When 0.9 M potassium/sodium tartrate, 0.1 M HEPES pH 7.5 was used as precipitant, similar trigonal crystals appeared in the drops together with a second, monoclinic ($P2_1$) crystal form (a = 107.4 Å, b = 153.4 Å, c = 162.5 Å, β = 94.2°). These crystals diffract as well as the trigonal although they contain 12 protomers in the asymmetric unit. A tantalum bromide ($Ta_6Br_{12}^{2+}$) derivative of the trigonal crystal form was obtained by overnight soaking of native crystals resulting in a unit cell with an enlarged c axis parameter ($a = b$ = 151.2 Å, c = 262.8 Å).

Data collection and phasing

Complete diffraction datasets were collected from a single crystal each at EMBL beam lines BW7A and BW7B (Outstation Hamburg, DESY) and ID14-2 and ID14-4 (Outstation Grenoble, ESRF). Three datasets were recorded from the $Ta_6Br_{12}^{2+}$ derivative (peak, inflection point and hard remote) around the theoretical Ta L-III edge. All data were processed with MOSFLM 6.01 (ref. 24) and scaled, merged and reduced with programs of the CCP4 suite[25]. A self-rotation calculation performed for the trigonal data indicated the presence of a local six-fold axis at Eulerian angles α = 12.8°, β = 56.6° and γ = 12.8°, in accordance with the presence of six monomers displaying local C6 symmetry. The structure was solved by multiple-wavelength anomalous diffraction (MAD) using the tantalum bromide derivative of the trigonal crystal form using SHELXS[26] and MLPHARE[25]. Six sites were found and refined, and phases were computed to 4.5 Å. A subsequent density modification step using DM within CCP4 (ref. 25) rendered an F_{obs} electron density map that allowed us to build a monomer mask and localize the local six-fold axis which was consistent with the self-rotation function. The original MAD-phases to 4.5 Å were used with the native structure factor amplitudes and a density modification run with non-crystallographic-symmetry (NCS) averaging and phase extension to 2.7 Å was computed. This led to a straightforward traceable native map. The model was built with Turbo-Frodo.

Structure refinement

Successive cycles of maximum-likelihood positional and temperature factor refinement with CNS version 1.0 (ref. 27) using progressively all data up to the full resolution of 2.4 Å and keeping NCS restraints followed. Computation of phased-combined maps and manual model building permitted the gradual completion of the model. At the final stages, solvent molecules and sulphate ions were introduced. Table 1 provides a summary for final model refinement. The structure of the monoclinic crystal form was solved by molecular replacement with AMoRe[28]. A whole TrwBΔN70 hexamer was used as a searching model and two clear solutions were obtained with correlation coefficient and R_{factor} equal to 67.4/36.3% (second highest solution 41.2/47.1%). The model was inspected and the new electron-density-based differences with the trigonal structure were corrected. Solvent molecule position assignment and model refinement with CNS proceeded similarly and applying NCS restraints. No ions were localized in these crystals. All residues excepting His 125 of each chain, clearly defined by electron density and involved in β-sheet interactions, are in allowed regions of the Ramachandran plot. Superimpositions were calculated with Turbo-Frodo, Lsqkab of the CCP4 suite[25] and Lsqman of the RAVE package (Uppsala Software Factory; http://alpha2.bmc.uu.se/~gerard/manuals). Figures were computed with Turbo-Frodo, SETOR[29] and GRASP[30]. Lee & Richards buried surface accessibility calculations (probe radius 1.4 Å) and close contacts (< 4 Å) were ascertained with CNS version 1.0. Structural similarity searches were performed with the DALI server (http://www.ebi.ac.uk/dali).

Received 21 September; accepted 8 November 2000.

1. Stachel, S. E. & Zambryski, P. *Agrobacterium tumefaciens* and the susceptible plant cell: a novel adaptation of extracellular recognition and DNA conjugation. *Cell* **47,** 155–157 (1986).
2. Heinemann, J. A. & Sprague, G. F. J. Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast. *Nature* **340,** 205–209 (1989).
3. Wu, L. J., Lewis, P. J., Allmansberger, R., Hauser, P. M. & Errington, J. A conjugation-like mechanism for prespore chromosome partitioning during sporulation in *Bacillus subtilis*. *Gene Dev.* **9,** 1316–1326 (1995).
4. Llosa, M., Bolland, S. & de la Cruz, F. Genetic organization of the conjugal DNA processing region of the IncW plasmid R388. *J. Mol. Biol.* **235,** 448–464 (1994).
5. Moncalián, G. *et al.* Characterization of ATP and DNA binding activities of TrwB, the coupling protein essential in plasmid R388 conjugation. *J. Biol. Chem.* **274,** 36117–36124 (1999).
6. Zechner, E. L. in *The Horizontal Gene Pool: Bacterial Plasmids and Gene Spread* (ed. Thomas, C. M.) 87–173 (Harwood Academic, London, 2000).
7. Cabezón, E., Sastre, J. I. & de la Cruz, F. Genetic evidence of a coupling role for the TraG protein family in bacterial conjugation. *Mol. Gen. Genet.* **254,** 400–406 (1997).
8. Begg, K. J., Dewar, S. J. & Donachie, W. D. A new *Escherichia coli* cell division gene, *ftsK*. *J. Bacteriol.* **177,** 6211–6222 (1995).
9. Walker, J. E., Saraste, M., Runswick, M. J. & Gay, N. J. Distantly related sequences in the α- and β-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.* **1,** 945–951 (1982).
10. Abrahams, J. P., Leslie, A. G. W., Lutter, R. & Walker, J. E. Structure at 2.8 Å resolution of F1-ATPase from bovine heart mitochondria. *Nature* **370,** 621–628 (1994).
11. Singleton, M. R., Sawaya, M. R., Ellenberger, T. & Wigley, D. B. Crystal structure of T7 gene 4 ring helicase indicates a mechanism for sequential hydrolysis of nucleotides. *Cell* **101,** 589–600 (2000).
12. Story, R. M., Weber, I. T. & Steitz, T. A. The structure of the *E.coli* recA protein monomer and polymer. *Nature* **355,** 318–325 (1992).
13. Sawaya, M. R., Guo, S., Tabor, S., Richardson, C. C. & Ellenberger, T. Crystal structure of the helicase domain from the replicative helicase-primase of bacteriophage T7. *Cell* **99,** 167–177 (1999).
14. Guenther, B., Onrust, R., Sali, A., O'Donnell, M. & Kuriyan, J. Crystal structure of the δ′ subunit of the clamp-loader complex of *E.coli* DNA polymerase III. *Cell* **91,** 335–345 (1997).
15. Lenzen, C. U., Steinmann, D., Whiteheart, S. W. & Weis, W. I. Crystal structure of the hexamerization domain of N-ethylmaleimide-sensitive fusion protein. *Cell* **94,** 525–535 (1998).
16. Subramanya, H. S., Bird, L. E., Brannigan, J. A. & Wigley, D. B. Crystal structure of a DExx box DNA helicase. *Nature* **384,** 379–383 (1996).
17. Subramanya, H. S. *et al.* Crystal structure of the site-specific recombinase, XerD. *EMBO J.* **16,** 5178–5187 (1997).
18. Egelman, E. H., Yu, X., Wild, R., Hingorani, M. M. & Patel, S. M. Bacteriophage T7 helicase/primase proteins form rings around single-stranded DNA that suggest a general structure for hexameric helicases. *Proc. Natl Acad. Sci. USA* **92,** 3869–3873 (1995).
19. Hacker, K. J. & Johnson, K. A. A hexameric helicase encircles one DNA strand and excludes the other during DNA unwinding. *Biochemistry* **36,** 14080-14087 (1997).
20. Yu, X., Hingorani, M. M., Patel, S. S. & Egelman, E. H. DNA is bound within the central hole to one or two of the six subunits of the T7 DNA helicase. *Nature Struct. Biol.* **3,** 740–743 (1996).
21. Soultanas, P. & Wigley, D. B. DNA helicases: 'inching forward'. *Curr. Opin. Struct. Biol.* **10,** 124–128 (2000).
22. Raney, K. D. & Benkovic, S. J. Bacteriophage T4 DDA helicase translocates in an unidirectional fashion on single-stranded DNA. *J. Biol. Chem.* **270,** 22236–22242 (1995).
23. Kaplan, D. L. The 3′-tail of a forked-duplex sterically determines whether one or two DNA strands pass through the central channel of a replication-fork helicase. *J. Mol. Biol.* **301,** 285–299 (2000).
24. Leslie, A. G. W. in *Crystallographic computing V* (eds Moras, D., Podjarny, A. D. & Thierry, J. C.) 27–38 (Oxford Univ. Press, Oxford, 1991).
25. CCP4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
26. Sheldrick, G. M. Patterson superposition and *ab initio* phasing. *Methods Enzymol.* **276,** 628–641 (1997).
27. Brünger, A. T. *et al.* Crystallography & NMR System: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).
28. Navaza, J. AMoRe: an automated package for molecular replacement. *Acta Crystallogr. A* **50,** 157–163 (1994).
29. Evans, S. V. SETOR: hardware lighted three-dimensional solid model representations of macromolecules. *J. Mol. Graphics* **11,** 134–138 (1993).
30. Nicholls, A., Bharadwaj, R. & Honig, B. GRASP: graphical representation and analysis of surface properties. *Biophys. J.* **64,** A166–A166 (1993).
31. Sastre, J. I., Cabezon, E. & de la Cruz, F. The carboxyl terminus of protein TraD adds specificity and efficiency to F-plasmid conjugative transfer. *J. Bacteriol.* **180,** 6039–6042 (1998).

Acknowledgements

We are most grateful to R. Huber for making tantalum bromide available to us, and to I. Usón for help with SHELX. This work was supported by grants from the Ministerio de Educación y Cultura of Spain, the Generalitat de Catalunya and the European Union. Synchrotron data collection was supported by EU grants and the ESRF.

Correspondence and requests for materials should be addressed to M.C. (e-mail: mcccri@ibmb.csic.es). The co-ordinates of TrwBΔN70 have been deposited with the Protein Data Bank (access codes 1e9r and 1e9s).

Three key residues form a critical contact network in a protein folding transition state

Michele Vendruscolo*, Emanuele Paci*†, Christopher M. Dobson* & Martin Karplus*†‡

** Oxford Centre for Molecular Sciences, New Chemistry Laboratory, University of Oxford, South Parks Road, Oxford OX1 3QT, UK*
† Laboratoire de Chimie Biophysique, ISIS, Université Louis Pasteur, 4 rue Blaise Pascal, 67000 Strasbourg, France
‡ Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138, USA

Determining how a protein folds is a central problem in structural biology. The rate of folding of many proteins is determined by the transition state, so that a knowledge of its structure is essential for understanding the protein folding reaction. Here we use mutation measurements—which determine the role of individual residues in stabilizing the transition state[1,2]—as restraints in a Monte Carlo sampling procedure to determine the ensemble of structures that make up the transition state. We apply this approach to the experimental data for the 98-residue protein acylphosphatase[3], and obtain a transition-state ensemble with the

native-state topology and an average root-mean-square deviation of 6 Å from the native structure. Although about 20 residues with small positional fluctuations form the structural core of this transition state, the native-like contact network of only three of these residues is sufficient to determine the overall fold of the protein. This result reveals how a nucleation mechanism involving a small number of key residues can lead to folding of a polypeptide chain to its unique native-state structure.

As the transition state corresponds to the least stable region on a potential energy surface, it is difficult to determine its structure by the methods commonly used for the stable portions of the surface (that is, reactants, products or intermediates). For certain small-molecule reactions direct observation of the transition state has been possible[4], but it is not evident how to achieve a corresponding result for protein folding. The protein engineering method, introduced by Fersht, Winter and co-workers in 1982 (refs 5, 6), can be used to provide experimental information about the residue interactions present in the transition state for folding[1,2]. The essential element of the method is measurement of the quantity ϕ_i^{exp}, which is the ratio of the change in stability, $\Delta\Delta G_{TSE}$, of the transition-state ensemble (TSE) to that of the native state, $\Delta\Delta G_{NSE}$, due to the mutation of residue i; the unfolded state is used as the reference.

Here we introduce a strategy that substantially extends the information that can be derived from a set of ϕ_i^{exp} values for a protein. The essence of the method is that the TSE is restrained to be the most stable region of an energy function, which includes the information from the ϕ_i^{exp} values, rather than being an unstable region, as for the true energy function of the protein. In this way the TSE can be obtained by standard methods for sampling conformational space, in a manner analogous to the determination of native structures from experimental NMR data[7].

Our approach uses the ϕ_i^{exp} values, interpreted in terms of native-like inter-residue contacts, as restraints to generate an ensemble of structures. It is applied here to determine the TSE of acylphosphatase (AcP) (Fig. 1a). A representation of the TSE of AcP consisting of 1,144 structures was generated using only the 24 measured ϕ_i^{exp} values[3] as restraints (the 'primitive' model, see Methods). The Pearson linear correlation coefficient ρ between the experimental values and those obtained from the TSE is 0.99. (See also Supplementary Information). As the chain connectivity requires that certain contacts are formed simultaneously, the very high correlation demonstrates the internal consistency of the experimental data. Moreover, this finding provides additional evidence for the fundamental assumptions on which the ϕ-value analysis is based[2].

Figure 2a shows the calculated values (ϕ_i^{calc}) for all the residues and compares them with the ϕ_i^{exp} values. For most of the residues for which experimental data are not available, the ϕ_i^{calc} interpolate smoothly between the ϕ_i^{exp}. There are, however, some exceptions. The most important of these are in the central region of strand β_4 (residues 79–82), for which the predicted values are relatively large. Although the hydrophobic mutation investigated in the region of AcP (F80L) destabilized the protein too much for ϕ_i^{exp} to be determined[3], ϕ values for the corresponding region of the procarboxypeptidase A2 activation domain (ADA2h), (ref. 8), a protein having the same overall topology as AcP, have been measured; the average ϕ_i^{exp} for ADA2h for the β_4 strand (0.22 ± 0.17) is reasonably

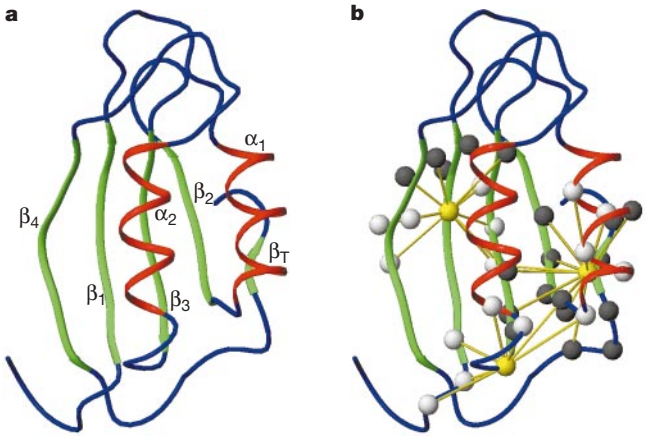

Figure 1 Native structure of acylphosphatase, AcP. **a**, The secondary structure elements are: β_1 (residues 7–13), α_1 (residues 22–33), β_2 (residues 36–42), β_3 (residues 46–53), α_2 (residues 55–56), β_4 (residues 77–85), β_T (residues 93–97); residues between these regions are parts of loops. **b**, the key residues 11, 54 and 94 found from the transition-state analysis are shown as gold spheres on the native structure (see text). They are connected by gold-coloured bonds to the residues (white and black spheres) forming a native contact with them (Y11 has long-range contacts with 47–52 and 78–81, P54 with 5–7 and 34–35, and F94 with 26–31, 36–39 and 50–52). Residues forming the transition-state core identified in the present work are shown as black spheres. They fall into two groups in spatial proximity; there is a larger group (residues 28, 35–39, 51–54 and 90–95) comprising parts of α_1, β_2, α_2 and β_T, and a smaller group (residues 11–13, 47 and 78, 79) comprising parts of the β_1, β_3 and β_4 strands in the native structure.

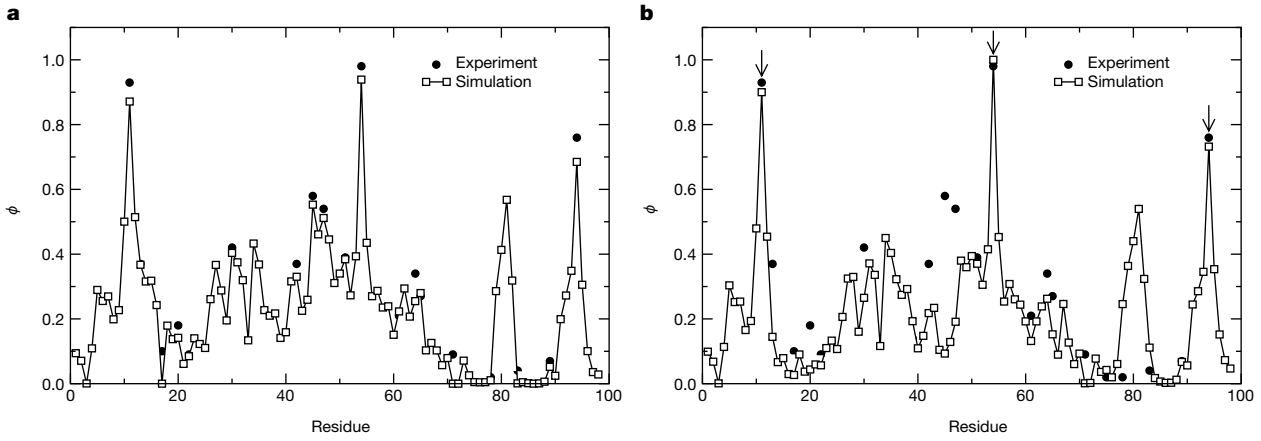

Figure 2 Comparison of ϕ_i^{calc} with ϕ_i^{exp}. **a**, When all 24 of the latter are used as restraints (filled circles); **b**, when only Y11, P54 and F94, are used as restraints (open squares, see text). Only four ϕ_i^{exp} are significantly underestimated; they are V13, T42, G45 and V47. The reason for the underestimation is that this subset of residues forms long-range contacts mainly to each other (for example, V13 with G45) and are, therefore, almost totally unrestrained in the primitive model when Y11, P54 and F94 are used as restraints; the only common contact is between residues Y11 and V47. Except for V13, which is close to Y11, they are not part of the folding core.

close to the average ϕ_i^{calc} (0.31 ± 0.17) for this region of AcP. Other regions of AcP, where the simulations predict significant non-zero values of ϕ_i^{calc} for which no ϕ_i^{exp} exist, include the hydrophobic residues L6 ($\phi_6^{calc} = 0.28 \pm 0.10$), A26 ($\phi_{26}^{calc} = 0.26 \pm 0.16$) and I35 ($\phi_{35}^{calc} = 0.37 \pm 0.13$); measurements on these residues are in progress (F. Chiti *et al.*, personal communication).

To probe the significance of different residues in determining the structure of TSE, we have performed a series of calculations of the TSE in which only subsets of the ϕ_i^{exp} are used as restraints. A notable result is that by use of only the three residues with the largest ϕ_i^{exp} (Y11, P54 and F94) as restraints, a correlation, coefficient of 0.86 is found between the ϕ_i^{exp} and the ϕ_i^{calc} (Fig. 2b); the cross-validated correlation coefficient (that is, with ϕ_{11}^{calc}, ϕ_{54}^{calc} and ϕ_{94}^{calc} excluded from the comparison) is 0.50. This value is significant in light of the sensitivity of ϕ values to local interactions, which makes their quantitative prediction more difficult than that of the overall fold.

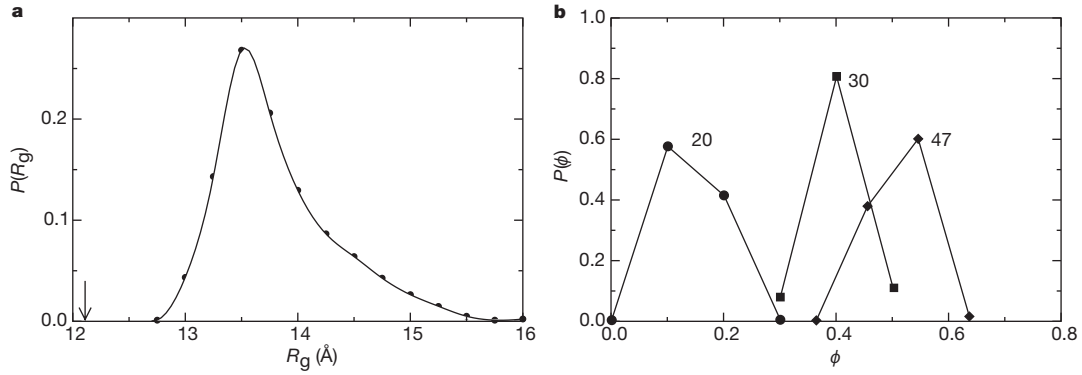

Figure 3 Properties of the transition-state ensemble, TSE, of AcP. **a**, Probability distribution $P(R_g)$ of the radius of gyration for the improved model with $\lambda = 0.85$ and $T = 0.1$ (see Methods). The arrow points to the R_g value of the native state. **b**, Probability $P(\phi)$ of a residue of having a given ϕ value in the TSE. The results for residues V20, A30 and V47 are shown (circles, squares and diamonds, respectively); they have a rather narrow unimodal distribution around the experimental values ($\phi_{20}^{exp} = 0.18$, $\phi_{30}^{exp} = 0.42$, $\phi_{47}^{exp} = 0.54$). The value $\phi_{47}^{calc} = 0.56 \pm 0.05$ arises because residue V47 has a contact probability close to 1 with residues 11–14, about 0.5 with residues G15, K40 and N41, and a very low probability of contact with any other residue. $P(\phi)$ remains unimodal for all the residues in the calculation in which only Y11, P54 and F94 are used as a bias in equation (3).

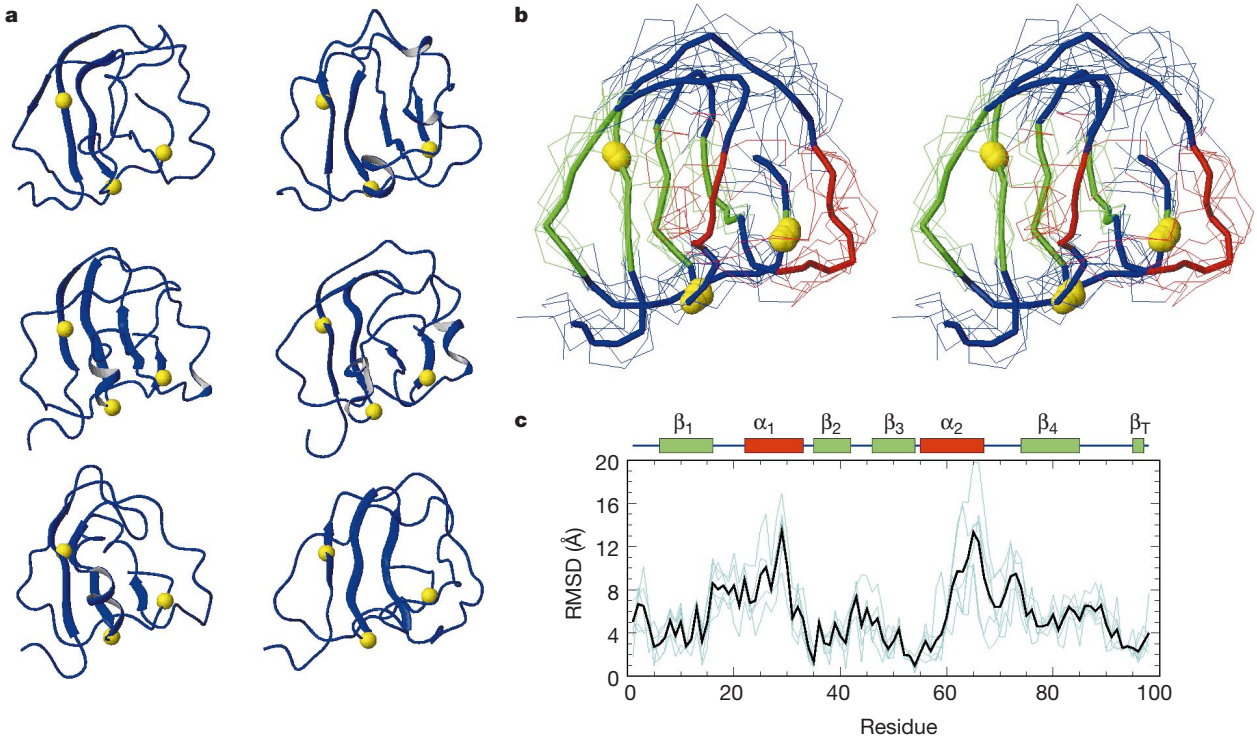

Figure 4 Representation of the TSE of AcP calculated with the improved model (see Methods). **a**, The six most representative conformations of the TSE generated from the structure calculations; more than one-quarter of the 867 generated structures are less than 4 Å in r.m.s. deviation from one of the six structures shown. The three key residues forming the structural core are shown as gold spheres and secondary structural elements are shown as ribbons with arrows indicating the directionality of the β-sytrands; a method which does not rely on hydrogen-bonding but on comparison with structural templates (M. Schaefer and M.K., unpublished results) was used to identify secondary structures. Preliminary results for molecular dynamics structures obtained with an all-atom potential and the same ϕ-value restraints indicate that somewhat more secondary structure, particularly for the helices, is present in the TSE (E.P. *et al.*, unpublished results). **b**, The average structure of the transition state (thick lines) derived from the six conformations shown in **a** (thin lines). The residues with a given secondary structure in the native state are indicated with the same colour code as that used in Fig. 1. The average structure is obtained by superimposing the three key residues (gold spheres) for all the structures in the TSE. **c**, Average r.m.s. deviation from the native state as a function of the residue number; the native secondary structural elements are indicated above the diagram.

Nevertheless, the predictions of 21 ϕ_i^{exp} by the 'leave-one-out' method (keeping the three key residues and 20 of the 21 remaining ones) yields a cross-validated correlation coefficient of 0.71. As a further test of the role of the three residues, we compare the ϕ_i^{calc} for all residues obtained with only the three residue restraints and with all the 24 ϕ_i^{exp} restraints; the correlation coefficient is 0.91 (cross-validated 0.83). Other small subsets of residues (not including Y11, P54 and F94) give very poor correlations; the cross-validated correlation coefficients are in the range −0.1 to 0.2. These results show that native-like contacts for the three key residues alone determine the overall fold of the TSE and give information about the additional ϕ values.

Residues Y11, P54 and F94 are part of the hydrophobic core of AcP (ref. 3), and make a large number of contacts with other residues (10, 7 and 14 contacts, respectively). These contacts, which are distributed throughout the structure, create a 'contact network' in the native state (Fig. 1b). There are other residues (for example, K7, V36 and V58) with an even larger numbers of contacts. These contacts, however, are more localized and their ϕ_i^{exp} are smaller, so that they are less important in determining the TSE. We note that the native-like environment for the three key residues can lead to a contact network that determines the fold of most of the protein (see Fig. 1b and legend).

As the primitive model has no stabilizing interactions other than those arising from the ϕ_i^{exp}, the conformational space accessible to the protein is expected to be too large. An 'improved' model (see equation (3) in Methods), in which the energy function is augmented by native-like attractive interactions such that the overall compactness of the TSE is consistent with experimental data on its average solvent exposure, yields a TSE with the same fold as the primitive model. The radius of gyration (R_g) is 13.5 ± 0.5 Å as compared to the native value of 12.6 Å; the R_g distribution of this TSE is shown in Fig. 3a. The degree of expansion is similar to that observed experimentally for 'molten globule' states in which much of the secondary structure is preserved but most tertiary contacts are not persistent[9]. The core of the transition state of AcP, defined as in Methods, consists of approximately 20 residues including the three key residues and many of their contacts (Fig. 1b).

Figure 4a shows the six most representative structures (that is, the centres of the clusters with most members) of the TSE. Figure 4b shows the average structure obtained by superposing all the members of the TSE; also shown are the structures corresponding to the dominant clusters (those in Fig. 4a), which yield the same average structure. Comparison of the average structure in Fig. 4b with Fig. 1a demonstrates that the fold of the TSE is native-like, but that it is significantly expanded relative to the native structure. There is a marked tendency for the β-sheets to be present, particularly β_2 and β_3, whereas the location of the helical regions is less well defined. Calculations incorporating additional local restraints, to account for the recent result of Taddei *et al.*[10] that stabilizing helix 2 increases the folding rate, suggest that this helix is in fact present in the TSE but that it is highly disordered relative to the rest of the structure.

The r.m.s. deviation of the TSE from the native state as a function of residue number is shown in Fig. 4c; it makes clear how much the average structure deviates from the native state (some parts by more than 15 Å), even though the fold is preserved. There are some non-native interactions (about 30% of the contacts are non-native). Overall the contact order[11] in the TSE (13.7) is significantly lower than in the native state (20.8), indicating that only about 70% of the native-like long-range contacts are formed. The distributions of the fraction of native contacts in the TSE obtained for AcP are relatively narrow; that is, the same residues make contact with each other in most members of the TSE (Fig. 3b). This finding for AcP supports the conclusions of Fersht *et al.*[12] for chymotrypsin inhibitor 2 that fractional ϕ values arise from the formation of an incomplete set of native contacts in all the members of the TSE rather than from the alternative possibility that some molecules in the TSE have most of their native contacts and others have only a few or none.

The present analysis introduces an approach for generating the molecular structure of the TSE from experimental ϕ values. It provides a check on the consistency of the experimental data, and determines the importance of specific structural elements and inter-residue interactions. If other parameters concerning the TSE are available from experiment, the present approach can be refined by including them in the energy function used for the structure determination. As the TSE is based on experiment, it complements existing theoretical studies[13–22]. The demonstration for AcP that the contacts of a small number of key residues determines the native-like fold provides a microscopic characterization of the nucleation step of the folding process[2]. These key residues form an extensive 'contact network' of native-like interactions in the TSE. This explains the role of these residues, which are not in direct contact with each other, as crucial elements in determining the transition state. The global contact network leads to a rather narrow TSE, whose core elements correspond to an expanded form of the native structure. We anticipate that the application of this method of analysis to a range of proteins will lead to a clearer understanding of the nature of the folding transition state and its relationship to the native structure. □

Methods

Definition of ϕ_i^{calc}

We have adopted a definition of ϕ_i^{calc} similar to one used previously in molecular-dynamics simulations[14,23]. For residue i,

$$\phi_i^{calc} = \frac{\langle N_i \rangle_{TSE}}{N_i^N} \quad (1)$$

where N_i is the number of native contacts formed by residue i; the average in the numerator, $\langle N_i \rangle_{TSE}$, is over all the conformations of the TSE, and in the denominator, N_i^N refers to the the native state structure. Contacts are defined to exist when the Cα positions are closer than 8.5 Å (ref. 24) (see below) and the residues are more than two neighbours apart along the chain (that is, contacts $i+1$ and $i+2$ are omitted). In equation (1) all native contacts give an equal contribution whereas non-native contacts are not included.

Models

Each amino acid is represented as a sphere of radius 4.25 Å with an inner hard-core radius of 2.35 Å to take into account excluded volume effects. Each sphere is centred at the position of the Cα atom and is connected by rigid bonds of length 3.8 Å to its neighbours along the chain[25]. Thus, two residues are in contact when they are closer than 8.5 Å, and non-neighbouring amino acids are not permitted to be closer than 4.7 Å. Comparison of the contacts derived from the present Cα model with those obtained from an all-atom model shows good agreement throughout the protein.

In the minimum interaction model ('primitive model') the energy of a given conformation, represented by its contact map S, is defined as:

$$E_{min}(S) = \frac{N_c^N}{N_\phi} \sum_i \left[\phi_i^{exp} - \phi_i^{calc}(S)\right]^2 \quad (2)$$

that is, the only attractive energy term between non-bonded residues is from the values of ϕ_i^{exp}; N_c^N is the total number of contacts in S^N, and N_ϕ is the number of available ϕ_i^{exp}. In the more realistic interaction model (the 'improved model'), an attractive (Gō-like[26]) interaction term for the native state contacts is introduced. It has the form:

$$E_\lambda(S) = \lambda E_{min}(S) + (1-\lambda) E_{Go}(S) \quad (3)$$

where the Gō-model energy, $E_{Gō}(S)$, is

$$E_{Gō}(S) = -\epsilon_1 \sum_{ij} S_{ij} S_{ij}^N + \epsilon_2 \sum_{ij} S_{ij}(1 - S_{ij}^N) \quad (4)$$

where S^N is the contact map of the native state. The element S_{ij} of a contact map is defined to be 1 if residues i and j are in contact and 0 otherwise. The phase diagram of E_{Go} is well known[26] and the two parameters ϵ_1 and ϵ_2 are fixed at $\epsilon_1 = 1$ and $\epsilon_2 = 0.1$ (M.V., E.P., C.M.D. and M.K., unpublished calculations). The parameter λ tunes the compactness of the TSE. For $\lambda = 0$ (the standard Gō-model) and a temperature $T = 0.1$ in arbitrary units (see below), we obtain a native-like value for the radius of gyration, $R_g = 12.7 \pm 0.1$ Å, and for $\lambda = 1$ (the 'primitive model') we obtain $R_g = 15.2 \pm 0.6$ Å. To choose a value for λ we performed Monte Carlo sampling at $T = 0.1$ to generate the TSE for several values of λ. For each value of λ, we computed the average accessible surface area (ASA) for the

conformations in the TSE using a probe of 3.1 Å that matches the standard one of 1.4 Å from all-atom calculations.

Using experimentally derived m values we estimate ASA_{TSE}^{exp} by using the empirical formula[14,27]

$$RT\frac{m_u}{m} = \frac{ASA_{TSE}^{exp} - ASA_{NSE}}{ASA_{RCE} - ASA_{NSE}} \quad (5)$$

where RCE and NSE stand for random-coil ensemble and native-state ensemble, respectively; m and m_u represent the derivative with respect to the denaturant concentration of the equilibrium constant and the rate of unfolding, respectively. The measured ratio is 0.23 for AcP (ref. 3). The intersection between the curves $ASA_{TSE}(\lambda)$ and ASA_{TSE}^{exp} gives the selected value, $\lambda = 0.85$. Use of only E_{Go} (equation (4)) in the Monte Carlo sampling yields a correlation coefficient of 0.09 between ϕ_i^{calc} and ϕ_i^{exp}; that is, there is essentially no correlation with the TSE.

Given the empirical energy function (equations (2)–(4)), a Monte Carlo method is used to determine the ensemble of structures, TSE, making up the transition state. Monte Carlo moves are bond-length-conserving crankshaft rotations that are restricted by steric restraints among neighbouring residues along the sequence and accepted according to the standard Metropolis criterion at the Monte Carlo temperature T (ref. 24). The choice of $T = 0.1$ is made by requiring that both for $\lambda = 0$ and for $\lambda = 1$ the protein is in a collapsed state.

Clustering

We assume that stable cores are formed by residues whose relative fluctuations are small for the given energy function. The 'distance' between residues used in clustering is related to the size of the fluctuations in their distance. Specifically, the distance f_{ij} between residues i and j is defined as a 'fluctuation ratio':

$$f_{ij} = \frac{\langle u_{ij}^2 \rangle}{\langle r_{ij} \rangle^2} \quad (6)$$

where the averages are taken over the conformations in the TSE, $u_{ij} = r_{ij} - \langle r_{ij} \rangle$, and r_{ij} are the cartesian distances between residues i and j in a given conformation in the TSE. The fluctuation ratio is related to the Berry parameter[28] used in the description of systems without a reference state when the Lindemann criterion[29] is unsuitable. The clustering using the distances f_{ij} was performed by using the SPC algorithm[30], and the resulting dendrograms were analysed with an algorithm provided by G. Getz and E. Domany (personal communication). The average fluctuation ratio over all pairs in the core is 0.019 and that over the entire protein is 0.049, more that twice as large; for comparison, the average fluctuation ratio in the native state with the energy function given in equation (4) is 0.010.

Received 8 June; accepted 8 December 2000.

1. Matouschek, A., Kellis, J. T., Serrano, L. & Fersht, A. R. Mapping the transition state and pathway of protein folding by protein engineering. *Nature* **340,** 122–126 (1989).
2. Fersht, A. R. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (Freeman, New York, 1999).
3. Chiti, F. *et al.* Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nature Struct. Biol.* **6,** 1005–1009 (1999).
4. Polanyi, J. C. & Zewail, A. H. Direct observation of the transition state. *Acc. Chem. Res.* **28,** 119–132 (1995).
5. Winter, G., Fersht, A. R., Wilkinson, A. J., Zoller, M. & Smith, M. Redesigning enzyme structure by site-directed mutagenesis. Tyrosyl tRNA ATP binding. *Nature* **299,** 756–758 (1982).
6. Fersht, A. R., Leatherbarrow, R. J. & Wells, T. N. Structure-activity relationships in engineered proteins. Analysis of use of binding energy by linear free energy relationships. *Biochemistry* **26,** 6030–6038 (1987).
7. Wüthrich, K. Protein structure determination in solution by nuclear magnetic resonance spectroscopy. *Science* **243,** 45–50 (1989).
8. Villegas, V., Martínez, J. C., Avilés, F. X. & Serrano, L. Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain. *J. Mol. Biol.* **283,** 1027–1036 (1998).
9. Kuwajima, K. The molten globule state of α-lactalbumin. *FASEB J.* **10,** 102–109 (1996).
10. Taddei, N. *et al.* Stabilisation of α-helices by site-directed mutagenesis reveals the secondary structure in the transition state for acylphosphatase folding. *J. Mol. biol.* **300,** 633–647 (2000).
11. Plaxco, K. W., Simons, K. & Baker, D. Contact order, transition state placement and the refolding rate of single domain proteins. *J. Mol. Biol.* **277,** 985–994 (1998).
12. Fersht, A. R., Itzhaki, L. S., elMasry, N. F., Matthews, J. M. & Otzen, D. E. Single versus parallel pathways of protein folding and fractional structure in the transition state. *Proc. Natl Acad. Sci. USA* **91,** 10426–10429 (1994).
13. Daggett, V., Li, A., Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. Structure of the transition state for folding of a protein derived from experiment and simulation. *J. Mol. Biol.* **257,** 430–440 (1996).
14. Lazaridis, T. & Karplus, M. "New View" of protein folding reconciled with the old through multiple unfolding simulations. *Science* **278,** 1928–1931 (1997).
15. Boczko, E. M. & Brooks, C. L. First-principles calculation of the folding free energy of a three-helix bundle protein. *Science* **269,** 393–396 (1995).
16. Muñoz, V. & Eaton, W. A. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl Acad. Sci. USA* **96,** 11311–11316 (1999).
17. Galzitskaya, O. V. & Finkelstein, A. V. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl Acad. Sci. USA* **96,** 11299–11304 (1999).
18. Alm, E. & Baker, D. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl Acad. Sci. USA* **96,** 11305–11310 (1999).
19. Micheletti, C., Banavar, J. R., Maritan, A. & Seno, F. Protein structures and optimal folding from a geometrical variational principle. *Phys. Rev. Lett.* **82,** 3372–3376 (1999).
20. Shoemaker, B. A., Wang, J. & Wolynes, P. G. Exploring structures in protein folding funnels with free energy functionals: The transition state ensemble. *J. Mol. Biol.* **287,** 675–694 (1999).
21. Clementi, C., Jennings, P. A. & Onuchic, J. N. How native-state topology affects the folding of dihydrofolate reductase. *Proc. Natl Acad. Sci. USA* **97,** 5871–5876 (2000).
22. Li, L., Mirny, L. A. & Shakhnovich, E. I. Kinetics, thermodynamics and evolution of non-native interactions in a protein folding nucleus. *Nature Struct. Biol.* **7,** 336–342 (2000).
23. Li, A. & Daggett, V. Characterization of the transition state of protein unfolding by use of molecular dynamics: Chymotrypsin inhibitor 2. *Proc. Natl Acad. Sci. USA* **91,** 10430–10434 (1994).
24. Vendruscolo, M. & Domany, E. Pairwise contact potentials are unsuitable for protein folding. *J. Chem. Phys.* **109,** 11101–11108 (1998).
25. Vendruscolo, M. & Domany, E. Efficient dynamics in the space of contact maps. *Fold. Des.* **3,** 329–336 (1998).
26. Zhou, Y. & Karplus, M. Interpreting the folding kinetics of helical proteins. *Nature* **401,** 400–403 (1999).
27. Tanford, C. Protein denaturation. C. Theoretical models for the mechanism of denaturation. *Adv. Protein Chem.* **24,** 1–95 (1970).
28. Berry, R. S., Beck, T. L., Davis, H. L. & Jellinek, J. in *Advances in Chemical Physics* Vol 70, (eds Prigogine, I. & Rice, S. A.) 75–135 (Wiley, New York, 1988).
29. Zhou, Y., Vitkup, D. & Karplus, M. Native proteins are surface-molten solids: Application of the Lindemann criterion for the solid versus liquid state. *J. Mol. Biol.* **285,** 1371–1375 (1999).
30. Domany, E. Superparamagnetic clustering of data. The definitive solution of an ill posed problem. *Physica A* **263,** 158–169 (1999).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank F. Chiti and A. Fersht for discussions and comments on this work. We also thank G. Ramponi and N. Taddei for continuing collaborations involving AcP. The Oxford Centre for Molecular Science is supported by BBSRC, EPSRC and MRC. M.V. is supported by an EMBP long-term fellowship. E.P. is supported in Oxford by an EC fellowship. C.M.D. is supported in part by a programme grant from the Wellcome Trust. M.K. Thanks Oxford University for inviting him to spend a year as Eastman Visiting Professor. Much of this work was done while he was in Oxford; the part done at Harvard was supported in part by the National Institutes of Health.

Correspondence and requests for materials should be addressed to C.M.D. (e-mail: chris.dobson@chem.ox.ac.uk) or M.K. (e-mail: marci@tammy.harvard.edu).