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, 79104 Freiburg, Germany

E-mail: stock@physik.uni-freiburg.de

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 50µs 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 life time of ∼10µs. 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.2–4 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.5,6 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,7–9 which leads to erroneous interpretations of the energy landscape.

On this account, a popular approach to construct reaction coordinates is to systematically reduce the dimensionality of the problem by introducing a transformation from high-dimensional MD data \(r = (r_1, \ldots, r_{3N})\) to a low-dimensional collective variable \(x = (x_1, \ldots, x_d)\).10,11 While several nonlinear mappings have been suggested,10,12 it is often convenient to use a linear transformation such as principal component analysis,13,14 or various versions of independent component analysis.15–18 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...

∗To whom correspondence should be addressed
the free energy landscape $\Delta G(x)$, this coordinate may be used to construct a Langevin model\textsuperscript{19} or in various enhanced sampling techniques.\textsuperscript{20} 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 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, e.g., 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 describable 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, e.g., 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} (Fig. 1). The timescale 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{24,25} and various enhanced sampling techniques.\textsuperscript{46–48}

On the basis of a 50 $\mu$s long unbiased MD trajec-

![Figure 1](image-url)
tory 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 life time of ~10 ps. 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 (~100 ns) opening-closing motion and the slow (~10 μs) 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), employing the Amber ff99sb*-ILDN force field and the TIP3P water model. Following earlier work by Hub and de Groot, we considered the M61 mutant of T4L (PDB 150L, chain D; residues 163-164 omitted as they are not resolved in the 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 Bussi 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 \( \mathbf{r} = (r_1, \ldots, r_N)^T \) is described by the covariance matrix

\[
\sigma_{mn} = \langle (r_m - \langle r_m \rangle)(r_n - \langle r_n \rangle) \rangle,
\]

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

\[
x_i = \mathbf{v}^{(i)} \cdot \mathbf{r}.
\]

Considering the first few PCs with highest eigenvalues, we may construct a reaction coordinate \( \mathbf{x} = (x_1, \ldots, x_d)^T \), 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, scale the resulting covariance matrix \( \sigma_{mn} \), and obtain the corresponding PCs via

\[
x_i = \mathbf{v}^{(i)} \cdot \mathbf{r}.
\]

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.\(^{63}\) Alternatively, it has recently been suggested\(^ {71}\) 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 \( (x = \frac{1}{N} \sum_{n=1}^{N} \sin \phi_n, \ y = \frac{1}{N} \sum_{n=1}^{N} \cos \phi_n) \) as obtained by the atan2 function, \( \langle \phi \rangle = \text{atan2}(y, x) \). When we project the data \( \phi \) 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 dPCA+.

### 2.2.2 Contact PCA

Numerous options exist to perform a PCA based on interatomic distances or contacts.\(^ {65-69}\) Following our recent work,\(^ {70}\) 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,

\[
D_{ij} = \min(|r_{i,k} - r_{j,l}|) \leq 4.5 \text{ Å}, \tag{4}
\]

with the indices \( k \) and \( l \) running over all heavy atoms of the selected residue pairs. In addition, we discard contacts between residues less than four residues apart in sequence, effectively omitting short-range contacts as, e.g., 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 Fig. 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,\(^ {72}\) 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).

### 2.3 Targeted Molecular Dynamics (TMD)

To test if some chosen reaction coordinate \( x \) accounts for the mechanism of the open-closed transition in T4L, we studied the molecule’s response to external pulling along \( x \). To this end, we used TMD simulations\(^ {31-33}\) which constrain the velocity in 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 \( x(t_{\text{end}}) = x_{\text{end}} \) by constraining coordinate \( x(t) \) to \( x_c(t) = x_0 + v_c t \) with a constant velocity \( v_c \) via the constraint function

\[
\Phi(x(t)) = (x(t) - x_c(t))^2 = 0. \tag{5}
\]

The constraint is realized via the force

\[
F_c = \lambda \frac{d\Phi(x(t))}{dx} = 2\lambda (x(t) - x_c(t)), \tag{6}
\]

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 \( \sqrt{m_i/m_{av}} \), where \( m_i \) and \( m_{av} \) 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,\(^ {49}\) using the “constraint” mode and a constraint velocity of \( v_c = 0.125 \text{ Å ns}^{-1} \) for pulling whole domains and \( v_c = 0.5 \text{ Å ns}^{-1} \) 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\(^ {1}\)

\[
\Delta G_{\text{TMD}} = \int_{x_0}^{x_{\text{end}}} \frac{dG}{dx} \, dx = \int_{x_0}^{x_{\text{end}}} \langle F_c \rangle_x \, dx \approx \sum_{i=1}^{N} \langle F_c \rangle_{x_i} \Delta x, \tag{7}
\]

where \( \langle F_c \rangle_x \) 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 \( v_c = 0 \)), using only the last 20 ns to evaluate \( \langle F_c \rangle_{x_i} \). 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). Unlike to $\tau$ 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 $n_{01}$ time steps $\delta t$ in initial state 0 before it jumps to state 1, the MFPT is defined as 

$$
\tau_{0\rightarrow 1}^{\text{mfpt}} = \frac{1}{n_{01}} \sum_{l=1}^{n_{01}} l \delta t = \frac{n_{01} + 1}{2} \delta t. \quad (8)
$$

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 $n_{ij}^{(k)}$ 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 $n_{ij} = \sum_{k=1}^{m} n_{ij}^{(k)}$. Hence the MFPT amounts to

$$
\tau_{i\rightarrow j}^{\text{mfpt}} = \frac{1}{n_{ij}} \sum_{k=1}^{m} \sum_{l=1}^{n_{ij}^{(k)}} l \delta t. \quad (9)
$$

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 geometric 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 a appropriate MFPT analysis can be carried out. We used an elliptical shape for the core regions (Fig. S2), which was chosen to match the shape of the minima of the free energy landscape$^{6}$ (Fig. 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$\mu$s MD trajectory. Despite large fluctuations, $R_G(t)$ clearly reveals several prominent open-closed transitions of T4L, e.g., at $t \approx 1, 2, 3$ and 7$\mu$s. On average, we find that the open state exists for about twice the time of the closed state. In agreement with experimental investigations, $^{43}$ we furthermore find that both states exists on a time scale of 5 to 20$\mu$s. Other choices of 1D order parameters, such as 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 Fig. 2a), showing two minima associated with the open and closed states, that are separated by an energy barrier $\Delta G_{\text{RC}} \lesssim k_B T$. Modeling the transition rate by the standard expression$^{73}$

$$
k = \frac{1}{\tau} = k_0 e^{-\Delta G_{\text{RC}}/k_B T}, \quad (10)
$$

it is obvious that a small energy barrier $\Delta G_{\text{RC}} \approx 1 k_B T$ 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. $^{75}$ 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 analysis$^{24,25}$), it is therefore 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 multi-event trajectories

As discussed in the Introduction, dimensionality reduction methods such as principal component analysis$^{13,14}$ (PCA) or independent component analysis$^{15,17,18}$ 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 Fig. 2b is overall quite similar to the evolution of the radius of gyration (Fig. 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 $2k_B T$ (Fig. Fig. 2b). As shown in Fig. S3, however, the free energy curves $\Delta G(x_i)$ of all higher PCs are struc-
tureless 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. 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 ∼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. 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 structure 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. 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.

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 (φ,ψ) backbone dihedral angles has proven quite powerful to model the conformational dynamics of peptides, RNA and small proteins. To obtain a first impression of the conformational distribution P(φ,ψ) of T4L, Fig. 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 loops flips of the residue, the latter are mostly located in the β1-β2 turn and the β2-β3 turn (Fig. S1). Performing dPCA+ using the covariance matrix (see Methods), the first principal components naturally focus on these large-amplitude torsional motions (see Fig. S5). 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). Focusing on the correlated (but not necessarily large-amplitude) motion of the backbone dihedral angles, the time evolution of the resulting first PC x1(t) in Fig. 2c is seen to monitor the overall opening-closing motion of T4L, quite similar to the evolution of the radius of gyration RG(t) in Fig. 2a. Interestingly, though, the transitions of x1(t) occur gradually rather than via jumps as found for RG(t). A closer analysis reveals that this is caused by the cumulative motion of the above mentioned ~20 residues that exhibit a small shift of their (φ,ψ) distribution. The effect can be most clearly seen for the residues of the long helix 3, which is bent by the opening-closing motion (Fig. S6).
Figure 3: Contact PCA of the 50 µs trajectory of T4L. (a) Cluster plot of the conformational distribution along the first three PCs, revealing seven conformational states of the system with major states 1 and 2. (b) Autocorrelation functions of the first seven PCs. (c) Eigenvector components of PC1 reveal the contacts that are formed and broken by the overall hinge-bending motion. (d) Indicated by lines in the closed (left) and open (right) structures of T4L, these contacts highlight important residues that are involved in the functional dynamics. Contacts colored in black lines can be formed at the mouth, red lines indicate contacts at the jaw joint.

However, the barrier of the corresponding free energy curve is rather small (Fig. 2c) and all higher PCs are structureless (Fig. S5), so that dPCA+ 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 (Fig. S7), in particular residues 4, 7, 8 and 104. The following side chain angles, $\chi_{n \geq 2}$, 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 (Fig. 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 (Fig. S7), while all higher PCs are structureless (Fig. S8).

To summarize, Figs. 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 Fig. 1b is expected to require a significant change of the contact network of T4L, which will be studied in the following.

3.4 Contact PCA may identify functionally important residues

As detailed in the Methods Section and illustrated by the contact map in Fig. 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 (Fig. 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 (Fig. S10), which is in line with previous studies.16,17,24,25,45 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 $3 k_B T$ (Fig. 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 (Fig. 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 Fig. 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 "Pac-Man". 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 vari-
ous PCs, Fig. 3b displays the time evolution of the cor-
responding autocorrelation functions with standard de-
viations shown in Fig. S9c. Decaying on a µs timescale,
the first three PCs seem to reflect the slowest motions
of the MD trajectory and are therefore promising can-
didates to construct a reaction coordinate. In what fol-
ows, 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

![Figure 4: TMD simulations show the response of T4L to pulling along (top) a distance coordinate accounting for the radius of gyration $R_G$ and (bottom) coordinate $p$ describing the locking transition of Phe4. Left panels compare the potential of mean force $\Delta G_{\text{TMD}}(x)$ of the TMD simulations (red lines) to the corresponding free energy profile $\Delta G_{\text{MD}}(x)$ of the unbiased simulations (gray lines). Right panels show the evolution of the system while performing the nonequilibrium pull runs, using PC1 and PC2 of contact PCA from unbiased MD simulations to represent the energy landscape of T4L (grayscale). The time evolution is indicated by trajectory points, employing colors from blue (short times) to red (long times).](image)

To test if a chosen reaction coordinate is causally re-
lated 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 $R_G$ (Fig. 2), we first
consider a pulling coordinate that mimics $R_G$. To con-
struct a distance coordinate $x$ that is suited for TMD
simulations, we combine all $C_\alpha$ atoms of the N- and
the C-terminal domain into one group each, respec-
tively, and pull the system along the vector connecting
the centers of mass of the two groups. Starting in the
open state, Fig. 4a compares the resulting potential of
mean force $\Delta G_{\text{TMD}}(x)$ [Eq. (7)] obtained by thermody-
namic integration to the corresponding free energy pro-
file $\Delta G_{\text{MD}}(x)$ of the unbiased MD simulations. While
the latter shows the above discussed double-well poten-
tial corresponding to the open and closed states of T4L,
$\Delta G_{\text{TMD}}(x)$ 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 (Fig. 4b) shows that we pull the system up a free
energy barrier rather than towards the minimum corre-
sponding to the closed state. Similarly, starting in the
open state, we do not manage to close the mouth by
pulling the same coordinate (Fig. 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 $C_\alpha$-atoms
of the mouth residues (20, 21, 22 and 137, 141, 142,
145, respectively). However, this again did not induce
the open-closed transition (Fig. 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 (Fig. 4a), this
does not mean that the desired conformational transi-
tion 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 min-
ima might force the system to go over peaks rather than
over passes of the energy landscape, which not necessar-
ily 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.) Us-
ing PC1 and PC2 of the contact PCA to represent the
energy landscape of T4L, Fig. S13 shows the evolution
of the system after pulling (starting in the open state)
and pushing (starting in the closed state), respectively.
Similarly, Fig. 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
we started the pull runs without undergoing a hinge-
bending transition. This response provides a further in-
dication 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 reac-
tion coordinates from other approaches, (such as committor
analysis) to assess the quality of reaction co-
ordinates.
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, like Arg8 to Glu64 (Fig. 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, Fig. 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. Due to 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

$$p = \frac{d_{60,67} \cdot d_{67,4}}{|d_{60,67}|^2},$$

(11)

where $d_{i,j}$ represents the distance vector between the $C_\alpha$ atoms of Lys60 or Phe67 and the center of mass of the carbon atoms of the Phe4 phenyl group. As illustrated in Fig. 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 $d_{60,67}$ (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, Fig. 2f reveals that $p(t)$ essentially follows the open-closed pattern displayed by the other order parameters such as $R_C$. 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 ($\sim 6 k_B T$). 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), Fig. 4c shows that the resulting potential of mean force $\Delta G_{TMD}(x)$ [Eq. (7)] exhibits two main minima separated by a barrier of $\sim 5 k_B T$, 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 (Fig. 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 (Fig. 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 Fig. 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 retrospective we found several mentions of it. For example, differences in the orientation of the Phe4 side chain were reported by Dixon et al., 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. 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 permutable, 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. Fig. 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 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)] account-
in the locking coordinate, reflecting the many possible conformations of Phe4 in the free state.

In 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 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 Fig. 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.

Table 1: Mean first passage times (in units of µs) of the four-state model of T4L.

<table>
<thead>
<tr>
<th>to\from</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>0.07</td>
<td>0.9</td>
<td>1.2</td>
</tr>
<tr>
<td>2</td>
<td>0.7</td>
<td>-</td>
<td>0.5</td>
<td>0.7</td>
</tr>
<tr>
<td>3</td>
<td>5.4</td>
<td>3.2</td>
<td>-</td>
<td>0.02</td>
</tr>
<tr>
<td>4</td>
<td>6.5</td>
<td>4.8</td>
<td>0.04</td>
<td>-</td>
</tr>
</tbody>
</table>

The free energy landscape ∆G(x, p) indicates that transitions between the main states 1 and 4 require subsequent closing and unlocking of T4L. To elucidate this process, Fig. S18 shows the time evolution of trajectory pieces during and shortly before/after the 1↔4 transitions. As expected from the appearance of the energy landscape, we find that most transitions proceed along the route 1→2→3→4 and back along 4→3→2→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 were 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 Fig. 6b, which provides the basis for the following discussion.

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

Hence we find that the relatively long MFPT of 6.5µs for the overall 1→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 ~1µs to close its mouth, but —due to frequent (0.07µs) back transitions— requires a MFPT of 0.7µs to actually accomplish the transition. This relatively fast process is the prerequisite for the subsequent slow locking step 2→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 Gln105) 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 10µs 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, Fig. 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, e.g., 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 timescale 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, Fig. 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 Fig. 6. This is in part caused by secondary structure elements like helix-1, helix-3 or β₃, which may lose 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, i.e., 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.³⁷ Mutating Phe67 and Phe104 to Ala, Xu et al.⁴⁰ 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 (Fig. 1). As a standard dimensionality reduction method, we have first 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.⁶¹ 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 (Fig. 2). Interestingly, we obtained a similar picture also for dihedral angle PCA, which reflects the fact that for T4L only very few (φ, ψ) 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 (Fig. 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 polyalanine⁶² or the folding of villin headpiece.⁶⁷ 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.¹¹–³³ Since for T4L, the first PC of various PCAs was found to behave similar to the radius of gyration R_g (Fig. 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 (Fig. 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 not necessarily leads to the desired target state.

On the basis of the most prominent contributions

Figure 7: Effects of mutating residues Phe4, Phe67 and Phe104 to either Ala or Asn. Shown are (left) the time evolution of the radius of gyration R_g and (right) the corresponding free energy profiles. Vertical gray lines indicate the positions of free energy minima of the original system shown in Fig. 2a).
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 (Fig. 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 (Fig. 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 data,\textsuperscript{43} we obtain a relatively long life time of $\sim 10$ ns for both states. To verify the significance of the key residues of the 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 (Fig. 7).

In extension of the usual two-state picture of the open-close transition, the above TMD/PCA study suggests a four-state model (Fig. 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 $\sim 1$ 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 timescale. While the concept of hierarchical protein energy landscapes has been discussed for some time,\textsuperscript{21-23} this study has made a first attempt to concretely specify the various subprocesses, timescales and hierarchical couplings for a two-domain protein.

**Acknowledgement** 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 by the bw-GRID, bwUniCluster and bwForCluster (RV bw16I016) initiatives of the State of Baden-Württemberg. This work has been supported by the Deutsche Forschungsgemeinschaft through Grant STO 247/10-1.

**Supporting Information Available:** Details of various PCAs and MFPT analyses (Figs. S1-S10), additional TMD simulations (Figs. S11-S15), SASA (Fig. S16), structural details of states 3 and 4 (Fig. S17), transition pathways in the four-state landscape (Fig. S18) and comparison of mutation studies with free energy landscapes of conPCA and four-state model of equilibrium simulation (Fig. S18). This material is available free of charge via the Internet at http://pubs.acs.org/.

**References**


