Protein translocation in narrow pores: Inferring bottlenecks from native structure topology

Marco Bacci,[1,*] Mauro Chinappi,[2,*] Carlo Massimo Casciola,[3] and Fabio Cecconi[4,†]

[1]*Dipartimento di Ingegneria Civile e Ambientale, Università degli Studi di Firenze, via Santa Marta 3, 50139, Firenze, Italy*

[2]*Center for Life Nano Science@Sapienza, Istituto Italiano di Tecnologia, Viale Regina Elena 291, 00161 Roma*

[3]*Dipartimento di Ingegneria Meccanica e Aerospaziale, Sapienza Università di Roma, via Eudossiana 18, 00184, Roma*

[4]*CNR-Istituto dei Sistemi Complessi, Via Dei Taurini 19, 00185 Roma, Italy*

(Received 13 February 2013; revised manuscript received 5 July 2013; published 14 August 2013)

Coarse-grained simulations of protein translocation across narrow pores suggest that the transport is characterized by long stall events. The translocation bottlenecks and the associated free-energy barriers are found to be strictly related to the structural properties of the protein native structure. The ascending ramps of the free-energy profile systematically correspond to regions of the chain denser in long range native contacts formed with the untranslocated portion of the protein. These very regions are responsible for the stalls occurring during the protein transport along the nanopore. The decomposition of the free energy in *internal energy* and *entropic* terms shows that the dominant energetic contribution can be estimated on the base of the protein native structure only. Interestingly, the essential features of the dynamics are retained in a reduced phenomenological model of the process describing the evolution of a suitable collective variable in the associated free-energy landscape.

DOI: 10.1103/PhysRevE.88.022712 PACS number(s): 87.15.A−, 87.14.E−, 87.16.dp

I. INTRODUCTION

Nanopore-based protocols for macromolecule detection offer the technology for developing sensors and devices able to operate at single-molecule level. Their working principle, based on the Coulter resistive method, is very simple [1]. A nanopore connects two chambers containing an electrolyte solution and an applied voltage across the chambers generates a net ion current. When one of the macromolecules dispersed in the solution engages the pore, the ion flux is altered and a drop in the current is commonly detected. The intensity and the duration of the drop depend on the physicochemical properties of the passing molecule and on the microscopic dynamics. Hence, nanopores can, in principle, provide precise information on a single-molecule level. Significant efforts have been devoted to nanopore DNA sequencing [2] and only recently possible proteomic applications have started being explored [3–12] using both solid-state and biological pores.

A widely used nanopore is the α-hemolysin (αHL) that reproducibly self-assembles into the lipid membrane resulting in a $\simeq$100-Å-long and $\simeq$20-Å-wide channel. The protein passage across a narrow pore, like α-hemolysin [3,5,7], occurs via a capture stage followed by single-file translocation. Folded proteins are extremely hard, if not impossible, to capture [5] in standard voltage-driven experiments with biopores, requiring a prior chemical unfolding in the bulk before translocating. Clearly, chemical unfolding has the disadvantage to erase native-structure information and can be detrimental to nanopore-based protein characterization. Nevertheless, it has been recently shown that proteins can be engineered with designed linkers able to promote capture and retention of even folded conformations [13,14]. In such conditions, translocation is accomplished concurrently with mechanical unfolding and takes the form of a *multistep* process revealed by different current levels that can be, in principle, exploited for protein recognition. The multistep dynamics, with the molecule stuck at specific conformations (stall events), has been addressed in theoretical and numerical studies, see Ref. [15] for maltose-binding protein (MBP) translocation across αHL-like channels and Ref. [16] for a rod-coil polymer passing through a model nanopore. Moreover, experimental studies [5–7] have shown that certain translocation can be characterized by long tails in the blockade time distributions with the suspect that, besides the capture, some other kind of rate-limiting mechanism could be at work during the migration determining the bottlenecks.

In the present paper, we show that (i) the stalling events are associated with specific regions of the protein particularly rich of *backward contacts*, i.e., the contacts an amino acid forms with the portion of the chain still outside the pore entrance, and (ii) the *backward contact profile* reflects directly into a step-like structure of the freeenergy G as a function of a suitably defined collective variable Q. Also, remarkably, (iii) the one-dimensional Langevin dynamics of Q based on the free-energy profile $G(Q)$ retains the essence of the phenomenon.

Simulations involving different proteins support the conjecture that these results are generic. The decomposition of the free energy into its entropic and energetic components confirms the crucial role of the protein regions that are denser of long-range interactions (native contacts) in hindering the translocation dynamics. The breaking of such contacts leads to an internal energy gain that dominates the entropy variation, with a final overall free-energy increase. It is worth stressing that, in contrast with unstructured polymers, the entropy variation for native-like proteins is not trivial, as it is the result of the competition between the disorder produced by contact breaking and the order associated with chain confinement inside the pore.

Method and results are first introduced for the MBP and then generalized to a MBP-mutant and to another globular protein (Bacillus Agaradherans family 5 Endoglucanase—PDB:id 1A3H [17]), selected for its substantially different native topology.

*These authors have contributed equally to this work.

†fabio.cecconi.@roma1.infn.it

II. SYSTEM SET-UP

The protein is described by a Gō-like model [18–20], a widely employed off-lattice coarse-grained approach where the protein is modeled as a sequence of beads, each one corresponding to a single amino acid. An important feature of the Gō-like force field is its ability to account for the influence of the protein native state on both folding and translocation pathways [15,18,19,21]. The model is described in Appendix A. Here, for the reader's convenience, we briefly report the reference quantities useful to discussing data and results. We select the angstrom as unit of length, the mass m_a of the bead representing an amino acid (in the description all the amino acids have the same mass) as mass unit, and the characteristic energy scale ϵ that appears in the Gō-like force field as energy unit. These three reference quantities define all the other units, in particular the force and the time unit. All the quantities discussed in the paper are represented as ratios with respect to the corresponding reference one.

It is worth noting that the model is built on the basis of the crystallographic protein structure, implicitly assuming that, except for compatible thermal distortions, it is the actual structure retained by the protein in solution prior to translocation. A basic parameter of the model is the cut-off radius R_c that determines the number of native (attractive) contacts between nonbonded amino acid and controls the stability of the native structure [15].

A cylindrical potential is tailored to mimic the confinement effect of a narrow pore (roughly corresponding to the actual dimensions of the αHL channel) that allows the protein translocation in single-file conformations only. This model of the nanopore is expected to reproduce the relevant phenomenology as far as nonspecific interactions are involved. The results we provide do not depend on the pore size as long as the migration remains single file (see the numerical experiments for a graphene-like pore, reported in the supplemental material [22]).

Translocation simulations are performed following the same protocol described in our previous works [15,21] and here briefly sketched. Protein dynamics is simulated by a Langevin equation to control the temperature ($k_BT = 0.75$), integrated via a Verlet algorithm generalized to include friction and stochastic forces [23], with friction coefficient $\gamma = 0.25$ and time step $h = 0.005$. The origin of the reference frame is the center of the left entrance of the pore. The x axis coincides with the pore axis and it is directed from left to right. The simulation procedure is the following: (i) Suitable initial conditions for translocation runs are prepared by constraining the protein to the proximity of the channel entrance. Specifically, the terminus to be pulled inside the pore (C or N) is restrained at point $(-1,0,0)$ by a stiff harmonic potential. We verify that the conformations obtained this way maintained a reasonable native similarity with the crystallographic structure. Thermalized protein conformations are sampled every 10^4 time units (10% of the selected translocation simulation time window) in order to obtain uncorrelated initial conditions for the translocation runs. (ii) In each translocation run the protein is pulled inside the pore by means of a constant force applied to the foremost residue present in the active region, i.e., $x \in [-2, L]$, where L is the pore length. The transport is considered accomplished when the last residue exits the pore. This pulling strategy can be thought of as a model of the average effect of the electric field in a voltage-driven translocation experiment or, even better, as a model of the action of a molecular motor like in the unfoldase-mediated translocation recently reported in Ref. [13]. For *in vivo* translocations, this pulling strategy may not be appropriate.

Throughout the paper, the C-pull and N-pull data (i.e., data concerning translocations pulled from the C- and the N-terminus) are always represented as leftright and rightleft translocations, respectively (see Fig. 1). For a given importing force F_x, the protein accomplishes the translocation in the allotted time window with a certain probability [15]. The present simulations are all run at critical force $F_x = F_c$, i.e., the force at which the translocation probability is one-half. In this condition, on one hand, the dynamics is slow enough to detect the occurrence of the transport stalls and, on the other, it generates a sufficient number of complete translocations to yield meaningful statistics. The critical force F_c is summarized

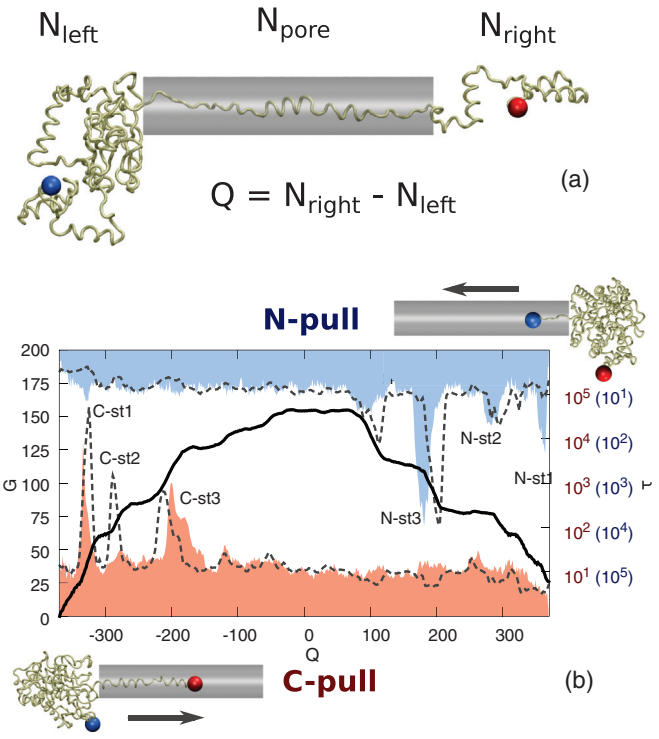

FIG. 1. (Color online) (a) Sketch of a translocating protein conformation. N_{left}, N_{right}, and N_{pore} are the number of monomers on the left side, on the right side, and inside the pore. In our presentation of data, translocations pulled from the C-terminus (red, right) are reported from left →right, while those pulled from the N-terminus (blue, left) from right → left. (b) Histograms: average residence time $\tau(Q)$ from nonequilibrium MD simulations (C-pull = Red, bottom; N-pull = Blue, top). Solid line: free-energy profile $G(Q)$ of MBP as a function of the collective variable $Q = N_{\text{right}} \to N_{\text{left}}$. Dashed lines: $\tau(Q)$ from the 1D Langevin model, Eq. (1). Langevin data are rescaled to match the uniform background value. For N-pulling simulations, the $\tau(Q)$ scale is reversed (values within parentheses on the right y axes).

in Table S1 in Ref. [22] for all the translocation simulations to be discussed.

III. COLLECTIVE VARIABLE AND STALLING POINTS

The high dimensionality of the protein conformation space calls for a reduced description of translocation in terms of a collective variable, a convenient choice is

$$Q = N_{\text{right}} - N_{\text{left}},$$

with N_{right} and N_{left} the number of residues outside the pore on its *right* and *left* side, respectively [Fig. 1(a)]. Q ranges from $-m$, all the chain in the CIS side, to m, full chain in the TRANS side, with m the total number of residues ($m = 370$ for MBP). This nomenclature is used consistently throughout, with C-pulling experiments proceeding from $Q = -m$ to $Q = m$ and N-Pulling simulations proceeding from $Q = m$ to $Q = -m$. The associated color code used in the figures is red for C- and blue for N-pulling. The average residence time $\tau(Q)$ the protein spends in configurations with a given Q value—Fig. 1(b), filled histograms—clearly indicates that the transport is not uniform, as most of the time is spent in specific states, here referred to as stalling points (events C-st1, C-st2, and C-st3 for C-pulling and N-st1, N-st2, N-st3, and N-st4 for N-pulling). In order to better understand the nature of such stalls, the free-energy landscape $G(Q)$ is reconstructed by umbrella sampling simulations combined with the WHAM algorithm [24,25], an approach that requires a continuous version of the collective variable Q (see Appendix B).

The free-energy profile $G(Q)$ [Fig. 1(b), solid curve] is characterized by a step-like shape with the ramps correlated to the stalling events. This noteworthy correspondence suggests that a significant amount of information on the translocation is encoded in $G(Q)$. To highlight the dynamical correspondence between stalls and free-energy ramps, we consider a Langevin model of the translocation over the profile $G(Q)$,

$$\dot{Q} = -\frac{1}{\gamma_e}\frac{\partial G}{\partial Q} + \frac{1}{\gamma_e}\frac{\partial W}{\partial Q} + \sqrt{\frac{2k_BT}{\gamma_e}}\xi\,, \tag{1}$$

with γ_e the effective friction coefficient, $W(Q)$ the work done by the importing force, and ξ a zero-average Gaussian white noise of unit variance. Numerical implementation of Eq. (1) requires an explicit relationship between $F_Q = \partial W/\partial Q$ and F_x, (the latter being the force used in the three-dimensional translocation simulations). Upon requiring that the work done by F_x and F_Q be the same, we obtain

$$F_Q = \frac{\partial W}{\partial x}\frac{\partial x}{\partial Q} \simeq F_x\left(\frac{\Delta x}{\Delta Q}\right);$$

thus, the knowledge of the factor $\Delta x/\Delta Q$ is needed. From the definition of Q and the observation that the distribution of the number of residues inside the pore N_{pore} is sharply peaked around its average $\bar{N}_{\text{pore}}$ (see Fig. S3 of Ref. [22]), it follows that $\Delta Q \simeq k(\bar{N}_{\text{pore}}/L)\Delta x$. Here, Δx is the displacement of the application point of F_x (corresponding to a shift in the pulled residue), with $k = 2$ if the pore is fully occupied (i.e., the protein straddles the pore) or $k = 1$ for a partial occupation (i.e., $N_{\text{left}} = 0$ or $N_{\text{right}} = 0$). Noteworthy, the Langevin dynamics Eq. (1) is robust in catching the stall points of the MD simulations as checked by exploring a range of possible values for the effective friction, importing force, and $\bar{N}_{\text{pore}}$ [see Fig. 1(b) dashed lines for $\tau(Q)$ obtained this way].

IV. STALLING POINTS AND NATIVE STRUCTURE

The sharp correspondence between the $\tau(Q)$ profile and the ascending ramps of $G(Q)$ calls for an interpretation of the step-like free-energy landscape in terms of specific structural features of the protein native state. In particular, the stalls take place in correspondence with residues forming a large number of long-range native contacts with the untranslocated chain portions. The number of such contacts is called here the *backward burial* of the residue. It clearly depends on the pulling direction and for C-pulling reads

$$B_C(i) = \sum_{j=1}^{i-\delta} \Delta_{ij}, \tag{2}$$

where Δ_{ij} is the contact matrix ($\Delta_{ij} = 1$ for residues i and j in native contact and 0 otherwise) and δ skips the closest contacts to include only interactions that are far enough in the sequence ($|i - j| > \delta = 20$). A similar expression holds for the N-pulling backward burial $B_N(i)$.

In order to highlight the connection with the free-energy profile $G(Q)$, it is instrumental to reexpress the backward burials as a function of the collective variable Q. Indeed, as shown in more detail in the Supplemental Material [22], the single-file nature of the translocation makes Q sharply related to the amino acid found at the pore entrance. Given the intrinsic noise in the functions $B_{C/N}$, a smoothed version $\tilde{B}_{C/N}$ is reported in the figures for better readability.

In Fig. 2, the peaks in $\tilde{B}_N(Q)$ and $\tilde{B}_C(Q)$ apparently correspond to the ascending ramps of $G(Q)$. To verify that the correspondence is not accidental, the MPB is "mutated," that is, in line with the Gō-model force-field, we remove the native contacts formed by two of the residues mainly involved in the C-st3 stall. In the bottom panel of Fig. 3, we highlight in red the contacts turned off in the mutated MBP. Mutation entails the lowering of the $G(Q)$-profile slope around the C-st3

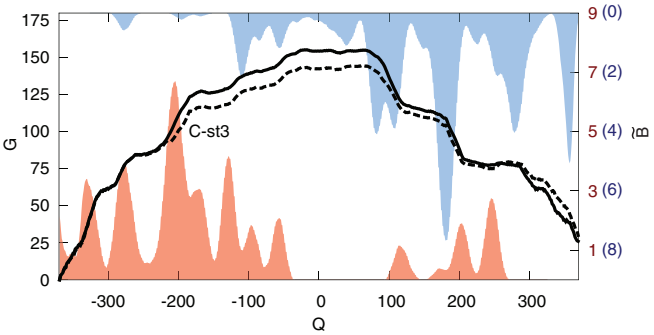

FIG. 2. (Color online) Smoothed backward burials $\tilde{B}_C(Q)$ (red, bottom) and $\tilde{B}_N(Q)$ (blue, top) for C- and N-pulling for MBP. The solid and dashed lines denote the MBP free-energy $G(Q)$ [already plotted in Fig. 1(b)] and the one of its mutant, respectively. For N-pulling simulations, the $\tilde{B}$ scale is reversed (values within parentheses on the right y axes).

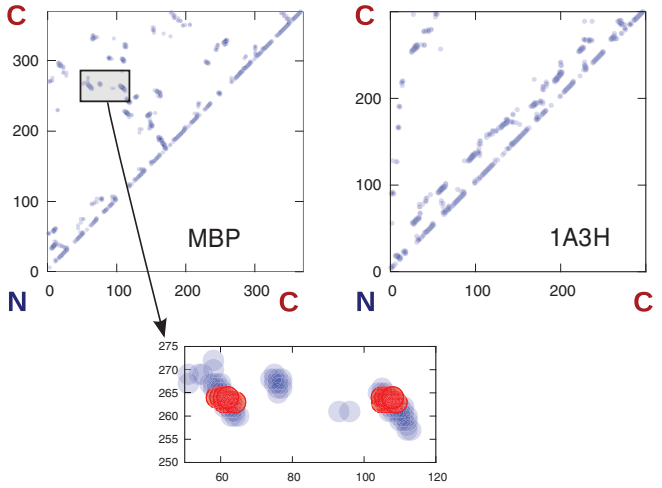

FIG. 3. (Color online) Top: Contact maps. Native interactions are reported in transparent light blue (lighter tone in the printed version) circles (left: MBP, $R_c = 7.5$; right: 1A3H, $R_c = 7.2$). Bottom: The red filled (darker tone in the printed version) circles correspond to the 14 contacts removed in order to obtain the mutated maltose-binding protein.

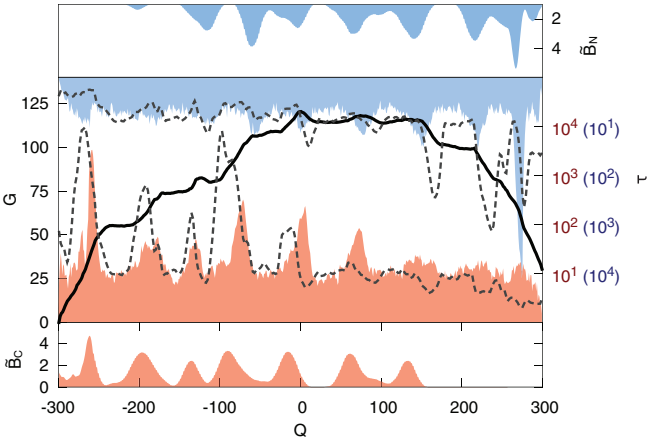

FIG. 4. (Color online) Protein 1A3H. The filled histograms in the central panel represent the average residence time $\tau(Q)$ from MD simulations, while the dashed lines refer to the Langevin dynamics. Upper and lower panels report the smoothed backward burials, $\tilde{B}_C(Q)$ (red, bottom) and $\tilde{B}_N(Q)$ (blue, top). For N-pulling simulations, the $\tau(Q)$ scale in the central panel is reversed (values within parentheses on the right y axes).

stall (dashed line in Fig. 2) and the shortening of the related residence time (Fig. S1 in the Supplemental Material [22]).

The correspondence among average residence time $\tau(Q)$, free-energy profile $G(Q)$, and backward burial $\tilde{B}(Q)$ appears to be a generic feature of protein-like structures, which is mainly determined by the protein topology. Indeed, such a view is supported by addressing the protein 1A3H, a 300-residue long globular protein. Figure 3 (upper panels) reports the contact maps for MBP and 1A3H, where residues in native contact are represented as circles. Such maps reveal the completely different native-state topology of the two proteins. In particular, MBP long-range interactions are more uniformly distributed and are mainly associated with either distal clusters or β sheets, whereas the 1A3H contacts are primarily formed by adjacent α helices, with also a small set of clustered contacts formed by the region close to the N-terminus with the rest of the structure.

Similarly to MBP, the 1A3H-dynamics is hindered by several stalls; the peaks of $\tau(Q)$ and $\tilde{B}(Q)$ closely match, and the larger clusters of contacts are again associated with the ramps in $G(Q)$, Fig. 4. Also in this case, the Langevin dynamics Eq. (1) stalls around the same Q values of the MD simulations. Here, however, the agreement is less satisfactory for the N-pulling process. This partial discrepancy is presumably due to the high critical force [22] of the N-pulling 1A3H translocation. In these conditions, the system is far from the quasistatic limit needed for accurate modeling in terms of a Langevin-like approach based on equilibrium free-energy landscapes.

We repeated the translocation protocol for MBP employing a graphene-like pore [22], narrower and much shorter than the αHL (length, 5 Å; diameter, 10 Å), which roughly reproduces the geometry of the pore recently obtained in Ref. [26]. The stall events occur at the same positions found for the αHL-like pore, with an overall equivalent picture of the process (see Fig. S2 of Ref. [22]). This suggests that the crucial feature responsible for the observed phenomenology is the peculiar and specific unfolding pathway induced by the single-file motion.

Given the strong link between stalls and free-energy profile $G(Q)$, it is natural to wonder about the energetic or entropic nature of the stalls. To answer the question, it is convenient to split $G(Q)$ into its energetic $V_{\rm tot}(Q)$ and entropic $TS(Q) = V_{\rm tot}(Q) - G(Q)$ contributions [Fig. 5(a)]. $V_{\rm tot}(Q)$ is defined

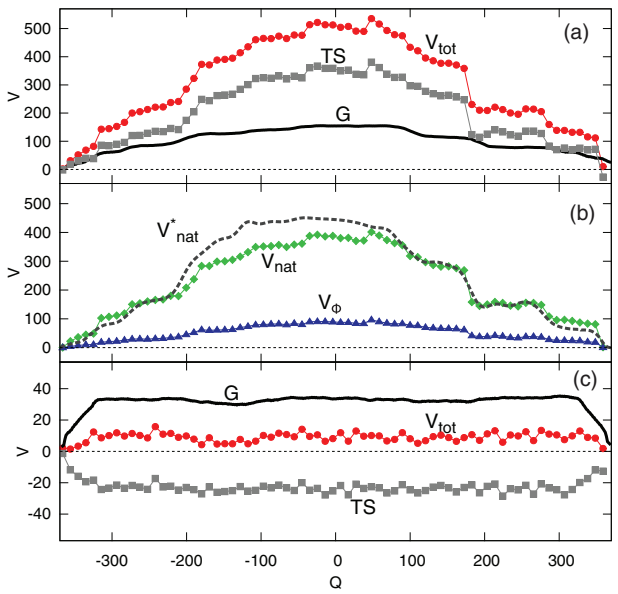

FIG. 5. (Color online) (a) Energetic (red circles) and entropic (gray squares) contributions to the free-energy profile $G(Q)$ (solid line) for wild type MBP. (b) Main contributions to $V_{\rm tot}$: $V_{\rm nat}$ (green diamonds) and V_Φ (blue triangles). The dashed line represent the heuristic estimate of $V^*_{\rm nat}$ – Eq. (4) – based on the *backward burial*. (c) Energetic and entropic contributions of $G(Q)$ for an unstructured polymer.

as the conditional averages over the protein conformations compatible with a selected Q; i.e.,

$$V_{\rm tot}(Q) = \frac{\int d^{3m}r\ V_{\rm tot}(\mathbf{r})\mathrm{e}^{-\beta V_{\rm tot}(\mathbf{r})}\delta[Q - Q(\mathbf{r})]}{\int d^{3m}r\ \mathrm{e}^{-\beta V_{\rm tot}(\mathbf{r})}\delta[Q - Q(\mathbf{r})]}, \qquad (3)$$

where $\mathbf{r} = (\mathbf{r}_1, \ldots, \mathbf{r}_m)$ is the configuration vector. The above conditional average can be taken directly over configurations from all the umbrella sampling runs without the need of reweighting. Indeed, the presence of the $\delta[Q - Q(\mathbf{r})]$ makes Eq. (3) unchanged upon the simultaneous shift $V_{\rm tot}(\mathbf{r}) \to V_{\rm tot}(\mathbf{r}) + V_{\rm umb}[Q(\mathbf{r})]$ in the Boltzmann weights at numerator and denominator; thus, the conditional average $V_{\rm tot}(Q)$ turns to be independent from the umbrella potential $V_{\rm umb}(Q)$. Since translocation implies a gradual unfolding, the energy increment due to contact breakage is accompanied by an entropy increase as the chain explores less compact and more disordered conformations, up to a maximum at $Q \simeq 0$ corresponding to the protein straddling the pore symmetrically, Fig. 5(a). The figure shows that the relevant contribution to G comes from $V_{\rm tot}$, suggesting a main energetic origin of the stalls. Therefore, the further decomposition of $V_{\rm tot}$ into its components is expected to contain crucial information on the stalls. In Fig. 5(b), we plot, as a function of Q, the two dominant contributions to $V_{\rm tot}$, namely, $V_{\rm nat}$ (the potential of the long-range native interactions) and V_Φ (the potential of the dihedral interactions responsible for the correct formation of secondary structures). In Fig. 5(c), we also report the free-energy decomposition into energetic and entropic contributions for an unstructured polymer, obtained by the MBP protein model where all the attractive interactions are removed. In this case, the shape of $G(Q)$ is very simple: symmetric, with a long plateau at the center when the polymer straddles the pore. Moreover, at variance with globular protein translocation, where entropy during the migration increases due to the unfolding, here the entropy decreases. The reason is that, in single-file translocation, the lateral chain fluctuations inside the pore are substantially frozen and, missing the structure, no entropic contribution is produced by unfolding outside the pore. The $V_{\rm nat}$ and V_ϕ contributions are reported in Fig. S5 of Ref. [22] for the sake of completeness.

For the folded portion of the chain, a heuristic argument based on the notion of *backward burial* allows an approximation of the long-range potential, $V_{\rm nat}(Q)$, that works remarkably well [22],

$$V^*_{\rm nat}(Q) = \sum_{q=-m}^{Q} \{\tilde{B}_C(q) - \tilde{B}_N(q)\}. \qquad (4)$$

The state Q is here assumed to be reached from the C-terminus (the summation starts from $q = -m$). The first contribution in Eq. (4) accounts for the number of contacts broken to reach state Q for a left-to-right translocation and the second one accounts for the number of contacts reformed by the translocated part of the chain; see Fig. 6. The same result is obtained by considering N-pull, right-to-left translocations, as shown in the Supplemental Material [22]. $V^*_{\rm nat}$ [dashed line in Fig. 5(b)] quantitatively reproduces the actual data.

Figure 7 reports the decomposition of the free-energy into energetic and entropic parts [Fig. 7(a)] and the main contributions to the energy [Fig. 7(b)] for 1A3H, to be compared with the corresponding plots in Fig. 5 for MBP. Also, in this case, a very good agreement between the estimate Eq. (4) and the actual data is observed.

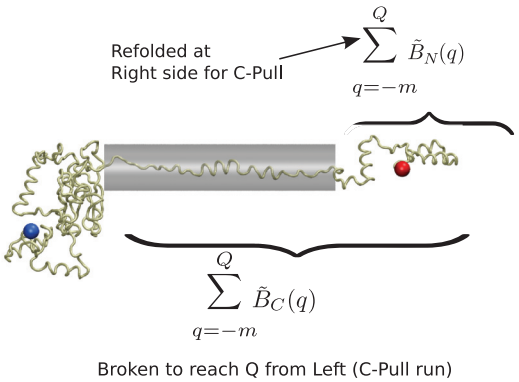

FIG. 6. (Color online) Sketch illustrating the empirical argument used to estimate the $V^*_{\rm nat}$ contribution, Eq. (4). The image refers to a C-pull (Left-to-Right) translocation. In the present exemplification the state Q is reached from the C-terminus ($Q = -m$). The first contribution of Eq. (4) accounts for the number of contacts broken to reach state Q while the second one accounts for the refolding on the right side of the pore.

The above considerations strongly suggest that the bottlenecks of the protein translocation are mainly determined by the resistance to rupture of certain clusters of long-range attractive contacts. This is a further confirmation that the essence of the translocation can be interpreted through the structural properties of the native conformation encoded in the contact map.

V. CONCLUSIONS

In this paper, we have shown, within the framework of coarse-grained native-centric protein modeling, that single-file translocation of a protein-like structure is characterized by stalling events. There is a tight correlation between the geometrical properties of the native structure and the stall

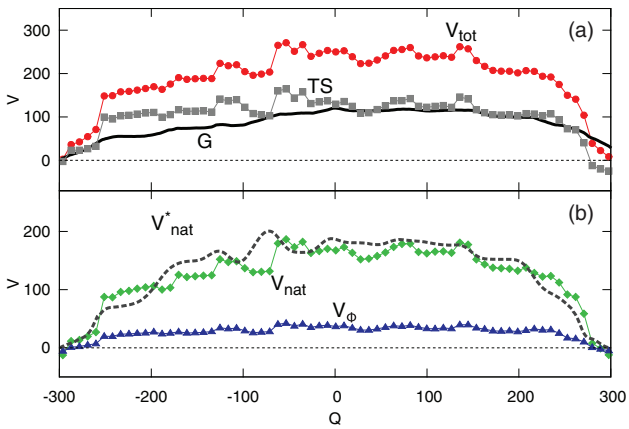

FIG. 7. (Color online) Upper panel. Energetic (red circles) and entropic (gray squares) contributions to the free-energy profile $G(Q)$ for 1A3H protein. Lower panel. Main contributions to $V_{\rm tot}$: $V_{\rm nat}$ (green diamonds) and V_Φ (blue triangles). The dashed line represents the heuristic estimation of $V^*_{\rm nat}$ [Eq. (4)] based on the *backward burial*.

occurrence to suggest that the stall sequence is specific for each protein and constitutes a sort of signature, potentially useful for protein misfolding detection via voltage driven translocation experiments. The identification of the features responsible for the bottlenecks of the transport allows us to develop a heuristic procedure able to estimate the native contribution to the free-energy profile by a summation Eq. (4) upon the *backward burial* trend. This feature appears to be generic and independent of the specific globular protein, possibly opening the way for systematic prescreening of the proteome that could also take advantage of the one-dimensional Langevin approach able to qualitatively reproduce the stall sequence. As a final comment, our results seem to suggest that, to some extent, "structure determines stall patterns," in the sense that a change in the stall pattern is a hallmark of structural modifications.

ACKNOWLEDGMENTS

Computing resources were provided by CASPUR (HPC Grant 2012). F.C. acknowledges the financial support from MIUR, PRIN 2009PYYZM5.

APPENDIX A: THE Gō AND PORE MODELS

The phenomenological off-lattice model proposed by Nobuhiro Gō (minimalist off-lattice native-centric C_α, Gō-model [27]), is a coarse-grained model where the protein is reduced to a sequence of beads of equal masses m_a coinciding with the C_α atoms of the main backbone chain. The model is characterized by a single energy scale ϵ (see below). Several versions and refinements have been suggested in the literature, our paper implements the approach of Ref. [18]. The force field is constituted by four terms: (I) peptide potential (or bond potential), V_p; (II) bending angle potential, V_θ; (III) twist angle potential, V_ϕ; and (IV) nonbonded long-range interaction, $V_{\rm nb}$. The peptide potential V_p reads

$$V_p(r_{i,i+1}) = \frac{k_p}{2}(r_{i,i+1} - R_{i,i+1})^2, \tag{A1}$$

with $r_{i,i+1} = |\mathbf{r}_{i+1} - \mathbf{r}_i|$, $R_{i,i+1} = |\mathbf{R}_{i+1} - \mathbf{R}_i|$, where $\mathbf{R}_i$ and $\mathbf{r}_i$ indicate the position of the ith C_α atom in the native and current conformation, respectively. The spring constant is $k_p = 1000\epsilon/d_m^2$ and $d_m = 3.8$ Å (average distance between two consecutive residues). The angular bending potential V_θ reads

$$V_\theta(\theta_i) = \tfrac{1}{2}k_\theta(\theta_i - \Theta_i)^2, \tag{A2}$$

where $k_\theta = 20\epsilon$ rad^{-2} and Θ_i, θ_i are the bond angles formed by three consecutive beads in the native and current conformations, respectively. Since k_θ is stiff, θ_i undergoes small fluctuation around its native value Θ_i. The dihedral potential V_ϕ is a function of the twist angles Φ_i and ϕ_i (i.e., the angle formed between the two planes determined by four consecutive amino acids along the chain, in native and current conformations), and it reads

$$V_\phi(\phi_i) = k_\phi^{(1)}[1 - \cos(\phi_i - \Phi_i)] + k_\phi^{(3)}[1 - \cos 3(\phi_i - \Phi_i)], \tag{A3}$$

where $k_\phi^{(1)} = \epsilon$ and $k_\phi^{(3)} = \epsilon/2$ are the dihedral constants.

Finally, the nonbonded (long-range) potential $V_{\rm nb}$ includes the pair-wise interaction selected to promote the native-like interactions found in the Protein Data Bank (PDB) structure. More specifically, residues i and j (with $|i - j| \geqslant 3$) attract each other via the 12-10 Lennard-Jones potential when they are considered in native contact. Otherwise, they repel one another with an excluded volume effect. Two residues are considered to be in native contact when their distance R_{ij} in the PDB structure is lower than a chosen cutoff radius R_c; thus, the pair-wise potential is

$$V_{\rm nb}(r_{ij}) = \epsilon \begin{cases} 5\left(\frac{R_{ij}}{r_{ij}}\right)^{12} - 6\left(\frac{R_{ij}}{r_{ij}}\right)^{10} & R_{ij} < R_c \\ \frac{10}{3}\left(\frac{\sigma}{r_{ij}}\right)^{12} & R_{ij} > R_c, \end{cases} \tag{A4}$$

where $\sigma = 4.5$ Å is a parameter defining the excluded volume of each residue.

The global potential acting on all the m residues of the protein is then

$$V_{\rm G\bar{o}} = \sum_{i=1}^{m-1} V_p(r_{i,i+1}) + \sum_{i=1}^{m-2} V_\theta(\theta_i) + \sum_{i=1}^{m-3} V_\phi(\phi_i) + \sum_{i,j\geqslant i+3} V_{\rm nb}(r_{ij}). \tag{A5}$$

The values of the parameters reported above are the typical ones used in similar Gō-like schemes, see, e.g., Refs. [18,28,29].

The confinement of the nanopore is described as a step-like soft-core repulsive cylindrical potential acting on the protein, where the axis of symmetry is taken as the x axis of the reference frame,

$$V_{\rm pore}(x,y,z) = V_0\left(\frac{y^2+z^2}{R_p^2}\right)^q \Theta[x(L-x)]. \tag{A6}$$

Here, $\Theta(s) = [1 + \tanh(\alpha s)]/2$ is a smooth steplike function limiting the action of the pore potential in the effective region $[0,L]$. L and R_p are pore length and radius, respectively. A convenient choice of the other parameters is $q = 1$, $\alpha = 3$ Å^{-2}, and $V_0 = 2\epsilon$, [21].

Therefore, the overall potential a residue is subjected to is expressed as

$$V_{\rm tot} = V_{\rm G\bar{o}} + \sum_{i=1}^{m} V_{\rm pore}(r_i). \tag{A7}$$

The unit system employed in the paper is specified in terms of the intrinsic scales of the coarse-grained model. Specifically, lengths are given in Å, while energy and mass are expressed as multiples of ϵ and m_a, which parametrize energy and mass, respectively. For the sake of definiteness, we mention here that all our simulations are run at $k_BT = 0.75$. All these units can be in principle converted to conventional ones. To this purpose, specific experimental data are needed to set the energy scale. As reported, e.g., in Refs. [21,25], thermal unfolding simulations can be performed to determine the unfolding temperature T_u in code units, then the energy scale ϵ in physical units is set by matching the computational and experimental unfolding temperatures. The knowledge of ϵ allows also the conversion of the code time unit into seconds. For the case of

MBP, where experimental thermal unfolding data are available, we performed this calculation in a previous paper [21], with a slightly different value of the cut-off radius R_c. This led to a value of the time unit of ∼10 ps, implying that a translocation occurring in 10^4–10^5 time units (as in the present work) would correspond to about 1 μs, i.e., more than one order of magnitude faster than the fastest experimentally observed translocations. This is a general feature of coarse-grained methods that typically do not reproduce actual time scales. In such conditions, the conversion of the code time unit into seconds does not provide further insights into the physics of the translocation. Indeed, the power of coarse-grained methods lies on their ability to describe significant qualitative features of the translocations, like stalling dynamics, which are still beyond the capabilities of the full-atom techniques. We refer the interested reader to Ref. [30] for a thorough discussion about coarse-grained approaches on polymers and colloids.

APPENDIX B: UMBRELLA SAMPLING

The free-energy profile $G(Q) = -k_B T \ln P(Q)$, where $P(Q)$ is the probability to find the protein in conformations characterized by a given value of the collective variable Q, is obtained via umbrella sampling combined with the multiple weighted histogram analysis method (WHAM) [24]. Here we provide a description of the parameters used in the procedure. We found convenient defining the continuous version of of the collective variable, $Q = N_{\text{right}} - N_{\text{left}}$ introduced in the text,

$$\bar{Q}(x_1, \ldots, x_m) = \frac{1}{2} \sum_{i=1}^{m} \{\tanh(a x_i) + \tanh[a(x_i - L)]\}, \tag{B1}$$

with x_i being the axial coordinate of the ith residue, m the number of residues of the protein, L the pore length, and $a = 3$ Å^{-1} a smoothing parameter.

The umbrella potential,

$$V_{\text{umb}}(x_1, \ldots, x_m) = \tfrac{1}{2} k_u [\bar{Q}(x_1, \ldots, x_m) - Q_w]^2, \tag{B2}$$

is superimposed to the system Hamiltonian in order to restrain the dynamics around the target value of the collective variable Q_w. The presence of the umbrella potential is clearly enhancing the exploration of just those Q states with low probability. The system evolves under the combined potential $V = V_{\text{tot}} + V_{\text{umb}}$ to estimate the probability density (pdf) $P_{\text{umb}}(Q)$ to find the biased system in configurations with Q values around Q_w. The unbiased probability $P(Q)$, i.e., without the umbrella potential, is recovered by the reweighting

$$P(Q) = P_{\text{umb}}(Q) e^{-\beta k_u / 2 (Q - Q_w)^2} Z_{\text{umb}} / Z,$$

where Z and Z_{umb} are the partition function of the original and the biased system, respectively.

To reconstruct the free-energy profile all over the pore size by the WHAM algorithm, we select a set of 200 umbrella windows equally spaced in the interval $[-m,m]$ centered around different values of Q_w ($w = 1, \ldots, 200$). We find it convenient to introduce, in Eq. (B2), a dependence of the type $k_u = k_u(Q_w)$ to reduce the number of windows while maintain a reasonable overlap between adjacent histograms. The elastic constant of the umbrella potential covers the range $k_u \in [0.22, 2]$; the highest values are chosen near the pore ends, where the protein can be too easily lost in the bulk under small thermal fluctuations.

Input configurations for the umbrella sampling runs were extracted among the conformations of translocation simulations (i.e., in the presence of the importing force) with the nearest Q to the window center, Q_w. Each simulation is run for a time suited to collect uncorrelated statistics, 10^3 points for each histogram with a decorrelation time equal to 150 internal time units (the latter being determined by preliminary simulations). Moreover, the first 10^4 time units (equal to 10% of the translocation simulation time window) are discarded for thermalization and to allow the possible refolding at the trans side. Since analogous results were obtained by using only one-half of the sampled data, the statistics was assumed to have reached convergence. The histograms that were collected from the biased simulations were finally combined with optimal weights according to the WHAM method [24] to reconstruct the free-energy profile $G(Q)$ that minimizes the resulting statistical error.

Finally, the statistical quality of the $G(Q)$ profile has been enhanced by combining half of the umbrella sampling runs obtained from C-terminus pulling initial conditions ($Q_w < 0$) with the complementary simulations from the N-terminus case ($Q_w > 0$).

[1] R. DeBlois and C. Bean, Rev. Sci. Instrum. **41**, 909 (1970).
[2] G. Schneider and C. Dekker, Nat. Biotechnol. **30**, 326 (2012).
[3] D. Rotem, L. Jayasinghe, M. Salichou, and H. Bayley, J. Am. Chem. Soc. **134**, 2781 (2012).
[4] L. Huang and D. Makarov, J. Chem. Phys. **129**, 121107 (2008).
[5] G. Oukhaled, J. Mathe, A. L. Biance, L. Bacri, J. M. Betton, D. Lairez, J. Pelta, and L. Auvray, Phys. Rev. Lett. **98**, 158101 (2007).
[6] A. G. Oukhaled, A. L. Biance, J. Pelta, L. Auvray, and L. Bacri, Phys. Rev. Lett. **108**, 88104 (2012).
[7] C. Madampage, O. Tavassoly, C. Christensen, M. Kumari, and J. Lee, Prion **6**, 110 (2012).
[8] D. Talaga and J. Li, J. Am. Chem. Soc. **131**, 9287 (2009).
[9] B. Cressiot, A. Oukhaled, G. Patriarche, M. Pastoriza-Gallego, J. Betton, L. Auvray, M. Muthukumar, L. Bacri, and J. Pelta, ACS nano **6**, 6236 (2012).
[10] M. Mohammad, R. Iyer, K. Howard, M. McPike, P. Borer, and L. Movileanu, J. Am. Chem. Soc. **134**, 9521 (2012).
[11] O. Tavassoly and J. S. Lee, FEBS Lett. **586**, 3222 (2012).
[12] B. Krasniqi and J. S. Lee, Metallomics **4**, 539 (2012).
[13] J. Nivala, D. B. Marks, and M. Akeson, Nat. Biotechnol. **31**, 247 (2013).
[14] D. Rodriguez-Larrea and H. Bayley, Nat. Nanotechnol. **8**, 288 (2013).
[15] M. Bacci, M. Chinappi, C. Casciola, and F. Cecconi, J. Phys. Chem. B **116**, 4255 (2012).

[16] H. W. de Haan and G. W. Slater, Phys. Rev. Lett. **110**, 048101 (2013).
[17] G. Davies, M. Dauter, A. Brzozowski, M. Bjørnvad, K. Andersen, and M. Schülein, Biochemistry **37**, 1926 (1998).
[18] C. Clementi *et al*., J. Mol. Biol. **298**, 937 (2000).
[19] C. Clementi, Curr. Opin. Struct. Biol. **18**, 10 (2008).
[20] O. K. Dudko, T. G. W. Graham, and R. B. Best, Phys. Rev. Lett. **107**, 208301 (2011).
[21] M. Chinappi, F. Cecconi, and C. M. Casciola, Philos. Mag. **91**, 2034 (2011).
[22] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.88.022712 for details on simulation protocol and additional data for N_{pore} statistics, contact maps, residence time, and $G(Q)$ contribution and stalls in graphene-like pores.
[23] T. Schlick, *Molecular Modeling and Simulation: An Interdisciplinary Guide* (Springer, Berlin, 2010), Vol. 21.
[24] B. Roux, Comput. Phys. Commun. **91**, 275 (1995).
[25] A. Ammenti, F. Cecconi, U. Marini Bettolo Marconi, and A. Vulpiani, J. Phys. Chem. B **113**, 10348 (2009).
[26] C. J. Russo and J. Golovchenko, Proc. Natl. Acad. Sci. USA **109**, 5953 (2012).
[27] N. Gō and H. A. Scheraga, Macromolecules **9**, 535 (1976).
[28] T. X. Hoang and M. Cieplak, J. Chem. Phys. **112**, 6851 (2000).
[29] F. Cecconi, P. De Los Rios, and F. Piazza, J. Phys. Chem. B **111**, 11057 (2007).
[30] J. T. Padding and A. A. Louis, Phys. Rev. E **74**, 031402 (2006).