Identification and Validation of Reaction Coordinates Describing Protein Functional Motion: Hierarchical Dynamics of T4 Lysozyme

Matthias Ernst, Steffen Wolf, and Gerhard Stock*

Biomolecular Dynamics, Institute of Physics, Albert Ludwigs University, Freiburg, 79104, Germany

Supporting Information

ABSTRACT: While adequately chosen reaction coordinates are expected to reveal the mechanism of a dynamical process, it proves to be notoriously difficult to model the complex structural rearrangements of a macromolecule by a low-dimensional collective coordinate. Adopting the hinge-bending motion of T4 lysozyme (T4L) as a prominent example and performing a 50 μs long unbiased molecular dynamics (MD) simulation of T4L, a general strategy to identify reaction coordinates of protein functional dynamics is developed. As a systematic method to reduce the dimensionality of the dynamics, first various types of principal component analyses are employed, and it is shown that the applicability and outcome of the approach crucially depends on the type of input coordinates used. In a second step, prospective candidates for a reaction coordinate are tested by studying the molecule’s response to external pulling along the coordinate, using targeted MD simulations. While trying to directly enforce the open-closed transition does not recover the two-state behavior of T4L, this transition is triggered by a locking mechanism, by which the side chain of Phe4 changes from a solvent-exposed to a hydrophobically buried state. The mechanism is found to stabilize the open and closed states of T4L and thereby causes their relatively long lifetime of ~10 μs. In extension of the usual two-state picture, a four-state model of the functional motion of T4L is proposed, which describes a hierarchical coupling of the fast nanosecond opening-closing motion and the slow microsecond locking transition.

1. INTRODUCTION

Classical molecular dynamics (MD) simulations offer an atomistic view of the structure, dynamics, and function of molecular systems. To obtain a concise interpretation of the ever-growing amount of simulation data, biomolecular processes such as folding and molecular recognition are often described in terms of the free energy surface:

$$\Delta G(x) = -k_B T \ln P(x)$$

where $P$ is the probability distribution of the molecular system along some (in general multidimensional) reaction coordinate $x$. Characterized by its minima (which represent the metastable conformational states of the system) and its barriers (which connect these states and define the kinetics of the system), the free energy landscape allows us to account for the pathways and their kinetics occurring in a biomolecular process. To obtain a qualitative illustration of the considered process, one-dimensional reaction coordinates such as the radius of gyration or the fraction of native contacts may be sufficient. When we aim for a more quantitative analysis, however, we often find that low-dimensional projections do not reproduce the correct connectivity and barriers between the conformational states, which leads to erroneous interpretations of the energy landscape.

One approach to construct reaction coordinates is to systematically reduce the dimensionality of the problem by introducing a transformation from high-dimensional MD data $r = (r_1, ..., r_M)$ to a low-dimensional collective variable $x = (x_1, ..., x_d)$. While several nonlinear mappings have been suggested, it is often convenient to use a linear transformation such as principal component analysis (PCA) or various versions of independent component analysis (ICA). Assuming a time scale separation between the slow motion of the first few components (representing the “system”) and the fast motion of the remaining components (representing the “bath”), the first components of the transformation may serve as a multidimensional reaction coordinate. Representing the free energy landscape $\Delta G(x)$, this coordinate may be used to construct a Langevin model or in various enhanced sampling techniques. As discussed in detail below, the outcome of the dimensionality reduction may critically depend on various issues, in particular the type of input coordinates.

As the name suggests, adequately chosen reaction coordinates should reveal the mechanism of the considered dynamical process. Although a careful principal component analysis usually provides an efficient preselection of main coordinates involved, the commonly considered first few principal components do not necessarily achieve this goal. For example, a complex structural rearrangement of a macromolecule may include numerous different steps (associated with, e.g., the formation and breaking of hydrogen bonds) that are difficult to model by a low-dimensional collective coordinate. This is particularly so for a hierarchical

Received: June 3, 2017
Published: September 15, 2017
energy landscape, which exhibits several tiers associated with coupled dynamical processes on various time scales.\textsuperscript{21–23} While no generally accepted approach exists to elucidate these more complicated scenarios, one may perform, for example, functional mode analysis to identify all atoms involved in the functional motion,\textsuperscript{24,25} employ various kinds of clustering techniques to define metastable conformational states that may clarify the process,\textsuperscript{26–30} or plainly (but tediously) resort to visual inspection of the structural evolution along the trajectory.

In this work, we suggest to identify useful reaction coordinates obtained by methods such as principal component analysis by assessing if they are causally related to the functional motion of the protein. To do so, we test suitable reaction coordinates by studying the molecule’s response to external pulling along the coordinate, using targeted MD simulations\textsuperscript{31–33} with a distance constraint. As an illustrative example, let us consider a problem that (along some given one-dimensional reaction coordinate) appears to be described by two states (say, A and B) which are connected via a single barrier. When we start in state A, pull the system up the barrier and then let it evolve freely, we expect the system to either relax back to state A (if the pulling is stopped before reaching the top of the barrier) or to state B (if we already pulled it over the barrier). If the pulling direction does not correspond to an appropriate reaction coordinate, however, we might find that the system behaves differently and, for example, goes back to state A, although it already passed the apparent barrier. In this way, pulling simulations may help us to identify suitable reaction coordinates that explain the underlying mechanism of the considered process.

As a prime example of protein functional dynamics, we consider T4 lysozyme (T4L) which has been extensively studied both in experiment\textsuperscript{34–43} as well as in computational work.\textsuperscript{24,25,44–48} The 164-residue enzyme is used by enterobacteria phage T4 to destroy bacterial cell walls by catalyzing the cleavage of glycosidic bonds. The interaction of T4L with the substrate involves a prominent hinge-bending motion of the two domains, which resembles the opening and closing of the mouth of a “Pac-Man\textsuperscript{35}” (Figure 1). The time scale of this motion was recently determined via fluorescence correlation spectroscopy to be in the range of 10 to 20 $\mu$s.\textsuperscript{43} Employing MD simulations, the functional motion of T4L has been analyzed using various methods including full correlation analysis,\textsuperscript{17} functional mode analysis,\textsuperscript{42,45} and various enhanced sampling techniques.\textsuperscript{46–48}

On the basis of a 50 $\mu$s long unbiased MD trajectory and numerous targeted MD simulations, we here pursue various strategies of dimensionality reduction to construct reaction coordinates that reveal the molecular mechanism underlying the hinge-bending motion of T4L. To this end we (i) perform various principal component analyses to determine important coordinates (and associated residues and motions), and (ii) employ targeted MD pulling simulations to select a suitable two-dimensional reaction coordinate. Apart from the obvious open-closed transition, we thus identify a locking transition where the side chain of Phe4 changes from a solvent-exposed to a hydrophobically buried state. This mechanism is found to stabilize the open and closed states of T4L and thereby causes their relatively long lifetime of $\sim 10$ $\mu$s. In extension of the usual two-state picture, we suggest a four-state model of the functional motion of T4L, which also accounts for the locking transition. Performing a mean first passage time analysis, we find a hierarchical coupling of the fast ($\sim 100$ ns) opening-closing motion and the slow ($\sim 10$ $\mu$s) locking transition. The significance of the key residues of the mechanism (Phe4, Phe67, and Phe104) is tested by additional MD simulations of in total six mutants, which are found to considerably modify the energy landscape and the overall kinetics of T4L.

2. METHODS

2.1. MD Simulations. All simulations were performed using the Gromacs package (version 4.6.7),\textsuperscript{49} employing the Amber ff99sb*-ILDN force field\textsuperscript{50–52} and the TIP3P water model.\textsuperscript{53} Following earlier work by Hub and de Groot,\textsuperscript{24} we considered the M6I mutant of T4L (PDB 150L,\textsuperscript{34} chain D; residues 163–164 omitted as they are not resolved in the

Figure 1. (a) Molecular structure of T4L (M6I mutant) adopted from PDB ID 150L,\textsuperscript{34} showing the N-domain ($\beta$-sheets $\beta$1–$\beta$3 and $\alpha$-helix h2) and the C-domain (helices h4 - h10) connected by a hinge formed by helix h3. Residues are color-coded from N-terminus (blue) to C-terminus (red). (b) Side view of the structure, illustrating the hinge-bending motion of the two domains, which resembles a “Pac-Man”\textsuperscript{35} with a mouth and a jaw joint. The open structure is shown in orange, the closed one in blue. (c) Contact map of T4L based on contacts formed in either open and/or closed state. The two domains of T4L are shown as light gray regions, helices 1 and 3 in darker gray shading. Contacts with a sequence difference of less than 4 have been omitted (white diagonal). The upper left region shows all contacts in black, while the lower right region color-codes the contribution to the first principal component of contact PCA: contacts with increasing and decreasing distance upon opening the mouth are shown in red and blue, respectively, while contacts of weak contributions are shown in yellow.
crystal structure). Employing a triclinic box with a NaCl salt concentration of 150 mmol L⁻¹ resulted in ~29 400 atoms in total. Use of the LINCS algorithm to constrain bonds including hydrogen atoms allowed for an integration time step of 2 fs. Particle-Mesh Ewald summation (PME) was used for the calculation of electrostatic interactions. All cut-offs (neighbor search, Verlet scheme, Lennard-Jones interactions, and the real space grid of PME) were set to 1.2 Å.

Following a steepest descent energy minimization of T4L in vacuo to remove sterically unfavorable interactions, the solvation box was constructed and energy minimization in solvent was performed. Equilibration included a 100 ps NVT run using position restraints and the Berendsen thermostat at 300 K, followed by a 1 ns NPT run using position restraints and the Berendsen barostat, a 5 ns free NPT simulation (of which the last 4 ns were used to calculate the averaged box volume), and a 10 ns free NVT simulation. Finally an NVT production run with a total length of 50 μs at 300 K was performed, saving atom coordinates every 1 ps. Visual inspection of the trajectories was carried out with VMD.

Mutations were introduced into the structure of lowest free energy of the closed state during the first 30 μs of the equilibrium run using PyMOL. The simulation protocol was the same as above with a production run length of 10 μs per mutation.

2.2. Principal Component Analysis (PCA). In a PCA, the correlated internal motion of a system with N degrees of freedom \( \mathbf{r} = (r_1, ..., r_N)^T \) is described by the covariance matrix

\[
\sigma_{ij} = \langle (r_i - \langle r_i \rangle)(r_j - \langle r_j \rangle) \rangle
\]

(2)

where \( \langle \ldots \rangle \) represents the average over all sampled conformations. Diagonalization of this covariance matrix results in N eigenvectors \( \{\mathbf{v}^{(i)}\} \) and eigenvalues \( \{\lambda_i\} \), which describe the modes of the collective motion and their respective amplitudes. The principal components (PCs) are then given by the projections of the coordinates \(\mathbf{r}\) onto the eigenvectors

\[
\mathbf{x}_i = \mathbf{v}^{(i)} \cdot \mathbf{r}
\]

(3)

Considering the first few PCs with highest eigenvalues, we may construct a reaction coordinate \( \mathbf{x} = (x_1, ..., x_N)^T \), which accounts for a large part of the system’s fluctuations. Rather than performing a PCA based on the covariance (eq 2) which points out coordinates with high variance, it may be advantageous to consider the correlation (i.e., the normalized covariance) which emphasizes correlated motion. To this end, we normalize each coordinate by its standard deviation, \( \tilde{x}_i = x_i / \sigma_x \), calculate the resulting covariance matrix \( \tilde{\sigma}_{ij} \) (which is equal to the correlation matrix of \( \{r_i\} \)) and its eigenvectors \( \tilde{\mathbf{v}}^{(i)} \), and obtain the corresponding PCs via \( \tilde{x}_i = \tilde{\mathbf{v}}^{(i)} \cdot \mathbf{r} \).

While this strategy appears straightforward, several issues need to be considered to obtain useful reaction coordinates via a PCA. First, the analysis depends crucially on the input coordinates used in the dimensionality reduction method. While Cartesian coordinates are convenient to handle, a Cartesian coordinate PCA is known to break down in the case of large-amplitude motion (as occurring, e.g., in a folding process), since structural dynamics of flexible molecules necessarily results in a mixing of overall and internal motion. To circumvent this problem, internal coordinates such as \( (\phi, \psi) \) backbone dihedral angles or distances and contacts between atoms and protein residues may be used. While backbone dihedral angle PCA has proven quite powerful to model the folding dynamics of peptides, RNA, and small proteins, distance-based PCAs also take into account the structure and dynamics of side chains and may therefore be advantageous to describe the functional dynamics of larger proteins. In both cases, it has been found important to perform a preselection of the coordinates to be included in the analysis. That is, coordinates reflecting irrelevant motion (e.g., uncorrelated motion of flexible terminal residues) or hardly any motion (such as essentially rigid secondary structures of the protein core) should be excluded in order to minimize noise in the analysis.

2.2.1. Dihedral Angle PCA. To perform a PCA on circular variables such as angles \( \phi \), we may change to sine/cosine-transformed coordinates \( (r_1 = \cos \phi, r_2 = \sin \phi) \) to obtain a linear coordinate space with the usual Euclidean distance as induced metric. Alternatively, it has recently been suggested to directly calculate the covariance matrix of angular variables by (i) associating the circular distance as the inner arc between two angles on the unit circle and (ii) defining the circular mean \( \langle \phi \rangle \) over various observations \( \{\phi_n\} \) as the projection onto the unit circle of the average of the sine and cosine projections \( x = \frac{1}{N} \sum_{n=1}^{N} \sin \phi_n \) and \( y = \frac{1}{N} \sum_{n=1}^{N} \cos \phi_n \) as obtained by the atan2 function, \( \langle \phi \rangle = \frac{\arctan2(y,x)}{\pi} \). When we project the data \( \mathbf{\phi} \) onto the resulting eigenvectors \( \mathbf{v}^{(i)} \) according to eq 3, we again need to account for the periodicity of the data. To this end, we introduce a cut of the periodic space (say, between \( \phi_{\min} = -\pi \) and \( \phi_{\max} = \pi \)) and shift all angles \( (\phi \to \phi + \delta) \) such that populated sections of the angle distribution remain connected.

In the case of backbone dihedral angles \( (\phi, \psi) \), the cut is naturally introduced at the (hardly crossed) barriers at \( \phi \approx 0^\circ \) and \( \psi \approx 20^\circ \). See ref 71 for details on this recently proposed method termed dPCAs.

2.2.2. Contact PCA. Numerous options exist to perform a PCA based on interatomic distances or contacts. Following our recent work, here we define a contact to be formed if the distance \( D_{ij} \) between the closest non-hydrogen atoms of two residues \( i \) and \( j \) is shorter than 4.5 Å,

\[
D_{ij} = \min(|r_{ik} - r_{lj}|) \leq 4.5 \text{Å}
\]

(4)

with the indices \( k \) and \( l \) running over all heavy atoms of the selected residue pair. In addition, we discard contacts between residues less than four residues apart in sequence, effectively omitting short-range contacts as, for example, in helical structure elements or turns.

Employing this definition, first the contacts of T4L need to be determined from a suitable reference structure. As neither the open nor the closed state of T4L contains all relevant contacts, we considered the contacts of both states, using the (energy minimized) crystal structure for the open state and the MD structure with the lowest radius of gyration for the closed state. The resulting contact matrix shown in Figure 1c reveals in total 402 contacts, including stabilizing hydrogen bonds of the \( \alpha \) and \( \beta \) secondary structures and numerous tertiary contacts. Using the MDAnalysis framework, we calculated the contact distance \( D_{ij} \) (now without threshold) of all identified contacts for all frames of the trajectory. This data set of distances is subsequently used as input for the contact PCA according to eq 2.
Figure 2. (Left) Time evolution of (a) the radius of gyration (red) and the first principal component obtained from (b) Cartesian PCA (orange), (c) backbone dihedral angle PCA+ (green), (d) contact PCA (purple), (e) mouth contact PCA (blue), and (f) the locking coordinate \(p\) (gray). (Right) Corresponding free energy curves along these reaction coordinates.

Simulations\(^{31-33}\) which constrain the velocity in the direction of \(x\) and provide an easy means to calculate the free energy profile \(\Delta G(x)\). Restricting ourselves to a 1D reaction coordinate \(x\), we drive the system from \(x(0) = x_0\) to \(x(x_{\text{end}}) = x_{\text{end}}\) by constraining coordinate \(x(t)\) to \(x_i(t) = x_0 + v_i t\) with a constant velocity \(v_i\) via the constraint function

\[
\Phi(x(t)) = (x(t) - x_i(t))^2 = 0
\]  

(5)

The constraint is realized via the force

\[
F_x = \lambda \frac{d\Phi(x(t))}{dx} = 2\lambda(x(t) - x_i(t))
\]  

(6)

where \(\lambda\) is a Lagrange parameter. In the present application, the pulling coordinate was chosen as the distance between the centers of mass of two atom groups of T4L. To account for different masses of the individual atoms, the force acting on atom \(i\) was scaled by a factor \(\sqrt{m_i/m_w}\), where \(m_i\) and \(m_w\) represent the mass of atom \(i\) and the average atom mass of the individual pull group, respectively. The TMD simulations were performed employing the PULL code as implemented in GROMACS 4.6.7,\(^{36,37}\) using the “constraint” mode and a constraint velocity of \(v_i = 0.125\ \text{Å ns}^{-1}\) for pulling whole domains and \(v_i = 0.5\ \text{Å ns}^{-1}\) for pulling selected side chains.

To calculate the free energy change \(\Delta G_{\text{TMD}} = G(x_{\text{end}}) - G(x_0)\) from the TMD simulations, we employed a thermodynamic integration scheme\(^1\)

\[
\Delta G_{\text{TMD}} = \int_{x_0}^{x_{\text{end}}} \frac{dG}{dx} dx = \int_{x_0}^{x_{\text{end}}} \langle F_x \rangle_x dx \approx \sum_{i=1}^{N} \langle F_x \rangle_x \Delta x
\]  

(7)

where \(\langle F_x \rangle_x\) represents an equilibrium time average of the constraining force at point \(x\). The averages were calculated from 200 ns long TMD runs at fixed position \(x_i\) (i.e., for \(v_i = 0\)), using only the last 20 ns to evaluate \(\langle F_x \rangle_x\). Typically \(N = 16\) equidistant points were used to approximate the integral.

2.4. Mean First Passage Time Analysis. Considering a system propagating in a discrete state space, the average transition time \(\tau\) between two states \(i\) and \(j\) can be estimated via the mean first passage time (MFPT).\(^73\) Unlike \(\tau\) calculated by counting transitions within a predefined lag time, the MFPT is more general as it avoids any Markovian assumption. In the simplest case of a system with two states (say, 0 and 1) and a trajectory with a single transition event \(0 \rightarrow 1\), which spends \(\tau_{01}\) time steps \(\delta t\) in initial state 0 before it jumps to state 1, the MFPT is defined as\(^73\)

\[
\tau_{01}^{\text{MFPT}} = \frac{1}{n_{01}} \sum_{l=1}^{n_{01}} \delta t = \frac{n_{01}}{2} + \frac{1}{2} \delta t
\]  

(8)

That is, the MFPT represents the sum of all first passage times divided by the number of time steps spent in the initial state. For a two-state system, the MFPT is therefore just half of the average time waited in the state.

We now generalize to the case of more than two states (\(i = 1, 2, \ldots\)) and consider a trajectory that exhibits multiple \((m > 1)\) transition events \(i \rightarrow j\). Defining \(n^{(k)}_{ij}\) as the number of time steps spent in initial state \(i\) occurring in event \(k\), the total number of time steps the system spends in state \(i\) before it jumps to state \(j\) is \(n_{ij} = \sum_{k=1}^{m} n^{(k)}_{ij}\). Hence the MFPT amounts to

\[
\tau_{i\rightarrow j}^{\text{MFPT}} = \frac{1}{n_{ij}} \sum_{l=1}^{m} \sum_{k=1}^{n^{(k)}_{ij}} \delta t
\]  

(9)

The outcome of a MFPT analysis depends crucially on the partitioning of the continuous MD data into discrete states. Using appropriate reaction coordinates (e.g., \(p\) and \(x\) defined in section 3.6), to this end we first applied density-based geostatistical clustering\(^28\) which yields well-defined microstates separated by local free energy barriers. As the projection on a low-dimensional space may induce spurious transitions in the vicinity of energy barriers, in a second step we identify core regions of the microstates and count transitions only if the core region of the other state is reached.\(^30,74\) Effectively, this procedure generates a state trajectory \(i(t)\) with clear-cut state boundaries, on the basis of which an appropriate MFPT analysis can be carried out. We used an elliptical shape for the core regions (Figure S2), which was chosen to match the shape of the minima of the free energy landscape (Figure 6a). Varying the size of the core regions resulted in no significant changes of the MFPT (see Table S1).

3. RESULTS AND DISCUSSION

3.1. 1D Order Parameters Can Provide a Qualitative Picture of Protein Motion. To give an overview of the functional dynamics of T4L, we first consider the radius of gyration \(R_G\) as a commonly used one-dimensional (1D) order parameter. Figure 2a displays the time evolution of \(R_G\), obtained from the 50 μs MD trajectory. Despite large
fl less 2b). As shown in Figure S3, however, the free energy curves involved in the motion along evolution of the radius of gyration (Figure 2a). This is in line with the root-mean-square distance (RMSD) to the (open or closed) reference structure gave similar results (data not shown).

While the radius of gyration provides a qualitative picture of the functional dynamics of T4L (such as the overall amplitude and the time scale), it may not tell much about the molecular mechanism causing the process. This is indicated by the free energy curve along $R_G$ (right panel of Figure 2a), showing two minima associated with the open and closed states, that are separated by an energy barrier $\Delta G_{\text{gg}} \lesssim 1 k_BT$. Modeling the transition rate by the standard expression

$$k = 1/\tau = k_0 e^{-\Delta G_{\text{gg}}/k_BT}$$

(10)

it is obvious that a small energy barrier $\Delta G_{\text{gg}} \approx 1k_BT$ cannot account for the long ($\sim 10 \mu s$) observed transition time $\tau$ of T4L. That is, the rate is almost solely caused by the prefactor $k_0$, which depends critically on the diffusion coefficients of the system. Rather we find that the projection of the many-dimensional molecular motion onto the 1D coordinate causes large fluctuations of $R_G$, which result in a small barrier. Hence the radius of gyration reflects the consequences rather than the origin of the functional motion of T4L. Although it is possible to further analyze 1D order parameters by identifying all atoms involved in its time evolution (e.g., via functional mode analysis24,25), it is not clear to what extent such an analysis sheds light on the underlying mechanism. For a complex system as such as T4L, 1D order parameters are apparently not adequate to uncover microscopic reaction pathways.

### 3.2. Cartesian PCA May Not Work for Multievent Trajectories.

As discussed in the Introduction, dimensionality reduction methods such as principal component analysis13,14 (PCA) or independent component analysis15,17,18 may provide an approach to identify suitable reaction coordinates in a systematic manner. With this end in mind, we first consider a standard PCA performed on the Cartesian coordinates of the backbone atoms of T4L. The time evolution of the resulting first PC, $x_1(t)$ shown in Figure 2b is overall quite similar to the evolution of the radius of gyration (Figure 2a). This is in line with the observation that similar atoms are prominently involved in the motion along $x_1(t)$ and along $R_G(t)$. Exhibiting less fluctuations than the latter, $x_1(t)$ results in a double-well free energy curve $\Delta G(x_1)$ with a barrier of about $2k_BT$ (Figure 2b). As shown in Figure S3, however, the free energy curves $\Delta G(x_i)$ of all higher PCs are structureless and hardly provide any further information on the reaction mechanism. In other words, the Cartesian PCA does not fulfill the promise of a systematic construction of a multidimensional reaction coordinate for T4L.

At first this finding appears surprising, since a previous Cartesian PCA study of a 460 ns long trajectory of T4L revealed a twist motion along PC2 and a torsional motion of the two domains along PC3.24 By studying short single-transition pieces of our trajectory, we could indeed qualitatively recover these findings, which rules out that the effect is caused by different force fields, etc. Hence the main difference between the previous and our study is that (with nowadays improved computational power) we use a $\sim 100$ times longer trajectory that covers several transition events instead of one. On the one hand, it is expected that longer MD simulations monitor a higher variance of the coordinates, which may reduce the resolution of the resulting free energy landscape. On the other hand, an additional issue exists for the Cartesian PCA, which requires a preceding rotational fit of the 3N Cartesian coordinates on a reference structure in order to separate internal and overall motion.76 Since the RMSD of the open and closed states differ by about 1 Å (the amplitude of the hinge-bending motion of the “mouth residues” can be up to 15 Å), the required fitting to a single reference structure (e.g., a crystal structure34 of the open state, PDB ID 150L) may introduce significant coupling of the internal motion and the overall rotation, thus obscuring the resulting structural analysis.61 This is less significant for short MD runs that stay relatively close to the reference structure or perform only a single transition to one specific target state which thus can be distinguished, but may become a serious problem for long trajectories showing numerous events and various transition pathways that may proceed via different intermediate structures. Similar findings have been obtained for the large-amplitude motion of various folding peptides and proteins (where this effect is expected), but also for the small-amplitude functional motion of BPTI.60,61

### 3.3. Dihedral Angle PCA Can Be Inadequate for Folded Proteins.

To circumvent the fitting problem of Cartesian coordinates, we resort to internal coordinates such as distances or angles. In particular, dihedral angle PCA (dPCA) using $(\phi, \psi)$ backbone dihedral angles has proven quite powerful to model the conformational dynamics of peptides, RNA, and small proteins.63-67 To obtain a first impression of the conformational distribution $P(\phi, \psi)$ of T4L, Figure S1 shows the Ramachandran plots of all residues $i$ that change significantly upon the open-closed transition. We find that T4L is in fact quite stable in terms of the backbone dihedral angles. Among these, about 130 residues hardly change at all, about 20 residues exhibit a small shift of the distribution, and less than 10 residues undergo a transition between two conformational states. Caused by local loop flips of the residue, the latter are mostly located in the $\beta_1-\beta_2$ turn and the $\beta_2-\beta_3$ turn (Figure S1). When dPCA+ is performed with the use of the covariance matrix (see Methods), the first principal components naturally focus on these large-amplitude torsional motions (see Figure SS). While showing one of the numerous facets of the conformational dynamics of T4L, however, these local motions are hardly correlated to the main functional dynamics and therefore of little interest here.

Nonetheless, it may be instructive to consider a PCA of the corresponding dihedral angle correlation matrix (see Methods). When focus is on the correlated (but not necessarily large-amplitude) motion of the backbone dihedral angles, the time evolution of the resulting first PC $x_1(t)$ in Figure 2c is seen to monitor the overall opening-closing motion of T4L, quite similar to the evolution of the radius of gyration $R_G(t)$ in Figure 2a. Interestingly, though, the transitions of $x_1(t)$ occur gradually rather than via jumps as found for $R_G(t)$. A closer analysis reveals that this is caused by the cumulative motion of the above-mentioned $\sim 20$ residues that exhibit a small shift of their $(\phi, \psi)$ distribution. The effect can be most clearly seen for the residues of the long helix 3, which is bent by the opening–closing motion (Figure S6). However, the barrier of the corresponding free energy curve is rather small (Figure 2c) and all higher PCs are structureless (Figure SS), so that dPCAs on
the backbone dihedral angle correlation matrix does not reveal much new information on the functional dynamics of T4L.

We also performed a PCA on the side chain dihedral angles of T4L. Considering the first dihedral angle of the side chain, \( \chi_1 \), we find that only relatively few residues show significant changes upon the open-closed transition (Figure S7), in particular residues 4, 7, 8, and 104. The following side chain angles, \( \chi_{2,3,4} \), on the other hand, typically show frequent changes between several rotameric states. Being mostly unrelated to the open-closed transition, these changes were found to result in more noise than signal in a PCA, and lead to a low resolution of the resulting energy landscapes or timetraces (Figure S7).

Using the correlation matrix, we find that the time trace of the first PC \( x_1(t) \) monitors the overall open-closed motion of T4L (Figure S7), while all higher PCs are structureless (Figure S8).

To summarize, Figures S5 and S8 demonstrate that the functional hinge-bending motion of T4L requires only minor changes of the system’s individual backbone and side chain dihedral angles, which are consequently of little use for a further analysis such as PCA. On the other hand, the open-closed motion in Figure 1b is expected to require a significant change of the contact network of T4L, which will be studied in what follows.

### 3.4. Contact PCA May Identify Functionally Important Residues

As detailed in the Methods section and illustrated by the contact map in Figure 1c, the structural analysis of the open and closed forms of T4L revealed 402 contacts, including stabilizing hydrogen bonds of the \( \alpha \) and \( \beta \) secondary structures and numerous tertiary contacts. By performing a PCA of the corresponding distance-based correlation matrix, we find that the free energy curves of (at least) the first three PCs are structured (Figure S9), which renders the contact PCA a promising approach to construct a reaction coordinate. Figure 3a shows a 3D density plot of the conformational distribution along these PCs, which clearly shows seven conformational states of the system. Structural analysis reveals that PC1 accounts for the overall hinge-bending motion, PC2 for a twist-like rearrangement in the N-terminal domain of T4L and PC3 for a rocking motion of helix 1 and the N-terminal domain (Figure S10), which is in line with previous studies. The main difference between state pairs 1-2, 3-4, and 5-6 is the opening or closing of the hinge (along PC1), while PCs 2 and 3 mirror side states, involving rearrangements in the N-terminal domain. Notably, the opening and closing motion can also occur when the system is in some of these side states, indicating independent motions.

It is instructive to first discuss some features of PC1. As found for the other kinds of PCA, the time evolution of PC1 is similar to the evolution of the radius of gyration and results in a double-well potential with a barrier of about 3 \( k_B T \) (Figure 2d). Moreover, the eigenvectors of the contact PCA readily illustrate the structural differences mapped out by the PCs. Discriminating the open and the closed state of T4L, for example, the eigenvector of PC1 directly reveals contacts formed and broken by the overall hinge-bending motion (Figure 3c). Upon closing, we find several forming contacts located at the tip of the mouth (around residues 21 and 141) and at the jaw joint (e.g., 8-67). On the other hand, upon closing we also find breaking of contacts, which are typically part of tight packing of hydrophobic side chains at the jaw joint (e.g., 4-60 and 4-63). Indicated in Figure 3d by colored lines in the molecular structures of the open and closed form of T4L, these contacts nicely illustrate the main structural determinants of the “Pacman”. Rather than simple formation of hydrogen bonds at the mouth, the close-open transition involves a complex pattern of contact breaking and formation at different positions within the protein.

To account for the dynamics represented by the various PCs, Figure 3b displays the time evolution of the corresponding autocorrelation functions with standard deviations shown in Figure S9c. Decaying on a \( \mu s \) time scale, the first three PCs seem to reflect the slowest motions of the MD trajectory and are therefore promising candidates to construct a reaction coordinate. In what follows, we will employ targeted MD simulations to study to what extent these coordinates indeed account for the sought-after opening-closing mechanism.

### 3.5. TMD Simulations Test Prospective Reaction Coordinates

To test if a chosen reaction coordinate is causally related to the functional dynamics of T4L, we want to study the molecule’s response to external pulling along it by performing targeted MD (TMD) simulations. As the first PC of all considered PCAs was found to behave similar to the radius of gyration, we employed a reaction coordinate of \( \mu_s\) (M), which was defined as the first nonzero eigenvector of the corresponding distance-based correlation matrix. For both opening and closing trials, the TMD simulations were performed in two steps: first, a fully unconstrained simulation was carried out for 10 ns to allow for equilibration under the external pulling force. The coordinates were then restrained to the values obtained from the last 8 ns of the unconstrained run and the pulling force was gradually turned off. During the next 5 ns of MD, the system was driven toward the closure state, and the pulling force was ramped up to its target value of \( \mu_s\) (M). Finally, the system was propagated for another 5 ns to equilibrate the new conformation. In what follows, we will employ targeted MD simulations to study to what extent these coordinates indeed account for the sought-after opening-closing mechanism.
of gyration $R_g$ (Figure 2), we first consider a pulling coordinate that mimics $R_g$. To construct a distance coordinate $x$ that is suited for TMD simulations, we combine all $C_α$ atoms of the N- and the C-terminal domain into one group each, respectively, and pull the system along the vector connecting the centers of mass of the two groups. Starting in the open state, Figure 4a compares the resulting potential of mean force $ΔG_{\text{TMD}}(x)$ [eq 7] obtained by thermodynamic integration to the corresponding free energy profile $ΔG_{\text{MD}}(x)$ of the unbiased MD simulations. While the latter shows the above-discussed double-well potential corresponding to the open and closed states of T4L, $ΔG_{\text{TMD}}(x)$ continues to increase throughout the TMD simulations. Projection of the pull run onto the free energy landscape of PC1 and PC2 obtained by contact PCA (Figure 4b) shows that we pull the system up a free energy barrier rather than toward the minimum corresponding to the closed state. Similarly, starting in the open state, we do not manage to close the mouth by pulling the same coordinate (Figure S11a,b). To test if we failed to trigger the hinge-bending transition because of pulling at too many residues in an inadequate way, we furthermore restricted the pull groups to the $C_α$ atoms of the mouth residues (20, 21, 22 and 137, 141, 142, 145, respectively) However, this again did not induce the open-closed transition (Figure S12).

Obviously, a reaction coordinate $x$ that basically pushes or pulls at the two protein domains to open or close T4L cannot account for the observed two-state behavior of the protein. Although the projection of the unbiased MD simulation trajectory on such coordinates shows a double-well free energy profile (Figure 4a), this does not mean that the desired conformational transition occurs when we force the system to go along $x$. While the unbiased MD typically uses minimum energy pathways on the multidimensional free energy surface, pulling along a coordinate directly connecting two minima might force the system to go over peaks rather than over passes of the energy landscape, which not necessarily leads to the desired target state.

As a further test of a reaction coordinate, we may let the system evolve freely after pulling. (That is, we stop the nonequilibrium TMD simulation at some point and continue with an unbiased MD simulation.) Using PC1 and PC2 of the contact PCA to represent the energy landscape of T4L, Figure S13 shows the evolution of the system after pulling (starting in the open state) and pushing (starting in the closed state), respectively. Similarly, Figure S14 shows the same projections for using the mouth residues only as pull groups. We see that in all cases, the system quickly reverts back to the state in which we started the pull runs without undergoing a hinge-bending transition. This response provides a further indication that pulling along a coordinate mimicking the radius of gyration is not causally related to the open-closed transition of T4L. It is this test of causality that mainly distinguishes our TMD-based analysis of reaction coordinates from other approaches (such as committor analysis) to assess the quality of reaction coordinates.

3.6. Locking Transition of Phe4. Similarly as done for the distance coordinate accounting for the radius of gyration, we may test the system’s response to pulling along various other choices of the reaction coordinate. Rather than directly using some PC discussed above, we found it instructive to analyze the contributions of individual residues to the leading eigenvectors of a PCA, especially of the contact PCA. Proceeding this way, we discovered a prominent role of residue 4, which is a phenylalanine that is located at the jaw joint of T4L and shows up in the side chain dPCA$^+$ as well as in the top four contacts of the contact PCA. Examining the structure of Phe4 in the trajectory, we find the side chain to exhibit two distinct orientations:

- In the open state, the side chain is buried inside a hydrophobic cavity (formed by residues Asn2, Glu5, Leu7, Arg8, Leu13, Ile29, Lys60, Ala63, Glu64, and Phe67) which restricts its motion. Additionally, some of these residues may interconnect via side chain hydrogen bonds, such as Arg8 to Glu64 (Figure S17).
- In the closed state, the side chain is outside of the cavity. It is fully exposed to the solvent and free to move.

Zooming into the region around Phe4, Figure 5 highlights the structural differences of the buried and exposed state of the side chain of Phe4. In particular, the illustration reveals that the transition between these states requires Phe4 to bypass the side chain of Phe67. Because of the large size of the phenyl rings, this motion is expected to amount to a high energy barrier. The transition of Phe4 from a solvent-exposed to a hydrophobically buried state thus constitutes a “locking mechanism”, which stabilizes the open and closed states of T4L.

To establish an order parameter that accounts for the locking transition, we define

$$p = \frac{d_{60,67}d_{5,4}}{d_{60,67}^2}$$  \hspace{1cm} (11)

where $d_{ij}$ represents the distance vector between the $Cα$ atoms of Lys60 or Phe67 and the center of mass of the carbon atoms of the Phe4 phenyl group. As illustrated in Figure 5d, the locking coordinate $p$ represents the orthogonal projection of vector $d_{67,4}$ (which accounts for the relative positions of the
bypassing residues Phe4 and Phe67 onto vector $d_{067}$ (which represents the hydrophobic cavity). In this way, $p > 0$ indicates that the side chain of Phe4 is solvent-exposed, while $p < 0$ indicates that the side chain is buried inside the hydrophobic cavity. Showing the time evolution of the locking coordinate and the corresponding free energy profile, Figure 2f reveals that $p(t)$ essentially follows the open-closed pattern displayed by the other order parameters such as $R_G$. Contrary to the other coordinates, though, the variance of the fluctuations in the open and the closed state is significantly reduced, which results in a relatively high barrier ($\sim 6k_BT$). Besides the two main states, moreover, the free energy curve along $p$ exhibits a shallow intermediate state which will be discussed below.

As a test if the locking coordinate $p$ is a suitable candidate for a reaction coordinate of T4L, we performed TMD simulations that pull the Phe4 side chain out of or into the hydrophobic cavity. To define a distance-type pulling coordinate that mimics $p$, we again used the centers of mass of two atom groups. These are (i) the carbon atoms of the phenyl ring of Phe4 and (ii) the side chain carbon atoms of Glu64 together with the $C_\alpha$ and $C_\beta$ atoms of Lys60. The latter two residues are part of the walls of the hydrophobic pocket that encloses the phenyl group of Phe4 in the open state. In other words, we separated the "key" residue Phe4 and the side chains forming the "lock" cavity into two pull groups. Starting in the hydrophobically buried state of Phe4 (corresponding to the open state of T4L), Figure 4c shows that the resulting potential of mean force $\Delta G_{TMD}(x)$ [eq 7] exhibits two main minima separated by a barrier of $\sim 5k_BT$, quite similar to the corresponding free energy profile $\Delta G_{MD}(x)$ of the unbiased MD simulations. Also the shallow intermediate state is reproduced by the TMD simulations. Projection of the pull run onto the free energy landscape along PC1 and PC2 (Figure 4d) confirms that we now successfully induced a hinge-bending transition. The same effect is observed when we start in the free state of the Phe4 side chain and pull it inside the cavity (Figure S11).

Even more revealing is the response of the free system to the TMD runs. That is, performing unbiased MD simulations subsequent to nonequilibrium pulling of Phe4 from the buried to the free state, we find that the system relaxes into the closed state of T4L (see Figure S15). This is remarkable, since we did not at all enforce the open–closed transition per se, but it nonetheless happened as a consequence of enforcing the locking transition. Hence, the open–closed transition of T4L appears to be indeed causally related to the locking transition of Phe4.

While Phe4 has hardly been in the focus of the extensive literature existing for T4L, in retrospect we found several mentions of it. For example, differences in the orientation of the Phe4 side chain were reported by Dixon et al., 38 who analyzed crystal structures and thermal stability of various mutations of T4L and speculated about a coupled motion of residues Phe4, Phe67, and Phe104 driving the hinge-bending motion. Introducing mutations which trapped the hinge-bending motion at distinct angles, Zhang et al.39 found a connection between the side chain dihedral angles of Phe4, Phe67, and Phe104 with the hinge-bending angles in crystal structures of the respective mutants. Phe4 also exhibited a change in position in the crystal structure of a circular permutant, 49 where helix-1 was cut off from the N-terminal domain and fused to the C-terminal domain. (In the wildtype, helix-1 functionally belongs to the C-terminal domain but is covalently attached to the N-terminal domain.)

It is interesting to note that the locking mechanism appears to explain the somewhat unusual finding that both open and closed forms of T4L occur with similar probability. In fact, solvent entropy is expected to drive the protein into its maximally contracted structure (i.e., the closed form). In the closed state, however, the solvent exposed ("free") side chain of Phe4 increases the hydrophobic solvent accessible surface area (cf. Figure S16), which in turn lowers the solvent entropy. On the other hand, in the open state the hydrophobically buried ("locked") side chain of Phe4 decreases the hydrophobic solvent accessible surface area. This results in a balance between the solvent water entropies in the open locked state and free closed state, which stabilizes both states. The balance maybe important for the protein’s function, as T4L only binds to its substrate in the open state.

3.7. Minimal Model of the Functional Dynamics of T4L. We are now in a position to combine all findings to construct a minimal model of the functional dynamics of T4L. Apart from coordinate $p$ [eq 11] accounting for the locking of Phe4, the model should also include a coordinate that describes the mouth opening of T4L. While this motion is qualitatively described by the first PC of several PCAs (see Figure 2), these coordinates typically also account for the motion of Phe4 which is already included in the locking coordinate $p$. To avoid redundancy, we therefore focus on residues of the mouth of T4L and perform a PCA using only contacts between residues 20–22 of the N-terminal domain and residues 137, 141, 142, and 145 of the C-terminal domain. Figure 2e shows the time evolution of the resulting first PC which again follows the overall open-closed pattern. Owing to its definition, however, it naturally describes the closed state with little variance and the open state with large variance (i.e., just the other way as $p(t)$).
In this way, the “opening coordinate” \( x(t) \) describes the relatively frequent (\( \sim 1 \text{ ns} \) time scale) attempts of mouth opening or closing in the open state of T4L, while the locking coordinate \( p(t) \) mainly account for the rare (\( \sim 1 \mu s \) time scale) locking transition.

Figure 6a shows the resulting free energy landscape \( \Delta G(x, p) \) of T4L, where the opening coordinate discriminates between the open \((x \geq -4)\) and closed \((x \leq -4)\) states, while the locking coordinate discriminates between closed and open states. The system exhibits two main metastable conformational states, the open locked state 1 and the closed free state 4, as well as two intermediate states 2 and 3. Minimal model of the functional dynamics of T4L, showing the metastable conformational states together with the corresponding mean first passage times (in units of \( \mu s \)).

The landscape exhibits four metastable conformational states labeled by numbers 1–4, including the prominent open locked state 1 and closed free state 4, as well as two sparsely populated intermediate closed states 2 and 3. As expected, we find that the open locked state 1 shows large variance in the opening coordinate, reflecting the wide distribution of mouth opening distances in the open state, while the closed free state 4 shows a large variance in the locking coordinate, reflecting the many possible conformations of Phe4 in the free state.

Intermediate state 2, the mouth is closed, but the side chain of Phe4 still is buried in the hydrophobic cavity. Hence there is strain in the system which drives it back to the open state, if the transition in \( p \) cannot be completed. In intermediate state 3, on the other hand, the side chain of Phe4 is found outside of the cavity and therefore in principle is free to move. While this also holds for state 4 and although these states are connected by only a minor barrier, there are nonetheless distinct structural differences between states 3 and 4. They manifest themselves in terms of the alignment of the Phe4 side chain, which needs to find a position on the protein surface that marks the entrance into the hydrophobic cavity (state 3), but frequently is sterically blocked by Arg8 (state 4). As displayed in Figure S17, the phenyl ring of Phe4 tries to reduce its nonpolar solvent accessible surface area in both states, by sticking to a hydrophobic patch at the outside of the protein and bridging a gap between Asn2, Ile3, Phe67, Asn68, and Val71. In state 4, Arg8 is found over the entrance to the hydrophobic binding pocket, forming a partially bidentate salt bridge with Glu64. In state 3, the side chain of Phe4 is positioned over the entrance, while Arg8 is displaced and the salt bridge with Glu64 perturbed, not allowing the highly stable bidentate connection. Besides this local change in connectivity, state 3 also exhibits a lower total number of intraprotein hydrogen bonds than state 4.

The free energy landscape \( \Delta G(x, p) \) indicates that transitions between the main states 1 and 4 require subsequent closing and unlocking of T4L. To elucidate this process, Figure S18 shows the time evolution of trajectory pieces during and shortly before/after the \( 1 \leftrightarrow 4 \) transitions. As expected from the appearance of the energy landscape, we find that most transitions proceed along the route \( 1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \) and back along \( 4 \rightarrow 3 \rightarrow 2 \rightarrow 1 \). To estimate the time scales associated with the individual steps of the process, we performed a mean first passage time (MFPT) analysis of the four-state system (see Methods). The resulting MFPT are comprised in Table 1, and shown to be quite robust against variation of the cores used for assigning the states (Table S1). The next-neighbor MFPTs are also included in the illustrative scheme in Figure 6b, which provides the basis for the following discussion.

Let us first consider the \( 1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \) pathway. We find a MFPT of 6.5 \( \mu s \) for the overall \( 1 \rightarrow 4 \) transition, which corresponds to an average waiting time of 13.0 \( \mu s \) in state 1 (see Methods). As expected from the discussion above, the locking transition \( 2 \rightarrow 3 \) with a MFPT of 3.2 \( \mu s \) represents the slowest step of the route. This is about 1 and 2 orders of magnitude slower than the other two steps \( 1 \rightarrow 2 \) and \( 3 \rightarrow 4 \), respectively. For the other direction of the process, \( 4 \rightarrow 3 \rightarrow 2 \rightarrow 1 \), we find a significantly shorter overall MFPT of 1.2 \( \mu s \), which is mostly caused by the relatively fast \( (0.5 \mu s) \) unlocking step \( 3 \rightarrow 2 \). Out of total 16 transitions between states 1 and 4, 11 follow directly the proposed “L”-shaped route \( 1 \leftrightarrow 2 \rightarrow 3 \leftrightarrow 4 \), while the remaining five transitions follow this route closely but may leave out states 2 or 3 (cf. Figure S18).

Hence we find that the relatively long MFPT of 6.5 \( \mu s \) for the overall \( 1 \rightarrow 4 \) transition is caused by a hierarchical process that consists of fast opening and slow locking motion. Starting in the open state 1, the system attempts every \( \sim 1 \text{ ns} \) to close its mouth, but due to frequent \( (0.07 \mu s) \) back transitions requires a MFPT of 0.7 \( \mu s \) to actually accomplish the transition. This relatively fast process is the prerequisite for the subsequent slow locking step \( 2 \rightarrow 3 \), before the system rapidly relaxes to the final state 4.

### 3.8. Mutation Studies

To test our hypothesis that Phe4 is a key residue for the functional dynamics of T4L, we performed additional MD simulations of various mutations of our reference system M6I (see Methods). Apart from Phe4, we chose to also mutate residues Phe67 and Phe104. This is because the side chain of Phe67 forms a steric barrier that Phe4 has to overcome when moving in or out of the hydrophobic pocket. By mutating one of these two side chains, we intended...
to directly modify the main barrier in the opening-closing process. Phe104 was chosen because it was found to show correlated motion with Phe4 and Phe67 and because it exhibits major contributions to the first contact PCA eigenvector. Moreover, it is part of an inner hydrophobic core (consisting of Leu7, Glu11, Ile29, Asp70, Ala71, Ala74, Ile100, Asn101, Val103, and sometimes Asp10, Phe67, or Glu105) and is solvent accessible at the inside of the mouth, probably with the task of properly aligning the substrate. We chose mutations Phe→Ala to change the large phenyl side chain to a small hydrophobic one, as well as Phe→Asn to change from a hydrophobic to a hydrophilic side chain of similar size. In total we carried out six MD runs with a trajectory length of 10 μs each, referred to as F4A, F4N, F67A, F67N, F104A, and F104N.

Showing the time evolution of the radius of gyration and the corresponding free energy profiles of all mutations, Figure 7 reveals that the various mutations may significantly change the energy landscape and the conformational dynamics of the system. As a thermodynamic effect, we find changes of the positions and values of minima and barriers of the free energy. This is due to the reorientation of the side chains in the vicinity of the mutations, which may result, for example, in new hydrophobic contacts (in particular for Ala mutations) or new hydrogen bonds (in particular for Asn mutations). As a consequence, we find that Ala mutations of Phe67 and Phe104 stabilize the closed state, while F67N stabilizes the open state. With respect to the kinetics of the system, we notice that the time scale of the hinge-bending motion may change considerably. For example, mutations of the Phe4 position appear to speed up the hinge-bending motion, while mutations of the Phe67 position rather seem to slow down the process (as far as we can tell from the quite limited statistics of our 10 μs trajectories).

A closer analysis reveals that also the nature of the open–closed transition can be affected by the mutations. As examples, Figure S19 shows for each mutation various two-dimensional free energy landscapes that feature a representative transition event. Particularly in the case of mutation of Phe4, we find clear deviations from the canonical "L"-shape observed in Figure 6. This is in part caused by secondary structure elements such as helix-1, helix-3, or βα, which may loose stability or alignment, such that some native contacts are not formed anymore, while nonnative contacts can form. Nonetheless, the overall hinge-bending motion is found to be quite robust with respect to all mutations considered, that is, the system never gets completely trapped in either the open or the closed state.

We note that our findings are in line with existing experimental results. For example, Remington et al. showed that deletion of Phe4 decreases the catalytic effectiveness of T4L. Mutating Phe67 and Phe104 to Ala, Xu et al. found a slight decrease of the overall stability of the protein (i.e., a decrease of the melting temperature by 5.7 K for F67A and 9.7 K for F104A) and small rearrangements of some side chains of neighboring residues.

4. CONCLUSIONS

We have outlined a general strategy to identify and validate reaction coordinates of protein functional dynamics and applied the approach to the prominent example of the hinge-bending motion of T4L (Figure 1). As a standard dimensionality reduction method, we have employed principal component analysis (PCA) and showed that the applicability and outcome of the method depends crucially on the input coordinates used. Although Cartesian coordinates are commonplace, they may lead to results of minor value, because structural dynamics of flexible molecules necessarily results in a mixing of overall and internal motion. In the case of our 50 μs trajectory of T4L, for example, Cartesian PCA did not yield more information than a standard 1D order parameter such as the radius of gyration (Figure 2). Interestingly, we obtained a similar picture also for dihedral angle PCA, which reflects the fact that for T4L only very few (φ, ω) dihedral angles change significantly upon the open–closed transition. On the other hand, we have identified numerous changes of interresidue contacts that account for this transition in detail. Hence for T4L a PCA on contact distances turned out to be the method of choice, which yields several structurally resolved principal components (PCs) and seven metastable states (Figure 3). Generally speaking, it is always advisable to first screen the raw MD data, in order to identify which internal coordinates change upon the reaction considered.

In favorable cases the first PCs directly constitute a suitable reaction coordinate that reveals the mechanism of the considered process. For example, the first few PCs of a backbone dihedral angle PCA have been shown to give a detailed picture of the conformational dynamics of polyanaline or the folding of a villin headpiece. The construction of a low-dimensional reaction coordinate may be less straightforward, however, when one wants to account for a complex structural rearrangement of a multidomain protein. As a test if a candidate for a reaction coordinate is causally related to the functional dynamics of the system, we have studied the molecule's response to external pulling along the coordinate, using targeted MD (TMD) simulations. Since for T4L the first PC of various PCAs was found to behave similar to the radius of gyration $R_G$ (Figure 2), we first considered a pulling coordinate that mimics $R_G$. Interestingly, these TMD simulations did not at all recover the two-state behavior of T4L observed in unbiased MD simulations, which clearly demonstrates that enforced opening of the mouth does not cause the open-closed transition of T4L (Figure 4). This is because an unbiased MD simulation typically uses minimum energy pathways on the multidimensional free energy surface, whereas pulling along a coordinate mimicking $R_G$ might force...
the system to go over peaks rather than over passes, which
necessarily leads to the desired target state.

On the basis of the most prominent contributions to the first
few contact PCA eigenvectors, we have tested the response of
T4L to pulling along various other choices of the reaction
coordinate. In this way, we have discovered a prominent role
of residue Phe4, which is located at the jaw joint of T4L. It
occurs in two states (Figure 5): In the open conformation, the side
chain of Phe4 is buried inside a hydrophobic cavity which
restricts its motion (locked state), while in the closed
conformation, the side chain is outside of the cavity and
exposed to the solvent (free state). By pulling apart Phe4 and
the residues forming the lock cavity, TMD simulations have been
shown to successfully induce the open-closed transition of
T4L (Figure 4). That is, although this transition is not enforced
per se, it happens as a consequence of enforcing Phe4 to bypass
Phe67. The locking mechanism explains the origin of the main
barrier of the process, which is caused by steric hindrance due
to the large size of the phenyl rings. Moreover, the mechanism
results in an entropic stabilization of both open and closed
states of T4L. In excellent agreement with recent experimental
results in an entropic stabilization of both open and closed
barrier of the process, which is caused by steric hindrance due
Phe67. The locking mechanism explains the origin of the main
scales, and hierarchical couplings for a two-domain protein.

In order to understand the complex nature of the
locking mechanism, we have performed additional MD simulations of
in total six mutants, which were found to considerably modify
the energy landscape and the overall kinetics of T4L (Figure 7).

In extension of the usual two-state picture of the open-close
transition, the above TMD/PCA study suggests a four-state model (Figure 6) which also accounts for the locking transition.
Performing a mean first passage time analysis of the various
steps of the process, we have found a hierarchical coupling of
the fast opening-closing motion and the slow locking transition.
That is, starting in the open state, the system attempts every
~1 ns to close its mouth, but (due to frequent back transitions)
requires hundreds of nanoseconds to accomplish this transition.
This relatively fast process is the prerequisite for the subsequent
slow locking step, which occurs on a microsecond time scale.
While the concept of hierarchical protein energy landscapes has
been discussed for some time,11–13 this study has made a first
attempt to concretely specify the various subprocesses, time
scales, and hierarchical couplings for a two-domain protein.

ASSOCIATED CONTENT

Supporting Information
The Supporting Information is available free of charge on the
ACS Publications website at DOI: 10.1021/acs.jctc.7b00571.

Details of various PCAs and MFPT analyses (Figures S1–S10), additional TMD simulations (Figures S11–S15), SASA (Figure S16), structural details of states 3 and 4 (Figure S17), transition pathways in the four-state landscape (Figure S18), and comparison of mutation studies with free energy landscapes of conPCA and four-
state model of equilibrium simulation (Figure S18) (PDF)

AUTHOR INFORMATION

Corresponding Author
*E-mail: stock@physik.uni-freiburg.de. Tel.: +49 761 2035750.
ORCID
Gerhard Stock: 0000-0002-3302-3044

Funding
This work has been supported by the Deutsche Forschungsge-
meinschaft through Grant STO 247/10-1.

Notes
The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Florian Sittel and Mithun Biswas for numerous
instructive and helpful discussions. Furthermore, we gratefully
acknowledge computing resources provided by the Black Forest
Grid initiative at the University of Freiburg, as well as the
bwGRiD, bwUniCluster, and bwForCluster (RV bw161016)
initiatives of the State of Baden-Württemberg.

REFERENCES

(1) Berendsen, H. J. C. Simulating the Physical World; Cambridge
(2) Onuchic, J. N.; Schulten, Z. L.; Wolynes, P. G. Theory of protein
1997, 48, 545–600.
(3) Dill, K. A.; Chan, H. S. From Levinthal to Pathways to Funnels:
1997, 4, 10–19.
(4) Wales, D. J. Energy Landscapes; Cambridge University Press:
(5) Best, R. B.; Hummer, G. Coordinate-dependent diffusion in
Theory Comput. 2013, 9, 135–146.
(7) Krivov, S. V.; Karplus, M. Hidden complexity of free energy
Construction of the free energy landscape of biomolecules via dihedral
(9) Maisuradze, G. G.; Liwo, A.; Scheraga, H. A. How Adequate are
One- and Two-Dimensional Free Energy Landscapes for Protein
(10) Rohrdanz, M. A.; Zheng, W.; Clementi, C. Discovering
Mountain Passes via Torchlight: Methods for the Definition of
Reaction Coordinates and Pathways in Complex Macromolecular
(11) Noe, F.; Clementi, C. Collective variables for the study of long-
time kinetics from molecular trajectories: theory and methods.
(12) Das, P.; Möll, M.; Stamati, H.; Kavraki, L. E.; Clementi, C.
Low-dimensional, free-energy landscapes of protein-folding reactions
2006, 103, 9885–9890.
(13) Jolliffe, I. T. Principal Component Analysis; Springer: New York,
2002.
(14) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential
425.
(15) Hyvärinen, A.; Karhunen, J.; Oja, E. Independent Component
(16) Lange, O. F.; Grubmüller, H. Generalized Correlation for
1053–1061.
(17) Lange, O. F.; Grubmüller, H. Full correlation analysis of
70, 1294–1312.
Noe, F. Identification of slow molecular order parameters for Markov

This document is the Accepted Manuscript version of a Published Work that appeared in final form in JCTC, copyright © 2017 American Chemical Society after peer review and technical editing by the publisher. To access the final edited and published work follow the linked title.


