Transition states for protein folding using molecular dynamics and experimental restraints

2007 J. Phys.: Condens. Matter 19 285211

(http://iopscience.iop.org/0953-8984/19/28/285211)

IOP Publishing — Journal of Physics: Condensed Matter

J. Phys.: Condens. Matter **19** (2007) 285211 (15pp) doi:10.1088/0953-8984/19/28/285211

Transition states for protein folding using molecular dynamics and experimental restraints

Lucy R Allen and Emanuele Paci

School of Physics and Astronomy and Astbury Centre for Structural Molecular Biology,
University of Leeds, Leeds LS2 9JT, UK

E-mail: e.paci@leeds.ac.uk

Received 30 October 2006, in final form 23 November 2006
Published 25 June 2007
Online at stacks.iop.org/JPhysCM/19/285211

Abstract

The mechanism through which a given sequence of amino acids finds its way to a global free energy minimum cannot yet be predicted by theory or numerical simulation. Much of the information available on the protein folding mechanism derives from the so-called ϕ values. These are believed to probe the structure of the rate limiting step, or transition state, for the folding of two-state proteins. In recent years experimental ϕ values have been widely used to benchmark the results of simulations, mostly of unfolding, which have been achieved using detailed sequence-dependent molecular models. A few years ago a novel technique was proposed which uses ϕ values as restraints so that only conformations which are transition-state-like are sampled in the simulation. This technique, albeit grounded on several approximations and assumptions, has provided an unprecedented structural representation of the transition state for folding. Here we explore various issues concerning the generation of ensembles of structures representing the transition state. One important result is that by allowing a large tolerance on the experimental restraints the information contained in the latter is lost; this suggests that an experimental error on the ϕ values which is too large might affect the results of restrained simulations and the picture provided by them.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

Proteins are heteropolymers which in general assume a specific three-dimensional conformation which is crucial for their biological function. Folding, i.e. the process through which a protein finds its folded conformation starting from a random conformation, is of fundamental importance. Our ability to describe it accurately in simple terms will facilitate immensely the task of predicting which sequences fold to specific structures and

0953-8984/07/285211+15$30.00 © 2007 IOP Publishing Ltd Printed in the UK

understanding why some sequences misfold under certain conditions. While novel techniques and sophisticated experiments are providing additional information on the folding pathways, a simple and coherent structural description of the folding mechanism remains elusive.

Measurement of so-called ϕ values [1] can give valuable information about the folding process, particularly for small, two-state proteins. Whilst ϕ values do not, by themselves, provide a structural picture of the transition state, they have been broadly used to benchmark the results of simulations (see e.g. [2] for a recent application) by assuming that they can be interpreted as the fraction of native structure present in the transition state at the level of individual residues. Recently a technique has been proposed [3, 4] which allows transition state structures to be generated at an atomistic level of detail. The method uses ϕ values as restraints in simulation so that only conformations which are transition-state-like are sampled. The technique has been widely used in the past few years [5–11]. The method is quite general and can be applied not only to transition states but also to other 'states': if some experimental observable, interpretable in structural terms, is available then these can be modelled with the same approach [12–16]. Here, we explore some technically important issues which have been so far overlooked or not thoroughly justified in previous applications of the method. We do so by focusing on the 86-residue two-state α helical protein Im9 using two different models, one coarse grained and structure based and another all atom and sequence based.

Definition of the restraints

ϕ values, measured using protein engineering methods, correspond to the ratio of the change in free energy of the transition state ($\Delta\Delta G_l^{\mathrm{TS}}$) and the change in free energy of the native state ($\Delta\Delta G_l^{\mathrm{NS}}$) on mutation of residue i:

$$\phi_I^{\mathrm{exp}} = \frac{\Delta\Delta G_l^{\mathrm{TS}}}{\Delta\Delta G_l^{\mathrm{NS}}}. \qquad (1)$$

Using ϕ values as restraints requires their determination as a function of the Cartesian coordinates of the protein.

One traditional way to interpret the ϕ value of a residue is to equate it to the number of contacts made by the side-chain of the residue at the transition state. The ϕ value of residue I at time t is calculated as

$$\phi_I^{\mathrm{calc}}(t) = \frac{N_I(t)}{N_I^{\mathrm{nat}}} \qquad (2)$$

where N_I is the number of contacts involving residue I (or its side-chain).

This definition of a ϕ value contains several approximations: one is that the free energy can be approximated by an enthalpy, another is that the enthalpy can be approximated by the number of contacts, yet another is that in the denatured state each residue makes no contacts. Such or similar approximations are necessary in order to interpret a ratio of free energy differences in terms of coordinates. Analogous assumptions are also used within more sophisticated empirical approaches to estimate the free energy change upon a mutation based on a structure (see e.g. [17]). One other commonly used approximation, which is not necessary but 'reasonable', is that the number of contacts in the numerator of equation (2) includes only native contacts. This assumption rules out ϕ values larger than unity, and relies on the fact that non-native interactions are believed to be rare and non-specific at the transition state. Below we use two approximations, i.e. including or disregarding the contribution of non-native contacts to the ϕ values, and compare the two.

Conformational averaging

Experimental ϕ values derive from free energies, which are not a property of each structure but of the whole ensemble of molecules in the test tube; strictly speaking, ϕ^{calc} should be estimated by averaging the number of contacts over an ensemble of molecules. In previous applications the restraint method has usually been applied to a single protein molecule. Ensemble-averaged ϕ values can be estimated by performing a number of independent simulations (or copies), as suggested by Davis *et al* [18]. Below we investigate the effect of increasing numbers of copies on the transition state ensemble (TSE), and empirically determine how many copies are required for good sampling.

Molecular models

The method of the restraints involves adding an additional, usually harmonic, term to the original Hamiltonian (see 'Methods'). Most TSEs determined using ϕ value restrained molecular dynamics (MD) have been produced using all-atom models: generally the initial restrained simulations are carried out with an implicit solvent model, with explicit solvent occasionally being used for later refinement of the TSE structures [19]. Whilst more realistic than coarse-grained models, all-atom simulations are computationally expensive: an exceedingly large amount of CPU time is required to sample to convergence all the structures compatible with a set of experimental restraints in the presence of a rugged energy landscape. In contrast, equilibrium properties can be determined by using computationally efficient coarse-grained models, such as native-centric Go models [20, 21]. The question of whether a native-centric Go model contains enough information to accurately predict experimental folding properties has been widely debated [22–24]. However, it is possible that, if used together with experimental data, for example in ϕ-value-restrained simulations, such models are adequate, since the shortcomings of the force field are compensated by the introduction of experimental information in the model. In this paper we address this issue by using experimental restraints to calculate TSEs using both an all-atom, sequence-based model and a coarse-grained, structure-based model. One of the questions we aim to answer is if the transition state is mostly or entirely characterized by the experimental restraints, or if the underlying force field plays a significant role.

Sampling

The determination of the ensemble of structures corresponding to the experimental restraints imposed through a stiff harmonic potential, at a given temperature, requires a sampling technique such as molecular dynamics or Monte Carlo. The method used in most previous applications has been described in detail in [4]. The procedure consists of driving the protein from an initial (usually native) conformation to one in which restraints are satisfied; when the deviation is close to zero a stiff harmonic term is added to the Hamiltonian to maintain the restraints. Sampling is then performed by performing MD simulations, increasing the temperature to overcome local barriers and sample a broader spectrum of low-energy conformations. Conformations are subsequently 'cooled down' by either minimizing their energy or by simulated annealing. The first issue that we can address using a simple model is the role of the sampling temperature: using replica exchange MD (REMD), we are able to compare directly TSE distributions obtained from replicas at different temperatures.

Another issue is that of the choice of the initial conformation: in principle we do not want to bias the structures to be in the neighbourhood of the native state. In practice, with all-atom, sequence-dependent models, if the procedure above is carried out starting from a random conformation an extremely broad non-converging ensemble of very non-native structures is found. The simpler coarse-grained model allows us to release this assumption: random initial conformations can be used (different ones in the case of the ensemble and replica exchange simulations) and the final result can be shown to be independent of the choice of these conformations. A further interesting issue is how precisely the restraints should be satisfied. In the limit of infinitely large bias, ϕ values will be satisfied exactly (unless the restraints are not simultaneously compatible, which is not likely to be the case when only a few ϕ^{exp} are available). If a finite bias is used, ϕ values will be satisfied approximatively; smaller biases might be more appropriate if one considers that the ϕ^{exp} are affected by an experimental error, which is often sizable [25].

The real transition state

For activated rate processes governed by stochastic dynamics, the transition state or, equivalently, the dividing surface between the reactants and products corresponds to the stochastic separatrix [26]. Starting from any point on the separatrix, the protein reaches the folded and the unfolded state for the first time with an equal probability, i.e., $p_{\mathrm{fold}} = 0.5$. The TS structures determined in restrained simulations do not necessarily correspond to structures on the reaction pathway, as they are by construction local minima of a potential energy function including an artificial energy term. Below we investigate the nature of both the transition state structures extracted from equilibrium unrestrained simulations (and satisfying $p_{\mathrm{fold}} = 0.5$), and also those calculated in simulations restrained using as ϕ^{exp} the ϕ values derived from those true (for the model) transition state structures.

2. Methods

Restraints

The definition of ϕ^{calc} given in equation (2) is used, with N_I, the number of contacts made by residue I, calculated as

$$N_I = \sum_{i\in I}^{M}\sum_{j\notin I}^{M}\theta(r_{ij}-r_c)\Delta_{ij}(Q). \tag{3}$$

M is the number of atoms (in the all-atom model) or residues (in the coarse-grained one) in the protein, r_{ij} is the distance between atoms or residues i and j, θ is the Heaviside function, and $\Delta_{ij}(Q) = 1$ if residues i and j are at least Q residues apart in the sequence and zero otherwise. For the cut-off distance we used 5.5 Å between side-chain heavy atoms of residues more than two residues apart in the sequence for the all-atom model, and 11 Å between the C_α atoms of residues more than four residues apart in the sequence for the coarse-grained model. Such criteria for defining the number of contacts give values in the same range for the two models; the results presented below depend weakly on this choice.

Conformations for which $\phi^{\mathrm{calc}} \simeq \phi^{\mathrm{exp}}$ are sampled by adding an additional term $E = \frac{\alpha}{2}\rho^2$ to the potential energy; α is a parameter which controls the strength of the bias, and ρ is the mean square deviation between the ϕ^{exp} and the ϕ^{calc} values:

$$\rho(t) = \frac{1}{N_\phi}\sum_i(\phi_i^{\mathrm{calc}}(t)-\phi_i^{\mathrm{exp}})^2 \tag{4}$$

where the sum runs over all the residues for which an experimental ϕ value is available.

Sampling

To sample conformations compatible with the restraints at a broad range of temperatures, REMD has been used [27]. Replica exchange involves performing a series of simulations at different temperatures in parallel, and occasionally swapping the conformations of one system with another; this allows the lowest temperature simulation to make use of the fact that trajectories at higher temperatures sample many different minima. Swapping of conformations is realized via a Monte-Carlo-like move, in which, in order to satisfy detailed balance, the probability of an attempted swap being accepted is related to the Boltzmann weights of the two states.

Conformational averaging

In the restrained REMD simulations, the bias is applied to each replica separately, so that $\phi^{\text{calc}} \simeq \phi^{\text{exp}}$ at each temperature. In the ensemble simulations it is the value of ϕ^{calc} averaged over all the copies which must satisfy the restraints, i.e.,

$$\langle \phi_i^{\text{calc}} \rangle = \frac{1}{N} \sum_J^N \phi_{iJ}^{\text{calc}} \simeq \phi^{\text{exp}} \tag{5}$$

where i is the residue number, J the copy label and N the number of molecules in the ensemble.

Protein

Im9 is an 86-residue four-helix bundle protein. The NMR solution structure [28] (PDB code 1IMQ) was used for construction of the Go model potential, and as a starting and reference structure for simulations. Eighteen experimental ϕ values measured by Friel *et al* [29] were used as restraints.

Coarse-grained model

The structure-based coarse-grained model used here is the Go model proposed by Karanicolas and Brooks [30]. This model has only C_α atoms which interact attractively only if they are in contact in the protein's native structure. It has some specific features, such as an additional repulsive term which results in a small energy barrier corresponding to a desolvation penalty, and pairwise interactions whose magnitudes depend on the residue type. The model is implemented in the CHARMM molecular mechanics simulation package [31]. Holonomic constraints were applied to the C_α–C_α bonds, making an integration timestep of 15 fs possible. Constant temperature was maintained by using Langevin dynamics with a friction coefficient of 0.1 ps^{-1}. Six replicas, at 300, 330, 360, 400, 450 and 500 K were used in all the REMD simulations, leading to exchange probabilities between 0.3 and 0.5. Six different initial conformations, in which the protein was in the unfolded state, were taken from an equilibrium simulation at 330 K. REMD was run for 150 ns, with swaps being attempted every 15 ps. The trajectory at 300 K was used for analysis. In the ensemble simulations all copies were kept at 300 K. For the restrained simulations, a harmonic bias was applied at $\rho = 0$. The strength of the bias, α, was set to 5×10^5 for all simulations unless specified otherwise in the text.

All-atom model

The all-atom, sequence-dependent force field used here is based on the CHARMM united-atom force field, with the implicit solvent EEF1 [32], implemented in the CHARMM program. Constraints were applied to bonds involving hydrogen atoms, and simulations were performed with a 2 fs timestep. Langevin dynamics, with a friction coefficient of 0.1 ps^{-1}, were used. REMD was carried out with ten replicas at 300, 310, 320, 331, 343, 356, 370, 385, 402 and 420 K, leading to exchange probabilities between 0.18 and 0.37. The starting configurations for replica exchange were found by initially heating the protein from its minimized PDB structure to 300 K over 4 ns, then equilibrating at 300 K for 5.5 ns and saving coordinates and velocities every 500 ps. The ϕ value restraints were applied to each of the ten structures, initially reducing ρ towards zero using a biased MD approach [4, 33], and then simply by using a harmonic potential centred at $\rho = 0$. Each structure was then heated to its target temperature over 2 ns, still under the harmonic potential, and equilibrated for 1 ns before replica exchange began. REMD was run for 100 ns, with swaps attempted every 5 ps. The trajectory at 300 K was used for analysis. The strength of the bias, α, was set to 5×10^5.

True transition state

Using the Go model the folding–unfolding reaction occurs spontaneously over long trajectories. Over a 30 μs trajectory, about 80 transitions could be observed. For such a model, the total number of contacts is already a reasonably good reaction coordinate in the sense that native and unfolded states correspond to two distinct ranges of values for this coordinate. Using a re-weighted contact map as suggested by Best and Hummer [34], an optimized reaction coordinate in which the two states are well separated can be obtained. This coordinate makes it easier to select transition conformations along the trajectory. The selection of these conformations which were found to have a $p_{\text{fold}} \simeq 0.5$ (about 25% of them) were taken as 'real' transition states for the model. From these conformations ϕ^{Go} values were calculated using the definition in equation (2). p_{fold} was calculated using the procedure described by Hubner *et al* [11]: each potential TS structure was used as the initial conformation for 100 MD runs of length 3.75 ns (2.5×10^5 steps) at 325 K. p_{fold} was then calculated as the fraction of runs that resulted in a folded structure. The results of equilibrium simulations were used to define a structure as folded if $Q_N \geqslant 0.52$ and RMSD $\leqslant 7$ Å (here and below the RMSD refers to residues 5–79, i.e., disregarding the termini. Q_N is the fraction of native contacts: a contact is defined as being present if two C_α atoms, more than four residues apart in sequence, are separated by 12 Å or less, and contacts are defined to be native if they are present in the energy minimized experimental native structure).

3. Results

3.1. Unbiased simulations

Using the C_α Go model, equilibrium simulations where many folding–unfolding events occur are feasible for several proteins; in principle the transition state can be extracted from these simulations [34]. With this model the protein Im9 has a mid-point temperature of approximately 325 K, at which the folding and unfolding times are about 400 ns. Histograms of the RMSD from the native structure and the fraction of native contacts (678 in the energy-minimized native structure) taken from a 30 μs simulation at 325 K are shown in figure 1. The protein is mainly two state, although the fraction of native contacts shows a splitting of both the native and the denatured state into two substates separated by relatively low free energy barriers.

3.2. ϕ-restrained simulations

Figure 2 shows the distributions of RMSD and Q_N for TSEs obtained with the Go model and experimental ϕ value restraints, and definitions of ϕ^{calc} which include (red/dashed line) and

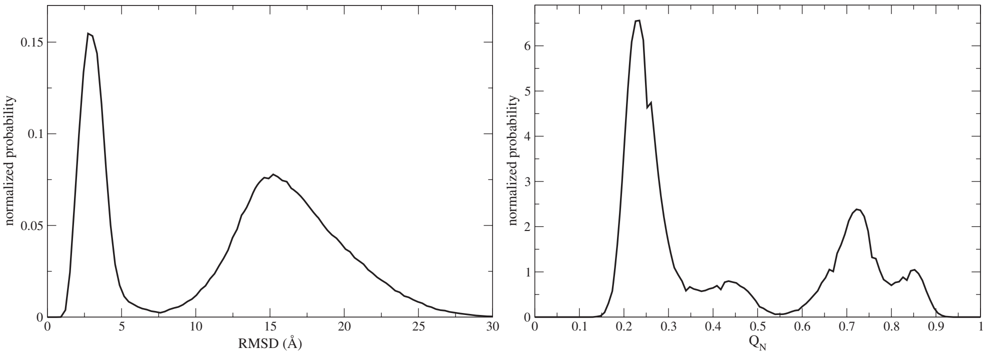

Figure 1. Distribution of the RMSD from the native structure and Q_N for an unrestrained simulation of Im9 at its mid-point.

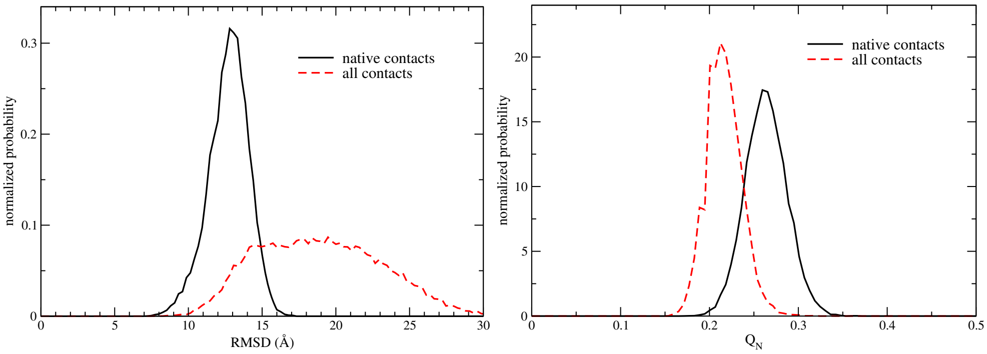

Figure 2. Distribution of the RMSD from the native structure (left) and Q_N (right) for TSEs calculated using experimental ϕ-value-restrained simulations, with ϕ interpreted as the fraction of native contacts formed (black/solid line), and fraction of all contacts formed (red/dashed). Restrained simulation data are taken from the replica at 300 K.

exclude (black/solid line) non-native contacts. Ensembles have been obtained from a set of REMD simulations between 300 and 500 K and the distributions shown are taken from the replica at 300 K. The distributions are convergent, i.e., they do not change as the simulation length (here 150 ns) is extended. Using all (native and non-native) contacts in the definition of $\phi^{\rm calc}$ broadens considerably the RMSD distribution and shifts both RMSD and Q_N towards values typical of the denatured protein. This is not completely surprising since there are many more ways of satisfying the restraints if both native and non-native contacts can contribute.

In figure 3 the values of the RMSD and Q_N for all the structures in the TSEs generated through restrained simulations are superimposed as dots onto a contour plot obtained from the equilibrium simulation at $T = T_{\rm m} = 325$ K. The Go model transition state can be approximately located in the region of no contours at $0.45 < Q_N < 0.6$ and 5 Å < RMSD < 7 Å. Both transition states from experimental ϕ values overlap more with the denatured state region of the contour plot than with the transition state region. Indeed, the restrained simulations probe the experimental, rather than the model's, transition state, and it is therefore not surprising that they do not overlap. However, the extent of the difference is surprising: given that the Go model appears to demonstrate the folding behaviour seen experimentally reasonably well for this protein, we would expect some correspondence between the two. One

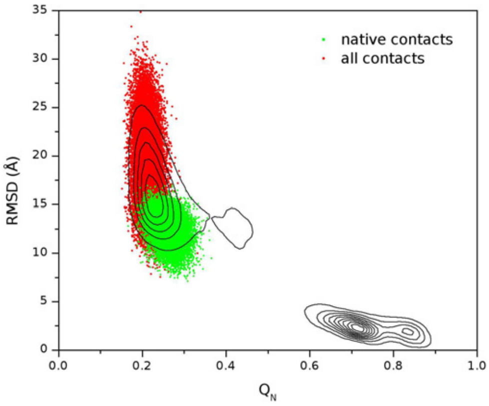

Figure 3. Contour plot showing the positions of the native and unfolded basins for the Go model of protein Im9. Data are from an unbiased simulation at 325 K. Along each line the probability of finding particular structures is constant. The difference in probability between lines is 2 × 10^{-3}. Superimposed are the TSE results at 300 K for simulations restrained with experimental ϕ values interpreted in terms of native contacts (in green/light grey) and all contacts (in red/dark grey).

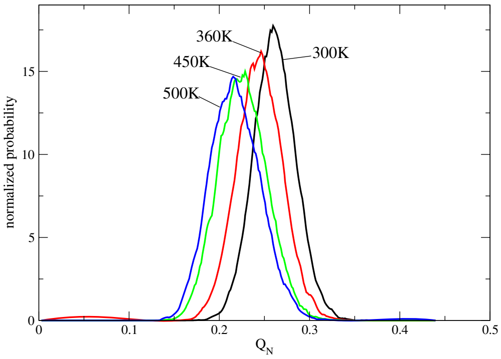

Figure 4. Distributions of Q_N for TSEs calculated using experimental ϕ-value-restrained MD at different temperatures. TSE calculated using native-contact-only definition of ϕ^{calc}.

possible reason is that the difference is due to the large cut-off used in the definition of a contact in equation (2), which may encourage sampling of more denatured states. We tested this by re-running the simulations with a lower cut-off (7.5 Å rather than 11 Å) and found that the calculated TS was still firmly in the denatured region. The finding that the model and restrained transition states are different is indeed interesting because it rules out the utility of using methods based on the calculation of p_{fold} to validate the transition-state nature of the structures found from simulations restrained with experimental ϕ values [11, 35] using a Go model.

3.3. Temperature

Whereas in unrestrained simulations temperature has a very large effect on the native contact distribution, in both sets of restrained simulations temperature appears to have only a slight effect on the distribution (figure 4). At higher temperatures the distribution is shifted towards a lower number of native contacts. However, the shift is small, and the shape of the curve

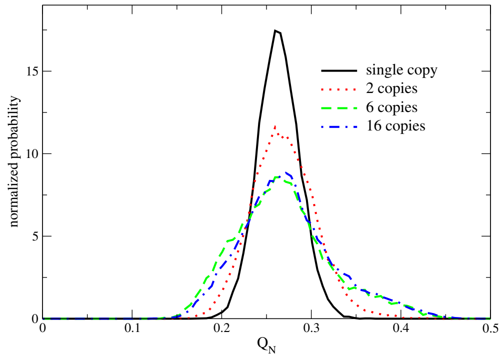

Figure 5. Distribution of Q_N for ensemble simulations restrained by experimental ϕ values. The calculated ϕ value averaged over all the copies satisfies the experimental restraints. Data for simulations with one, two, six and 16 copies are shown.

does not change; this suggests that the temperature at which sampling is performed is not particularly significant for biased simulations. A closer look at the structural properties of the ensembles shows minor structural changes in the structures, which are negligible after energy minimization. The average secondary structure is similar at all temperatures, with all four helices present (albeit frayed) in a significant proportion of the structures. As temperature increases the percentage of structures with intact helices decreases: for example, helix 3 is present in 80% of structures at 300 K and only 40% at 500 K. Another test on the similarity of the structures at the various temperatures was performed by clustering structures from the two extreme temperatures together. Using a rigorous clustering procedure where all the pairwise RMSDs are computed [36] we found that the most populated clusters (using a cut-off of 8 Å to define clusters) contain a mixture of structures from simulations at both temperatures (around 20–40% from the 500 K replica and the rest from the 300 K replica). These results demonstrate that the high sampling temperature used to generate transition state ensembles (e.g. those reviewed in [7]) is unlikely to have affected the structural properties therein described.

3.4. Conformational averaging

As described in section 2, the fact that experimental ϕ values are an average property of the ensemble of molecules probed in the experiment can be taken into account by performing N simulations simultaneously (copies), and imposing the bias on the ϕ value calculated as an average over the N copies [14]. Since we found that sampling temperature does not affect the result, and full convergence can be achieved by performing sufficiently long simulations, we performed these simulations at one temperature only (300 K).

Figure 5 shows the effect of increasing numbers of copies on the Q_N distribution. Perhaps unsurprisingly, the distribution initially broadens with ensemble size as the number of ways of satisfying the restraints increases. However, the broadening is not large and the position of the maximum of the distribution does not change, suggesting that the overall structural picture is not significantly affected by conformational averaging. The distributions converge above six copies: there is no difference in breadth between the six-copy and 16-copy distributions. This indicates that no more than six copies need to be used to represent the effect of conformational averaging effectively in this case. Our results suggest that conformational averaging is not important in determining conformations satisfying ϕ^{exp} values for Im9. This result is in agreement with the MC simulations of Paci *et al* [6], which show that, in the cases of Im9

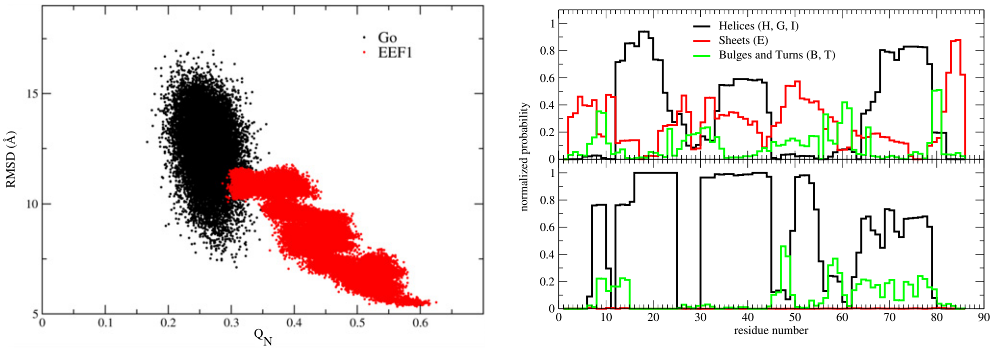

Figure 6. (Left) Scatter plot on the Q_N/RMSD plane for the TSE determined using the Go and EEF1 models. (Right) Average secondary structure for the two TSEs.

and its three-state homologue Im7, ϕ^{calc} values were unimodal for all residues for numbers of copies between 1 and 100. Other analogous studies on different proteins (see, e.g., [18, 19]) suggest that for other proteins, in particular those with β structure, conformational averaging might influence considerably the transition state picture. As we will show later, the effect of conformational averaging depends not only on the protein but also on the set of ϕ^{exp} used as restraints.

3.5. Comparison between TSEs obtained using different force fields and degrees of coarse-graining

The comparison of the putative TSEs obtained using the coarse-grained Go model and the atomistic, implicit solvent EEF1 force field is complicated by the fact that the atomic degrees of freedom differ, and that the restraints are effectively applied differently, in one case to C_α contacts and in the other to side-chain atoms.

In figure 6 the TSE structures obtained with the two models are projected onto the Q_N/RMSD plane. We stress here that in both cases identical sets of ϕ^{exp} have been used and that restraints are similarly satisfied for both models ($\langle\rho(\text{EEF1})\rangle = 0.012$ and $\langle\rho(\text{Go})\rangle = 0.013$). It is clear, however, that the two ensembles are different. With the Go model the TSE seems to be less native, while with the EEF1 model it appears somewhat less converged and reminiscent of the fact that we used native initial structures for the restrained simulations. Also in figure 6 the average secondary structures for the two models are compared. Surprisingly, the native secondary structure is very well conserved in the Go model, whereas with EEF1 the helices are conserved in a smaller fraction of structures, and helix 3 is never present. There is also some β structure in the EEF1 structures. We further tested the (dis)similarity of the two sets of structures by clustering both sets together as described in section 3.3. We found that the two sets were separated in the clustering, i.e., the most populated clusters contained either only EEF1 structures or only Go model structures (meaning that the statistically more probable structures differ by more than 8 Å RMSD).

3.6. True transition state for the Go model

As mentioned in section 2, in a 30 μs long unbiased simulation at the melting temperature T_{m}, about 80 folding and unfolding events could be observed. Using an optimized reaction coordinate following the recipe proposed by Best and Hummer [34] we could select about 100

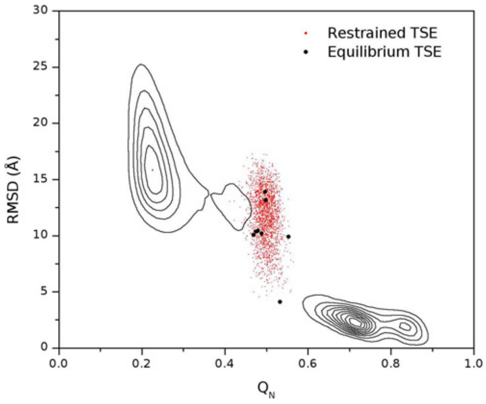

Figure 7. Q_N/RMSD scatter plot showing transition state conformations satisfying the condition $p_{\text{fold}} \simeq 0.5$ (black dots) and calculated using REMD restrained by 'simulation' ϕ values (red/grey cloud). Unrestrained data at 325 K shown as a contour plot for comparison. Along each line the probability of finding particular structures is constant. The difference in probability between lines is 2×10^{-3}.

conformations in the transition region. We tested the p_{fold} of 24 of them and eight were found to have $0.45 < p_{\text{fold}} < 0.55$; these structures are shown as black dots in figure 7, where the usual projection on the Q_N/RMSD coordinates is used, and the contours represent the result from the unbiased simulation at T_{m}. From the eight true transition structures we computed ϕ values (ϕ^{Go}) for all 86 residues using equation (2). These ϕ^{Go} were in turn used as restraints to re-determine the transition state ensemble. The cloud of red/grey dots represents the structures thus obtained. The two samples occupy the same region in the projection over the Q_N/RMSD pair of variables ($Q_N \approx 0.5$ and 5 Å $<$ RMSD $<$ 15 Å). However, the red/grey dots appear to extend beyond the TS region, to RMSDs of 17 Å. This could either be because restraints are not sufficient to define the TS region, or because Q_N/RMSD are not ideal coordinates to describe the folding reaction. To remove this doubt, we calculated p_{fold} for a random sample of 30 of the structures found in the ϕ^{Go} restrained simulations. For all 30 structures p_{fold} was between 0.07 and 0.18: clearly, these structures are not true transition states, and are closer to the denatured state than the native state.

The availability of true ϕ^{Go} values for the model, and thus the possibility of using the restraint method self-consistently, allowed us to investigate conformational averaging further. When used as restraints, ϕ^{exp} and ϕ^{Go} define TSEs located in vastly different regions of the Go model energy landscape: the TS generated using ϕ^{exp} lies firmly in the relatively flat denatured state region, whereas using ϕ^{Go} as restraints results in structures in the region of a large energy barrier. This variance in underlying energy landscape could result in the effects of conformational averaging being significantly different according to which set of ϕ are used as restraints. To test this, we repeated the ensemble simulations, this time using the ϕ^{Go} as restraints. The results, shown in figure 8, differ significantly from those presented earlier. The effect of including only one extra copy is a splitting of the distribution such that the native state is now also sampled. As the number of copies increases the distribution becomes more and more like the equilibrium distribution. This is because the copies can satisfy the ϕ^{Go} values as an average without suffering the energetic penalty of being located at an energy maximum, as has to be the case for a single copy. These results show that, depending on the choice of ϕ values, conformational averaging can lead to an almost total loss of the information contained in the experimental restraints. The loss of information appears to be

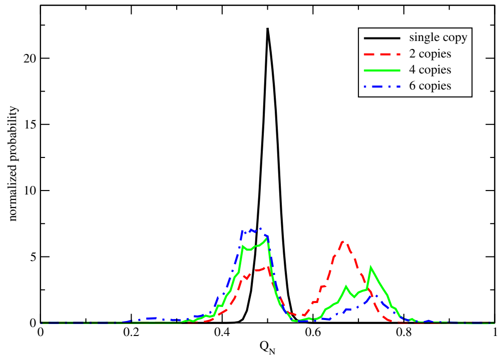

Figure 8. Distribution of Q_N for ensemble simulations using simulation ϕ values as restraints. The calculated ϕ value averaged over all the copies satisfies the restraints. Data for simulations with one, two, four and six copies are shown.

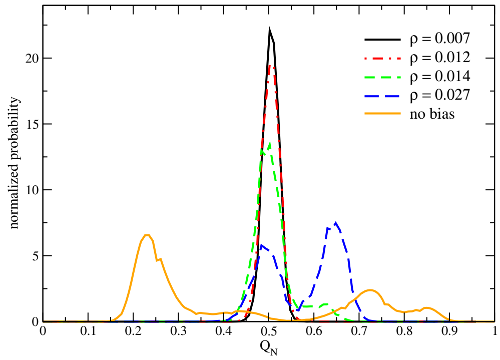

Figure 9. Distribution of Q_N for TSEs calculated using simulation ϕ values, various values of α and $T = 300$ K, with the unrestrained distribution at $T_{\rm m} = 325$ K as a comparison.

particularly large if the TS described by the ϕ values coincides with the TS for the underlying force field.

Another interesting issue that can be tested using the $\phi^{\rm Go}$ values is the effect of changing bias strength, α, on the ensemble obtained from restrained simulations. We interpret our results in terms of ρ rather than α, as the value of α will vary significantly with different proteins and force fields. The value of ρ depends only on how well the restraints are satisfied; as α increases ρ decreases (the square root of ρ is also the average root mean square deviation between experimental and calculated ϕ values). Figure 9 shows the Q_N distributions for TSEs generated using different bias strengths. At $\rho \leqslant 0.012$ only the transition state region is sampled, i.e. the bias is sufficiently large for the restraints to be satisfied. When the restraints are more poorly satisfied ($\rho \geqslant 0.014$), however, a broader region is also sampled, indicating that the bias is not large enough here. With $\rho \geqslant 0.014$ the TSE 'spills' into the native state because the temperature of the restrained simulations (300 K) is well below $T_{\rm m}$ (325 K). At 330 K, in the restrained simulations for the same larger values of ρ, a 'spilling' in both the native and the denatured state is evident.

The ideal bias is large enough to restrain the protein within the transition state region; excessively large biases, however, will prevent efficient sampling by introducing large energy barriers. The value $\rho = 0.012$ is a good value in this case because it provides a well defined ensemble of structures (and, in fact, corresponds to the value of $\alpha = 5 \times 10^5$ used throughout the present work). Assuming the deviation from the experimental ϕ is approximately equal for each residue, this means (equation (4)) that each ϕ^{calc} should be within about 0.11 of the ϕ^{exp}. The experimental uncertainty involved in measurement of ϕ values is often greater than this [25]. These results suggest that experimental ϕ values with an error bar larger than 0.1 might not be useful in structurally defining the TSE.

4. Discussion

We have dissected the problem of determining protein structures compatible with experimental restraints [3, 4]. Most of our conclusions apply to circumstances in which the available experimental restraints are sparse, which is the case for non-native states which are intrinsically disordered. The method has been broadly used in the past few years and has been demonstrated to be valuable for interpreting experimental data in terms of structure and heterogeneity [5–16, 37, 38]. The specific case analysed here is that of the transition state of the all-α protein Im9, for which a large number of reliable experimental ϕ values have been determined. The method provides an ensemble of structures which are 'compatible' with a microscopic interpretation of the ϕ values. Among the issues that we have explored are the importance of the sampling method and sampling temperature, conformational averaging, alternative microscopic definitions of ϕ^{calc}, the underlying force field and the magnitude of the bias required.

One finding is that the replica exchange method enhances considerably the restrained sampling, whilst providing samples at different temperatures. The sampling temperature does not have a significant effect on the structural properties of the calculated ensemble. This indicates that the calculation of TSEs using a range of temperatures to enhance the sampling, as broadly done previously following the protocol suggested by Paci *et al* [4], does not affect the final result when sampling is performed to full convergence.

On the other hand, inclusion of non-native contacts in the definition of ϕ^{calc} values changes the structures of the putative TSE considerably; distributions of order parameters such as Q_N and RMSD are significantly broadened and shifted towards more denatured structures. The importance of non-native interactions in the TS varies greatly between proteins, and for the Im proteins they are thought to be particularly significant [39]. Whilst ideally non-native interactions should be included in restrained simulations using ϕ values, in practice considering native and non-native contacts to be equivalent leads to the loss of information contained in the experimental ϕ values.

Conformational averaging changes the distributions of Q_N and RMSD: they broaden slightly as increasing numbers of copies are simultaneously simulated. Convergence of distributions occurs above six copies. This result is specific to Im9: other simulations indicate that conformational averaging may be more important for other proteins. The result also depends strongly on the set of ϕ^{exp} used as restraints.

The underlying force field and the degree of coarse-graining strongly influence the features of the calculated putative TSE. The structures produced using a C_α Go model are considerably different from those obtained, using the same identical ϕ values as restraints, with an all-atom implicit-solvent model. The underlying model is important since the information contained in a sparse set of ϕ^{exp} is clearly not sufficient to picture the TS.

Finally, using ϕ values calculated from the true transition state for the Go model (i.e., a set of structures which have $p_{\text{fold}} \simeq 0.5$), we determined the corresponding ensemble using restrained simulations, thus for the first time self-consistently testing how close the restrained ensemble is to the real one without relying on a microscopic interpretation of the ϕ^{exp}. We found that the structures determined from restrained simulations at 300 K have a p_{fold} close to zero. Although these structures are clearly not true transition states, they still share the relevant structural features of the transition state. This, however, is only the case if only a single copy of the molecule is used: conformational averaging results in a loss of the information contained in the ϕ^{exp}, with the distribution of Q_N approaching the equilibrium distribution as more copies are added. It is also only the case if the restraints are satisfied with a very small tolerance (i.e., if the harmonic constant α in the restraint term in the Hamiltonian is very large). We found that $\alpha = 5 \times 10^5$, corresponding to $\rho = 0.012$, was just sufficient. This corresponds to a deviation of ~ 0.1 on each ϕ^{calc} from ϕ^{exp}. If the tolerance on the restraints is larger, the ensemble is no longer well defined and structures belonging to the native and denatured basins will also be sampled. This is an important result, and is particularly relevant if one considers that the statistical error on the experimental ϕ values can be larger than 0.1.

One other issue we have not addressed here is that of how accurate the assumption that experimental ϕ values are equivalent to the fraction of native contacts formed at the transition state is. At best this will introduce a small systematic error, which should still be considered behind the experimental statistical error when using ϕ values as restraints. All these considerations are also relevant when using the method to sample other protein states using different experimental restraints.

Acknowledgments

We thank Xavier Periole for interesting discussions and for critically reading the manuscript, Robert Best for support in using the ENSEMBLE module of CHARMM and Michael Schaefer for providing the source of his CLUSTER program. LRA acknowledges EPSRC for a PhD studentship.

References

[1] Fersht A R 1999 *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (New York: Freeman)

[2] Sato S, Religa T L, Daggett V and Fersht A R 2004 Testing protein-folding simulations by experiment: B domain of protein *A. Proc. Natl Acad. Sci. USA* **101** 6952–6

[3] Vendruscolo M, Paci E, Dobson C M and Karplus M 2001 Three key residues form a critical contact network in a protein folding transition state *Nature* **409** 641–5

[4] Paci E, Vendruscolo M, Dobson C M and Karplus M 2002 Determination of a transition state at atomic resolution from protein engineering data *J. Mol. Biol.* **324** 151–63

[5] Paci E, Clarke J, Steward A, Vendruscolo M and 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** 394–9

[6] Paci E, Friel C T, Lindorff-Larsen K, Radford S E, Karplus M and Vendruscolo M 2004 Comparison of the transition states ensembles for folding of Im7 and Im9 determined using all-atom molecular dynamics simulations with value restraints *Proteins* **54** 513–25

[7] Paci E, Lindorff-Larsen K, Karplus M, Dobson C M and Vendruscolo M 2005 Transition state contact orders correlate with protein folding rates *J. Mol. Biol.* **352** 495–500

[8] Lindorff-Larsen K, Paci E, Serrano L, Dobson C M and Vendruscolo M 2003 Calculation of mutational free energy changes in transition states for protein folding *Biophys. J.* **85** 1207–14

[9] Lindorff-Larsen K, Vendruscolo M, Paci E and Dobson C M 2004 Transition states for protein folding have native topologies despite high structural variability *Nat. Struct. Mol. Biol.* **11** 443–9

[10] Lindorff-Larsen K, Røgen P, Paci E, Vendruscolo M and Dobson C M 2005 Protein folding and the organization of the protein topology universe *Trends Biochem. Sci.* **30** 13–9

[11] Hubner I A, Shimada J and Shakhnovich E I 2004 Commitment and nucleation in the protein G transition state *J. Mol. Biol.* **336** 745–61

[12] Vendruscolo M, Paci E, Karplus M and Dobson C M 2003 Rare fluctuations of native proteins sampled during equilibrium hydrogen exchange *J. Am. Chem. Soc.* **125** 15686–7

[13] Vendruscolo M, Paci E, Karplus M and Dobson C M 2003 Structures and relative free energies of partially folded states of proteins *Proc. Natl Acad. Sci. USA* **100** 14817–21

[14] Best R B and Vendruscolo M 2004 Determination of protein structures consistent with NMR order parameters *J. Am. Chem. Soc.* **126** 8090–1

[15] Gsponer J, Hopearuoho H, Whittaker S B-M, Spence G R, Moore G R, Paci E, Radford S E and Vendruscolo M 2006 Determination of an ensemble of structures representing the intermediate state of the bacterial immunity protein Im7 *Proc. Natl Acad. Sci. USA* **103** 99–104

[16] Best R B and Vendruscolo M 2006 Structural interpretation of hydrogen exchange protection factors in proteins: characterization of the native state fluctuations of CI2 *Structure* **14** 97–106

[17] Guerois R, Nielsen J E and Serrano L 2002 Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations *J. Mol. Biol.* **320** 369–87

[18] Davis R, Dobson C M and Vendruscolo M 2002 Determination of the structures of distinct transition state ensembles for a-sheet peptide with parallel folding pathways *J. Chem. Phys.* **17** 9510–7

[19] Geierhaas C D, Best R B, Paci E, Vendruscolo M and Clarke J 2006 Structural comparison of the two alternative transition states for folding of TI I27 *Biophys. J.* **91** 263–75

[20] Taketomi H, Ueda Y and Gō N 1975 Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions *Int. J. Pept. Protein Res.* **7** 445–59

[21] Shea J E and Brooks C L III 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

[22] Paci E, Vendruscolo M and Karplus M 2002 Validity of G models: Comparison with a solvent-o shielded empirical energy decomposition *Biophys. J.* **83** 3032–8

[23] Prieto L, de Sancho D and Rey A 2005 Thermodynamics of Go-type models for protein folding *J. Chem. Phys.* **123** 154903

[24] Cavalli A, Vendruscolo M and Paci E 2005 Comparison of sequence-based and structure-based energy functions for the reversible folding of a peptide *Biophys. J.* **88** 3158–66

[25] Sanchez I E and Kiefhaber T 2003 Origin of unusual small ϕ-values in protein folding: evidence against specific nucleation sites *J. Mol. Biol.* **334** 1077–85

[26] Du R, Pande V S, Grosberg A Yu, Tanaka T and Shakhnovich E I 1998 On the transition coordinate for protein folding *J. Chem. Phys.* **108** 334–50

[27] Sugita Y and Okamoto Y 1999 Replica-exchange molecular dynamics method for protein folding *Chem. Phys. Lett.* **314** 141–51

[28] Osborne M J, Breeze A L, Lian L Y, Reilly A, James R, Kleanthous C and Moore G R 1996 Three-dimensional solution structure and 13C nuclear magnetic resonance assignments of the colicin E9 immunity protein Im9 *Biochemistry* **35** 9505–12

[29] Friel C T, Capaldi A P and Radford S E 2003 Structural analysis of the rate-limiting transition states in the folding of Im7 and Im9: similarities and differences in the folding of homologous proteins *J. Mol. Biol.* **326** 293–305

[30] Karanicolas J and Brooks C L III 2003 The structural basis for biphasic kinetics in the folding of the WW domain from a formin-binding protein: lessons for protein design? *Proc. Natl Acad. Sci. USA* **100** 3954–9

[31] Brooks B R, Bruccoleri R E, Olafson B D, States D J, Swaminathan S and Karplus M 1983 CHARMM: a program for macromolecular energy, minimization and dynamics calculations *J. Comput. Chem.* **4** 187–217

[32] Lazaridis T and Karplus M 1999 Effective energy function for protein dynamics and thermodynamics *Proteins* **35** 133–52

[33] Paci E and Karplus M 1999 Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations *J. Mol. Biol.* **288** 441–59

[34] Best R B and Hummer G 2005 Reaction coordinates and rates from transition paths *Proc. Natl Acad. Sci. USA* **102** 6732–7

[35] Hubner I A, Oliveberg M and Shakhnovich E I 2004 Simulation, experiment, and evolution: understanding nucleation in protein S6 folding *Proc. Natl Acad. Sci. USA* **101** 8354–9

[36] Cavalli A, Haberthür U, Paci E and Caflisch A 2003 Fast protein folding on downhill energy landscape *Prot. Sci.* **12** 1801–3

[37] Korzhnev D M, Salvatella X, Vendruscolo M, Di Nardo A A, Davidson A R, Dobson C M and Kay L E 2004 Low-populated folding intermediates of Fyn SH3 characterized by relaxation dispersion NMR *Nature* **430** 586–90

[38] Salvatella X, Dobson C M, Fersht A R and Vendruscolo M 2005 Determination of the folding transition states of barnase by using PhiI-value-restrained simulations validated by double mutant PhiIJ-values *Proc. Natl Acad. Sci. USA* **102** 12389–94

[39] Capaldi A P, Kleanthous C and Radford S E 2002 Im7 folding mechanism: misfolding on a path to the native state *Nat. Struct. Biol.* **9** 209–16