Article

Reconstruction of ARNT PAS-B Unfolding Pathways by Steered Molecular Dynamics and Artificial Neural Networks

Stefano Motta, Alessandro Pandini,* Arianna Fornili, and Laura Bonati*

Cite This: *J. Chem. Theory Comput.* 2021, 17, 2080−2089

Read Online

ACCESS | 📊 Metrics & More | 📖 Article Recommendations | 🆂🅸 Supporting Information

ABSTRACT: Several experimental studies indicated that large conformational changes, including partial domain unfolding, have a role in the functional mechanisms of the basic helix loop helix Per/ARNT/SIM (bHLH-PAS) transcription factors. Recently, single-molecule atomic force microscopy (AFM) revealed two distinct pathways for the mechanical unfolding of the ARNT PAS-B. In this work we used steered molecular dynamics simulations to gain new insights into this process at an atomistic level. To reconstruct and classify pathways sampled in multiple simulations, we designed an original approach based on the use of self-organizing maps (SOMs). This led us to identify two types of unfolding pathways for the ARNT PAS-B, which are in good agreement with the AFM findings. Analysis of average forces mapped on the SOM revealed a stable conformation of the PAS-B along one pathway, which represents a possible structural model for the intermediate state detected by AFM. The approach here proposed will facilitate the study of other signal transmission mechanisms involving the folding/unfolding of PAS domains.
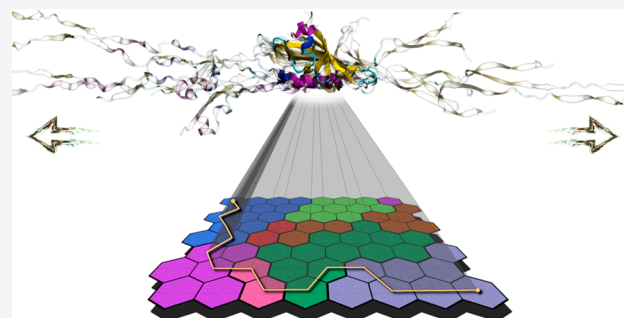
INTRODUCTION

An increasing number of proteins have been reported showing the ability to switch between different fold arrangements. During their life cycle, these proteins can undergo large conformational changes from an ordered state to an alternative one involving secondary structure shifts or exposure of new surfaces. This behavior is often associated with an expanded functional role of the protein that, thanks to the structural changes, can for example modify the interactions with partners.[1,2]

Experimental evidence has suggested that considerable conformational changes, and perhaps even a switch between folded and partially unfolded states of a domain, have a role in the functional mechanisms of the basic helix loop helix Per/ARNT/SIM (bHLH-PAS) family of transcription factors. Members of the bHLH-PAS family have a broad range of functions in developmental and physiological processes, and some are involved in cancer.[3] The bHLH-PAS proteins generally act as heterodimers that consist of a signal-regulated subunit (for example, the aryl hydrocarbon receptor (AhR), involved in toxin metabolism, and the hypoxia-inducible factor-α (HIFα) proteins contributing to maintenance of cellular oxygen homeostasis) and a more ubiquitous subunit (for example, the aryl hydrocarbon receptor nuclear translocator (ARNT), which participates in both AhR and HIFα mechanisms by dimerizing with them). These proteins exhibit a relatively well-conserved N-terminal domain structure, including the bHLH and the PAS regions. The latter contains two structurally conserved domains: PAS-A, critical for

dimerization selectivity, and PAS-B, responsible for sensing diverse exogenous and endogenous signals.

The PAS domain is present in several proteins also outside the bHLH-PAS family and exhibits a high level of plasticity, fundamental for its role of sensor for different signals, including oxygen, ligands, light, and redox potential.[4] A typical PAS fold is composed by a central five-stranded antiparallel β-sheet, a long α-helix, and some shorter helices, which surround a buried internal cavity[5] (Figure 1). An additional N-terminal helix (A$'\alpha$) was observed in some PAS domains.

In most PAS domains, signals apparently propagate to and through the central β-sheet and ultimately toward spatially remote effector domains, where they modulate biological activity. Signal reception and propagation often involve significant fold changes. In the photoactive yellow protein (PYP), a partial unfolding of the PAS domain was observed upon blue-light absorption.[6,7] This perturbation causes the dissociation of the A$'\alpha$ helix from the β-sheet surface followed by a conformational change of A$'\alpha$.[8] A similar mechanism was also reported for the PAS domains of N. crassa Vivid[9] and LOV2.[10] In the sensor histidine kinase (CitA), ligand binding
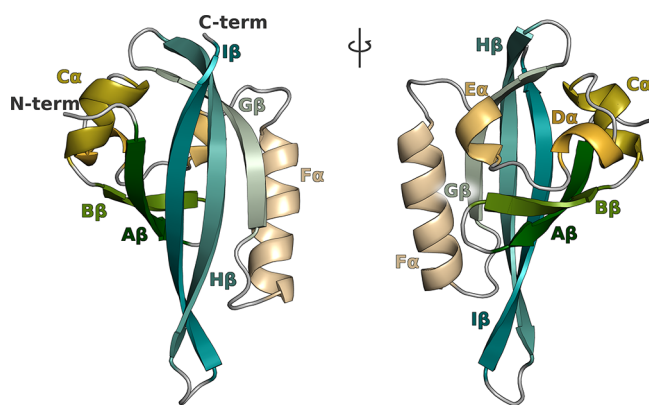
Figure 1. Typical PAS-domain fold is here represented using the three-dimensional structure of the ARNT PAS-B domain (PDBID: 1X0O). Secondary structure elements are a five-stranded antiparallel β-sheet (the N-terminal Aβ, Bβ and the C-terminal Gβ, Hβ, and Iβ) flanked by a long α-helix (Fα, called "helical-connector") and several shorter α-helices (Cα, Dα, and Eα, often called "helical bundle").

was found to induce a considerable contraction of the PAS domain, followed by partial unfolding of an N-terminal helix.[11] These similarities suggest that a common signal transduction mechanism involving changes in the β-sheet is conserved among the PAS proteins, despite the wide range of stimuli they sense. The malleability of the β-sheet of the PAS domains is further confirmed by the β-strand slippage observed for PYP in AFM-based pulling experiments.[6]

Evidence of the flexibility of PAS domains has also been found within the bHLH-PAS family, and substantial conformational changes have been suggested to have a role in the function of proteins from this family. Despite the buried nature of the HIF-2α PAS-B cavity, ligand binding was found to occur with rapid association rates, typical of solvent accessible ligand binding sites rather than internal cavities.[12] The comparable rate constants for compounds with different chemical properties and size suggested a common energetic barrier to overcome, which could be associated with the HIF2α PAS-B conversion from a "closed" structure to a binding-competent "open" state.[12] Similarly, the AhR PAS-B was suggested to have a more open cavity when associated with the HSP90 chaperone protein and undergoing a substantial conformational change upon ligand binding and HSP90 displacement.[13] The hypothesis that the open conformation of the AhR PAS-B displays partially unfolded Aβ and Bβ strands was supported by mutagenesis and coimmunoprecipitation experiments showing that a set of residues in these sheets are involved in both HSP90 association and ligand binding (suggesting that they are exposed in the open state and buried into the binding cavity in the closed one).[14] In addition, other HSP90 clients have been shown to undergo an unfolding of two beta strands to allow chaperone recognition.[15] Particularly relevant in this framework is the experimental evidence regarding ARNT. Several studies evidenced the remarkable flexibility of the ARNT PAS-B β-sheet. In a recent study, a single-site mutant in this β-sheet was shown to populate an alternative conformation with a three-residue register shift in the Iβ,[16] and it was observed that the interconversion between the two states proceeds through a chiefly unfolded transition state.[17] Moreover, single-molecule atomic force microscopy (AFM), used to investigate the mechanical unfolding of the ARNT PAS-B domain, revealed two distinct pathways via a kinetic partitioning mechanism.[18] A

simple two-state pathway was observed for the majority of the unfolding events (67%), whereas in the other 33% of cases a well-defined intermediate state was found in which the C-terminal β-hairpin is detached from the domain. It was suggested that the observed low mechanical stability of the PAS-B domain may help PAS proteins to recruit protein partners and lower the free-energy barrier for the formation of the binding interface.

In this work we use steered molecular dynamics (SMD) simulations to investigate the unfolding process of the ARNT PAS-B domain, with the aim of gaining insight into the involvement of the PAS-domain unfolding in the signal transmission mechanisms of the bHLH-PAS family. This system represents an ideal starting point for our studies given the availability of AFM experiments[18] for a direct validation of computational results.

SMD applies external forces to manipulate biomolecules in order to probe mechanical functions, as well as to accelerate processes that are otherwise too slow to simulate. This method complements single-molecule AFM experiments and provides invaluable insights into mechanical unfolding processes at an atomistic level.[19−22] Multiple SMD replicas can be run to sample different unfolding events.[23−25] The analysis of the resulting trajectories may lead to the discovery of different pathways for the process under study,[26] but the interpretation of the results is not trivial. Often visual inspection of the trajectories may highlight the conformational differences in the sampled states, but there are no standard protocols to detect and classify multiple pathways. In this work, we developed an original approach based on the use of a self-organizing map[27] (SOM) to identify different pathways in the unfolding of the ARNT PAS-B. A SOM is considered a type of artificial neural network, with an explicit visual representation of data on a two-dimensional map, which has been widely used for the analysis of different types of data,[28−30] including protein conformations extracted from MD simulations.[27,31] Here, we applied a SOM-based approach not only to detect the different conformational states observed during the unfolding but also to reconstruct and classify the unfolding pathways sampled by the SMD replicas. Using this strategy, we were able to identify two different types of unfolding pathways for the ARNT PAS-B domain, which are in good agreement with the available AFM data, as well as to indicate a possible atomistic model for the intermediate state revealed by the AFM experiments in one of the pathways.

■ METHODS

Steered MD Simulations. The structure of the ARNT PAS-B domain was downloaded from the Protein Data Bank[32] (PDBID: 1X0O[33]). The structure was preprocessed for simulation with the Schrodinger's Protein Preparation Wizard tool,[34] and residue protonation states were determined by PROPKA[35] at pH = 7.0. The system was then solvated in a triclinic box with TIP3P water molecules (size of the box: 45 × 8 × 8 nm). The size of the box was set to accommodate the extended ARNT PAS-B polypeptide (40 nm long) on the x-dimension. No counterions were added since the system was already neutral. Simulations were run using GROMACS 2018.3[36] with the Amber ff14SB force field.[37] A multistage equilibration protocol, as described in ref 38, was applied to remove unfavorable contacts and provide a reliable starting point for the SMD runs: the system was first subjected to 2000 steps of steepest descent energy minimization, with positional

restraints (2000 kJ mol^{-1} nm^{-2}) on all resolved atoms. Subsequently a 200 ps NVT MD simulation was used to heat the system from 0 to 100 K with restraints lowered to 400 kJ mol^{-1} nm^{-2}, and then the system was heated to 300 K in 400 ps during a NPT simulation with further lowered restraint (200 kJ mol^{-1} nm^{-2}). Finally, the system was equilibrated during an NPT simulation for 2 ns with backbone restraints lowered to 50 kJ mol^{-1} nm^{-2}. In the NVT simulations the temperature was controlled by the Berendsen thermostat[39] with a coupling constant of 0.2 ps, while in the NPT simulations the V-rescale thermostat[40] (coupling constant of 0.1 ps) was used and the pressure was set to 1 bar with the Parrinello−Rahman barostat[41] (coupling constant of 2 ps). A time step of 2.0 fs was used, together with the LINCS[42] algorithm, to constrain all the bonds. The particle mesh Ewald method[43] was used to treat the long-range electrostatic interactions with the cutoff distance set at 12 Å. Short-range repulsive and attractive dispersion interactions were simultaneously described by a Lennard-Jones potential, with a cutoff at 12 Å, applying long-range dispersion corrections for energy and pressure.[44]

With structures properly equilibrated, SMD simulations[22] were performed by harmonically restraining the *x*-component of the distance between the center of mass of the first and last four residues of the protein backbone. A force constant of 500 kJ mol^{-1} nm^{-2} was used, and the equilibrium value of the distance was changed from the initial to the final value at a constant velocity (0.2 nm ns^{-1}). The system was simulated for 200 ns to steer the protein to a fully extended conformation. In order to assess the reproducibility of the unfolding pathways, the SMD simulations were run in 50 replicas, for a total simulation time of 10 μs. The force applied to the harmonic spring was monitored during each replica. To test the dependence of the results from the pulling speed, 10 replicas of 1200 ns at a lower velocity (0.02 nm ns^{-1}) were also performed, for a total simulation time of 12 μs.

H-bonds were computed using Chimera-X[45] software with default values for geometrical parameters.

Per-residue nonbonded energy decomposition analysis was performed using the gmx energy command. Pairwise nonbonded contributions of each residue were then summed up to obtain the cumulative interaction energy of each residue with the remaining ones.

Self-Organizing Maps. A SOM[46,47] is considered an unsupervised artificial neural network where neurons are arranged in a grid map enforcing topological relationships. Multidimensional input data can be effectively visualized in a low-dimensional representation using a SOM. Each neuron is a feature vector with the same dimension of the input data vectors. Training of the map is an iterative process in which the following are true:

1. The map is initialized with random values for the neuron vectors;

2. Input data vectors are assigned to the neuron with closer feature values, also called the best matching unit (BMU);

3. The feature values of the winning neuron and its neighbors are adjusted toward the values in the input vector. The magnitude of the modification decreases with the distance from the BMU and along the training.

The resulting SOM can be interpreted as an approximation of the data space where similar samples are mapped close

together. In this work we used an 8 × 8 sheet-shaped SOM (without periodicity across the boundaries) with hexagonal lattice shape. The input features to train the SOM were selected from the set of pairwise distances among Cβ atoms. Only distances between Cβ atoms closer than 1.0 nm in the native folded conformation were included. This choice was motivated by the evidence that the Cβ contact matrix with 1.0 nm cutoff is within the optimal range for accurate reconstruction of a protein conformation from a set of pairwise atomic distances.[48] In the native structure of ARNT there are 839 Cβ distances below 1.0 nm. This set of distances was used to build the data set of SMD conformations for SOM training. Input conformations for the SOM training were taken every 100 ps from the SMD simulations at 0.2 nm ns^{-1} pulling speed. In a second step, the neurons are further grouped in a small, but representative, number of clusters by agglomerative hierarchical clustering using Euclidean distance and complete linkage. The optimal number of clusters, *N*, was selected based on the Silhouette profiles. In the present case, selecting the clustering with *N* = 8 produced less neurons with negative silhouette scores compared to other clustering methods with comparable average silhouette width. All the analyses were performed in the R statistical environment[49] using the *kohonen* package.[50]

Mapping of Pathways on the SOM. For each SMD simulation, the unfolding pathway was mapped on the SOM by monitoring the BMU of each frame of the simulation and tracking the route covered on the map. The resulting SOM pathways were clustered by agglomerative hierarchical clustering using average linkage. The distance between the SOM pathway of simulation A and that of simulation B is defined as

$$\sum_{i=1}^{n} d(A_i, B_i)$$

where *n* is the number of frames of each simulation and $d(A_i, B_i)$ is the distance between the BMUs of frame *i* in the two simulations. The distance between two BMUs is defined as the mean square deviation of the two BMU vectors.

SOM can also provide a higher-level representation of the dynamics in the form of a state graph, where important steps in the process are represented by discrete nodes connected according to transition probabilities. To build this graph, an approximate transition matrix between neurons was estimated from the counts of transitions between the starting neuron A and the ending neuron B in all the simulations and repeated for all the possible combinations of A and B. The matrix was then transformed into a row stochastic matrix (dividing each element by the sum of the row). A graph was then built with nodes represented by neurons and edges set to the negative logarithm of the transition probability between the corresponding neurons. The distance between two nodes in the graph was calculated for their shortest path. The distance value was calculated as the negative logarithm of the product of the pairwise transition probabilities between neurons along the path.

Simulations run at 0.02 nm ns^{-1} pulling speed were retraced on the SOM trained with conformations at 0.2 nm ns^{-1} pulling speed, where each frame of the simulations was assigned to the closest neuron. A single SOM was used to analyze both sets of trajectories for an easier comparison and to avoid changes in map topology. We verified that the SOM derived from the

highest speed simulations is consistent with the low-speed ones by calculating the distances between the two sets of frames and the closest neurons (Figure S1). Comparable distance distributions were observed for the two sets of simulations, indicating that the low-speed frames are well represented by the SOM even if they were not used for training. All the analyses were performed in the R statistical environment[49] using the *igraph* package.[51]

RESULTS

Mapping Conformations Sampled by Steered MD on SOM. To study the unfolding pathway of ARNT PAS-B under mechanical forces, we performed multiple replicas of SMD simulations using the backbone atoms of the last four residues at the N-terminal and C-terminal ends as pulling groups (see Methods). The simulations were run until the protein was extended at 40 nm, consistent with the extension of the fully elongated polypeptide.[18]

To reconstruct the possible paths of unfolding sampled by different replicas, we used a self-organizing map (SOM), a specific architecture of artificial neural networks, consisting of a grid of neurons (hexagons of the map). The map was trained using protein conformations from the steered MD simulations. After training, each input conformation is assigned to a single neuron so that similar conformations are represented by the same neuron, and similar neurons are close to each other. In this work, a square 8 × 8 SOM without periodic boundaries conditions was used with the distances among $C\beta$ atoms as neuron features (see Methods). Different SOM parameters (SOM shape and neuron features) were tested to obtain an efficient grouping of the protein conformations on the map (data not shown). Figure 2 represents the three-dimensional structures of the centrotypes of each neuron (conformation closest to the neuron vector).

In the trained map, the completely folded and completely unfolded conformations (first and last frames of the SMD simulations) are at the top left and bottom right corners of the map, respectively. Intermediate conformations along the unfolding path populate the other regions of the map.
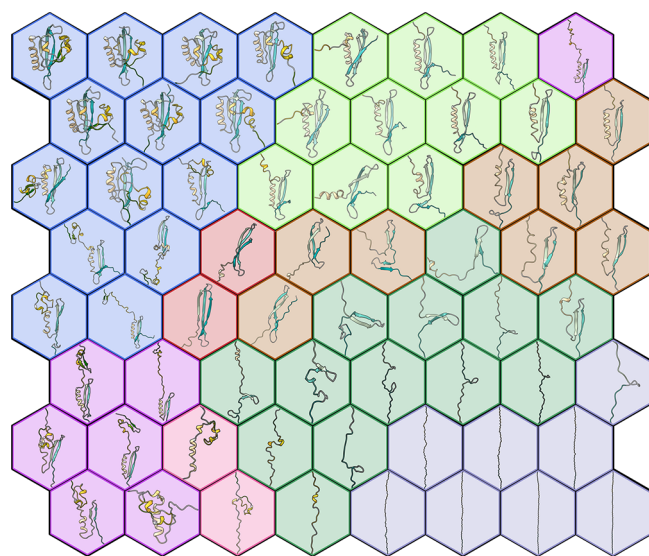

Figure 2. 3D structures of the neuron centrotypes represented on the SOM. Colors assigned to neurons refer to cluster analysis that is treated later in the text.

To evaluate the quality of the method in separating different conformations, we compared the distance root mean square deviation (dRMSD) values calculated between structures from the same neuron (blue in Figure 3) with the values calculated between structures belonging to neighboring neurons (red). The intraneuron differences are significantly smaller, indicating an effective segregation of similar conformations in each neuron.

To obtain a more "coarse-grained" view of the process and highlight putative macrostates sampled during the unfolding, the neurons were further grouped by agglomerative hierarchical clustering (clusters A−H in Figure 4). An optimal number of eight clusters was selected based on the analysis of the silhouette profiles (Figure S3) and the visual inspection of the resulting clusters.

A diagram of the secondary structure composition of frames within each neuron is reported in Figure S4. The average composition in secondary structure and the radius of gyration for each cluster are reported in Table S1.

In addition to canonical descriptors of secondary structures, annotations using structural alphabets are particularly effective in capturing local conformations and their dynamical changes.[52] Local structures from the ensembles in each neuron were encoded with a structural alphabet (SA),[53] and a per-fragment profile of divergence from the distribution of local conformations in the folded state was calculated (see the Supporting Information). These profiles were mapped on each of the associated SOM neurons (Figure S5). The analysis highlighted local patterns of unfolding in intermediate neurons; in particular, clusters B and E show changes in conformational dynamics mainly at the N-terminal fragments of the protein, and cluster C does so more frequently at the C-terminal region of the protein.

SOM Highlights Different Unfolding Pathways. In order to analyze the unfolding pathways followed during the simulations, we mapped each trajectory onto the SOM by tracking the position of each frame on the map. All simulations started from the upper left corner and ended in the lower right corner, but we classified the simulations into two distinct types of pathways: pathway 1, visiting neurons in the lower left corner of the map (left in Figure 5), and pathway 2, going through neurons in the top right corner of the map (right in Figure 5).

The mapping of each trajectory replica is shown in Figures S6 and S7. Pathway 2 was the most probable and was sampled in 74% of the replicas.

The above classification was assessed performing hierarchical clustering of the simulation pathways (using average linkage, as detailed in Methods) and analyzing the resulting dendrogram (Figure S8). This diagram mostly agrees with our classification except for the identification of two additional minor pathways (pathway 1b for replicas 1, 26, 42 and pathway 2b for replicas 8, 35, 39) that are grouped in different small branches of the dendrogram. We grouped replicas of pathway 1b to pathway 1, and replicas of pathway 2b to pathway 2, for their similarity to the characteristics of each pathway in the first part of the unfolding process.

Simulations following either pathway 1 or 2 visit most of the eight clusters on the SOM. Two clusters remain exclusive and differentiate the type of path: cluster B was only sampled by simulations following pathway 2, while cluster C was only sampled by simulations following pathway 1. Interestingly also cluster E was only sampled by replicas 8, 35, and 39 and cluster
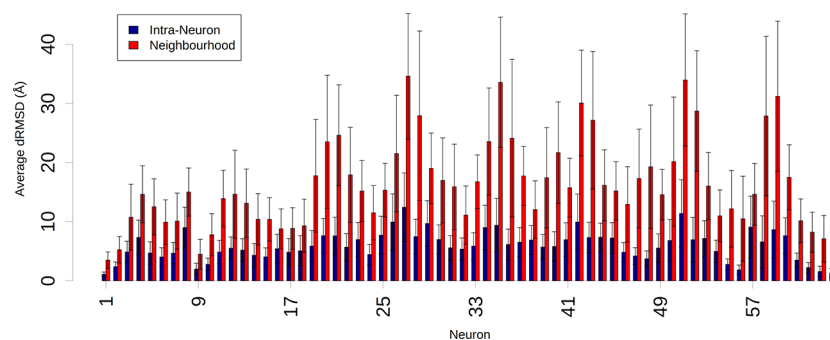
Figure 3. Distribution of dRMSD values calculated between structures from the same neuron (blue) and between structures assigned to neighboring neurons (red). The neurons numbering is shown in Figure S2.
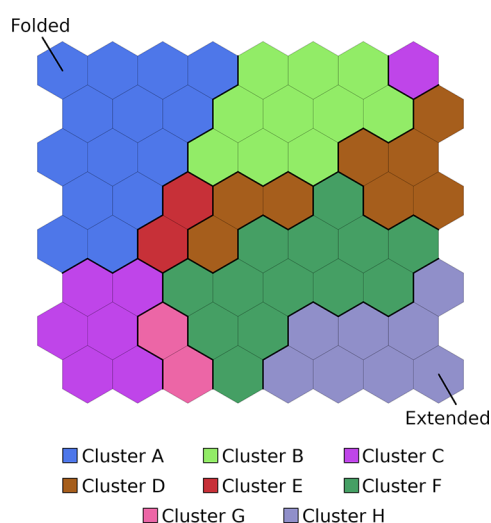

Figure 4. Self-organizing map. Conformations sampled in the different replicas are first assigned to different neurons (tiles of the map), and then neurons are further grouped by hierarchical clustering (colors of the tiles). See Methods for a detailed description.
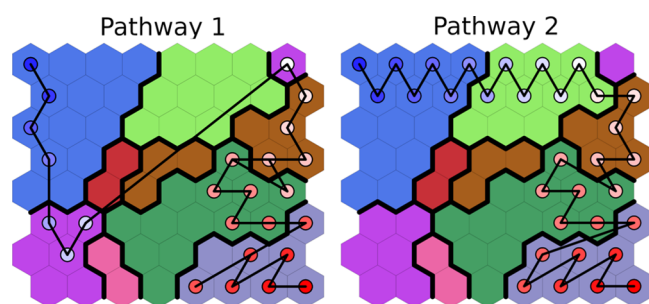

Figure 5. Tracing of two replicas representative of pathways 1 and 2 on the SOM.

Figure 6. Graph representation of the transition matrix for the SMD unfolding of the PAS-B domain. Nodes are colored according to the cluster colors of Figure 2.

(see the Methods section). The two unfolding pathways are clearly visible in this representation.

Analysis of the structural changes for the different pathways indicates that simulations following pathway 1 start with the detachment of the N-terminal region (Aβ, Bβ, and helical bundle) from the rest of the protein (a in Figure 7). This region, however, preserves its internal contacts and does not completely unfold until the C-terminal Iβ is completely unfolded (b). Then the simulations continue with the unfolding of either the N-terminal region (most of pathway 1 replicas) or the C-terminal region (only in replicas 1, 26, and 42 that visit cluster F). Pathway 2 differs from pathway 1 in the order of events leading to the Iβ unfolding: while in pathway 1 the N-terminal region detaches from the core of the protein but remains "folded", in pathway 2 the N-terminal region completely unfolds (f in Figure 7) before the Iβ (g).

The different timing for the unfolding of Iβ is also evident from the time evolution of protein backbone H-bonds: Figure S9 reports the extension of the protein corresponding to the last frame where each H-bond was detected in the different replicas. In simulations following pathway 1, the H-bonds

G by replicas 1, 26, and 42. The transition of these replicas through the E or G clusters explains their assignment to separate branches of the dendrogram in the hierarchical clustering of the simulation pathways.

From pairwise transitions between neurons, we reconstructed an approximate transition matrix. This matrix was then visualized as a graph in which nodes represent the SOM neurons, and the edges are weighted by the negative logarithm of the transition probability between pairs of neurons (Figure 6). The sum of edge weights along a pathway is proportional to the logarithm of the combined probability along that pathway
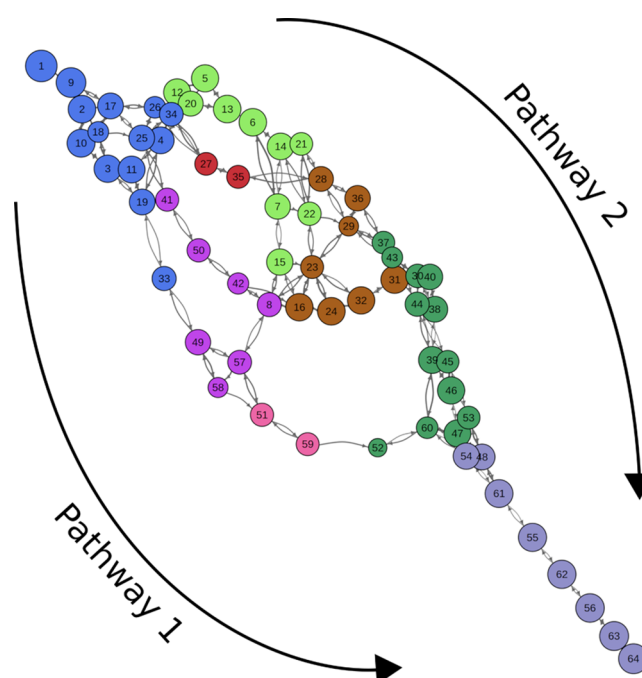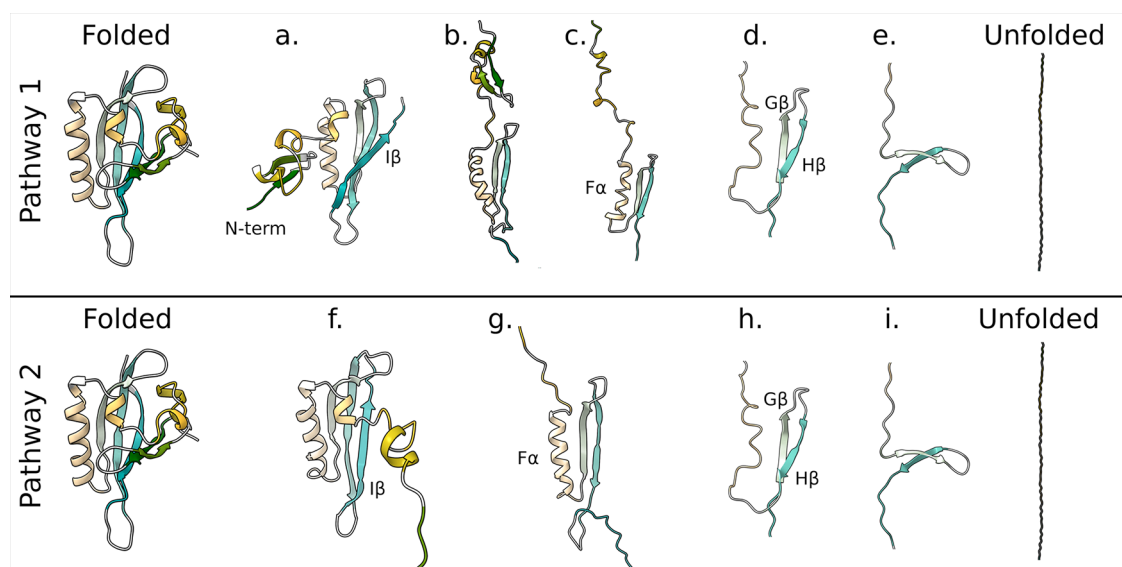
Figure 7. Three-dimensional representation of unfolding pathways. Pathway 1 starts with the detachment of the N-terminal region (a), then unfolding of Iβ (b), elongation of the N-terminal region (c), unfolding of Fα (d), and finally unfolding of Gβ and Hβ (e). Pathway 2 starts with detachment and unfolding of the N-terminal region (f) before the unfolding of the Iβ (g). Then the pathway proceeds as in the case of pathway 1 (h and i). Secondary structure colors are consistent with Figure 1.

between Hβ and Iβ (dark blue) break in the first part of the simulation (before the extension of the protein reaches 10.5 nm), while H-bonds in the N-terminal part (Cα (dark green) and bundle-Bβ (orange)) break at a later time. An opposite behavior is observed in pathway 2 simulations, where the N-terminal H-bonds break in the first part of the simulation and the Hβ-Iβ ones break in the second part. This is consistent with the AFM results that showed an intermediate with Iβ unfolded before 10.5 nm in one-third of the experiments.[18]

To assess the presence of an unfolding barrier along one of the two pathways, we mapped the SMD forces on the SOM by calculating the average force over the frames belonging to each neuron (Figure 8). Most of the frames belonging to cluster C (sampled only by pathway 1 trajectories) are associated with higher forces compared to the adjacent clusters, which could

Figure 8. Average SMD forces mapped on the SOM. Neuron 41, discussed in text, is labeled. Cluster boundaries are highlighted for cluster B (green) and C (magenta).

be interpreted as the presence of a force peak consistent across different replicas and suggests the presence of an intermediate state just before those frames. This observation led us to consider frames belonging to neuron 41 (within cluster C and often visited before the high-force frames) as an intermediate state along pathway 1. From the analysis of the accessible pathways on the graph representation in Figure 6, it can be noted that neuron 41 is often visited along pathway 1. The only alternative route along pathway 1 is through neuron 49 that however shares fundamental features of neuron 41, like the folding of the helical bundle. The frames in neuron 41 have a length extension (9.4 ± 1.0 nm) consistent with that of the intermediate observed in AFM experiments (10.5 nm). On the contrary, frames belonging to cluster B (exclusive of pathway 2) are associated with low forces, without clear peaks along the pathway, which do not suggest the presence of a stable intermediate conformation along the path.

Conformations belonging to Neuron 41 are highly structured, with the PAS domain split into two lobes: the N-terminal lobe composed of Aβ, Bβ, Cα, and Dα and the C-terminal lobe composed of Fα, Gβ, and Hβ (Figure 9).

Surprisingly, each lobe remained well folded preserving almost all the backbone H-bonds. The N-terminal lobe, in particular, showed great resistance to the unfolding process despite the lower number of native hydrogen bond interactions that stabilize its fold (10 compared to the 22 found in the C-terminal β-strands). Indeed, even more surprising are the simulations passing through cluster F, where the N-terminal lobe remains folded until the complete unfolding of the C-terminal lobe. With the aim of detecting the residues that mostly stabilize the N-terminal lobe in this type of conformation, we performed a per-residue energy decomposition analysis on conformations belonging to neuron 41 (Figure S10). This analysis highlighted the residues that give the most stable interactions: a group of hydrophobic residues (F363, F373, Y386, L391) in the core of the protein and two pairs of residues involved in electrostatic interactions (Glu362-Arg379 and Glu390-Gln387). Together, the above interactions
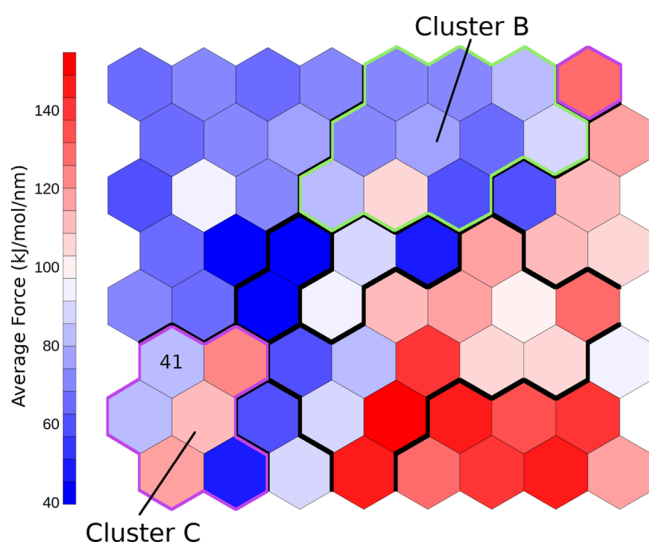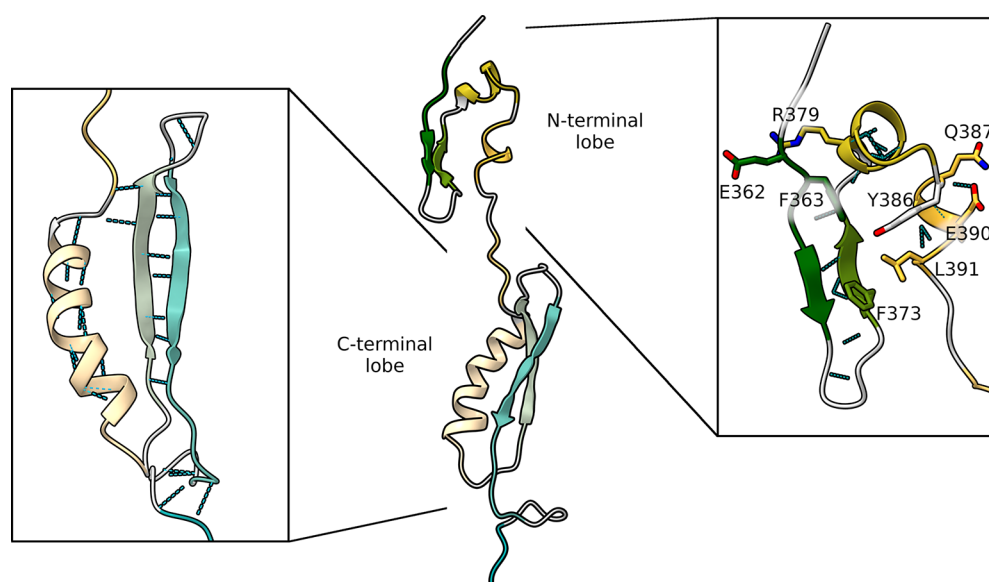
Figure 9. Three-dimensional representation of the putative intermediate state during pathway 1 unfolding. The two insets show the backbone hydrogen bonds in the C-terminal lobe (left) and the group of residues that mostly contribute to stabilizing the N-terminal lobe (right). Secondary structures are colored according to Figure 1.
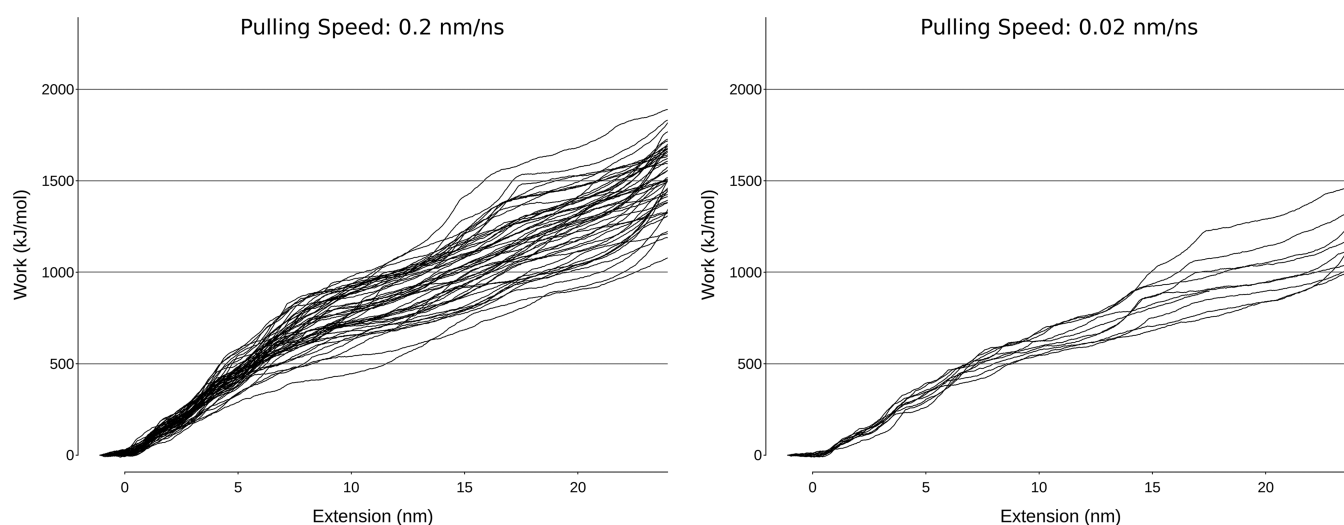

Figure 10. Work profiles for replicas at higher (0.2 nm ns^{-1} on the left) and lower (0.02 nm ns^{-1} on the right) pulling speeds.

hinder the unfolding process of the N-terminal lobe (Figure 9, right inset).

Retracing of Low Pulling Speed SMD on the Trained SOM Confirmed the Two Pathways. To verify the effect of the SMD pulling speed on the conformations sampled in the unfolding process, 10 replicas at lower pulling speed were run for 1200 ns, up to an extension of about 24 nm. No alternative pathway was detected aside from the ones already recorded at a faster pulling speed. While these simulations did not cover the full unfolding for time limitations, they were long enough to compare the initial stages of the unfolding in the two sets of replicas.

The work performed in the 0−24 nm extension range (the region sampled at both pulling speeds and where most of the unfolding process takes place) is reported as a function of the extension length in Figure 10.

Despite the difference in pulling speed, the total work in the low-speed simulations (1000−1500 kJ/mol) falls within the range of the higher speed ones (1000−2000 kJ/mol),

indicating that the pulling speed of 0.2 nm ns^{-1} is still able to capture lower energy pathways.

Mapping the low-speed simulations over the SOM (Figure 11) shows that they are very similar to the initial part of the higher speed simulations (Figures S4 and S5). In particular, six low-speed simulations visit cluster C (pathway 1), while the other 4 go through cluster B (pathway 2).

At a lower pulling speed we found a relative difference in the frequency of sampling of the two pathways, but this is expected due to the limited sample (10 replicas). These data provide limited statistics. However, further exploration of the process at low speed would require significant computational resources, and it is beyond the scope of the present work.

■ DISCUSSION

In this work we present a novel computational approach to analyze large conformational changes in SMD simulations
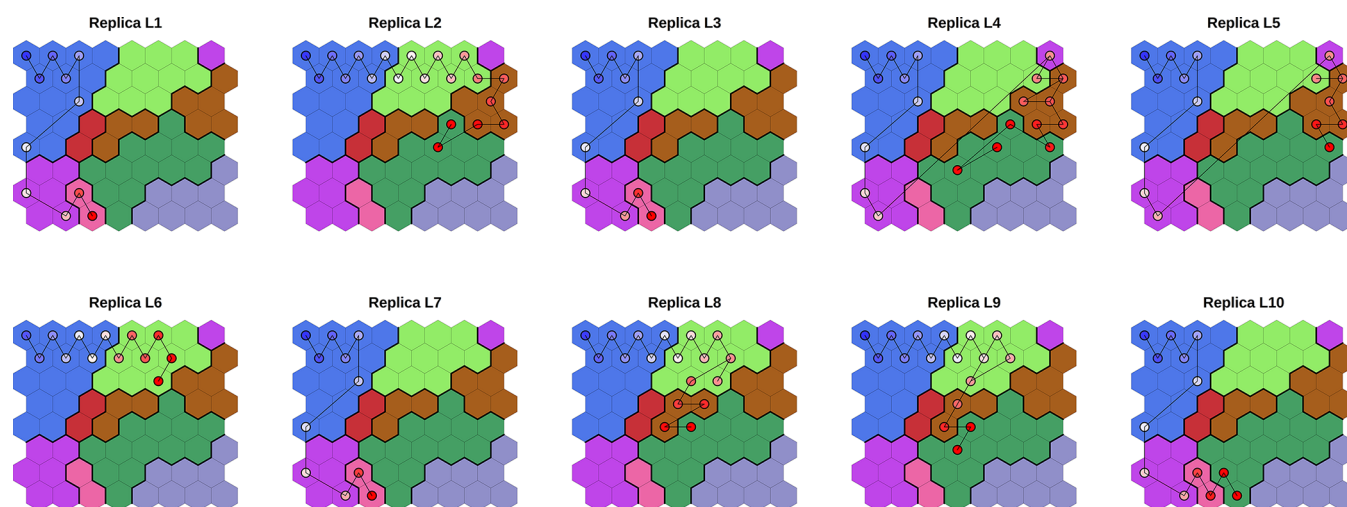
Figure 11. Evolution of low pulling speed (0.02 nm ns^{-1}) SMD simulations, plotted over the 0.2 nm ns^{-1} SOM. For each replica, frames were assigned to the nearest neuron and represented as circles colored in a blue−white−red color scale according to their time in the simulation. Consecutive frames are connected by solid lines.

using SOMs. The approach is applied to the characterization of the mechanical unfolding of the ARNT PAS-B domain.

In SMD, the use of multiple replicas is crucial to sample all the possible unfolding routes. However, this usually leads to a large amount of complex information, which is difficult to translate into a unified and simple representation of the process. A strategy that is often used to extract information from MD trajectories is cluster analysis, which generates a reduced set of nonredundant structures that are representative of the main features of the simulations. A variety of clustering approaches exist, many of which are optimized for the analysis of conformational ensembles of proteins.[27,54,55] Popular choices include hierarchical clustering, linkage, and k-means.[54] Among the different solutions, artificial neural networks have emerged as particularly successful for many applications in bioinformatics, chemometrics, and computational chemistry.[27] SOMs, in particular, are a powerful data-analysis method that combines the advantages of an adaptive learning process with the ability to produce a topological mapping.

In this work, SOMs were used to obtain a geometrical clustering of the ARNT PAS-B conformations. The use of SOM bidimensional visualization facilitates the identification of major conformational states and unfolding pathways.

The use of Cβ atom distances instead of Cartesian coordinates as descriptors of protein conformations made the calculation of the dissimilarity matrix superposition-independent. This removed the effect of structural alignment errors, which can be particularly evident in unfolding simulations due to the large conformational changes. The reference set for superposition was limited to the Cβ distances within 1.0 nm in the native conformation with the advantage that the similarity measure was driven by detecting nativelike contacts and interactions. Finally, the use of a 2D map without periodic boundary conditions automatically segregated the end points of the process (folded and unfolded states) in two corners of the map. These features enhanced the readability of the map, and the final SOM well represented the most important conformational states.

Taking advantage of the topological mapping of the SOMs, in this work we introduced for the first time the idea of tracing

the pathways followed by different SMD replicas on the map to obtain an immediate visualization of differences among the sampled pathways. This approach led us to identify two groups of pathways that undergo unfolding of secondary structures in a different order. Pathway 1 starts with a rigid detachment of the N-terminal region, followed by unfolding of the Iβ and subsequently of the N-terminal region, while in pathway 2 the whole N-terminal region unfolds at first, followed by Iβ.

Pulling forces associated with each frame of the simulations are informative of the process, but detecting patterns in the force profiles from the different replicas is difficult due to the different stages in which they reach peak forces. Through analysis of average forces mapped on the SOM, we were able to highlight conformations along pathway 1 with maximum forces, while neurons along pathway 2 are all characterized by lower forces. This suggested the presence of a stable intermediate conformation along pathway 1.

Our findings are in agreement with previous AFM experiments that indicate the existence of two unfolding pathways, with one of them characterized by an intermediate state at about 10.5 nm extension.[18] Not only did we find a probability distribution for the two pathways (24% pathway 1 vs 76% pathway 2) comparable to the experimental one (33% pathway 1 vs 67% pathway 2), but also the intermediate state identified on neuron 41 has the features highlighted by the experiments, i.e., an extension of about 10 nm and an unfolded Iβ region.[18] The description of the intermediate state, emerging from our analysis, complemented or added to the experimental findings by offering a possible structural model.

It has been suggested that other proteins in the bHLH-PAS family, acting as transcription factors in key developmental and physiological processes, may exert their functions through major conformational changes within their PAS domains.[12,13] However, for these proteins no experimental data on the unfolding process and no evidence on the structure of their putative partially unfolded states have been obtained. The approach proposed in this work for simulating the unfolding processes of PAS domains and analyzing the metastable states along the different pathways may contribute to shed light into the role of PAS-domain folding/unfolding events in the signal transmission mechanisms of the bHLH-PAS proteins.

■ ASSOCIATED CONTENT

ⓢ Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.0c01308.

Analysis of local conformational changes with a Structural Alphabet; Boxplot reporting distances between each frame of simulations and the closest neuron on the SOM; SOM with neuron numbering labeled on each unit; choice of the best number of clusters; secondary structure composition for each neuron; local backbone conformational changes according to Structural Alphabet encoding; trace paths on the trained SOM for simulations following pathway 1 and 2; dendrogram of hierarchical clustering of the pathways followed by different replicas; boxplots reporting the extension of the protein corresponding to the last frame where each native H-bond was recorded in the different replicas; per-residue nonbonded energy from decomposition analysis; average percentage of secondary structure type in structures from each cluster and average radius of gyration for each cluster (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Alessandro Pandini − *Department of Computer Science, Brunel University London, Uxbridge UB8 3PH, United Kingdom; The Thomas Young Centre for Theory and Simulation of Materials, London SW7 2AZ, United Kingdom*; Email: alessandro.pandini@brunel.ac.uk

Laura Bonati − *Department of Earth and Environmental Sciences, University of Milano-Bicocca, Milan 20126, Italy;* ⓞ orcid.org/0000-0003-3028-0368; Email: laura.bonati@unimib.it

Authors

Stefano Motta − *Department of Earth and Environmental Sciences, University of Milano-Bicocca, Milan 20126, Italy;* ⓞ orcid.org/0000-0002-0812-6834

Arianna Fornili − *School of Biological and Chemical Sciences, Queen Mary University of London, London E1 4NS, United Kingdom; The Thomas Young Centre for Theory and Simulation of Materials, London SW7 2AZ, United Kingdom;* ⓞ orcid.org/0000-0001-6554-3979

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jctc.0c01308

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

■ ABBREVIATIONS

bHLH-PAS: basic helix loop helix Per/ARNT/SIM
AFM: atomic force microscopy
SMD: steered molecular dynamics
SOM: self-organizing map
AhR: aryl hydrocarbon receptor
HIFα: hypoxia-inducible factor-α
ARNT: aryl hydrocarbon receptor nuclear translocator
PYP: photoactive yellow protein
BMU: best matching unit
dRMSD: distance root mean square deviation

■ REFERENCES

(1) Luo, X.; Yu, H. Protein Metamorphosis: The Two-State Behavior of Mad2. *Structure* **2008**, *16*, 1616−1625.

(2) Tuinstra, R.; Peterson, F.; Kutlesa, S.; Elgin, E. S.; Kron, M.; Volkman, B. Interconversion between Two Unrelated Protein Folds in the Lymphotactin Native State. *Chemtracts* **2008**, *21*, 94−95.

(3) Kolonko, M.; Greb-Markiewicz, B. BHLH−PAS Proteins: Their Structure and Intrinsic Disorder. *Int. J. Mol. Sci.* **2019**, *20*, 3653.

(4) Taylor, B. L.; Zhulin, I. B. PAS Domains: Internal Sensors of Oxygen, Redox Potential, and Light. *Microbiol. Mol. Biol. Rev.* **1999**, *63*, 479−506.

(5) Kewley, R. J.; Whitelaw, M. L.; Chapman-Smith, A. The Mammalian Basic Helix-Loop-Helix/PAS Family of Transcriptional Regulators. *Int. J. Biochem. Cell Biol.* **2004**, *36*, 189−204.

(6) Zhao, J. M.; Lee, H.; Nome, R. A.; Majid, S.; Scherer, N. F.; Hoff, W. D. Single-Molecule Detection of Structural Changes during Per-Arnt-Sim (PAS) Domain Activation. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 11561−11566.

(7) Blankenburg, L.; Schroeder, L.; Habenstein, F.; Błasiak, B.; Kottke, T.; Bredenbeck, J. Following Local Light-Induced Structure Changes and Dynamics of the Photoreceptor PYP with the Thiocyanate IR Label. *Phys. Chem. Chem. Phys.* **2019**, *21*, 6622−6634.

(8) Rajagopal, S.; Anderson, S.; Srajer, V.; Schmidt, M.; Pahl, R.; Moffat, K. A Structural Pathway for Signaling in the E46Q Mutant of Photoactive Yellow Protein. *Structure* **2005**, *13*, 55−63.

(9) Zoltowski, B. D.; Schwerdtfeger, C.; Widom, J.; Loros, J. J.; Bilwes, A. M.; Dunlap, J. C.; Crane, B. R. Conformational Switching in the Fungal Light Sensor Vivid. *Science* **2007**, *316*, 1054−1057.

(10) Harper, S. M.; Neil, L. C.; Gardner, K. H. Structural Basis of a Phototropin Light Switch. *Science* **2003**, *301*, 1541−1544.

(11) Sevvana, M.; Vijayan, V.; Zweckstetter, M.; Reinelt, S.; Madden, D. R.; Herbst-Irmer, R.; Sheldrick, G. M.; Bott, M.; Griesinger, C.; Becker, S. A Ligand-Induced Switch in the Periplasmic Domain of Sensor Histidine Kinase CitA. *J. Mol. Biol.* **2008**, *377*, 512−523.

(12) Key, J.; Scheuermann, T. H.; Anderson, P. C.; Daggett, V.; Gardner, K. H. Principles of Ligand Binding within a Completely Buried Cavity in HIF2α PAS-B. *J. Am. Chem. Soc.* **2009**, *131*, 17647−17654.

(13) Soshilov, A. A.; Motta, S.; Bonati, L.; Denison, M. S. Transitional States in Ligand-Dependent Transformation of the Aryl Hydrocarbon Receptor into Its DNA-Binding Form. *Int. J. Mol. Sci.* **2020**, *21*, 2474.

(14) Soshilov, A. A.; Denison, M. S. Ligand Displaces Heat Shock Protein 90 from Overlapping Binding Sites within the Aryl Hydrocarbon Receptor Ligandbinding Domain. *J. Biol. Chem.* **2011**, *286*, 35275−35282.

(15) Verba, K. A.; Wang, R. Y.-R.; Arakawa, A.; Liu, Y.; Shirouzu, M.; Yokoyama, S.; Agard, D. A. Atomic Structure of Hsp90-Cdc37-Cdk4 Reveals That Hsp90 Traps and Stabilizes an Unfolded Kinase. *Science* **2016**, *352*, 1542−154.

(16) Evans, M. R.; Card, P. B.; Gardner, K. H. ARNT PAS-B Has a Fragile Native State Structure with an Alternative β-Sheet Register Nearby in Sequence Space. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 2617−2622.

(17) Evans, M. R.; Gardner, K. H. Slow Transition between Two *β*-Strand Registers Is Dictated by Protein Unfolding. *J. Am. Chem. Soc.* **2009**, *131*, 11306−11307.

(18) Gao, X.; Qin, M.; Yin, P.; Liang, J.; Wang, J.; Cao, Y.; Wang, W. Single-Molecule Experiments Reveal the Flexibility of a Per-ARNT-Sim Domain and the Kinetic Partitioning in the Unfolding Pathway under Force. *Biophys. J.* **2012**, *102*, 2149−2157.

(19) Genchev, G. Z.; Källberg, M.; Gürsoy, G.; Mittal, A.; Dubey, L.; Perisic, O.; Feng, G.; Langlois, R.; Lu, H. Mechanical Signaling on the Single Protein Level Studied Using Steered Molecular Dynamics. *Cell Biochem. Biophys.* **2009**, *55*, 141−152.

(20) Sotomayor, M.; Schulten, K. Single-Molecule Experiments in Vitro and in Silico. *Science (Washington, DC, U. S.)* **2007**, *316*, 1144−1148.

(21) Paci, E.; Karplus, M. Unfolding Proteins by External Forces and Temperature: The Importance of Topology and Energetics. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 6521−6526.

(22) Isralewitz, B.; Gao, M.; Schulten, K. Steered Molecular Dynamics and Mechanical Functions of Proteins. *Curr. Opin. Struct. Biol.* **2001**, *11*, 224−230.

(23) Ozer, G.; Valeev, E. F.; Quirk, S.; Hernandez, R. Adaptive Steered Molecular Dynamics of the Long-Distance Unfolding of Neuropeptide Y. *J. Chem. Theory Comput.* **2010**, *6*, 3026−3038.

(24) Lee, W.; Zeng, X.; Zhou, H. X.; Bennett, V.; Yang, W.; Marszalek, P. E. Full Reconstruction of a Vectorial Protein Folding Pathway by Atomic Force Microscopy and Molecular Dynamics Simulations. *J. Biol. Chem.* **2010**, *285*, 38167−38172.

(25) Milles, L. F.; Schulten, K.; Gaub, H. E.; Bernardi, R. C. Molecular Mechanism of Extreme Mechanostability in a Pathogen Adhesin. *Science (Washington, DC, U. S.)* **2018**, *359*, 1527−1533.

(26) He, C.; Genchev, G. Z.; Lu, H.; Li, H. Mechanically Untying a Protein Slipknot: Multiple Pathways Revealed by Force Spectroscopy and Steered Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **2012**, *134*, 10428−1043.

(27) Pandini, A.; Fraccalvieri, D.; Bonati, L. Artificial Neural Networks for Efficient Clustering of Conformational Ensembles and Their Potential for Medicinal Chemistry. *Curr. Top. Med. Chem.* **2013**, *13*, 642−651.

(28) Miljković, D. Brief Review of Self-Organizing Maps. MIPRO 2017; pp 1061−1066.

(29) Oja, M.; Kaski, S.; Kohonen, T. Bibliography of Self-Organizing Map (SOM) Papers: 1998−2001 Addendum. *Neural Comput. Surv.* **2002**, 1−156.

(30) Kaski, S.; Kangas, J.; Kohonen, T. Bibliography of Self-Organizing Map (SOM) Papers: 1981−1997. *Neural Comput. Surv.* **1998**, *1*, 1−176.

(31) Fraccalvieri, D.; Pandini, A.; Stella, F.; Bonati, L. Conformational and Functional Analysis of Molecular Dynamics Trajectories by Self-Organising Maps. *BMC Bioinf.* **2011**, *12*, 158.

(32) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(33) Card, P. B.; Erbel, P. J. a; Gardner, K. H. Structural Basis of ARNT PAS-B Dimerization: Use of a Common Beta-Sheet Interface for Hetero- and Homodimerization. *J. Mol. Biol.* **2005**, *353*, 664−677.

(34) Madhavi Sastry, G.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and Ligand Preparation: Parameters, Protocols, and Influence on Virtual Screening Enrichments. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 221−234.

(35) Bas, D. C.; Rogers, D. M.; Jensen, J. H. Very Fast Prediction and Rationalization of PKa Values for Protein-Ligand Complexes. *Proteins: Struct., Funct., Genet.* **2008**, *73*, 765−78.

(36) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1−2*, 19−25.

(37) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696−3713.

(38) Motta, S.; Minici, C.; Corrada, D.; Bonati, L.; Pandini, A. Ligand-Induced Perturbation of the HIF-2 A: ARNT Dimer Dynamics. *PLoS Comput. Biol.* **2018**, *14*, No. e1006021.

(39) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81*, 3684−369.

(40) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity-Rescaling. *J. Chem. Phys.* **2007**, *126*, No. 014101.

(41) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52*, 7182−7190.

(42) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18*, 1463−1472.

(43) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577−8593.

(44) Allen, M. P. Computer Simulation of a Biaxial Liquid Crystal. *Liq. Cryst.* **1990**, *8*, 499−51.

(45) Goddard, T. D.; Huang, C. C.; Meng, E. C.; Pettersen, E. F.; Couch, G. S.; Morris, J. H.; Ferrin, T. E. UCSF ChimeraX: Meeting Modern Challenges in Visualization and Analysis. *Protein Sci.* **2018**, *27*, 14−25.

(46) Kohonen, T. The Self-Organizing Map. *Proc. IEEE* **1990**, *78*, 1464−1480.

(47) Kohonen, T. Essentials of the Self-Organizing Map. *Neural Networks* **2013**, *37*, 52−65.

(48) Duarte, J. M.; Sathyapriya, R.; Stehr, H.; Filippis, I.; Lappe, M. Optimal Contact Definition for Reconstruction of Contact Maps. *BMC Bioinf.* **2010**, *11*, 283.

(49) R-Development-Core-Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2014.

(50) Wehrens, R. Self- and Super-Organizing Maps in R: The Kohonen Package. *JSS J. Stat. Softw.* **2007**, *21* (5), 1−19.

(51) Csardi, G.; Nepusz, T. The Igraph Software Package for Complex Network Research. *InterJournal Complex Sy* **2006**, 1695.

(52) Craveur, P.; Joseph, A. P.; Esque, J.; Narwani, T. J.; Noël, F.; Shinada, N.; Goguet, M.; Leonard, S.; Poulain, P.; Bertrand, O.; et al. Protein Flexibility in the Light of Structural Alphabets. *Front. Mol. Biosci.* **2015**, *2*, 1−20.

(53) Pandini, A.; Fornili, A.; Fraternali, F.; Kleinjung, J. GSATools: Analysis of Allosteric Communication and Functional Local Motions Using a Structural Alphabet. *Bioinformatics* **2013**, *29*, 2053−2055.

(54) Shao, J.; Tanner, S. W.; Thompson, N.; Cheatham, T. E. Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *J. Chem. Theory Comput.* **2007**, *3*, 2312−2334.

(55) Keller, B.; Daura, X.; Van Gunsteren, W. F. Comparing Geometric and Kinetic Cluster Algorithms for Molecular Simulation Data. *J. Chem. Phys.* **2010**, *132*, 074110.