Electronic Supplementary Material (ESI) for Nanoscale.

Supporting Information for "Translocation intermediates of Ubiquitin through an α-Hemolysin nanopore: implications for detection of post-translational modifications"

Emma Letizia Bonome[†], Fabio Cecconi[‡], and Mauro Chinappi[‖]

E-mail:

†Dipartimento di Ingegneria Meccanica e Aerospaziale, Sapienza Università di Roma, Via Eudossiana 18, 00184, Roma, Italia.

‡ CNR-Istituto dei Sistemi Complessi UoS Sapienza, Roma, 00185, Italia.

‖Dipartimento di Ingegneria Industriale, Università di Roma Tor Vergata, Roma, 00133, Italia.

E-mail: mauro.chinappi@uniroma2.it

Table S1: List of the simulations performed for the C- and N-pulling cases. The first column reports the simulation code $\mathrm{TFS_x}$ where T is the pulling terminus (C or N), F the code for the forcing ($F_1 = 0.75$ nN, $F_2 = 0.65$ nN and $F_3 = 0.6$nN, S is the starting conformation (native conformation Ω or a pre-unfolded conformation Υ), and x $(1 \dots 5)$ indicates the replica number. The second column reports the translocation time. If the Ubq does not completely translocate during the simulation, the second column reports the symbol "–" and the total simulation time T_w is indicated in parenthesis. The third column reports the stalls observed along the trajectory.

Simulation	Translocation time (ns)	Stalls				
$\mathrm{C1}\Omega_1$	14.2		*II*	*III*	*IV*	
$\mathrm{C1}\Omega_2$	18.3	*I*			*IV*	
$\mathrm{C1}\Omega_3$	17.5	*I*			*IV*	
$\mathrm{C1}\Omega_4$	– (28)	*I*		*III*	*IV*	
$\mathrm{C1}\Omega_5$	18.5	*I*		*III*		
$\mathrm{C2}\Omega_1$	– (50)	*I*	*II*	*III*		
$\mathrm{C2}\Upsilon_1$	33.7				*IV*	*V*
$\mathrm{C2}\Upsilon_2$	30.9				*IV*	*V*
$\mathrm{C2}\Upsilon_3$	29.2				*IV*	
$\mathrm{C2}\Upsilon_4$	28.3				*IV*	*V*
$\mathrm{C2}\Upsilon_5$	28.5				*IV*	*V*
$\mathrm{C3}\Upsilon_1$	– (44)				*IV*	*V*
$\mathrm{C3}\Upsilon_2$	– (44)					*V*
$\mathrm{C3}\Upsilon_3$	38.5				*IV*	*V*
$\mathrm{C3}\Upsilon_4$	43.4			*III*		
$\mathrm{C3}\Upsilon_5$	– (44)				*IV*	*V*
$\mathrm{N1}\Omega_1$	29.4	*I*		*II*		*III*
$\mathrm{N1}\Omega_2$	33.24			*II*		
$\mathrm{N1}\Omega_3$	46.44	*I*		*II*		*III*
$\mathrm{N1}\Omega_4$	– (48)	*I*				
$\mathrm{N1}\Omega_5$	22.68	*I*		*II*		*III*
$\mathrm{N2}\Omega_1$	– (40)	*I*		*II*		
$\mathrm{N2}\Upsilon_1$	42			*II*		*III*
$\mathrm{N2}\Upsilon_2$	42			*II*		*III*
$\mathrm{N2}\Upsilon_3$	23.9			*II*		*III*
$\mathrm{N2}\Upsilon_4$	21.3			*II*		
$\mathrm{N2}\Upsilon_5$	28.4			*II*		*III*

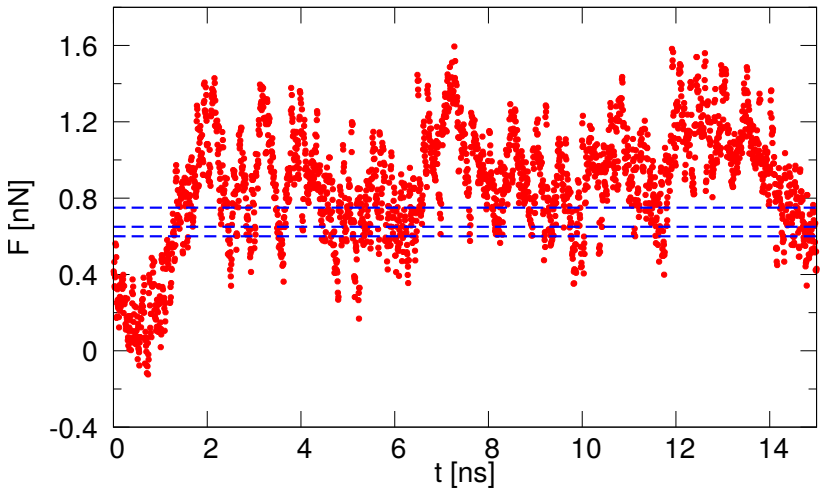

Figure S1: Constant velocity steered molecular dynamics (cvSMD) for C-terminus pulling. Time evolution of the $z-$component of $\mathbf{F}_{\mathrm{v}}$. The force profile shows distinct peaks corresponding to specific unfolding of secondary structure elements, e.g., $\beta 5$ at 2 ns, $\beta 4$ and $\beta 3$ at 3 ns, HA at 6.5 ns. The horizontal blue dashed lines correspond to the selected values of the force at which we run the cfSMD simulations, $F_1 = 0.75$ nN, $F_2 = 0.65$ nN and $F_3 = 0.6$ nN.

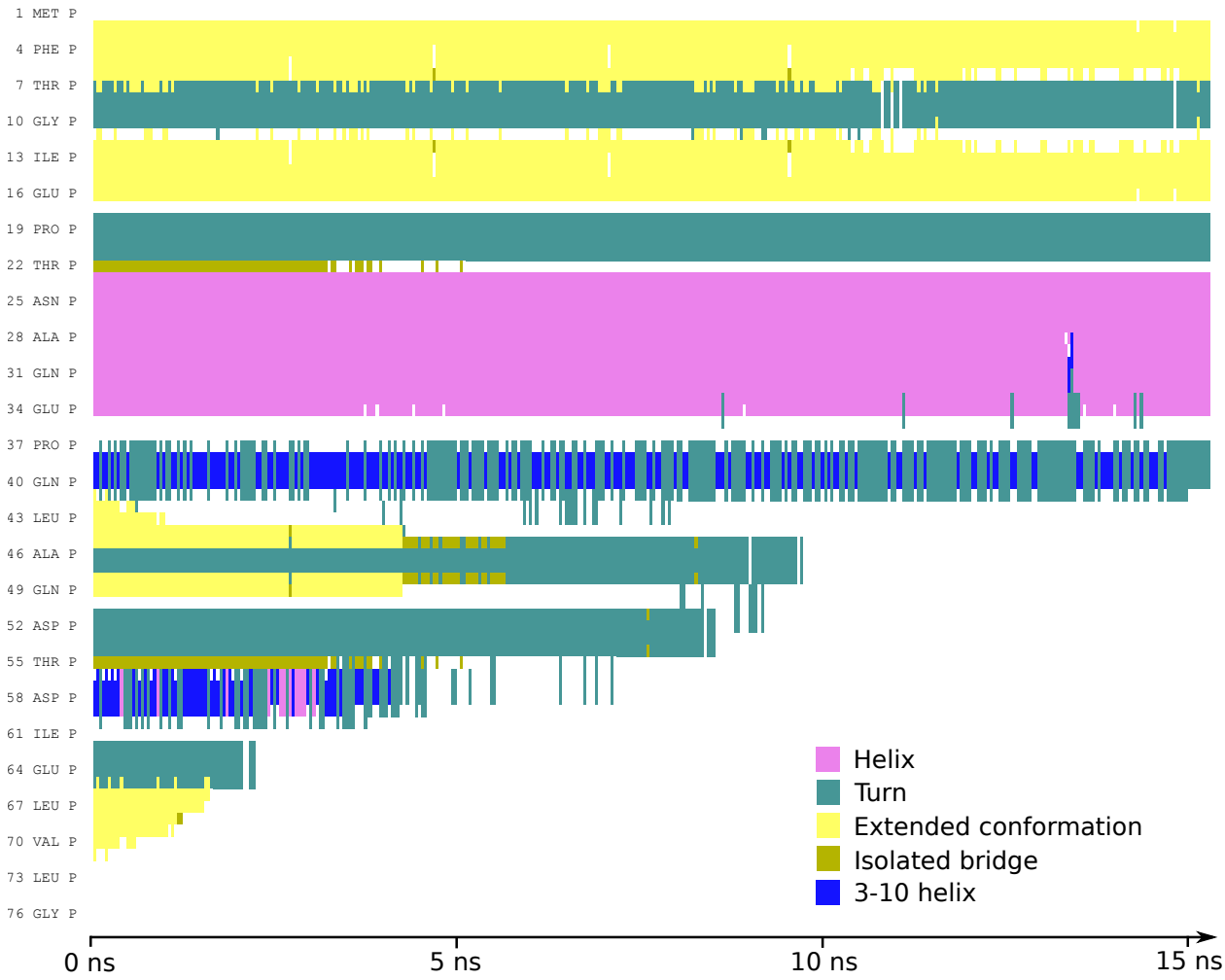

Figure S2: Time evolution of secondary structure elements for the C-terminus simulation $C2\Omega_1$ applied force $F_2 = 0.65$ nN (figure 2A of the manuscript). The color key and the one-letter structure code is the one used by the STRIDE software[1]. The analysis is performed using VMD[2]. Color-code: yellow corresponds to "Extended conformation", i.e. the main component of beta sheets; aqua corresponds to "Turn", another beta sheet component, blue and pink correspond to $3-10$ and α helices, respectively, green to the isolated bridge while white stands for random coil, i.e. no secondary structure.

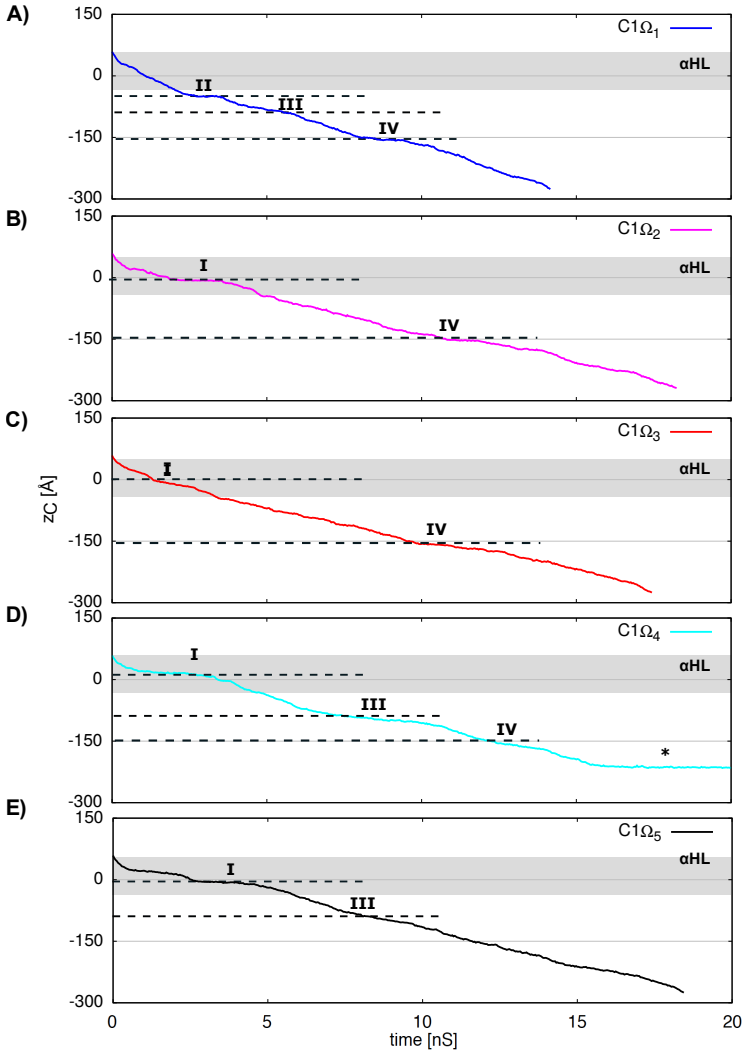

Figure S3: Constant force steered Molecular Dynamics Simulation (cfSMD) for C-terminus pulling. A-E) Time evolution of C-terminus $z-$coordinate for simulation starting from the native conformation Ω, pulling force $F_1 = 0.75$ nN. Simulations $C1\Omega_1$, $C1\Omega_2$, $C1\Omega_3$, $C1\Omega_5$ complete the translocation in a time window $T_{\rm w} < 20$ ns, while in the $C1\Omega_4$ the protein get stuck in the pore in a conformation, identified with $*$ similar to the stall IV reported in figure 2F of the paper. The translocation intermediates corresponding to the stalls are indicated using the roman numerals. Stalls I and III are present in several simulations, while the stall II appears only in one of them. As discussed in the paper, a possible explanation is that, while stall I and III stalls correspond to rearrangements of the native secondary structure stall II is associated to a contingent interaction of the unfolded residues originally belonging to $\beta 4$ and $\beta 5$ (residue 48-49, 66-71) with the interior pore surface. Consequently, we expect that, stall I and III are quite reproducible among the replicas while stall II not.

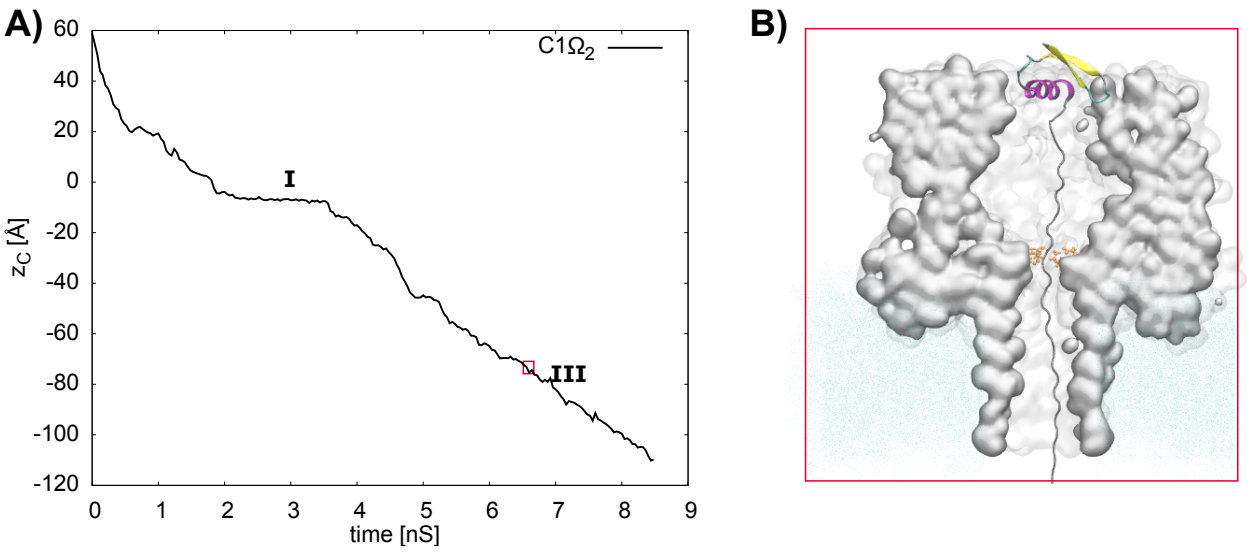

Figure S4: Υ conformation for C-terminus. A) Time evolution of C-terminus $z-$coordinate for a simulation starting from native structure, pulling force $F_1 = 0.75$ nN. From this translocation pathway we have selected a conformation after stall III, red rectangle at $z_{\rm C} \simeq -80$ Å, $t \simeq 6.5$ ns. This conformation, indicated as Υ, is used as initial condition for two new sets of simulations $C2\Upsilon_{\rm x}$ and $C3\Upsilon_{\rm x}$, $x = 1 \ldots 5$, performed to explore the whole translocation pathway.

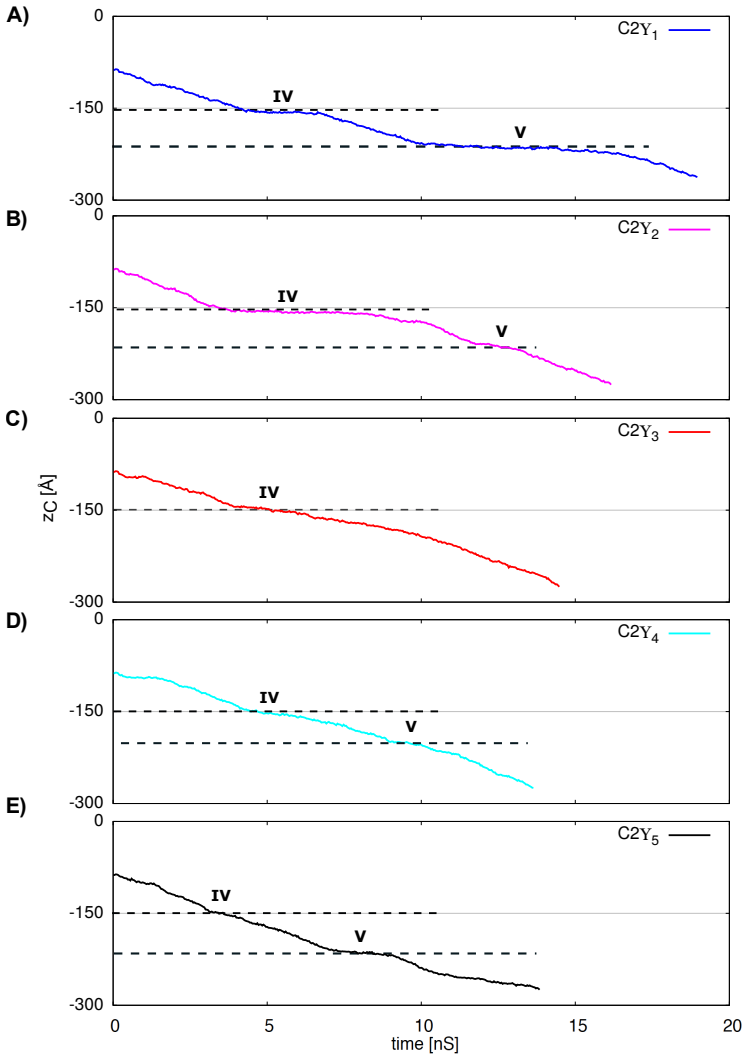

Figure S5: Constant force steered Molecular Dynamics Simulation (cfSMD) for C-terminus pulling. A-E) Time evolution of C-terminus $z-$coordinate for simulation starting from conformation Υ, pulling force $F_2 = 0.65$ nN.

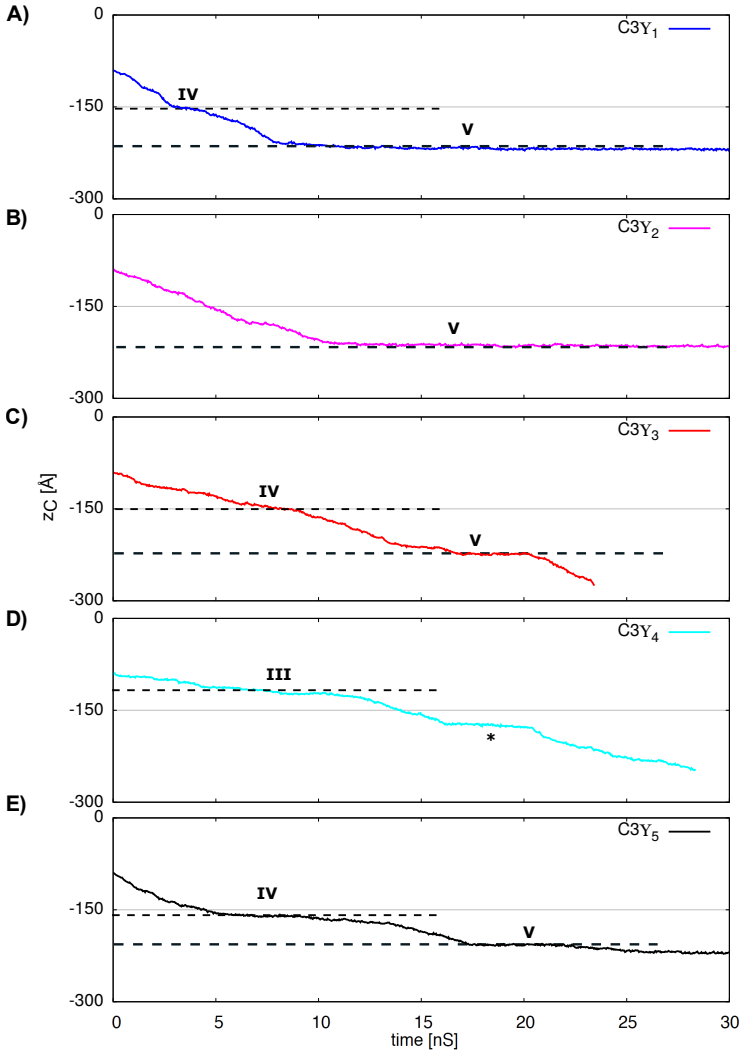

Figure S6: Constant force steered Molecular Dynamics Simulation (cfSMD) for C-terminus pulling. A-E) Time evolution of C-terminus $z-$coordinate for simulation starting from conformation Υ, pulling force $F_3 = 0.60$ nN. The simulations $C3\Upsilon_1$, $C3\Upsilon_2$, $C3\Upsilon_5$ do not complete the translocation and the protein remains blocked in the nanopore in the intermediate V reported in the figure 2 of the paper. The time window is $T_{\rm w} = 44$ ns (only the first 30 ns are shown). The simulation $C3\Upsilon_4$ presents the same stall $*$ previously analyzed in the figure S3D where the two βstrands, $\beta 2$ and $\beta 1$, get stuck at the cis entrance and unfold before entering the vestibule.

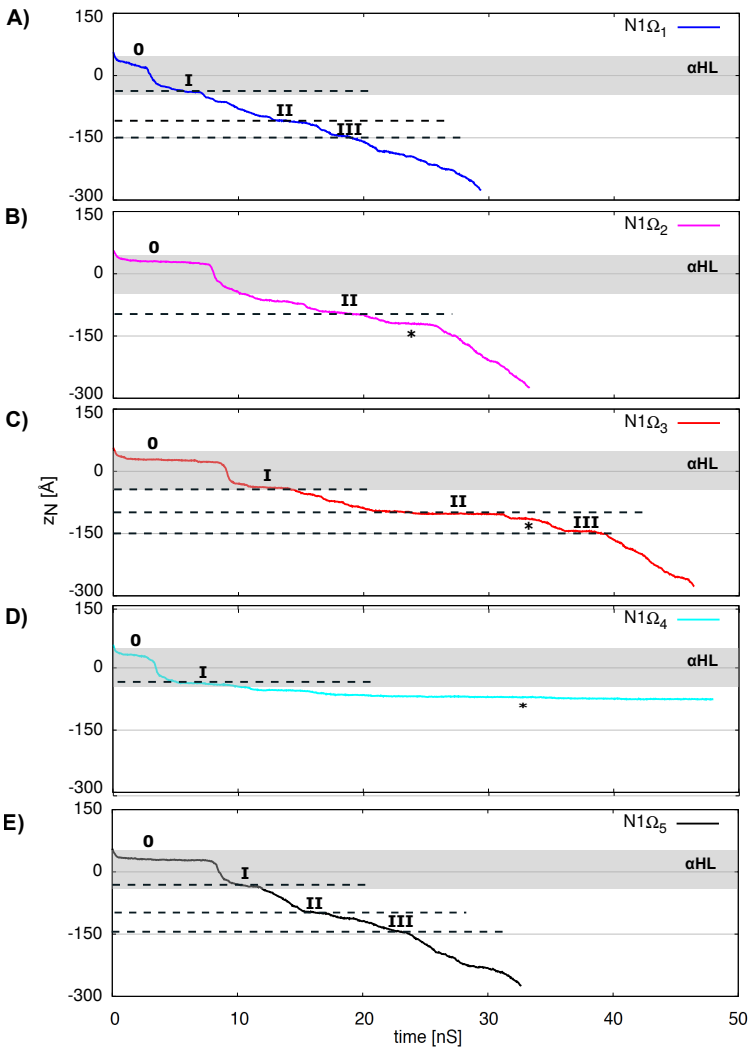

Figure S7: Constant force steered Molecular Dynamics Simulation (cfSMD) for N-terminus pulling. A-E) Time evolution of N-terminus $z-$coordinate for simulation $N1\Omega_{\mathrm{x}}$, $x \in (1,5)$, i.e. simulation pulled from native initial condition and force $F_1 = 0.75$ nN. Simulations $N1\Omega_1$, $N1\Omega_2$, $N1\Omega_3$, $N1\Omega_5$ complete the translocation in a time window $T_{\mathrm{w}} = 50$ ns, while in the $N1\Omega_4$ the protein get stuck in the pore in a conformation, identified with $*$ similar to the stall II reported in figure 3 on the paper. All the simulations show the same translocation intermediates already reported in the figure 3 of the paper. Also a first stall (0) is apparent. In this conformation, all the secondary structure elements are still folded although the tertiary structure is slightly deformed and the protein get stuck at the cis entrance.

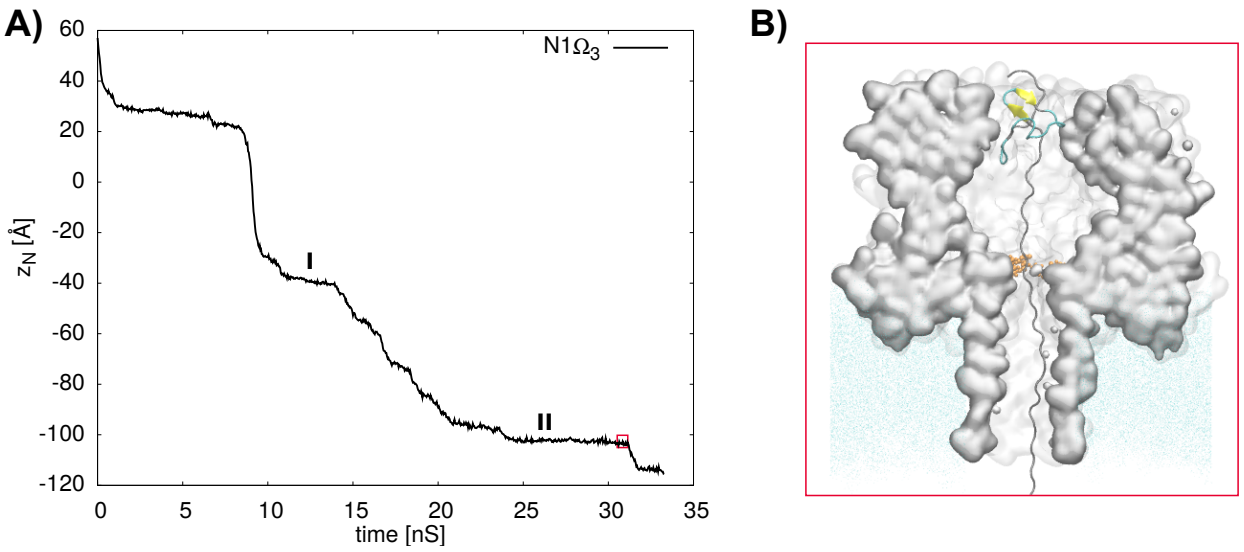

Figure S8: Υ conformation for N-terminus pulling. A) Time evolution of N-terminus for simulation starting from the native conformation Ω, pulling force $F_1 = 0.75$ nN, simulation code $N1\Omega_3$. From this translocation pathway, we selected a conformation after stall II (red square at $z_{\rm N} \simeq -120$ Å, $t \simeq 31$ ns. This conformation, indicated as Υ in the simulation code, is used as initial condition for a new sets of simulations $N2\Upsilon_{\rm x}$ x $(1 \ldots 5)$ conceived to explore the whole the translocation pathway.

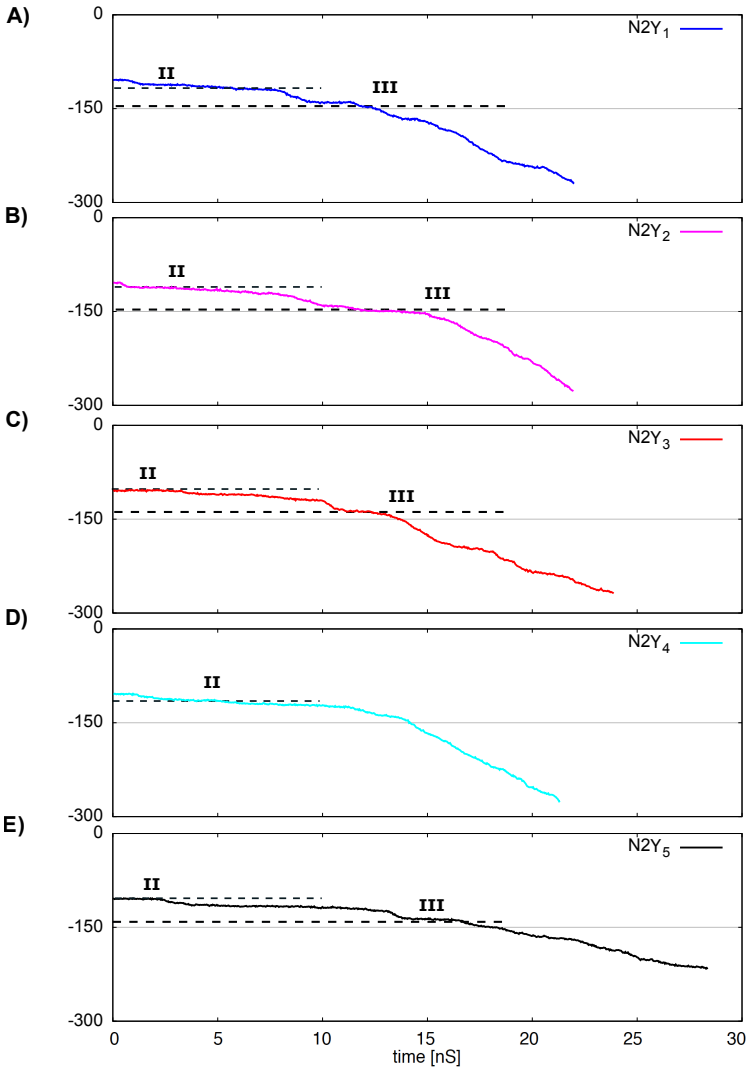

Figure S9: Constant force steered Molecular Dynamics Simulation (cfSMD) for N-terminus pulling. A-E) Time evolution of N-terminus $z-$coordinate starting from conformation Υ. Pulling force $F_2 = 0.65$ nN. All the simulations show the same translocation intermediates, II and III, reported in the figure 3 on the paper.

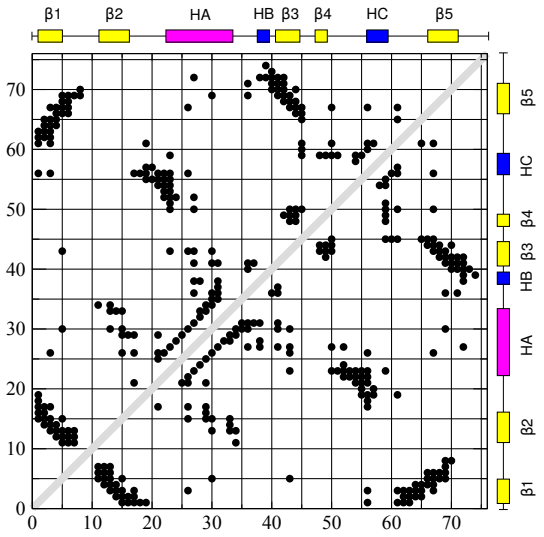

Figure S10: Heavy-atom contact map generated from the crystallographic structure of the ubiquitin, pdb entry 1UBQ[5]. A cutoff of $R_c = 5$ Å selects $M = 190$ native contacts for 1UBQ structure.

S1. Gō-model with heavy map: C-pulling

The Ubq is modelled by a Gō-like force field proposed by Clementi et al.[3]. The details about force-field parameterization and implementation can be found in[3,4]. The chain is represented by taking into account only Cα atom positions (beads), since we are mainly interested in the structural rearrangements of the backbone along the translocation pathway. We recall that Gō-models are such that energy function takes its minimum on the coordinates of the crystallographic structure of the native state, in the present work such coordinates are extracted from the pdb entry 1UBQ[5]. A simple way to achieve that the native structure is a minimum of the potential energy is by introducing the notion of native interactions, or *contacts*. In this work, we consider two residues i, j in native interaction if they share a couple of heavy-atoms, i.e. all atoms but Hydrogens and Nitrogens, within a cutoff distance $R_c < 5$ Å. The resulting contact map is reported in Fig. S10, showing the 190 contacts.

The interactions between the beads are associated to peptide bonds, angular bending, torsional deformation and native contacts, see[3,4], leading to the energy function for a N

residue protein

$$\begin{aligned}\Phi_{\mathrm{G\bar{o}}} &= \sum_{\mathrm{i}=1}^{N-1} V_{\mathrm{p}}(r_{\mathrm{i,i+1}}) + \sum_{\mathrm{i}=1}^{N-2} V_{\theta}(\theta_{\mathrm{i}} - \theta_{\mathrm{i}}^{0}) + \sum_{\mathrm{i}=1}^{N-3} V_{\varphi}(\varphi_{\mathrm{i}} - \varphi_{\mathrm{i}}^{0}) \\ &+ \sum_{\mathrm{i,j}\geq\mathrm{i}+3} V_{\mathrm{nb}}(r_{\mathrm{ij}}). \end{aligned} \tag{1}$$

The peptide bond term, V_{p}, enforcing the chain connectivity, is a stiff harmonic potential allowing only small oscillations of the bond lengths around their crystallographic values. Likewise, the bending potential V_{θ} allows only small fluctuations of the bending angles θ_{i} around their native values θ_{i}^{0}. Dihedral potential V_{ϕ} (associated to torsional deformation) further contributes to the correct formation of the native secondary structure characterized also by angles φ_{i}^{0}. Finally, the long-range potential V_{nb}, which favors the formation of the correct native tertiary structure, is a collection of two-body 12-10 Lennard-Jones contributions that are attractive between residues forming native contacts and repelling for non-native couples.

S1.1. Pore geometry

The confining effect of the αHL nanopore is described by the following potential acting only in the pore region, $0 < x < L$,

$$V_{\mathrm{p}}(x,y,z) = V_0 \begin{cases} 0 & y^2 + z^2 \leq R^2(x) \\ \\ \left[\left(\dfrac{y}{R(x)}\right)^2 + \left(\dfrac{z}{R(x)}\right)^2 - 1\right]^m & y^2 + z^2 > R^2(x)\ . \end{cases} \tag{2}$$

To fit the vestibule-barrel shape of the αHL, the pore radius is modulated as

$$R(x) = \frac{R_{\mathrm{v}} + R_{\mathrm{b}}}{2} - \frac{R_{\mathrm{v}} - R_{\mathrm{b}}}{2} \tanh[\alpha(x - x_{\mathrm{c}})] \tag{3}$$

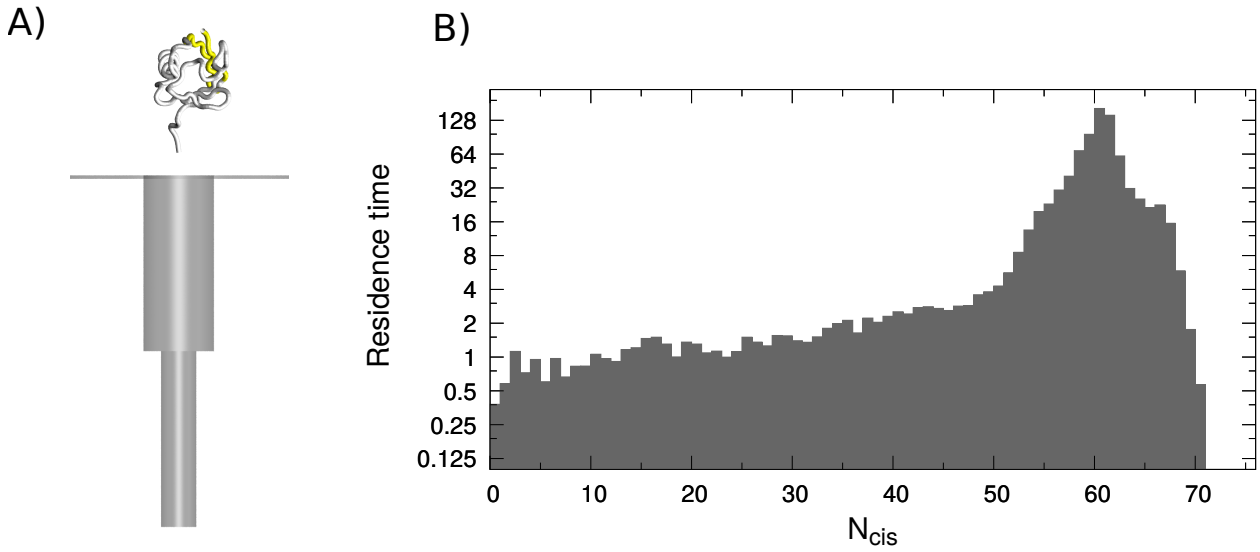

Figure S11: A) Sketch of the coarse grained simulation set-up. B) Histogram of the residence time for the variable $N_{\rm cis}$ obtained averaging over 2000 C-pulling runs starting from native conformation for pulling force $F = 1.6$. The large peak is associated to the unfolding of $\beta 5$.

with $L \simeq 100$ Å, $R_{\rm v} = 10$ Å (for the vestibule), $R_{\rm b} = 4$ Å (for the barrel), $x_{\rm c} = L/2$ and $m = 4$.

A repulsive force, $F_{\rm w}(x)$, orthogonal to planes $x = 0$, $x = L$ and vanishing for $y^2 + z^2 < R(x)^2$, models the presence of the impenetrable membrane hosting the pore

$$F_{\rm w}(x) = \begin{cases} -\dfrac{e^{\lambda x}}{x + c} & x \leq 0 \\ 0 & 0 < x < L \\ \dfrac{e^{-\lambda(x-L)}}{x - L + c} & x \geq L \end{cases} \tag{4}$$

with $c = 10^{-4}$ Å being a regularisation cutoff to avoid numerical overflow and $\lambda = 6$ Å^{-1}.

S1.2. Simulation protocol

The importing mechanism that drives the protein into the pore is simplified to a constant pulling force $(F, 0, 0)$ acting only on the C-terminus bead ($\mathbf{r}_{76}$). Moreover, the pulled terminus is constrained to slide along the pore axis for all time, i.e., $y_{76}(t) = z_{76}(t) = 0$. Simulations

were performed by using a coarse grained molecular dynamics at constant temperature $T = 0.8$ implemented with a Langevin dynamics that evolves the position $\mathbf{r}_\mathrm{i}$ of the $i = 1, \ldots, N$ residues

$$M_\mathrm{aa}\ddot{\mathbf{r}}_\mathrm{i} = -\gamma\dot{\mathbf{r}}_\mathrm{i} - \nabla_{\mathrm{r_i}}\left(\Phi_\mathrm{G\bar{o}} + V_\mathrm{p}\right) + \mathbf{F}_{76} + \mathbf{F}_\mathrm{w}(x_\mathrm{i}) + \mathbf{Z}_\mathrm{i}\,, \tag{5}$$

where M_aa denotes the average amino acid mass, $\mathbf{Z}_\mathrm{i}$ is a random force with zero average and correlation $\langle Z_{\mathrm{i},\mu}(0)Z_{\mathrm{i},\nu}(t)\rangle = 2\gamma k_\mathrm{B}T\delta_{\mu,\nu}\delta(t)$, with $\mu, \nu = x, y, z$ and k_B being the Boltzmann's constant ($k_\mathrm{B} = 1$), $\mathbf{F}_{76}$ the importing force and $\Phi_\mathrm{G\bar{o}}$, V_p and $\mathbf{F}_\mathrm{w}$ given by (1), (2) and (4), respectively. We used $\gamma = 5.0$ and a time step $h = 0.005$. Each translocation run started by positioning the native structure with the C-terminus at $\mathbf{x}_{76} = (10, 0, 0)$ Å and thermalizing for $t_\mathrm{eq} = 10^4$ time steps with the C-terminal blocked. Then, the forcing is turned on and the simulation is stopped when all the residues reached the trans side ($x_\mathrm{i} > L$). Like in the all-atom case, the unfolding of the first translocation intermediate requires a quite large force, hence, once unfolded, the remaining translocation intermediates are less evident. Therefore, we employed an approach similar to the one described for all-atom MD. First, we performed high force runs ($F_{76} = 2.2$) to explore the complete translocation pathway. Then, we selected a conformation just after the unfolding of the first structural cluster (e.g. after stall III) and, thus, we used it as starting conformation for runs at lower force ($F_{76} = 1.6$).

To characterize the dynamics of the translocation we measure the time course of the two collective variables

$$N_\mathrm{cis}(t) = N - \sum_{\mathrm{i}=1}^{N}\Theta(x_\mathrm{i})\,, \tag{6}$$

and

$$N_\mathrm{cis,v}(t) = N - \sum_{\mathrm{i}=1}^{N}\Theta(x_\mathrm{i} - L/2)\,, \tag{7}$$

$\Theta(s)$ is the unitary step function. N_cis and $N_\mathrm{cis,v}$ correspond to the number of residues that have not yet entered the pore vestibule (N_cis) and the barrel ($N_\mathrm{cis,v}$), respectively. In essence, at the initial condition $N_\mathrm{cis} = N_\mathrm{cis,v} = 76$, i.e. all the 76 amino acids are outside the pore at the cis side. As long as the translocation proceeds, both variables decrease. Peaks in the

histogram of these variables indicate the presence of translocation bottlenecks.

S2. Electrolyte accessibility estimator

In experiments, nanopore clogging is usually characterized in term of the current blockade $(I_0 - I(t))/I_0$ or, in alternative, in term of the residual current $I(t)/I_0$, where $I(t)$ is the current trace associated to the nanopore-molecule interaction and I_0 is the open pore current. The residual current $I(t)/I_0$ can be written in term of pore resistances, indeed, using Ohm's law,

$$I_{\rm res}(t) = \frac{I(t)}{I_0} = \frac{R_0}{R(t)} \ . \tag{8}$$

In a quasi-1D continuum model, the resistance of a pore can be modelled as

$$R = \int_0^L \frac{\rho(z)}{A(z)}\, dz \ , \tag{9}$$

where the $z-$axis coincides with the pore axis, the pore goes from $z = 0$ to $z = L$, $\rho(z)$ is the electrolyte resistivity, here assumed to be homogeneous, and $A(z)$ is the area of the pore section available to the electrolyte passage. Access resistances are neglected in expression (9).

The integral in eq. (9) can be approximated dividing the system in $N_{\rm z}$ slabs of size Δz obtaining

$$R = \sum_{\rm i=1}^{N_{\rm z}} \frac{\rho}{A_{\rm i}} \Delta z \ , \tag{10}$$

while the available effective area $A_{\rm i}$ can be estimated as $A_{\rm i} = V_{\rm i}/\Delta z$ where $V_{\rm i}$ is the volume occupied by the electrolyte in the i-th slab.

Inspired by these continuum quasi-1D arguments, we defined the electrolyte accessibility estimator as

$$c(t) = \frac{R_0}{R(t)} \ . \tag{11}$$

In particular, we calculated eq. (10) for each frame and then, we used this value in eq. (11). The volume V_{i} appearing in eq. (10) is estimated as the number of electrolyte molecules (water or ions) in the i-th slab times a reference volume V_{ele} that corresponds to the typical volume of a water molecule. Note that, in the expression (11), V_{ele} simplifies, so its exact value is not relevant for our purposes. The time evolution of $c(t)$ is highly noisy, hence, for the sake of clarity, in figure 5 of the main text we reported a running average performed over 10 values from consecutive snapshots separated by 0.04 ns.

The above model is based on several hypotheses that are violated by the actual αHL pore shape. In particular, the continuum assumption is not justified at nanoscale, moreover, the model implicitly assumes a smooth variation of A_{i} along the pore axis (quasi-1D) and a homogeneous electrolyte resistivity ρ. Nevertheless, although a quantitative agreement with the residual current is not expected, we are confident that the trend in the current levels would be the same. In particular, we expect that the smaller current (larger blockade) would correspond to stall III while stall IV and V would be associated to progressively larger currents (smaller blockades). As a final comment, for readers reference, it is worth mentioning that similar quasi-1D models have been employed in other all-atom MD studies[6,7].

S3. Structural analysis

To check if there were systematic differences among the translocation intermediates over the different replicas (e.g. if stall V from simulations C3Υ_3 and C3Υ_5 differs) we performed a structural clustering analysis as follows:

1. For a given stall, we selected all the corresponding MD frames for all the simulations where the stall appears.

2. For each frame, we selected only the folded portion of the Ubq. For instance, in the case of stall III for C-pulling, we selected the residues from 1 to 35, corresponding to

the HA, $\beta 1$ and $\beta 2$. Hence, at this stage, we have a set of N structures corresponding to a given stall.

3. We then calculated the RMSD distance among all the N structures obtaining an $N \times N$ distance matrix.

4. The distance matrix was used as input of a clustering algorithm (complete linkage method for hierarchical clustering from cluster library[8] in the R software[9])

5. As usual in clustering analysis, we calculated the average silhouette for different partitions of the data (number of clusters).

For all the stalls, we get a low silhouette (maximum value $\simeq 0.35$), hence, indicating no clear partitions among the different structures. For stall involving folded structures, the average RMSD among translocation intermediate configurations taken from different replicas are also very small (just a few Angstrom). As an example, for stall III in the C-pulling runs, the maximum RMSD is $\simeq 5.5$ Å corresponding to the conformations from C1Ω_4 (red in fig. S12) and C3Υ_4 (blue). The main variations are a different orientation of HA and a partial unfolding of β_1.

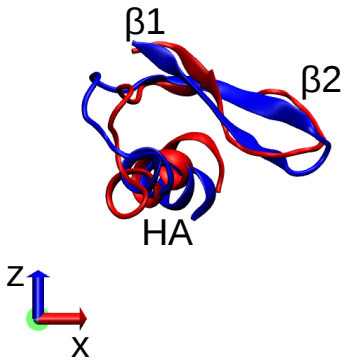

Figure S12: Maximum difference between the conformations belonging to stall III in the C-pulling runs. Red structure is taken from C1Ω_4 while blue from C3Υ_4.

References

(1) Heinig, M.; Frishman, D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic acids research* **2004**, *32*, W500–W502.

(2) Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. **1996**, *14*, 33–38.

(3) Clementi, C.; Nymeyer, H.; Onuchic, J. N. Topological and energetic factors: what determines the structural details of the transition state ensemble and en-route intermediates for protein folding? an investigation for small globular proteins1. *Journal of molecular biology* **2000**, *298*, 937–953.

(4) Cecconi, F.; Guardiani, C.; Livi, R. Testing simplified proteins models of the hPin1 WW domain. *Biophysical journal* **2006**, *91*, 694–704.

(5) Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J. Structure of ubiquitin refined at 1.8 Åresolution. *Journal of molecular biology* **1987**, *194*, 531–544.

(6) Si, W.; Aksimentiev, A. Nanopore sensing of protein folding. *ACS nano* **2017**, *11*, 7091–7100.

(7) Di Muccio, G.; Rossini, A. E.; Di Marino, D.; Zollo, G.; Chinappi, M. Insights into protein sequencing with an α-Hemolysin nanopore by atomistic simulations. *Scientific Reports* **2019**,

(8) Maechler, M.; Rousseeuw, P.; Struyf, A.; Hubert, M.; Hornik, K. cluster: Cluster Analysis Basics and Extensions. 2018.

(9) R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2015.