We present a method for determining the structure of the transition state ensemble (TSE) of a protein by using $\phi$ values derived from protein engineering experiments as restraints in molecular dynamics simulations employing a realistic all-atom molecular mechanics energy function. The method uses a biasing potential to select an ensemble of structures having $\phi$ values in agreement with the experimental data set. An application to acylphosphatase (AcP), a protein for which $\phi$ values have been measured for 24 out of 98 residues, illustrates the approach. The properties of the TSE determined in this way are compared with those of a coarse-grained model obtained using a Monte Carlo (MC) sampling method based on a C$^\alpha$ representation of the structure. The two TSEs determined at different structural resolution are consistent and complementary. While the C$^\alpha$ model allows better sampling of the conformation space occupied by the transition state, the all-atom model offers a more detailed description of the structural and energetic properties of the conformations included in the TSE. The combination of low-resolution C$^\alpha$ results with all-atom molecular dynamics simulations provides a powerful and general method for determining the nature of TSEs from protein engineering data.

Keywords: protein folding; molecular dynamics; $\phi$ values; AcP; EEEF1 potential
an adequate sample of the pseudo-equilibrium distribution of conformations corresponding to the TSE; that is, to sample efficiently the equilibrium distribution of the pseudo-energy function that includes the $\phi$ value restraints. In our initial work, the TSE of acylphosphatase (AcP) was determined by MC sampling of a C$^\alpha$ model, in which the residues were represented by spheres connected by rigid bonds of length 3.8 Å. The resulting TSE predicted the $\phi$ values for residues for which mutations had not been performed and provided a check for the self-consistency of the measured $\phi$ values. Of particular importance was the finding that the members of the TSE have the same overall topology as the native state and that this topology is defined by a very small number of residues that form key interactions during folding.

The simplicity of the model used in the early simulations was important for a full sampling of the TSE, but much atomic detail was missing because of the use of a C$^\alpha$ representation and a simplified potential function. It is the purpose of the present study to complement the C$^\alpha$ determination of the TSE with results based on molecular dynamics simulations using an all-atom model of the protein$^7$ and an implicit model for the solvent.$^8$ The potential function, called EEF1, provides a potential-of-mean-force description of the solvent and has been shown to discriminate between native structures and misfolded decoys.$^9$ It has been used to simulate high-temperature unfolding of proteins$^{10}$ and has been shown to be particularly useful for obtaining a realistic description of large-scale conformational changes, such as those induced by external forces,$^{11}$ from simulations carried out on a time-scale that is short compared to the solvent relaxation time.$^{12}$

Here, we generate the TSE structure using a biased molecular dynamics (BMD) method in which a time-dependent bias, based on the experimental $\phi$ values, is added to the EEF1 energy function. The bias forces the structures of the protein in a simulation away from the native state towards a region of conformation space where the pairwise contact probabilities between residues correspond to those estimated from the experimental $\phi$ values. Using this procedure we have generated a set of low-energy structures that represent the TSE of AcP, enabling us to analyze its properties and make comparison with the TSE obtained in the previous MC simulation.$^5$ Although the microscopic definition of $\phi$ values used here is based on atomic distances and not on free energy differences ratios, a recent analysis has shown that this approach is a good approximation for transition states that are relatively close to the native state, such as that of AcP.$^4$

Recently, Li & Shakhnovich$^{13}$ have applied our approach$^5$ to determine the TSE of C12 using the protein engineering data of Itzhaki et al.$^{14}$ They used MC sampling, as done by Vendruscolo et al.$^5$ and replaced the C$^\alpha$ model by an all-atom Go-type model. Here, we use a molecular mechanics energy function, which has the full complexity inherent in the energy surface of a protein. Moreover, as the present method has been introduced to provide an experimental description of the structure of the transition state, it is complementary to methods that seek to predict such structures purely by means of molecular dynamics simulations.$^{15,16}$ In this sense we believe that the present approach will contribute to enhance further the synergy between theory and experiment that has contributed crucially to the recent progress in understanding the fundamental elements of the process of protein folding.$^{17,18}$

Theory and Methods

System and model

AcP is a 98 residue single-domain protein having an $\alpha/\beta$ structure, with $\beta$-strands and two $\alpha$-helices (Figure 1). It has been shown to fold with two-state kinetics under a wide range of conditions.$^{19,20}$ A total of 26 $\phi$ values are available for AcP from horse muscle from the work of Chiti et al.$^{21}$ Two of these correspond to different mutations of the same residue, Y11F ($\phi = 0.93$) and Y11I ($\phi = 0.83$); only the first of those values is used in the following analysis. We disregarded the $\phi$ value for one mutation, Y25A, because its experimental $\phi$ value is larger than unity and thus not directly interpretable.$^2$ This leaves a set of 24 experimental

![Figure 1. NMR structure of AcP. The secondary structure elements are: $\beta_1$ (residues 7–13), $\alpha_1$ (residues 22–33), $\beta_2$ (residues 36–42), $\beta_3$ (residues 46–53), $\alpha_2$ (residues 55–65), $\beta_4$ (residues 77–85), $\beta_5$ (residues 93–97); residues between these regions are parts of loops. The secondary structure is represented as ribbon and computed with DSSP.$^{40}$ $\beta$-Strands are represented in green, $\alpha$-helices are in red and loops in blue. The Figure was drawn with the program MOLMOL.$^{41}$]
\( \Phi \) values, the same set as that used by Vendruscolo et al.\(^5\)

The solution structure of horse muscle AcP, which is the protein studied by Chiti et al.,\(^21\) has been determined by NMR\(^22\) (entry 1APS in PDB) and five structures are listed for the native state, all compatible with the experimental restraints. The \( C^a \) RMSD between any pair of structures is between 2 Å and 3 Å (for \( C^a \)). We have used the same NMR model (model 1) as the reference structure in the present study, as that is the one chosen by Vendruscolo et al.\(^5\)

In the molecular dynamics simulations the polar hydrogen model\(^7\) in the CHARMM program\(^6\) was used for the protein and an implicit Gaussian exclusion model for the solvent.\(^6\) High-frequency vibrational modes involving bonds to hydrogen atoms were frozen out by means of holonomic constraints through the SHAKE algorithm;\(^23\) this leads to good conservation of the constants of motion with a time-step of 2 fs.\(^24\)

**Microscopic definition of \( \Phi \) values**

For a configuration at time \( t \), we define the calculated \( \Phi \) value of residue \( I \) as:

\[
\Phi_{I}^{\text{calc}}(t) = \frac{\mathcal{N}_{I}^{\text{calc}}(t)}{\mathcal{N}_{I}^{\text{nat}}} \tag{1}
\]

where \( \mathcal{N}_{I} \) the number of native contacts made by residue \( I \), is given by:

\[
\mathcal{N}_{I} = \sum_{i \in I} \sum_{j \in I} \psi(r_{ij} - r_{c}) \Delta_{ij}(Q) \tag{2}
\]

with:

\[
\psi(r) = \frac{1}{1 + e^{\beta r}} \tag{3}
\]

where \( M \) is the number of atoms in the protein (or the number of selected atoms, e.g. \( C^a \) atoms, heavy side-chain atoms), \( r_{ij} \) the distance between the atoms \( i \) and \( j \) and \( \Delta_{ij}(Q) \) is equal to 1 if atoms \( i \) and \( j \) are closer than a cut-off distance \( r_c \) in the reference structure and belong to residues at least \( Q \) residues apart in the sequence, and is equal to zero, otherwise. The definition given in equation (2) is flexible and appropriate for different choices of the atoms to be considered. For example, if only \( C^a \) atoms are chosen, \( \mathcal{N}_{I} \) counts the number of residues which are in contact with \( I \) based on the \( C^a-C^a \) distance. This definition was used in earlier calculations.\(^4\) Here, all the heavy atoms of the side-chains are included and \( \mathcal{N}_{I} \) counts all the contacts made by each of its side-chain atoms with any side-chain atom of any other residue. A definition based on side-chains is appropriate, since experimental \( \Phi \) values are primarily a measure of the loss of side-chain contacts at the transition state, relative to the native state. 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.\(^23\)

In the definition of \( \mathcal{N}_{I} \) we use a step function (equation (3)) because it is continuous and allows the evaluation of the gradient of \( \mathcal{N}_{I}^{\text{calc}}(t) \) with respect to the atomic coordinates. The specific value of \( \beta \), within a range 2–8 Å\(^{-1}\) has little effect on the results, although too large a \( \beta \) would cause the trajectory to be unstable; we used \( \beta = 5 \text{ Å}^{-1} \).

For the set of structures that represent the TSE, the experimental \( \Phi \) values (\( \Phi_{I}^{\text{exp}} \)) can be compared with those calculated by averaging over the TSE; that is:

\[
\langle \Phi_{I}^{\text{calc}} \rangle = \frac{\langle \mathcal{N}_{I}^{\text{TSE}} \rangle}{\mathcal{N}_{I}^{\text{nat}}} \tag{4}
\]

Equation (4) is similar to the previously used definitions.\(^30,32,26\) Simply counting contacts, calculating their energies is an approximation.\(^7\) Shea et al.\(^26\) have found in their model calculations that this approximation for estimating \( \Phi \) values from structures is a good one under certain conditions, which appear to be satisfied by AcP.\(^7\) A more detailed relation between experimental \( \Phi \) 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 residue.

The comparison between \( \Phi_{I}^{\text{calc}} \) and \( \Phi_{I}^{\text{exp}} \) provides a measure of the compatibility of the experimental TSE and an ensemble of structures generated by simulation. A good correlation between the two sets of values indicates that the TSE determined by molecular dynamics is consistent with the experimental data. If enough \( \Phi_{I}^{\text{exp}} \) values are available, as appears to be the case for AcP,\(^8\) the calculated TSE should resemble closely the true transition state. In this case, a detailed study of the properties of the transition state and the interactions between different regions of the polypeptide chain can be carried out.

**The bias for sampling the transition state**

The procedure we use here is conceptually related to that used to generate structures of the native state compatible with NOE measurements from NMR experiments, in that pseudo-energy terms involving experimental restraints are added to the protein force field.\(^25\) The sets of structures representing the NSE determined in this way are quite similar, provided sufficient NMR data are available; in the case of AcP, for example, the mean pairwise RMSD is 2.6(±0.3) Å for the five structures in the PDB. For the TSE, we expect a much larger RMSD range for the contributing structures, as only a subset of the specific native state interactions is present at this stage in folding. The \( C^a \) representatives of the TSE for AcP determined by Vendruscolo et al.\(^5\) had an average pairwise RMSD of 5.7(±1.4) Å. Thus, what is required in the present simulation is a method to sample a broad set of structures compatible with a set of
experimental restraints, rather than the search for an essentially unique structure.

Unlike the native state which is (or is close to) the minimum energy structure, the TSE consists of unstable structures corresponding to the highest free energy region on the reaction coordinate between the denatured state and the native state. The probability of observing directly structures corresponding to this region is negligible. Thus, in order to make molecular dynamics sampling of the TSE possible, a “pseudo-energy” based on experimental data is added to the force field to make the members of the TSE the most stable structure. The restraints used here are based on measured values; these values provide information of a more stochastic nature about the TSE than do NOE distance restraints for the native state. The comparison with the fuller sampling, obtained from the C model and the MC approach, shows, however, that the two TSEs are very similar (see below).

The method for determining the TSE is implemented by introducing a small energy perturbation that forces the system to follow trajectories which, starting at the native state, lead to decreasing deviations between the experimental and calculated values. For this purpose we introduce the quantity:

$$\rho(t) = \frac{1}{N_0} \sum_{i=1}^{N_0} (\phi_{i}^{calc}(t) - \phi_{i}^{exp})^2$$

where $E$ is the list of $N_0$ available experimental $\phi$ values, $\phi_{i}^{exp}$. The quantity $\rho$ is the mean-square deviation between $\phi_{i}^{exp}$ and $\phi_{i}^{calc}$ and is related to the cost function that was used to define the “minimal model” described by Vendruscolo et al.

To sample regions of configurational space corresponding to low values of $\rho$, the BMD method is used. It supplements the molecular potential energy function with a bias of the form:

$$W(r, t) = \begin{cases} \frac{\alpha}{2} (r - r_0)^2 & \text{if } \rho(t) \geq \rho_a \\ 0 & \text{if } \rho(t) < \rho_a \end{cases}$$

where:

$$\rho_a(t) = \min_{0 \leq t \leq 1} \rho(\tau)$$

The bias depends on the time through $\rho$. The BMD approach has been discussed in detail elsewhere, and we do not repeat the description here. It can be considered to introduce a “Maxwell demon” or a soft “molecular ratchet” to bias the trajectories toward the desired state.

Details of the simulations

Simulations were initiated from the experimental native state structure (Figure 1) after a short energy minimization (400 steepest descent steps) to remove bad contacts. The system was heated from 0 K to 300 K during a time-period of 400 ps and then equilibrated in the canonical ensemble for 3 ns. The simulation yielded an average RMSD of 3.4 Å from the initial structure. The last conformation of the native simulation was used as the initial conformation for the BMD simulation.

Generation of the TSE

The TSE is obtained from BMD simulations with the coupling parameter $\alpha$ in equation (6) chosen as follows: initially $\alpha$ was set equal to 10,000 and the execution was doubled every 400 ps for a total simulation time of 4 ns. A harmonic term in the potential was then applied to keep $\rho$ near zero and a 1 ns simulation was performed. The conformations sampled are characterized by $\rho = 5 \times 10^{-3}$ and a correlation between $\phi_{TSE}$ and $\phi_{calc}$ of 0.97. However, this simulation samples only a fraction of the structures of the TSE. For this reason, we also performed higher temperature simulations to explore a larger fraction of the conformation space, compatible with the condition $\phi_{calc} = \phi_{exp}$. Series of simulations 1 ns in length were performed at five temperatures between 300 K and 780 K (300, 360, 500, 640 and 780 K). The temperatures used in the sampling do not correspond to physical temperatures because there is an additional term in the energy function that is introduced to constrain the sampling to the region of conformation space compatible with the experimental $\phi$ values. By varying the temperature we can modify the role played by the native interactions. For example, at low temperature the native features of the transition state are likely to be over-emphasized even in the presence of the restraints, whilst the states sampled at very high temperatures will be too non-native. Here, the role of the temperature is analogous to the scale factor used in simulated annealing refinement of X-ray data where the relative contributions to the effective energy of the molecular mechanics energy function and the X-ray data can be varied. It might seem appropriate to use the $P(1/2)$ criterion to test the TSE ensemble. The conceptual and practical problems with the use of $P(1/2)$ when experimental restraints are applied to determine the TSE are described in the Appendix.

Selection of representative structures

The structural analysis of the TSE was carried out on a reduced set of conformations selected as follows. Five thousand conformations were extracted, one every 1 ps of each of the five 1 ns trajectories and the C$^\alpha$ RMSD between all pairs of structures was computed. The structure with the largest number of neighbors within a 3 Å cutoff was selected as the first cluster center, with the cluster including all the structures within the cutoff. The procedure was repeated with the remaining structures until all the structures were attributed to a cluster; some clusters consisted of a single structure. The central structure of clusters
Table 1. Properties of the Transition State for AcP

<table>
<thead>
<tr>
<th>$N_c$</th>
<th>$N_s$</th>
<th>$N(\phi_{exp})$</th>
<th>$\langle \phi_{calc} \rangle$</th>
<th>$\langle \Delta \phi \rangle$</th>
<th>$\sigma$</th>
<th>RMSD (Å)</th>
<th>$R_g$ (Å)</th>
<th>$\Delta R_g$ (%)</th>
<th>S (Å$^2$)</th>
<th>$\Delta S$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td>1270</td>
<td>24</td>
<td>0.30</td>
<td>0.26 ± 0.02</td>
<td>0.99</td>
<td>6.4 (1.0)</td>
<td>13.9 (0.5)</td>
<td>3.8</td>
<td>7300 (700)</td>
<td>23</td>
</tr>
<tr>
<td>41</td>
<td>314</td>
<td>3</td>
<td>0.30</td>
<td>0.26 ± 0.01</td>
<td>0.71</td>
<td>6.7 (1.0)</td>
<td>14.4 (0.4)</td>
<td>7.8</td>
<td>7800 (400)</td>
<td>32</td>
</tr>
</tbody>
</table>

$N_c$ is the number of structures derived from the 5000 generated structures, which make up the set of the transition state discussed here; $N_s$ is the number of representative cluster centers. The average $\langle \phi_{exp} \rangle$ is computed using $N(\phi_{exp}) = 22$ residues, while the average $\langle \phi_{calc} \rangle$ is computed for all the 87 residues which have a non-zero number of side-chain native contacts (also when only three key residues are used to determine the transition state). RMSD is the average root-mean-square deviation from the native structure. RMSD$_{max}$ is the maximum RMSD between any pair of structures in the ensemble, and $\Delta R_g$ and $\Delta S$ are the variations in the radius of gyration and solvent-accessible surface relative to the native state, respectively.

Results

Native state properties

The simulations of the NSE of AcP (using only EE1 without restraints) indicated that the native protein undergoes relatively large structural fluctuations. This is in accord with the fact that the pairwise RMSD between the five models in the PDB is about 3 Å. The structures from the native state simulation are within an RMSD of 2.0–3.5 Å of the chosen native structure, depending on the model used in the restrained MD simulations. One possibility is based on the comparison of the solvent-accessible surface area using the approximate relationship between the area and the experimental $m$ value. 3,10,35

Key residues

The nucleation mechanism for protein folding involves the formation of a specific set of native interactions between certain amino acid residues. 2,35 These native interactions determine the fold or “architecture” of the native state. The amino acid residues that are part of the folding nucleus are key ones in the sense that the formation of their interactions determines a large fraction of the network of interactions in the native state. In a previous paper, we have shown that key residues can be determined by finding the minimal set of $\phi$ values required a bias for finding the structures corresponding to the TSE. 5 We also proposed a measure, the “betweenness” that detects the centrality of a residue in the network of interactions that are present in the TSE, residues with the largest betweenness are those that are critical in determining the topology of the network. Hence, the betweenness provides an additional criterion for the key residues. Both methods show that in AcP residues Y11, P54 and F94 are key ones. We thus repeated the calculations described above using as experimental restraints only the experimental $\phi$ values of these three key residues.

The transition state ensemble

In Table 1 are shown average properties of the TSE generated as described above, using as experimental restraints either all 24 experimental $\phi$ values available (although only 22 were used, since V17 and G45 make no side-chain contacts in the native state) or only the three key residues. The correlation (on 22 data points) between $\phi_{exp}$ and $\phi_{calc}$ is 0.99 for the “all-residue” case and 0.71 for the “key-only” one. The average (C$^\alpha$) RMSD from the native structure is about 6.5 Å for both the “all-atom” and key-only ensembles. The distribution of pairwise RMSD between members of the ensembles is peaked around 5.4 Å (5.8 Å), while the maximum is 7.8 Å (10.0 Å) in the all-residue (key only); this gives a measure of the broadness of the TS ensemble. The radius of gyration $R_g$ and the solvent-accessible surface suggest that the TS ensemble is moderately expanded relative to the native state. The key-only model of the TSE has very similar properties to the all-residue one, while being moderately broader and more expanded.

The two TSE ensembles generated have similar properties to the ensemble generated by MC for a C$^\alpha$ model. 5 The solvent-accessible surfaces of the TSE obtained by MD and MC are identical within the standard deviation (the solvent-accessible
surfaces were obtained using only Cα atoms in both cases and attributing a radius of 3.1 Å to the spherical probe). This is important because the accessible surface was used in the MC simulation to determine the TSE. The \( R_g \) and the RMSD from the native state structure are similar in the MD and MC simulations, although in the latter these quantities are slightly larger and more broadly distributed.

The prediction of \( \phi \) values for residues which have not been measured experimentally is one of the important applications of the present method. In Figure 2, we show a plot of the calculated \( \phi \) values. Shaded areas, in blue and green, respectively, represent \( \pm 1 \) standard deviations; the extent of the white region is thus a measure of the uncertainty of the estimated values. The agreement between experimental and calculated \( \phi \) values is very good, as expected from the high correlation between the two sets of data. Residues whose calculated \( \phi \) values vary substantially are those for which the prediction has the largest error. That is, the experimental \( \phi \) values do not restrain sufficiently the number of possible conformations and the underlying energy function plays an important role. Thus, the present procedure is able to pick out the residues whose experimental \( \phi \) values could add valuable information to our knowledge of the transition state. In the present case, the only residues which show a considerable variation in \( \phi^{cal} \), and at the same time are not totally exposed to the solvent and make a non-negligible number of native contacts, are Arg23 and Glu27. These residues are at the C-terminal end and in the middle of helix \( \alpha_1 \), respectively, and make most of their contacts with Val96 in strand \( \beta_T \).

In Figure 3, the calculated \( \phi \) profile obtained with 22 experimental \( \phi \) values is compared with that obtained using only three residues as constraints in the simulations. The two sets of results are quite similar and they have a coefficient of correlation of 0.74; if the three key residues are not

---

**Figure 2.** \( \phi \) values for AcP. Black diamonds are the experimental data, the continuous red line represents the average \( \phi^{cal} \) values in the TSE and the white area the precision of the calculation (see the text).

**Figure 3.** Calculated \( \phi \) values from the key-only calculations compared to the all-residue calculations. Circles indicate the key residues.
included in the calculation of the coefficient of correlation, the result is 0.65. For residues other than the three key ones for which experimental \( \phi \) values have been measured, the predictive power of the present technique can be tested. For example for residue V13 we obtain \( \phi^{\text{calc}} = 0.37 \pm 0.14 \) to be compared to \( \phi^{\exp} = 0.37 \pm 0.03 \), for V20 \( \phi^{\text{calc}} = 0.11 \pm 0.16 \) to be compared to \( \phi^{\exp} = 0.18 \pm 0.21 \), and for I75 \( \phi^{\text{calc}} = 0.02 \pm 0.04 \) to be compared to \( \phi^{\exp} = 0.02 \pm 0.10 \). There is one residue (L89) for which our prediction is inaccurate. In this case, using only the \( \phi^{\exp} \) values of the key residues as constraints, we get \( \phi^{\text{calc}} = 0.68 \pm 0.17 \) while the experimental result is \( \phi^{\exp} = 0.07 \pm 0.03 \). This difference arises from the fact that \( \phi^{\exp} \) values of residues in non-key positions have only a local effect on the structure. Residue L89 forms native interactions mainly with residue V36, for which \( \phi^{\exp} = 0.22 \). These interactions are localized in the long \( \beta_1-\beta_2 \) loop and do not influence global conformational features.

In Figure 4, the profile of \( \phi^{\text{calc}} \) obtained from the MD simulations is compared with that obtained from the C\(^\alpha\) model sampled by an MC procedure. The agreement is very good, although there are significant differences for some residues for which there are no \( \phi^{\exp} \) values. The results support the validity of the approximation in the MC approach; ignoring the detailed interactions and the side-chains, both in the model and in the definition of \( \phi^{\text{calc}} \), does not strongly affect the prediction of the \( \phi \) values for residues which have not been mutated. The validity of such an approximation, which may appear surprising, is a direct consequence of the important role that the native topology, as determined from protein engineering data, plays in determining the conformation space available in the TSE. Correspondingly, the agreement suggests that the multiple-temperature MD approach used here allows the sampling of a region of the conformation space that is comparable to that sampled in the MC approach.

The most important differences between the MD and MC \( \phi \) values are found for residues E56 and K57. They are both in the \( \alpha_2 \) helix, for which the \( \phi^{\text{calc}} \) is large (0.82 for both residues, while those from the MC simulation are both below 0.3). It appears that even in the least native members of the TSE, K57 and residues 33–35 of \( \alpha_1 \), and the beginning of helix \( \alpha_2 \) are packed in a manner similar to that of the native structure. Taddei et al.\(^{38}\) engineered multiple variants of AcP, designed to stabilize its helices. From the analysis of the effect of the mutations on the folding and unfolding kinetics they suggested that the second helix is more structured than the first in the transition state.\(^{38}\) The present determination of the TSE of AcP is consistent with these conclusions, despite the fact that neither helix is an essential part of the folding nucleus. In other words, the process of helix formations is quite distinct from the process of nucleation of the native structure, although stabilization of the helices may affect the folding and the unfolding rates.

One case where the MC and MD results are in good agreement, even though the values are different from those expected from a simple interpolation of the experimental values, occurs for the residues in the middle of the sheet \( \beta_2 \). Both simulations give quite low \( \phi^{\text{calc}} \) (\( \phi^{\text{calc}}_{\text{calc}} = 0.01 \), \( \phi^{\text{calc}}_{\text{calc}} = 0.03 \) and \( \phi^{\text{calc}}_{\text{calc}} = 0.01 \) for the MD simulation) for the three residues in the \( \beta_4 \) strand for which measurements exist (\( \phi^{\exp}_{\text{exp}} = 0.02 \), \( \phi^{\exp}_{\text{exp}} = 0.04 \) and \( \phi^{\exp}_{\text{exp}} = 0.00 \)), while residues S81 and N82 have generally large \( \phi^{\text{calc}} \) (between 0.5 and 0.7). These finding are, however, in accord with the \( \phi \) value measurements for the corresponding residues in ADA2h,\(^{39}\) a protein that has a structure very similar to that of AcP. Similarly, residues E27, D28 and A30 have quite large \( \phi^{\text{calc}} \), while consideration of only F22 and E29 (for which \( \phi^{\exp} < 0.15 \)) might suggest that the whole helix \( \alpha_1 \) (see Figure 1) is absent in the transition state.
The structure of the transition state

As described in Theory and Methods, the NSE and the TSE structures were clustered to simplify the description of the results. The structures corresponding to the centers of these representative clusters are shown in Figure 5. Figure 5(a) shows the experimental structure of AcP for comparison; Figure 5(b)–(e) show the cluster centers representing the NSE. The RMSD between the structures is in the range 1.3–4.0 Å, comparable to those between different NMR models. The structures in Figure 5(f)–(o) are representative of all the structures of the TSE. They are all at least 4 Å from each other and all the structures of the ensemble are within 4 Å RMSD of one of these structures; also, structures come from all the simulations between 360 K and 780 K, see Theory and Methods).

In all the cases the overall fold of the TSE is essentially that of the native protein but the structure is expanded and the secondary elements are less well formed. The higher temperature sample has clearly less native-like secondary structure with respect to the lower temperature one. It appears that the hairpin $\beta_2-\beta_3$ is the most persistent part of the native $\beta$ sheet.

Figure 6 shows the representative structures of the MC TSE (top) and the ten structures in Figure 5(f)–(o) (bottom). Mean structures are also shown as thick lines. Both have the native-like topology, although the MD structure is less expanded horizontally. The differences are probably due to the fact that the $C^\alpha$ model is defined on the basis of experimental constraints and chain connectivity. As a consequence the TSE obtained by the MC method is likely to show too large a value for the backbone mobility, since it is not limited by the presence of hydrogen bonds and the requirement of side-chain packing.

Role of individual residues in the overall stability

Three residues, Y11, P54 and F94, were identified as key residues in the transition state of AcP by showing that their native contacts determine the fold of the transition state. To examine whether these residues are particularly stabilizing, we use the fact that the EEF1 solvation free energy can be decomposed into a sum of pairwise interactions, i.e. the effective energy, which includes the solvated contribution, can be written as a sum over residues $i$ as:

$$E_{\text{EEF1}} = \sum_i \frac{1}{2} E_i = \sum_{i,j} E_{ij}$$  \hspace{1cm} (7)$$

where $E_i$ is the energy gained when residues $i$ and $j$ interact and $\Delta_{ij}$ is a matrix whose elements are equal to 1, for $j \geq i + 2$ and zero, otherwise.
The single residue energies $E_i$, where local contributions up to $i + 4$ have been disregarded, are shown in Figure 7. In the energy minimized native structure, residues Y11, Y25, W38, Q50, Q52, and F94 have values of $E_i$ less than $-15$ kcal/mol. Mutation of proline residues gives $\phi$ values that are difficult to interpret. Residues that strongly contribute to the stabilizing energy of the TSE are Y11, W38, Q50, Q52 and F94. The key residue not in this list, P54, also makes a non-negligible stabilizing contribution. Our finding that P54 plays a key role in the folding process of AcP is supported by the fact that the two nearby residues Q50 and Q52 are also characterized by large $\phi$ values and that the chain fragment Q50–P54 occupies a central position in the native fold of AcP. P54, together with Y11 and F94, occupy particularly suitable positions to seed the process of nucleation, as also shown by the fact that these three residues occupy a central position in the network of interactions that are present in the TSE (see Figure 8).37

Figure 6. Stereo plots of representative structures of the MC TSE (top, taken from Vendruscolo et al.7) and the ten structures in Figure 5(f)–(o) (bottom). Mean structures are shown as thick lines. The residues with a given secondary structure in the native state are indicated with the color code used in Figure 1. Average structures are obtained by superimposing the three key residues (Y11, P54 and F94, shown as yellow beads) for all the structures in the TSE.

Figure 7. Single residue energies (see equation (7)) in the native and the transition states.
Interestingly, the three key residues Y11, P54 and F94, although their contribution to the protein stability is not the largest, contribute to the energetic stability of both the NSE and the TSE. In the TSE residue Y11 interacts strongly with the segment V47-V51 in β3, S65 in α2 and F80 in β4; residue P54 makes strong interactions with the segment P5-K7 and G34-V35 which stabilize the the N-terminal and two loops α2–β3 and α1–β2; and residue F94 interacts mainly with A26-E27-R31 in α1, and V35-G37-V39 in β2 and V51 in β3. Thus, it appears that the importance of the key residues is due both to their stabilizing energy and to their network of contacts with the other residues.

The effective energy $E_{ij}$ are shown as “energy maps” in Figure 9. Strongly interacting structural elements are easily identified on such energy maps. For instance, in the native energy map (Figure 9, left) pairs of residues involved in secondary structure elements appear clearly and they form local stable regions. In the energy map for the transition state (Figure 9, right) all the native regions are still present. However, in certain cases, they are “smeared out” relative to the native state. This is the case for the interactions involving the various β-strands (β1 with β4 and β2 with β3); see the legend to Figure 1 for the secondary structure residue number. The analysis of the energy maps also reveals that parts of the structure which do not preserve the native secondary structure, so that they appear disordered, still interact strongly, although less specifically. An example is the region involving strands β1 and β4. Thus, the energy maps show that the interactions favoring the native topology are still present in the much less structured transition state. It is also interesting to note that two key residues (Y11 and F94) “cross” these smeared-out regions in the energy map, suggesting that one feature of the key residues is that they have some interactions that are non-native but help to stabilize the native topology. Specifically, the greater flexibility of residues in...
the TSE relative to the NSE leads to favorable interactions with residues which are neighbors of those which make strong interactions in the native state.

Conclusions

Here, we introduce a molecular dynamics approach based on a realistic all-atom protein model with implicit solvent for using protein engineering (φ value) data as experimental restraints to obtain a detailed structural description of the TSE. The relative importance of experimental restraints and force field is modulated by varying the temperature at which simulations are performed. The results complement those obtained earlier with a similar method based on MC sampling and a Cα model. The approach is similar in spirit to the use of NOE restraints and simulated annealing for the determination of the native structure of a protein. The resulting ensemble is more coarse-grained than the native state structure, in part because fewer data are available and in part because the TSE is much more disordered but it enables important features of the transition state to be identified and analyzed. The approach is general and can be applied to any protein for which φ values have been measured. Applications to other systems will be published in the near future.

Acknowledgements

We thank Jane Clarke, Fabrizio Chiti and Niccolò Taddei for stimulating discussions. E.P. acknowledges financial support from Forschungskredit der Universität Zürich. M.V. was supported by an EMBO long-term fellowship and by a Royal Society University Research Fellowship. The research of C.M.D. is supported by the Wellcome Trust. Work at Harvard (M.K.) was supported in part by a grant from the NIH. This work was begun while M.K. was Eastman Visiting Professor at Oxford University (1999–2000).

References

Appendix: The \( P(1/2) \) Criterion for Transition States: an Evaluation

An approach for determining or validating transition states has been proposed for protein folding.\(^1\) The method, called \( P(1/2) \), defines the TSE as made up of the set of structures such that half the trajectories, starting from those structures with varying initial conditions (i.e. distribution of velocities), will fold and half will unfold. The origin of this concept goes back to methods for determining transmission coefficients, starting from a known transition state\(^2\) and the identification of simpler transition states in protein dynamics (e.g. ring flips in the bovine pancreatic trypsin inhibitor\(^3\)).

The basic problem with using \( P(1/2) \) to validate the TSE obtained here is that the true force field of a protein is unknown. The situation can be understood by considering an NMR structure determination with simulated annealing using NOE and other data (e.g. coupling constants) as restraints on the structure in the presence of an empirical energy function. It would not be a test of the structure, but rather of the energy function used for the simulated annealing calculation, if one studied whether the structure was stable during molecular dynamics simulations in the absence of the experimental restraints. Exactly the same argument applies to a \( P(1/2) \) test of the TSE determined with restraints based on experimental data (e.g. the \( \phi \) values used here). That is, refolding and unfolding simulations using the empirical energy function with the transition state determined experimentally is a test of the energy function, not of the transition state. The lack of knowledge of the true energy function and the difficulty of searching conformation space with it is exactly the reason for using experimental data to determine the transition state. The energy function introduces the connectivity of the protein and is convenient for eliminating unrealistic interactions.

The above argument does not apply to theoretical studies where the transition state of a known force field is determined by simulations. In this case the approach is consistent in that the same force field is used both to generate and to validate the transition state. For example, \( P(1/2) \) analyses with all-atom molecular mechanics potentials have been made in high-temperature unfolding study of C12 with explicit solvent\(^4\) and in unfolding simulations of an SH3 domain with implicit solvent\(^5\). In both systems the transition state was determined from unfolding trajectories; in the first case by using the criterion that the protein is expected to unfold rapidly after the transition.

state is reached and in the second by selecting coordinate sets by using experimental $\phi$ values. Thus, it is entirely appropriate to test the chosen TSE by applying the $P(1/2)$ criterion to it. However, a practical difficulty with doing so with all-atom molecular mechanics energy functions is that refolding to the native state is never fully achieved; e.g. in the study by Gsponer & Caflisch, the best refolding corresponded to a reduction of the RMSD from the native to 2.5 Å from values between 3.5 Å and 5.3 Å.

Another possibility is to determine the transition state from experimental data and then to analyse the results using the $P(1/2)$ criterion with a model force field. In one example of such an approach, Li & Shakhnovich implemented the $P(1/2)$ criterion using the force field corresponding to a Go model. In their simulations a full unfolding and folding has been achieved as required in the $P(1/2)$ test. However, we note that the Go potential is characterized by a well that is much deeper than the physical well in a native protein, so that the choice of a temperature for the $P(1/2)$ simulation is not straightforward. This is a critical aspect of $P(1/2)$ calculations with the Go model because at lower temperatures, all the trajectories would fold. As a matter of fact, once the TSE has been determined experimentally for a given protein (CI2 in their case), as done by Li et al. using the same criterion as described by Vendruscolo et al. and here, the $P(1/2)$ test can be used to determine the temperature for the Go model that corresponds to the "physiological" transition temperature for folding that particular protein.

As pointed out above, due to the fact that we do not know the true force field, the $P(1/2)$ criterion cannot, strictly speaking, be used to validate the TSE determined from experimental restraints. However, we could still ask whether the force field used in the simulations is a good approximation to the true one for determining transition states. To explore this question in a preliminary way, we calculated several trial trajectories, each 10 ns in length starting with different structures in the TSE. None folded or unfolded completely during the simulation time as measured by the RMSD from the native state; in fact six out of ten were trapped in a local minimum (the RMSD fluctuated by 0.4 Å for the simulations which had RMSD values between 5 Å and 7 Å), one folded slightly (the RMSD was reduced from 5.4 Å to 4.2 Å), and three unfolded slightly (the RMSD increased from about 7 Å to 8–10 Å). It would require very long simulations to obtain less ambiguous results. Therefore, we are not able to conclude whether or not the transition state of EEF1 coincides with the transition state of the true force field, which corresponds to that determined experimentally here.

References


Edited by C. R. Matthews

(Received 25 March 2002; received in revised form 20 August 2002; accepted 29 August 2002)