparallel apart from the two central strands which are parallel [16]. The ϕ value analysis (see Section 32.2.3 for a definition of ϕ value) of protein L and protein G indicates that for proteins with symmetric native structure more than one folding pathway may be consistent with the native state topology and the selected route depends on the sequence [16]. Our MD simulation results for the two antiparallel three-stranded β-sheet peptides (whose sequence identity is also 15%) go beyond the experimental findings for protein G and L. The MD trajectories demonstrate the existence of more than one folding pathway for each peptide sequence [12]. Interestingly, Jane Clarke and collaborators [17] have recently provided experimental evidence for two different unfolding pathways using the anomalous kinetic behavior of the 27th immunoglobulin domain (β-sandwich) of the human cardiac muscle protein titin. They have interpreted the upward curvature in the denaturant-dependent unfolding kinetics as due to changes in the flux between transition states on parallel pathways. In the conclusion of their article [17] they leave open the question ''whether what is unusual is not the existence of parallel pathways, but the fact that they can be experimentally detected and resolved.''

α-Helices Richardson et al. [18] have analyzed the structure and stability of the synthetic peptide $Y(MEARA)_6$ by circular dichroism (CD) and differential scanning calorimetry (DSC). This repetitive sequence was ''extracted'' from a 60-amino-acid domain of the human CstF-64 polyadenylation factor which contains 12 nearly identical repeats of the consensus motif MEAR(A/G). The CD and DSC data were insensitive to concentration indicating that $Y(MEARA)_6$ is monomeric in solution at concentrations up to 2 mM. The far-UV CD spectrum indicates that the peptide has a helical content of about 65% at 1 °C. The DSC profiles were used to determine an enthalpy difference for helix formation of 0.8 kcal mol^{-1} per amino acid. The length of $Y(MEARA)_6$ makes it difficult to study helix formation by MD simulations with explicit water molecules. Therefore, multiple MD runs were performed with the same implicit solvation model used for the β-sheet peptides [13].

The simulation results indicate that the synthetic peptide $Y(MEARA)_6$ assumes a mainly α-helical structure with a nonnegligible content of π-helix [149]. This is not inconsistent with the currently available experimental evidence [18]. A significant π-helical content was found previously by explicit solvent molecular dynamics simulations of the peptides $(AAQAA)_3$ and $(AAKAA)_3$ [19], which provides further evidence that the π-helical content of $Y(MEARA)_6$ is not an artifact of the approximations inherent to the solvation model.

An exponential decay of the unfolded population is common to both $Y(MEARA)_6$ [149] and the 20-residue three-stranded antiparallel β-sheet [14] previously investigated by MD at the same temperature (360 K) [11]. The free energy surfaces of $Y(MEARA)_6$ and the antiparallel β-sheet peptide differ mainly in the height and location of the folding barrier, which in $Y(MEARA)_6$ is much lower and closer to the fully unfolded state. The main difference between the two types of secondary structure formation consists of the presence of multiple pathways in the α-helix and only two predominant pathways in the three-stranded β-sheet. The helix can nucleate everywhere, with a preference for the C-terminal third of the sequence in $Y(MEARA)_6$. Furthermore, two concomitant nucleation sites far apart in the sequence are possible. Folding of the three-stranded antiparallel β-sheet peptide beta3s started with the formation of most of the side chain contacts and hydrogen bonds between strands 2 and 3, followed by the 1–2 interstrand contacts. The inverse sequence of events, i.e., first formation of 1–2 and then 2–3 contacts was also observed, but less frequently [11].

The free energy barrier seems to have an important entropic component in both helical peptides and antiparallel β-sheets. In an α-helix, it originates from constraining the backbone conformation of three consecutive amino acids before the first helical hydrogen bond can form, while in the antiparallel β-sheet it is due to the constraining of a β-hairpin onto which a third strand can coalesce [11]. Therefore, the folding of the two most common types of secondary structure seems to have similarities (a mainly entropic nucleation barrier and an exponential folding rate) as well as important differences (location of the barrier and multiple vs. two pathways). The similarities are in accord with a plethora of experimental and theoretical evidence [20] while the differences might be a consequence of the fact that $Y(MEARA)_6$ has about 7–9 helical turns whereas the three-stranded antiparallel β-sheet consists of only two ''minimal blocks'', i.e., two β-hairpins.

32.2.1.2 Non-Arrhenius Temperature Dependence of the Folding Rate

Small molecule reactions show an Arrhenius-like temperature dependence, i.e., faster rates at higher temperatures. Protein folding is a complex reaction involving many degrees of freedom; the folding rate is Arrhenius-like at physiological temperatures, but deviates from Arrhenius behavior at higher temperatures [20].

To quantitatively investigate the kinetics of folding, MD simulations of two model peptides, Ace-$(AAQAA)_3$-$NHCH_3$ (α-helical stable structure) and Ace-$V_5{}^{D}PGV_5$-NH_2 (β-hairpin), were performed using the same implicit solvation model [13]. Folding and unfolding at different temperature values were studied by 862 simulations for a total of 4 μs [21]. Different starting conformations (folded

and random) were used to obtain a statistically significant sampling of conformational space at each temperature value. An important feature of the folding of both peptides is the negative activation enthalpy at high temperatures. The rate constant for folding initially increases with temperature, goes through a maximum at about T_m, and then decreases [21]. The non-Arrhenius behavior of the folding rate is in accord with experimental data on two mainly alanine α-helical peptides [22, 23], a β-hairpin [24], CI2, and barnase [25], lysozyme [26, 27], and lattice simulation results [28–30]. It has been proposed that the non-Arrhenius profile of the folding rate originates from the temperature dependence of the hydrophobic interaction [31, 32]. The MD simulation results show that a non-Arrhenius behavior can arise at high values of the temperature in a model where all the interactions are temperature independent. This has been found also in lattice simulations [28, 29]. The curvature of the folding rate at high temperature may be a property of a reaction dominated by enthalpy at low temperatures and entropy at high temperatures [30]. The non-Arrhenius behavior for a system where the interactions do not depend on the temperature might be a simple consequence of the temperature dependence of the accessible configuration space. At low temperatures, an increase in temperature makes it easier to jump over the energy barriers, which are rate limiting. However, at very high temperatures, a larger portion of the configuration space becomes accessible, which results in a slowing down of the folding process.

32.2.1.3 Denatured State and Levinthal Paradox

The size of the accessible conformational space and how it depends on the number of residues is not easy to estimate. To investigate the complexity of the denatured state four molecular dynamics runs of beta3s were performed at the melting temperature of the model (330 K) for a total simulation time of 12.6 μs [33]. The simulation length is about two orders of magnitude longer than the average folding or unfolding time (about 85 ns each), which are similar because at the melting temperature the folded and unfolded states are equally populated. The peptide is within 2.5 Å Cα root mean square deviation (RMSD) from the folded conformation about 48% of the time. Figure 32.1 shows the results of a cluster analysis based on Cα RMSD. There are more than 15 000 conformers (cluster centers) and it is evident that a plateau has not been reached within the 12.6 μs of simulation time. However, the number of significantly populated clusters (see Ref. [12] for a detailed description) converges already within 2 μs. Hence, the simulation-length dependence of the total number of clusters is dominated by the small ones. At each simulation interval between an unfolding event and the successive refolding event additional conformations are sampled. More than 90% of the unfolded state conformations are in small clusters (each containing less than 0.1% of the saved snapshots) and the total number of small clusters does not reach a plateau within 12.6 μs. Note that there is also a monotonic growth with simulation time of the number of snapshots in the folded-state cluster. After 12.6 μs (and also within each of the four trajectories) the system has sampled an equilibrium of folded and unfolded states despite a large part of the denatured state ensemble has not yet been explored. In fact, the average folding time converges to a value around 85 ns which

shows that the length of each simulation is much larger than the relaxation time of the slowest conformational change. Interestingly, in the average folding time of about 85 ns beta3s visits less than 400 clusters (diamond in Figure 32.1). This is only a small fraction of the total amount of conformers in the denatured state. It is possible to reconcile the fast folding with the large conformational space by analyzing the effective energy, which includes all of the contributions to the free energy except for the configurational entropy of the protein [11, 34]. Fast folding of beta3s is consistent with the monotonically decreasing profile of the effective energy (inset in Figure 32.1). Despite the large number of conformers in the denatured state ensemble, the protein chain efficiently finds its way to the folded state because native-like interactions are on average more stable than nonnative ones.

In conclusion, the unfolded state ensemble at the melting temperature is a large collection of conformers differing among each other, in agreement with previous high temperature molecular dynamics simulations [8, 35]. The energy ''bias'' which makes fast folding possible does not imply that the unfolded state ensemble is made up of a small number of statistically relevant conformations. The simulations provide further evidence that the number of denatured state conformations is orders of magnitudes larger than the conformers sampled during a folding event. This result also suggests that measurements which imply an average over the unfolded state do not necessarily provide information on the folding mechanism.

32.2.1.4 **Folding Events of Trp-cage**

Very small proteins are ideal systems to validate force fields and simulation methodology. Neidigh et al. [36] have truncated and mutated a marginally stable 39-residue natural sequence thereby designing a 20-residue peptide, the Trp-cage, that is more than 95% folded in aqueous solution at 280 K. The stability of the Trp-cage is due to the packing of a Trp side chain within three Pro rings and a Tyr side chain. Moreover, the C-terminal half contains four Pro residues which dramatically restrict the conformational space, i.e., entropy, of the unfolded state [36, 37].

Four MD studies have appeared in the 12 months following the publication of the Trp-cage structure [38–41]. All of the simulations were started from the completely extended conformation and used different versions of the AMBER force field and the generalized Born continuum electrostatic solvation model [42]. Two simulations were run with conventional constant temperature MD at 300 K [40] and 325 K [38], a third study used replica exchange MD with a range of temperatures from 250 K to 630 K [41], and in the fourth paper distributed computing simulations at 300 K with full water viscosity were reported [39].

An important problem of the three constant temperature studies is that the Trp-cage seems to fold to a very deep free energy minimum and no unfolding events have been observed [38–40]. Moreover, only one folding event is presented by Simmerling et al. [38] and Chowdhury et al. [40]. The poor statistics does not allow to draw any conclusions on free energy landscapes or on the folding mechanism of the Trp-cage.

Another potential problem is the discrepancy between the most stable state

sampled by MD and the NMR conformers. Only in two of the four MD studies NOE distance restraints were measured along the trajectories and about 20% were found to be violated [40, 41]. Moreover, as explicitly stated by the authors, the native state sampled by distributed computing contains a π-helix (instead of the α-helix) and the Trp is not packed correctly in the core [39]. These discrepancies are significant because the Trp-cage has a very small core and a rather rigid C-terminal segment.

32.2.2 Unfolding Simulations of Proteins

32.2.2.1 High-temperature Simulations

Since the early work of Daggett and Levitt [43] and Caflisch and Karplus [44], several other high-temperature simulation studies have been concerned with exploring protein unfolding pathways. Several comprehensive review articles exist on this simulation protocol [45] which has been widely used since. Recent MD simulations at temperatures of 100 °C and 225 °C of a three-helix bundle 61-residue protein, the engrailed homeodomain (En-HD), by Daggett and coworkers [46, 47] have been used to analyze a folding intermediate at atomic level of detail. The unfolding half-life of the En-HD at 100 °C has been extrapolated to be about 7.5 ns, a time scale that can be accessed by MD simulations with explicit water molecules.

Also, unfolding simulations in the presence of explicit urea molecules have shown that the protein (barnase) remains stable at 300 K but unfolds partially at moderately high temperature (360 K) [48]. The results suggested a mechanism for urea induced unfolding due to the interaction of urea with both polar and nonpolar groups of the protein.

32.2.2.2 Biased Unfolding

Because of the limitations on simulation times and height of the barriers to conformational transitions in proteins, a number of methods, alternative to the use of high, nonphysical temperatures, have been proposed to accelerate such transitions by the introduction of an external time-dependent perturbation [49–54]. The perturbation induce the reaction of interest in a reasonable amount of time (the strength of the perturbation is inversely proportional to the available computer time). These methods have been used for studying not only protein unfolding at native or realistic denaturing conditions, but also large conformational changes between known relevant conformers [51]. Their goal is to generate pathways which are realistic, in spite of the several orders of magnitude reduction in the time required for the conformational change. They are not alternatives to methods to compute free energy profiles along defined pathways. The external perturbation is usually applied to a function of the coordinates which is assumed to vary monotonically as the protein goes from the native to the nonnative state of interest. For certain perturbations the unfolding pathways obtained have been shown to depend on the nature of the perturbation and the choice of the reaction coordinate [52];

this is even more the case when the perturbation is strong and the reaction is induced too quickly for the system to relax along the pathway.

A perturbation which is particularly ''gentle'' since it exploits the intrinsic thermal fluctuations of the system and produces the acceleration by selecting the fluctuations that correspond to the motion along the reaction coordinate has also been used to unfold proteins [55, 56]. This perturbation has been employed, in particular, to expand α-lactalbumin by increasing its radius of gyration starting from the native state, and generate a large number of low-energy conformers that differ in terms of their root mean square deviation, for a given radius of gyration. The resulting structures were relaxed by unbiased simulations and used as models of the molten globule (see Chapter 23) and more unfolded denatured states of α-lactalbumin based on measured radii of gyration obtained from nuclear magnetic resonance experiments [57]. The ensemble of compact nonnative structures agree in their overall properties with experimental data available for the α-lactalbumin molten globule, showing that the native-like fold of the α-domain is preserved and that a considerable proportion of the antiparallel β-sheet in the β-domain is present. This indicated that the lack of hydrogen exchange protection found experimentally for the β-domain [58] is due to rearrangement of the β-sheet involving transient populations of nonnative β-structures in agreement with more recent infrared spectroscopy measurements [59]. The simulations also provide details concerning the ensemble of structures that contribute as the molten globule unfolds and shows, in accord with experimental data [60], that the unfolding is not cooperative, i.e., the various structural elements do not unfold simultaneously.

32.2.2.3 **Forced Unfolding**

Unfolding by stretching proteins individually has become routinely mainly thanks to the advent of the atomic force microscopy technology [61]. This peculiar way of unfolding proteins opened new perspectives on protein folding studies. Experiments are usually performed on engineered homopolyproteins, and the I27 domain from titin has become the reference system for this type of studies. Experiments measure force-extension profiles, and show typical ''saw-tooth'' profiles, where peaks are due to the sudden unfolding of individual domains, sequentially in time, causing a drop in the recorded force. These profiles are generally interpreted assuming that the unfolding event is determined by a single barrier which is decreased by the external force. For a detailed description of the experimental techniques and of the most recent results on forced unfolding of single molecules (by atomic force microscopy and optical tweezers) see Chapter 31.

To provide a structural interpretation of the typical saw-tooth-like spectra measured in single molecule stretching experiments, various simulation techniques have been proposed, where detailed all-atom models of proteins are stretched by pulling two atoms apart [62, 63], differing mainly in the way the solvent is treated.

In some cases simulation can effectively explain the force patterns measured (see Ref. [64] for a review). For all the proteins experimentally unfolded by pulling, only a simple saw-tooth pattern has been recorded related to the sudden unfolding

when the protein was pulled beyond a certain length, i.e., a simple two-state behavior. Simulations showed a more complex behavior [63] with possible intermediates on the forced unfolding pathways for certain proteins.[1]

Simulation has been used [65] to compare forced unfolding of two protein classes (all-β-sandwich proteins and all-α-helix bundle proteins). In particular, simulations suggested that different proteins should show a significantly different forced unfolding behavior, both within a protein class and for the different classes and dramatic differences between the unfolding induced by high temperature and by external pulling forces. The result was shown to be correlated to the type of perturbation, the folding topologies, the nature of the secondary and tertiary interactions and the relative stability of the various structural elements [65]. Improvements in the AFM technique combined with protein engineering methods have now confirmed (see Chapter 31) that chemical (or thermal) and forced unfolding occur through different pathways and that forced unfolding is related to crossing of a free energy barrier which might not be unique, but might change with force magnitude or upon specific mutations [66].

It should be borne in mind, however, that the forced unfolding of proteins is a nonequilibrium phenomenon strongly dependent on the pulling speed, and, since time scales in simulations and experiments are very different, the respective pathways need not to be the same. Recently, through a combination of experimental analysis and molecular dynamics simulations it has been shown that, in the case of mechanical unfolding, pathways might effectively be the same in a large range of pulling speeds or forces [67, 68], thus providing another demonstration of the robustness of the energy landscape (i.e., the funnel-like shape of the free energy surface sculpted by evolution is not affected by the application of even strong perturbations [69, 70]). In two recent papers [71, 72] it has been shown that proteins resist differently when pulled in different directions. In both cases the experiments have been complemented with simulations, with either explicit or implicit solvents. In both cases the behavior observed experimentally is qualitatively reproduced. This fact strongly suggests, although does not prove it, that in this particular case the forced unfolding mechanisms explored in the simulations is the same as that which determines the experimentally measured force.

Difference between solvation models is discussed in detail in Section 32.3.4 in the context of forced unfolding simulations, the disadvantage of an explicit solvation model [62, 72, 73] relative to an implicit [63, 65, 67, 68, 71] is not only that of being much slower, but also to provide an environment which relaxes slowly relative to the fast unraveling of the protein under force. Moreover, properly hydrating with explicit water a partially extended protein requires a large quantity of water, thus requiring a very large amount of CPU time for a single simulation. Implicit solvent models, on the other hand, allow unfolding to be performed at much lower

1) The presence of ''late'' intermediates on the forced unfolding pathway was first observed [63] in the 10th domain of fibronectin type III from fibronectin (FNfn3). A more complex pattern than equally spaced peaks in the force extension profile was predicted to arise from the presence of a kinetically metastable state. Most recent experimental results (J. Fernandez, personal communication) confirm the behavior predicted by the simulation.

forces (or pulling speeds) and multiple simulations to be used to study the dependence of the results on the initial conditions and/or on the applied force.

32.2.3 Determination of the Transition State Ensemble

The understanding of the folding mechanism has crucially advanced since the development of a method which provide information on the transition state [74] (see Chapter 13). The method allows the structure of the transition state at the level of single residue to be probed by measuring the change in folding and unfolding rates upon mutations. The method provides a so-called ϕ-value for each of the mutated residues which is a measure of the formation of native structure around the residue: a ϕ-value of 1 suggests that the residue is in a native environment at the transition state while a ϕ-value of 0 can be interpreted as a loss of the interactions of the residue at the transition state. Fractional ϕ-values are more difficult to interpret, but have been shown to arise from weakened interactions [75] and not from a mixture of species, some with fully formed and some with fully broken interaction.

As we discussed in Section 32.2.2.1, the use of high temperature makes it possible to observe the unfolding of a protein by MD on a time scale which can be simulated on current computers. Valerie Daggett and collaborators [76] first had the idea of performing a very high-temperature simulation and looking for a sudden change in the structure of the protein along the trajectory, indicating the escape from the native minimum of the free energy surface. The collections of structures around the ''jump'' were assumed to constitute a sample of the putative transition state. Assuming that the experimental ϕ-values correspond in microscopic terms to fraction of native contacts, they found a good agreement between calculated and experimental values. This approach was initially applied to the protein CI2, a small two-state proteins which has been probably the most thoroughly studied by experimental ϕ-value analysis; it has been subsequently improved and extended to the study of several proteins for which experimental ϕ-values were available (see Ref. [77] for review and other references).

Another related method has been used recently [78] to unfold a protein by high-temperature simulation (srcSH3 in the specific case) and determine a putative transition state by looking for conformations where the difference between calculated and experimental ϕ-values was smallest.

Both methods presented above have the advantage of providing structures extracted from an unfolding trajectory and thus the fast refolding or complete unfolding from these structures (a property of transition states) has been reported [78, 79]. But both approaches only provide few transition state structures, because a long simulation is required to generate each member of the transition state ensemble, while the transition state can be a quite broad ensemble for some proteins [80].

In a recent development, it has been shown that the amount of information that can be obtained from experimental measurements can be expanded further by

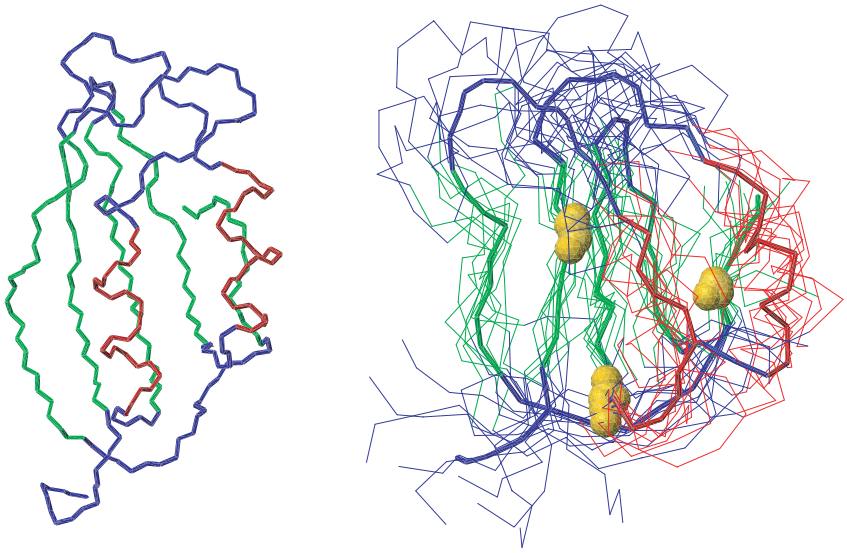

Fig. 32.2. Comparison between the native state structure (left) and the most representative structures of the transition state ensemble of AcP, determined by all-atom molecular dynamics simulations [82]. Native secondary structure elements are show in color (the two α-helices are plotted in red and the β-sheet in green). The three key residues for folding are shown as gold spheres [81, 82]. Figure from Ref. [69].

using the data to build up phenomenological energy functions to bias computer generated trajectories. With this approach (see Section 32.3.2 for a more detailed description of the technique), conformations compatible with experimental data are determined directly during the simulations [81, 82], rather than being obtained from filtering procedures such as those discussed above [78]. The incorporation of experimental data into the energy function creates a minimum in correspondence of the state observed experimentally and therefore allows for a very efficient sampling of conformational space. The transition state for folding of acylphosphatase (see Figure 32.2) was determined in this way [81, 82], showing that the network of interactions that stabilize the transition state is established when a few key residues form their native-like arrangement.

Based on this computational technique, a general approach in which theory and experiments are combined in an iterative manner to provide a detailed description of the transition state ensemble has been recently proposed [83]. In the first iteration, a coarse-grained determination of the transition state ensemble (TSE) is carried out by using a limited set of experimental ϕ-values as constraints in a molecular dynamics simulation. The resulting model of the TSE is used to determine the additional residues whose ϕ-value measurement would provide the most information for refining the TSE. Successive iterations with an increasing number of ϕ-value measurements are carried out until no further changes in the properties of the TSE are detected or there are no additional residues whose ϕ-values can be

measured. The method can be also used to find key residues for folding (i.e., those that are most important for the formation of the TSE).

The study of the transition state represents probably the most interesting example of how experiment and molecular dynamics simulations complement each other in understanding and visualizing the folding mechanisms in terms of relevant structures involved. Simulations are performed with approximate force fields and unfolding induced using artificial means (such as high temperature or other perturbations). At this stage in the development of MD simulations, the experiment provides evidence that what is observed in silico is consistent with what happens in vitro. On the other hand, and particularly in the case in which the full ensemble of conformations compatible with the experimental results is generated [82, 83], the simulation suggests further mutations to increase the resolution of the picture of the transition state, and allows detailed hypothesis of the mechanisms, such as the identity and structure of the residues involved in the folding nucleus [150].

32.3 MD Techniques and Protocols

32.3.1 Techniques to Improve Sampling

A thorough sampling of the relevant conformations is required to accurately describe the thermodynamics and kinetics of protein folding. Since the energetic and entropic barriers are higher than the thermal energy at physiological temperature, standard MD techniques often fail to adequately sample the conformational space. As already mentioned in this chapter, even for a small protein it is currently not yet feasible to simulate reversible folding with a high-resolution approach (e.g., MD simulations with an all-atom model). The practical difficulties in performing such brute force simulations have led to several types of computational approaches and/or approximative models to study protein folding. An interesting approach is to unfold starting from the native structure [84–86] but detailed comparison with experiments [47] is mandatory to make sure that the high-temperature sampling does not introduce artifacts. In addition, a number of approaches to enhance sampling of phase space have been introduced [87, 88]. They are based on adaptive umbrella sampling [89], generalized ensembles (e.g., entropic sampling, multicanonical methods, replica exchange methods) [90], modified Hamiltonians [91–93], multiple time steps [94], or combinations thereof.

32.3.1.1 Replica Exchange Molecular Dynamics

Replica exchange is an efficient way to simulate complex systems at low temperature and is the simplest and most general form of simulated tempering [95]. Sugita and Okamoto have been the first to extend the original formulation of replica exchange into an MD-based version (REMD), testing it on the pentapeptide Met-

enkephalin in vacuo [96]. The basic idea of REMD is to simulate different copies (replicas) of the system at the same time but at different temperatures values.

We recently applied a REMD protocol to implicit solvent simulations of a 20-residue three-stranded antiparallel β-sheet peptide (beta3s) [97]. Each replica evolves independently by MD and every 1000 MD steps (2 ps), states i, j with neighbor temperatures are swapped (by velocity rescaling) with a probability $w_{ij} = \exp(-\Delta)$ [96], where $\Delta \equiv (\beta_i - \beta_j)(E_j - E_i)$, $\beta = 1/kT$ and E is the potential energy. During the 1000 MD steps the Berendsen thermostat [98] is used to keep the temperature close to a given value. This rather tight coupling and the length of each MD segment (2 ps) allow the kinetic and potential energy of the system to relax. High temperature simulation segments facilitate the crossing of the energy barriers while the low-temperature ones explore in detail the conformations present in the minimum energy basins. The result of this swapping between different temperatures is that high-temperature replicas help the low-temperature ones to jump across the energy barriers of the system. In the beta3s study eight replicas were used with temperatures between 275 and 465 K [97].

The higher the number of degrees of freedom in the system the more replicas should be used. It is not clear how many replicas should be used if a peptide or protein is simulated with explicit water. The transition probability between two temperatures depends on the overlap of the energy histograms. The histograms' width depends on $1/\sqrt{N}$ (where N is the size of the system). Hence, the number of replicas required to cover a given temperature range increases with the size. Moreover, in order to have a random walk in temperature space (and then a random walk in energy space which enhances the sampling), all the temperature exchanges should occur with the same probability. This probability should be at least of 20–30%. To optimize the efficiency of the method, one should find the best compromise between the number of replicas to be used, the temperature space to cover and the acceptance ratios for temperature exchanges. In the literature there is no clear indication about the selection of temperatures and empirical methods are usually applied (weak point of the method). The choice of the boundary temperatures depends on the system under study. The highest temperature has to be chosen in order to overcome the highest energy barriers (probably higher in explicit water) separating different basins; the lowest temperature to investigate the details of the different basins.

Sanbonmatsu and Garcia have applied REMD to investigate the structure of Met-enkephalin in explicit water [99] and the α-helical stabilization by the arginine side chain which was found to originate from the shielding of main-chain hydrogen bonds [100]. Furthermore, the energy landscape of the C-terminal β-hairpin of protein G in explicit water has been investigated by REMD [101, 102]. Recently, a multiplexed approach with multiple replicas for each temperature level has been applied to large-scale distributed computing of the folding of a 23-residue miniprotein [103]. Starting from a completely extended chain, conformations close to the NMR structures were reached in about 100 trajectories (out of a total of 4000) but no evidence of reversible folding (i.e., several folding and unfolding events in the same trajectory) was presented [103].

32.3.1.2 Methods Based on Path Sampling

A very promising computational method, called transition path sampling (reviewed in Ref. [104]) has been recently used [105] to study the folding of a β-hairpin in explicit solvent. The method allows in principle the study of rare events (such as protein folding) without requiring knowledge of the mechanisms, reaction coordinates, and transition states. Transition path sampling focuses on the sampling not of conformations but of trajectories linking two conformations or regions (possibly basins of attraction) in the conformational space. Other methods focus on building ensemble of paths connecting states; the stochastic path approach [106] and the reaction path method [107] have been also used to study the folding of peptides and small proteins in explicit solvent. The stochastic path ensemble and the reaction path methods introduce a bias in the computed trajectories but allow the exploration of long time scales. All the methods mentioned above are promising but rely on the choice of a somewhat arbitrary initial unfolded conformation beside the final native one.

32.3.2 MD with Restraints

A method to generate structures belonging to the TSE ensemble discussed in Section 32.2.3 consists in performing molecular dynamics simulations restrained with a pseudo-energy function based on the set of experimental ϕ-values. The ϕ-values are interpreted as the fraction of native contacts present in the structures that contribute to the TSE. With this restraint the TSE becomes the most stable state on the potential energy surface rather than being an unstable region, as it is for the true energy function of the protein. This procedure is conceptually related to that used to generate native state structures compatible with measurements from nuclear magnetic resonance (NMR) experiments, in that pseudo-energy terms involving experimental restraints are added to the protein force field [108, 109]. The main difference is that an approach is required to sample a broad state compatible with some experimental restraints, rather than a method to search for an essentially unique native structure.

The method is based on molecular dynamics simulations using an all-atom model of the protein [110, 111] and an implicit model for the solvent [112] with an additional term in the energy function:

$$\rho = \frac{1}{N_\phi} \sum_{i \in E} (\phi_i - \phi_i^{\exp})^2 \tag{1}$$

where E is the list of the N_ϕ available experimental ϕ-values, $\phi_i^{\exp}$. The ϕ_i-value of amino acid i in the conformation at time t is defined as

$$\phi_i(t) = \frac{N_i(t)}{N_i^{\text{nat}}} \tag{2}$$

where $N_i(t)$ is the number of native contacts of i at time t and N_i^{nat} the number of native contacts of i in the native state.

Molecular dynamics simulations are then performed to sample all the possible structures compatible with the restraints. The structures thus generated are not necessarily at the transition state for folding for the potential used. They provide instead a structural model of the experimental transition state, including all possible structures compatible with the restraints derived from the experiment. The experimental information provided by the ϕ-values might not be enough to restrain the sampling to meaningful structures (e.g., when only few mutations have been performed). In such circumstance, other experimentally measured quantities, such as the m-value, which is related to the solvent accessible surface, must be used to restrain the sampling or to a posteriori select meaningful structures.

This type of computational approach relies on the assumption implicit in Eq. (2). This consists in approximating a ϕ-value, measured as a ratio of free energy variations upon mutation, as a ratio of side-chain contacts. A definition based on side-chains is appropriate since experimental ϕ-values are primarily a measure of the loss of side-chain contacts at the transition state, relative to the native state. Although simply counting contacts, rather than calculating their energies, is a crude approximation [113], it has been shown that there is a good correlation between loss of stability and loss of side-chain contacts within about 6 Å on mutation [114]. Also, Shea et al. [115] have found in their model calculations that this approximation for estimating ϕ-values from structures is a good one under certain conditions. A more detailed relation between experimental ϕ-values and atomic contacts could in principle be established by using the energies of the all-atom contacts made by the side chain of the mutated amino acid.

The same approach can be extended to generate the structures corresponding to other unfolded or intermediate states as the site-specific information provided by the experiment is steadily increasing (see Chapters 20 and 21).

32.3.3 Distributed Computing Approach

As mentioned in the introduction, the problem of simulating the folding process of any sequence from a random conformation is mainly a problem of potentials and computer time. Duan and Kollman [116] have showed that a huge effort in parallelizing (on a medium-scale, 256 processors) an MD code and exploiting for several months a several million dollars computer (a Cray T3E) could lead to the simulation of 1 μs of the small protein villin headpiece. Even approaching the typical experimental folding times (which is, however, larger than 1 ms for most proteins), a statistical characterization of the folding process is still impossible in the foreseeable future.

Developing a large-scale parallelization method seems the most viable approach, as the cost of fast CPUs decreases steadily and their performances approach those of much more expensive mainframes. Time being sequential, MD codes are not massively parallelizable in an efficient way. A good scaling is usually obtained for

large systems with explicit water and a relatively small number of processors (between 2 and 100, depending on the program and the problem studied). One approach has been proposed that allows the scalability of a MD simulation to be pushed to the level of being able to use efficiently a network of heterogeneous and loosely connected computer [117]. The approach (called distributed computing) exploits the stochastic nature of the folding process. In general protein folding involves the crossing of free energy barriers. The approach is most easily understood assuming that the proteins have a single barrier and a single exponential kinetic (which is the case for a large number of small proteins [118]). The probability that a protein is folded after a time t is $P(t) = 1 - \exp(-kt)$, where k is the folding rate. Thus, for short times, and considering M proteins or independent simulations, the probability of observing a folding event is Mkt. So, if M is large, there is a sizable probability of observing a folding event on simulations much shorter than the time constant of the folding process [119]. The folding rate could then in principle be estimated by running M independent simulations (starting from the completely extended conformation with different random velocities) for a time t and counting the number N of simulations which end up in the folded state as $k = N/(Mt)$. Simulations have been reported where the folding rate estimated in this way (assuming that partial refolding counts as folding) is in good agreement with the experimental one (see, for example, Ref. [39]).

However, it has been argued [120] that even for simple two-state proteins, folding has a series of early conformational steps that lead to lag phases at the beginning of the kinetics. Their presence can bias short simulations toward selecting minor pathways that have fewer or faster lag steps and so miss the major folding pathways. This fact has been clearly observed by comparing equilibrium and fast folding trajectories simulations [121] for a 20-residue three-stranded antiparallel β-sheet peptide (beta3s). It was found that the folding rate is estimated correctly by the distributed computing approach when trajectories longer than a fraction of the equilibrium folding time are considered; in the case of the 20-residue peptide studied within the frictionless implicit solvation model used for the simulations, this time is about 1% of the average folding time at equilibrium. However, careful analysis of the folding trajectories showed that the fastest folding events occur through high-energy pathways, which are unlikely under equilibrium conditions (see Section 32.2.1.1). Along these very fast folding pathways the peptide does not relax within the equilibrium denatured state which is stabilized by the transient presence of both native and nonnative interactions. Instead, collapse and formation of native interactions coincides and, unlike at equilibrium, the formation of the two β-hairpins is nearly simultaneous.

These results demonstrate that the ability to predict the folding rate does not imply that the folding mechanisms are correctly characterized: the fast folding events occur through a pathway that is very unlikely at equilibrium. However, extending the time scale of the short simulations to 10% of the equilibrium folding time, the folding mechanism of the fast folding events becomes almost indistinguishable from equilibrium folding events. It must be stressed that this result is not general but concerns the specific peptide studied; the explicit presence of sol-

vent molecule (and the consequent friction), might decrease the differences between equilibrium and shortest folding events. Unfortunately, this kind of validation of the distributed computing approach is not possible for a generic protein in a realistic solvent, as equilibrium simulations are not feasible.

An alternative method to use many processors simultaneously to access time scales relevant in the folding process by MD simulations has recently been proposed by Settanni et al. [122]. The method is based on parallel MD simulations that are started from the denatured state; trajectories are periodically interrupted, and are restarted only if they approach the transition (or some other target) state. In other words, the method choses trajectories along which a cost function decreases. The effectiveness of such an approach was shown by determining the transition state for folding an SH3 domain using as cost function the deviation between experimental and computed ϕ-values (Eq. (1) in Section 32.3.2). The method can efficiently use a large number of computers simultaneously because simulations are loosely coupled (i.e., only the comparison between final conformations, needed periodically to choose which trajectory to restart, involve communications between CPUs). This method can also be extended to complex nondifferentiable cost functions.

32.3.4
Implicit Solvent Models versus Explicit Water

Incorporating solvent effects in MD and Monte-Carlo simulations is of key importance in quantitatively understanding the chemical and physical properties of biomolecular processes. Accurate electrostatic energies of proteins in an aqueous environment are needed in order to discriminate between native and nonnative conformations. An exact evaluation of electrostatic energies considers the interactions among all possible solute–solute, solute–solvent, and solvent–solvent pairs of charges. However, this is computationally expensive for macromolecules. Continuum dielectric approximations offer a more tractable approach [123–127]. The essential concept in continuum models is to represent the solvent by a high dielectric medium, which eliminates the solvent degrees of freedom, and to describe the macromolecule as a region with a low dielectric constant and a spatial charge distribution. The Poisson equation provides an exact description of such a system. The increase in computation speed for a finite difference solution of the Poisson equation [128–131] with respect to an explicit treatment of the solvent is remarkable but still not enough for effective utilization in computer simulations of macromolecules. The generalized Born (GB) model was introduced to facilitate an efficient evaluation of continuum electrostatic energies [42]. It provides accurate energetics and the most efficient implementations are between five and ten times slower than in vacuo simulations [132–134]. The essential element of the GB approach is the calculation of an effective Born radius for each atom in the system which is a measure of how deeply the atom is buried inside the protein. This information is combined in a heuristic way to obtain a correction to the Coulomb law for each atom pair [42]. For the integration of energy density, necessary to obtain the effective Born radii, both numerical [42, 132, 135] and analytical [134, 136, 137]

implementations exist. The former are more accurate but slower than the latter [135]. Moreover, analytical derivatives that are required for MD simulations are not given by numerical implementations.

For efficiency reasons empirical dielectric screening functions are the most common choice in MD simulations with implicit solvent. One kind of solvation model is based on the use of a dielectric function that depends linearly on the distance r between two charges ($\varepsilon(r) = \alpha r$) [138, 139] or has a sigmoidal shape [140, 141]. Although very fast, these options suffer from their inability to discriminate between buried and solvent exposed regions of a macromolecule and are therefore rather inaccurate. A distance and exposure dependent dielectric function has been proposed [142]. Recently, an approach based on the distribution of solute atomic volumes around pairs of charges in a macromolecule has been proposed to calculate the effective dielectric function of proteins in aqueous solution [143].

The simulation results presented in Section 32.2.1 were obtained using an implicit solvent model based on a fast analytical approximation of the solvent accessible surface (SAS) [13] and the CHARMM force field [110]. The former drastically reduces the computational cost with respect to an explicit solvent simulation. The SAS model is based on the approximation proposed by Lazaridis and Karplus [112] for dielectric shielding due to the solvent, and the surface area model for the hydrophobic effect introduced by Eisenberg and McLachlan [144]. Electrostatic screening effects are approximated by a distance-dependent dielectric function and a set of partial charges with neutralized ionic groups [112]. An approximate analytical expression [145] is employed to calculate the SAS because an exact analytical or numerical computation of the SAS is too slow to compete with simulations in explicit solvent. The SAS model is based on the assumptions that most of the solvation energy arises from the first water shell around the protein [144] and that two atomic solvation parameters are sufficient to describe these effects at a qualitative level of accuracy. Within these assumptions, the SAS energy term approximates the solute–solvent interactions (i.e., it should account for the energy of cavity formation, solute–solvent dispersion interactions, and the direct (or Born) solvation of polar groups). The two atomic solvation parameters were optimized by performing 1 ns MD simulations at 300 K on six small proteins [13]. It is important to underline that the structured peptides discussed in Section 32.2.1 were not used for the calibration of the SAS atomic solvation parameters. The SAS model is a good approximation for investigating the folded and denatured state (large ensemble of conformers) of structured peptides. Its limitations, in particular for highly charged peptides and large proteins, have been discussed [13].

The most detailed and physically sound approaches (e.g., explicit solvent and particle mesh Ewald treatment of the long-range electrostatic interactions [146]) are still approximations and might introduces artifacts (see, for example, Ref. [147]). All solvation models, even those computationally most expensive, are approximations and their range of validity is difficult to explore. It is likely that most proteins will unfold fast relative to the experimental time scale if one could afford long (e.g., 100 ns) explicit water MD simulations even at room temperature. Some evidence of this instability has been recently published [148].

32.4 Conclusion

It is a very exciting time for studying protein folding using multidisciplinary approaches rooted in physics, chemistry, and computer science. The time scale gap between folding in vitro and in silico is being continuously reduced and this will bring interesting surprises. We expect an increasing role of MD simulations in the elucidation of protein folding thanks to further improvements in force fields and solvation models.

References

1 Karplus, M. & McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nature Struct. Biol.* **9**, 646–652.

2 Dill, K. A. & Chan, H. S. (1997). From Levinthal to pathways to funnels. *Nature Struct. Biol.* **4**, 10–19.

3 Mirny, L. & Shakhnovich, E. (2001). Protein folding theory: From lattice to all-atom models. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 361–396.

4 Daggett, V. & Fersht, A. R. (2003). Is there a unifying mechanism for protein folding? *Trends Biochem. Sci.* **28**, 18–25.

5 Creighton, T. E. (1992). *Protein Folding*. W. H. Freeman & Co., New York.

6 Merz Jr, K. M. & LeGrand, S. M. (1994). *The Protein Folding Problem and Tertiary Structure Prediction*. Birkhäuser, Boston.

7 Pain, R. H., ed. (2000). *Mechanisms of Protein Folding*. Oxford University Press, Oxford.

8 Shea, J. E. & Brooks III, C. L. (2001). From folding theories to folding proteins: A review and assessment of simulation studies of protein folding and unfolding. *Annu. Rev. Phys. Chem.* **52**, 499–535.

9 Galzitskaya, O. V., Higo, J. & Finkelstein, A. V. (2002). α-helix and β-hairpin folding from experiment, analytical theory and molecular dynamics simulations. *Curr. Protein Pept. Sci.* **3**, 191–200.

10 Gnanakaran, S., Nymeyer, H., Portman, J., Sanbonmatsu, K. Y. & Garcia, A. E. (2003). Peptide folding simulations. *Curr. Opin. Struct. Biol.* **13**, 168–174.

11 Ferrara, P. & Caflisch, A. (2000). Folding simulations of a three-stranded antiparallel β-sheet peptide. *Proc. Natl Acad. Sci. USA* **97**, 10780–10785.

12 Ferrara, P. & Caflisch, A. (2001). Native topology or specific interactions: What is more important for protein folding? *J. Mol. Biol.* **306**, 837–850.

13 Ferrara, P., Apostolakis, J. & Caflisch, A. (2002). Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins* **46**, 24–33.

14 de Alba, E., Santoro, J., Rico, M. & Jimenez, M. A. (1999). De novo design of a monomeric three-stranded antiparallel β-sheet. *Protein Sci.* **8**, 854–865.

15 Schenck, H. L. & Gellman, S. H. (1998). Use of a designed triple-stranded antiparallel β-sheet to probe β-sheet cooperativity in aqueous solution. *J. Am. Chem. Soc.* **120**, 4869–4870.

16 McCallister, E. L., Alm, E. & Baker, D. (2000). Critical role of β-hairpin formation in protein G folding. *Nature Struct. Biol.* **7**, 669–673.

17 Wright, C. F., Lindorff-Larsen, K., Randles, L. G. & Clarke, J. (2003). Parallel protein-unfolding pathways

revealed and mapped. *Nature Struct. Biol.* **10**, 658–662.

18 Richardson, J. M., McMahon, K. W., MacDonald, C. C. & Makhatadze, G. I. (1999). MEARA sequence repeat of human CstF-64 polyadenylation factor is helical in solution. A spectroscopic and calorimetric study. *Biochemistry* **38**, 12869–12875.

19 Shirley, W. A. & Brooks III, C. L. (1997). Curious structure in "canonical" alanine-based peptides. *Proteins* **28**, 59–71.

20 Karplus, M. (2000). Aspects of protein reaction dynamics: Deviations from simple behavior. *J. Phys. Chem. B* **104**, 11–27.

21 Ferrara, P., Apostolakis, J. & Caflisch, A. (2000). Thermodynamics and kinetics of folding of two model peptides investigated by molecular dynamics simulations. *J. Phys. Chem. B* **104**, 5000–5010.

22 Clarke, D. T., Doig, A. J., Stapley, B. J. & Jones, G. R. (1999). The α-helix folds on the millisecond time scale. *Proc. Natl Acad. Sci. USA* **96**, 7232–7237.

23 Lednev, I. K., Karnoup, A. S., Sparrow, M. C. & Asher, S. A. (1999). α-Helix peptide folding and unfolding activation barriers: A nanosecond UV resonance raman study. *J. Am. Chem. Soc.* **121**, 8074–8086.

24 Muñoz, V., Thompson, P. A., Hofrichter, J. & Eaton, W. A. (1997). Folding dynamics and mechanism of β-hairpin formation. *Nature* **390**, 196–199.

25 Oliveberg, M., Tan, Y. J. & Fersht, A. R. (1995). Negative activation enthalpies in the kinetics of protein folding. *Proc. Natl Acad. Sci. USA* **92**, 8926–8929.

26 Segawa, S. & Sugihara, M. (1984). Characterization of the transition state of lysozyme unfolding. I. Effect. *Biopolymers* **23**, 2473–2488.

27 Matagne, A., Jamin, M., Chung, E. W., Robinson, C. V., Radford, S. E. & Dobson, C. M. (2000). Thermal unfolding of an intermediate is associated with non-Arrhenius kinetics in the folding of hen lysozyme. *J. Mol. Biol.* **297**, 193–210.

28 Karplus, M., Caflisch, A., Sali, A. & Shakhnovich, E. (1995). Protein dynamics: From the native to the unfolded state and back again. In *Modelling of Biomolecular Structures and Mechanisms* (Pullman, A., Jortner, J. & Pullman, B., eds), pp. 69–84, Kluwer Academic, Dordrecht, The Netherlands.

29 Karplus, M. (1997). The Levinthal paradox: Yesterday and today. *Folding Des.* **2**, S69–S75.

30 Dobson, C. M., Sali, A. & Karplus, M. (1998). Protein folding: A perspective from theory and experiment. *Angew. Chem. Int. Ed.* **37**, 868–893.

31 Scalley, M. L. & Baker, D. (1997). Protein folding kinetics exhibit an Arrhenius temperature dependence when corrected for the temperature dependence of protein stability. *Proc. Natl Acad. Sci. USA* **94**, 10636–10640.

32 Chan, H. S. & Dill, K. A. (1998). Protein folding in the landscape perspective: Chevron plots and non-Arrhenius kinetics. *Proteins* **30**, 2–33.

33 Cavalli, A., Haberthür, U., Paci, E. & Caflisch, A. (2003). Fast protein folding on downhill energy landscape. *Protein Sci.* **12**, 1801–1803.

34 Dinner, A. R., Sali, A., Smith, L. J., Dobson, C. M. & Karplus, M. (2000). Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem. Sci.* **25**, 331–339.

35 Wong, K. B., Clarke, J., Bond, C. J. et al. (2000). Towards a complete description of the structural and dynamic properties of the denatured state of barnase and the role of residual structure in folding. *J. Mol. Biol.* **296**, 1257–1282.

36 Neidigh, J. W., Fesinmeyer, R. M. & Andersen, N. H. (2002). Designing a 20-residue protein. *Nature Struct. Biol.* **9**, 425–430.

37 Gellman, S. H. & Woolfson, D. N. (2002). Mini-proteins Trp the light

fantastic. *Nature Struct. Biol.* **9**, 408–410.

38 Simmerling, C., Strockbine, B. & Roitberg, A. E. (2002). All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.* **124**, 11258–11259.

39 Snow, C. D., Zagrovic, B. & Pande, V. S. (2002). The Trp cage: Folding kinetics and unfolded state topology via molecular dynamics simulations. *J. Am. Chem. Soc.* **124**, 14548–14549.

40 Chowdhury, S., Lee, M. C., Xiong, G. & Duan, Y. (2003). Ab initio folding simulation of the Trp-cage mini-protein approaches NMR resolution. *J. Mol. Biol.* **327**, 711–717.

41 Pitera, J. W. & Swope, W. (2003). Understanding folding and design: replica-exchange simulations of "Trp-cage" miniproteins. *Proc. Natl Acad. Sci. USA* **100**, 7587–7592.

42 Still, W. C., Tempczyk, A., Hawley, R. C. & Hendrickson, T. (1990). Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112**, 6127–6129.

43 Daggett, V. & Levitt, M. (1993). Protein unfolding pathways explored through molecular dynamics simulations. *J. Mol. Biol.* **232**, 600–619.

44 Caflisch, A. & Karplus, M. (1994). Molecular dynamics simulation of protein denaturation: Solvation of the hydrophobic cores and secondary structure of barnase. *Proc. Natl Acad. Sci. USA* **91**, 1746–1750.

45 Daggett, V. & Fersht, A. (2003). Opinion: The present view of the mechanism of protein folding. *Nature Rev. Mol. Cell Biol.* **4**, 497–502.

46 Mayor, U., Johnson, C. M., Daggett, V. & Fersht, A. R. (2000). Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation. *Proc. Natl Acad. Sci. USA* **97**, 13518–13522.

47 Mayor, U., Guydosh, N. R., Johnson, C. M. et al. (2003). The complete folding pathway of a protein from nanoseconds to microseconds. *Nature* **421**, 863–867.

48 Caflisch, A. & Karplus, M. (1999). Structural details of urea binding to barnase: A molecular dynamics analysis. *Structure* **7**, 477–488.

49 Hao, M.-H., Pincus, M. R., Rachovsky, S. & Scheraga, H. A. (1993). Unfolding and refolding of the native structure of bovine pancreatic trypsin inhibitor studied by computer simulations. *Biochemistry* **32**, 9614–9631.

50 Harvey, S. C. & Gabb, H. A. (1993). Conformational transition using molecular dynamics with minimum biasing. *Biopolymers* **33**, 1167–1172.

51 Schlitter, J., Engels, M., Kruger, P., Jacoby, E. & Wollmer, A. (1993). Targeted molecular dynamics simulation of conformational change. Application to the TR transition in insulin. *Mol. Simulations* **10**, 291–308.

52 Hünenberger, P. H., Mark, A. E. & van Gunsteren, W. F. (1995). Computational approaches to study protein unfolding: Hen egg white lysozyme as a case study. *Proteins* **21**, 196–213.

53 Ferrara, P., Apostolakis, J. & Caflisch, A. (2000). Computer simulations of protein folding by targeted molecular dynamics. *Proteins* **39**, 252–260.

54 Ferrara, P., Apostolakis, J. & Caflisch, A. (2000). Targeted molecular dynamics simulations of protein unfolding. *J. Phys. Chem. B* **104**, 4511–4518.

55 Paci, E., Smith, L. J., Dobson, C. M. & Karplus, M. (2001). Exploration of partially unfolded states of human α-lactalbumin by molecular dynamics simulation. *J. Mol. Biol.* **306**, 329–347.

56 Marchi, M. & Ballone, P. (1999). Adiabatic bias mlecular dynamics: A method to navigate the conformational space of complex molecular systems. *J. Chem. Phys.* **110**, 3697–3702.

57 Wilkins, D. K., Grimshaw, S. B., Receveur, V., Dobson, C. M., Jones, J. A. & Smith, L. J. (1999). Hydrodynamic radii of native and denatured proteins measured by pulse NMR techniques. *Biochemistry* **38**, 16424–16431.

58 Kuwajima, K. (1996). The molten globule state of α-lactalbumin. *FASEB J.* **10**, 102–109.

59 Troullier, A., Reinstädler, D., Dupont, Y., Naumann, D. & Forge, V. (2000). Transient nonnative secondary structures during the refolding of α-lactalbumin by infrared spectroscopy. *Nature Struct. Biol.* **7**, 78–86.

60 Schulman, B., Kim, P. S., Dobson, C. M. & Redfield, C. (1997). A residue-specific NMR view of the non-cooperative unfolding of a molten globule. *Nature Struct. Biol.* **4**, 630–634.

61 Rief, M., Gautel, M., Oesterhelt, F., Fernandez, J. M. & Gaub, H. E. (1997). Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science* **276**, 1109–1112.

62 Lu, H., Isralewitz, B., Krammer, A., Vogel, V. & Schulten, K. (1998). Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation. *Biophys. J.* **75**, 662–671.

63 Paci, E. & Karplus, M. (1999). Forced unfolding of fibronectin type 3 modules: An analysis by biased molecular dynamics simulations. *J. Mol. Biol.* **288**, 441–459.

64 Isralewitz, B., Gao, M. & Schulten, K. (2001). Steered molecular dynamics and mechanical functions of proteins. *Curr. Opin. Struct. Biol.* **11**, 224–230.

65 Paci, E. & Karplus, M. (2000). Unfolding proteins by external forces and high temperatures: The importance of topology and energetics. *Proc. Natl Acad. Sci. USA* **97**, 6521–6526.

66 Williams, P. M., Fowler, S. B., Best, R. B. et al. (2003). Hidden complexity in the mechanical properties of titin. *Nature* **422**, 446–449.

67 Fowler, S., Best, R. B., Toca-Herrera, J. L. et al. (2002). Mechanical unfolding of a titin Ig domain: Structure of unfolding intermediate revealed by combining AFM, molecular dynamics simulations, NMR and protein engineering. *J. Mol. Biol.* **322**, 841–849.

68 Best, R. B., Fowler, S., Toca-Herrera, J. L., Steward, A., Paci, E. & Clarke, J. (2003). Mechanical unfolding of a titin Ig domain: Structure of transition state revealed by combining atomic force microscopy, protein engineering and molecular dynamics simulations. *J. Mol. Biol.* **330**, 867–877.

69 Vendruscolo, M. & Paci, E. (2003). Protein folding: Bringing theory and experiment closer together. *Curr. Opin. Struct. Biol.* **13**, 82–87.

70 Vendruscolo, M., Paci, E., Karplus, M. & Dobson, C. M. (2003). Structures and relative free energies of partially folded states of proteins. *Proc. Natl Acad. Sci. USA* **100**, 14817–14821.

71 Brockwell, D. J., Paci, E., Zinober, R. C. et al. (2003). Pulling geometry defines the mechanical resistance of a β-sheet protein. *Nature Struct. Biol.* **10**, 731–737.

72 Carrion-Vazquez, M., Li, H., Lu, H., Marszalek, P. E., Oberhauser, A. F. & Fernandez, J. M. (2003). The mechanical stability of ubiquitin is linkage dependent. *Nature Struct. Biol.* **10**, 738–743.

73 Lu, H. & Schulten, K. (2000). The key event in force-induced unfolding of titin's immunoglobulin domains. *Biophys. J.* **79**, 51–65.

74 Fersht, A. R., Matouschek, A. & Serrano, L. (1992). The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.* **224**, 771–782.

75 Fersht, A. R., Itzhaki, L. S., elMasry, N. F., Matthews, J. M. & Otzen, D. E. (1994). Single versus parallel pathways of protein folding and fractional structure in the transition state. *Proc. Natl Acad. Sci. USA* **91**, 10426–10429.

76 Li, A. & Daggett, V. (1994). Characterization of the transition state of protein unfolding by use of molecular dynamics: Chymotrypsin inhibitor 2. *Proc. Natl Acad. Sci. USA* **91**, 10430–10434.

77 Daggett, V. (2002). Molecular dynamics simulations of the protein

unfolding/folding reaction. *Acc. Chem. Res.* **35**, 422–429.

78 Gsponer, J. & Caflisch, A. (2002). Molecular dynamics simulations of protein folding from the transition state. *Proc. Natl Acad. Sci. USA* **99**, 6719–6724.

79 DeJong, D., Riley, R., Alonso, D. O. & Daggett, V. (2002). Probing the energy landscape of protein folding/unfolding transition states. *J. Mol. Biol.* **319**, 229–242.

80 Fersht, A. R. (1999). *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding.* W. H. Freeman & Co., New York.

81 Vendruscolo, M., Paci, E., Dobson, C. M. & Karplus, M. (2001). Three key residues form a critical contact network in a protein folding transition state. *Nature* **409**, 641–645.

82 Paci, E., Vendruscolo, M., Dobson, C. M. & Karplus, M. (2002). Determination of a transition state at atomic resolution from protein engineering data. *J. Mol. Biol.* **324**, 151–163.

83 Paci, E., Clarke, J., Steward, A., Vendruscolo, M. & Karplus, M. (2003). Self-consistent determination of the transition state for protein folding. Application to a fibronectin type III domain. *Proc. Natl Acad. Sci. USA* **100**, 394–399.

84 Daggett, V. & Levitt, M. (1993). Realistic simulations of native-protein dynamics in solution and beyond. *Annu. Rev. Biophys. Biomol. Struct.* **22**, 353–380.

85 Caflisch, A. & Karplus, M. (1995). Acid and thermal denaturation of barnase investigated by molecular dynamics simulations. *J. Mol. Biol.* **252**, 672–708.

86 Lazaridis, T. & Karplus, M. (1997). "New View" of protein folding reconciled with the old through multiple unfolding simulations. *Science* **278**, 1928–1931.

87 Frenkel, D. & Smit, B. (1996). *Understanding Molecular Simulation*, 2nd edition, Academic Press, London.

88 Berne, B. J. & Straub, J. E. (1997). Novel methods of sampling phase space in the simulation of biological systems. *Curr. Opin. Struct. Biol.* **7**, 181–189.

89 Bartels, C. & Karplus, M. (1997). Multidimensional adaptive umbrella sampling: Applications to main chain and side chain peptide conformations. *J. Comput. Chem.* **18**, 140–1462.

90 Mitsutake, A., Sugita, Y. & Okamoto, Y. (2001). Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers* **60**, 96–123.

91 Wu, X. & Wang, S. (1998). Self-guided molecular dynamics simulation for efficient conformational search. *J. Phys. Chem. B* **102**, 7238–7250.

92 Apostolakis, J., Ferrara, P. & Caflisch, A. (1999). Calculation of conformational transitions and barriers in solvated systems: Application to the alanine dipeptide in water. *J. Chem. Phys.* **110**, 2099–2108.

93 Andricioaei, I., Dinner, A. R. & Karplus, M. (2003). Self-guided enhanced sampling methods for thermodynamic averages. *J. Chem. Phys.* **118**, 1074–1084.

94 Schlick, T., Barth, E. & Mandziuk, M. (1997). Biomolecular dynamics at long timesteps: bridging the time-scale gap between simulation and experimentation. *Annu. Rev. Biophys. Biomol. Struct.* **26**, 181–222.

95 Marinari, E. & Parisi, G. (1992). Simulated tempering: A new Monte Carlo scheme. *Europhys. Lett.* **19**, 451–458.

96 Sugita, Y. & Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314**, 141–151.

97 Rao, F. & Caflisch, A. (2003). Replica exchange molecular dynamics simulations of reversible folding. *J. Chem. Phys.* **119**, 4035–4042.

98 Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690.

99 Sanbonmatsu, K. Y. & Garcia, A. E. (2002). Structure of Met-enkephalin in

explicit aqueous solution using replica exchange molecular dynamics. *Proteins* **46**, 225–234.

100 Garcia, A. E. & Sanbonmatsu, K. Y. (2002). Alpha-helical stabilization by side chain shielding of backbone hydrogen bonds. *Proc. Natl Acad. Sci. USA* **99**, 2782–2787.

101 García, A. E. & Sanbonmatsu, K. Y. (2001). Exploring the energy landscape of a hairpin in explicit solvent. *Proteins* **42**, 345–354.

102 Zhou, R., Berne, B. J. & Germain, R. (2001). The free energy landscape for β hairpin folding in explicit water. *Proc. Natl Acad. Sci. USA* **98**, 14931–14936.

103 Rhee, Y. M. & Pande, V. S. (2003). Multiplexed-replica exchange molecular dynamics method for protein folding simulation. *Biophys. J.* **84**, 775–786.

104 Bolhuis, P. G., Chandler, D., Dellago, C. & Geissler, P. L. (2002). Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem*, **53**, 291–318.

105 Bolhuis, P. G. (2003). Transition-path sampling of β-hairpin folding. *Proc. Natl Acad. Sci. USA* **100**, 12129–12134.

106 Elber, R., Meller, J. & Olender, R. (1999). Stochastic path approach to compute atomically detailed trajectories: Application to the folding of C peptide. *J. Phys. Chem. B*, **103**, 899–911.

107 Eastman, P., Gronbech-Jensen, N. & Doniach, S. (2001). Simulation of protein folding by reaction path annealing. *J. Chem. Phys.* **114**, 3823–3841.

108 Wüthrich, K. (1989). Protein structure determination in solution by nuclear magnetic resonance spectroscopy. *Science* **243**, 45–50.

109 Clore, G. M. & Schwieters, C. D. (2002). Theoretical and computational advances in biomolecular NMR spectroscopy. *Curr. Opin. Struct. Biol.* **12**, 146–153.

110 Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization and dynamics calculations. *J. Comput. Chem.* **4**, 187–217.

111 Neria, E., Fischer, S. & Karplus, M. (1996). Simulation of activation free energies in molecular dynamics system. *J. Chem. Phys.* **105**, 1902–1921.

112 Lazaridis, T. & Karplus, M. (1999). Effective energy function for protein dynamics and thermodynamics. *Proteins* **35**, 133–152.

113 Paci, E., Vendruscolo, M. & Karplus, M. (2002). Native and nonnative interactions along protein folding and unfolding pathways. *Proteins* **47**, 379–392.

114 Cota, E., Hamill, S. J., Fowler, S. B. & Clarke, J. (2000). Two proteins with the same structure respond very differently to mutation: The role of plasticity in protein stability. *J. Mol. Biol.* **302**, 713–725.

115 Shea, J.-E., Onuchic, J. N. & Brooks III, C. L. (1999). Exploring the origins of topological frustration: Design of a minimally frustrated model of fragment B of protein A. *Proc. Natl Acad. Sci. USA* **96**, 12512–12517.

116 Duan, Y. & Kollman, P. A. (1998). Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **282**, 740–744.

117 Shirts, M. & Pande, V. (2000). COMPUTING: Screen savers of the world unite! *Science* **290**, 1903–1904.

118 Jackson, S. E. (1998). How do small single-domain proteins fold? *Folding Des.* **3**, R81–R91.

119 Pande, V. S., Baker, I., Chapman, J. et al. (2003). Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. *Biopolymers* **68**, 91–109.

120 Fersht, A. R. (2002). On the simulation of protein folding by short time scale molecular dynamics and distributed computing. *Proc. Natl Acad. Sci. USA* **99**, 14122–14125.

121 Paci, E., Cavalli, A., Vendruscolo, M. & Caflisch, A. (2003). Analysis of

the distributed computing approach applied to the folding of a small β peptide. *Proc. Natl Acad. Sci. USA* **100**, 8217–8222.

122 Settanni, G., Gsponer, J. & Caflisch, A. (2004). Formation of the folding nucleus of an SH3 domain investigated by loosely coupled molecular dynamics simulations. *Biophys. J.* **86**, 1691–1701.

123 Roux, B. & Simonson, T. (1999). Implicit solvent models. *Biophys. Chem.* **78**, 1–20.

124 Gilson, M. K. (1995). Theory of electrostatic interactions in macromolecules. *Curr. Opin. Struct. Biol.* **5**, 216–223.

125 Tomasi, J. & Persico, M. (1994). Molecular interactions in solution: An overview of methods based on continuous distributions of the solvent. *Chem. Rev.* **94**, 2027–2094.

126 Cramer, C. J. & Truhlar, D. G. (1999). Implicit solvation models: Equilibria, structure, spectra, and dynamics. *Chem. Rev.* **99**, 2161–2200.

127 Orozco, M. & Luque, F. J. (2000). Theoretical methods for the description of the solvent effect in biomolecular systems. *Chem. Rev.* **100**, 4187–4226.

128 Warwicker, J. & Watson, H. C. (1982). Calculation of the electric potential in the active site cleft due to α-helix dipoles. *J. Mol. Biol.* **157**, 671–679.

129 Gilson, M. K. & Honig, B. H. (1988). Energetics of charge-charge interactions in proteins. *Proteins* **3**, 32–52.

130 Bashford, D. & Karplus, M. (1990). pKa's of ionizable groups in proteins: Atomic detail from a continuum electrostatic model. *Biochemistry* **29**, 10219–10225.

131 Davis, M. E., Madura, J. D., Luty, B. A. & McCammon, J. A. (1991). Electrostatics and diffusion of molecules in solution – simulations with the University-of-Houston-brownian dynamics program. *Comput. Phys. Comm.* **62**, 187–197.

132 Scarsi, M., Apostolakis, J. & Caflisch, A. (1997). Continuum electrostatic energies of macromolecules in aqueous solutions. *J. Phys. Chem. B* **101**, 8098–8106.

133 Bashford, D. & Case, D. A. (2000). Generalized Born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* **51**, 129–152.

134 Lee, M. S., Feig, M., Salsbury, F. R. & Brooks III, C. L. (2003). New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. *J. Comput. Chem.* **24**, 1348–1356.

135 Lee, M. S., Salsbury, F. R. & Brooks III, C. L. (2002). Novel generalized Born methods. *J. Chem. Phys.* **116**, 10606–10614.

136 Qiu, D., Shenkin, P. S., Hollinger, F. P. & Still, W. C. (1997). The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem. A* **101**, 3005–3014.

137 Dominy, B. N. & Brooks III, C. L. (1999). Development of a generalized Born model parametrization for proteins and nucleic acids. *J. Phys. Chem. B* **103**, 3765–3773.

138 Warshel, A. & Levitt, M. (1976). Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* **103**, 227–249.

139 Gelin, B. R. & Karplus, M. (1979). Side-chain torsional potentials: effect of dipeptide, protein, and solvent environment. *Biochemistry* **18**, 1256–1268.

140 Mehler, E. L. (1990). Comparison of dielectric response models for simulating electrostatic effects in proteins. *Protein Eng.* **3**, 415–417.

141 Wang, L., Hingerty, B. E., Srinivasan, A. R., Olson, W. K. & Broyde, S. (2002). Accurate representation of B-DNA double helical structure with implicit solvent and counterions. *Biophys. J.* **83**, 382–406.

142 Mallik, B., Masunov, A. & Lazaridis, T. (2002). Distance and exposure dependent effective dielectric

function. *J. Comput. Chem.* **23**, 1090–1099.

143 Haberthür, U., Majeux, N., Werner, P. & Caflisch, A. (2003). Efficient evaluation of the effective dielectric function of a macromolecule in aqueous solution. *J. Comput. Chem.* **24**, 1936–1949.

144 Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature* **319**, 199–203.

145 Hasel, W., Hendrickson, T. F. & Still, W. C. (1988). A rapid approximation to the solvent accessible surface areas of atoms. *Tetrahedron Comput. Methodol.* **1**, 103–116.

146 Darden, T. A., York, D. M. & Pedersen, L. (1993). Particle mesh Ewald: An N log(N) method for computing Ewald sums. *J. Chem. Phys.* **98**, 10089–10092.

147 Weber, W., Hunenberger, P. H. & McCammon, J. A. (2000). Molecular dynamics simulations of a polyalanine octapeptide under Ewald boundary conditions: Influence of artificial periodicity on peptide conformation. *J. Phys. Chem. B* **104**, 3668–3675.

148 Fan, H. & Mark, A. E. (2003). Relative stability of protein structures determined by X-ray crystallography or NMR spectroscopy: a molecular dynamics simulation study. *Proteins* **53**, 111–120.

149 Hiltpold, A., Ferrara, P., Gsponer, J. & Caflisch, A. (2000). Free energy surface of the helical peptide Y(MEARA)6. *J. Phys. Chem. B* **104**, 10080–10086.

150 Lindorff-Larsen, K., Vendruscolo, M., Paci, E. & Dobson, C. M. (2004). Transition states for protein folding have native topologies despite high structural variability. *Nature Struct. Mol. Biol* **11**, 443–449.