JCTC Journal of Chemical Theory and Computation

Transition Networks for the Comprehensive Characterization of Complex Conformational Change in Proteins

Frank Noé,[†,‡] Dieter Krachtus,[†,‡] Jeremy C. Smith,[†] and Stefan Fischer*[,‡]

Computational Molecular Biophysics, Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Im Neuenheimer Feld 368, 69120 Heidelberg, Germany, and Computational Biochemistry, Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Im Neuenheimer Feld 368, 69120 Heidelberg, Germany

Received June 24, 2005

Abstract: Functionally relevant transitions between native conformations of a protein can be complex, involving, for example, the reorganization of parts of the backbone fold, and may occur via a multitude of pathways. Such transitions can be characterized by a transition network (TN), in which the experimentally determined end state structures are connected by a dense network of subtransitions via low-energy intermediates. We show here how the computation of a TN can be achieved for a complex protein transition. First, an efficient hierarchical procedure is used to uniformly sample the conformational subspace relevant to the transition. Then, the best path which connects the end states is determined as well as the rate-limiting ridge on the energy surface which separates them. Graph-theoretical algorithms permit this to be achived by computing the barriers of only a small number out of the many subtransitions in the TN. These barriers are computed using the Conjugate Peak Refinement method. The approach is illustrated on the conformational switch of Ras p21. The best and the 12 next-best transition pathways, having rate-limiting barriers within a range of 10 kcal/mol, were identified. Two main energy ridges, which respectively involve rearrangements of the switch I and switch II loops, show that switch I must rearrange by threading Tyr32 underneath the protein backbone before the rate-limiting switch II rearrangement can occur, while the details of the switch II rearrangement differ significantly among the low-energy pathways.

1. Introduction

Conformational changes are critical to the function of many proteins. Well-known examples of functional transitions include the rearrangement of subunits in the hemoglobin tetramer upon oxygen binding,[1] the lever-arm motion in myosin during muscle contraction,[2] and the molecular switch in Ras p21 (see Figure 2) that signals cell division.[3,4] Such functional changes in conformation are often complex, involving the rearrangement of backbone segments or the packing at domain interfaces. Understanding the mechanism of these transitions is particularly challenging, because the nature and order of their subtransitions are difficult to predict and may, in principle, occur in different ways.

X-ray crystallography and nuclear magnetic resonance spectroscopy can provide atomic-detail structures for stable end states of conformational transitions and sometimes long-lived intermediates. However, the transitions themselves are difficult to characterize experimentally because, although the time required for a complete structural change can be relatively long (μs or longer), the transition states involved

* Corresponding author e-mail: stefan.fischer@iwr.uni-heidelberg.de.
† Computational Molecular Biophysics.
‡ Computational Biochemistry.

10.1021/ct050162r CCC: $33.50
Published on Web 03/11/2006

are very short-lived. Computer simulation can help to gain insight into these processes.

Because complex conformational transitions usually occur on long time scales, they are not accessible to unbiased molecular dynamics (MD) simulation with presently available computing power. Consequently, alternative computational approaches must be used. Variations of molecular dynamics have been proposed to overcome this time scale problem. For example, multiple time-step methods[5] are quite successful in certain multiple-time scale contexts. However, they do not achieve sufficient speedup for the present purposes.[6,7] Other methods bias the underlying energy potential[8] or reduce the dimension of the conformational space.[9] These methods face the difficulty that a good guess of the energy surface along the whole transition must in principle be known a priori, which is usually not possible for complex transitions in proteins. Steered and Targeted Molecular Dynamics[10,11] incorporate a constraint into the energy function that directs the system toward the desired product structure. While these methods are successful in cases where the reaction follows a pathway that is compatible with these constraints,[12] they lead to unnatural structures and unrealistic energy barriers in other cases.[13] A further variant is conformational flooding,[14] which approximates the local shape of the underlying energy surface explored by an MD trajectory by computing its principal components and then escapes the local energy minimum by adding a multivariate Gaussian function to the energy function the form of which depends on these principal components. Although this method allows the trajectory to overcome high energy barriers, it is designed to explore yet unknown new states rather than performing a transition into a predefined state.

Pathway methods are a different approach to simulating molecular transitions. Starting from an initial guess, the transition pathway is allowed to relax on the energy surface by constrained molecular dynamics[15,16] or by local minimization methods.[17−21] These methods have been applied successfully in cases where the transition does not involve too complex rearrangements of the protein, such that a number of reasonable initial guesses of the pathway can easily be formulated.[22−24] This is particularly the case when the conformational distance between the transition end states is small. In contrast, when the transition involves rearrangements of the protein fold, a guess for the initial path is more difficult to make. Moreover, such transitions can follow multiple pathways, as the energy landscape is likely to include broad energy ridges with many saddle-points of similar energies. Therefore, the determination of a single reaction pathway (even if it is the lowest-energy one) does not yield a comprehensive description of the transition.[13]

To represent multiple pathways, the *transition network* approach may be used. Transition networks are a discrete and simplified representation of configurational space and encode the possible transition pathways in a network of subtransitions. Each subtransition occurs between two conformations that are relatively close in conformational space. Each conformation in the network can be reached and left through at least one, but usually several, subtransitions. Each subtransition has an associated energy barrier that can be used to determine an associated rate constant or a mean passage time (i.e. "cost"). See Figure 1 for an illustration.

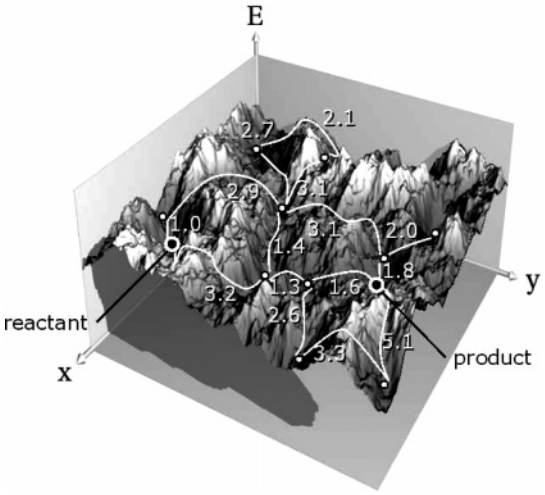

Figure 1. Transition network on a schematic two-dimensional energy surface. The network vertices (white bullets) correspond to low-energy intermediates between the reactant and product end states of the transition (black bullets). The network edges (white lines) correspond to subtransitions between the vertices and are associated with weights (white numbers), which are the rate-limiting energy barriers along each subtransition.

The construction of transition networks is documented in a large number of studies which have addressed the analysis of energy surfaces by mapping its local minima and saddle points.[25−46] These stationary points can be generated by local optimization starting from conformational ensembles that are generated by high-temperature molecular dynamics,[29,33,36,39,47] by a mode-following guided parallel search starting from a deep initial minimum,[35,42,48] or by Discrete Path Sampling (DPS).[43,45,49] The kinetics between groups of stationary points may be recovered using Master-Equation dynamics (MED),[28,29,32,36,39,35,38−40,42,43,45,46] Kinetic Monte Carlo (KMC),[45] or, again, by Discrete Path Sampling (DPS).[43,45,49] Typical applications of the above methodology are the rearrangement of atomic or molecular clusters[29−31] and the rearrangement or folding of peptides[27,28,32,33,36,35,38,40,43,45] and of model proteins.[34,42,50]

The applicability of the above approaches to complex transitions between native conformations of a protein is limited by two main difficulties. The first involves the generation of the minima which serve as TN vertices: It is a priori unclear how a conformational ensemble can be generated that adequately covers the volume of conformational space that is relevant for the transition. In particular, the direct manipulation of the backbone torsion angles or high-temperature dynamics are likely to disrupt the native structure, while search-based procedures may get lost in the huge number of possibly distant low-energy minima. Discrete Path Sampling is likely to be successful in identifying a connected channel between the end states, but it is unclear how it can identify a collection of considerably different channels. The second problem involves the computation of energy barriers. The determination of global properties of the network, such as the kinetics or the optimal path between

two end states,[51] requires the barriers of the subtransitions in the network to be known. Dense transition networks for complex macromolecular transitions typically have so many edges and the computation of each subtransition barrier is so CPU-demanding that the computation of all subtransition barriers cannot be afforded.

In the present contribution we address these two problems. In section 3.1 we present a procedure for efficiently sampling the relevant degrees of freedom of a complex transition between two native conformations of a protein so as to yield a representative set of low-energy conformers. Then, in section 3.2 we present a graph-theoretical approach that allows for determining global network properties (such as the best transition pathway or the rate-limiting ridge dividing the energy surface into reactant and product basins) based on the computation of only a limited subset of all subtransition barriers.

The methods introduced here are used to identify likely pathways and the order of events in an example system, the molecular switch of Ras p21 (section 4). This switch plays an important role in the signal transduction pathways that control proliferation, differentiation, and metabolism.[3,4] The conformational transition, which occurs in the GDP-bound state, involves a complex rearrangement of the backbone fold around the nucleotide binding site (Figure 2). The complexity of the transition suggests that it may occur via many different pathways, thus making it an excellent case for testing the present transition network approach. Among the questions that have been raised by a previous study of this conformational transition[13] and that are addressed here are as follows: (1) Is the rearrangement of switch I characterized by the side chain of Tyr32 threading underneath the backbone or by moving it through the solvent (see Figure 2)? (2) Is there a coupling between the switch I and switch II transitions, i.e. is the relative order of events in the two switch regions strictly defined? (3) Is there a well-defined unfolding pathway of switch II?

The present study reports on methodological advances which permit the generation and analysis of a comprehensive set of pathways for a complex conformational transition between two native conformations of a protein. The methodology is applicable to complex conformational changes in many other proteins whose functional time scale and complexity precludes the use of direct simulation.

2. Theory

Transition networks (TN) can in principle be used to model any dynamical system that can be appropriately described by a (possibly large) number of states and interstate transition rules. Here, we focus on TN for molecular systems and in particular for conformational changes in proteins. A TN is a discrete model which abstracts dynamic properties from the full system and captures its relevant kinetic behavior. Formally, a TN is a weighted graph, $\mathcal{G} = (\mathcal{V},\mathbf{X}^S,\mathbf{E}^S,\mathcal{E},\mathbf{X}^{TS},\mathbf{E}^{TS})$, where the list of vertices, $\mathcal{V} = (1,...,|\mathcal{V}|)$, represents the stable states of the protein, $\mathbf{X}^S = (\mathbf{x}_1^S,...,\mathbf{x}_{|\mathcal{V}|}^S)$ are the corresponding configuration vectors, and $\mathbf{E}^S = (E_1^S,..., E_{|\mathcal{V}|}^S)$ are corresponding state energies. The list of edges, $\mathcal{E} = ((u, v),...,(w, y))$, specifies between which pairs of states direct subtransitions are considered. $\mathbf{X}^{TS} = (\mathbf{x}_{uv}^{TS},..., \mathbf{x}_{wy}^{TS})$ are the configuration vectors of the corresponding transition states and $\mathbf{E}^{TS} = (E_{uv}^{TS},..., E_{wy}^{TS})$ are the associated transition state energies.

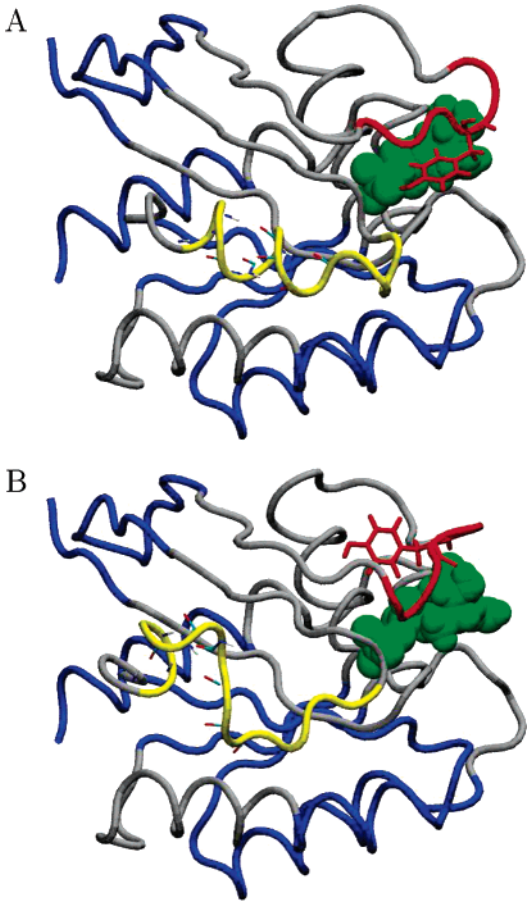

Figure 2. The conformational switch in Ras p21. (A) The GTP-bound and (B) GDP-bound conformers of the Ras p21 transition. The blue regions are very similar in the crystallographic end states and were kept fixed during the simulations. During the transition, switch I (residues 30−35, in red) rearranges such that Tyr32 (shown in red) is repositioned on the opposite site of the backbone and opens the nucleotide binding site to prepare for the release of GDP (shown in green van der Waals spheres). Switch II (residues 61−71, in yellow) unfolds from a helix to a coil structure.

Each TN vertex, v, corresponds to a region R_v of the configurational space, containing a group of geometrically similar molecular configurations. What is appropriate as a definition of "group" depends on the application. For the current work, each given vertex v corresponds to an *attraction basin*, i.e. the set of configurations that can be mapped to the same local minimum $\mathbf{x}_v^S$ on the potential energy surface $U(\mathbf{x})$ by a direct minimization.[25,52] Each vertex, v, is associated with a state energy E_v^S, generally, the free energy of the basin R_v. Depending on the complexity of the system used, different approximations to the free energy may be necessary. For systems where all basins can be mapped, it has been shown that free energies calculated from harmonic approximations to the potential in each basin are able to reproduce thermodynamic properties.[42,43]

The TN edges represent subtransitions between pairs of neighboring vertices. To any given edge (u, v), there is an associated transition state configuration, $\mathbf{x}_{uv}^{TS}$. The energy E_{uv}^{TS} associated with the edge is generally the free energy of the transition state.

The absolute height of the vertex and edge energies can be shifted by subtracting an arbitrary constant value E_0 without affecting the results. To avoid numerical problems when using exponentials of E_{uv}^{TS}, it is desirable to choose E_0 in such a way as to keep E_{uv}^{TS} small.

Figure 1 shows a schematic representation of a transition network.

2.1. Best Paths. Given the weighted graph $\mathcal{G}$, one can search for the "best" path connecting two particular vertices v_R and v_P (e.g. corresponding to experimentally determined "reactant" and "product" structures). For this, consider a subtransition along one edge between two vertices, $u \rightarrow v$, with associated energies E_u^S and E_{uv}^{TS}. We can express the flux from state u into state v, k_{uv}, as a product of the probability of being in state u, p_u, with the rate constant for the transition $u \rightarrow v$, k'_{uv}:[53]

$$k_{uv} = p_u k'_{uv} \tag{1}$$

Using free energies for vertices and edges, the probability p_u at equilibrium is given by the ratio of the partition functions for u and the full phase space:

$$p_u = \frac{\exp(-E_u^S/k_B T)}{\sum_{w=1}^{|\mathcal{V}|} \exp(-E_w^S/k_B T)} \tag{2}$$

Using the phenomenological form of transition state theory,[54] and assuming that all subtransitions have a similar dynamic prefactor, γ, the rate constant can be expressed as

$$k'_{uv} = \gamma \exp\frac{-(E_{uv}^{TS} - E_u^S)}{k_B T} \tag{3}$$

where k_B is Boltzmann's constant, and T is the temperature. Substituting eqs 2 and 3 into eq 1, we see that the equilibrium flux for the transition $u \rightarrow v$ is proportional to the Boltzmann weight of E_{uv}^{TS}:

$$k_{uv} \propto \exp\left(-\frac{E_{uv}^{TS}}{k_B T}\right) \tag{4}$$

The mean time between two subsequent transition events from u to v, τ_{uv}, is given by the inverse of the flux: $\tau_{uv} = k_{uv}^{-1}$. We take the *edge cost*, $\bar{w}_{uv}$, as proportional to τ_{uv}, setting the proportionality factor to unity:

$$\bar{w}_{uv} = \exp\left(\frac{E_{uv}^{TS}}{k_B T}\right) \tag{5}$$

For a path connecting vertices $v_1 = v_R$ and $v_m = v_P$ via a series of m vertices, $P = (v_1, v_2, ..., v_m)$, travelling over edges $((v_1, v_2), ..., (v_{m-1}, v_m))$, the best path is defined as that which minimizes the cumulative edge costs

$$C(P) = \sum_{k=1}^{m-1} \bar{w}_{v_k v_{k+1}} \tag{6}$$

This definition of a best path is similar to the previously proposed notion of the continuous pathway with "maximum flux" or "minimum resistance".[15,55] To determine the best path in practice, the edge energy vector $\mathbf{E}^{TS}$ is transformed into a cost-vector $\bar{\mathbf{w}}$ using eq 5. $\bar{\mathbf{w}}$ has size $|\mathcal{E}|$ and assigns a cost $\bar{w}_{uv}$ to each edge (u, v) in $\mathcal{E}$. Subsequently, the Dijkstra algorithm[51] is used to identify a best path between the two end states through the weighted network defined by $(\mathcal{V}, \mathcal{E}, \bar{\mathbf{w}})$. This path minimizes the path cost $C(P)$ given in eq 6.

This best path furnishes a preliminary understanding of the transition,[13] and it may be used as a guess for a reaction coordinate for free energy calculations[56] or as a starting point for discrete path sampling.[49] However, it dominates the transition only if the barriers of alternative pathways are considerably higher. To obtain an idea of the number of different accessible pathways and their associated structures, it is useful to determine the set of k different pathways, $(P_1, P_2, ..., P_k)$ with costs $(C_1 \leq C_2 \leq ... \leq C_k)$, where P_1 is the path with the lowest cost, C_1, P_2 is the path with the second-lowest cost, C_2, etc. This so-called "k best path problem" is well-known in graph theory.[57] To precisely define it, one must define in which way two paths must differ in order to be treated as different. In a transition network, it is clearly not very meaningful to distinguish two pathways which differ only in two low-energy, non-rate-limiting barriers. Therefore, two paths are treated as different only if their rate-limiting steps (i.e. their highest-energy edges) do not coincide. The k best paths are determined in k steps: The second best path is found by using the Dijkstra algorithm after "blocking" the edge (u, v) associated with the highest energy barrier in the previously found best path (by setting its $E_{uv}^{TS} = \infty$). The third best path is found by blocking the highest edges of the best and second best paths, etc.

2.2. Energy Ridge. The collection of rate-liming transition states from all different (as defined above) paths from a defined reactant to a defined product belongs to a $(D-1)$-dimensional transition surface that divides the D-dimensional conformation space into a reactant and a product side. In terms of topography, this transition surface corresponds to an *energy ridge*, as illustrated in Figure 3. On a geographical landscape, it is analogous to a water-shed, i.e. the mountain ridge that separates water flows toward distinct oceans. The particular interest of the energy ridge is that it allows for a feeling to be quickly obtained for how degenerate the transition is, i.e., how many significantly different paths are likely to be accessible. For instance, if one transition state in the ridge has a significantly lower energy than the other transition states in the ridge, then the transition mechanism is dominated by a well-defined bottleneck. In contrast, if the ridge contains many different transition states with similar energies, the transition mechanism is not well defined.

In graph-theoretical terms, an energy ridge is a *cut*. The name "cut" stems from the fact that deletion of its edges

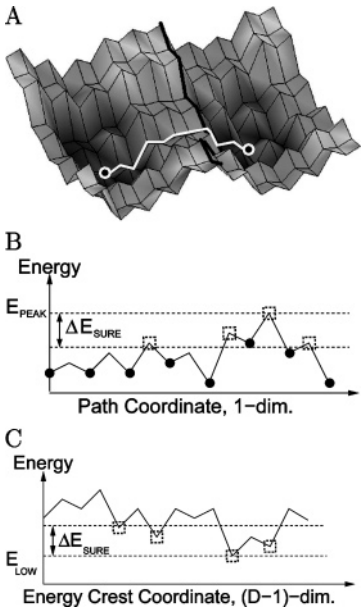

Figure 3. Schematic representation of the graph-theoretical concepts introduced in sections 2 and 3.3. (A) The best path (white line) connecting the transition end states (black bullets) and the energy ridge (black line) separating them. (B) Profile of vertex and edge energies along the best path through the network. The best path is requested to be correct in all edges with energies in the range $[E_{peak} - \Delta E_{sure}, E_{peak}]$ (indicated by squares). (C) Profile of the energy ridge cutting the TN into two conformational regions. The energy ridge is guaranteed to be correct in all edges with energies in the range $[E_{low}, E_{low} + \Delta E_{sure}]$.

dissociates the network into two disconnected subnetworks. Formally, the cut C is a set of M edges $C = \{(u_1, v_1),...,(u_M, v_M)\}$ with the property that each vertex u_i belongs to one set, U (e.g. "reactant side"), each vertex v_i belongs to another set, V (e.g. "product side"), and (U, V) partition the set of all vertices (i.e., $U \cup V = \mathcal{V}$ and $U \cap V = \varnothing$).

When the best and all next-best paths each have a dominant (rate-limiting) step, the energy ridge is identical to the cut whose total flux k_{UV} across it is minimal. k_{UV} is given by the sum of all localized fluxes $k_{u_iv_i}$ in the direction $u \rightarrow v$ across edges (u_i, v_i) in the cut

$$k_{UV} = \sum_{(u_i,v_i)\in C} k_{u_iv_i} \tag{7}$$

where $k_{u_iv_i}$ is the equilibrium flux from eq 4. By dismissing the proportionality constant, we obtain the normalized total flux, $k_{UV,0}$:

$$k_{UV,0} = \sum_{(u_i,v_i)\in C} \exp\left(-\frac{E^{TS}_{u_iv_i}}{k_B T}\right) \tag{8}$$

Note that the cut that minimizes $k_{UV,0}$ (the *rate-limiting cut*) and the cut associated with the topographic energy ridge are not always identical. For example, consider a case where the topographic ridge is very broad, its cut containing many edges of similar energy, whereas another cut contains only a single edge of slightly lower energy than those of the topographic ridge. Then the cut with the single edge has a lower $k_{UV,0}$ than the cut of the topographic ridge, because the many individual fluxes across the broad topographic ridge add up to a larger total flux. In the current context, however, this theoretical difference is not of importance.

The rate-limiting cut can be found by defining the vector of weights $\mathbf{w} = (w_{\mathcal{E}_1},..., w_{\mathcal{E}_{|\mathcal{E}|}})$, where for each edge (u, v) in the network $w_{uv} = \exp(-E^{TS}_{u_iv_i}/k_BT)$, and using the algorithm of Nagamochi and Ibaraki.[58] However, this algorithm is computationally expensive (scaling as $O(|\mathcal{V}|^3)$ or $O(|\mathcal{V}|^2 + |\mathcal{V}||\mathcal{E}| \log|\mathcal{V}|)$, depending on the implementation). Since the computation of the cut has to be repeated many times (see section 3.3.2), we used the topographical energy-ridge cut rather than the rate-limiting cut.

The topographical energy-ridge cut is determined by an algorithm that can be likened to flooding the energy landscape by stepwise filling up its basins. The ridge that last divides the reactant and product "lakes" before they become connected is the energy ridge. In the network, the edges which define the energy ridge are identified iteratively, starting from an edgeless network, $\mathcal{G}$, consisting only of the vertices, $\mathcal{V}$. In each iteration, a new edge $e \in \mathcal{E}$ is added to the network in order of increasing edge energy. At each iteration, the topology of $\mathcal{G}$ allows the identification of connected subgraphs (i.e. sets of vertices in which each vertex has at least one link to another vertex in the set). Each vertex is assigned an identifier that is unique for the connected subgraph it belongs to. The subgraph containing the reactant vertex is always assigned the identifier '0', while the subgraph containing the product vertex is always assigned '1'. Whenever an edge would be added that connects two vertices with identifiers '0' and '1', this edge is not added but marked as part of the energy ridge. The full ridge is determined when all edges have been iterated.

3. Methods

3.1. Efficient Sampling Procedure for Complex Conformational Changes in Proteins. In this section, a method is described for generating a representative sample of low-energy minima covering the conformational (sub)space relevant to a conformational transition. The method consists mainly of two stages: (1) generation of a sample of low-energy minima that are sparsely distributed over a large conformational subspace and (2) finding new low-energy minima between the minima found in (1) so as to densely map out the low-energy regions of conformational space.

In the first stage, the sampling can be limited to conformations that are likely to be relevant to the transition, thus avoiding sampling of the full conformational space (which would include, for example, mostly unfolded structures of the protein). For most transitions between native protein conformations, a good estimation can be made as to which structural regions are likely to require significant sampling. Small deformations in the remaining domains, which are structurally similar in the two end states, are then sampled by a simple interpolation between the end states. This partitioning and sampling procedure is described in sections 3.1.1−3.1.3 and is illustrated in Figure 4. Section 3.1.4 describes strategies for finding a uniformly dense set of low-

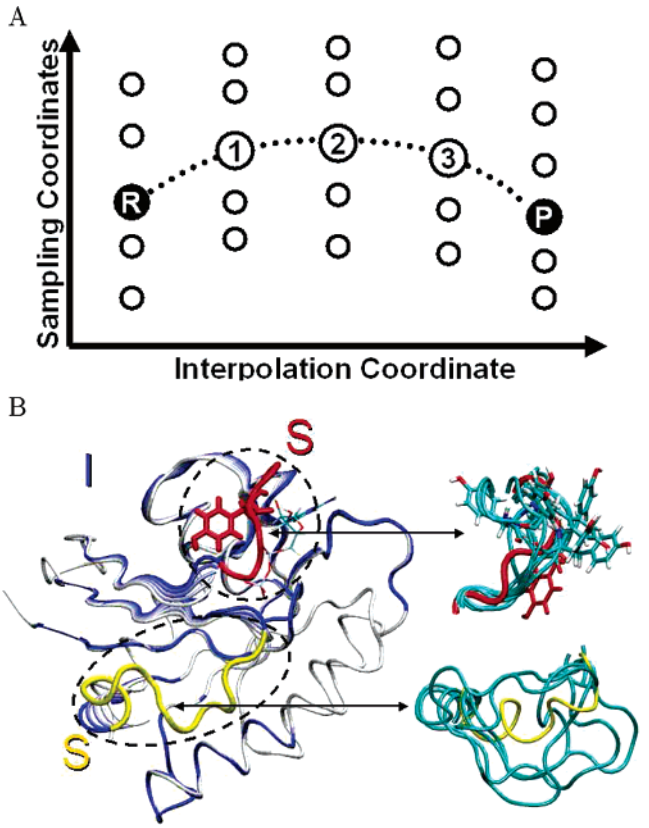

Figure 4. Illustration of the sampling procedure. (A) Some intermediate conformations are generated by interpolating the positions of a subset of protein atoms (the **I**-region) between their end state positions "R" and "P" (here 3 intermediate structures are generated, shown in circles). From each structure along the interpolation, a large set of conformations is generated by sampling the torsional angles of the **S**-region of the protein. The full set of conformations is defined by all combinations of the five (including both end states) conformations for the **I**-region with each sample of the **S**-region. (B) The interpolation (**I**) and sampling (**S**) regions in Ras p21. Left: The atoms of the **I**-region are interpolated between the transition end states (shown in white and dark blue), here producing three intermediates (shown in shades of blue). Right: The single-bond torsion angles of the **S**-region (switch I in red, switch II in yellow) are sampled uniformly (examples of several **S** conformations are overlaid). The **S** region encompasses switch I (residues 30−35) and switch II (residues 61−70).

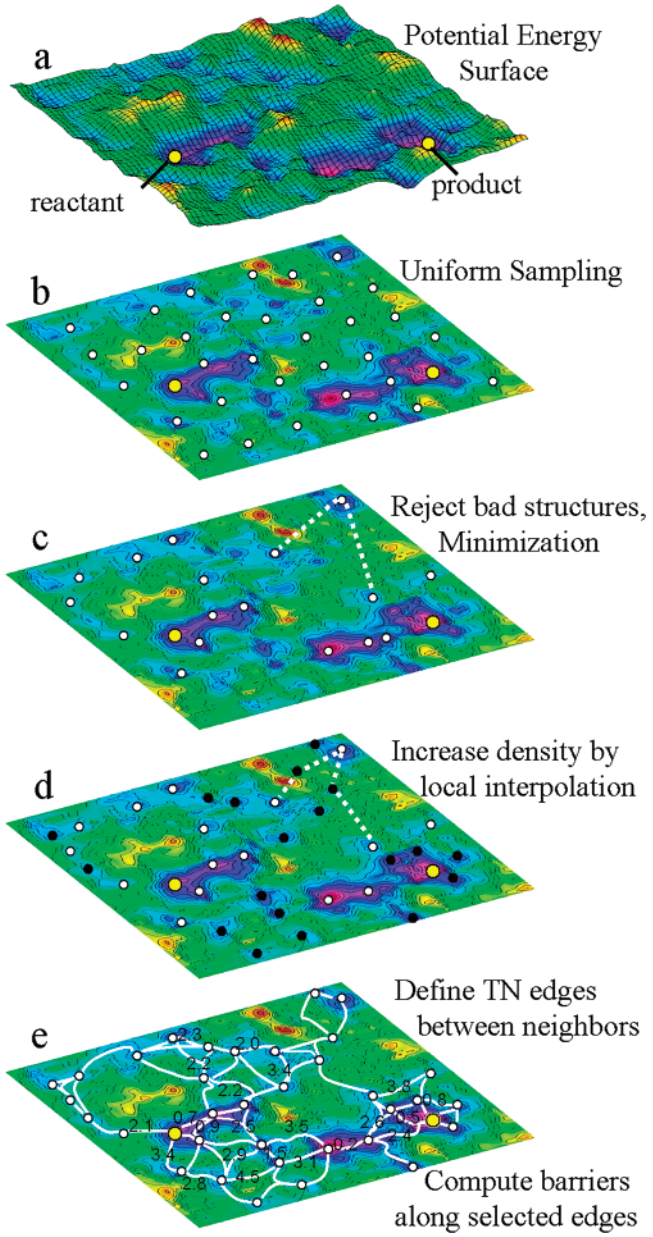

Figure 5. Overview of the steps in generating a transition network (TN). (a) The potential energy surface and the minimized reactant and product end states of the transition. (b) Conformers (white bullets) are uniformly spread over the part of conformational space that is relevant to the transition (see Figure 4 and section 3.1.3). (c) Structures without steric clashes are accepted (see Appendix A) and energy-minimized (see section 3.1.4). Interpolation between pairs of available conformers is used to explore nearby low-energy regions (dashed line). (d) The minimized interpolation intermediates (black bullets) increase the density of low-energy minima (see section 3.1.4). These minima form the TN vertices. (e) Pairs of neighboring minima are associated, forming the TN edges. The subtransitions of selected edges are computed by CPR, yielding the rate-limiting energy barriers for the TN edges (see section 3.2).

energy minima. Figure 5 gives a schematic overview of all the steps involved in the generation of the TN.

3.1.1. The Sampling (S) and Interpolation (I) Regions. Functionally relevant conformational changes in proteins are usually relatively local in the sense that most native contacts are preserved. While there might be complex rearrangements in certain regions, involving sometimes even refolding of parts of the backbone (such as in Ras p21), the remainder of the protein only deforms flexibly. This allows a sampling subspace with a considerably reduced dimensionality to be defined. Thus, the protein can be partitioned into interpolation (**I**) and sampling (**S**) regions. The changes in the **I** atoms are sampled by simply interpolating between the atomic positions in the two transition end states (see Figure 4a for an illustration). For each such interpolated structure, the rotable torsion angles of the **S**-region (including backbone ϕ/ψ torsions and single-bond side-chain torsions) are sampled

uniformly (Figure 4b). Finally, all degrees of freedom are locally relaxed during the energy minimizations that follow this combined **I**/**S** sampling procedure.

3.1.2. Interpolation of I-Atoms. To obtain a smooth variation of the positions of atoms in the **I** region near the boundary to the **S** region, the coordinates of the **I** region are generated by interpolating between the end states of the transition. For this, a combined interpolation procedure is used: First, so as to preserve the backbone fold, the backbone atoms are interpolated in Cartesian coordinates, and then the side-chain atoms are built onto the interpolated backbone, using internal coordinate values that are interpolated between the internal coordinates of the end states. This interpolation method has been shown to produce less distorted structures than Cartesian or internal coordinate interpolation alone.[13]

For practical convenience, the combined interpolation is done for all atoms of the protein (including the atoms of the **S** region). Because the **S** region has by definition very different conformations in the end states, the interpolated structures involve distorted internal coordinates in the **S** part of the backbone. To start the **S** sampling with reasonable values of the internal coordinates of the **S** region, each interpolated structure is energy minimized with positional harmonic constraints on the **I** atoms (force constant 1 kcal $mol^{-1}Å^{-1}$). In the example treated here, $n_{interpol} = 3$ interpolated structures of Ras p21 were generated in this way, yielding 5 structures along the interpolation including the end states.

3.1.3. Conformational Sampling of the S Region. For each of the structures along the interpolation between the two end states, many conformers of the **S** region are generated (Figure 5b). Sampling of the **S** region is performed uniformly in the space of flexible torsion angles, comprising the ϕ/ψ backbone and single-bond side-chain angles. The stiff internal degrees of freedom (i.e. bond lengths, valence angles and backbone ω angles) were not sampled here. The flowchart in Figure 6 summarizes the following algorithm for sampling backbone conformations.

If the sampling region is located at one of the termini of the polypeptide chain, there are no closure constraints on the backbone, allowing the ϕ/ψ angles to be sampled directly by setting them to random values. However, when the sampling region is within the polypeptide chain (as is the case for the switch I and switch II loops in Ras p21), this "free" sampling is not possible, as it would involve violation of the backbone closure (i.e., some backbone bond lengths and angles would not be preserved) or disruption of the native fold of the protein. Therefore, random backbone conformations are generated using a variant of the so-called window method.[59,60] This procedure allows backbone variation of a series of $r \geq 3$ consecutive residues (the "window") while preserving the position and orientation of the backbone at the boundaries of that window. Out of the windows' $2r$ torsion angles (ϕ and ψ), $2r - 6$ can be freely chosen and rotated randomly. The remaining six torsion angle values are determined by the window method (see ref 59 for a detailed description). In each sampling step, the location and length of the window in the **S** region and the rotated torsion angles are randomly chosen.

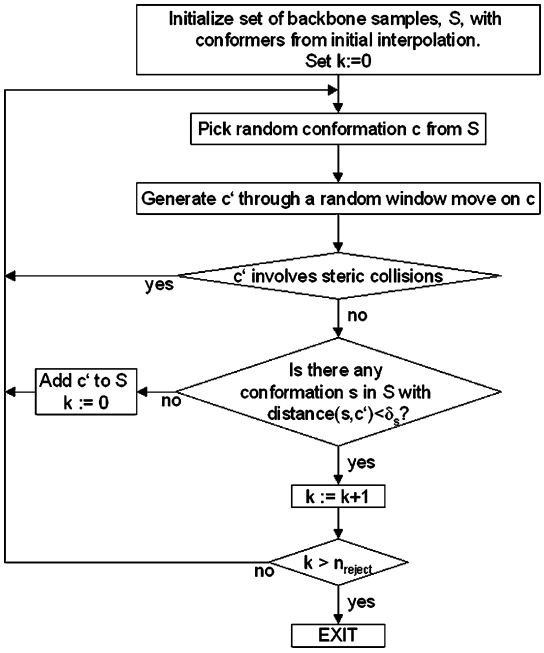

Figure 6. Flowchart for the backbone sampling procedure (section 3.1.3).

A backbone conformer is considered "valid" if it does not produce steric collisions. For this, it is checked whether the resulting backbone atoms and atoms whose positions are directly dependent on the backbone configuration (i.e., backbone O and H, C_β, and proline side chains) can be placed without collision among themselves and with the **I** region of the protein. Since a large number of **S** conformers have to be tested for collisions, an efficient strategy is used to perform these collision checks (see Appendix A).

To obtain a conformational sample that is approximately uniform, conformers that are valid (i.e., have no collisions) are "accepted" (i.e., added to a "conformational repository") only if they significantly differ from already-accepted conformers as measured by the ϕ/ψ dihedral RMS difference, which must exceed a chosen value, δ_s (see Table 2 for suggested parameter values). The choice of δ_s determines the density of sampling. It is set according to how finely the details of the transition should be probed. Backbone conformers are generated until the sampling density defined by δ_s has been reached. The criterion used here for this is that no "valid" structures are "accepted" anymore for a number of n_{reject} successive attempts. This yields a number of n_i^{back} backbone conformations for each interpolation step i ($i \in \{0,1,...,n_{interpol} + 1\}$, where $i = 0$ and $i = n_{interpol} + 1$ are associated with the end states and $1,...,n_{interpol}$ are the interpolated intemediates).

To obtain a complete conformation, the side chains of the **S** region are built onto a randomly picked backbone out of the n_i^{back} generated backbones, using randomly chosen single-bond torsion angles. The resulting conformer is accepted if it does not involve atom collisions (see Appendix A), giving for each interpolation step i a number n_i^{full} of sterically valid conformations of the **S** region. $n_i^{full} = k^{side} n_i^{back}$, where k^{side} is the desired average number of side chain conformers per backbone conformer. An efficient

Table 1: Frequently Used Symbols

symbol	meaning
$U(\mathbf{x})$	energy function
$\mathcal{V}=(1,...,\lvert\mathcal{V}\rvert)$	list of vertices in network; $\lvert\mathcal{V}\rvert$ is the number of vertices
$\mathrm{X}^S=(\mathbf{x}_1^S,...,\mathbf{x}_{\lvert\mathcal{V}\rvert}^S)$	conformers corresponding to vertices (here: minima on $U(\mathbf{x})$)
E_0	energy of the minimized reactant structure (see section 3.4)
$\mathbf{E}^S=(E_1^S,...,E_{\lvert\mathcal{V}\rvert}^S)$	vertex energies (here: $E_i^S=U(\mathbf{x}_i^S)-E_0$)
$\mathcal{E}=((u_1,v_1),...,(u_{\lvert\mathcal{E}\rvert},v_{\lvert\mathcal{E}\rvert}))$	list of edges in the network, each edge connecting two vertices of V; $\lvert\mathcal{E}\rvert$ is the number of edges
$\mathbf{X}^{TS}=(\mathbf{x}^{TS}_{u_1v_1},...,\mathbf{x}^{TS}_{u_{\lvert\mathcal{E}\rvert}v_{\lvert\mathcal{E}\rvert}})$	for each edge in $\mathcal{E}$, highest saddle point on minimum-energy path connecting the conformers $\mathbf{x}_{u_i}^S$, $\mathbf{x}_{v_i}^S$
$\mathbf{E}^{TS}=(\mathbf{E}^{TS}_{u_1v_1},...,\mathbf{E}^{TS}_{u_{\lvert\mathcal{E}\rvert}v_{\lvert\mathcal{E}\rvert}})$	edge (i.e. saddle-point) energies: $E^{TS}_{u_iv_i}=U(\mathbf{x}^{TS}_{u_iv_i})-E_0$
$\mathcal{G}=(\mathcal{V},\mathbf{X}^S,\mathbf{E}^S,\mathcal{E},\mathbf{X}^{TS},\mathbf{E}^{TS})$	transition network composed of vertices $\mathcal{V}$ connected by edges $\mathcal{E}$
$\mathbf{E}^{TS,min}=(\mathbf{E}^{TS,\mathrm{min}}_{u_1v_1},...,\mathbf{E}^{TS,\mathrm{min}}_{u_{\lvert\mathcal{E}\rvert}v_{\lvert\mathcal{E}\rvert}})$	lower bounds to the (yet unknown) edge energies
$\mathbf{E}^{TS,\mathrm{max}}=(\mathbf{E}^{TS,\mathrm{max}}_{u_1v_1},...,\mathbf{E}^{TS,\mathrm{max}}_{u_{\lvert\mathcal{E}\rvert}v_{\lvert\mathcal{E}\rvert}})$	upper bounds to the (yet unknown) edge energies
$w=(w_{u_1v_1},...,w_{u_{\lvert\mathcal{E}\rvert}v_{\lvert\mathcal{E}\rvert}})$	Boltzmann weights of the edge energies: $w_{u_iv_i}=\exp(-E^{TS}_{u_iv_i}/k_BT)$
$\bar{\mathbf{w}}=(\bar{w}_{u_1v_1},...,\bar{w}_{u_{\lvert\mathcal{E}\rvert}v_{\lvert\mathcal{E}\rvert}})$	inverse Boltzmann weights of the edge energies: $w_{u_iv_i}=exp(E^{TS}_{u_iv_i}/k_BT)$
$\bar{\mathbf{w}}^{\mathrm{min}}=(\bar{w}^{\mathrm{min}}_{u_1v_1},...,\bar{w}^{\mathrm{min}}_{u_{\lvert\mathcal{E}\rvert}v_{\lvert\mathcal{E}\rvert}})$	inverse Boltzmann weights of the lower edge energy bounds: $w^{\mathrm{min}}_{u_iv_i}=exp(E^{TS,\mathrm{min}}_{u_iv_i}/k_BT)$
$\bar{\mathbf{w}}^{\mathrm{max}}=(\bar{w}^{\mathrm{max}}_{u_1v_1},...,\bar{w}^{\mathrm{max}}_{u_{\lvert\mathcal{E}\rvert}v_{\lvert\mathcal{E}\rvert}})$	same as $\bar{\mathbf{w}}^{\mathrm{min}}$, for upper bounds

Table 2: Parameters for the Sampling Algorithm

parameter	meaning	value
n_{interpol}	number of steps along the interpolation between the transition end states, including the end states ($n_{\mathrm{interpol}}\geq 2$)	5
δ_s	shortest permitted RMS distance between two accepted backbone conformers (here: in ϕ/ψ torsion angle space)	50°
n_{reject}	sampling has converged when n_{reject} newly generated conformers are successively rejected because they are closer than δ_s to already-accepted conformers	1000
k^{side}	average number of side chain conformers per backbone conformer.	10
E_{tol}	largest permitted interaction energy between each atom-pair, for a structure to be considered valid	20 kcal/mol
n_{min}	number of structures drawn from the conf. repository for minimization	15000
E_{low}	largest permitted energy difference above minimized reactant structure to accept a conformer	40 kcal/mol
$\delta^{\mathrm{interpol}}_{\mathrm{min}}$	shortest and longest permitted RMS distance between a pair of minima to generate additional conformers by interpolation between them	$\delta^{\mathrm{interpol}}_{\mathrm{min}}=0.75$ Å
$\delta^{\mathrm{interpol}}_{\mathrm{max}}$		$\delta^{\mathrm{interpol}}_{\mathrm{max}}=2$ Å
$\delta^{\mathrm{connect}}_{\mathrm{min}}$	shortest permitted RMS distance between any pair of minima $\mathbf{x}_u^S$, $\mathbf{x}_v^S$, to avoid redundancy in the TN	0.75 Å
$\delta^{\mathrm{connect}}_{\mathrm{max}}$	longest permitted RMS distance between any pair of minima $\mathbf{x}_u^S$, $\mathbf{x}_v^S$, to form an edge (u, v) in the TN	1.5 Å
$n^{\mathrm{connect}}_{\mathrm{max}}$	maximum number of neighbors for each vertex	20

protocol for building side chains on large **S** regions is described in Appendix B.

For Ras p21, the switch I and II **S** regions were sampled independently, using $\delta_s=50°$ and $n_{\mathrm{reject}}=1000$. For each interpolation step, i ($i\in\{0,...,4\}$), this yielded $n_i^{\mathrm{back1}}\approx 30$ backbone conformers for switch I and $n_i^{\mathrm{back2}}\approx 10^4$ backbone conformers for switch II. An average of $k_{\mathrm{side}}=10$ side-chain conformations per backbone conformer were generated, yielding $n_i^{\mathrm{full1}}\approx 300$ and $n_i^{\mathrm{full2}}\approx 10^5$ collision-free conformations of switch I and II, respectively. Combining pairs of these switch I and II conformers yielded 3×10^7 fully built protein structures for each interpolation step i. Thus, the total number of collision-free and significantly different structures is $n^{\mathrm{full}}=1.5\times 10^8$, forming a large conformational repository from which structures can be drawn and further energy optimized. The conformations in this repository are distributed uniformly within the sterically accessible regions of the conformational subspace spanned by the torsional coordinates of **S** and the interpolation coordinate of **I**.

3.1.4. Constructing a Uniformly Dense Set of Low-Energy Minima. To obtain a representative collection of low-energy minima, a number of n_{min} conformers is drawn randomly from the conformational repository and energy-minimized on the potential $U(\mathbf{x})$ (see Figure 5c). Only minima which reach a low-energy region defined by $U(\mathbf{x})<E_{\mathrm{low}}$ are accepted, where E_{low} is a predefined constant. Minimization of many conformers is expensive, so it is desirable to reject structures early which are not likely to fall into low-energy minima. An efficient method to do this is proposed in Appendix C. For Ras p21, $n_{\mathrm{min}}=15\,000$ conformers were randomly retrieved from the conformational repository. Out of these, 189 reached the desired low-energy region below E_{low}, which was taken here as 40 kcal/mol above the energy of the minimized reactant structure (obtained by quenched molecular dynamics, see section 3.4). These were minimized to a gradient RMS of 10^{-3} kcal mol^{-1} Å^{-1}. They form a sparse set of low-energy conformations in the desired region of conformational space (see Table 4).

The density of conformers in the low-energy regions can, in principle, be increased by minimizing more structures from the conformational repository. Given the low yield of this approach (see above: $189/15\,000\approx 1.25\%$), this is com-

Table 3: Sampling of the **S** Regions in Ras P21

symbol	meaning (see section 3.1.3)	value
n_i^{back1}	number of backbone conformers (residues 30−35), for each interpolation step i ($i \in \{0,...,4\}$)	≈30
n_i^{back2}	same as n_i^{back1}, for residues 61−70	$\approx 10^4$
n_i^{full1}	number of fully built conformers with side chains (residues 30−35), for interpolation step i	≈300
n_i^{full2}	same as n_i^{full1}, for residues 61−70	$\approx 10^5$
n^{full}	total size of the conformational repository	$1.5 \cdot 10^8$

Table 4: Size and Density of the Network during Sampling

	minima[a]	accepted[b]	neighbors[c]
after first sampling[d]	15002	189	3
after increasing density[e]	35836	10831	267
TN vertices[f]	n/a	6242	117

[a] The total number of generated energy-minima. [b] The number of accepted minima with energy below E_{low}. [c] The average number of neighbors around accepted minima within a distance-range from $\delta_{\text{min}}^{\text{connect}}$ to $\delta_{\text{max}}^{\text{connect}}$. [d] After the sampling of the **I** and **S** regions (sections 3.1.2 and 3.1.3). [e] After increasing the density of low-energy minima (section 3.1.4). [f] After removing redundancy by not allowing neighbors closer than $\delta_{\text{min}}^{\text{connect}}$ (section 3.2).

putationally inefficient. Instead, additional conformers are built by interpolation between the already-found low-energy conformers. This can be done in various ways. The strategy used here was to select each pair of low-energy conformers separated by a distance in the range $\delta_{\text{min}}^{\text{interpol}} = 0.75$ Å and $\delta_{\text{max}}^{\text{interpol}} = 2$ Å (measured as Cartesian RMSD of the C_α atoms in the **S** region) and to generate an interpolation pathway between them using the method described in section 3.1.2. Two structures were generated, one-third and two-thirds of the way along each interpolation, respectively, and energy minimized as described in section 3.1.4 (see Figure 5d). This procedure was efficient in finding low-energy minima, increasing the number of conformers below E_{low} from 189 to 10 831 (see Table 4). This considerably increased the average number of neighbors for each minimum from 3 to 267 ("neighborhood" being defined by a cutoff distance $\delta_{\text{max}}^{\text{connect}}$, see section 3.2).

During minimization, it is possible that some conformers end up in similar minima. This produces conformational redundancy, which was subsequently removed. For this, minima were considered in the order of increasing energy, accepting only those minima whose nearest-neighbor distance to any already-accepted minimum was at least $\delta_{\text{min}}^{\text{connect}} = 0.75$ Å. This led to a final number of $|\mathcal{V}| = 6242$ diverse minima.

3.1.5. Verification of the Set of Minima. The available set of minima is approximately uniformly distributed in a conformational subspace which depends on the original definitions of **S** and **I**. There are two questions regarding the adequacy of this set of minima: (1) is the set dense enough and (2) are relevant parts of conformational space sampled.

The density of the set may be increased by reducing the parameter $\delta_{\text{min}}^{\text{connect}}$ and conducting further interpolations. Clearly, there is a tradeoff between the density of minima and the computational requirements. The important question

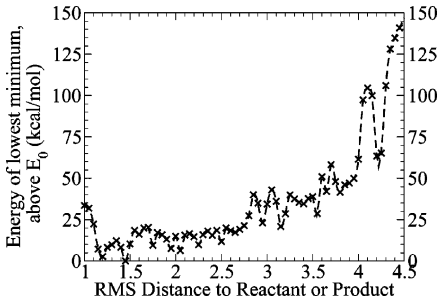

Figure 7. Dependence of energies of minima on the distance from the crystallographic end states. For each minimum, the distance to the reactant or product structure (whichever is closer) is calculated. The minima with distances between 1 and 4.5 Å were grouped according to their distances, each group being 0.1 Å wide. The lowest energy of the minima within each group is plotted versus the distance of that group. The plot shows that the energies increase considerably with increasing distance from the crystallographic end states, indicating that the relevant portions of conformational space have been sampled.

is how sensitive the analyses are to the density of minima. In the present study, density variation, within reason, has little effect as the CPR path calculations used to compute the edge energies ensure that no important intervening barriers are missed (see section 3.2). Moreover, the purpose of the present calculations is to generate a coarse-grained model of the Ras p21 energy function which is analyzed in the Results section based on qualitative properties of large sets of pathways. Local features of the network do not play a role in this analysis, and therefore the density of minima was not further increased.

A more critical question is whether there are important parts of the conformational space that are not sampled at all. A logical check of this is to examine whether any low-energy minima exist in regions of conformational space continguous with regions already explored. If so, energetically accessible pathways might exist that lead out of the available set of minima into other regions of the conformational space which were not included in the initial sampling. This can be checked by calculating the lowest-energy minima within shells that are increasingly distant from the reactant and product vertices and ensuring that the found set of minima defines an energetic basin that is unlikely to be left.

To examine this, we have analyzed all 35 836 minima that were generated (see Table 4) and computed their distances to the reactant or product structure (whichever of the two was closer). For each distance-window between 1 and 4.5 Å, the lowest energy of all minima within that window was recorded. The result, which is shown in Figure 7, shows that there is a strong increase in energy with increasing distance, reaching about 150 kcal/mol above E_0 at 4.5 Å. This result shows that the existence of low-energy exit pathways from the initial set of minima is unlikely, and therefore a sufficient volume of conformational space has been sampled.

3.2. Construction of the Ras p21 Transition Network. The final number of $|\mathcal{V}| = 6242$ diverse minima served as the vertices of the transition network (see Table 4).

Due to the large number of possible conformations of the system studied in this article, it is possible to map only a subset of the dynamically accessible minima, so that the computation of meaningful free energies is not feasible. Here, the contributions from local vibrations around the minima are neglected, and the potential energy $U(\mathbf{x}_v^S)$ is used directly for E_v^S. Nevertheless, the important free energy contributions from bulk solvent are accounted for in the calculation of $U(\mathbf{x})$ through a continuum solvent method (see section 3.4).

The absolute height of the vertex can be shifted by subtracting an arbitrary constant value E_0 without affecting the results. Thus, we define E_v^S as $E_v^S := U(\mathbf{x}_v^S) - E_0$. Here, E_0 is chosen as the minimized reactant energy.

The "reactant" and "product" vertices were redefined by selecting the lowest energy minima within the vicinity of the crystallographic reactant and product structures after quenched MD (see section 3.4). The "vicinity" was defined here to be within both a ϕ/ψ-RMSD of 50° and a Cartesian RMSD of 1.5 Å for the C_α-atoms of the switch regions. The resulting Cartesian RMSD over all C_α atoms between the crystal structures and the so-chosen reactant and product conformers was 1.4 and 1.5 Å, respectively.

Given the complete set of vertices, $\mathcal{V}$, the edges of the transition network are generated by defining connections according to distance-based criteria. Each vertex is connected to up to $n_{\max}^{\text{connect}}$ of its nearest neighbors that are within a distance $\delta_{\max}^{\text{connect}}$ (see Figure 5e). For Ras p21, $\delta_{\max}^{\text{connect}} = 1.5$ Å (measured as RMS distance between the C_α-atoms of the **S** region), and $n_{\max}^{\text{connect}} = 20$ were used. The resulting transition network had $|\mathcal{E}| = 47\,404$ edges and was fully connected (i.e. any given pair of vertices is connected by some pathway).

To determine the energy barrier associated with a given edge, (u, v), a Minimum Energy Path (MEP) between the two minima $\mathbf{x}_u^S$ and $\mathbf{x}_v^S$ corresponding to that edge was computed using the Conjugate Peak Refinement (CPR) method.[18] An initial guess for the path was generated by interpolation between the edge end structures, using the procedure described in section 3.1.2. Starting from this guess, CPR identifies all the first-order saddle points that are local energy maxima along the Minimum Energy Path. Here, the CPR calculation was stopped as soon as the highest (i.e. rate-limiting) of these saddle points along the edge was determined. Its structure is assigned to $\mathbf{x}_{uv}^{TS}$, the corresponding edge energy is taken as $E_{uv}^{TS} = U(\mathbf{x}_{uv}^{TS}) - E_0$ (see section 2).

3.3. Efficient Determination of Best Paths and Energy Ridges. Even though the subtransition pathways are short compared with a whole pathway between the transition end states, finding the highest saddle-point along an edge can still be very CPU intensive (here, the average time on a single 3 MHz CPU was about 2 h per subtransition). Therefore, it is computationally infeasible to do this for all edges in the network. The problem thus arises that global properties of the network (such as the best path or the dividing energy ridge) must be determined using incomplete information on the barriers. We solve this problem by devising a strategy such that only a small number of edge energies need to be computed to determine the best path and the energy ridge. The strategy relies on the introduction of lower and upper bounds, $E_{uv}^{TS,\min}$, $E_{uv}^{TS,\max}$ on each edge energy, which bracket the (yet unknown) true edge energy, E_{uv}^{TS}.

The "safe" lower bound for the edge energy is $E_{uv}^{TS,\min} = \max\{E_u^S, E_v^S\}$, because the energy of each barrier is at least as high as the highest of the two minima it connects. The safe upper bound is $E_{uv}^{TS,\max} = \infty$, but for numerical reasons it is taken as $E_{uv}^{TS,\max} = E_{uv}^{TS,\min} + M$, where M is a large but finite number (here, M=100 kcal/mol). A tighter upper bound could also be obtained by performing a very short (i.e. unconverged) CPR path refinement on the edge (u, v) and using the highest energy along the resulting path as $E_{uv}^{TS,\max}$. An alternative method to both the lower and upper bounds, based on statistical estimates, is used here (see Appendix D).

To implement lower and upper bounds, two graphs are defined, which have the same topology as the actual transition network: one using the lower bounds for the edge energies, $\mathcal{G}^{\min} = (\mathcal{V}, \mathbf{X}^S, \mathbf{E}^S, \mathcal{E}, \mathbf{E}^{TS,\min})$, and the other using the with upper bounds for the edge energies, $\mathcal{G}^{\max} = (\mathcal{V}, \mathbf{X}^S, \mathbf{E}^S, \mathcal{E}, \mathbf{E}^{TS,\max})$. The best paths (and energy ridges) through $\mathcal{G}^{\min}$ and $\mathcal{G}^{\max}$ can be computed, using the corresponding inverse Boltzmann weight vectors, $\bar{\mathbf{w}}^{\min}, \bar{\mathbf{w}}^{\max}$ and Boltzmann weight vectors $\mathbf{w}^{\min}, \mathbf{w}^{\max}$. This is addressed in the next sections.

3.3.1. Best Paths. The flowchart in Figure 8 summarizes the procedure for finding the best path. The best path through the transition network is found iteratively as follows: In each iteration, the best path through the graph $\mathcal{G}^{\min}$, $P_{\text{best}}^{\min} = (v_R, \ldots, v_P)$, is determined as described in section 2.1. The edge with the highest unknown energy along $P_{\text{best}}^{\min}$, (u, v), is identified, and its true energy E_{uv}^{TS} is computed by CPR. The network $\mathcal{G}^{\min}$ is updated by setting $E_{uv}^{TS,\min}$ to the true edge energy, E_{uv}^{TS} (the weights $\bar{\mathbf{w}}^{\min}$ are also updated accordingly). This procedure is repeated until all edge energies along the resulting best path have been computed, yielding the true best path.

A preliminary estimate of the energy barrier of the best path can be easily obtained. For this, in each iteration of the above algorithm, the best path is calculated on both graphs $\mathcal{G}^{\min}$ and $\mathcal{G}^{\max}$, yielding best paths $P_{\text{best}}^{\min}$ and $P_{\text{best}}^{\max}$, respectively. The rate-limiting barrier of the true best path is bounded by those of $P_{\text{best}}^{\min}$ and $P_{\text{best}}^{\max}$. During successive iterations, these bounds converge to the true value (see Figure 9).

Often, one is not interested in the details of how the best path travels in the low-energy regions, since it is the highest-energy edges along the whole path that are rate determining. Computation time can thus be saved if only the high-energy edges of the best path, i.e., those with barrier energies within a range ΔE_{sure} of the highest-energy barrier along the path, E_{peak}, are required to be correct (see Figure 3b). To achieve this, the computation proceeds as above until the energy of the rate-limiting barrier, E_{peak}, is identified. This is the case when $P^{\min}$ and $P^{\max}$ have the same value for E_{peak}. After that, whenever a barrier is computed whose energy E_{uv}^{TS} is below the threshold $E_{\text{peak}} - \Delta E_{\text{sure}}$, the transition networks are updated by setting $E_{uv}^{TS,\min} = E_{uv}^{TS,\max} := \max\{E_u, E_v\}$, i.e., as

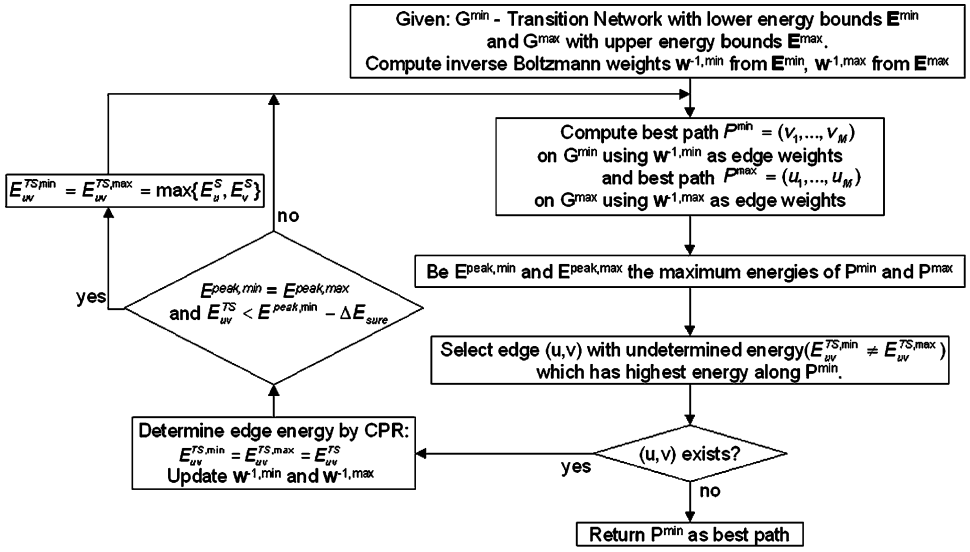

Figure 8. Flowchart for the best path finding procedure (section 3.3.1).

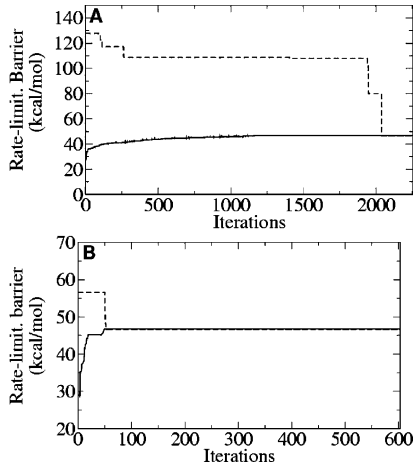

Figure 9. Convergence behavior of the best-path. The energy of the rate-limiting transition state along the best path $P_{\text{best}}^{\min}$ (solid line) or $P_{\text{best}}^{\max}$ (dashed line) is plotted as a function of iterations of the algorithm (see section 3.3.1). (A) Using "safe" bounds, i.e. setting the lower and upper bounds to 0 and ∞. (B) Using guessed bounds, i.e. setting the barrier bounds based on statistics (see Figure 13).

if the transition were barrierless. This saves computational effort as it makes sure that (u, v) is a part of the best path identified in the next Dijkstra computation, thereby avoiding to spend time in identifying paths avoiding (u, v) that might have a lower energy barrier than E_{uv}^{TS}.

The computing time can be drastically reduced, at the expense of possibly failing to identify the true best path, if the "safe" lower edge energy bound $E_{uv}^{TS,\min} = \max\{E_u, E_v\}$ is replaced with a statistical estimate (described in more detail in Appendix D). In this case the lower energy bound is not necessarily correct as it might overestimate the real barrier. That is, edges which are not included in the resulting best path and have been rejected based on their lower estimate $E_{uv}^{TS,\min}$ might in fact have a true edge energy $E_{uv}^{TS} < E_{uv}^{TS,\min}$. The maximum overestimation possible, $\text{err}_{\max}$, is given by the maximal difference found between any estimated lower bound and the corresponding safe lower bound:

$$\text{err}_{\max} = \text{MAX}[E_{uv}^{TS,\min} - \max\{E_u^S, E_v^S\}]_{\text{all pairs}(u,v)} \qquad (9)$$

Thus, $\text{err}_{\max}$, also gives the maximum possible error on the rate-limiting barrier of the path. If the graph $\mathcal{G}^{\max}$ is used to obtain a preliminary estimate of the energy barrier of the best path, this estimate is usually improved by also replacing the safe upper energy bound $E_{uv}^{TS,\max} = \infty$ by an statistical estimate. As the identification of the best path does not depend on $\mathcal{G}^{\max}$, an incorrect upper bound $E_{uv}^{TS,\max}$ cannot lead to a wrong result but may give a wrong upper bound for the preliminary estimate of the best path's energy barrier.

3.3.2. Energy Ridges. As defined in section 2.2, the energy ridge is the rate-limiting cut that divides the TN into a reactant and product side. The algorithm that computes the energy ridge while determining only a limited number of edge barriers uses a strategy similar to the one used to find the best-path. The energy-ridge cut is determined (as described in section 2.2) on the graph $\mathcal{G}^{\max}$ (i.e. using upper bounds for the yet unknown edge energies). The lowest-energy unknown edge, (u, v), in the resulting cut, $C^{\max}$, is computed by CPR, and $\mathcal{G}^{\max}$ is updated with the value of E_{uv}^{TS}. This process is repeated in successive iterations. When an energy-ridge cut $C^{\max}$ is identified whose edge barriers are all determined, it is identical to the true energy ridge.

In practice, it is sufficient to compute only the low-energy barriers of the energy ridge, since the higher-energy barriers are not populated. Thus, one is only interested in finding the energy barriers of the energy ridge that are up to an energy difference ΔE_{sure} above the energy of the lowest barrier in the ridge, E_{low} (see Figure 3c). To find these barriers, the algorithm given above proceeds until the value of E_{low} is identified. From this moment on, whenever a barrier is computed which has $E_{uv}^{TS} > E_{\text{low}} + \Delta E_{\text{sure}}$, $\mathcal{G}^{\max}$ is updated by setting $E_{uv}^{TS} = \infty$. This fools the algorithm so that it leaves these high-energy barriers in the ridge and thereby saves the computational cost of identifying the high-energy regions of the full ridge.

Finding the energy ridge depends on the upper bounds on the edge energies; therefore, the performance of the algorithm is considerably increased if estimates (see Appendix D) are used instead of infinite upper bounds. However, in contrast to the error involved in the determination of best paths (eq 9), no upper bound for the error from using such estimates can be derived here.

3.4. Test Case: Protein Model and Energy Function. The method is tested here on Ras p21. The conformational change in Ras occurs after the γ-phosphate of bound GTP is cleaved off. GTP hydrolysis can be catalyzed by the binding of a GTPase-activating protein (GAP)[61] or it can take place as a result of the weak intrinsic GTPase activity of Ras in absence of GAP. The conformational transition studied here is in the absence of GAP, as would occur after intrinsic hydrolysis of GTP. Crystallographic structures exist for both the GTP-bound (Protein Data Bank structure 5p21[62]) and GDP-bound states (1q21[63]) of Ras p21. The γ-phosphate was deleted from 5p21, to yield the reactant state. The 1q21 structure served as product state. The HBUILD facility in CHARMM[64] was used to place the missing hydrogens.

All calculations were performed using the extended-carbon potential function (PARAM19).[65] Contributions from bulk solvent to the free energy of the conformational substates were included with the Generalized Born model of continuum solvation, using version 2 of the Analytical Continuum Electrostatics (ACE) method.[66] Nonbonded interactions were smoothly brought to zero by multiplying them with a switching function between 8 and 12 Å.

The structure of a protein may be affected by the crystal environment. Therefore, both the reactant and the product structures were first relaxed using molecular dynamics simulations with ACE. For this 20 ps of heating were followed by 100 ps of equilibration and a 10 ns production run. One structure every 100 ps (making up 100 structures in total) was selected and energy minimized with ACE to a gradient RMS of 10^{-3} kcal mol^{-1} Å^{-1}. The structures with the lowest energies were selected as reactant and product structures. The potential energy of these structures was lower than that obtained by a direct minimization of 5p21 and 1q21 by 30−45 kcal mol^{-1}. Structurally, the differences compared to 5p21 and 1q21 were rather small, consisting mainly of exposed side-chain rearrangements, while the backbone fold of the switch regions was preserved. The RMS coordinate deviations from the directly minimized crystallographic end states were <1.8 Å for the nonfixed atoms (<2.4 Å for the switch regions).

To remove insignificant degrees of freedom, residues which were not involved in the conformational switch and whose atoms had similar positions in both end states were fixed (residues 1−4, 42−53, 77−95, 110−115, 124−143, 155−167), leaving 1001 atoms free to move. To obtain the same positions for the fixed atoms in the two end states, the product structure was oriented onto the reactant structure so as to minimize the RMS deviation between the fixed atom coordinate sets. Then, the reactant and product values of these coordinates were averaged. The averaged coordinates of the fixed atoms were used for all calculations. Furthermore, insignificant differences in the side chains of nonswitch regions were removed from the end states as described in ref 13. Finally, both end states were minimized to a gradient RMS of 10^{-3} kcal mol^{-1} Å^{-1}.

Table 5: Number of Edges Computed with CPR To Determine the Best Path[a]

ΔE_{sure}[b]	safe bounds[c]	guessed bounds[d]
∞[e]	2252	603
30	2246	589
25	2224	565
20	2208	505
15	2115	321
10	2069	212
5	2059	114
0	2059	106
path length[f]	23	24
energy barrier[g]	45.7	45.7

[a] Assuming no energy barrier has been previously computed. [b] Energy range below the highest barrier for which the barriers of the best path are to be determined (see Figure 3B). [c] Using $E_{uv}^{\min} = \max\{E_u, E_v\}$, $E_{uv}^{\max} = \infty$ as bounds on the unknown energy barriers. [d] Using statistical estimates (Appendix D) to guess the $E_{uv}^{\min}$ and $E_{uv}^{\max}$ bounds. [e] ∞ means that the whole best path with all its edges is to be determined. [f] Number of edges along the fully determined best path. [g] Rate-limiting energy barrier along the best path, in kcal/mol relative to the reactant.

4. Results and Discussion

4.1. Performance of Best Path Calculations. Best paths between the reactant and product structures of the Ras p21 conformational switch were computed using the iterative algorithm described in section 3.3.1. The performance was evaluated, first using safe values for the upper and lower bounds on the edge barriers (i.e. $E_{uv}^{TS,\min} = \max\{E_u, E_v\}$ and $E_{uv}^{TS,\max} = \infty$). Alternatively, statistical estimates for the bounds (described in Appendix D) were used. The partial computation of best paths, using different values for the energy interval ΔE_{sure} (see section 3.3.1 and Figure 3B) was also examined.

Table 5 shows how many edges need to be computed with CPR in order for the best path to be determined under these different conditions (starting the count from scratch for each setting). To determine the full best path ($\Delta E_{\text{sure}} = \infty$) using safe bounds on the energy barriers, 2252 edges had to be computed (only 5% of the total number of 47 404 edges). It was possible to reduce this number by a factor of 4 (to 603) when statistical estimates of the bounds were used. This faster convergence behavior is also apparent when comparing Figure 9A,B and demonstrates that the computation time can be greatly reduced by introducing a relatively small uncertainty. In the worst case, the error on the rate-limiting barrier resulting from the present estimates of $E_{uv}^{TS,\min}$ could have been as much as 5.25 kcal/mol (from eq 9). But in the present case, statistical estimation actually resulted in a best path with the same rate-limiting energy barrier as found when safe bounds were used. Moreover, except for one additional, insignificant low-energy step, the estimated best path is equal to the true best path.

The computational savings are even larger when only the highest-energy barriers of the best path are determined. The number of edges that need to be computed with CPR when

Table 6: Number of Edges Computed with CPR To Determine Ridge 2[a]

ΔE_{sure}[b]	safe bounds[c]	guessed bounds[c]
∞[d]	1092	667
20	1092	622
15	1092	509
10	897	383
5	862	293
0	805	214
ridge size[e]	174	162
energy barrier[f]	45.7	45.7

[a] The energy ridge of the switch II rearrangement, assuming no energy barrier has been previously computed. [b] Energy range above the lowest barrier for which ridge barriers are to be determined (see Figure 3C). [c] Same meaning as in Table 5. [d] ∞ means that all barriers of the ridge are determined. [e] Number of edges in the fully determined energy ridge. [f] Lowest edge barrier of the energy ridge, in kcal/mol relative to the reactant.

only the highest barrier along the best path is requested to be certain ($\Delta E_{sure} = 0$) in conjunction with statistical estimates on E_{uv}^{min} is 106, reducing the number of edge computations by a factor of 6 (from 603). This shows that statistical estimates help to quickly isolate the rate-limiting step of the reaction.

4.2. Performance of Energy Ridge Computations. The highest energy ridge in the TN (here: termed as ridge 2, as it is associated with rearrangements in the switch II) was computed with and without the use of statistical estimates for the barrier bounds to test the performance of the algorithm given in section 3.3.2. The results are shown in Table 6, where the counting is started from scratch for each setting, assuming that no energy barrier has been computed yet. 1092 energy barriers were computed to determine the full energy ridge ($\Delta E_{sure} = \infty$) with safe bounds, which amounts to $\approx$2% of the total number of 47 404 TN edges. By using statistical estimates, this number was reduced to 667. When only the lowest energy barrier of the ridge ($\Delta E_{sure} = 0$) was computed, the computational savings were comparatively slight (805 energy barriers were computed for this), but using both $\Delta E_{sure} = 0$ and statistical estimates reduced the number of computed barriers to 214. The "safe" and the "estimated" ridge 2 agree in their lower-energy edges (up to 5 kcal/mol above the lowest edge). For edges of higher energy, only about 25% of the edges in the "estimated" ridge belong to the safe energy ridge.

4.3. Structural Mechanism of Ras p21. Structural analysis shows that in the best path about half of the switch II helix first unfolds before the rearrangement of the switch I, in which Tyr32 passes underneath the backbone. Subsequently, the rearrangement of switch II completes. This latter step is rate limiting, having the highest potential energy barrier along the best path (E_{peak}=45 kcal/mol relative to the reactant). From the time scale of the Ras p21 conformational switch[67] (on the order of 10^4 s in absence of GAP), it follows that the highest free energy barrier along the path cannot exceed 23 kcal/mol.[13] This indicates that a significant contribution from entropy, due possibly to an increase in backbone flexibility, reduces the high enthalpic barrier found here.

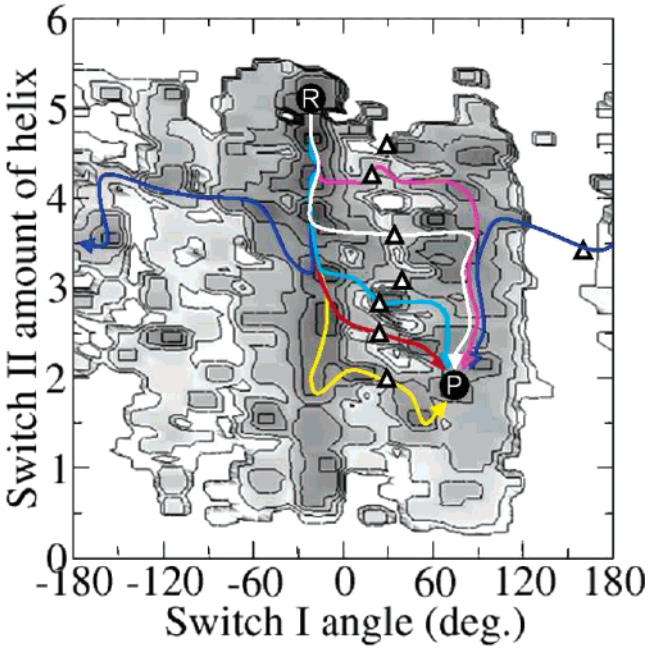

Figure 10. Two-dimensional representation of the potential energy surface of Ras p21. The horizontal axis measures the orientation of Tyr32 on the switch I loop (α = dihedral angle P_β,C_{32},N_{32},OH_{32}, in degrees). The vertical axis measures the helicity of switch II (number of α-helical H-bonds). The contour levels show the energy of the TN vertices (dark gray = 0−10 kcal/mol, light gray > 60 kcal/mol). Reactant and product structures are labeled as 'R' and 'P'. The best transition pathway is shown in white, and the next-best transition pathways (with a rate-limiting step up to 10 kcal/mol higher than the white) in yellow, red, magenta, and cyan. Triangles mark the rate-limiting transition state of the switch I rearrangement (corresponding to the lowest-energy points shown in Figure 11b,c) and belong to ridge 1 (see section 4.3). Ridge 1 can be split into two energy ridges: one along $\alpha \approx 30°$ (where Tyr32 passes underneath the backbone, Figure 11b) and another along $\alpha \approx 150°$ (where Tyr32 passes through the solvent, Figure 11c). The best path with Tyr32 moving through the solvent is shown in dark blue.

The next-best pathways (i.e. pathways having a different rate-limiting transition state, see section 2.1) with rate-limiting barriers within 10 kcal/mol above that of the best path were also computed. There are 12 such pathways in the present TN. The order of events in these pathways is similar to the events described above for the best path, i.e., switch II unfolds (to a varying degree) in the first part of the transition, and the subsequent switch I transition occurs with Tyr32 passing underneath the backbone (see Figure 10). The differences between these pathways are mainly in the precise order of events in the switch II rearrangement.

In the lowest best path, the passage of Tyr32 underneath the backbone is associated to an important barrier of about 25 kcal/mol. This raises the question whether the Tyr32 must necessarily pass underneath the backbone. An obvious alternative would be for Tyr32 to pass the other way (i.e. through the solvent). To better analyze the motion of Tyr32, the energy ridge corresponding to its reorientation (abbreviated as ridge 1, since it is the rate-limiting step of switch I

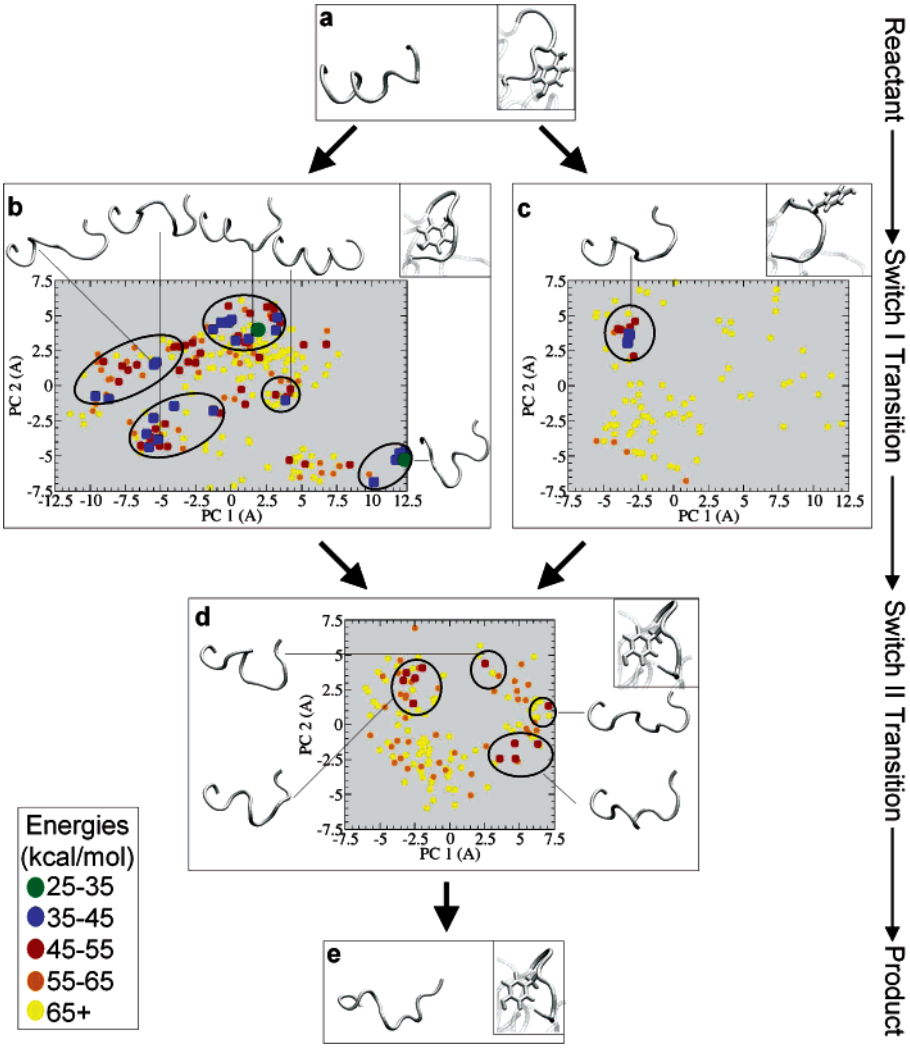

Figure 11. Two-dimensional projection of the energy ridges of the Ras p21 transition. Three major ridges were identified: two for the switch I rearrangement (both belonging to ridge 1) and one for the switch II rearrangement (ridge 2). Transition states from each ridge were projected on their two first principal components (computed from the C_α-coordinates). Each panel (b,c,d) shows one ridge and the corresponding conformation of the switch I loop (box in top right corner of each panel). The projected points cluster (ellipsoids) according to their different switch II conformations (typical backbone conformation shown for each cluster). The energy of each transition state is coded by color. (a) Reactant state: switch I has Tyr32 pointing to the 'right', switch II is a helix. From here, the conformational change proceeds through panels b or c. (b) Energy ridge of the switch I-transition (ridge 1), with Tyr32 passing underneath the backbone. There is a large variety of alternative switch II-conformations at this step of the transition. (c) Ridge 1 with Tyr32 moving through the solvent. (d) Energy ridge of the switch II-transition (ridge 2), which is globally rate-limiting. The transition of switch I is already completed and Tyr32 is pointing to the 'left'. Various isoenergetic ways for the switch II rearrangement coexist. (e) Product state: switch I is pointing to the 'left' and switch II helix has fully unfolded.

rearrangement) was determined. Tyr32 passes from an orientation where its side chain points toward GDP and $-30° < \alpha < -10°$, to an orientation where $60° < \alpha < 110°$. α is an artificial dihedral angle, defined over atoms P_β,C_{32},N_{32},-OH_{32} (P_β is the β-phosphorus of GDP). Ridge 1 was computed with $\Delta E_{sure} = 30$ kcal/mol and using safe values for the barrier bounds (see section 3.3). The resulting energy ridge consists of 92 transition states. In 11 of them, Tyr32 goes through the solvent, and the associated barrier is at least 40 kcal/mol, clearly indicating that passage underneath the backbone is the preferred mechanism. In Figure 10, ridge 1 appears split in two regions.

To visualize the two energy ridges, Figure 11 shows a two-dimensional projection of the transition states contained in ridge 1 and ridge 2. Ridge 1 was split into two sets: one set containing the transition states that involve the passage of Tyr32 underneath the backbone and the other set containing the transition states having Tyr32 passing through the solvent. In the case where passage of Tyr32 is underneath the backbone, there are 7 different transition states in ridge 1 up to 10 kcal/mol above the lowest transition state in ridge 1. These differ considerably in the amount of unfolding of the switch II helix: some still form a perfect helix, while in others the helix is fully unfolded (see Figure 11b). In the unfavorable case that Tyr32 passes through the solvent the conformation of the partially unfolded switch II helix is well defined, as can be seen from its similar structure in all next-higher transition states (Figure 11c).

After the switch I rearrangement has completed, the transition pathways must cross ridge 2, which contains the globally rate-limiting transition states. Ridge 2 contains 14 transition states within 10 kcal/mol above the lowest transition state in ridge 2 (which here is identical to the highest transition state in the lowest best path). These alternative transition states are highly scattered in Figure 11d, showing that the structure of switch II varies considerably. Thus, there are many different ways in which switch II can rearrange toward the product structure, and the coupling between switch I and II is weak enough to allow for many different orders of the conformational events in both switch regions. This means that the Ras p21 conformational switch is highly degenerate, thus confirming a significant entropic contribution to the free energy profile of the conformational switch.[13]

5. Conclusions

We have introduced here an efficient method for mapping the low-energy minima involved in a complex conformational transition in a protein. The method was shown to be effective in identifying minima belonging to very different conformational pathways. Furthermore, the resulting set of minima is dense in the low-energy regions.

A transition network is constructed to connect the available set of low-energy minima. The graph-theoretical methods have allowed to determine global properties of the network while only requiring computation of a small subset of the subtransition barriers in the network. When applied to the conformational switch of Ras p21, the globally best pathway connecting the transition end states and the energy ridge separating them could be determined while computing less than 5% of the total number of subtransitions in the network.

The energetically best pathway and the two main energy ridges of the Ras p21 switch give insight in the mechanism of the transition and provide answers to the three questions asked in the Introduction: (1) The rearrangement of switch I always occurs such that Tyr32 is threaded underneath the protein backbone. (2) This rearrangement of switch I must be finished before the rate-limiting rearrangement of switch II can start. (3) The order of conformational events in either switch I or II and the details of rearrangement in switch II vary substantially. This confirms that complex conformational transitions in proteins such as Ras may occur via multiple pathways.

The methodological advances presented here allow comprehensive analysis of the mechanism of complex transitions in proteins. To allow for comparison with certain experiments, it will be desirable to obtain free energy TN that allow calculation of thermodynamic and kinetic properties. This might be achieved by estimating vibrational free energies for the TN states[42,43] and merging vertices which are separated by low-energy barriers so as to account for intrastate configurational entropy.[32,42,43]

Acknowledgment. We thank Prof. G. Reinelt and M. Oswald for valuable discussions concerning the graph-theoretical aspects of this work. We kindly acknowledge Deutsche Forschungsgemeinschaft (DFG) for financial support.

Appendices: A. Checking Conformers for Steric Collisions

During the sampling procedures described in section 3.1.3, new conformers are validated by checking that they do not produce very high potential energies, mostly due to atom collisions (see Figure 5c). For each pair of atoms i, j, the criterion used is that the combined Lennard-Jones and Coulomb interaction energy should not exceed a tolerance value E_{tol} ($E_{\mathrm{tol}} = 20$ kcal mol^{-1} in this study). This check needs to be repeated so often that it is a computational bottleneck for the sampling method. Therefore, to avoid computing all pairwise interaction energies for each conformer, we first precompute the minimum distance, $d_{i,j}^{\mathrm{min}}$ allowed for each atom pair. This is obtained as the solution of

$$\epsilon_w\left[\left(\frac{\sigma_{ij}}{d_{i,j}^{\mathrm{min}}}\right)^{12}-\left(\frac{\sigma_{ij}}{d_{i,j}^{\mathrm{min}}}\right)^{6}\right]+\frac{q_i q_j}{4\pi\epsilon_r\epsilon_0 d_{i,j}^{\mathrm{min}}}-E_{\mathrm{tol}}=0 \quad (10)$$

where ϵ_w is the van der Waals well depth, σ_{ij} is the effective van der Waals radius for atoms i and j, q_i and q_j are the partial charges of atoms i and j, and $\epsilon_r\epsilon_0$ is the dielectric constant. The above equation is solved for $d_{i,j}^{\mathrm{min}}$ with Newton's root-finding method. For the E_{tol} used in this study, there was always a unique solution for $d_{i,j}^{\mathrm{min}}$. If smaller E_{tol} are used, eq 10 may have two solutions, in which case, the smaller solution must be used so as to ensure that $d_{i,j}^{\mathrm{min}}$ reflects the repulsive interaction. The resulting $d_{i,j}^{\mathrm{min}}$ values are stored. A given conformation is treated as valid if all nonbonded atom distances, $d_{i,j}$ (excluding 1−4 pairs) fulfill the criterion $d_{i,j}^{\mathrm{min}} \leq d_{i,j}$. The number of distance computations is kept small by embedding the protein coordinates in a lattice and computing distances only between atoms which have been changed in a given sampling step and atoms which are in the same or adjacent lattice cells.

B. Efficient Side-Chain Sampling Method

Given a set of backbone conformations that is uniformly distributed in ϕ/ψ-torsional space, a uniformly distributed set of full (backbone and side chain) conformations can be build by repeating following steps: (1) randomly selecting a backbone conformation, (2) building all side chains on this backbone conformation, using random torsion angles, and (3) accepting the conformation if it does not produce collisions. This trivial method is not very efficient in practice, first because some backbone conformers may never allow a given side chain to be built without collisions, and second because for a given backbone conformer it is unlikely that placing all side chains at once produces a conformation without collisions. Here, a more efficient method is used that consists of the following steps: (1) For each backbone conformation c, a weight w_c is computed which is equal to the probability that a set of noncolliding side chains can be built on this backbone, when a uniform distribution of side chain torsion angles is used. (2) A random backbone conformation is selected according to the probability $p_c = w_c/\sum_k w_k$. (3) Onto the selected backbone, each side chain is build by itself in a number of conformations that do not produce collisions with the backbone and the nonsampled

regions of the protein. (4) Side-chain conformations from step 3 are combined randomly to form a fully build protein conformation, which is accepted if is does not have any collisions. Steps 2−4 are repeated until a desired number of conformations have been generated.

The weight w_c is computed as follows: For each backbone conformation c, an acceptance probability $p_{c,i}$ for each side chain i is calculated by generating a large number of random rotamers for that side chain (in the absence of the other side chains of the **S** region) and counting the number of noncolliding rotamers. If any $p_{c,i} = 0$ (i.e. some side chain cannot be placed at all without producing collisions), then backbone conformation c is permanently rejected and $w_c = 0$. Otherwise, the probability q_c to find a noncolliding combination of the individually valid side-chain conformations is computed. This is done by generating a large number N_c of random combinations of valid side-chain rotamers and counting the number n_c of noncolliding combinations, $q_c = n_c/N_c$. The weight w_c is obtained as $w_c = q_c \prod_i p_{c,i}$.

C. Efficient Method for Early Rejection of High-Energy Minima

This describes a method for the early rejection of energy minima with $U(\mathbf{x}) > E_{\text{tol}}$ during the minimization. Early rejection is based on statistics that are collected during a number of preliminary full minimizations, correlating the energy difference between partially minimized and final structures with the gradient of the partially minimized structure.

For Ras p21, 100 samples were retrieved from the sample repository and minimized to a gradient RMS of 10^{-5} kcal mol^{-1} Å^{-1}. Each of these minimization trajectories delivered a series of gradients ($\mathbf{g}_0,\mathbf{g}_1,...,\mathbf{g}_n$) and associated potential energies (U_0, U_1,..., U_n), where the pair ($\mathbf{g}_n$, U_n) corresponds to the fully minimized structure. All pairs ($\mathbf{g}_i$, U_i) from all 100 minimization trajectories were used to derive correlation statistics between $\mathbf{g}_i$ and $\Delta U = U_i - U_n$, i.e., the energy difference from the fully minimized structure. These statistics, shown in Figure 12, were used to obtain for each range of gradient a corresponding value of ΔU that was higher than 90% of the ΔU's in that range. This yields an upper estimate of ΔU, given a certain gradient $\mathbf{g}$. This estimate was used to reject structures during minimizations if their minimum energy, predicted from this upper estimate, considerably exceeded an energy tolerance threshold: $U(\mathbf{x}) - \Delta U > E_0 + E_{\text{low}} + 10$ kcal mol^{-1}. E_0 is the minimized reactant energy, and E_{low} was set to 40 kcal mol^{-1} (see Table 2).

D. Barrier Estimation. A method is given for the statistical estimation of lower and upper bounds for the energy barriers of subtransitions. For this, one correlates available information on the edges $e = (u, v)$, such as distance between its vertices $\delta_{uv} = |\mathbf{x}_u - \mathbf{x}_v|$, with the computed energy barriers $B_{uv} = E^{TS}_{uv} - \max\{E^S_u, E^S_v\}$. Using a certain confidence interval, one obtains upper and lower estimates, $B^{\min}_{uv}(\delta_{uv})$ and $B^{\max}_{uv}(\delta_{uv})$, which are used to replace the strict edge-weight bounds by $\max\{E^S_u, E^S_v\} + B^{\min}_{uv}(\delta_{uv})$ and $\max\{E_u, E_v\} + B^{\max}_{uv}(\delta_{uv})$.

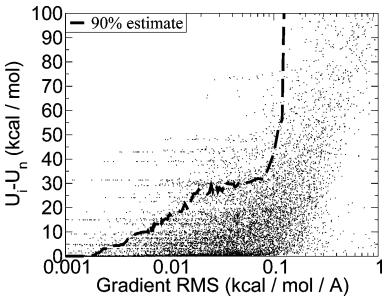

Figure 12. Using the gradient during minimizations to predict the expected energy at the minimum. Based on the minimizations of 100 different conformers, each minimization going through a series of intermediates with gradients ($\mathbf{g}_0,...,\mathbf{g}_n$) and energies ($U_0,..., U_n$), the difference between the energy of an intermediate and the final (minimum) energy, $U_i - U_n$ is plotted against the current gradient $\mathbf{g}_i$. 90% of the points are below the dashed line, which can be used to estimate how much more the energy may decrease during a minimization, based on the current gradient value, thus allowing nonpromising minimizations to be stopped early.

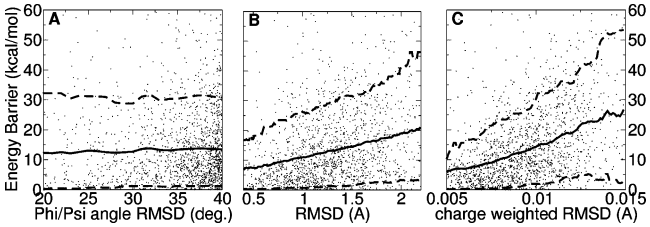

Figure 13. Predicting lower and upper bounds to the energy barrier of subtransitions. The energy barrier is plotted versus the distance between the end states of a given subtransition in Ras p21, using different distance metrics: (A) RMSD in ϕ/ψ-dihedral space of the **S**-regions, (B) all-atom RMSD in Cartesian space, and (C) same, but with each atomic distance weighted by the absolute atomic charge. Solid line: average barrier. 90% of the points lie below the upper dashed line, 10% below the lower dashed line. These were used as lower and upper estimates for the estimation of optimistic and pessimistic best paths (see Figure 9).

For Ras p21, after computing the first ∼2000 energy barriers, these barriers were correlated with the distance between the corresponding minima so as to yield a distance-dependent barrier estimate. Figure 13 shows a plot the first ∼2000 barriers against three different distance measures. The average value and the boundaries of a 90% confidence interval are given. Clearly, the ϕ/ψ-RMSD is not a useful measure here as it is not correlated with the energy barrier. The Cartesian RMSD gives a better correlation, while the charge-weighted RMSD, $d_C(\mathbf{x},\mathbf{y})$, defined as

$$d_C(\mathbf{x},\mathbf{y}) = \sqrt{\frac{\sum_{i=1}^{N}(\mathbf{x}_i - \mathbf{y}_i)^2 q_i^2}{N}}$$

where N is the number of atoms and q_i is the charge on atom i, here gives the best correlation of the three distance measures. The 90% confidence interval was used to derive $B^{\min}_{uv}(\delta_{uv})$ and $B^{\min}_{uv}(\delta_{uv})$.

References

(1) Olsen, K.; Fischer, S.; Karplus, M. A continuous path for the $T \rightarrow R$ allosteric transition of hemoglobin. *Biophys. J.* **2000**, *78*, 394A.

(2) Fischer, S.; Windshuegel, B.; Horak, D.; Holmes, K. C.; Smith, J. C. Structural mechanism of the recovery stroke in the myosin molecular motor. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6873.

(3) Coleman, M. L.; Marshall, C. J.; Olson, M. F. Ras and Rho GTPases in G1-Phase Cell-Cycle Regulation. *Nat. Rev. Mol. Cell Bio.* **1993**, *62*, 851.

(4) Vojtek, A. B.; Der, C. J. Increasing Complexity of the Ras Signaling Pathway. *J. Biol. Chem.* **1998**, *32*, 19925.

(5) Streett, W. B.; Tildesley, D. J.; Saville, G. Multiple time step methods in molecular dynamics. *Mol. Phys.* **1978**, *35*, 639.

(6) Peskin, C. S.; Schlick, T. Molecular dynamics by the backward: Euler's method. *Commun. Pure Appl. Math.* **1989**, *42*, 1001.

(7) Schlick, T.; Barth, E.; Mandziuk, M. Biomolecular dynamics at long time steps: Bridging the time scale gap between simulation and experimentation. *Annu. Rev. Biophys. Biomol. Struct.* **1997**, *26*, 181.

(8) Sanz-Navarro, C. F.; Smith, R. Numerical calculations using the hyper-molecular dynamics method. *Comput. Phys. Comm.* **2001**, *137*, 206.

(9) Amadei, A.; Linssen, A. B.; Berendsen, H. J. C. Essential dynamics of proteins. *Proteins* **1993**, *17*, 412.

(10) Krammer, A.; Lu, H.; Isralewitz, B.; Schulten, K.; Vogel, V. Forced unfolding of the fibronectin type III module reveals a tensile molecular recognition switch. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 1351.

(11) Schlitter, J.; Engels, M.; Krüger, P. Targeted molecular dynamics: A new approach for searching pathways of conformational transitions. *J. Mol. Graphics* **1994**, *12*, 84.

(12) Böckmann, R.; Grubmüller, H. Nanoseconds molecular dynamics simulations of primary mechanical energy transfer steps in F_1-ATP synthase. *Nat. Struct. Biol.* **2002**, *9*, 196.

(13) Noé, F.; Ille, F.; Smith, J. C.; Fischer, S. Automated computation of low-energy pathways for complex rearrangements in proteins: Application to the conformational switch of ras p21. *Proteins* **2005**, *59*, 534.

(14) Grubmüller, H. Predicting slow structural transitions in macromolecular systems: conformational flooding. *Phys. Rev. E* **1995**, *52*, 2893.

(15) Huo, S.; Straub, J. E. The maxflux algorithm for calculating variationally optimized reaction paths for conformational transitions in many body systems at finite temperature. *J. Chem. Phys.* **1997**, *107*, 5000.

(16) Huo, S.; Straub, J. E. Direct computation of long time processes in peptides and proteins: Reaction path study of the coil-to-helix transition in polyalanine. *Proteins* **1999**, *36*, 249.

(17) Czerminski, R.; Elber, R. Self-avoiding walk between two fixed points as a tool to calculate reaction paths in large molecular systems. *Int. J. Quantum Chem.* **1990**, *24*, 167.

(18) Fischer, S.; Karplus, M. Conjugate peak refinement: an algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom. *Chem. Phys. Lett.* **1992**, *194*, 252.

(19) Jónsson, H.; Mills, G.; Jacobsen, K. W. Nudged Elastic Band Method for Finding Minimum Energy Paths of Transitions. In *Classical and Quantum Dynamics in Condensed Phase Simulations*; Berne, B. J., Ciccotti, G., Coker, D. F., Eds.; World Scientific: Singapore, 1998; pp 385−404.

(20) Henkelman, G.; Uberuaga, B. P.; Jonsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.* **2000**, *113*, 9901.

(21) Henkelman, G.; Jóhannesson, G.; Jónsson, H. Methods for Finding Saddle Points and Minimum Energy Paths. In *Progress on Theoretical Chemistry and Physics*; Schwartz, S. D., Ed.; Kluwer Academic: New York, 2000; pp 269−300.

(22) Fischer, S.; Michnick, S.; Karplus, M. A mechanism for rotamase catalysis by the fk506 binding protein (fkbp). *Biochemistry* **1993**, *32*, 13830.

(23) Bondar, A. N.; Elstner, M.; Suhai, S.; Smith, J. C.; Fischer, S. Mechanism of primary proton transfer in bacteriorhodopsin. *Structure* **2004**, *12*, 1281.

(24) Gruia, A. D.; Bondar, A. N.; Smith, J. C.; Fischer, S. Mechanism of a molecular valve in the halorhodopsin chloride pump. *Structure* **2005**, *13*, 617.

(25) Stillinger, F. H.; Weber, T. A. Hidden structure in liquids. *Phys. Rev. A* **1982**, *25*, 978.

(26) Stillinger, F. H. A topographic view of supercooled liquids and glass formation. *Science* **1995**, *267*, 1935.

(27) Czerminski, R.; Elber, R. Reaction path study of conformational transitions and helix formation in a tetrapeptide. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 6963.

(28) Czerminski, R.; Elber, R. Reaction path study of conformational transitions in flexible systems: Application to peptides. *J. Chem. Phys.* **1990**, *92*, 5580.

(29) Berry, R. S.; Breitengraser-Kunz, R. Topography and dynamics of multidimensional interatomic potential surfaces. *Phys. Rev. Lett.* **1995**, *74*, 3951.

(30) Wales, D. J. Structure, Dynamics, and Thermodynamics of Clusters: Tales from Topographic Potential Surfaces. *Science* **1996**, *271*, 925.

(31) Ball, K. D.; Berry, R. S.; Kunz, R. E.; Li, F.-Y.; Proykova, A.; Wales, D. J. From topographies to dynamics on multidimensional potential energy surfaces of atomic clusters. *Science* **1996**, *271*, 963.

(32) Becker, O. M.; Karplus, M. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *J. Chem. Phys.* **1996**, *106*, 1495.

(33) Levy, Y.; Becker, O. M. Effect of conformational constraints on the topography of complex potential energy surfaces. *Phys. Rev. Lett.* **1998**, *81*, 1126.

(34) Miller, M. A.; Wales, D. J. Energy landscape of a model protein. *J. Chem. Phys.* **1999**, *111*, 6610.

(35) Mortenson, P. N.; Wales, D. J. Energy landscapes, global optimization and dynamics of the polyalanine $Ac(ala)_8NHMe$. *J. Chem. Phys.* **2001**, *114*, 6443.

(36) Levy, Y.; Becker, O. M. Energy landscapes of conformationally constrained peptides. *J. Chem. Phys.* **2001**, *114*, 993.

(37) Brooks, C. L., III; Onuchic, J. N.; Wales, D. J. Taking a walk on a landscape. *Science* **2001**, *293*, 612.

(38) Levy, Y.; Jortner, J.; Becker, O. M. Dynamics of hierarchical folding on energy landscapes of hexapeptides. *J. Chem. Phys.* **2001**, *115*, 10533.

(39) Levy, Y.; Jortner, J.; Becker, O. M. Solvent effects on the energy landscapes and folding kinetics of polyalanines. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 2188.

(40) Mortenson, P. N.; Evans, D. A.; Wales, D. J. Energy landscapes of model polyalanines. *J. Chem. Phys.* **2001**, *117*, 1363.

(41) Levy, Y.; Becker, O. M. Conformational polymorphism of wild-type and mutant prion proteins: energy landscape analysis. *Proteins* **2002**, *47*, 458.

(42) Evans, D. A.; Wales, D. J. Free energy landscapes of model peptides and proteins. *J. Chem. Phys.* **2003**, *118*, 3891.

(43) Evans, D. A.; Wales, D. J. The free energy landscape and dynamics of met-enkephalin. *J. Chem. Phys.* **2003**, *119*, 9947.

(44) Wales, D. J.; Doye, J. P. K. Stationary points and dynamics in high-dimensional systems. *J. Chem. Phys.* **2003**, *119*, 12409.

(45) Evans, D. A.; Wales, D. J. Folding of the gb1 hairpin peptide from discrete path sampling. *J. Chem. Phys.* **2004**, *121*, 1080.

(46) Despa, F.; Wales, D. J.; Berry, R. S. Archetypal energy landscapes: Dynamical diagnosis. *J. Chem. Phys.* **2005**, *122*, 024103.

(47) Levy, Y.; Jortner, J.; Becker, O. M. Dynamics of hierarchical folding on energy landscapes of hexapeptides. *J. Chem. Phys.* **2001**, *115*, 10533.

(48) Miller, M. A.; Doye, J. P. K.; Wales, D. J. Energy landscapes of model polyalanines. *J. Chem. Phys.* **2002**, *117*, 1363.

(49) Wales, D. J. Discrete path sampling. *Mol. Phys.* **2002**, *100*, 3285.

(50) Apaydin, M. S.; Brutlag, D. L.; Guestrin, C.; Hsu, D.; Latombe, J.-C.; Varma, C. Stochastic Road map Simulation: An Efficient Representation and Algorithm for Analyzing Molecular Motion. *J. Comput. Bio.* **2003**, *10*, 257.

(51) Dijkstra, E. A note on two problems in connection with graphs. *Num. Math.* **1959**, *1*, 269.

(52) Stillinger, F. H.; Weber, T. A. Packing structures and transitions in liquids and solids. *Science* **1984**, *228*, 983.

(53) McQuarrie, D. A.; Simon, J. D. *Molecular Thermodynamics*; University Science Books: Sausalito, CA, 1999.

(54) Hänggi, P.; Talkner, P.; Borkovec, M. Reaction rate theory: Fifty years after kramers. *Rev. Mod. Phys.* **1990**, *62*, 251.

(55) Berkowitz, M.; Morgan, J. D.; McCammon, J. A.; Northrup, S. H. Diffusion-controlled reactions: A variational formula for the optimum reaction coordinate. *J. Chem. Phys.* **1983**, *79*, 5563.

(56) Elber, R. Reaction Path Studies of Biological Molecules. In *Recent Developments in Theoretical Studies of Proteins (Advanced Series in Physical Chemistry, Vol. 7)*; World Scientific: Singapore, 1996.

(57) Eppstein, D. Finding the *k* shortest paths. *SIAM J. Comp.* **1998**, *28*, 652.

(58) Nagamochi, H.; Ibaraki, T. Computing edge connectivity in multigraphs and capacitated graphs. *SIAM J. Discr. Math.* **1992**, *5*, 54.

(59) Hoffmann, D.; Knapp, E.-W. Polypeptide folding with off-lattice Monte Carlo dynamics: the method. *Eur. Biophysics J.* **1996**, *24*, 387.

(60) Mezei, M. Efficient Monte Carlo sampling for long molecular chains using local moves, tested on a solvated lipid bilayer. *J. Chem. Phys.* **2003**, *118*, 3874.

(61) Lowy, D. R.; Willumsen, B. M.; Function and regulation of ras. *Annu. Rev. Biochem.* **1993**, *62*, 851.

(62) Pai, E. F.; Krengel, U.; Petsko, G. A.; Goody, R. S.; Kabsch, W. and Wittinghofer, A. Refined crystal structure of the triphosphate conformation of H-ras p21 at 1.35 Å resolution: implications for the mechanism of GTP hydrolysis. *EMBO J.* **1990**, *9*, 2351.

(63) Tong, L. A.; de Vos, A.; Milburn, M. V.; Kim, S. H. Crystal structures at 2.2 Å resolution of the catalytic domains of normal ras protein and an oncogenic mutant complexed with GDP. *J. Mol. Biol.* **1991**, *217*, 503.

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

(65) Neria, E.; Fischer, S.; Karplus, M. Simulation of activation free energies in molecular systems. *J. Chem. Phys.* **1996**, *105*, 1902.

(66) Schaefer, M.; Karplus, M. A Comprehensive Analytical Treatment of Continuum Electrostatics. *J. Chem. Phys.* **1996**, *100*, 1578.

(67) Neal, S. E.; Eccleston, J. F.; Webb, M. R. Hydrolysis of GTP by $p21^{NRAS}$, the *NRAS* protooncoge ne product, is accompanied by a conformational change in the wild-type protein. Use of a single fluorescent probe at the catalytic site. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 3562.

CT050162R