Sparsely populated folding intermediates of the Fyn SH3 domain: Matching native-centric essential dynamics and experiment

Jason E. Ollerenshaw*, Hüseyin Kaya[‡§¶], Hue Sun Chan[‡§], and Lewis E. Kay*[‡§‖]

Departments of *Chemistry, [‡]Biochemistry, and [§]Medical Genetics and Microbiology, University of Toronto, Toronto, ON, Canada M5S 1A8

Edited by Michael Levitt, Stanford University School of Medicine, Stanford, CA, and approved September 2, 2004 (received for review June 21, 2004)

A complete description of how a protein folds requires the characterization of intermediate conformations traversed during the folding transition. We have calculated dynamics trajectories of a simplified model of the Fyn SH3 domain with a native-centric potential energy function. Analysis of the resulting site-resolved energy trajectory identifies an ensemble of intermediate conformations for folding and another for unfolding. The model's folding intermediate is structured in the three β-strands that make up the protein's core and is strikingly similar to intermediates detected in a recent NMR study of Fyn SH3 folding and to folding transition states elucidated in mutagenesis studies of SH3 domains. The unfolding intermediate is formed by dissociation of the folded protein's two terminal β-strands from its core. The presence of such an intermediate is consistent with the results of a protein-engineering study on the src SH3 domain showing that these strands separate before the rate-limiting step of unfolding. Despite the presence of these conformations intermediate between the native and fully unfolded states, the computed heat capacity vs. temperature profile of the model protein indicates that its thermodynamics satisfies the usual calorimetric criterion for two-state folding. This observation highlights the fact that, if not properly interpreted, methods such as calorimetry that do not probe multiple sites in a molecule can lead to an oversimplified view of folding. The close agreement between results from this simplified model and experimental work underscores the important contributions that computational methods can make in providing insights into protein folding.

Understanding protein folding at the atomic level is a critical but elusive goal in structural biology. A protein's folded state can often be studied by x-ray crystallography or NMR spectroscopy, and recent developments in NMR methodology have made it possible to also characterize the unfolded state (1, 2). However, the transition between these states is difficult to study because intermediate conformations are most often only transiently populated. Experimentally, structural characterization of intermediates is therefore limited to indirect measurements. For example, a protein's folding transition state can be probed by studying the effects of single-residue mutations on folding and unfolding rates (3). Measurements of a protein's native state hydrogen-exchange rates can be used to identify partially unfolded substructures that cooperatively unfold. In some cases, such substructures may then be interpreted to form sequentially during the overall folding process (4). Computational techniques are not limited to indirect measurements, in principle permitting the examination of every conformation a protein passes through as it folds (5, 6). These methods are instead limited both by the accuracy of the force field that is used to evolve the system and by resources: many transitions between unfolded and folded states must be studied to accurately characterize a protein's folding reaction, but it is currently a major accomplishment to observe even a single transition in an all-atom molecular dynamics simulation.

Simplified computational models have emerged as a valuable investigative tool, allowing efficient generation of model protein-folding data with relatively limited computational resources. These models have fewer independent particles and simpler potential energy functions than are used in all-atom molecular dynamics simulations, facilitating more thorough sampling of conformational space. Results from simplified models have strongly influenced the way we conceptualize protein folding, recasting the problem in terms of energy landscapes and folding funnels (7–9). The relative effects of nonlocal interactions, local conformational preferences, and desolvation barriers on the folding transition have been investigated by testing a variety of representations of proteins (10–14). Advances in our understanding of protein folding have led to simplified models that behave more like real proteins, that exhibit thermodynamic folding cooperativity, and that can produce two-state-like folding and unfolding kinetics (14, 15).

In this work we have studied the folding and unfolding pathways of the SH3 domain from Fyn tyrosine kinase, a 59-residue domain that adopts a β-sandwich fold (16). By using the continuum (off-lattice) native-centric construct recently shown to exhibit protein-like thermodynamic cooperativity (14), transient intermediates are identified here in both the folding and unfolding transitions of Fyn SH3. By applying covariance analysis to potential energy data collected on a site-resolved basis during simulation of the unfolded protein, we have identified a group of residues that collectively form native-like structure before any other part of the protein. This partially structured intermediate is unstable and would not be expected to accumulate during the folding reaction and, in this respect, is different from those folding intermediates that are traditionally defined and identified by their accumulated populations (17, 18). Instead, the intermediates observed in the present study are more akin to the sparsely populated ''hidden'' intermediates inferred from native-state hydrogen-exchange experiments (4, 19, 20). Similarly, a partially unstructured intermediate on the protein's unfolding pathway was identified in trajectories of the folded protein. A description of each of these intermediates was obtained by collecting characteristic structures from the simulation data. Most importantly, the structures that are produced share many features with those generated on the basis of experiment (21–26). The present study demonstrates the utility of supplementing experiment with simulation and illustrates the important role that simplified protein models can play in increasing our understanding of protein folding.

Methods

Native-Centric Topological Modeling. Dynamics trajectories of a model Fyn SH3 molecule were computed with a topological modeling program published by Kaya and Chan (14). All pa-

This paper was submitted directly (Track II) to the PNAS office.

[¶]Present address: Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138.

[‖]To whom correspondence should be addressed. E-mail: kay@pound.med.utoronto.ca.

rameter values were taken from their work. No desolvation barrier was included in the present potential energy function.

The model is coarse-grained, specifying protein conformation entirely by the positions of the C^{α} atoms. Motion and energetics are simulated by using Langevin dynamics with a native-centric potential energy function. Local energy terms are based on distances, angles, and dihedrals between adjacent C^{α} atoms in the protein chain. Native-like nonlocal contacts between residues are subject to a 10–12 Lennard-Jones potential, whereas nonnative nonlocal interactions are repulsive. Minima in all energy terms are positioned to match distances, angles, and dihedrals observed in a native reference structure of the protein, and native contacts between residues are identified by using this structure. In this work, an x-ray crystal structure [PDB ID code 1SHF (27), chain B] served as the reference, and 145 native contacts were obtained by using the CSU program (28).

Two versions of the modeling program were used. One records a histogram containing the value of the potential energy function at every time step for use in simulated calorimetry. A second version calculates a separate potential energy for each residue by assigning an equal portion of each interaction energy to each of the involved residues. The energy of each residue is averaged over blocks (periods) of 400 simulation steps and recorded at the end of each averaging period for use in a covariance analysis. The Cartesian coordinates of each residue's C^{α} atom are similarly averaged and recorded in parallel to the energy data for use in a structural analysis of intermediates.

Each simulation is 200 million time steps in length and runs in ~36 h on a 1.7-GHz AMD processor (Advanced Micro Devices, Sunnyvale, CA). Data from the first million steps is discarded to allow for equilibration. The simulation's energy autocorrelation time was measured to be 5,200 ± 1,500 steps over a wide range of Langevin dynamics temperatures. The energy-scaling factor ε (as defined in ref. 14) was the same for all kinetic simulations.

Separation of Data from the Folded and Unfolded States. Prior to covariance analysis, simulation data were separated into subsets, depending on the state (folded or unfolded) of the model protein by using the following procedure: First, average state energies, $E_{\text{folded}} < E_{\text{unfolded}}$, are extracted from potential energy data by numerically minimizing

$$R = \sum_{t=1}^{N_{\text{samp}}} \min([E(t) - E_{\text{folded}}]^2, [E(t) - E_{\text{unfolded}}]^2),$$

where $E(t)$ is the model's overall potential energy measured during sampling period t, N_{samp} is the number of sampling periods, and the min function returns the lesser of its arguments. Second, the data are scanned for folding transitions, which are identified when the overall potential energy subsequently crosses thresholds at $E = E_{\text{folded}}$ and $E = E_{\text{unfolded}}$. Data collected during transitions are discarded, and sampling periods between transitions are assigned to the appropriate data subset.

Intermediate Structural Ensembles. Ensembles of structures representative of the endpoints of energy-fluctuation modes from the covariance analysis were collected as described in *Theoretical Considerations*. Each ensemble constitutes 1% of the structures observed during the relevant simulation trajectory. The structures were aligned by using a best-fit rotation (29), taking a selected element of secondary structure as a reference point. A subset of 50 structures with approximately the same mean and standard deviation in each particle coordinate as the full ensemble was then selected by using a Monte Carlo procedure. These 50 structures were plotted to visually represent intermediate states. A detailed description of this procedure is included in the supporting information, which is published on the PNAS web site.

Simulated Calorimetry. The thermodynamic folding cooperativity of the native-centric Fyn SH3 model was tested with a multiple-histogram simulated calorimetry method (30), because we found that single-simulation histogram techniques (31, 32) were not adequate for the present model. Here, a model protein's density of states as a function of energy, $n(E)$, is determined approximately from histograms of potential energy values observed during multiple simulations run at N_{temp} different temperatures. The data are combined according to

$$n(E) = \frac{\sum_{i=1}^{N_{\text{temp}}} g_i^{-1} h(E, T_i)}{\sum_{j=1}^{N_{\text{temp}}} N_j g_j^{-1} \exp[-E/T_j - f(T_j)]}, \qquad [1]$$

where $h(E, T_i)$ is the number of conformations with energy E observed during a simulation at temperature T_i (energy histogram), $N_i = \Sigma_E h(E, T_i)$, $f(T_i)$ is the free energy at temperature T_i, and the Boltzmann constant k_B has been set equal to 1. The factor $g_i = 1 + 2\tau_{\text{ac},i}$, where $\tau_{\text{ac},i}$ is the energy autocorrelation time at temperature T_i, is included to account for the fact that successive samples from a simulation are not fully independent. To obtain $\tau_{\text{ac},i}$, an exponential decay is fit to the simulation's energy autocorrelation function: $\langle E_t E_{t+\tau} \rangle \sim \exp(-\tau/\tau_{\text{ac}})$. When folding transitions are present in a data set, the autocorrelation function must be fit to a sum of two exponentials to separate $\tau_{\text{ac},i}$ from the slower folding correlation time. The free energies are defined by $\exp[f(T_i)] = \Sigma_E\, n(E)\, \exp(-E/T_i)$ and can be determined by iteration between this equation and Eq. **1** (30).

This $n(E)$ is then used to evaluate the protein's heat capacity as a function of temperature, $C_P(T)$, which can be interpreted as the results of a calorimetry experiment. After a baseline subtraction (11), the $C_P(T)$ curve is analyzed to determine the calorimetric enthalpy of folding, $\Delta H_{\text{cal}} = \int C_P(T) dT$, and the van't Hoff enthalpy of folding, $\Delta H_{\text{vH}} = 2T_{\max}\sqrt{C_P(T_{\max})}$, where $T_{\max}$ is the temperature at which $C_P(T)$ is maximal. The ratio of these enthalpies, $\kappa = \Delta H_{\text{vH}}/\Delta H_{\text{cal}}$, is a measure of the cooperativity of the protein's folding transition; $\kappa \approx 1$ is consistent with a two-state folding model. (Note that $\kappa = \kappa_2^{(s)}$ as defined in ref. 11.)

Results and Discussion

Theoretical Considerations. Covariance analysis, also known as essential dynamics or principal component analysis, identifies correlations between multiple fluctuating variables in sampled numerical data (33). It is routinely applied to atomic coordinate data from molecular dynamics simulations to find favored modes of conformational fluctuation (34) and to native-centric modeling of the dynamics of the folded state (35–37). We have attempted a similar analysis of C^{α} coordinate trajectory data from our native-centric Fyn SH3 domain model (which is described in *Methods*) by using an algorithm from the literature (33, 34). Before covariance analysis, data must be processed to remove overall translations and rotations of the model protein, which strongly interfere with the identification of coordinated internal motions. However, the present investigation of the folding/unfolding transition requires the consideration of highly disordered conformations. Their presence precludes a straightforward removal of overall rotations and consequently makes it difficult to gain insight through coordinate-based principal component analysis.

This difficulty led us to a different approach. Here, covariance

analysis is applied in the energy regime rather than the coordinate regime. In the analysis below, we use the technique to treat site-resolved energy trajectory data from the Fyn SH3 model. Motions irrelevant to folding cannot interfere with such an analysis because they do not change the model's energy. The simulation produces energy data of the form $\vec{\mathbf{E}}(t) = \sum_{i=1}^{N_{\text{res}}} e_i(t)\hat{\mathbf{e}}_i$, where N_{res} is the number of residues in the protein, $e_i(t)$ is the energy of residue i measured during sampling period t, $\hat{\mathbf{e}}_i$ is a unit vector for the energy of residue i, and $t = 1, \ldots, N_{\text{samp}}$, where N_{samp} is the number of sampling periods. Variations in residue energies during the simulation cause fluctuations in the displacement of $\vec{\mathbf{E}}(t)$ from its average position. According to García's derivation (33), the object of covariance analysis is to determine the most probable direction of this displacement by finding the unit vector $\hat{\mathbf{v}}$ that maximizes

$$f(\hat{\mathbf{v}}) = \frac{1}{N_{\text{samp}}} \sum_{t=1}^{N_{\text{samp}}} [(\vec{\mathbf{E}}(t) - \vec{\mathbf{E}}_{\text{av}})\cdot\hat{\mathbf{v}}]^2, \qquad [2]$$

where $\vec{\mathbf{E}}_{\text{av}} = N_{\text{samp}}^{-1} \sum_{t=1}^{N_{\text{samp}}} \vec{\mathbf{E}}(t)$. This expression can be simplified to $f(\hat{\mathbf{v}}) = (\mathbf{C}\hat{\mathbf{v}})\cdot\hat{\mathbf{v}}$ where $\mathbf{C}$ is the residue energy covariance matrix:

$$\mathbf{C} = \frac{1}{N_{\text{samp}}} \sum_{t=1}^{N_{\text{samp}}} (\vec{\mathbf{E}}(t) - \vec{\mathbf{E}}_{\text{av}})(\vec{\mathbf{E}}(t) - \vec{\mathbf{E}}_{\text{av}})^{\dagger}. \qquad [3]$$

In the supporting information we show that $f(\hat{\mathbf{v}})$ is maximal for vectors $\hat{\mathbf{v}}$ that satisfy $\mathbf{C}\hat{\mathbf{v}} = \lambda\hat{\mathbf{v}}$, i.e., for eigenvectors of the covariance matrix. Because $\mathbf{C}$ is symmetric, it will always be possible to find N_{res} orthonormal eigenvectors $\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \ldots, \hat{\mathbf{v}}_{N_{\text{res}}}$ with corresponding eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_{N_{\text{res}}}$. Note that $f(\hat{\mathbf{v}}_n) = \lambda_n$ for these vectors, so that by Eq. **2**, eigenvalue λ_n gives the mean square projection of $(\vec{\mathbf{E}}(t) - \vec{\mathbf{E}}_{\text{av}})$ along eigenvector $\hat{\mathbf{v}}_n$. The eigenvalues will therefore always be positive. We sort the eigenvectors in order of decreasing eigenvalue, so that $f(\hat{\mathbf{v}}_1)$ is the global maximum of $f(\hat{\mathbf{v}})$ and $f(\hat{\mathbf{v}}_2), \ldots, f(\hat{\mathbf{v}}_{N_{\text{res}}})$ are lesser maxima of decreasing significance.

The eigenvectors and eigenvalues contain valuable data on the model protein's exploration of conformational space. Each $\hat{\mathbf{v}}_n$ describes a mode of energy fluctuation favored by the model protein to an extent indicated by λ_n. This energy fluctuation must be driven by some fluctuation in the protein's conformation, the nature of which can be examined when the eigenvector is expressed in the basis of residue energies: $\hat{\mathbf{v}}_n = \sum_{i=1}^{N_{\text{res}}} v_{n,i}\hat{\mathbf{e}}_i$. Each component $v_{n,i}$ of $\hat{\mathbf{v}}_n$ reveals how the conformational fluctuation affects a different residue in the model protein. The magnitude of $v_{n,i}$ indicates how strongly the fluctuation affects the potential energy of residue i, and the relative signs of $v_{n,i}$ and $v_{n,j}$ indicate whether the fluctuation leads to a correlated or anticorrelated change in the energies of residues i and j. All covariance analysis results presented here were tested for convergence by using a procedure described in supporting information.

Although analysis of the eigenvectors efficiently identifies the model protein's dominant modes of conformational fluctuation, it is preferable to consider coordinate data when drawing conclusions about structure. We therefore wish to relate the covariance analysis results to coordinate data recorded in parallel to the energy data. Energy data $\vec{\mathbf{E}}(t)$ can be transformed from the residue energy basis $\{\hat{\mathbf{e}}_i\}$ to a basis of eigenvectors $\{\hat{\mathbf{v}}_n\}$ by using the same unitary transformation U that diagonalizes $\mathbf{C}$ (see supporting information). Thus, $\mathbf{U}\vec{\mathbf{E}}(t) = \sum_{n=1}^{N_{\text{res}}} \varepsilon_n(t)\hat{\mathbf{v}}_n$, where $\varepsilon_n(t) = \sum_{i=1}^{N_{\text{res}}} (\hat{\mathbf{v}}_n\cdot\hat{\mathbf{e}}_i)e_i(t)$ is the projection of the model protein's energy along eigenvector $\hat{\mathbf{v}}_n$ during sampling period t. Extremal points on the $\varepsilon_n(t)$ trajectory correspond to maximal conformational excursions associated with mode $\hat{\mathbf{v}}_n$. By retrieving coordinate data from those 1% of sampling periods t with the greatest values of $\varepsilon_n(t)$ we can collect an ensemble of structures characteristic of one endpoint of the fluctuation. The other endpoint can be characterized by selecting periods for which $\varepsilon_n(t)$ is smallest. The structures are then aligned relative to some element of the model protein's secondary structure, and a representative subset is selected for display and analysis. A detailed description of this procedure is presented in supporting information.

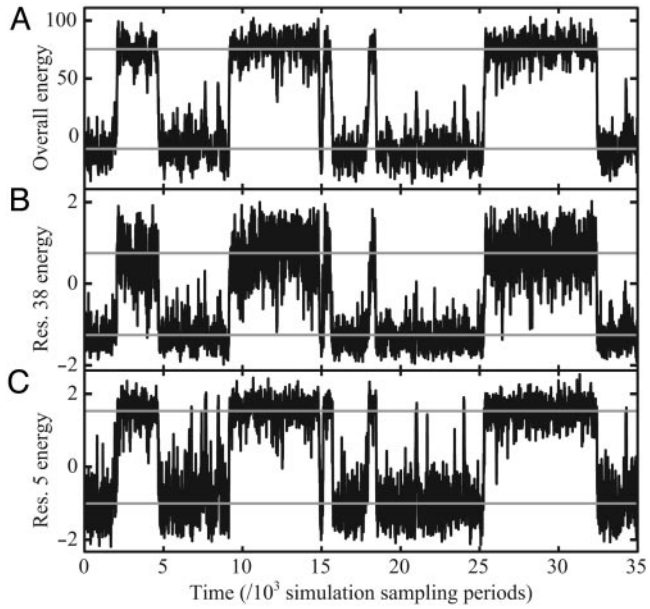

Fig. 1. Potential energy data from a portion of a dynamics trajectory of a simplified model of the Fyn SH3 domain (see *Methods*). The data are plotted as a function of simulation time for the model protein as a whole (*A*), residue 38 in the hydrophobic core (*B*), and residue 5 in the N-terminal β-strand (*C*). The upper and lower gray horizontal lines in each graph indicate the average energy for the unfolded and folded states, respectively. In *B*, fluctuations attributed to transient formation of native-like structure while the protein is unfolded can be seen, for example, at a simulation time near 12.5×10^3 sampling periods. In *C*, fluctuations due to transient dissociation of strands β1 and β5 from the folded protein's core are apparent at a simulation time near 7.5×10^3 sampling periods. For the purposes of this figure only, the data have been smoothed by boxcar averaging with an 11-point window.

Covariance Analysis Identifies Transient Protein-Folding Intermediates. Site-resolved energy data were recorded during a single simulation of the native-centric Fyn SH3 model near its folding midpoint temperature. Fig. 1 shows typical data for a small portion of the trajectory. The model's overall potential energy (Fig. 1*A*) exhibits only two stable states, folded and unfolded, and each of the 145 observed transitions between these states is rapid. The potential energy of each residue in the model protein similarly shows two distinct states, and abrupt transitions between folded and unfolded occur synchronously for all residues. Despite the concerted nature of the transitions, significant variations in energetics exist between different sites. Some residues in the hydrophobic core (e.g., residue 38, Fig. 1*B*) exhibit large energy fluctuations while the protein is unfolded, with energies briefly reaching average values for the folded configuration. The fluctuations of several core residues seem to be correlated (e.g., the residues in strand β3), suggesting a cooperative process. Likewise, residues in the N- and C-terminal strands (e.g., residue 5, Fig. 1*C*) undergo large, coordinated energy fluctuations while the protein is folded.

Covariance analysis was used to quantitatively characterize correlations in the site-resolved energy data. Because the model is native-centric, any major energy change must involve the

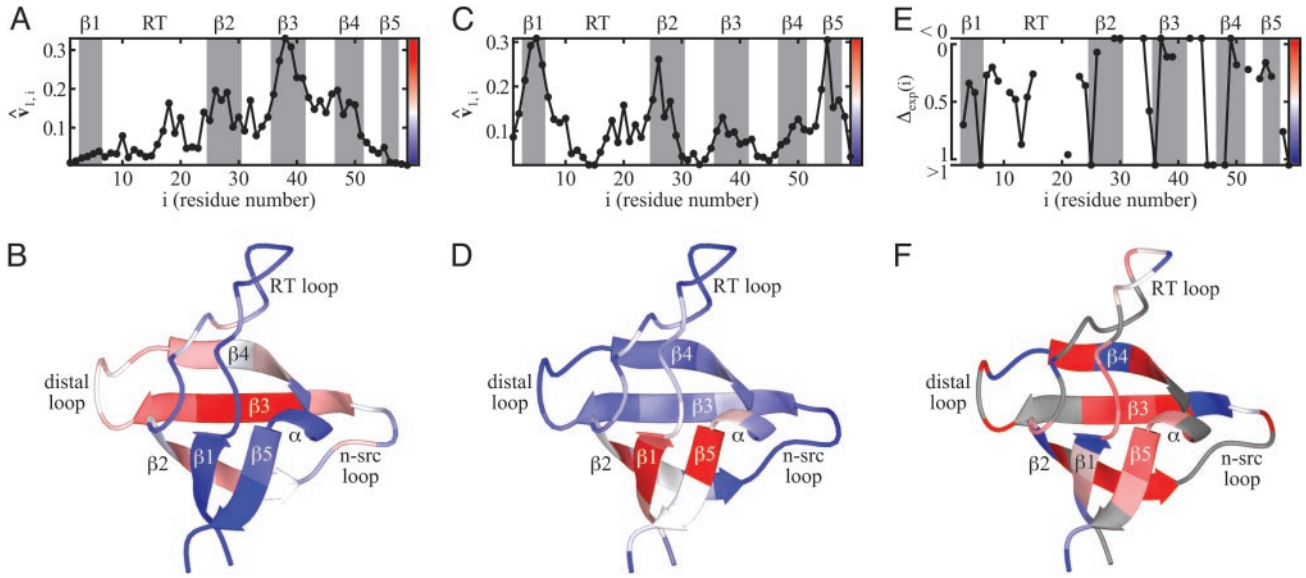

Fig. 2. Results of the energy-based covariance analysis and comparison with experiment. The most significant coordinated energy fluctuations observed in unfolded (*A* and *B*) and folded (*C* and *D*) subsets of the simulation data are represented by the components of the first covariance matrix eigenvector for these data sets. The components, which indicate the extent to which each residue in the protein participates in the fluctuation, are plotted as a function of residue number (*A* and *C*) and mapped as colors onto the protein's native structure (*B* and *D*). Experimentally determined Δ_{exp} values for the G48V mutant of Fyn SH3 (see text and ref. 21) are plotted as a function of residue number (*E*) and on the structure of the Fyn SH3 domain (*F*). Residues for which no NMR data could be obtained are omitted from *E* and colored gray in *F*.

formation or dissolution of native-like structure. The analysis, therefore, identifies cooperative, multiresidue interactions related to folding. The energy fluctuations of many residues change significantly in character when the protein folds or unfolds, so the data were divided into folded and unfolded subsets comprising 96,184,400 and 97,263,600 simulation steps, respectively (240,461 and 243,159 sampling periods), with separate covariance analyses performed on each.

Covariance analysis of the unfolded energy trajectory identified a transient intermediate on the protein's folding pathway. The most significant energy fluctuation mode found by the analysis, $\hat{\mathbf{v}}_1$, accounts for 22% of the unfolded protein's mean square energy fluctuations (i.e., $\lambda_1/\sum_{i=1}^{N_{res}}\lambda_i = 0.22$, where λ_i is an eigenvalue of the energy covariance matrix defined in Eq. **3**). The components of $\hat{\mathbf{v}}_1$ (Fig. 2*A*) indicate that the fluctuation is mostly localized in strands β2, β3, and β4 and the distal loop. This is especially clear when the components are mapped onto a 3D structure of Fyn SH3 (Fig. 2*B*). The mode apparently involves the cooperative formation of the three central strands into a core of native-like structure in an otherwise unfolded protein. Strand β3 makes the greatest contribution to the structure, whereas β2 and β4 are less strongly involved. The data also indicate that the RT loop is weakly involved at the point where it contacts β4.

As described in the preceding section, it is possible to obtain a structural representation of the endpoints of conformational fluctuations associated with a given mode $\hat{\mathbf{v}}_n$ by selecting structures that correspond to energy extrema in $\varepsilon_n(t)$. Fig. 3*A* shows an ensemble of 50 structures representing the minimum $\varepsilon_1(t)$ endpoint of the $\hat{\mathbf{v}}_1$ mode from the trajectory of the unfolded state. These structures correspond, therefore, to the lowest energy states along the folding excursion characterized by $\hat{\mathbf{v}}_1$. Strands β2, β3, and β4 and the loops connecting them form a well defined native-like sheet structure, as would be expected from the eigenvalue components (Fig. 2 *A* and *B*). The ensemble of Fig. 3*A* includes structures from 1% of the unfolded trajectory data (961,600 simulation steps, 2,404 sampling periods; see supplemental information). The ensemble establishes that strands β2–β4 must frequently come together to form local native-like structure while the protein is unfolded, producing a partially structured intermediate on the protein's folding pathway.

A separate covariance analysis of the energy data subset corresponding to the folded state has revealed a transient intermediate on the protein's unfolding pathway. Mode $\hat{\mathbf{v}}_1$, accounting for 32% of the folded protein's mean square energy fluctuations, has the components shown in Fig. 2 *C* and *D*. This mode affects strand β1, strand β5 (especially residue 55, though eigenvector components are elevated throughout the strand), and that portion of strand β2 that contacts β1 in the native structure. A picture emerges in which strands β1 and β5 frequently dissociate from the rest of the β-sheet while the protein as a whole remains folded, forming a transient unfolding intermediate. The presence of the intermediate is confirmed by an ensemble of structures collected from the simulation data corresponding to maximum $\varepsilon_1(t)$ values from the trajectory of the folded protein (Fig. 3*B*). In these structures, which comprise 1% of the folded trajectory (corresponding to 972,400 simulation steps or 2,431 sampling periods), strands β1 and β5 are disordered but the core strands remain intact.

Energy fluctuation modes beyond $\hat{\mathbf{v}}_1$ were investigated in both the folded and unfolded covariance analysis results. From the analysis of data derived from the unfolded protein, mode $\hat{\mathbf{v}}_2$, which accounts for 13% of the mean square energy fluctuations in the unfolded state, shows an energy anticorrelation between residues in the core of the protein and residues in the RT loop. The anticorrelation is a consequence of the protein's two-state nature; these regions cannot both assume their native conformations without a folding transition, so a balance in their energies is observed. Modes beyond $\hat{\mathbf{v}}_2$ for the unfolded data and beyond $\hat{\mathbf{v}}_1$ for the folded data each account for only a small fraction of the model protein's mean square energy fluctuations and cannot be interpreted easily in terms of conformational fluctuations.

Comparison to Experimental Measurements. It is illuminating to compare the findings described above with the results of our

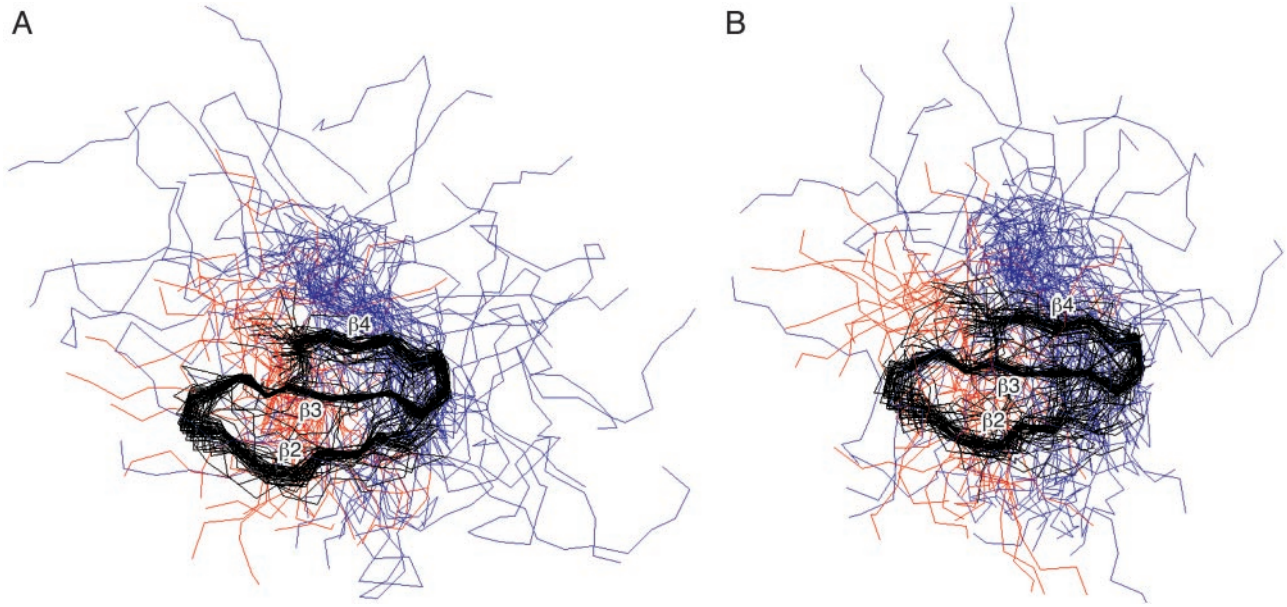

Fig. 3. Covariance analysis structural ensembles. Transient intermediates on the protein's folding (*A*) and unfolding (*B*) pathways are each represented by 50 structures collected from the simulation data as described in the text. The N-terminal (strand β1, RT loop), core (strands β2, β3, β4; n-src, distal loops), and C-terminal (3_{10} helix, strand β5) portions of the protein are drawn in blue, black, and red, respectively. Each of the structures in the ensembles is aligned with respect to strand β3.

recent relaxation dispersion NMR spectroscopy study of the folding of a pair of Fyn SH3 point mutants (G48M and G48V), which are less stable than the wild-type protein but have a significantly greater folding rate. For both mutants, a low population folding intermediate was detected in equilibrium with the folded and unfolded states (21). A comparison of chemical shifts measured for the intermediate, folded, and unfolded states established that strands β2, β3, and β4 and the distal loop connecting β3 and β4 are reasonably well structured in the intermediates of both the G48M and G48V mutants, with significantly less structure in other regions of the protein. The formation of structure in the central strands of one of the intermediates is illustrated in Fig. 2*E* in which Δ_{exp} values for the G48V mutant are plotted vs. residue number. The value of $\Delta_{exp}(i) = [\delta_F(i) - \delta_I(i)]/[\delta_F(i) - \delta_U(i)]$, where $\delta_k(i)$ is the ^{15}N chemical shift of residue i in state k [k = folded (F), intermediate (I), or unfolded (U)], provides a measure of the extent to which amide site i in the protein is folded in the I state. Values of $\Delta_{exp}(i)$ close to zero indicate native-like structure at this position, whereas values close to unity are consistent with an unfolded-like conformation (21). The $\Delta_{exp}(i)$ values are plotted on the structure of the folded state of the Fyn SH3 domain in Fig. 2*F*. Comparing Fig. 2 *B* and *F*, a very significant correlation clearly exists between the folding intermediate identified by simulation in our simplified Fyn SH3 model and by experiment.

Our results also have interesting similarities to those of protein-engineering studies of several homologous SH3 domains, in which folding transition states were characterized by using Φ-values from kinetic studies of mutants. Northey *et al.* (22) identified in the Fyn SH3 domain a "core folding nucleus" of residues in strands β2, β3, and β4 that are important for transition-state stabilization. Multiple residues in the distal β-hairpin (comprising strands β3 and β4 and the distal loop) and one in strand β2 are strongly implicated in the transition state for src SH3 (23). Residues in the distal β-hairpin are also highly structured in the transition state for the α-spectrin SH3 domain (24). Another study concluded that formation of the distal loop in α-spectrin SH3 in an obligatory step in forming the transition state, because it is necessary to bring together strands β3 and β4 (25). With the exception of certain transition-state features that are not conserved among these homologous proteins [i.e., structuring of the C-terminal residues in the RT loop in src SH3 (23) and of the 3_{10} helix in α-spectrin SH3 (24)], a strong correlation is observed between experiment and our simulation results.

The nature of the model's unfolding intermediate (Fig. 3*B*) is also consistent with previous experimental results. In a study on the src SH3 domain it was found that creating a disulfide crosslink between strands β1 and β5 greatly decreases the protein's unfolding rate, which indicates that dissociation of the N- and C-terminal strands is an early step on the unfolding pathway (26).

Comparison with the Two-State Model of Protein Folding. The folding behavior of Fyn SH3 (38) and its homologs src SH3 (39) and spectrin SH3 (40) have previously been characterized as consistent with the two-state model of protein folding. In general, evidence of two-state folding is drawn from measurements that do not resolve multiple sites within a protein molecule; these include data from calorimetry, CD spectroscopy, and fluorescence emission spectroscopy (41). With this in mind we have investigated the folding of our native-centric Fyn SH3 model with simulated calorimetry techniques.

A simulated calorimetry method (see *Methods* and ref. 30) was applied to data from 15 Fyn SH3 simulations run at a wide range of temperatures and the heat capacity of the model domain was calculated as a function of temperature, $C_P(T)$ (Fig. 4). The resulting heat-capacity profile has a single sharp peak and κ = 0.987 (the calorimetric two-state criterion is $\kappa \approx 1$), indicating that the thermodynamic folding behavior of the model is well described by a two-state model by conventional standards (42).

It is interesting to consider the covariance analysis results in light of this finding. The folding and unfolding intermediates that we have observed are consistent with two-state folding because they do not accumulate. Indeed, in-depth theoretical analyses have shown that the usual calorimetric two-state criterion does not imply that conformations with intermediate energies are nonexistent and that the criterion, in fact, allows for a minute population of such conformations (11, 15, 43), although it does

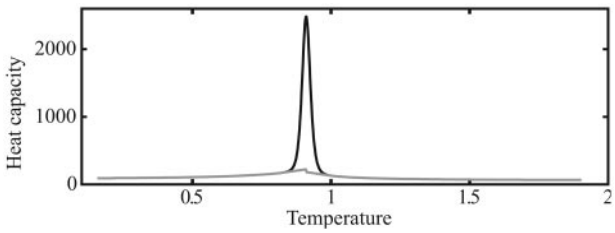

Fig. 4. Simulated calorimetry results. The model protein's heat capacity as a function of temperature, $C_P(T)$, was calculated from the model's density of states. A van't Hoff to calorimetric enthalpy ratio of $\kappa = 0.987$ was calculated after subtraction of an empirical baseline (gray line).

imply that such a population cannot be large (4, 11, 44). Nevertheless, the sparsely populated intermediates revealed in the present model study expose subtleties in the folding mechanism that a simple two-state picture would not lead us to consider. Similar observations have been made in site-resolved experimental studies of protein folding, including our NMR study of fast-folding Fyn SH3 mutants (21), in which intermediates on the order of 1% of the total population of protein molecules in solution were observed. These results emphasize that experimental measures that do not probe multiple sites within a molecule can mislead us toward an oversimplified view of the folding reaction.

Conclusion

In summary, we have identified folding and unfolding intermediates in a native-centric computational model of the Fyn SH3 domain. The folding intermediate, observed while the protein as a whole was unfolded, is a partially structured state formed by association of the protein's three central β-strands. It is strikingly similar to Fyn SH3 folding intermediates recently characterized by relaxation dispersion NMR methods (21) and to folding transition states elucidated in point mutagenesis studies of several homologous SH3 domains (22–25). The unfolding intermediate is a partially unstructured state formed when the two terminal β-strands dissociate from the folded domain; structures of this sort are also inferred from experiment (26). Because our model is native-centric, to a significant degree, the experimentally observed intermediates are a consequence of native-like interactions (10–14). The present investigation complements recent simulation studies on homologous SH3 domains (13, 45–48). Our finding of sparsely populated intermediates, indicating early formation of the central β-sheet during the folding process, is in general agreement with previous simulation results (45–47). The present approach is limited in that it uses only a reduced chain representation, solvation (13, 14, 45, 48, 49) is not explicitly treated, and aspects of kinetic cooperativity (15) are yet to be addressed. Nonetheless, the simplicity of our model allows for broad conformational sampling and energy covariance and thermodynamic analyses over many cycles of reversible folding/unfolding transitions, which are currently difficult to achieve in higher-resolution models. The model's folding/unfolding behavior was found to display apparent two-state thermodynamics; techniques such as calorimetry or CD that measure only global parameters would, therefore, give no information about the presence of intermediates along the folding trajectory. This emphasizes the importance of methods (both computational and experimental) that can provide information on a per-residue basis to supplement the more traditional approaches that have been used to study protein-folding dynamics.

J.E.O. received a postgraduate scholarship from the Natural Sciences and Engineering Research Council of Canada. This research was supported by grants from the Canadian Institutes of Health Research and the Canada Research Chair Program (to H.S.C. and L.E.K.) and by the Protein Engineering Network of Centres of Excellence and the Premier's Research Excellence Awards (to H.K. and H.S.C.).

1. Dyson, H. J. & Wright, P. E. (1998) *Nat. Struct. Biol.* **5,** Suppl., 499–503.
2. Shortle, D. (1993) *Curr. Opin. Struct. Biol.* **3,** 66–74.
3. Matouschek, A., Kellis, J. T., Serrano, L. & Fersht, A. R. (1989) *Nature* **340,** 122–126.
4. Bai, Y., Sosnick, T. R., Mayne, L. & Englander, S. W. (1995) *Science* **269,** 192–197.
5. Duan, Y. & Kollman, P. A. (1998) *Science* **282,** 740–744.
6. Mayor, U., Guydosh, N. R., Johnson, C. M., Grossman, J. G., Sato, S., Jas, G. S., Freund, S. M. V., Alonso, D. O. V., Daggett, V. & Fersht, A. R. (2003) *Nature* **421,** 863–867.
7. Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. (1995) *Proteins Struct. Funct. Genet.* **21,** 167–195.
8. Thirumalai, D. & Woodson, S. A. (1996) *Acc. Chem. Res.* **29,** 433–439.
9. Dill, K. A. & Chan, H. S. (1997) *Nat. Struct. Biol.* **4,** 10–19.
10. Micheletti, C., Banavar, J. R., Maritan, A. & Seno, F. (1999) *Phys. Rev. Lett.* **82,** 3372–3375.
11. Kaya, H. & Chan, H. S. (2000) *Proteins Struct. Funct. Genet.* **40,** 637–661.
12. Clementi, C., Nymeyer, H. & Onuchic, J. N. (2000) *J. Mol. Biol.* **298,** 937–953.
13. Cheung, M. S., García, A. E. & Onuchic, J. N. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 685–690.
14. Kaya, H. & Chan, H. S. (2003) *J. Mol. Biol.* **326,** 911–931.
15. Chan, H. S., Shimizu, S. & Kaya, H. (2004) *Methods Enzymol.* **380,** 350–379.
16. Kuriyan, J. & Cowburn, D. (1993) *Curr. Opin. Struct. Biol.* **3,** 828–837.
17. Kim, P. S. & Baldwin, R. L. (1982) *Annu. Rev. Biochem.* **51,** 459–489.
18. Kim, P. S. & Baldwin, R. L. (1990) *Annu. Rev. Biochem.* **59,** 631–660.
19. Ozkan, S. B., Dill, K. A. & Bahar, I. (2002) *Protein Sci.* **11,** 1958–1970.
20. Vu, N. D., Feng, H. Q. & Bai, Y. W. (2004) *Biochemistry* **43,** 3346–3356.
21. Korzhnev, D. M., Salvatella, X., Vendrusculo, M., Di Nardo, A. A., Davidson, A. R., Dobson, C. M. & Kay, L. E. (2004) *Nature* **430,** 586–590.
22. Northey, J. G. B., Di Nardo, A. A. & Davidson, A. R. (2002) *Nat. Struct. Biol.* **9,** 126–130.
23. Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I. & Baker, D. (1999) *Nat. Struct. Biol.* **6,** 1016–1024.
24. Martínez, J. C. & Serrano, L. (1999) *Nat. Struct. Biol.* **6,** 1010–1016.
25. Martínez, J. C., Pisabarro, M. T. & Serrano, L. (1998) *Nat. Struct. Biol.* **5,** 721–729.
26. Grantcharova, V. P., Riddle, D. S. & Baker, D. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 7084–7089.
27. Noble, M. E. M., Musacchio, A., Saraste, M., Courtneidge, S. A. & Wierenga, R. K. (1993) *EMBO J.* **12,** 2617–2624.
28. Sobolev, V., Sorokine, A., Prilusky, J., Abola, E. E. & Edelman, M. (1999) *Bioinformatics* **15,** 327–332.
29. Kabsch, W. (1978) *Acta Crystallogr. A* **34,** 827–828.
30. Ferrenberg, A. M. & Swendsen, R. H. (1989) *Phys. Rev. Lett.* **63,** 1195–1198.
31. Socci, N. D. & Onuchic, J. N. (1995) *J. Chem. Phys.* **103,** 4732–4744.
32. Ferrenberg, A. M. & Swendsen, R. H. (1988) *Phys. Rev. Lett.* **61,** 2635–2638.
33. García, A. E. (1992) *Phys. Rev. Lett.* **68,** 2696–2699.
34. Amadei, A., Linssen, A. B. M. & Berendsen, H. J. C. (1993) *Proteins Struct. Funct. Genet.* **17,** 412–425.
35. Jacobs, D. J., Rader, A. J., Kuhn, L. A. & Thorpe, M. F. (2001) *Proteins Struct. Funct. Genet.* **44,** 150–165.
36. Keskin, O., Bahar, I., Flatow, D., Covell, D. G. & Jernigan, R. L. (2002) *Biochemistry* **41,** 491–501.
37. Micheletti, C., Lattanzi, G. & Maritan, A. (2002) *J. Mol. Biol.* **321,** 909–921.
38. Plaxco, K. W., Guijarro, J. I., Pitkeathly, M., Campbell, I. D. & Dobson, C. M. (1998) *Biochemistry* **37,** 2529–2537.
39. Grantcharova, V. P. & Baker, D. (1997) *Biochemistry* **36,** 15685–15692.
40. Viguera, A. R., Martínez, J. C., Filimonov, V. V., Mateo, P. L. & Serrano, L. (1994) *Biochemistry* **33,** 2142–2150.
41. Fersht, A. R. (1999) *Structure and Mechanism in Protein Science* (Freeman, New York).
42. Privalov, P. L. (1979) *Adv. Protein Chem.* **33,** 167–241.
43. Englander, S. W., Mayne, L., Bai, Y. & Sosnick, T. R. (1997) *Protein Sci.* **6,** 1101–1109.
44. Klimov, D. K. & Thirumalai, D. (2002) *J. Comput. Chem.* **23,** 161–165.
45. Shea, J.-E., Onuchic, J. N. & Brooks, C. L., III (2002) *Proc. Natl. Acad. Sci. USA* **99,** 16064–16068.
46. Guo, W., Lampoudi, S. & Shea, J.-E. (2003) *Biophys. J.* **85,** 61–69.
47. Settanni, G., Gsponer, J. & Caflisch, A. (2004) *Biophys. J.* **86,** 1691–1701.
48. Fernandez-Escamilla, A. M., Cheung, M. S., Vega, M. C., Wilmanns, M., Onuchic, J. N. & Serrano, L. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 2834–2839.
49. Papoian, G. A., Ulander, J., Eastwood, M. P., Luthey-Schulten, Z. & Wolynes, P. G. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 3352–3357.