

Translocation of structured polynucleotides through nanopores

Ulrich Gerland^{1,3}, Ralf Bundschuh² and Terence Hwa¹

¹ Department of Physics and Center for Theoretical Biological Physics, University of California at San Diego, La Jolla, CA 92093-0319, USA

² Department of Physics, The Ohio State University, 174 W 18th Av., Columbus, OH 43210-1106, USA

E-mail: Ulrich.Gerland@physik.uni-muenchen.de

Received 26 November 2003

Accepted for publication 13 January 2004

Published 12 February 2004

Online at stacks.iop.org/PhysBio/1/19 (DOI: 10.1088/1478-3967/1/1/002)

Abstract

We investigate theoretically the translocation of structured RNA/DNA molecules through narrow pores which allow single but not double strands to pass. The unzipping of basepaired regions within the molecules presents significant kinetic barriers for the translocation process. We show that this circumstance may be exploited to determine the full basepairing pattern of polynucleotides, including RNA pseudoknots. The crucial requirement is that the translocation dynamics (i.e. the length of the translocated molecular segment) needs to be recorded as a function of time with a spatial resolution of a few nucleotides. This could be achieved, for instance, by applying a mechanical driving force for translocation and recording force-extension curves (FECs) with a device such as an atomic force microscope or optical tweezers. Our analysis suggests that, with this added spatial resolution, nanopores could be transformed into a powerful experimental tool to study the folding of nucleic acids.

1. Introduction

A series of recent experiments studied the translocation of DNA and RNA molecules through narrow pores, which allow single but not double strands to pass [1–8], see [9] for a review. These investigations pursued two main goals: (i) to probe in a well-defined model system the physics of biopolymer translocation across membranes, a process which is ubiquitous in cell biology, and (ii) to explore the potential of nanopores as a single-molecule tool. In the experiments so far, a membrane protein, α -hemolysin, was used as the pore. An electric field acting on the negatively charged DNA/RNA backbone drives the molecules through the pore, and translocation is monitored by measuring the induced ionic current, which is strongly reduced while a DNA/RNA chain blocks the pore. Until very recently [5, 8], the experiments have focused on the translocation of unstructured, mostly homopolymeric molecules, a problem which has also received considerable theoretical interest [10–18]. For such unstructured molecules,

the main results regarding the above goals were that (i) the basic physics of translocation is adequately described by a drift-diffusion process, in which monomers hop randomly in and out of the pore with a directional bias due to the applied voltage [13], and (ii) nanopores could possibly be developed into rapid sequencing devices, since the ionic current during blockage displays a weak sequence-dependence [2, 3].

In contrast, for structured polynucleotides, both the basic physics and the potential applications of translocation still remain largely unexplored. Experimentally, important first steps have been taken by studying the translocation of simple hairpin (i.e. stem-loop) structures [5] and the unzipping of double-stranded DNA through a nanopore [8]. However, a general theoretical framework to describe translocation of these as well as more complex RNA/DNA structures is currently lacking. Here, we first construct such a framework and then use it to investigate the potential of nanopores as a single-molecule tool for the study of biopolymer folding.

In this paper, we are interested in the generic physical aspects of the translocation process that depend neither on the specific properties of a particular protein pore, nor on the

³ Present address: Physics Department, Ludwig-Maximilians-Universität, Theresienstr. 37, 80333 München, Germany.

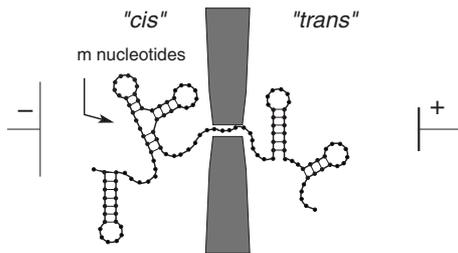


Figure 1. Sketch of a structured polynucleotide that is driven across a nanopore which allows single but not double strands to pass. Here, the driving force causing translocation from the cis to the trans side is exerted by an electric field that acts on the negatively charged backbone of the molecule.

detailed way in which the driving force for translocation is applied. As in previous theoretical studies [10–16], we use a coarse-grained model which treats the pore basically as a separator between a cis and a trans part of the molecule with a characteristic friction coefficient; see the sketch in figure 1. Presumably this description will apply directly to solid-state nanopores [19, 20], which can now be fabricated with sizes down to ~ 2 nm, not much larger than the ~ 1.5 nm aperture of the α -hemolysin pore and slightly smaller than the ~ 2.2 nm diameter of double-stranded DNA or stems in RNA. Also, we do not consider the full three-dimensional (tertiary) structure of the molecules, but focus on the basepairing pattern, i.e. the secondary structure including possible pseudoknots, which are the only structural features present when there are no divalent metal ions in the solution. Unless stated otherwise, the term ‘structure’ refers here to this basepairing pattern. While both our theoretical framework and our conclusions apply equally to RNA and single-stranded DNA, the RNA case is particularly interesting, since structured RNAs have a multitude of functions in molecular biology and RNA folding is an active field of research [21–24].

2. General theoretical framework

Figure 1 depicts schematically the driven translocation of a structured polynucleotide from the cis to the trans side of the pore. We seek here a convenient reduced description of this translocation process, rather than modeling the full three-dimensional polymer dynamics explicitly. Our approach is similar in spirit to the existing models for the case of unstructured polymers [10, 12, 13], where the translocation dynamics is formulated in terms of a single variable, e.g. the number of nucleotides, m , on the cis side, see figure 1. The dynamics, $m(t)$, is stochastic and can be described by ‘hopping rates’, $k_-(m)$ and $k_+(m)$, for forward and backward motion of the nucleotide chain through the pore with a stepsize of one monomer. The external force on the molecule leads to an imbalance in the hopping rates, $k_-(m) > k_+(m)$, and hence a mean drift towards the trans side. For unstructured molecules, the one-dimensional description is permissible if the relaxation of the polymer degrees of freedom on both sides of the pore is faster than the hopping process. This assumption does

not hold for arbitrarily long polymers, since the relaxation time increases with the polymer length [13, 15]; however, for lengths of the order of a thousand bases, the one-dimensional description is adequate under typical experimental conditions [13]. The residual effect of the polymer ends is then only to introduce an entropic barrier for translocation, which leads to a weak m -dependence of the hopping rates.

For structured molecules, the translocation dynamics is considerably more complicated, since the dynamics of the ‘reaction coordinate’, $m(t)$, is then coupled to the dynamics of the basepairing patterns on both sides: the structure on the cis side, $\mathcal{S}_{\text{cis}}(t)$, affects the forward rate, while the structure on the trans side, $\mathcal{S}_{\text{trans}}(t)$, affects the backward rate,

$$m \xrightarrow{k_-(m, \mathcal{S}_{\text{cis}}(t))} m-1 \quad m \xrightarrow{k_+(m, \mathcal{S}_{\text{trans}}(t))} m+1. \quad (1)$$

In two limiting cases, however, the process can be modeled by a one-dimensional Brownian walk as for unstructured molecules, but with a complex sequence/structure-dependent free energy landscape $\mathcal{F}(m)$ along the coordinate m . (A) If the dynamics of the basepairing patterns $\mathcal{S}_{\text{cis}}(t)$ and $\mathcal{S}_{\text{trans}}(t)$ is much faster than the hopping process, the landscape is determined by the ensemble free energy of all basepairing patterns on the cis and trans side. (B) In the opposite limit, the basepairing pattern on the cis side is essentially frozen and is unzipped basepair by basepair as it is driven through the pore. The landscape is then determined by the basepairing energetics of the particular molecular structure prior to translocation, see below. In both cases, the free energy naturally decomposes into three parts,

$$\mathcal{F}(m) = \mathcal{F}_{\text{cis}}(m) + \mathcal{F}_{\text{trans}}(m) + \mathcal{F}_{\text{ext}}(m), \quad (2)$$

where $\mathcal{F}_{\text{cis}}(m)$ and $\mathcal{F}_{\text{trans}}(m)$ denote the intrinsic binding free energies of the cis and trans parts of the molecule, while $\mathcal{F}_{\text{ext}}(m)$ describes the effect of the external force. Given $\mathcal{F}(m)$, the simplest form for the hopping rates $k_{\pm}(m)$ which satisfies the detailed balance condition $k_+(m)/k_-(m+1) = e^{-\beta[\mathcal{F}(m+1) - \mathcal{F}(m)]}$ (with $\beta = 1/k_B T$) is

$$k_{\pm}(m) = k_0 e^{-\beta \cdot \max\{\mathcal{F}(m \pm 1) - \mathcal{F}(m), 0\}}. \quad (3)$$

Here, k_0 denotes a microscopic rate constant, which can in principle be tuned by adjusting the properties of the pore. It can be interpreted as a friction coefficient and corresponds approximately to the bare hopping rate for unstructured molecules at zero external force (typical experimental estimates for k_0 are on the order of 10^5 s^{-1} [6]). The dynamics of the translocation process, as described by equations (2) and (3) is dominated by energetic barriers due to basepairing, whereas the above-mentioned entropic barrier is completely negligible for structured molecules. These energetic barriers lead to arrests during translocation, as clearly observed already in the experiments with simple hairpins [5] and double-stranded DNA [8].

3. Pulling through a pore

3.1. Qualitative aspects

We now make use of the theoretical framework constructed above to investigate which information on structured

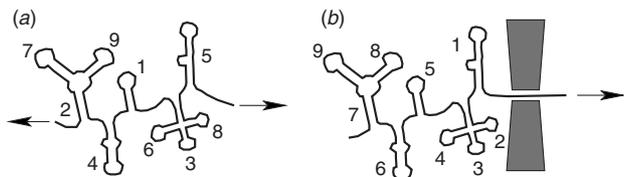


Figure 2. Unzipping a structured molecule by pulling on its ends is fundamentally different from unzipping by driven translocation through a narrow pore. (a) For pulling on the ends, the stems (i.e., contiguously basepaired segments) in the molecule unfold in an order determined by their relative stability and the topology of the structure (a possible order 1–9 is indicated). (b) In contrast, the pore forces the stems to unfold in a linear order along the sequence, as again indicated by the numbering 1–9.

molecules could be derived from pore translocation experiments. To this end, it is useful to compare unzipping by driven translocation through a nanopore with the more conventional method of unzipping by applying a force on the ends of a biopolymer, see e.g. [25–27]. As illustrated in figure 2, the two approaches differ fundamentally: pulling on the ends induces a spontaneous unfolding order for the individual structural elements, which is a function of their relative stabilities and the topology of the structure. In contrast, the nanopore prescribes a linear order along the sequence, and unfolds an RNA molecule much as enzymes such as the ribosome do in cells. This difference suggests that the two approaches can also yield different types of information about the molecule under study. As demonstrated by Onoa *et al* [27], clever use of the pulling on the ends approach can reveal detailed information on the (un)folding pathway of an RNA molecule with known structure. However, when the structure of an RNA molecule is unknown, pulling on the ends can provide, by itself, little information beyond a count of the number of structural elements that unfold separately [27, 28]. In the following we therefore focus on the question of how much structural information may in principle be obtained with the nanopore approach.

Let us suppose that we were able to observe the trajectories $m(t)$ of the molecules during the translocation process. We could then assign a position within the sequence to each arrest during translocation. Since an arrest is caused by a kinetic barrier, i.e. a stem trapped at the entrance to the pore, we could thereby identify the positions of the stems in the structure. Such information can indeed be sufficient to reconstruct almost the entire basepairing pattern of a molecule, as we demonstrate explicitly using an example below. If the translocation dynamics is in the strongly driven limit (B) where the structure on the cis side is essentially frozen, then the reconstructed structure would correspond to the initial structure of the molecule before translocation. We concentrate on this limit in the following, including a discussion of its attainability. However, it may be noteworthy that in the slow translocation limit (A) one would also obtain useful structural information, namely on the average structure of the molecule (with respect to the thermodynamic ensemble of all structures [29]). As long as the molecule is ‘well-designed’

this average structure will be dominated by the ground-state, i.e. the minimum binding free energy structure⁴.

How could one possibly observe the trajectories $m(t)$ during translocation? For the purpose of structure determination, we will need $m(t)$ with a spatial resolution below the typical length of a stem in an RNA structure (5–10 basepairs). This may be achievable through a refinement of the current nanopore technology, such that careful analysis of the ionic current allows a count (or even sequencing) of the bases that have passed the pore [2, 3]. With artificial solid-state pores [19, 20] it is also conceivable to use a tunneling current through leads within the membrane as a probe to count (or sequence) the bases as they pass through the pore. Here, we explore yet another option, namely pulling the molecule mechanically through the pore, with a device that can record force-extension curves, e.g. an atomic force microscope or optical tweezers. The explicit discussion of this case with an exemplary RNA sequence serves us to gauge the more general capability of nanopores as single-molecule tools for the study of biopolymer folding.

3.2. Quantitative aspects

Mechanical unfolding of a biopolymer yields characteristic sawtooth-shaped signatures in the force-extension curve (FEC) indicating the opening of structural elements within the molecule, see e.g. [25, 27]. From the relative positions of these sawteeth one can determine length changes within the molecule with an extremely high resolution of about 1 nm. In the usual setup where the molecule is unfolded by pulling on its ends, such length changes can only be used to infer the ‘stored length’ of a structural element, but not its precise position along the backbone of the molecule, cf figure 2. In contrast, for mechanical pulling through a pore, the relative positions of the resulting sawteeth will correspond directly to the relative positions of the structural elements in the sequence⁵. One conceivable way to prepare the initial condition where an RNA molecule is almost entirely on the cis side, with one end threaded through the pore and attached to a pulling device on the trans side, is to start with an attached molecule on the trans side and to apply a voltage pulse across the pore that suffices to drive the molecule as far as possible to the cis side.

To apply our general model to the particular case of mechanical pulling in the strongly driven (fast pulling) limit, we need to specify the form of the three terms in the free energy landscape (2). The second term, i.e. the binding free energy on the trans side, may be set to zero,

$$\mathcal{F}_{\text{trans}}(m) = 0, \quad (4)$$

⁴ The worst case for the purpose of structure determination corresponds to the regime where the typical timescale for the translocation of, say, a single hairpin is comparable to the timescale for structural rearrangements involving the formation of new stems: in this case, the structure on the cis side may relax after a stem is unzipped, so that one would observe only the signatures of the relaxed structure rather than the original structure. This regime should be avoided by a proper choice of the driving force and the friction coefficient of the pore (Bundschuh and Gerland, to be published).

⁵ The absolute position can be inferred by adding a known structural element, e.g. a strong C-G hairpin, to one end of the RNA, which can then function as a reference point.

since the reformation of structure after translocation is suppressed at high tensions in the RNA single strand⁶. The third term, $\mathcal{F}_{\text{ext}}(m)$, describes the effect of the mechanical stress on the RNA, which stretches the single-stranded trans part of the molecule. The elastic response of this single-strand may be modeled by a freely jointed chain (FJC) polymer model. Assuming for simplicity a constant pulling speed v , the third term then takes the form

$$\mathcal{F}_{\text{ext}}(m) = \mathcal{F}_{\text{FJC+spring}}(v \cdot t; N - m). \quad (5)$$

Here, the function $\mathcal{F}_{\text{FJC+spring}}(R_t; n)$ denotes the combined free energy of a single-stranded RNA of n bases in series with a linear spring, stretched to a total extension $R_t = v \cdot t$ [28]. (The linear spring takes into account the stiffness of the force-measuring device; see the appendix for details.) By assumption, the first term, $\mathcal{F}_{\text{cis}}(m)$, represents the binding free energy of the remaining part of the initial structure on the cis side. $\mathcal{F}_{\text{cis}}(m)$ can be calculated for any initial structure, based on the free energy rules for RNA secondary structure [30] with a natural extension for pseudoknotted structures; see the appendix. Our assumption of a frozen structure on the cis side is most likely an oversimplification for realistic pulling speeds, since small fluctuations in the secondary structure are known to occur already on timescales of the order of tens of microseconds [31]. However, since the pore pulling approach is sensitive only to stem positions, we expect that it is unaffected by small fluctuations and sensitive only to major rearrangements which significantly change the secondary structure. Such rearrangements are typically slow, sometimes even on the timescale of hours [32, 33].

3.3. Reconstruction of secondary structures

To illustrate the problem and the method, we use an exemplary RNA, the well-studied self-splicing intron of *Tetrahymena thermophila* [21] with a sequence of 419 bases (Genbank # V01416). In its correctly folded active state, the basepairing pattern of this ribozyme contains a pseudoknot (see figure 3(a)), while its best characterized long-lived folding intermediate [32, 34] has a known alternative structure without pseudoknot (see figure 3(b)) [32]. We will investigate whether one can in principle use the pulling-through-a-pore approach not only to discriminate between these two different conformations in individual molecules, but also to reconstruct both structures from the FECs.

To obtain FECs for these structures, we performed Monte Carlo simulations of the stochastic process defined by equations (1)–(3), and used equation (6) from the appendix to calculate the force and extension time traces. We performed all calculations at the same pulling speed ($v = 0.1$ nm/time step, which roughly corresponds to $10 \mu\text{m s}^{-1}$ given typical values for k_0 , see above), and the same stiffness of the force-measuring device ($\lambda = 0.5$ pN nm⁻¹). Figure 4 displays

⁶ For instance, Liphardt *et al* [26] observed refolding rates for a single hairpin around 1 s^{-1} at the unfolding force $f_{1/2} \approx 14$ pN. At a pulling speed of say $1 \mu\text{m s}^{-1}$, the translocation of an RNA molecule with a thousand bases would therefore be terminated before refolding of a structural element on the trans side occurs.

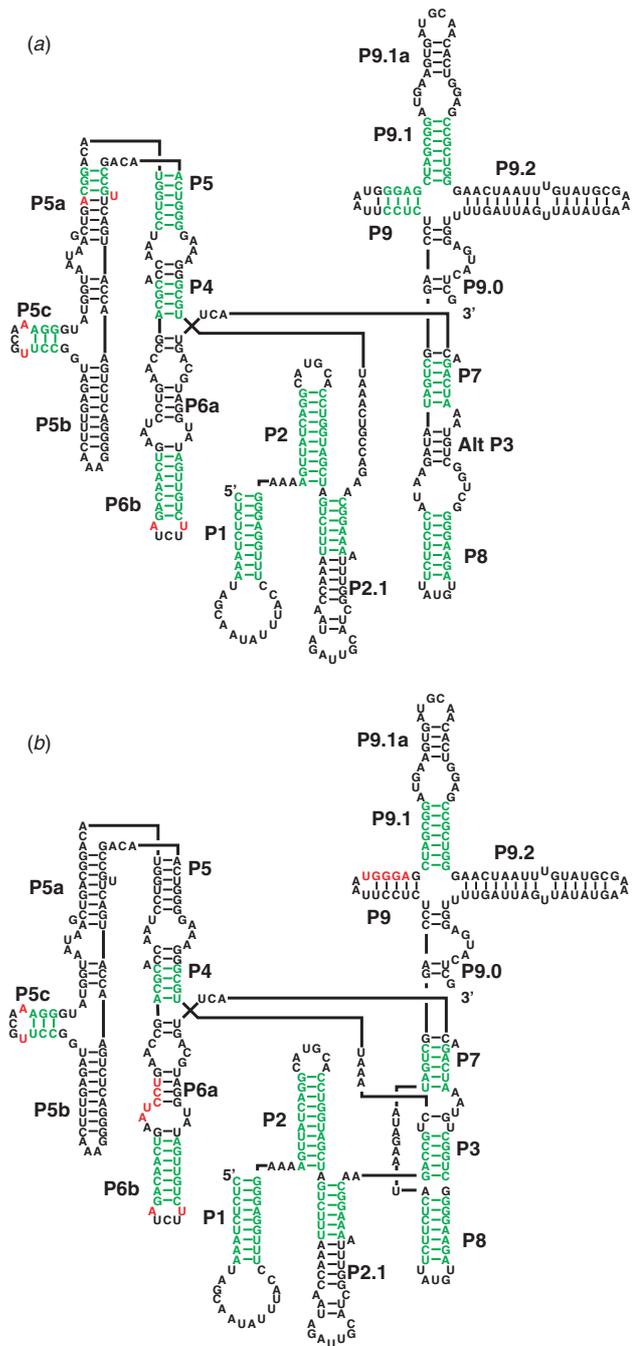


Figure 3. Secondary structure of the *Tetrahymena thermophila* Group I intron: (a) Long-lived folding intermediate [32]. (b) Native state with pseudoknot. The basepairs shown in green are correctly reconstructed from the force-extension curves, see figure 4, using the procedure described in the main text, while the bases shown in red are involved in incorrect basepair predictions (the procedure yields no prediction for the bases shown in black); see also figure 5.

three such FECs (solid lines) for each of the two structures of figure 3, corresponding to unzipping from the 3' end. These FECs show the sawtooth-like behavior which is characteristic for the sequential opening of structural elements (a very similar behavior was observed in the experiments of Onoa *et al* [27]

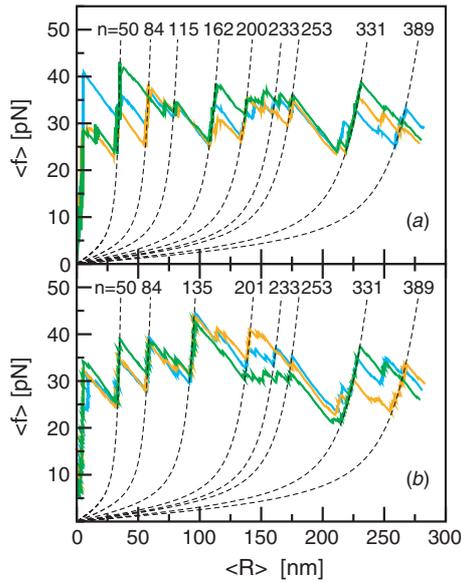


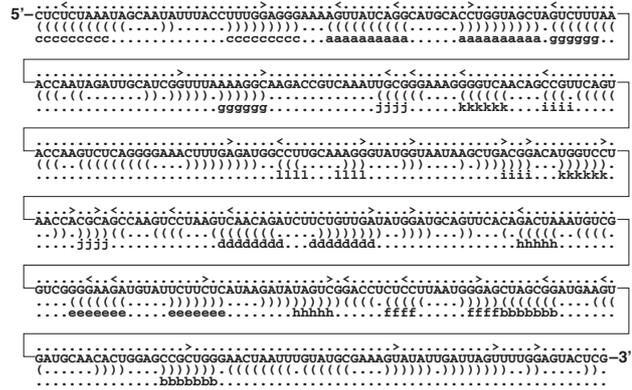
Figure 4. Force-extension traces (solid lines) as obtained with our stochastic model for mechanical pulling through a nanopore. (a) and (b) each show three different runs with the same initial conditions (and pulling speed of $v = 0.1$ nm/time step) for the structures in figure 3(a) and (b), respectively. The force $\langle f \rangle$ and extension $\langle R \rangle$ are calculated using equation (6) in the appendix. The dashed lines are freely jointed chain FECs whose lengths are fitted to some of the positions that correspond to translocation arrests.

where the molecule was rapidly unzipped by pulling on its ends). The rising parts of the sawteeth correspond to stretching of single strand on the trans side as a stacked region is ‘trapped’ in front of the pore on the cis side. When a stacked region opens, some single strand is freed to pass the pore, which leads to relaxation of the tension and causes the downstrokes in the FECs. Note that the FECs do not share all of their sawteeth, which reflects the importance of thermal fluctuations for this type of single-molecule experiments (this property is manifest also in the experiment of Onoa *et al* [27]).

The most relevant information contained in the FECs are the positions of the translocation arrests, during which the required force for the opening of basepairs is built up. To extract these positions, we use FECs of freely jointed chains with different lengths. The dashed lines in figure 4 show some examples of such FECs where the chain length n coincides with the length of the RNA single strand on the trans side during such an arrest. With an automated procedure described in the appendix we obtain all of these positions (above a threshold for the duration of an arrest).

Since the bases around the position of an arrest are very likely basepaired with another segment of the sequence further to the 5′ end, we represent this information by a closing angular bracket, ‘)’, above that position in the RNA sequence (written from 5′ to 3′), see figure 5. Of course, the molecule can also be pulled through the pore in the other direction, i.e. from the 5′ end. This yields information on the positions of segments that have downstream binding partners. The same procedure then leads to the opening brackets, ‘(’, also shown in figure 5.

(A) Non-pseudoknotted intermediate state:



(B) Pseudoknotted native state:

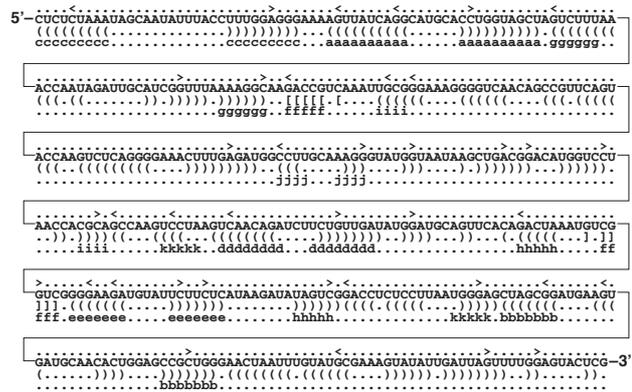


Figure 5. Reconstruction of the basepairing pattern from the FECs. First row: parentheses extracted from the FECs, which indicate the position of basepaired regions. Second row: RNA sequence. Third row: parentheses indicating the basepairs in the full structures shown in figure 3. Fourth row: stems predicted from the parentheses in the first row by sequence alignment. See main text for details.

Bracket representations are a widely used short hand notation for RNA secondary structures. For the structures in figure 3, such a representation is shown in the third row of figure 5. Note that two types of brackets have to be used for the pseudoknotted native structure, in order to make the association between opening and closing brackets unambiguous. We observe that the angular brackets extracted from the FECs can be viewed as an incomplete bracket representation of the RNA secondary structure. Can we complete it using only the given sequence of the RNA molecule?

This task is a sequence alignment problem, which consists of matching each opening (closing) bracket with an associated downstream (upstream) binding sequence. Several circumstances conspire to make this, somewhat surprisingly, a nontrivial problem: (i) stems, i.e. contiguous basepaired regions, are usually short, typically 5–10 basepairs, (ii) structural elements often lead to a different number of angular brackets in the two pulling directions, i.e. not every opening bracket has a corresponding closing bracket and *vice versa*, and (iii) sequence segments containing several U’s have

many possible binding partners, since U's can pair with A's and G's.

To overcome this problem, we developed a probabilistic sequence alignment algorithm (see appendix), which identifies the most likely set of stems that is consistent with all angular brackets and where all paired sequence segments contain at least one angular bracket on each side. The output of this algorithm is shown in the fourth rows of figure 5, where lower case letters indicate paired sequence segments and the alphabetic order represents the confidence level (confidence is largest for 'a'). In figure 3 the bases involved in this reconstructed set of stems are colored, with green (red) indicating (in)correct basepairing. We observe that the two different basepairing patterns (for the same sequence) are clearly distinguished and the large scale secondary structure is captured in both cases. In particular, the pseudoknot in the native structure is correctly identified. The only incorrectly predicted stem is the least significant one ('k') in the pseudoknotted structure.

While these results seem satisfactory as a proof of principle, we stress that our reconstruction algorithm can certainly be improved upon, e.g. by allowing for mismatches in longer stems, which should help to fill in many of the missed basepairs. Also, one could make use of the known basepairing energies in the reconstruction.

4. Discussion and outlook

Our theoretical study has led us to a simple coarse-grained model, equations (2), (3), for the translocation of structured polynucleotides, which is applicable in the two opposite limits of very slow and very rapid translocation. This model is a useful starting point for a more detailed description that remains valid in the entire parameter regime. Here, we have applied the model to demonstrate that the physics of the translocation process can in principle be exploited to use nanopores for secondary structure determination (including pseudoknots) on the single-molecule level. Indeed, the nanopore technique would be a useful addition to the existing repertoire of structure determination methods: RNA secondary structure can be predicted computationally to some extent [29, 39, 40] based on experimentally determined free energy rules [30]; however this approach is unreliable for RNA molecules exceeding ~ 100 bases and cannot take pseudoknots properly into account. The inclusion of pseudoknots, which are often crucial to the function of RNA enzymes [32, 41], is not only computationally expensive [42, 43], but is also limited by a lack of experimental information on the corresponding binding free energies. Experimentally, x-ray crystallography [35] or NMR [36] provide detailed structures, but these techniques are cumbersome and limited to small molecules or isolated domains of larger RNAs. Structural information for larger RNAs can currently only be obtained from comparative sequence analysis [38], which requires large sets of homologous RNA sequences, or from indirect biochemical methods [32].

Throughout this paper, we have focused on basepairing only, which is permissible under ionic conditions that disfavor

tertiary interactions, e.g. low sodium and no magnesium. However, once the translocation of a molecule is well characterized under these conditions, it becomes interesting to switch to the native ionic conditions and examine the effects of tertiary interactions. Generally, one can expect more cooperativity in the presence of tertiary interactions, i.e. larger domains will open in a single step, as observed by Onoa *et al* [27]. This suggests a hierarchical approach to structure determination with nanopores: first unzip under low ionic conditions to obtain the secondary structure, and then repeat in the presence of magnesium to identify how the secondary structure elements are grouped into larger tertiary structure domains (such as the P4-P6 domain in the *Tetrahymena* ribozyme). It is worthwhile stressing the advantage of RNA as a model system to separately study the effect of secondary and tertiary structure. In contrast, the secondary structure of proteins is not stable in the absence of tertiary structure, and hence one may expect that single-domain proteins will unfold and translocate across a pore in a single step.

Nanopores could in principle also be used to probe the kinetics of large-scale secondary structure rearrangements in single molecules. For instance, it would be useful to attach larger objects to both ends of a molecule that is already threaded through the pore, allowing the same molecule to be driven forth and back through the pore, over and over again. By varying the time interval between successive reversals of the driving force, one could then probe structural relaxation over a broad range of time scales. More generally, nanopores may emerge as a new tool to probe intra- and inter-molecular interactions in single biomolecules. For instance, one could probe the biophysics of combined binding and folding in the context of RNA-protein interactions.

Acknowledgments

We are grateful to D Branton, S Ling, J Liphardt and D Lubensky for stimulating discussions. This research was supported by the National Science Foundation through grant No. 0211308, 0216576, and 0225630.

Appendix

Calculation of the free energy landscape

Given a secondary structure of the molecule, we obtain $\mathcal{F}_{\text{cis}}(m)$ by eliminating all basepairs involving the terminal $N-m$ bases, and calculating the binding free energy of the remaining structure according to the free energy rules for RNA secondary structure [30]. (We take the free energy parameters as supplied with the Vienna RNA package (version 1.3.1) at room temperature $T = 25$ °C. The salt concentrations at which these parameters were measured are $[\text{Na}^+] = 1\text{M}$ and $[\text{Mg}^{++}] = 0\text{M}$.) For pseudoknotted structures, the free energy rules currently include no prescription; however, the following extension appears reasonable: we first eliminate basepairs in stems that give rise to the pseudoknot(s) and calculate the free energy of the remaining structure according to the standard rules. We then add the free energies of the eliminated stems

separately, including the free energy for the loops created by these stems, again according to the standard free energy rules (however, the bases in these loops that are involved in other stems are removed before calculating the loop free energy).

The trans part of the molecule is tethered at both ends, by the pore and the pulling device, respectively. The pulling device can be described by a linear spring, while the configurational entropy of the RNA single strand can be modeled by a freely jointed chain (FJC) with extensible segments. (For the few bases that are inside the pore, we neglect the effect of the confinement on the entropy.) We denote by R_t the total extension of the trans part in series with the linear spring. The free energy (5) can then be expressed in terms of the total end-to-end distance distribution $W_{\text{FJC+spring}}$,

$$\mathcal{F}_{\text{FJC+spring}}(R_t; n) = -k_B T \log W_{\text{FJC+spring}}(R_t; n),$$

which can in turn be written as the convolution of the individual end-to-end distance distributions of the FJC and the spring [28],

$$W_{\text{FJC+spring}}(R_t; n) = \int_0^\infty dR W_{\text{FJC}}(R; n) W_{\text{spring}}(R_t - R).$$

Here, $W_{\text{spring}}(R_s) = \exp(-\beta\lambda R_s^2/2)/\sqrt{2\pi/\beta\lambda}$, where λ denotes the inherent stiffness of the pulling device. We calculate the end-to-end distance distribution of the freely jointed chain, $W_{\text{FJC}}(R; n)$, as described previously [37]. The polymer parameters we use were obtained from a fit [45] to FECs of single-stranded DNA [44] (base-to-base length 0.7 nm, Kuhn length 1.9 nm, and stretch modulus 815 pN), since we are unaware of corresponding data for the chemically very similar RNA.

Calculation of FECs

We obtain several trajectories $m(t)$ with a Monte Carlo simulation of equations (2)–(3) with $m(0) = N$, $R_t(0) = 0$ as initial condition and incrementing R_t at the constant rate v . The simulation is stopped when all bases have translocated ($m = 0$). From the time trace $m(t)$, we calculate the force-extension curve $f(R)$ using

$$\langle f \rangle = \frac{\partial}{\partial R_t} \mathcal{F}_{\text{FJC+spring}}(R_t = vt; N - m(t)) \quad (6)$$

and $\langle R \rangle = vt - \langle f \rangle/\lambda$. Here, $\langle f \rangle$ and $\langle R \rangle$ are both thermal averages over the polymer and spring degrees of freedom at fixed total extension R_t and fixed basepairing pattern.

Extraction of parentheses positions from a FEC

For every point on a FEC, we determine the length n of the freely jointed chain whose FEC passes closest to the point (using the polymer parameters for single-stranded RNA as given above). We take a histogram of the resulting lengths n over three independent FECs for each structure. In this histogram, the lengths n that correspond to start positions of stably basepaired regions appear as peaks, since the length of single-stranded RNA on the trans side remains approximately constant while the force required to unzip the basepairs builds up. (A similar procedure was applied in [46] to identify the

positions of proteins bound to double-stranded DNA as it is being unzipped.) We keep all n -values where the histogram exceeds a threshold of 30 counts (a count is made every Monte Carlo time-step). Since thermal noise makes the molecule fluctuate back and forth by a few bases while the force is building up for the next stem to open, we pick out of each contiguous stretch in the remaining n -values only the largest. Finally, we increment the extracted n -values by one and mark the corresponding position in the sequence with a parenthesis.

Reconstruction of the basepairing pattern

The FECs do not reveal which opening and closing parentheses are paired with each other. However, given the sequence of the RNA, we can match the parentheses by sequence complementarity. (To keep the number of false basepair predictions to a minimum, we consider only stems where we have at least one parenthesis at each end.) Here, we summarize the essential steps in our sequence alignment algorithm, while a detailed presentation and characterization will be given elsewhere (Bundschuh and Gerland, to be published). First, we find all possible gapless local alignments between a subsequence containing a parenthesis and subsequences to the open side of the parenthesis, using the scoring scheme 2 for GC, 1 for AU, and 0 for GU. We keep only those alignments with a score larger than 5 and where the matching sequence segment also contains a matching parenthesis. We consider the remaining alignments as possible stems in the secondary structure. To pick the most likely set of mutually consistent stems, we assign an alignment E-value to each stem [47]. We then iteratively include the most likely stem into the structure prediction, and remove all other stems it excludes due to overlapping basepairs from the list of allowed stems.

Glossary

Basepairing energetics. The binding free energy associated with the formation of a basepairing pattern, which can be decomposed, to a good approximation, into a sum over separate contributions from each stem and each loop.

FEC. Force-extension curve, usually for single (bio-)molecules or intermolecular bonds. FECs are obtained experimentally using devices such as atomic force microscopes or optical tweezers to record both the applied force and the displacement while stretching the molecule.

Hopping rates. Used here to denote the kinetic rates for forward/backward translocation of a polymer by one monomer. For structured RNA/DNA, these rates depend on the basepairing energetics on the cis and the trans side of the pore.

Pseudoknot. Non-nested basepairing pattern in an RNA (or single-stranded DNA) molecule, i.e. a basepairing pattern that includes a stem inside a loop.

Translocation dynamics. The detailed time dependence of the translocation process, where the reaction coordinate is the number of monomers on one side of the pore, e.g. the cis side.

References

- [1] Kasianowicz J J, Brandin E, Branton D and Deamer D W 1996 *Proc. Natl Acad. Sci. USA* **93** 13770–3
- [2] Akeson M, Branton D, Kasianowicz J J, Brandin E and Deamer D W 1999 *Biophys. J.* **77** 3227–33
- [3] Meller A, Nivon L, Brandin E, Golovchenko J A and Branton D 2000 *Proc. Natl Acad. Sci. USA* **97** 1079–84
- [4] Henrickson S E, Misakian M, Robertson B and Kasianowicz J J 2000 *Phys. Rev. Lett.* **85** 3057–60
- [5] Vercoutere W, Winters-Hilt S, Olsen H, Deamer D, Haussler D and Akeson M 2001 *Nat. Biotechnol.* **19** 248–52
- [6] Meller A, Nivon L and Branton D 2001 *Phys. Rev. Lett.* **86** 3435–38
- [7] Bates M, Burns M and Meller A 2003 *Biophys. J.* **84** 2366–72
- [8] Sauer-Budge A F, Nyamwanda J A, Lubensky D K and Branton D 2003 *Phys. Rev. Lett.* **90** 238101
- [9] Meller A 2003 *J. Phys.: Condens. Matter* **15** R581–R607
- [10] Sung W and Park P J 1996 *Phys. Rev. Lett.* **77** 783–6
- [11] Di Marzio E A and Mandell A J 1997 *J. Chem. Phys.* **107** 5510–4
- [12] Muthukumar M 1999 *J. Chem. Phys.* **111** 10371–74
- [13] Lubensky D K and Nelson D R 1999 *Biophys. J.* **77** 1824–38
- [14] Muthukumar M 2001 *Phys. Rev. Lett.* **86** 3188–91
- [15] Chuang J, Kantor Y and Kardar M 2002 *Phys. Rev. E* **65** 011802
- [16] Ambjörnsson T, Apell S P, Konkoli Z, Di Marzio E A and Kasianowicz J J 2002 *J. Chem. Phys.* **117** 4063–73
- [17] Flomenbom O and Klafter J 2003 *Phys. Rev. E* **68** 041910
- [18] Metzler R and Klafter J 2003 *Biophys. J.* **85** 2776–9
- [19] Li J, Stein D, McMullan C, Branton D, Aziz M J and Golovchenko J A 2001 *Nature* **412** 166–9
- [20] Storm A J, Chen J H, Ling X S, Zandbergen H W and Dekker C 2003 *Nat. Mater.* **2** 537–40
- [21] Cech T R 1993 *The RNA World* ed R F Gesteland and J F Atkins (Plainview, NY: Cold Spring Harbor Laboratory Press) pp 239–69
- [22] Brion P and Westhof E 1997 *Ann. Rev. Bioph. Biom.* **26** 113–37
- [23] Tinoco I Jr and Bustamante C 1999 How RNA folds *J. Mol. Biol.* **293** 271–81
- [24] Thirumalai D, Lee N, Woodson S A and Klimov D K 2001 *Ann. Rev. Phys. Chem.* **52** 751–62
- [25] Rief M, Gautel M, Oesterhelt F, Fernandez J M and Gaub H E 1997 *Science* **276** 1109–12
- [26] Liphardt J, Onoa B, Smith S B, Tinoco I Jr and Bustamante C 2001 *Science* **292** 733–7
- [27] Onoa B, Dumont S, Liphardt J, Smith S B, Tinoco I Jr and Bustamante C 2003 *Science* **299** 1892–5
- [28] Gerland U, Bundschuh R and Hwa T 2003 *Biophys. J.* **84** 2831–40
- [29] McCaskill J S 1990 *Biopolymers* **29** 1105–19
- [30] Walter A E, Turner D H, Kim J, Lyttle M H, Muller P, Mathews D H and Zuker M 1994 *Proc. Natl. Acad. Sci. USA* **91** 9218–22
- [31] Bonnet G, Krichevsky O and Libchaber A 1998 *Proc. Natl Acad. Sci. USA* **95** 8602–6
- [32] Pan J and Woodson S A 1998 *J. Mol. Biol.* **280** 597–609
- [33] Zhuang X, Bartley L E, Babcock H P, Russell R, Ha T, Herschlag D and Chu S 2000 *Science* **288** 2048–51
- [34] Zhuang X, Kim H, Pereira M J B, Babcock H P, Walter N G and Chu S 2002 *Science* **296** 1473–6
- [35] Cate J H, Gooding A R, Podell E, Zhou K, Golden B L, Kundrot C E, Cech T R and Doudna J A 1996 *Science* **273** 1678–85
- [36] Colmenarejo G and Tinoco I Jr 1999 *J. Mol. Biol.* **290** 119–35
- [37] Gerland U, Bundschuh R and Hwa T 2001 *Biophys. J.* **81** 1324–32
- [38] Gutell R R, Power A, Hertz G Z, Putz E J and Stormo G D 1992 *Nucl. Acids Res.* **20** 5785–95
- [39] Hofacker I L, Fontana W, Stadler P F, Bonhoeffer S, Tacker M and Schuster P 1994 *Monatsh. Chem.* **125** 167–88 Online at <http://www.tbi.univie.ac.at/>
- [40] Zuker M and Stiegler P 1981 *Nucl. Acid. Res.* **9** 133–48
- [41] Wadkins T S, Perrotta A T, Ferre-D'Amare A R, Doudna J A and Been M D 1999 *RNA* **5** 720–7
- [42] Rivas E and Eddy S R 1999 *J. Mol. Biol.* **285** 2053–68
- [43] Isambert H and Siggia E D 2000 *Proc. Natl Acad. Sci. USA* **97** 6515–20
- [44] Maier B, Bensimon D and Croquette V 2000 *Proc. Natl Acad. Sci. USA* **97** 12002–7
- [45] Montanari A and Mézard M 2001 *Phys. Rev. Lett.* **86** 2178–81
- [46] Koch S J, Shundrovsky A, Jantzen B C and Wang M D 2002 *Biophys. J.* **83** 1098–105
- [47] Karlin S and Altschul S F 1990 *Proc. Natl Acad. Sci. USA* **87** 2264–8