Title

Single molecule force experiments investigating the role of substrate structure in a viral DNA translocase reveal a novel mechanism of translocation

Author

Tong, Alexander B

Publication Date

2022

Peer reviewed|Thesis/dissertation

Single molecule force experiments investigating the role of substrate structure in a viral DNA translocase reveal a novel mechanism of translocation

by

Alexander B Tong

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Chemistry

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Carlos Bustamante, Chair
Professor Andreas Martin
Professor David Wemmer

Spring 2022

Abstract

Single molecule force experiments investigating the role of substrate structure in a viral DNA translocase reveal a novel mechanism of translocation

by

Alexander B Tong

Doctor of Philosophy in Chemistry

University of California, Berkeley

Professor Carlos Bustamante, Chair

Molecular machines are small assemblies harnessed to perform mechanical work in cells. These motors convert the chemical energy of ATP to move cargoes, transcribe RNA, and pump protons, to name a few examples. The study of these nanomachines is at the border of physics and biology, interesting to both fields for its implications of energy transfer at these length scales and their importance to cellular function, respectively. Optical tweezers is a natural choice as the tool to investigate these machines, as the technique has very high spatio-temporal resolution, enough to resolve the quick, small steps that these motors can take, and the applied force can be used to investigate the energetics and force generation properties of the system.

The DNA packaging motor of *Bacillus subtilis* phage φ29 is a model system for studying biological nanomachines. The role of this homopentameric ring motor is to package the DNA genome into capsids during viral replication. This ATPase is a member of the additional strand, conserved glutamate (ASCE) superfamily, meaning the study of this motor can reveal insights to the function of other members in the family. Benefits to studying φ29 include the simplicity of performing experiments, only requiring the motor-capsid complex, DNA, and ATP, its relatively large step size (0.85nm), and moderate speed (40nm/s) which makes observation of motor stepping well-suited for optical tweezers experiments. Single-molecule optical tweezers studies of this motor have revealed great detail about its mechanochemical cycle and packaging mechanism, including interesting symmetry breaking from this homomeric ring, in which one of the five subunits is *special* in that it does not perform a mechanical task, but a regulatory one. This motor operates under a *dwell-burst* cycle, where the motor, starting from an ADP-filled ring, first exchanges ADP for ATP in an ordered, sequential fashion (during the dwell). After all five have exchanged, the special subunit hydrolyzes its ATP, which then causes the other four subunits to sequentially hydrolyze their ATPs and translocate 0.85nm (2.5bp) of substrate each (during the burst), resulting in the packaging of 10bp. The delineation between the one special and four translocatory subunits shows a division of labor, where one subunit takes on a different role from the others, despite their identical protein sequences. The similarity between the amount of DNA packaged

in one cycle (10bp) and the pitch of DNA (10.4bp) is not just a coincidence, previous studies suggest that the event that assigns the identity of the special subunit is that it contacts two adjacent phosphates every pitch, and that this identity is preserved cycle after cycle. Despite this knowledge, the actual motions of the motor subunits that translate into substrate translocation are unknown.

One proposed translocation model for p29 involves a dehydration-induced B-form DNA to A-form DNA transition. This *scrunchworm* model explains the 0.85nm step size as the difference in contour length of 10bp B-form and 10bp A-form DNA. We can test if this model is correct by making the motor package dsRNA, a substrate that can only adopt an A-form structure. In addition, this substrate has a different pitch than DNA, so we can see if the burst size changes if the pitch of the substrate changes. We found that, when challenged with dsRNA, the motor is indeed able to package it, invalidating the scrunchworm model. Packaging of dsRNA and RNA:DNA hybrid combinations reveals that the motor conforms its burst size to the periodicity of its substrate. Hence, we conclude that the motor's burst size is determined by the pitch of the substrate. However, we observe that the step size of the motor is still the 0.85nm it is on DNA, with one step shortened to accommodate the shorter total burst size, suggesting that the step size of the motor is innate and not determined by the substrate. Because the structure of the ATP-full motor was found to exist in an open lock-washer state by cryo-EM, we propose a model of translocation in which the ring is planar at the start of the dwell and successively opens up upon nucleotide exchange during the dwell. Thus, at the end of the dwell, and at the beginning of the burst, the motor adopts a lock-washer conformation. The open-ring structure is the mechanism by which the burst size of the motor adapts to the pitch of its substrate. During the burst, nucleotide hydrolysis successively closes the ring, translocating the substrate and reverting the motor to a planar conformation. The step size is determined by the motion of the "hinges" formed by the adjacent subunits in the spiral. This translocation mechanism is an interesting example of a motor adapting to the periodicity of its substrate, while taking advantage of its repeating chemical moieties by using them as rest points like the rungs of a ladder.

Preface: Studying Molecular Motors

Molecular nanomachines are a foundational part of life; these tiny assemblies are responsible for the execution of mechanical work on the nanoscale. Nanomachines in nature include “walking” motors such as myosin and kinesin[1], rotating motors such as pumps and flagella[2,3] and nucleic acid motors such as FtsK and RNA polymerase that are involved in the upkeep, expression, and migration of genetic material[4-6]. From a biology perspective, these motors carry out very important processes, whose dysfunction can lead to disease and are potential targets for therapeutics. From a physics perspective, performing work on a molecular scale involves different phenomena than on a macroscopic scale, leading to the observation that motors in nature can have energetic efficiencies nearing 100%, unobtainable by our macroscopic motors[7,8]. Study of how the molecular nanomachines in nature work has given researchers insight into how to attempt to build novel nanomachines in order to perform tasks at the nanoscale that would otherwise be impossible[9,10].

The research of biological nanomachines naturally falls under the field of biophysics, and due to the mechanical nature of their function, single-molecule force spectroscopy is an obvious fit as a tool for their study. Single-molecule techniques obtain data for individual motors, which can illustrate variability that would otherwise be averaged out in a bulk experiment. In addition, the often higher time resolution inherent to the techniques used in single molecule compared to bulk allows for direct observation of the machine’s action in real time. In addition, single-molecule experiments can capture rare events, like off-pathway motor pausing. Force spectroscopy also allows for measurement of the work done in a system, which can then be compared to the input energy to get the efficiency[11], and the measurement of the force dependence can show the energetics of a reaction pathway or reveal the existence of multiple pathways in a process[12].

Chapter 1: Introduction to the φ29 DNA packaging motor

Many molecular nanomachines take the form of a ring ATPase, a cyclic arrangement of proteins that work together to do a common task and harness the hydrolysis of ATP as an energy source. Motors that fall under this category include helicases, which split nucleic acids by translocating one strand[13], protein unfoldases, which translocate and unfold polypeptide chains for degradation[14,15], or DNA translocases, which segregate genomic material into a daughter cell or a viral capsid[6,16]. The members of the ring are often identical or are at least very similar in structure, which evokes questions about how the subunits work together to perform a common task. Do the subunits coordinate ATP hydrolysis along the ring, and if so, how? What specific protein residues are involved in that process? What are the conformational changes that translate into mechanical work? While the ATPase must be a motor to be able to be studied via force spectroscopy, not all ring ATPases have a mechanical function, but the answers to how the motors solve the coordination problem are potentially universal and applicable to the non-motor ring ATPases, too. The way these motors convert the chemical energy of

ATP to force in an extremely efficient manner can reveal laws governing economic energy transduction on this length scale, useful for future nanomachine design.

The bacteriophage φ29's DNA packaging motor is a ring ATPase that has been extensively studied as a model system. Its role is to package the genome into the capsid, a key step in the viral life cycle. A member of the additional strand conserved glutamate (ASCE) superfamily[17], which includes other ring ATPases such as membrane scission protein Vps4, the DNA polymerase clamp loader, and the previously mentioned translocases[18], the motor packaging system can be probed to understand how these ATPases work. Using single molecule force spectroscopy, we can answer the questions posed earlier regarding how the often identical members of these rings communicate and coordinate to do perform their task, and what conformational changes are coupled to force generation.

To study the φ29 motor using optical tweezers, the two ends of the packaging complex, the capsid and the DNA, are held between two optical traps by polystyrene beads to monitor the progress of DNA packaging over time (Figure 1.1). Decrease in the tether length corresponds to DNA packaging by the motor. The biochemistry behind the packaging system is detailed in Appendix 1. Performing an optical tweezers experiment involves capturing the two beads in the traps, and then forming a tether between the two traps by rubbing the two beads together. For example, one bead could have the packaging complex attached by the DNA end through a biotin-streptavidin interaction, and the other bead could be coated in antibodies to the capsid. When the two beads are brought together, the antibody grabs the capsid, and a connection is formed between the two beads. The length of this tether is monitored over time to follow motor packaging. The control of the instrument is detailed in Appendix 2, including optimizations made to increase throughput and decrease the effort to perform these single-molecule experiments and a documentation of the LabVIEW code that controls the machine.

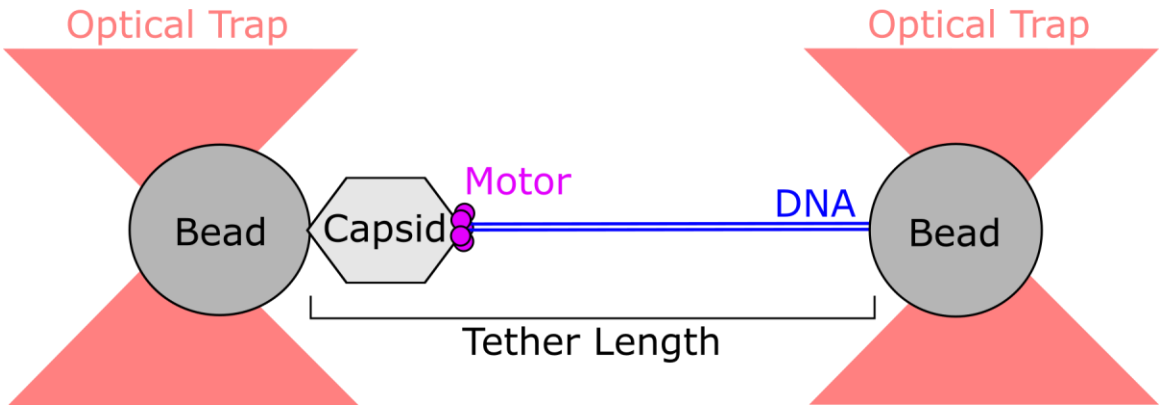


Figure 1.1 φ29 packaging experiments in single-molecule optical tweezers The packaging system (capsid, motor, DNA) are held between two beads captured in optical traps. The deflections of the bead from the center of the traps are used to calculate the force applied to the motor, and the distance between the beads (tether length) is monitored over time. DNA packaging results in a decrease in the tether length, and one optical trap moves closer to the other to

keep up with the motor's action. The capsid is attached to the left bead via an antibody, and the DNA is attached to the right bead by a biotin-streptavidin interaction.

Previous studies in the φ29 packaging system using single molecule optical tweezers have revealed many details of this motor's operation. First off, the motor packages at a rate of ~120bp/s and can package against up to 60pN of opposing force[19]. Observation of packaging trajectories reveal this motor operates via a *dwell-burst* cycle, where the motor waits for an amount of time in a *dwell* and then packages 10bp of DNA in one *burst* of translocation[20]. The length of the dwell is sensitive to [ATP], [ADP], and follows a Gamma distribution with shape factor 5, suggesting at least five rate-limiting events happen during this time. The motor is a homopentamer, and combined with the sensitivity of the dwell time to nucleotide concentration, suggests that the motor exchanges ADP for ATP during the dwell, and presumably the rate-determining events are the five ATP binding events. This result also means that the ATP exchange process is ordered, where one subunit exchanges, then the next, and so on along the ring (as opposed to the exchange events happening in parallel). At low applied forces (~10pN), the motor burst looks like it occurs in one transition, but at higher applied loads (~40pN) the individual steps of the burst are slowed by the force and the burst breaks down into four steps of 2.5bp each[20]. The discrepancy between the number of subunits in the motor (5) and the number of steps it takes (4) is rationalized by a division of labor between the five subunits, where four subunits perform translocation and one plays some other kind of role. To determine the role of this *special* fifth subunit, experiments where the motor packages in a medium containing small amounts of ATPγS, a non-hydrolyzable ATP analog, shows that the motor pauses upon trying to hydrolysis of one ATPγS molecule[21]. This result shows that the burst is also ordered, where the stalling of one subunit is enough to halt the entire motor (as opposed to the hydrolysis events occurring independently, where the stalling of one subunit does not prevent the hydrolysis of the other subunits). By tracking when the motor pauses due to ATPγS (after how many 2.5bp steps), it was found that each step correlates to one hydrolysis, and the dwell contains two hydrolyses, one of the special subunit and then one of a translocatory subunit[21]. Hence, the special subunit's hydrolysis occurs either at the start or the end of the burst, more likely a signal to start the burst. Combining all of these results yields the mechanochemical cycle of the packaging motor, where sequential ATP exchange occurs during the dwell and sequential hydrolysis-coupled translocation happens during the burst (Figure 1.2).

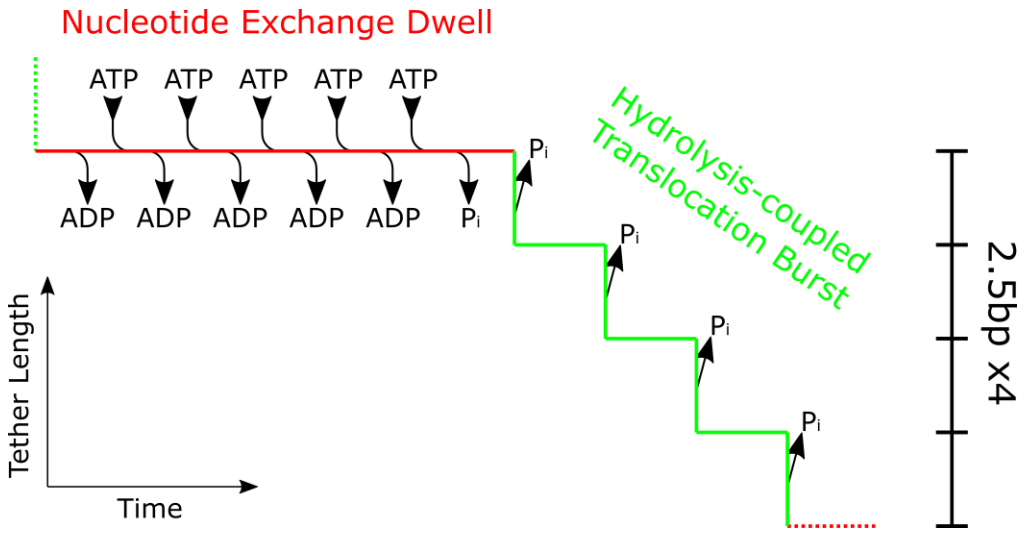


Figure 1.2 The φ29 mechanochemical cycle
From the start of the dwell (red line), ADP is exchanged for ATP sequentially in all five subunits, marked by the leaving of ADP and arrival of ATP. The hydrolysis of the special subunit (marked by the loss of P_i), denotes the start of the burst. In the burst (green staircase), hydrolysis-coupled translocation packages the DNA in four steps of 2.5bp each, and the dwell starts again. Modeled after *Liu, 2014*[22].

But how does the motor physically work? i.e., how does it interact with its substrate to translocate it? To investigate the sensitivity of the motor to its substrate, the motor was challenged with packaging other substrates, such as base-less DNA or bulged DNA, but most importantly DNA with a methyl phosphonate backbone. This charge-neutralized backbone only gave the motor significant trouble if it was 11bp long, which combined with the 10bp step size leads to the conclusion that the important contact between the motor and the substrate is an electrostatic interaction with a pair of phosphates every 10bp (while the motor prefers a pair, one single phosphate is sufficient for grip)[23]. When there is an 11bp gap, the motor no longer has phosphates to grab, and the motor pauses strongly at this point. In addition, this effect is only seen when the gap is introduced on a certain strand of the DNA (the one going 5' to 3' in the direction of packaging), charge neutralization of the opposite strand is well-tolerated[23]. So, this strand is called the *tracking strand* as it is the one presumably tracked by the motor. A cartoon representation of this process is shown in Figure 1.3. However, the further details of how the motor generates its power stroke are unknown.

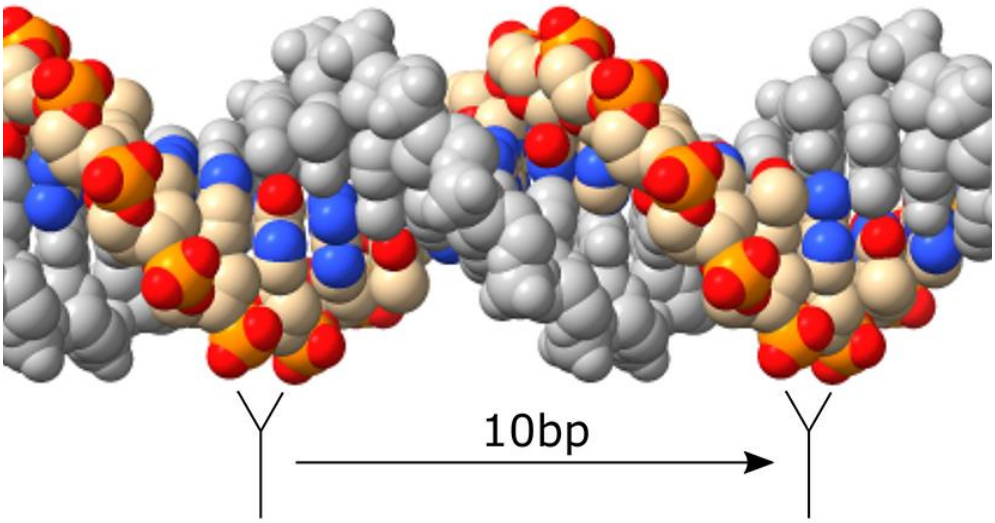


Figure 1.3 The footprint of the φ29 packaging motor
A DNA is shown with the tracking strand in color and the other in grey. The motor tracks the colored strand with a 2-phosphate footprint; two example gripping points of the motor are denoted by the two Ys. Note how the two gripping points are on the same side of the DNA, just translated one pitch away. The DNA structure was generated from web.x3dna.org.

The substrate-engaged structure of the motor was solved by cryo-EM (Figure 1.4a,PDB:7jqq)[24]. The complex was made to be in a late stage of packaging where the motor dwell slows and can also enter long-lived pausing events due to the high degree of capsid filling[22]. In addition to becoming longer, the shape of the dwell distribution shifts from Gamma with shape 5 at low filling to Gamma with shape 2 at high filling, which suggests that the rate-limiting step has become some other regulatory signal that delays the start of the burst at high filling. So, this structure is most likely an ATP-full motor waiting at the end of the dwell for the signal to start the burst. Surprisingly, the motor's N-terminus is arranged in a spiraled lock-washer shape that follows one strand of the DNA (this strand is the same as the tracking strand) (Figure 1.4b). Other ring ATPase translocases that adopt lock washer structures have been theorized to act via a *hand over hand* mechanism, in which the conformational changes that lead to translocation resemble climbing up a spiral staircase, where the spiral structure climbs up a helical substrate by translating the lowest subunit upwards one pitch to become the highest subunit[25-27].

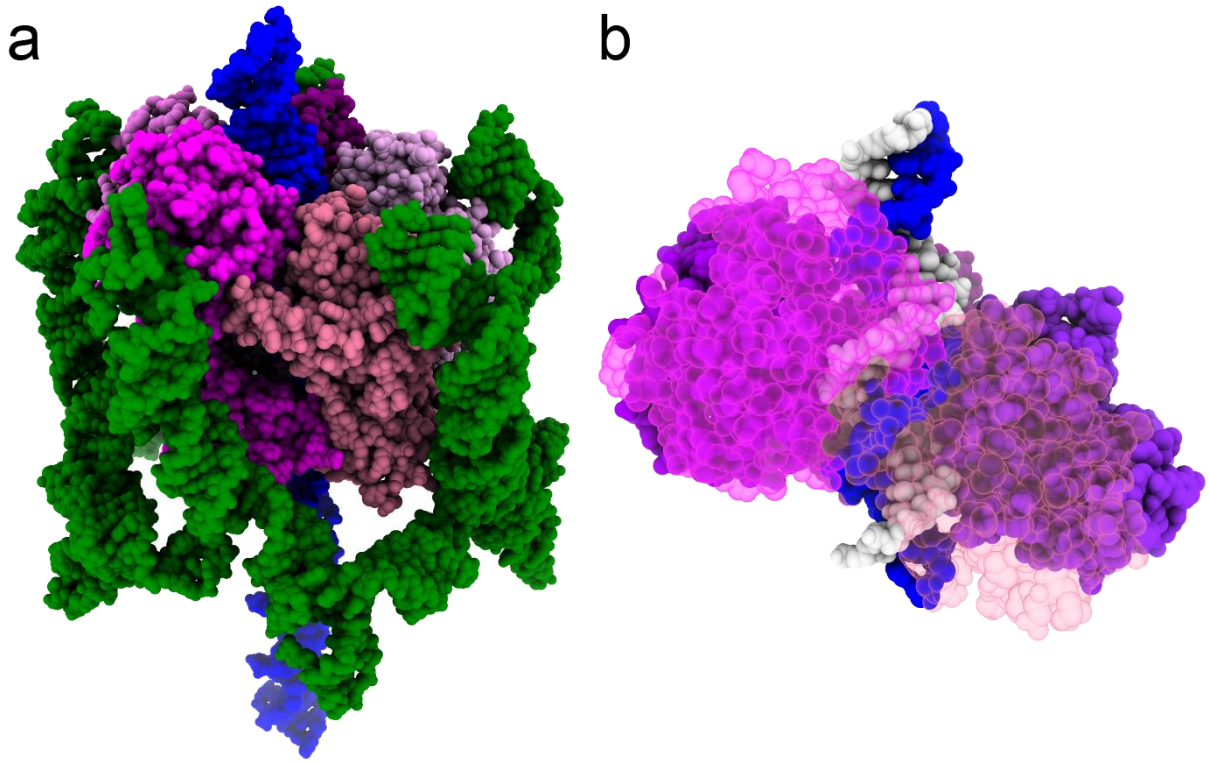


Figure 1.4 The structure of the φ29 packaging complex (PDB:6jqq)
a) The structure of the packaging complex, with pRNA scaffold in green (see Appendix 1), DNA in blue, and motor subunits in magenta colors. b) Cutaway just showing the N-terminus of the motor and the DNA, with the tracking strand colored blue and the two subunits facing the camera made transparent. Observe how the subunits form a lock-washer shape, and span one pitch of the DNA.

At this point, the motor's mechanochemical cycle is well-characterized, but the specific residues that are involved in these steps are unclear. Structure can be used to direct mutagenesis at key sites to see the effect of disrupting or removing the function of that residue. A study in φ29 targeted the putative arginine finger R146, the residue involved in coordinating the gamma phosphate of ATP during catalysis [28]. It was found that disruption of this residue was not tolerated as a homopentamer, but complexes with altered activity can be found by mixing mutants with wild-type ATPase. The conclusion of the work is that this residue is involved in both stimulating ADP release and ATP hydrolysis, demonstrating specific roles for this residue[29]. It was later determined that the residue is more likely a gamma phosphate sensor rather than the arginine finger itself (this work was done before the cryo-EM structure), which still is consistent with the conclusion. There are many other unstudied residues involved with not just the catalysis of ATP hydrolysis, but also non-catalytic signaling between motor subunits, interaction between the motor and DNA, and the communication between the motor and the other elements of the system (for example, the *connector* is a component that the motor sits on and is thought to be responsible for the slowing of packaging observed at high filling, and so must somehow communicate with the motor). Such details are important for fully understanding the biology behind the signaling and catalysis of these ATPases.

A number of models have been proposed as the mechanism of translocation for this motor. The most peculiar observation made on this motor is the division of labor and the non-integer step size, trying to rationalize this observation has spawned three translocation models alone. The first is one in which the motor translocates DNA by the movement of a “steric paddle”, which interacts with the substrate in a non-charge dependent manner, which explains the step size as the size of the conformational change (motion of the paddle) made by each subunit and the insensitivity of the motor to the charge of the backbone outside of the pair of phosphates it uses every 10bp[20]. The second is a “lever-latch” mechanism where the motor opens, grabs its target phosphate 10bp down the helix, and closes in four steps since there are four “hinges” between the five subunits[20]. The third is a *scrunchworm* model based on a B-form DNA to A-form DNA transition caused by motor-based dehydration of the DNA. Since the two polymers have different pitches (10bp of A-form DNA has the same length of 7.5bp of B-form DNA, a difference in distance of 2.5bp!), DNA is brought into the motor by dehydration, and released into the capsid by rehydration[30]. And finally, there is the hand-over-hand mechanism that has been proposed for other ring ATPases with lock-washer structures, wherein the stepping is the result of the bottom member of the open ring moving to become the top, climbing its substrate like a spiral staircase[25-27]. Cartoons of these four models are shown in Figure 1.5.

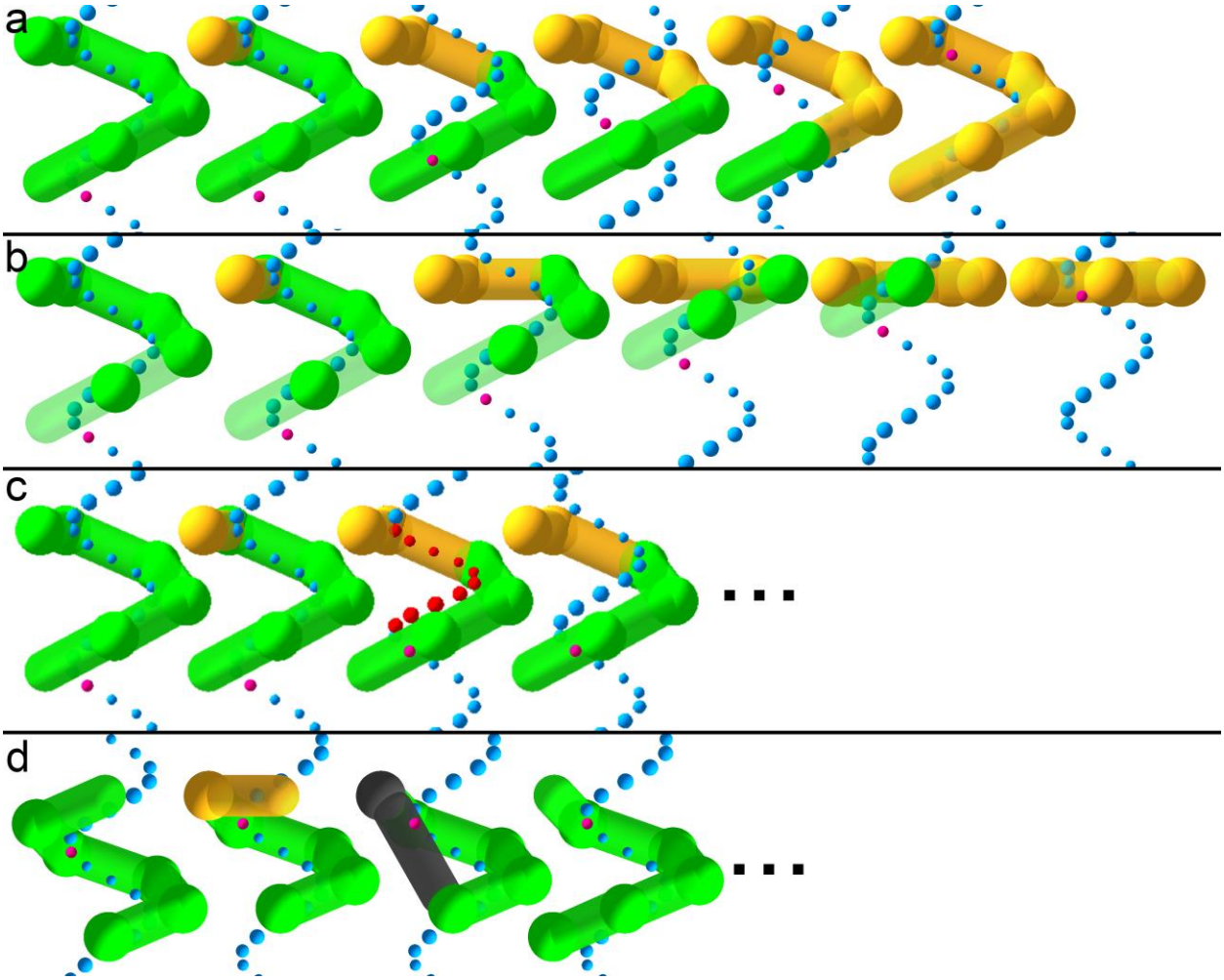


Figure 1.5 Models of packaging for φ29
Cartoon representations of the four proposed models of translocation for this motor. Each subunit of the motor is represented as a sphere connected to a cylinder, with color representing nucleotide state (ATP/ADP/apo is green/yellow/black, respectively). Snapshots go left-to-right in time, for example, a subunit that is green in one frame and yellow in the next has just undergone ATP hydrolysis. The open lock-washer structure of the motor is shown on the left in all subpanels. The phosphates of the DNA's tracking strand are shown as blue spheres, and one phosphate is colored pink as a guide for the upwards translocation of DNA. a) In the steric paddle model, DNA translocation is carried out by "paddles" in each of the four translocatory subunits that move DNA in 2.5bp increments each. b) In the lever-latch model, the open ring closes to translocate the DNA. c) In the scrunchworm model, the power stroke involves a transition into A-form DNA (red circles), the conversion of 10bp causing one step. The process repeats three more times to finish the burst (not shown). d) In the hand-over-hand model, the subunit most proximal to the capsid (highest) is rearranged to become the lowest member, causing DNA translocation. This process repeats three more times to complete the burst (not shown).

So which model is right? We can attempt to disprove the third model by attempting to package an A-form nucleic acid. Since there is no known structure of a double-stranded nucleic acid that has an even shorter structure than the A-form, there would be no structure with a shorter pitch to "scrunch" to and the motor would be unable to package it. dsRNA is a natural choice for this alternate substrate, since it can only adopt an A-form structure. Additionally, the size of the burst of the motor matches the size of the pitch of DNA (10bp vs. 10.4bp), with the rounding down a consequence from the DNA phosphates being used as a gripping point in between bursts[22,23]. Another question is, then, how would the motor adapt to this substrate of a shorter pitch? Would it continue packaging the same length of substrate (3.4nm), or instead package one pitch of dsRNA (3.0nm)? If the burst size changes, does the step size change, too?

Chapter 2: Tricking the φ29 motor to package dsRNA with a chimeric substrate

The *in vitro* reconstitution of the φ29 packaging system is straightforward and just involves adding the components in series. The system consists of the viral capsid, the pRNA (a structural RNA that forms the scaffold that attaches the ATPase to the capsid), the ATPase gp16, and the genomic DNA. The components self-assemble and the addition of ATP starts packaging. More information and relevant protocols are in Appendix 1. Getting the motor to package dsRNA isn't as simple as adding lengths of dsRNA instead of the φ29 genome. This is because, in the test tube, the motor will only initiate on DNA carrying a terminal protein, gp3, on the end. φ29 DNA polymerase uses gp3 as the primer to initiate replication, so genomic φ29 DNA carries this protein. We don't have a way to introduce gp3 to an arbitrary DNA, so we first tried to attach a length of dsRNA to genomic DNA carrying a gp3, so that the motor would initiate on the gp3-DNA and eventually make its way to the dsRNA region and try to package it. A cartoon representation of this chimeric DNA-dsRNA substrate is in Figure 2.1a.

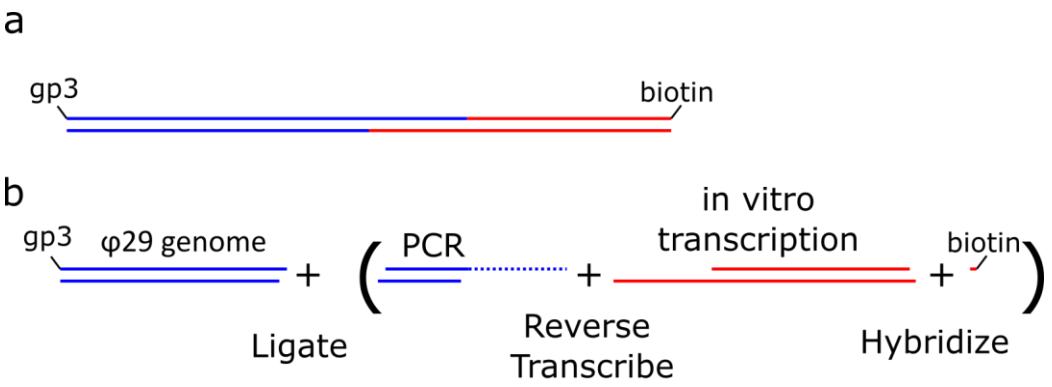


Figure 2.1 Chimeric substrate for φ29 packaging
a) Schematic of the chimeric packaging substrate, a gp3-DNA-hybrid-dsRNA-biotin construct, with DNA strand in blue and RNA strand in red, 5' end on top-left. b) The component parts are (from left to right) a genomic gp3-DNA, a PCR product, a dsRNA fragment, and a biotin oligo. Assembly is denoted with + signs, with method written below and parentheses denoting order. The dotted blue is added via reverse transcription.

The parts that create the chimeric substrate are shown in Figure 2.1b. First, the genomic φ29 genome is cut by BstEII and PspGI restriction endonucleases, to cut the genome into a 2.7kb and 6.8kb DNA with sticky ends and terminal gp3s. There is also a 9.8kb fragment, which is the central part of the 19.3kb genome. The motor is known to selectively package starting from a specific end of the genome, biologically relevant because it allows the correct end of the genome to be injected into the target cell upon infection, so potentially only one of these fragments will be useful for initiating packaging[31]. This *left end selection* is removed if a shorter form of the pRNA, a truncation of the 174nt RNA (*174b*) to 120nt (*120b*), is used[31]. While the experiments are performed using the full-length pRNA, we do not observe left end selection in our hands, so left end selection is not a concern (data not shown). A possible explanation is degradation of the 174b RNA to 120b by RNAse contamination. Coming back to the 2.7kb left, 9.8kb middle, and 6.8kb right fragments, these now need to be separated while preserving the integrity of the gp3. Conventional methods such as agarose gel extraction won't work, as there is a guanidinium extraction step that will destroy the gp3. So, to separate these three pieces, a custom-built continuous elution agarose gel electrophoresis device was used, which uses the separation properties of gel electrophoresis to elute the DNA pieces by size. For more details on this device, see Appendix 3. A similar commercial machine is the Bio-Rad Model 491, which has a similar function but is designed for polyacrylamide gel electrophoresis instead of agarose gel electrophoresis. The result of this step is the purification of DNA with gp3 on one end and a sticky end on the other, which will be eventually ligated to the rest of the construct.

The dsRNA part of the substrate is formed by hybridization of two ssRNA pieces. Each is transcribed using in vitro transcription (T7 MegaScript kit, "long transcript" protocol), with one arm being 4kb in length and the other 2.7kb in length. The T7 promoter for transcription is introduced by polymerase chain reaction (PCR) to lambda phage DNA.

The two pieces are annealed by mixing together, heating them, and slowly cooling them over two hours. The annealed product is run in an agarose gel and purified by gel extraction. This final dsRNA product has overhangs on both sides, one small overhang for the annealing of a biotin oligo, and the larger overhang for reverse transcription to join to the DNA piece. The 1.3kb section will act as “glue” to make sure the DNA and the dsRNA parts stay together, since the joints are not necessarily ligated[32].

The next step is to prepare an adapter DNA that will allow hybridization to the dsRNA for reverse transcription on one end and ligation to the genomic DNA on the other. The adapter is lambda phage DNA, either 1kb or 3kb in length and with either a BstEII or PspGI site (for eventual ligation to the genomic DNA fragment) introduced on one end and a BstXI site (whose overhang is the suitable overhang to prime reverse transcription on the dsRNA piece) introduced on the other. The longer length of the adapter with a PspGI site is to lengthen the dsDNA section, since the genomic DNA fragment is only 2.7kb long. The DNA section should be more than ~3.6kb in length to ensure that, when preparing the complex for tweezing, the motor is stalled on a DNA section after prepackaging for 30 seconds (the motor packages at ~120bp/s at room temperature; see Appendix 1, Table A1.3). After PCR, the DNA is cut by BstXI and purified by agarose gel electrophoresis followed by gel extraction to prepare it for reverse transcription.

To combine the dsRNA, the adapter DNA, and the biotin oligo, the three are mixed and T4 RNA ligase 2 is added to ligate the biotin oligo to the dsRNA and the adapter DNA to the dsRNA overhang. Next, Protoscript II reverse transcriptase is used to extend the DNA using the ssRNA section as a template. After the reverse transcription finishes, the nick that remains is potentially ligated by the T4 RNA ligase 2, although that specific nick geometry (5’-DNA to RNA-3’ over RNA) is apparently impossible to ligate[32]. Finally, the chimeric substrate is cut with BstEII or PspGI and purified by continuous elution gel electrophoresis. The 5kb or 7kb product is then ligated to the 7kb or 3kb genomic DNA piece by E.coli DNA ligase and is ready for use. Another round of purification is skipped to try to keep the gp3 intact. When the stalled complex is made for tweezing, any packaging complexes that initiate on the unligated DNA fragments do not attach to the streptavidin bead since their free end is not biotinylated, so they will not interfere with the experiment (see Figure 1.1, 2.1b).

The finished chimeric substrate is a length of double-stranded nucleic acid, starting with >6kb of DNA, then 1.3kb of DNA-RNA hybrid, and then 2.7kb of dsRNA (Figure 2.1a). This chimeric substrate replaces the genomic DNA, which allows us to examine how the motor deals with packaging these substrates with differing pitches in the optical tweezer. Single-molecule trajectories showing the packaging of this substrate by φ29 are shown in Figure 2.2. The data show a change in the velocity of the motor at the expected location of the DNA-hybrid junction. However, for whatever reason, this method of packaging initiation proved extremely inconsistent. Nonetheless, the data was enough to show that the motor could package these substrates, that they were packaged differently, and hence this was a project worth pursuing. In addition, the simple observation that the motor could package the A-form hybrid is enough to disprove the scrunchworm model of motor operation. The next chapter will discuss basic analyses of

p29 translocation data, which will then be applied to translocation data of an A-form substrate.

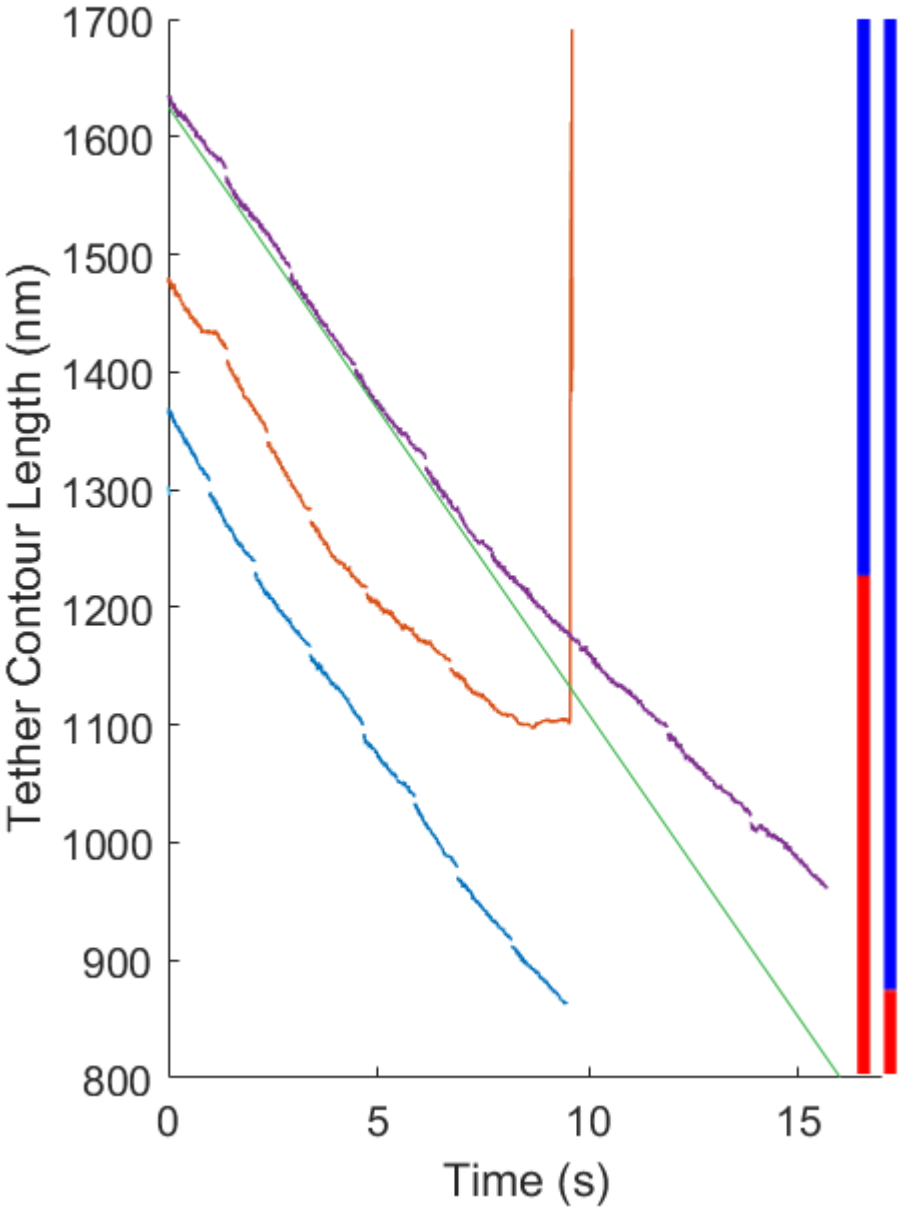


Figure 2.2 Packaging of the chimeric substrate
Packaging traces of the chimeric substrate. Applied force is 7-12pN with saturating (0.25mM) [ATP]. Different colors are different packaging trajectories, the gaps in between data segments are when the mirror moves to reset the increasing force (see Appendix 2, Figure A2.5). The expected location of the motor on the substrate is on the right. The velocity of the traces is seen to kink (change velocity) around 1250nm, corresponding to the expected position of the DNA-hybrid junction. The vertical line in the orange trace corresponds to a break

in the tether. The straight green line is parallel to the dsDNA packaging region and serves as a guide for the change in packaging velocity

Chapter 3: Analysis of φ29 packaging velocity, burst size, and dwell time distribution

A basic analysis that can be done on phage packaging data (or any translocation data) is a quantification of the motor's velocity. The most primitive version is the global velocity: the length of the trace (nm) divided by the duration of the trace (s). The φ29 motor is consistent in its translocation velocity due to its constant burst size (10bp) and peaked dwell time distribution (Gamma, shape 5), making traces look like straight lines until viewing at a resolution that can separate the dwell-burst cycle of the motor (Figure 3.1), so this makes the global velocity not a terrible method for determining velocity for a pause-free trace. However, this method does not handle pauses well, and is poor for gathering statistics on any variability in velocity.

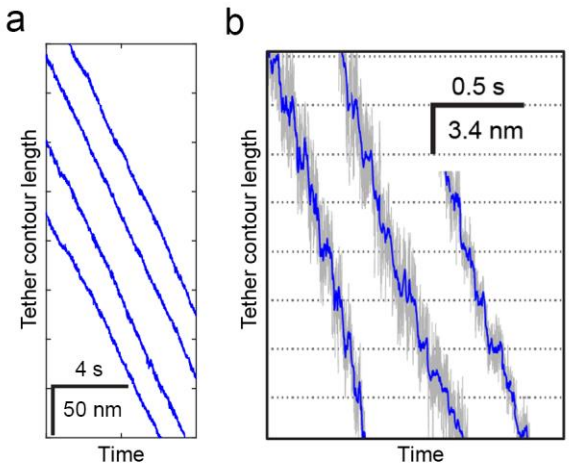


Figure 3.1 φ29 translocation data at different levels of zoom
a) At low zoom, the traces look straight. b) At higher zoom and filtering (grey 2.5kHz, blue 250Hz), the motor translocation looks like a staircase with a 3.4nm step.

Of course, we can do better than that method. One way is by instead dealing with small time slices of a trace. For example, slice a 10s trace into ten 1s segments and get the velocity of each of those sections. In addition, instead of taking a naïve linear fit by drawing a line connecting the start and end point, we can do an actual linear fit of the data. These changes should better handle pauses in the data (as these sections will separate to zero velocity sections and can be then identified as a separate population after binning) and be more accurate in determining the velocity of each section due to considering more points. An elegant way to take care of this fitting process is by filtering the data with a Savitzky-Golay differentiating filter of order 1, which just requires the (computationally inexpensive) application of a constant 1d filter[33]. The result that the least squares fitting of a line to a set of points can be summarized into a filter independent of the actual points is not obvious, but useful here to improve code runtime.

The filter transforms each point in the raw data into a velocity, which can then be binned into a velocity histogram (Figure 3.2). The tunable parameter is the choice of filter width (expressed in time = points / sampling frequency), where lengthening the width will generally make the spread of the velocity histogram smaller at the cost of time resolution. The lower limit on filter width is either dictated by noise (where the signal is drowned out) or the details of motor operation (for example, the phage motor cycle is ~100ms, widths shorter than this will separately identify the dwell and burst, which is not the intention), and the upper limit is pause duration, where the filter can no longer separate paused sections from active ones. When taking statistics on data obtained by this method, it should be noted that there is degeneracy in the data, as adjacent time points in the velocity come from shared data since the filter is applied without downsampling. For binning, I use all of the data to make the bins look smoother, but for statistics, take care to use the number of points divided by the filter width for the "true" number of independent data. For example, the standard error of the velocity is $\sigma / sqrt(N/filter\_width)$.

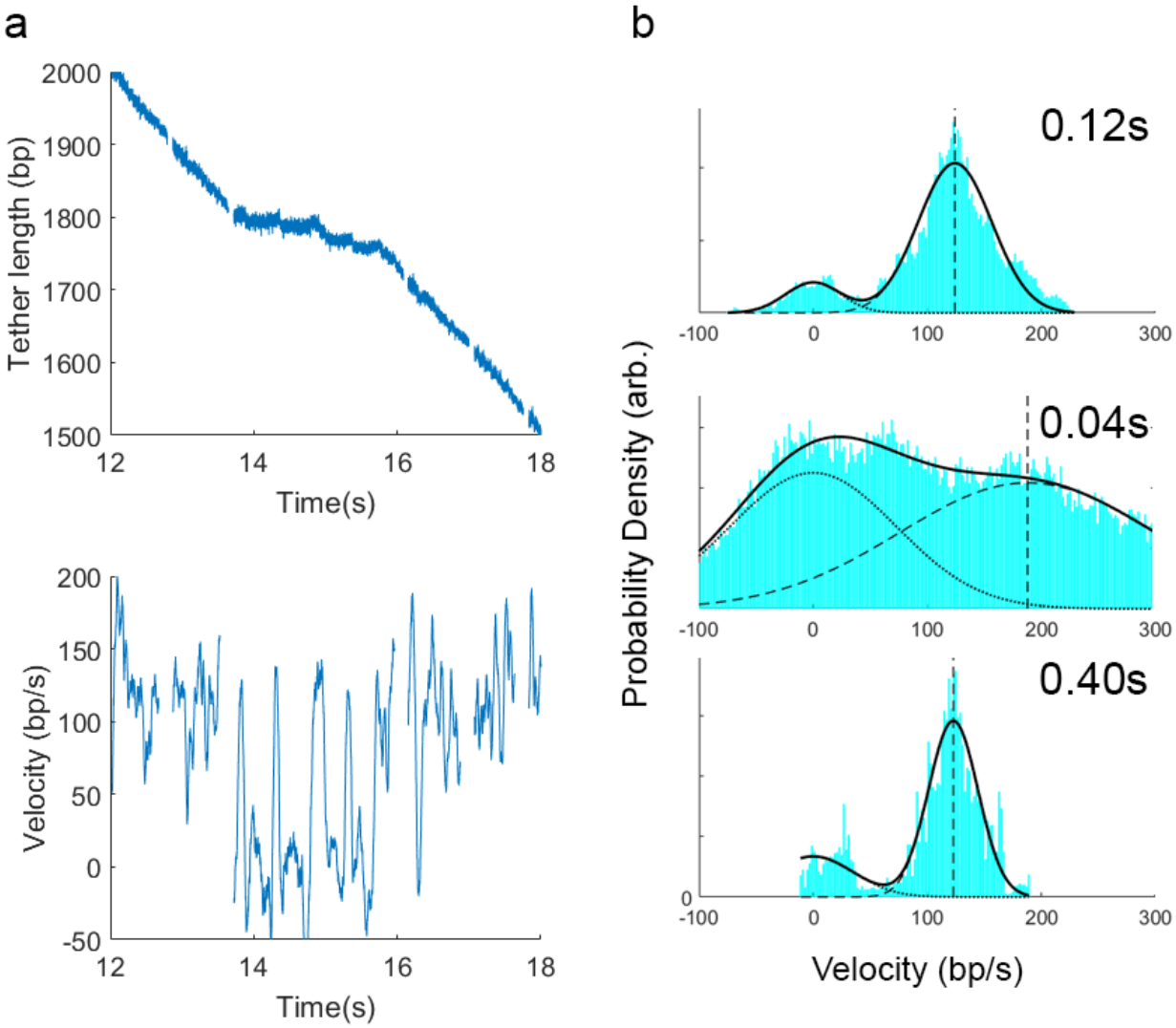


Figure 3.2 Velocity determination by filtering
a) Above is the raw data, a trace taken at 7-12pN force and 0.25mM ATP. Filtering with the Savitzky-Golay method yields the bottom velocity-time trace. Note how the velocity goes to zero during the section where the motor is paused (14-16s), and is about 120bp/s otherwise. b) By binning the velocity data obtained in a), we see that the velocity is biphasic, and we fit two Gaussians to the data, one centered at zero (for pauses) and one at a positive value (the translocation velocity). The top graph is for a "good" value for the filter width (inset, seconds), shown below are what happens if the filter value is picked too narrow or too wide. When too narrow (0.04s filter width), the motor dwell-burst cycle is not homogenized, so it looks like there are more pauses (dwells being miscategorized as pauses), with a corresponding increase to the detected velocity. Note how the distribution widens, too, since fewer points are being used to get each point. If the width is too wide (0.40s filter width), the distribution of the velocity looks nice and narrow, but the pause population is no longer centered around zero, meaning there is a significant population of data that comes from a mix of paused and translocating populations. The good value of 0.12s is chosen to be just wider than the dwell-burst cycle of the motor.

The next level of detail is to detect the dwell-burst cycling of the motor. The residence time for a given trace peaks in a repeating manner because of the dwell-burst cycle (Figure 3.3ab). The periodicity can be found by taking the autocorrelation of the residence time histogram (*pairwise distribution*, Figure 3.3c). Since this motor has a defined burst size, the pairwise distribution peaks at 10bp and multiples, as the motor dwells every 10bp (Figure 3.3c). Signal and noise in an optical tweezers experiment can vary from tether to tether, a convenient metric to determine signal to noise ratio in phage data is by quantifying the strength of the periodicity of the PWD, and by this metric data can be sorted by signal strength when necessary[20]. The peaks in the residence time histogram (RTH) can be used to directly identify the location of the dwells, since their positions are well-separated (compared to noise) and the trace is monotonic, so the loss of time information in making the RTH is not a problem (if the trace went backwards, peaks would overlap with each other) (Figure 3.3b). The actual algorithm involves Gaussian-filtering the RTH to make the peaks smooth, so a simple peak detection algorithm (find where the first derivative changes sign) will robustly detect the locations. To remove potential misdetections that come from minor peaks (for example, from the burst in φ29), a "prominence" requirement is used, which will reject smaller peaks in the RTH. So, the parameters for this fit are the filtering of the data to generate the RTH, the width of the Gaussian filter to be used on the RTH, and the prominence cutoff for the peaks. This process only gives the spatial locations of the dwells, to get the temporal information, the data then needs to be fit to a "staircase" (a piecewise constant function) with steps located at the detected positions. This can be done with a hidden Markov model (HMM, see Appendix 4), where the underlying model is a chain that transits from the first step incrementally to the last one, a fit of this model to the data is shown in Figure 3.3d. From this staircase, we can extract dwell times, the time the motor spends between steps. Because the step locations are already set, and we just want to find the locations of the dwells, I call this an HMM *dwellfinder*.

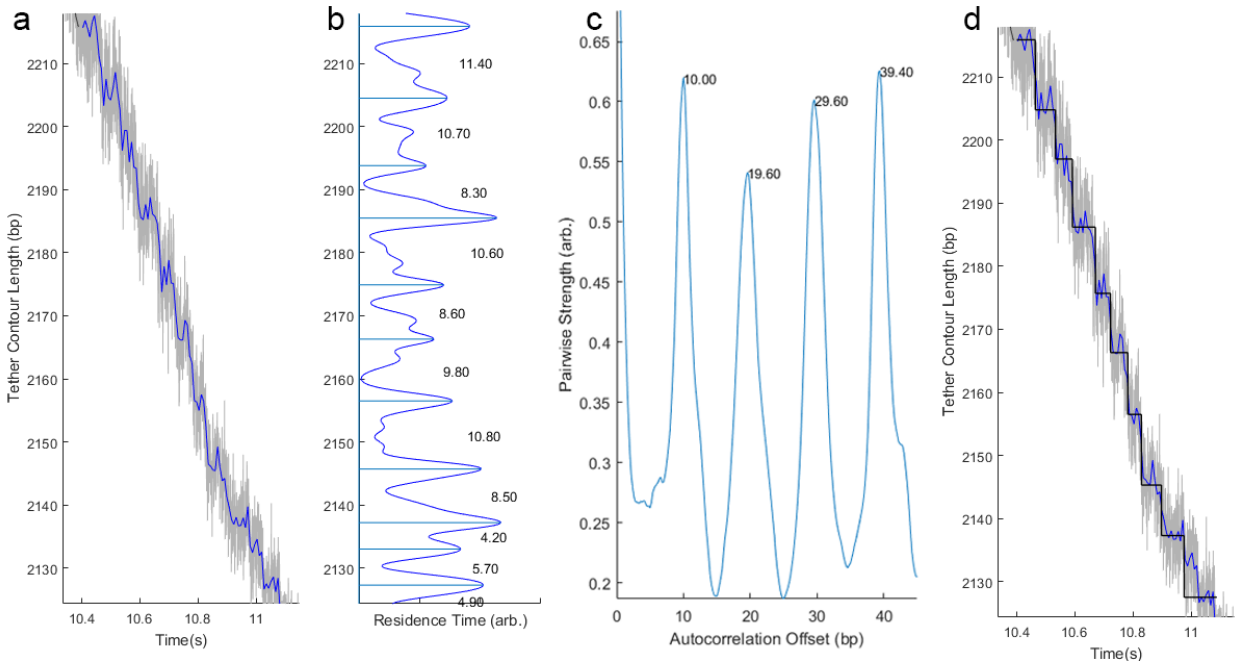


Figure 3.3 Pairwise distribution
a) Plotting of a trace at 2.5kHz (grey) and filtered to 100Hz (blue). b) The residence time histogram of the trace in a), with accepted peaks marked with lines, and step sizes (distance between peaks, bp) displayed as text. c) Pairwise distribution of the residence time histogram in b, with peak locations written. Note how the peak locations are at multiples of 10bp d) The trace in a) with the staircase found by the HMM dwellfinder overlaid in black.

Alternatively, stepfinding could be done via a point-of-change method, such as the Kalafut-Visscher algorithm[34]. This algorithm works by finding the location to place a step in the data that minimizes the quadratic error of the resulting staircase (the staircase heights are the mean of the data in the region) (Figure 3.4). Steps are added until the reduction in quadratic error is less than some cutoff. This value can be determined by the Schwarz information Criterion (SIC), but in actuality it ends up being a tunable factor to control when the algorithm stops finding steps. For example, a common value used for φ29 data is five times the cutoff determined by the SIC. Practically, this means we can continually add steps to reduce the average determined step size, which makes this algorithm less fit for determining step size and better for creating a staircase that has a mean step size that matches that found by a more robust metric, like the pairwise distribution, or for comparing the found step size between two different conditions.

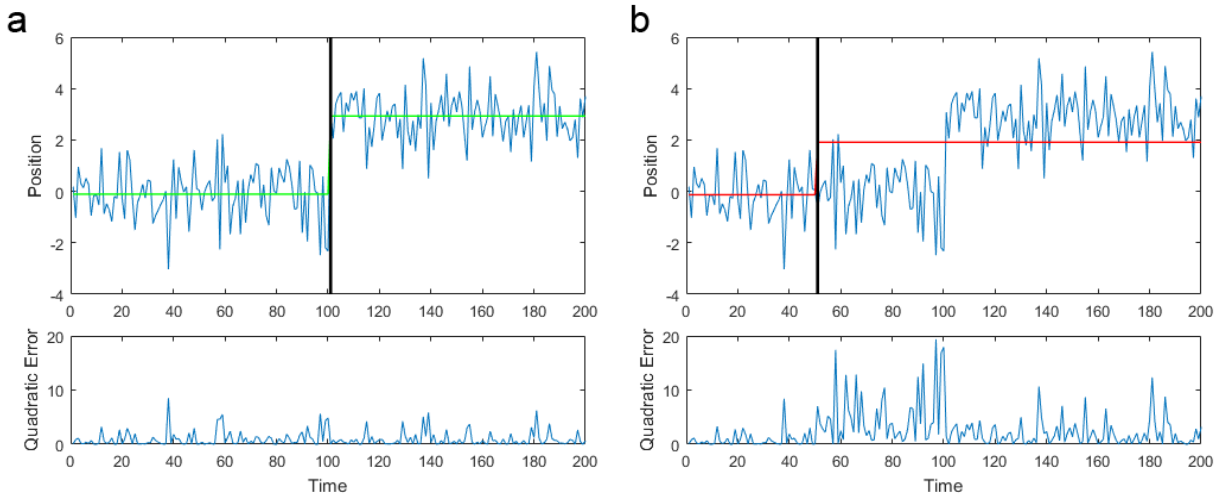


Figure 3.4 Kalafut-Visscher stepfinding algorithm
The Kalafut-Visscher algorithm finds the point in the data where, if a step was placed there, would result in the greatest decrease in quadratic error between the trace and the fit staircase. Shown is blue is a simulated trace with a step, which is Gaussian-distributed around position 0 before time 100, and centered around 3 after time 100. a) The proper step is at time 101, (black vertical line), which results in the green staircase (the staircase step heights are the means of the data in the range). The quadratic error by point is plotted below. b) Compare with an incorrect step, in this case at time 51, which results in the red curve with a higher quadratic error between the data and the fit curve. Hence, given the choice of these two steps, the algorithm would choose the step to create the green staircase. In actuality, the algorithm considers placing a step at every point in the trace, and chooses the one that best reduces the quadratic error.

Once we have found the dwells, we can then characterize their distribution (Figure 3.5). The distribution is peaked, so there are multiple rate-determining events occurring within a dwell (i.e., the distribution is not single-exponential). A way to characterize how many is by fitting the distribution to a Gamma distribution, the distribution found by the sum of multiple exponential distributions happening in series. The *shape factor* of the distribution corresponds to how many exponential distributions in series are there, and can be non-integer. The shape factor is then the minimum number of rate-limiting transitions for a process (each corresponds to one single-exponential). An equivalent metric is $n_{min} = \bar{\tau}^2/\sigma_\tau^2$, the mean squared divided by the variance, which is an estimator for the shape factor of a Gamma distribution.

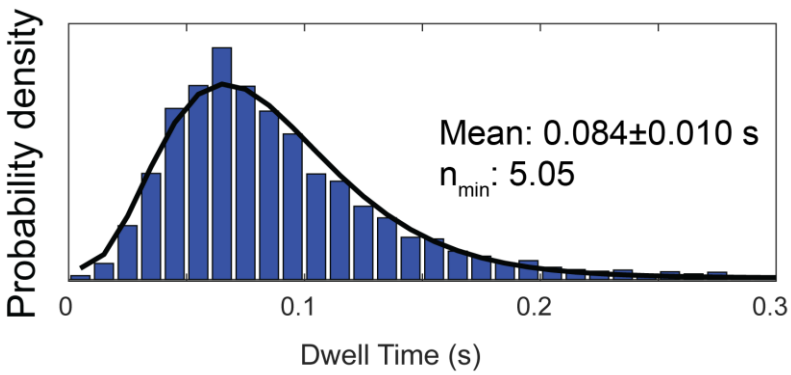


Figure 3.5 Dwell time distribution for φ29 and its fit to a Gamma distribution
The dwell time distribution of the φ29 motor fits well to a Gamma distribution. The mean and shape (n_{min}) of the fit are displayed.

Chapter 4: Tricking the φ29 motor to package dsRNA by initiation *in situ*

Applying the analysis methods described in the previous chapter suggested that the motor translocated slower on the chimeric substrate, that its burst size was shorter, and its dwell times were similar to those of DNA packaging. However, the poor throughput of the experiment performed this way meant that gathering enough data for meaningful statistics was impossible.

There is another strategy to perform the packaging experiment, where the initiation of the motor is done in the tweezer when the two beads are brought together. The act of bringing the motor-prohead complex and the DNA together is enough to initiate packaging, even when the DNA does not carry a gp3. This method is called *in situ* initiation. It should be noted that the knowledge to use this method was lost in the lab, but revived by me, and the documentation of the experimental protocols in Appendix 1 is an attempt to prevent this from happening again. The breakthrough was the incubation time: the complex needs 2 weeks of time sitting at 4°C before it works. The existing protocol said to wait “2 hours or more” but for whatever reason, now it takes 2 weeks or more. I have no way to rationalize what could be happening in the test tube that requires a timescale on the order of weeks, but it works. Complexes are also good to use for, say, 2 months or more. For more details, see Appendix 1. The packaging of the chimeric substrate can be repeated with this initiation method. The motor readily packaged the substrate, allowing for the collection of more data (Figure 4.1). In addition to the change in velocity at the DNA-hybrid junction, a change at the presumed hybrid-dsRNA junction is also observed.

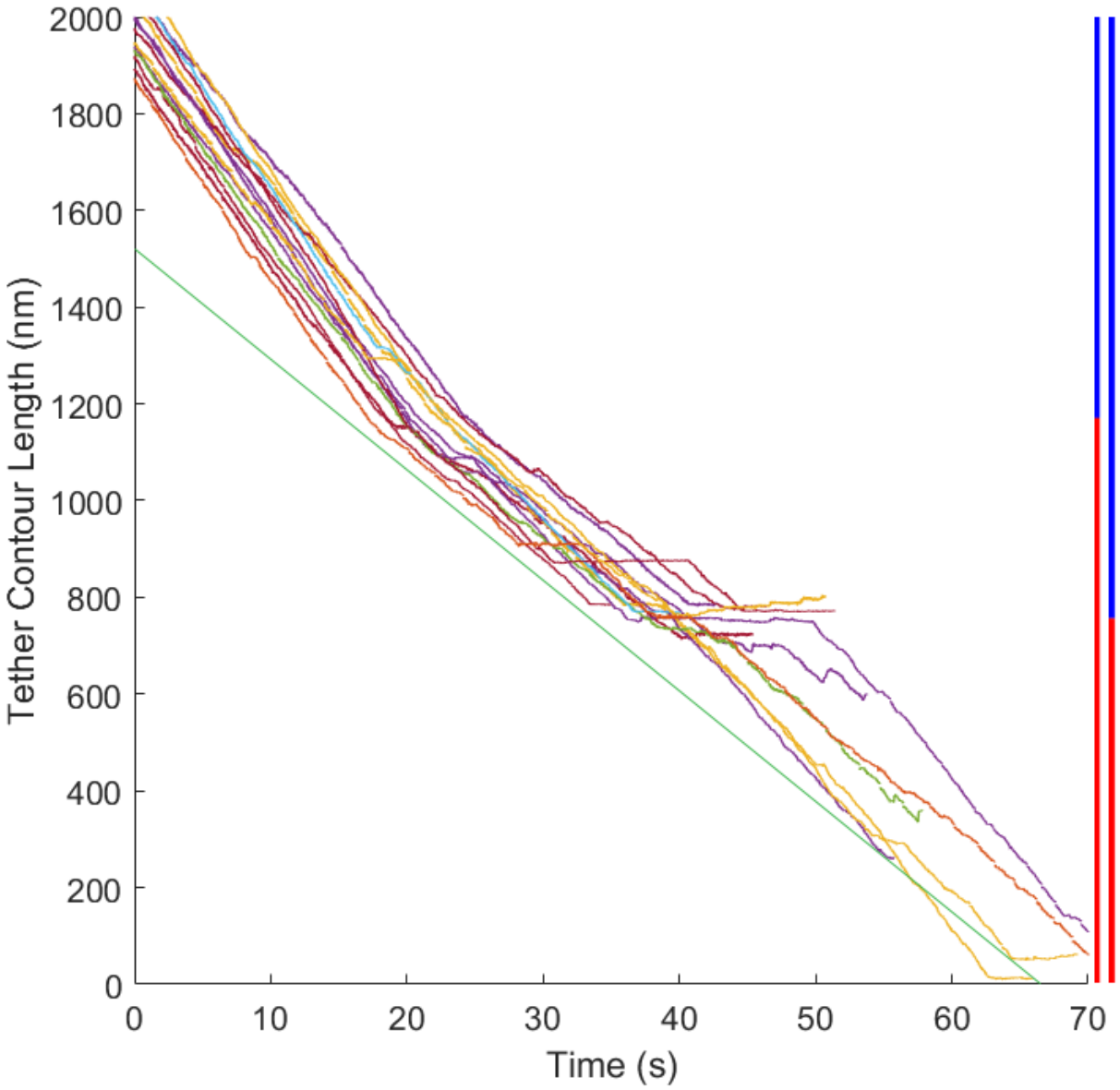


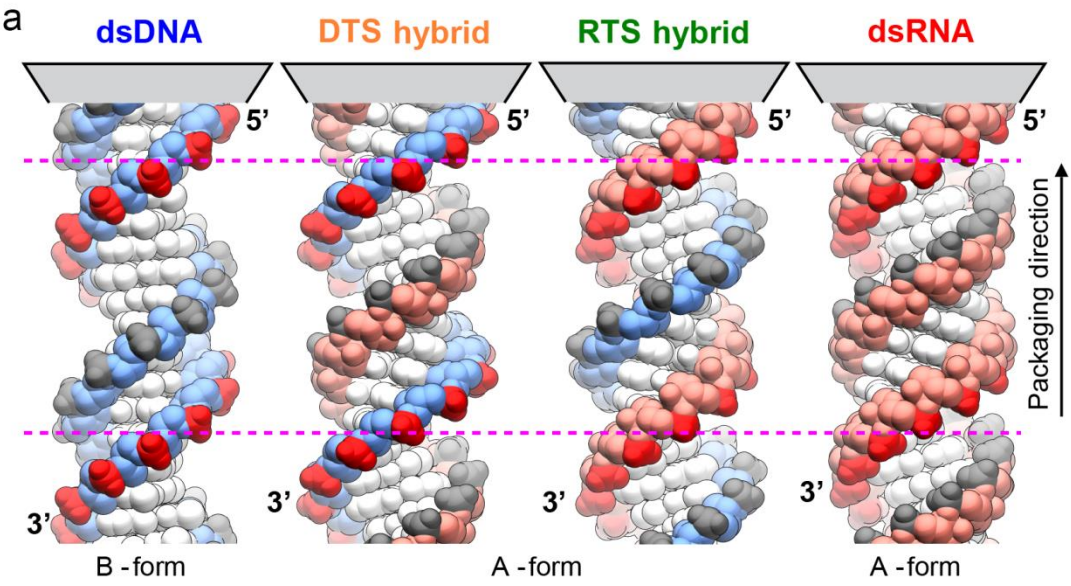
Figure 4.1 Packaging of the chimeric substrate
Packaging of the chimeric substrate, each trace colored a separate color and displayed at 250Hz. On the right is the predicted position on the substrate. Note the change in velocity at the DNA-hybrid and hybrid-dsRNA junction. The straight green line is parallel to the hybrid packaging region as a guide. Some traces have a propensity to pause at the hybrid-dsRNA junction. Data is taken at 7-12pN and 0.25mM ATP.

Note: From here on, DNA will be called dsDNA, to match the prefix on dsRNA, and lengths will no longer be in terms of bp of B-form dsDNA but rather nm, since the four substrates have different helical rises. The 10bp burst size of the motor on dsDNA is 3.4nm, and the step size of 2.5bp is 0.85nm. The content will also closely match that contained in Castillo & Tong 2021[35].

Since the DNA no longer needs to be tagged with gp3, the experiments can be done more simply now by just making a length of the alternate substrate (dsRNA or DNA:RNA hybrid) with a biotin on one end. Using this strategy, data on the alternative substrates became as easy to obtain as data on DNA. 4kb lengths of DNA, hybrid, and dsRNA were made with a biotin on one end for attaching to beads. Also, since the motor treats the two strands of the duplex differently, two hybrids were made, one in which the DNA is the tracking strand (*DTS hybrid*) and another where the tracking strand is RNA (*RTS hybrid*) to see how the motor handles differences in the identity of the tracking strand.

The structures of the four packaged substrates are shown in Figure 4.2a, and their relevant helical parameters are in Figure 4.2b[36-39]. The four substrates are color-coded, with dsRNA, DTS hybrid, RTS hybrid, and dsRNA corresponding to blue, orange, green, and red, respectively.

a



b

Structural data	dsDNA	hybrid	dsRNA
Periodicity (bp/turn)	10.4	10.8	11
Helical pitch (nm/turn)	3.52	3.11	3.02

Figure 4.2 The four nucleic acid substrates
a) Structures of the four double helical substrates generated by web.x3dna.com. The bases are colored white, the ribose sugars as red and deoxyribose sugars as blue, and the phosphates are red if they are on the tracking strand and grey if they are on the non-tracking strand. b) Helical parameters for the four substrates[36-39].

Single-molecule packaging traces of these four substrates are shown in Figure 4.3. The three A-form substrates are packaged at a slower velocity than dsDNA (compare with dotted blue line). In addition, the substrates that have an RNA tracking strand exhibit slipping events that happen at finite velocity (denoted by arrows).

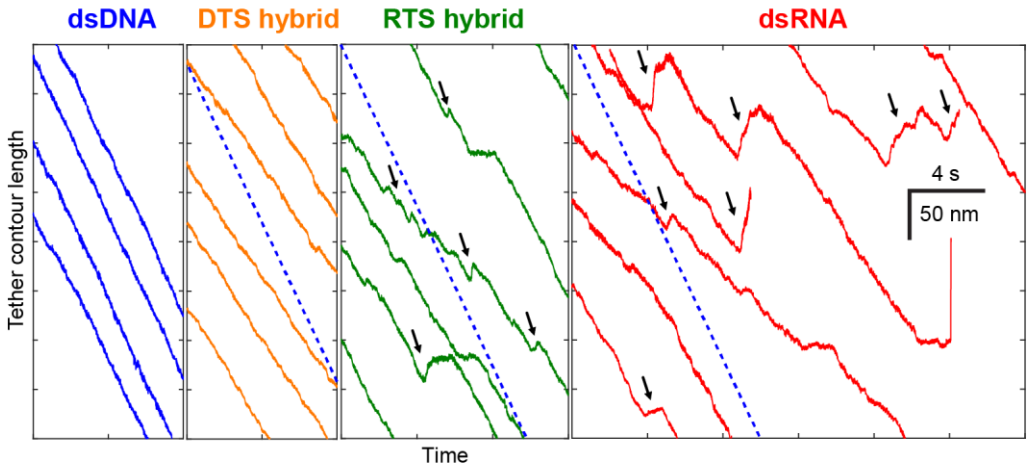


Figure 4.3 Single-molecule packaging trajectories of the four substrates Packaging traces of the four traces are displayed at 250Hz. The dotted blue line represents the speed of dsDNA packaging to compare to that of the A-form substrates. Arrows denote slipping events on RTS hybrid and dsRNA. Data is taken at 7-12pN and 0.25mM ATP.

To measure the velocity of packaging on these substrates, the velocity filtering method discussed in Chapter 4 was applied (Figure 4.3a). The three A-form substrates packaged slightly slower than dsDNA, with the four substrates packaging at 36.6 ± 0.5, 29.4 ± 0.2, 33.9 ± 0.3, and 30.6 ± 0.3 nm/s, respectively. Oddly, the motor packages RTS hybrid faster than it does DTS hybrid, even though the DNA tracking strand is more “normal” in this case. To speculate on a reason for this observation, the non-tracking strand of the substrate may play some role in the packaging process, namely, interactions with the motor and this strand may be involved in the signaling that ends the burst. Indeed, the cryoEM structure shows potential interactions with the special subunit and the non-tracking strand[24]. The motor also had a tendency to pause on the A-form substrates, evidenced by the population with zero velocity (dotted curve), with the propensity increasing from DTS hybrid to RTS hybrid and dsRNA. The slipping events manifest as a population with negative velocity in the distribution, and affirm that these events are not present when packaging dsDNA and DTS hybrid (the two substrates where the tracking strand is DNA) but do occur when packaging RTS hybrid and dsRNA (the two substrates where the tracking strand is RNA), with increased frequency on dsRNA (Figure 4.3b).

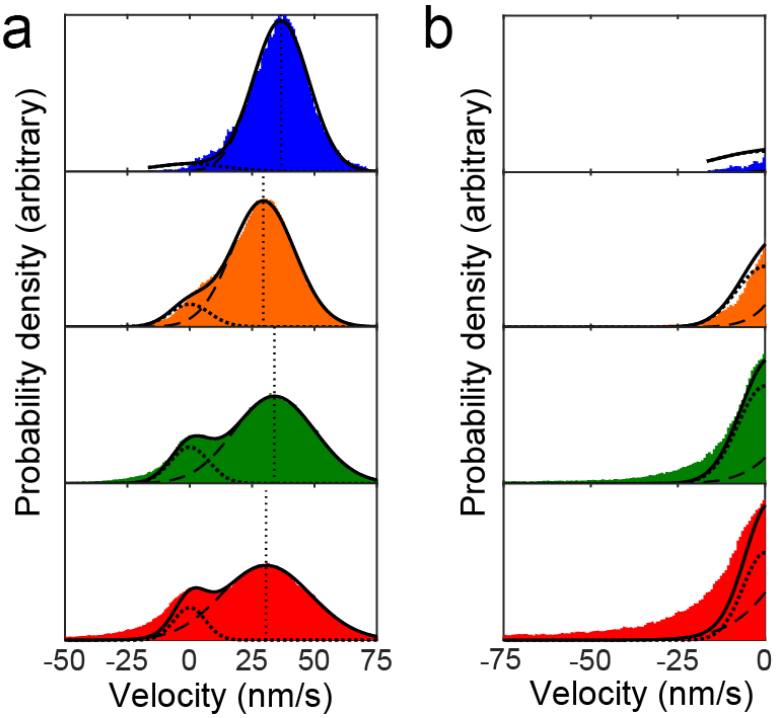


Figure 4.3 The packaging velocity of the four substrates
a) Velocity is measured by Savtizky-Golay filtering and binning. The resulting distribution is fit to a sum of two exponentials, one centered at 0 (representing pauses) and one with a positive velocity (representing the packaging velocity). b) Zooming into the negative velocities shows the relative amounts of slipping on the four substrates.

To determine the burst size of the motor on these substrates, the pairwise distribution method described in Chapter 4 was applied (Figure 4.4a). The signal of the periodicity was found to be low for RTS hybrid and dsRNA, repeating the analysis with only the top 25% of the data yields the curves in Figure 4.4b. The periodicity of the packaging of dsDNA was found to be 3.4nm as previously described, the periodicities on the two hybrids were found to be 3.0nm and the periodicity of dsRNA packaging was found to be 2.7nm. The hybrid burst size matches well with the helical pitch of 3.1nm, and the dsRNA pitch matches well to the length of 10bp of dsRNA (Figure 4.2b). The observation that the motor packages 10bp of dsRNA means that either the pitch of the dsRNA in this case is slightly under 11bp, causing the motor to round down to translocate 10bp as the burst size must be an integer, or that the size of the burst is defined by the translocation of 10bp rather than one pitch. Either way, these lengths are commensurate to the pitches of these substrates, so we conclude that the motor's burst size is determined by the pitch of its substrate. The motor having some sort of mechanism to adapt to the pitch is sensible, as dsDNA on its own exhibits sequence-dependence in its pitch size, so it is understandable that the motor can deal with those fluctuations. Representative traces showing the periodicity of the motor stepping is shown in Figure 4.4c. To get statistics on the step size, Kalafut-Visscher stepfinding was employed to yield an average step size of the four substrates of 3.39 ± 0.03, 3.09 ± 0.02, 3.02 ± 0.02, 2.70 ± 0.02 nm, respectively.

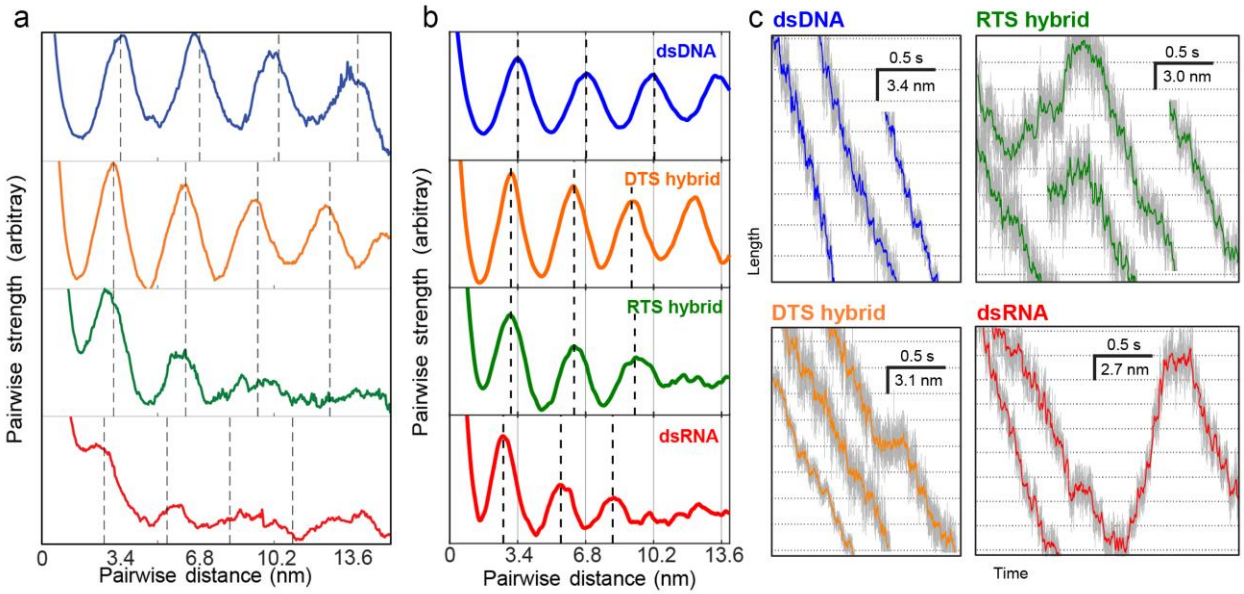


Figure 4.4 The burst size of the four substrates
a) The pairwise distribution of the packaging traces taken at 7-12pN applied force and 0.25mM ATP. The dsDNA and DTS Hybrid show a strong periodicity, while the RTS hybrid and dsRNA signal is weaker due to noise and slipping. b) Selection of the top 25% of the data improves the strength of the signal, revealing the periodicity of the RTS and dsRNA substrates. c) Representative packaging trajectories showing the periodicity of the stepping, with data at 2.5kHz in grey and 250Hz colored. The spacing of the dotted horizontal lines matches the periodicity determined by PWD.

It should be noted that signal-to-noise ratio (SNR) in optical tweezers is dependent on a number of factors, one of which is the effective spring constant of the tether[40]. Essentially, a stiffer tether yields a greater SNR; you can imagine that a floppy tether will dampen the response of the observed system and introduce noise, which will decrease SNR. The stretching of a DNA can be well-characterized by the extensible worm-like chain model (XWLC), which boils down the polymer's to its persistence length (unit nm) and its stretch modulus (unit pN). An experimental force-extension curve for DNA, hybrid, and dsRNA are shown in Figure 4.5. The polymer has very little force response until the force is brought very close to its contour length (the length the polymer would be if it is completely extended) where the force sharply increases. As the force continues to increase past 10pN, the curve levels out to a line with constant slope; in this regime the polymer is fully extended and the increase in length is from stretching the polymer. At forces beyond 40pN, the curve gets shallower as the regime approaches the melting or overstretching transition, where the helix unwinds at forces of ~60pN. The stiffness of the nucleic acid tether is the slope of this curve at the force the experiment is performed at. Since we are performing this packaging experiment at ~10pN, we are in the regime where the polymer is fully extended, and the increase in extension comes from the stretching of the bases. So, the stiffness of the polymer can be written as simply the stretch modulus divided by the tether length (by definition, the

stretch modulus is the force at which the polymer stretches to double its size). In packaging experiments, the tether length is continuously decreasing, leading to the phenomenon that as the tether length decreases, the SNR increases. What this means for the packaging of these A-form substrates is that the SNR on these substrates will be worse than on dsDNA, since the stretch moduli of hybrid (700pN) and dsRNA (450pN) are lower than that of dsDNA (900pN) obtained by fitting the curves to the XWLC. The differing signal qualities of the substrates will influence how analysis is done, including the previously noted trace selection for pairwise distribution and a choice later in finding the step size of the motor on the A-form substrates.

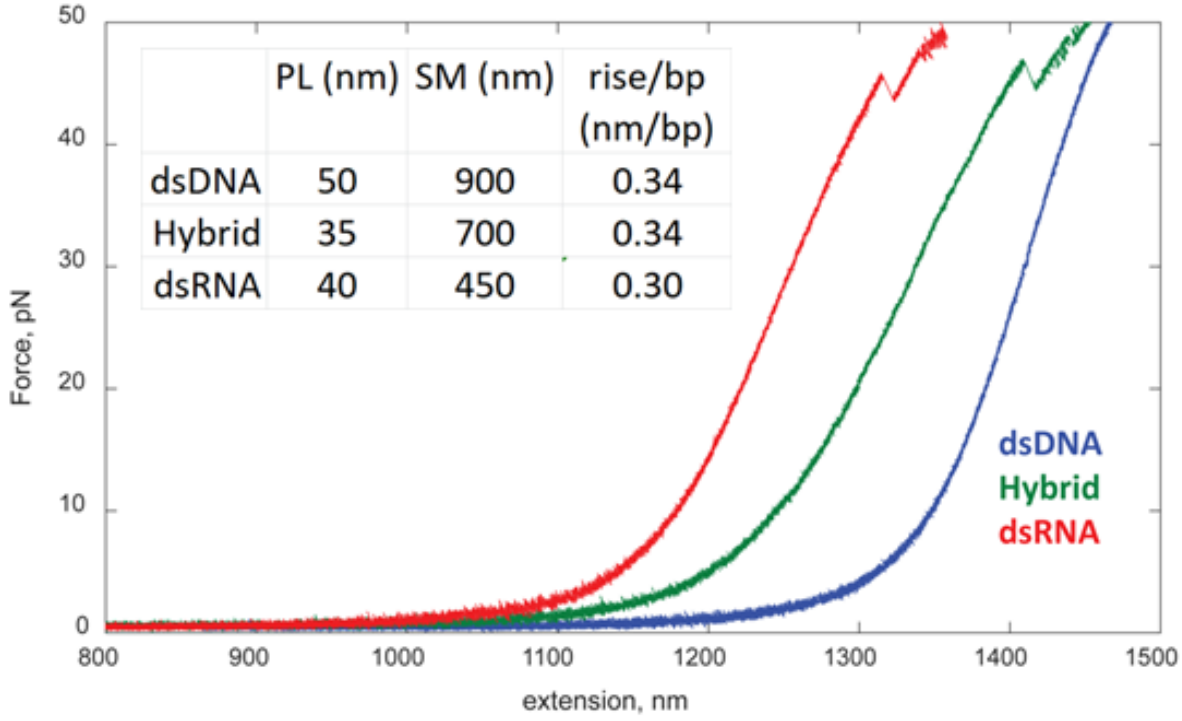


Figure 4.5 Stretching of the substrates
A 4kb length of each of the substrates is tethered in the tweezer and pulled to get its force-extension curve. Inset: Extensible worm-like chain parameters (persistence length, stretch modulus, and rise per basepair) obtained by fitting for the three substrates.

It is a good control to check that the motor's dwell is not significantly altered by the introduction of the alternate substrates. By applying the HMM dwellfinding method described in Chapter 4, the dwell times of the motor were measured and their distribution is shown in Figure 4.6. All dwell time distributions follow a Gamma distribution with shape around 5, consistent with the dwell of the motor being defined by the five ATP exchange events. Indeed, we see that the dwell time of the DTS hybrid is larger than the others, consistent with its slower observed overall velocity. That said, it is reasonable to say that the dwell of the motor is mostly unchanged, with potentially some kinetic rates altered by the DTS hybrid.

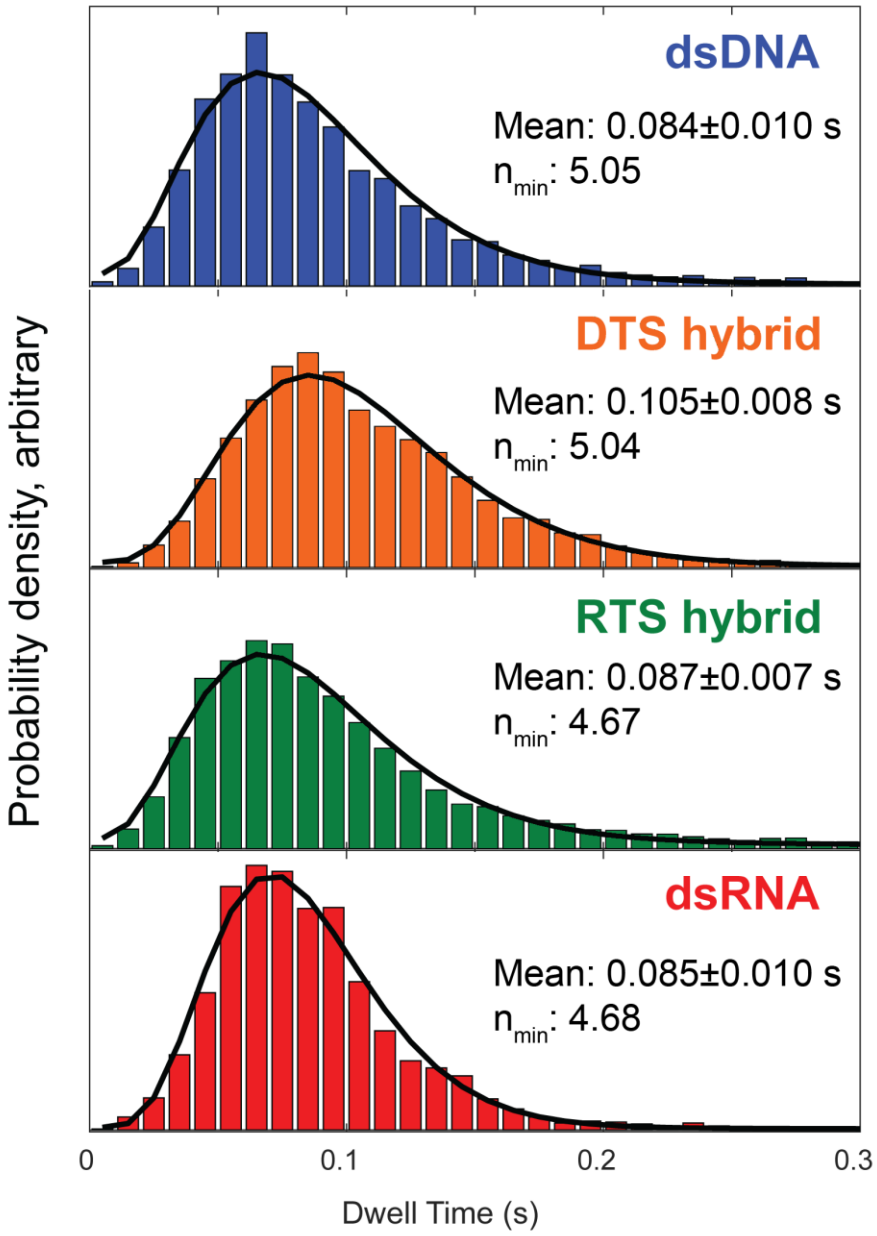


Figure 4.6 The dwell times of the four substrates
The dwell time of the substrates is found via the HMM dwellfinder (see Chapter 3) and fit via maximum likelihood estimation to a Gamma distribution plus a single-exponential (to capture motor pausing). The dsDNA did not need fitting to the single-exponential, as it did not have a sizable paused population. Inset text shows the mean and n_{min} of the fit.

Chapter 5: The step size of φ29 on DTS Hybrid and burst-sized slipping

To obtain the step size of the motor, the size of the individual translocations that make up a burst, data needs to be taken at high force (30-35pN). The application of force opposing the direction of motion slows down the mechanical transitions of the motor, i.e.

the steps that compose the burst, and also decreases the noise of the tether extension by holding the tether tauter.

When packaging dsDNA, the separation of the burst into distinct steps is moderately evident, but can still be difficult to observe. Historically, the analysis of these steps was either done by pairwise distribution, where the packaging traces at high force are shown to have a 2.5bp periodicity [20], or by stepfinding, where a stepfinding algorithm attempts to detect the steps directly[22]. To find the step size on the A-form substrates, which have lower SNR than dsDNA, will be even harder, then. In addition, the slipping events are exacerbated at high applied force, making the taking of data more difficult on RTS hybrid and dsRNA. Hence, the decision was made to only quantitatively determine the step size for DTS hybrid.

How could the step size change to fit the smaller burst size of the motor on DTS hybrid? The two simplest cases are either that the step size proportionally decreases too (still subdividing the 3.0nm burst into four equal steps of 0.75nm each), or that the step size does not change (the smaller burst size results from the shortening of one of these steps, yielding three 0.85nm steps and one 0.45nm step). To differentiate between these two possibilities, it should first be noted that pairwise distribution does not work well for a non-uniform step size, as there will be signal from all combinations of step sizes. An example on simulated data is shown in Figure 5.1. The PWD of the evenly spaced steps (blue) show peaks at multiples of 0.75nm, while there isn't an obvious pattern in the orange PWD (Figure 5.1ab). With perfect resolution, the smaller step will create a PWD with peaks at twice as many locations (0.45, 0.85, 1.30, 1.70, … nm), requiring twice the spatial resolution to resolve them compared to consistent 0.75nm-spaced peaks (Figure 5.1c). If resolution is not enough to separate the peaks in the PWD, they blend together to create the flat valley in Figure 5.1b.

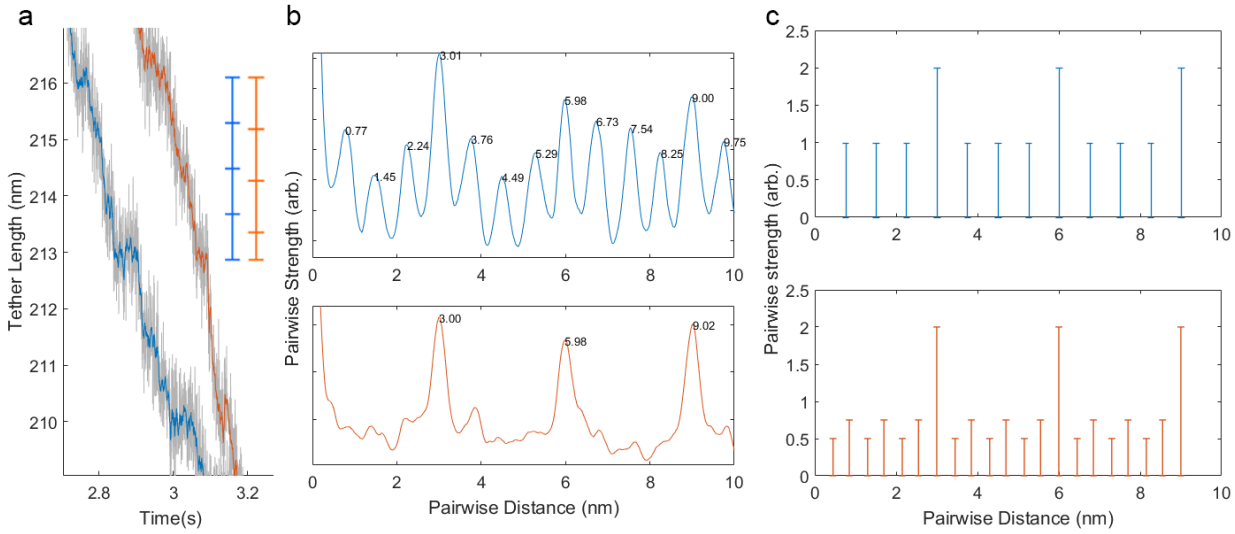


Figure 5.1 Pairwise distribution of a uniform and non-uniform step size
a) Simulated data showing a trace stepping at 0.75nm x4 (blue) or 0.85nm x3 + 0.45nm (orange). Data is at 2.5kHz (grey) and filtered to 250Hz (colored) with 0.5nm Gaussian noise to roughly match that of actual tweezers data. The ATP

loading dwells are slightly longer than the sub-dwells between the steps of the burst. b) PWD of the data in (a), with peaks labeled. The peaks at 3nm and multiples are taller than the rest because the dwells every 3nm are longer than the sub-dwells. c) Theoretical PWD of the data at perfect spatial resolution. Peaks are infinitely thin, so are shown here as vertical lines. The blue PWD shows peaks at multiples of 0.75nm (0.75, 1.50, 2.25, and 3.00 nm), while the orange PWD has peaks at 0.45, 0.85, 1.30, 1.70, 2.15, 2.55, and 3.00 nm.

Because of the issues that could arise with determining the motor's step size on the A-form substrates via PWD, stepfinding was used instead. The Kalafut-Visscher algorithm was deemed unfit for the job, because the arbitrary nature of choosing an endpoint for stepfinding can yield any result as a step size. Instead, a stepfinding algorithm based on a hidden Markov model was used (see Appendix 4). The benefits of this algorithm over the Kalafut-Visscher one is fewer tunable parameters. The hidden Markov model needs a starting model guess, in this case, a flat step size distribution, but otherwise has no tunable parameters, so I see it more fair to use in this case. Running the HMM stepfinder on the two datasets here results in the two step size distributions in Figure 5.2a. Note how both peak at 0.85nm (black dotted line), and that the DTS hybrid distribution definitely does not peak at 0.75nm (pink dotted line), the step size if the burst was subdivided into four equal steps. Taking the difference between the two curves yields probability density centered near 0.45nm and 1.3nm, which are the sizes expected for the shorter step and the shorter step plus a regular step (0.45nm+0.85nm) which can show up if the two are not separated. Hence, this density corresponds to the detection of the correction steps, adding further support for the model that the step size of the motor on DTS hybrid is three steps of 0.85nm and one step of 0.45nm. Hence, we conclude that the step size of the motor is not dependent on the structure of the substrate, but rather determined by a conformational change in the motor.

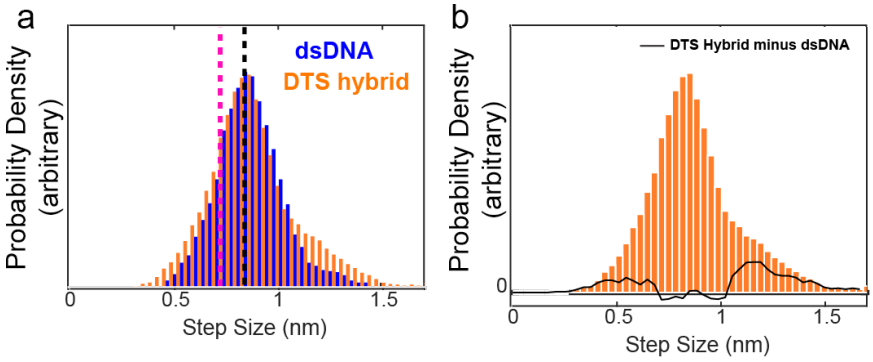


Figure 5.2 Step size distribution of φ29 on dsDNA and DTS hybrid
a) The step size distribution of dsDNA (blue) and DTS hybrid (orange) found by HMM stepfinding of traces taken at 30-35pN, 0.25mM ATP. The two distributions share a peak at 0.85nm. The dotted vertical lines correspond to 0.75nm (magenta) and 0.85nm (black), respectively, the two major step sizes of the two possibilities. b) By subtracting the DTS hybrid and dsDNA curves, we can isolate the extra step density found in the DTS hybrid on the left and right of the central peak (black curve). Subtracting the two curves should better handle systematic errors

introduced by the algorithm (compared to subtracting a Gaussian centered at 0.85nm, for example).

Since the step size on DTS hybrid is three steps of 0.85nm and one of 0.45nm, it is reasonable to believe that the adaptation by the motor is the same on the other substrates. So, we allege that the step size on RTS hybrid is the same three steps of 0.85nm and one of 0.45nm as on DTS hybrid, and the step size on dsRNA is three steps of 0.85nm and one of 0.15nm. Knowing this result, we show representative traces of the φ29 motor translocating on dsDNA and DTS hybrid (Figure 5.3). In addition, we see sections of the motor translocating on RTS hybrid and even dsRNA that exhibit the predicted stepping pattern of three 0.85nm steps and one correction step. When the correction step occurs compared to the dwell is unknown, under these conditions the duration of the dwell is the similar to the duration of the time between steps, but we can assume that it is either the first or last step in the burst.

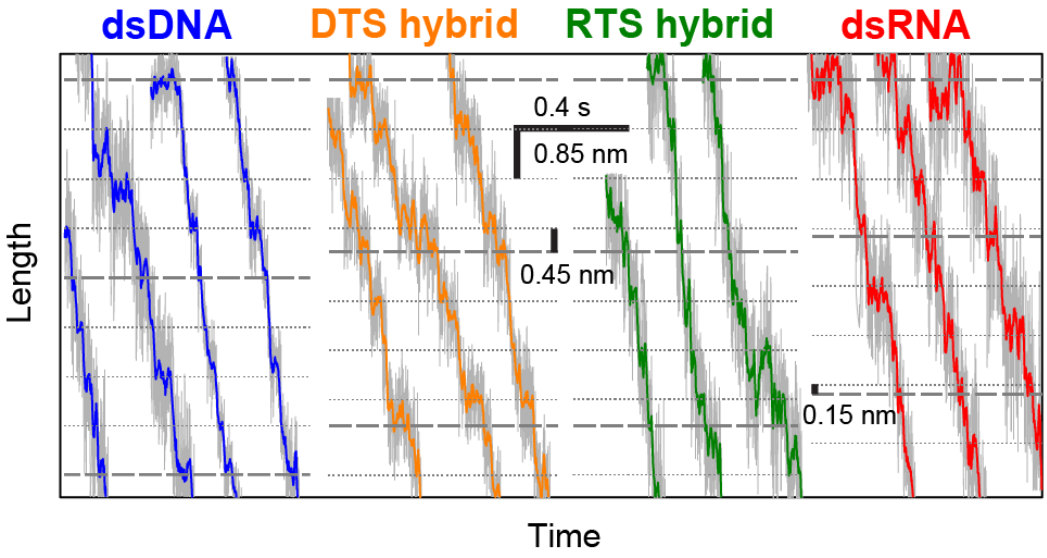


Figure 5.3 Stepping on the four substrates
Single-molecule packaging trajectories taken on the four substrates at 30-35pN force and 0.25mM ATP. Raw data is shown at 2.5kHz (grey) and 250Hz (colored). Dotted horizontal lines match the predicted stepping periodicity of the substrate (four 0.85nm steps for dsDNA, three 0.85nm and one correction for the others).

The slipping events on dsRNA and RTS hybrid are of finite velocity, different from the near-instantaneous slips that the motor has been shown to do in the past at low [ATP] (Chemla 2005). Upon further inspection, we observe that the events look like the bursts of the motor—they are stepwise with size equal to the burst size on dsRNA; hence, we term this event *burst-sized slipping* (Figure 5.4). The events can be divided up into the steps that compose the event, which we can find by the Kalafut-Visscher stepfinding algorithm (see Chapter 4). We define the *start dwell* as the dwell that immediately precedes an event, and the *end dwell* the dwell that immediately ends an event; the dwells in between we term *slipping dwells*.

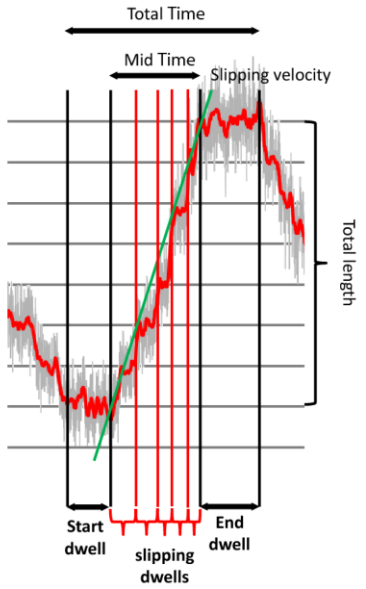


Figure 5.4 Burst-sized slipping on dsRNA
A burst-sized slipping event is shown at 2.5kHz (grey) and 250Hz (red). It can be divided into three sections: The start dwell, where the motor pauses before slipping, the slipping time, where the motor slips in a stepwise manner, and the end dwell, where the motor pauses before packaging resumes. Observe how the slipping is of constant, finite velocity (green line). Data is taken at 10pN constant force and 0.25mM ATP.

Why is the motor slipping in a burst-sized manner? First, the events only happen when the tracking strand is RNA, suggesting they are a result of poor contact between the motor and the tracking strand. Indeed, observe the differences between the orientation of the phosphates in the tracking strands, where they are turned more “inwards” on the RNA tracking strands compared to on the DTS hybrid and dsDNA (Figure 4.2a, red phosphates). To further investigate what the motor is going through during this event, we compare the characteristics of these events at “normal” conditions (0.25mM ATP and 10pN applied force) to them at lower [ATP] (0.025mM) and separately at high force (30pN) to see if the rate of ATP loading and magnitude of applied force have an effect on the events. Low [ATP] increases the frequency (events per second) and density (events per nm) of these events, but the characteristics of the events were unchanged compared to “normal” conditions (Figure 5.5). From these results we note that the slower exchange of ADP for ATP during the dwell leads to an increase in the chance to enter an event, suggesting that there is a kinetic competition between the end of the dwell and the entry to a burst-sized slipping event. The insensitivity of the event to [ATP] suggests that the normal action of the motor (ATP loading) is halted during the event, and so we call the burst-sized slipping state a packaging-incompetent state. Increased applied force also increases the frequency and density of events, and increases the velocity of slipping. So, force can also bias the motor to enter this state. We conclude then that the burst-sized slipping event occurs when the motor enters a state of the motor that is unable to bind nucleotide, and has worse grip for the substrate. The motor then slips successively in spurts of one burst at time, briefly reattaching each turn of the

helix as the motor reencounters the tracking strand of the substrate. The observation that the attainment of this state can be avoided by saturating the ring with ATP suggests that the motor's grip for the motor is dependent on the number of ATP molecules bound, where the motor fully-bound with ATP cannot slip. Indeed, this correlation between ATP bound and grip has been observed when φ29 packages dsDNA[19].

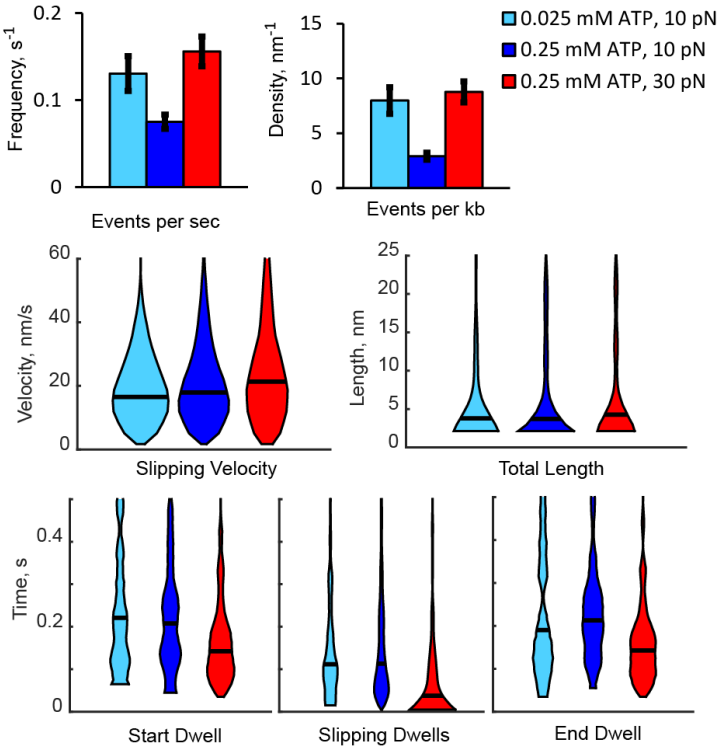


Figure 5.5 Statistics of burst-sized slipping on dsRNA
Statistics of the event frequency and the statistics shown in Figure 5.4 are quantified. Frequencies are shown as mean ± sem, and the rest are shown as violins with horizontal line at the median. Three conditions are shown, 0.025mM ATP and 10pN constant force (cyan), 0.25mM ATP and 10pN constant force (blue), and , 0.25mM ATP and 30pN constant force (red).

Chapter 6: The helical inchworm model of translocation

The results of packaging A-form substrates contain a number of interesting observations. The first is that the burst size adapts to the pitch of the helix, confirming the source of the motor's burst size on dsDNA is indeed the pitch of the substrate. The linear translocation of 3.4nm of DNA keeps the relative positioning between the motor and the tracking strand of the DNA constant, and preserves the identity of the special subunit (the one that attaches to the DNA) across cycles. The adaptation of the burst size to the substrate's pitch suggests that the motor may have a mechanism for measuring the pitch, which could come from the opening of the motor ring to span one pitch as seen in cryo-EM. This tracking of one strand of the helix assumedly grants the motor grip strength, increasing the surface area by which the motor interacts with the duplex. The burst-sized slipping phenomenon and the observation that the motor's grip

depends on its degree of ATP saturation suggests that this interaction with the tracking strand is only progressively achieved with ATP loading, suggesting that there is a non-lock washer state of the motor, namely a planar state. Additionally, if the motor were to stay in its lock-washer state during the burst, the motor and the tracking strand would become off-register after just one step (Figure 1.5a). Hence, we conclude that the ATP-full motor has the lock-washer shape observed in the cryo-EM structure, and the ADP-full motor instead exhibits a planar structure. This sequential opening and closing of the ring allows the motor to translocate its substrate while keeping contact with the tracking strand at all times (Figure 6.1). In this *helical inchworm* model, starting from a planar, ADP-full motor at the start of the dwell, ATP exchange sequentially opens the ring, following the pitch of the DNA. As more subunits contact the tracking strand of the DNA, the grip of the motor for the substrate increases. After exchange has completed, the special subunit receives a signal to hydrolyze its ATP, starting the burst. This signaling probably does not come directly from the subunit that just exchanged ATP, as the interface between the special and this subunit is the crack in the ring[24]. Instead, we propose that the tension stored in the ring by the opening is what signals the special subunit to hydrolyze its ATP. Next, the translocating subunits hydrolyze their ATPs, translocating 0.85nm of dsDNA into the capsid with each step, reverting the ring to a planar structure. Note that the amount the ring opens and closes is dependent on the "hinges" that are formed between the adjacent subunits (cylinders in Figure 6.1), which explains why there are four steps on this motor (there are only four hinges in a lock washer made up of 5 subunits). If redrawn with the position of the DNA held constant, the model can be seen as the motor inchworming its way up a helical staircase (the DNA), which is where the model gets its name.

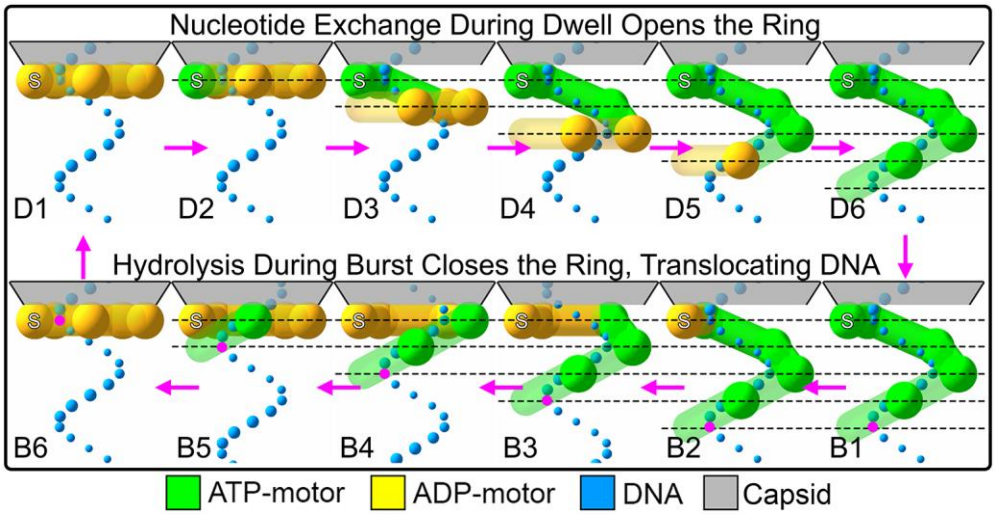


Figure 6.1. The helical inchworm mechanism on dsDNA
The helical inchworm mechanism of translocation is shown. The DNA-contacting surface of the motor subunits are represented by a sphere connected to a cylinder, and are colored by nucleotide state (green for an ATP-bound subunit, yellow for ADP-bound). The capsid is above in grey, and the phosphates of the tracking strand are shown as blue dots. Horizontal dotted lines have 0.85nm spacing. The twelve states of the motor are labeled D1-D6 and B1-B6, denoting

the stage of the Dwell or Burst. State D1 and B6 are identical, and state D6 and B1 are identical. Pink arrows show the relation between the states. In the Burst images, one phosphate is colored pink to guide the eye to the movement of the DNA. At the start of the dwell, the ring is planar and ADP-full. As the dwell progresses, ADP is exchanged for ATP, opening the motor to track one pitch of the DNA. Once open, the hydrolysis of the special subunit occurs, followed by hydrolysis-coupled closing of the ring, concurrent with DNA translocation. At the end of the burst, the ring is flat again, and the cycle repeats.

When the motor tries to translocate dsRNA, the cycle is mostly the same, with the differences occurring in state D6, where the final opening step is shortened to find the phosphate of the substrate one pitch down from the motor, and in state B5, where this shorter step now closes to complete the translocation of one pitch of substrate (Figure 6.2).

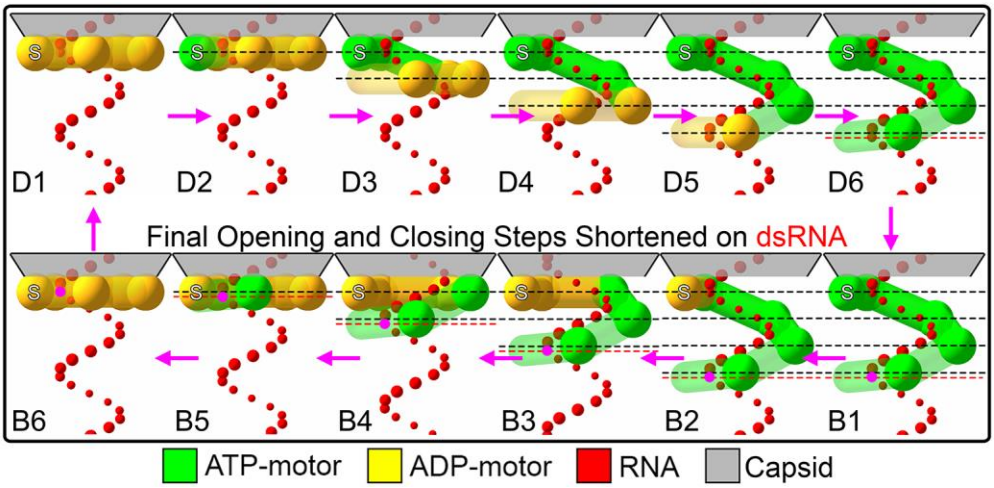


Figure 6.2. The helical inchworm mechanism on dsRNA
The adaptations of the helical inchworm mechanism to translocate dsRNA is shown. The components are the same as Figure 6.1, except the DNA is replaced with RNA (red). The changes from the motor's operation on dsDNA manifest in state D6, where the motor has to open less (red line, spacing 0.15nm) to span one pitch of the substrate, and in state B5, where the smaller opening in D6 now closes, resulting in the final 0.15nm step.

And so, given the four models that were originally discussed at the end of Chapter 1, we assert that a slightly modified version of the lever-latch mechanism is the one that this motor uses, and call it the helical inchworm. In addition to the model offering sensible explanations to why the pentameric motor only makes four steps (a five-membered lock-washer has four subunit interfaces), how it translocates exactly one pitch of substrate (it's how much the ring opens), and why the step size of the motor is what it is (the subunit interface hinges have a set opening amount), it is interesting to contrast this model against the presiding model for most other lock-washer ring ATPase translocases, the hand-over-hand model (for a refresher, see the end of Chapter 1). A common factor between these hand-over-hand motors is that their substrate is unstructured,

translocating ssDNA or polypeptide[25-27]. In these cases, the helical structure of the motors enforces a helical structure onto the unstructured substrate, creating an ordered structure that the motor can grip and translocate. We believe that this difference is the reason why the φ29 dsDNA translocase acts differently, as its substrate has a well-defined helical structure, and does not need to impose order on its substrate. Instead, it can take advantage of the preexisting structure, using the repeating phosphates every 10bp like rungs on a ladder. To continue the metaphor, the φ29 motor can be personified as a person climbing a ladder, with rungs defined by the pitch of the substrate, while the hand-over-hand motors can be personified by a person climbing a rope, where they must deform the rope in order to grip it better. Along the same lines, climbing a ladder is easier than a rope because it is easier to grip, which symbolizes the higher forces that the φ29 motor can achieve (60pN) compared to that of protein translocases (15pN)[19,41].

While we believe this model to be the most parsimonious one given the plethora of observations made on this motor, we can go further and more directly prove its existence. In single-molecule studies, we can find direct evidence for the open-to-planar ring cycling by the simultaneous measurement of translocation by optical tweezers and ring opening by single molecule Förster resonance energy transfer (smFRET) using a "fleezer," an optical tweezer that combines trapping and fluorescence[42]. By introducing two dyes, e.g. one on the N-terminus of the motor (which should move with ring opening) and one on the C-terminus (whose position should stay constant), we can monitor the relative position of the two domains via FRET and use this as a readout for ring opening and closing. A schematic of this experiment is shown in Figure 6.3a. Alternatively, we can solve the structure of the motor in other states of the cycle; at the moment, we only have the motor in state D6 (named as in Figure 6.1), but other states of the motor are potentially solvable via cryoEM (Figure 6.3b). For example, studies in the 26S proteasome have concluded that it operates via the hand-over-hand mechanism by the arrangement of several substates of the motor[26,27]. However, it is difficult to trap the φ29 motor in e.g. an ADP-full state to capture state D1, as the motor in this state is not stable.

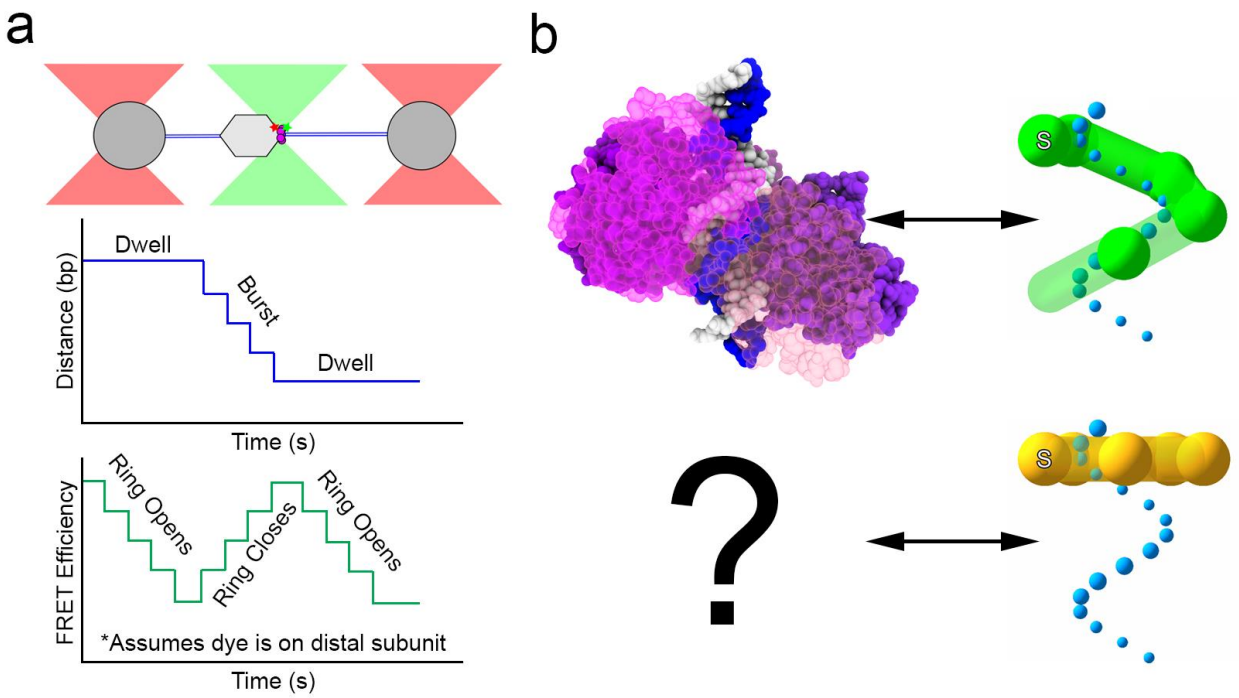


Figure 6.3 In search of more definitive proof of the helical inchworm model
Two potential experiments to obtain definitive proof of the helical inchworm model. a) Single molecule optical tweezers combined with FRET can be used to view the ring opening in closing in real time. Using a dye on the motor's N-terminus and a reference dye somewhere else, the ring opening can be correlated to change in FRET. The optical tweezers channel reads out the dwell-burst state of the motor, while the FRET channel shows the ring open and close, with ring closure linked to the motor burst. A cartoon of the experiment is above, with the expected data is below. b) Finding missing states of the model via structure would also provide evidence for the model. The most useful would be to capture the motor in the ADP-full state, where the N-terminal ring is presumed to be in a planar structure. This state combined with the already solved ATP-full lock-washer will prove that the motor cycles between planar and lock-washer structures.

Chapter 7: Single-molecule studies on RNA polymerase

There are motors other than the φ29 DNA translocase that are interesting to study. One such motor is the RNA polymerase (RNAP), the motor that moves along DNA to transcribe a copy as RNA. As the machine responsible for the first step of the central dogma of molecular biology, its importance needs no introduction, and learning about how it works can reveal the myriad complex methods by which gene expression is controlled. Motor pausing, usually seen as detrimental as it halts the activity of the motor, in RNAP has a beneficial role as it can give time for transcription factors to bind, control the folding of regulatory RNA secondary structures, or signify a paused motor in need of rescue[43,44]. The resolution in single-molecule optical tweezers is sufficient to observe individual nucleotide addition cycles, and so the pausing events are directly

observable as longer dwells in the elongation trace[45]. The applied force can be used to *assist* or *oppose* the motion of the motor, and can be applied to the transcript to affect the folding of the nascent RNA. Diagrams of the two geometries of RNAP experiments are shown in Figure 7.1.

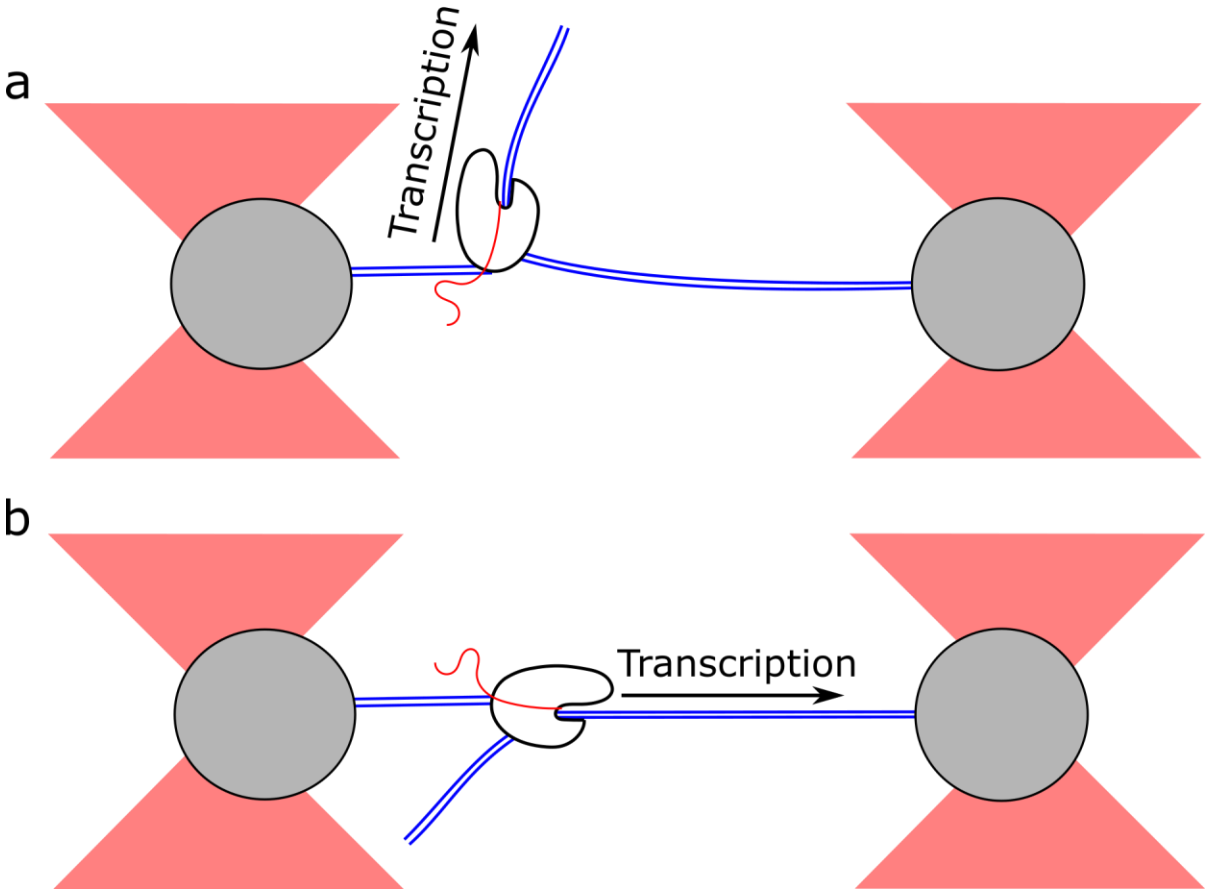


Figure 7.1 Single molecule force spectroscopy experiments with RNAP Tethering geometry for the assisting force and opposing force setups. Shown are the traps (red cones), beads (grey spheres), RNAP (white bean-shaped blob), DNA handle (left blue lines), DNA template (right and middle blue lines), and RNA transcript (red line). An arrow shows the direction the RNAP travels as it transcribes. a) In the assisting force setup, transcription increases the length of the tether, resulting in the force acting to aid transcription. b) In the opposing force setup, transcription shortens the length of the tether, resulting in the force acting against the movement of the motor.

The current model for RNAP elongation is a Brownian ratchet. The RNAP switches between the *pre-translocated* and *post-translocated* state by thermal fluctuations, and the movement is rectified by addition of an NTP. In each cycle, starting from the RNAP in the pre-translocated state, the RNAP must *translocate*, the cognate NTP must *bind*, and then the *catalysis* of the phosphodiester bond must occur to end up in the pre-translocated state again. A cartoon of this nucleotide addition cycle is in Figure 7.2a. The three steps of translocation, binding, and catalysis make up the steps that must

occur each dwell. The relative rates of each of these three steps will define the distribution of the regular (non-paused) dwell times. If all three steps are of similar rate, the shape will be Gamma-distributed with shape factor 3, but if only one is rate-limiting, the dwells should be single-exponentially distributed. In a single molecule optical tweezers experiment, the distribution is often single-exponential due to the applied conditions. The most common experiment is to tether the RNAP in the assisting force geometry and provide saturating NTP concentrations, since these conditions create a polymerase that transcribes consistently with less pausing. The assisting force biases the RNAP into the post-translocated state, making the translocation step very fast, and the saturating NTPs makes the binding step very quick, too, leaving catalysis as the sole rate-limiting step. When only one step is rate-limiting, the nucleotide addition cycle can be collapsed to a single step with rate constant k_n. Pausing can be added to this kinetic model as an off-pathway state with entry rate k_p (Figure 7.2b). Note how there is a kinetic competition between pausing and nucleotide addition, in other words, pausing efficiency ($E_P = k_p/(k_p + k_n)$) decreases with increasing k_n. Performing experiments in assisting force mode can also prevent or rescue an RNAP from *backtracking*, an off-pathway state where the polymerase moves backwards on the template, extruding its transcript.

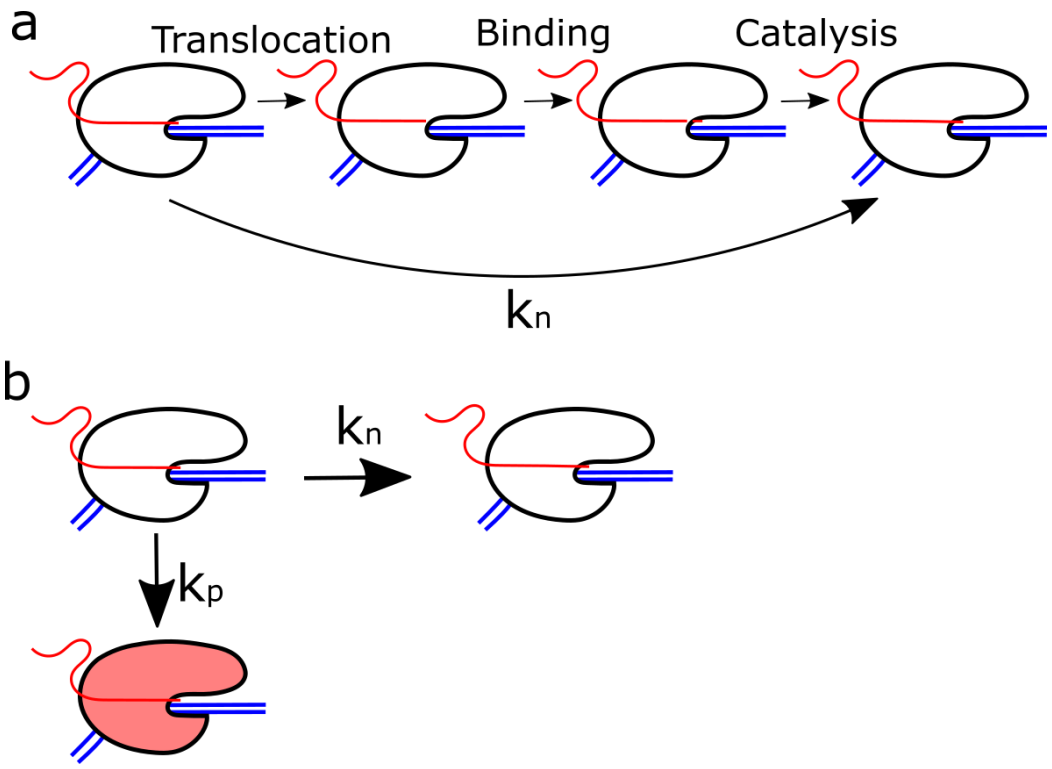


Figure 7.2 RNAP elongation cycle
a) The nucleotide addition cycle of RNAP consists of three steps, translocation, binding, and catalysis. Translocation involves the movement of the motor to the right, freeing the catalytic site (armpit of the bean). Binding is the arrival of the cognate NTP (red dash), which then is incorporated into the transcript via the Catalysis step. The three steps are often combined to one step with rate k_n. b) At

every step, the RNAP may either transcribe at a rate k_n or enter a paused state (red-shaded RNAP) at a rate k_p.

The velocity analysis method discussed in Chapter 4 can be applied to a RNAP trace to obtain the *pause-free velocity* (PFV) of the motor ($1/k_n$). However, unlike the peaked dwell time distribution of φ29, the single-exponential distribution of the dwell times means that the average velocity determined by filtering will have a wide distribution unless heavy filtering is employed (Figure 7.3a). In addition, the shortest pauses of the motor are not that much longer than the dwells, so the required filtering also smooths out these short pauses, leading to errors in the fitting process for identification of the PFV (Figure 7.3b).

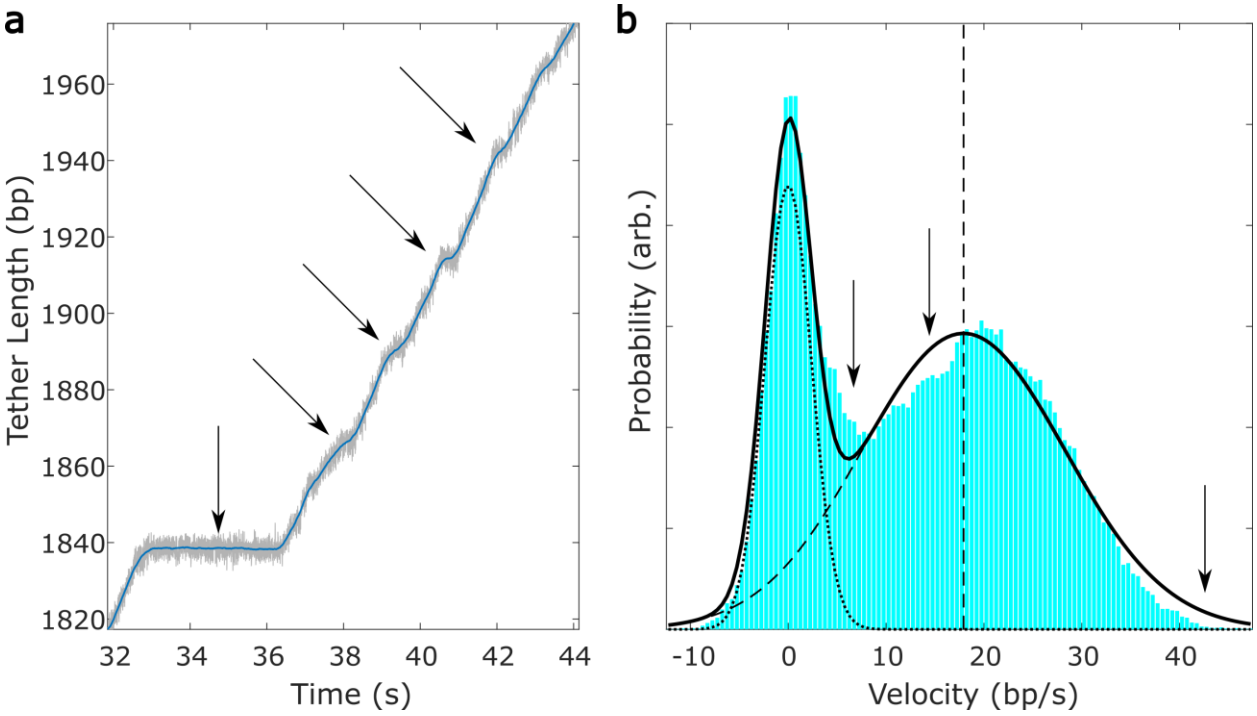


Figure 7.3 Velocity filtering of RNAP data
a) RNAP translocation data is shown at 1kHz (grey) and filtered to 3Hz (colored), the amount used for velocity smoothing. A vertical arrow shows a long pause, while the diagonal arrows point to quick pauses. b) Velocity distribution of RNAP data is shown, fit to the sum of two Gaussians (black curve), one with mean zero to represent the paused population (dotted curve), and one with positive mean (vertical dashed line) to represent the pause-free velocity (dashed curve) (see Chapter 4). Arrows show areas of poor fitting, the leftmost comes from short pauses that are quicker than the filter width, which show up as regions of small positive velocity instead of zero velocity (e.g. the short pauses in subpanel a). The misidentified pauses shift the fitting of the Gaussian with a positive mean, resulting in poor fitting at the other two identified regions.

The identities of these very short pauses are so-called *elemental* pauses, near-ubiquitous pauses that act as gateways into longer-lived pausing[46]. The sources of the

pausing are common sequence motifs, for example, the $G_{-10}Y_{-1}G_{+1}$ pause (denoted by the nucleotide identity at positions relative to the catalytic +1 site; Y = C or T), which has a 3% chance to show up in any given random sequence. Another method to quantify speed is a *crossing time analysis*, where the time it takes for the motor to cross a set distance is measured. Crossing time analysis methods have been used to quantify motor velocities using single-molecule methods with lower resolution, like magnetic tweezers[47]. In that study, the times it takes the motor to cross 10bp are measured, and the resulting distribution is fit to a Gamma distribution with shape factor 10, since the crossing time should be the sum of ten single-exponentials with the same mean. If the motor pauses, the time is dominated by the one paused step and that population shows up as a slower single-exponential tail. In optical tweezers, however, we have the resolution to detect single-bp steps under certain conditions[45], resulting in single-nucleotide dwell time distributions for the motor. One strategy is to use low [NTP] and high force, which prolongs the dwell time and reduce tether noise, respectively[45]. Another is to use a *molecular ruler*, a region of repeated sequence containing a pause, which can be used to improve spatial accuracy by detection of these repeated pauses[48].

We can apply the crossing time analysis method using a step size of 1bp on traces in RNAP in any condition, but is this method accurate in any conditions, without the tricks of low [NTP] or a molecular ruler? It should be noted that the length correction by molecular ruler is less than 3%[48] and the method differs from the study mentioned earlier in that the step size is assumed, instead of having to be detected by the algorithm with no outside information. So, perhaps this method can work. To measure the time it takes for the motor to traverse 1bp, we use the HMM dwellfinder (see Appendix 4) with a step size of 0.34nm (1bp). In plain words, the algorithm looks to find the best fitting step locations for a staircase with a 1bp step size.

Applying the crossing time analysis method to real data shows a dwell time distribution that fits well to sum of exponentials (Figure 8.5). Maximum likelihood estimation is used for the fitting, and the Akaike information criterion is used to determine the number of populations. For this dataset, two exponentials are fit to the distribution. The fastest rate corresponds to k_n (~17bp/s) and characterizes 94% of the steps taken, the second-fastest rate we associate with elemental pausing (~2bp/s) and happens about 6% of the time (Figure 7.5b). Slower pauses are observed, but not in enough number for the algorithm to fit a third population, however by eye it seems like these pauses may belong to a same third population of pause. This method allows the separation and characterization of the nucleotide addition cycle and the various pauses of the enzyme of any translocation data.

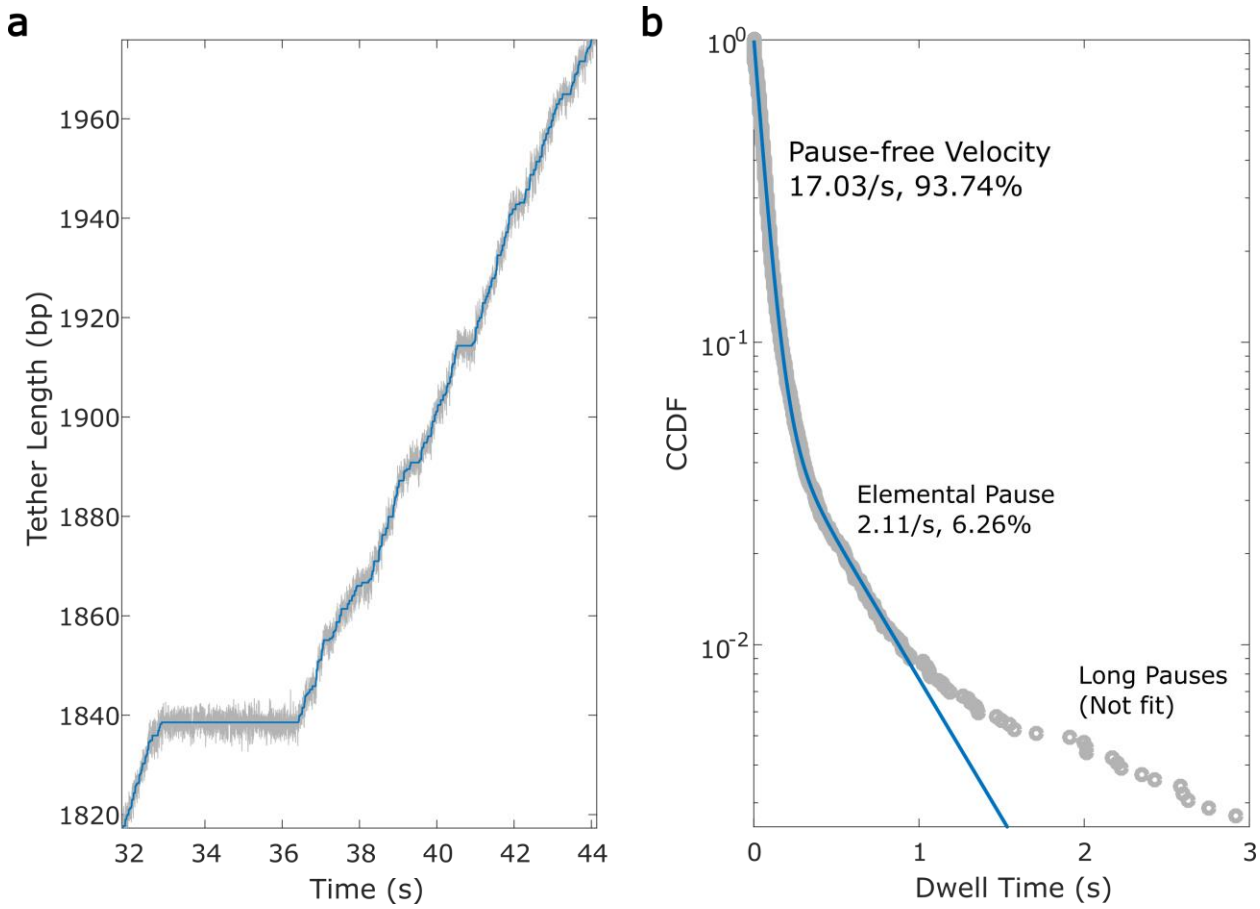


Figure 7.4 Crossing Time Analysis of RNAP data
a) An RNAP transcription trace is shown at 1kHz (grey) with the 1bp fit staircase (blue). b) The dwell time distribution is plotted as complementary CDF (grey circles) fit to a sum of two exponentials (blue line) on a semilog axis (where single-exponentials look like straight lines). The two fit exponentials are labeled with their rate and population, and the unfit tail of data is shown.

References

1 Nebenführ, A. & Dixit, R. Kinesins and Myosins: Molecular Motors that Coordinate Cellular Functions in Plants. *Annual Review of Plant Biology* **69**, 329-361, doi:10.1146/annurev-arplant-042817-040024 (2018).

2 Forgac, M. Vacuolar ATPases: rotary proton pumps in physiology and pathophysiology. *Nature Reviews Molecular Cell Biology* **8**, 917-929, doi:10.1038/nrm2272 (2007).

3 Nakamura, S. & Minamino, T. Flagella-Driven Motility of Bacteria. *Biomolecules* **9**, doi:10.3390/biom9070279 (2019).

4 Wu, W.-J., Yang, W. & Tsai, M.-D. How DNA polymerases catalyse replication and repair with contrasting fidelity. *Nature Reviews Chemistry* **1**, 0068, doi:10.1038/s41570-017-0068 (2017).

5 Gelles, J. & Landick, R. RNA Polymerase as a Molecular Motor. *Cell* **93**, 13-16, doi:10.1016/S0092-8674(00)81140-X (1998).
6 Crozat, E., Rousseau, P., Fournes, F. & Cornet, F. The FtsK family of DNA translocases finds the ends of circles. *J Mol Microbiol Biotechnol* **24**, 396-408, doi:10.1159/000369213 (2014).
7 Brown, A. I. & Sivak, D. A. Theory of Nonequilibrium Free Energy Transduction by Molecular Machines. *Chemical Reviews* **120**, 434-459, doi:10.1021/acs.chemrev.9b00254 (2020).
8 Kinosita, K., Jr., Yasuda, R., Noji, H. & Adachi, K. A rotary molecular motor that can work at near 100% efficiency. *Philos Trans R Soc Lond B Biol Sci* **355**, 473-489, doi:10.1098/rstb.2000.0589 (2000).
9 Banavar, J. R., Cieplak, M., Hoang, T. X. & Maritan, A. First-principles design of nanomachines. *Proceedings of the National Academy of Sciences* **106**, 6900-6903, doi:doi:10.1073/pnas.0901429106 (2009).
10 Courbet, A. *et al.* Computational design of mechanically coupled axle-rotor protein assemblies. *Science* **376**, 383-390, doi:doi:10.1126/science.abm1183 (2022).
11 Tafoya, S., Large, S. J., Liu, S., Bustamante, C. & Sivak, D. A. Using a system's equilibrium behavior to reduce its energy dissipation in nonequilibrium processes. *Proceedings of the National Academy of Sciences* **116**, 5920-5924, doi:doi:10.1073/pnas.1817778116 (2019).
12 Desai, V. P. *et al.* Co-temporal Force and Fluorescence Measurements Reveal a Ribosomal Gear Shift Mechanism of Translation Regulation by Structured mRNAs. *Mol Cell* **75**, 1007-1019.e1005, doi:10.1016/j.molcel.2019.07.024 (2019).
13 Gao, Y. *et al.* Structures and operating principles of the replisome. *Science* **363**, eaav7003, doi:doi:10.1126/science.aav7003 (2019).
14 Baker, T. A. & Sauer, R. T. ClpXP, an ATP-powered unfolding and protein-degradation machine. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1823**, 15-28, doi:10.1016/j.bbamcr.2011.06.007 (2012).
15 Bard, J. A. M. *et al.* Structure and Function of the 26S Proteasome. *Annual Review of Biochemistry* **87**, 697-724, doi:10.1146/annurev-biochem-062917-011931 (2018).
16 Sun, S., Rao, V. B. & Rossmann, M. G. Genome packaging in viruses. *Curr Opin Struct Biol* **20**, 114-120, doi:10.1016/j.sbi.2009.12.006 (2010).
17 Snider, J., Thibault, G. & Houry, W. A. The AAA+ superfamily of functionally diverse proteins. *Genome biology* **9**, 216-216, doi:10.1186/gb-2008-9-4-216 (2008).
18 Khan, Y. A., White, K. I. & Brunger, A. T. The AAA+ superfamily: a review of the structural and mechanistic principles of these molecular machines. *Crit Rev Biochem Mol Biol* **57**, 156-187, doi:10.1080/10409238.2021.1979460 (2022).
19 Chemla, Y. R. *et al.* Mechanism of force generation of a viral DNA packaging motor. *Cell* **122**, 683-692, doi:10.1016/j.cell.2005.06.024 (2005).
20 Moffitt, J. R. *et al.* Intersubunit coordination in a homomeric ring ATPase. *Nature* **457**, 446-450, doi:10.1038/nature07637 (2009).

21 Chistol, G. *et al.* High degree of coordination and division of labor among subunits in a homomeric ring ATPase. *Cell* **151**, 1017-1028, doi:10.1016/j.cell.2012.10.031 (2012).
22 Liu, S. *et al.* A viral packaging motor varies its DNA rotation and step size to preserve subunit coordination as the capsid fills. *Cell* **157**, 702-713, doi:10.1016/j.cell.2014.02.034 (2014).
23 Aathavan, K. *et al.* Substrate interactions and promiscuity in a viral DNA packaging motor. *Nature* **461**, 669-673, doi:10.1038/nature08443 (2009).
24 Woodson, M. *et al.* A viral genome packaging motor transitions between cyclic and helical symmetry to translocate dsDNA. *Science Advances* **7**, eabc1955, doi:10.1126/sciadv.abc1955 (2021).
25 Gao, Y. *et al.* Structures and operating principles of the replisome. *Science* **363**, doi:10.1126/science.aav7003 (2019).
26 de la Peña, A. H., Goodall, E. A., Gates, S. N., Lander, G. C. & Martin, A. Substrate-engaged 26S proteasome structures reveal mechanisms for ATP-hydrolysis–driven translocation. *Science* **362**, eaav0725, doi:10.1126/science.aav0725 (2018).
27 Dong, Y. *et al.* Cryo-EM structures and dynamics of substrate-engaged human 26S proteasome. *Nature* **565**, 49-55, doi:10.1038/s41586-018-0736-4 (2019).
28 Nagy, G. N. *et al.* Structural Characterization of Arginine Fingers: Identification of an Arginine Finger for the Pyrophosphatase dUTPases. *Journal of the American Chemical Society* **138**, 15035-15045, doi:10.1021/jacs.6b09012 (2016).
29 Tafoya, S. *et al.* Molecular switch-like regulation enables global subunit coordination in a viral ring ATPase. *Proceedings of the National Academy of Sciences* **115**, 7961-7966, doi:doi:10.1073/pnas.1802736115 (2018).
30 Harvey, S. C. The scrunchworm hypothesis: transitions between A-DNA and B-DNA provide the driving force for genome packaging in double-stranded DNA bacteriophages. *J Struct Biol* **189**, 1-8, doi:10.1016/j.jsb.2014.11.012 (2015).
31 Zhao, W., Jardine, P. J. & Grimes, S. An RNA Domain Imparts Specificity and Selectivity to a Viral DNA Packaging Motor. *J Virol* **89**, 12457-12466, doi:10.1128/jvi.01895-15 (2015).
32 Bullard, D. R. & Bowater, R. P. Direct comparison of nick-joining activity of the nucleic acid ligases from bacteriophage T4. *Biochem J* **398**, 135-144, doi:10.1042/bj20060313 (2006).
33 Gorry, P. A. General least-squares smoothing and differentiation by the convolution (Savitzky-Golay) method. *Analytical Chemistry* **62**, 570-573, doi:10.1021/ac00205a007 (1990).
34 Kalafut, B. & Visscher, K. An objective, model-independent method for detection of non-uniform steps in noisy signals. *Computer Physics Communications* **179**, 716-723, doi:10.1016/j.cpc.2008.06.008 (2008).
35 Castillo, J. P., B. Tong, A., Tafoya, S., Jardine, P. J. & Bustamante, C. A DNA packaging motor inchworms along one strand allowing it to adapt to alternative double-helical structures. *Nature Communications* **12**, 3439, doi:10.1038/s41467-021-23725-5 (2021).
36 Conn, G. L., Brown, T. & Leonard, G. A. The crystal structure of the RNA/DNA hybrid r(GAAGAGAAGC). d(GCTTCTCTTC) shows significant differences to that

found in solution. *Nucleic Acids Res* **27**, 555-561, doi:10.1093/nar/27.2.555 (1999).
37 Hantz, E. *et al.* Solution conformation of an RNA--DNA hybrid duplex containing a pyrimidine RNA strand and a purine DNA strand. *Int J Biol Macromol* **28**, 273-284, doi:10.1016/s0141-8130(01)00123-4 (2001).
38 Anosova, I. *et al.* The structural diversity of artificial genetic polymers. *Nucleic Acids Research* **44**, 1007-1021, doi:10.1093/nar/gkv1472 (2015).
39 Neidle, S. in *Principles of Nucleic Acid Structure* (ed Stephen Neidle) 38-80 (Academic Press, 2008).
40 Moffitt, J. R., Chemla, Y. R., Izhaky, D. & Bustamante, C. Differential detection of dual traps improves the spatial resolution of optical tweezers. *Proceedings of the National Academy of Sciences* **103**, 9006-9011, doi:doi:10.1073/pnas.0603342103 (2006).
41 Maillard, R. A. *et al.* ClpX(P) generates mechanical force to unfold and translocate its protein substrates. *Cell* **145**, 459-469, doi:10.1016/j.cell.2011.04.010 (2011).
42 Whitley, K. D., Comstock, M. J. & Chemla, Y. R. High-Resolution "Fleezers": Dual-Trap Optical Tweezers Combined with Single-Molecule Fluorescence Detection. *Methods Mol Biol* **1486**, 183-256, doi:10.1007/978-1-4939-6421-5_8 (2017).
43 Chen, F. X., Smith, E. R. & Shilatifard, A. Born to run: control of transcription elongation by RNA polymerase II. *Nature Reviews Molecular Cell Biology* **19**, 464-478, doi:10.1038/s41580-018-0010-5 (2018).
44 Mayer, A., Landry, H. M. & Churchman, L. S. Pause & go: from the discovery of RNA polymerase pausing to its functional implications. *Curr Opin Cell Biol* **46**, 72-80, doi:10.1016/j.ceb.2017.03.002 (2017).
45 Righini, M. *et al.* Full molecular trajectories of RNA polymerase at single base-pair resolution. *Proc Natl Acad Sci U S A* **115**, 1286-1291, doi:10.1073/pnas.1719906115 (2018).
46 Saba, J. *et al.* The elemental mechanism of transcriptional pausing. *eLife* **8**, e40981, doi:10.7554/eLife.40981 (2019).
47 Dulin, D., Berghuis, B. A., Depken, M. & Dekker, N. H. Untangling reaction pathways through modern approaches to high-throughput single-molecule force-spectroscopy experiments. *Curr Opin Struct Biol* **34**, 116-122, doi:10.1016/j.sbi.2015.08.007 (2015).
48 Gabizon, R., Lee, A., Vahedian-Movahed, H., Ebright, R. H. & Bustamante, C. J. Pause sequences facilitate entry into long-lived paused states by reducing RNA polymerase transcription rates. *Nature Communications* **9**, 2930, doi:10.1038/s41467-018-05344-9 (2018).
49 Bustamante, C., Chemla, Y. R. & Moffitt, J. R. High-Resolution Dual-Trap Optical Tweezers with Differential Detection: Instrument Design. *Cold Spring Harbor Protocols* **2009**, pdb.ip73, doi:10.1101/pdb.ip73 (2009).
50 Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**, 257-286, doi:10.1109/5.18626 (1989).

Appendix 1: φ29 protocols

The φ29 DNA packaging system consists of the viral capsid, the pRNA, the ATPase gp16, and the DNA. The materials are received from Paul Jardine's lab, but I will record as much as I know about their preparation in Table A1.1.

Component	Details	Storage Concentration	Working Concentration
Capsid 900-	Capsid purified from phage infection. Contains (mostly) 174b pRNA and some head spikes.	8g/L ("8x")	1g/L ("1x")
RNA-free capsid	Capsid formed by expression in E.coli or by RNAse treatment of 900-	8g/L ("8x")	1g/L ("1x")
174b pRNA	Full-length pRNA, including "Domain II" which can affect initiation[31]	--	--
120b pRNA	Domain II-less pRNA. For proper motor assembly, only the first 117nt of the pRNA is required.	~1ug/uL	67ng/uL
gp16	The DNA packaging motor	0.8g/L ("8x")	0.1g/L
DNA	The genomic DNA of φ29. The ends are attached to gp3, required for packaging initiation in bulk.	400ng/uL	400ng/uL

Table A1.1: φ29 Packaging Materials Information

When the four components are added together, the system self-assembles and packaging can initiate upon the addition of ATP. The pRNA ring forms on the capsid, and the gp16s attach to the pRNA ring. The motor recognizes gp3, a protein attached to the ends of the φ29 genome that is required for initiation in the test tube. The protocol for a bulk packaging experiment is shown in Table A1.2. The packaging is measured by the degree of DNA protection: the added DNAse will degrade any unpackaged genome, but any DNA packaged into the capsid is kept intact. The strength of the 19.3kb band compared to the input shows the degree of packaging. As written, the capsid-to-DNA ratio is 2:1, meaning only half of the complexes need to be active to achieve full

packaging of the DNA. This ratio can be reduced to make a more strenuous packaging assay.

Bulk Packaging	Vol. (uL)	Details
Water	10.5	Total volume of 20uL after addition of ATP
10x TMS	1	Final concentration is 0.5x TMS = 25mM Tris-HCl pH 7.8, 50mM NaCl, 5mM $MgCl_2$
gp3-DNA 400ng/uL	2.5	1ug total
1x 900-	2	
1x gp16	2	
Mix, Wait 5m at room temperature		
5mM ATP	2	Replace with 2uL water for Negative control
Mix, Wait 15m at room temperature		
1% DNAse	2	Replace with 2uL water for Positive control
Mix very gently, wait 10m @ RT		
proteinase K+EDTA	2	Equal parts pK + 500mM EDTA
Incubate 30m at 65°C		
Run on a gel (0.8% agarose), packaged DNA will show up as a 19.3kb band		

Table A1.2: Bulk packaging protocol
Protocols are read as steps top-to-bottom, as add steps (for components) or if noted

To adapt the bulk packaging protocol for single-molecule experiments, the motor is stalled mid-packaging in the test tube so it can be resumed in the optical tweezers. Stalling of the motor is done via addition of (ATPγS), a slowly hydrolyzable analog of ATP whose effects on the motor has been thoroughly studied[21]. Hence, we call this a stalled complex (SC), and a protocol is shown in Table A1.3.

SC complex	Vol. (uL)	Details
Water	4.5	Total volume of 20uL after addition of ATP
10x TMS	1	
RNAseOUT	0.5	
300ng/uL DNA-gp3	2	
1x 900-	4	
Mix		
1x gp16	4	
Mix, Wait 5m at room temperature		
5mM ATP	2	
Wait 30s at room temperature		
5mM ATPγS	2	

Store on ice for up to 12h

Table A1.3: Stalled Complex Initiation protocol

To tether the packaging complex in the optical tweezer, the two ends of the complex (capsid, free DNA end) need to be grabbed. To grab the DNA end, we will introduce a biotin there and use it to attach to streptavidin. To introduce the biotin, first the genome is cut with BstEII, which results in a 12.3 and 7kb fragment. The generated overhangs are filled with biotinylated NTPs by Klenow DNA polymerase fragment (exo-), creating a DNA with gp3 on one end and biotins on the other. Care must be taken to preserve the terminal gp3, or else the motor will not package the DNA, so the enzymes are heat inactivated and the excess NTPs are dialyzed away instead of a harsher method such as silica spin column purification. To grab the capsid, an antibody to gp8, the major capsid protein of φ29, is used.

Optical tweezers experiments are performed by grabbing two ~1um polystyrene microspheres ("beads"), each deposited with different material such that, when they are rubbed together with the optical tweezers, a tether is formed between the beads. To grab the biotinylated end of the genome, streptavidin-coated beads are used (SA), and to grab the capsid, antibody-coated beads are used (A8). The stalled complex is deposited on the SA, and the antibody-capsid connection is made in the tweezer. A protocol for preparing a stalled complex for tweezing is shown in Table A1.6, and a diagram of the attachment strategy is shown in Figure A1.1a, and a diagram of the microfluidics chamber the experiments are performed in is in Figure A1.1b.

Bead 1	Vol (uL)
0.5x TMS	10
5mM ATP	2
5mM ATPγS	2
RNAseOUT	0.5
ApaL1	0.5
SA†	4
SC	1
Wait 20m at room temperature	
0.5x TMS	1000
100mM ATP	1
100mM ATPγS	1

Bead 2	
0.5x TMS	1000
A8†	2

Middle	Vol (uL)
0.5x TMS	1000
100mM ATP	5
Dextrose 0.5g/mL	10
GODCAT*	1

Table A1.6: Stalled Complex Initiation Tweezing protocol
***GODCAT = 100mg/mL glucose oxidase and 20mg/mL catalase, forms an oxygen scavenging system along with the 5mg/mL dextrose. †Beads (SA, A8) are 0.05% w/v, passivated by vortexing for 10m with 2mg/mL BSA.**

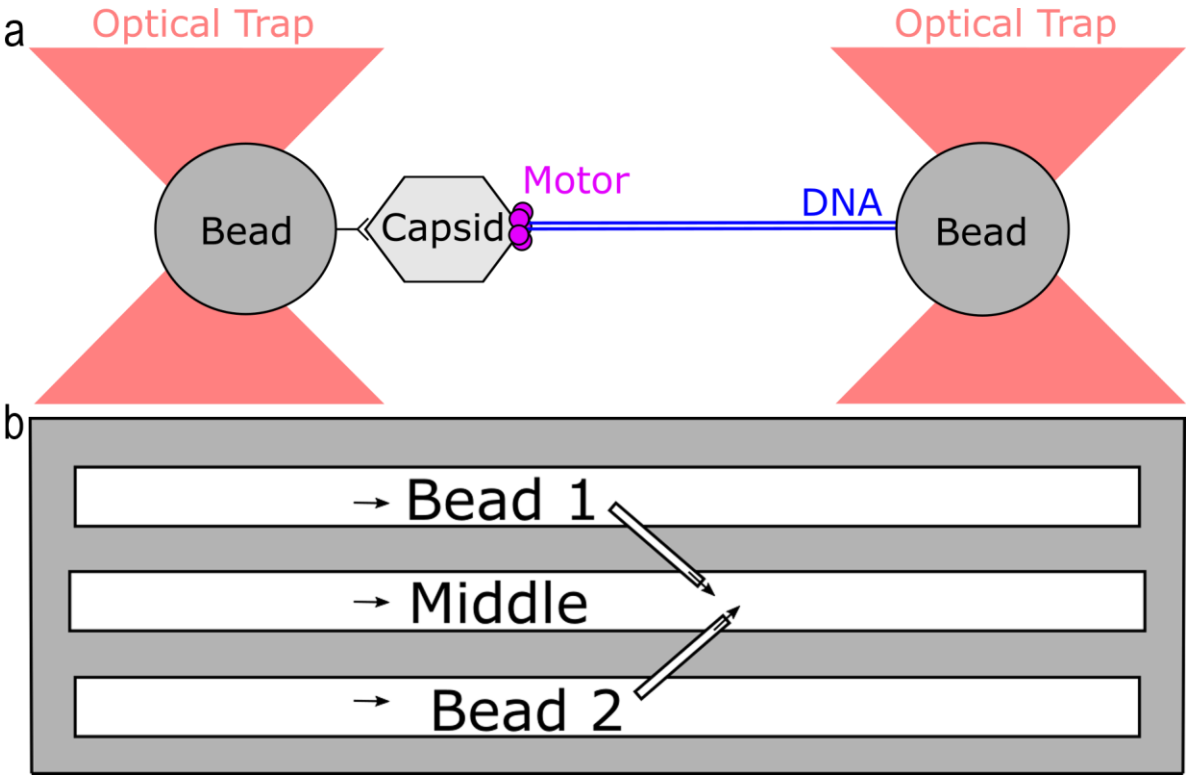


Figure A1.1 Stalled Complex Initiation Tether Geometry
a) The tethering geometry of the stalled complex initiation method is shown. The left bead is antibody-coated, the right bead is covered in Capsid-Motor-DNA complexes. When the beads are rubbed together, the antibody attaches to the capsid and a tether is formed. b) The microfluidics chamber is shown. The two beads are dispersed in separate solutions and flown through the top and bottom channel. The traps are formed in the middle channel, where the two beads can be separately trapped as they flow into the middle chamber through the two dispenser tubes.

Packaging experiments can also be initiated via the *in situ* method, where the motor complex and the substrate are brought together in the tweezer, and the forced vicinity of the two components is enough for the motor to start packaging it (assumedly by diffusion of the free DNA end into the pore of the motor). Comparing to Figure A1.1a, the left bead is covered in motor-capsid complexes and the right bead is covered in DNA; rubbing the beads together initiates the motor on the DNA. The recipe for the complex for use in in situ initiation is shown in Table A1.7, and the protocol for bringing that complex to the tweezer is in Table A1.8.

ISI Complex	Vol. (uL)
67ng/uL 120b pRNA	5
8x RNA-free capsid	2
Incubate 10m at room temperature	
10X TMS	1

RNAseOUT	0.5
8x gp16	2
Mix, Wait 5m	
5mM ATPγS	1
Wait 2 weeks at 4°C	

Table A1.7 *In Situ* Initiation Complex Protocol
The minimum incubation time for the ISI complex has varied from as little as two days to as much as two weeks.

Bead 1	Vol. (uL)
0.5x TMS	5
A8	1.5
5mM ATPγS	1
RNAseOUT	0.5
Mix	
ISI	1
Incubate 15m at room temperature	
0.5x TMS	1000
100mM ATPγS	0.5

Bead 2	Vol. (uL)
0.5x TMS	5
SA	3
Mix	
1ng/uL DNA*	2uL
Incubate 15m at room temperature	
0.5x TMS	1000

Middle	Vol. (uL)
0.5x TMS	1000
100mM ATP	5
Dextrose	10
GODCAT	1

Table A1.8 *In Situ* Initiation Tweezing Protocol
***The substrate to be packaged should be a DNA with a biotin on one end and nothing on the other. If the substrate has a terminal gp3, for example, it can be removed by silica column purification.**

Appendix 2: Documentation and optimizations to the optical tweezers control interface

The single-molecule experiments are performed on in a dual-trap optical tweezers[49] that we colloquially call *HiRes*. A 1064nm laser is split into two beams and focused into a microfluidics chamber, where the focused beams form two optical traps with which two polystyrene beads can be trapped. One beam is steerable via a piezoelectric mirror. Past the chamber, the beams are split again and measured on two position-sensitive detectors (PSDs), which read any displacements of the beam caused by the position of the beads with respect to the center of the traps. The PSD output, the displacement of the bead from the center of the trap, and the force applied on the bead are all proportional to each other.

The instrument is controlled by a LabVIEW interface. I inherited the main structure of the program, but made ease-of-use and quality-of-life changes to improve experimental

throughput. In short, the user interacts with the Main *VI* (virtual instrument, the filetype of LabVIEW code) in order to do manual manipulation of the chamber and traps in order to capture two beads and form a tether (See Figure A1.1b), after which the program breaks out into specialized VIs for the experimental protocol. The major advantage of this structure is code modularization, as each protocol VI can be made without adding bloat to the Main VI. The user should have easy access to any information they might need to check while tweezing, a photo of the tweezing computer area is in Figure A2.1, and a screenshot of the desktop in Figure A2.2.

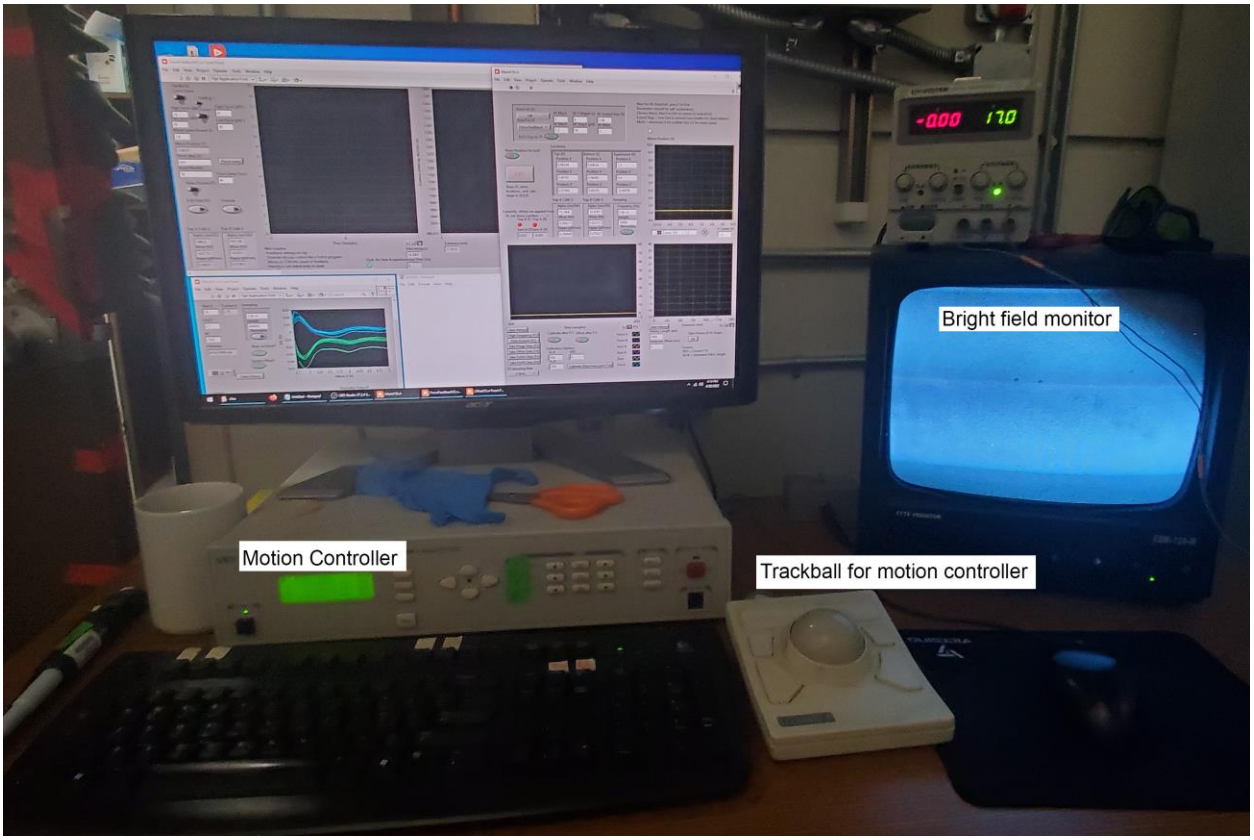


Figure A2.1 Computer (and other hardware) setup for optical tweezers control In addition to the computer monitor, there is a monitor that shows the bright field image (to see inside the chamber), the motion controller (used as a live readout for absolute chamber position) and a trackball which is used to move the chamber using the motion controller.

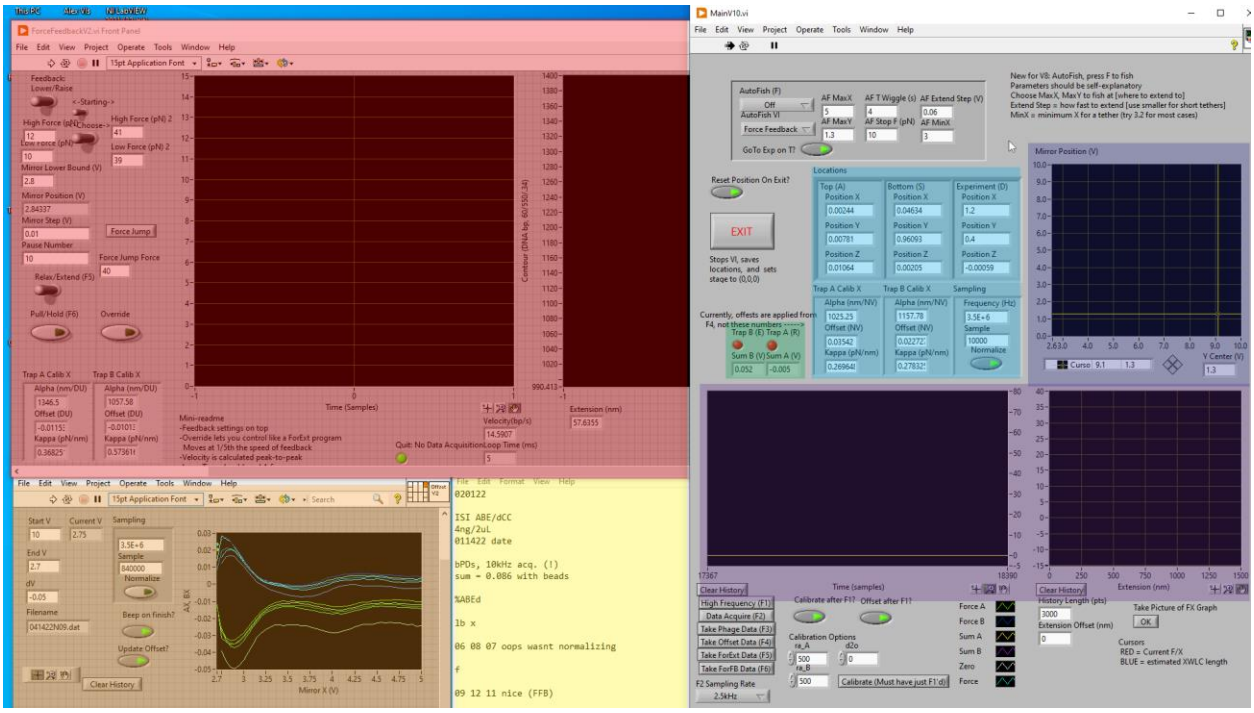
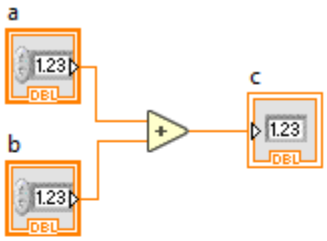


Figure A2.2 Computer desktop during tweezing
An example setup of the tweezing computer's desktop is shown, with the Main VI open on the right and various other windows open. Colored rectangular regions are drawn to denote areas of interest.
On the Main VI, the green shaded area shows the status of the two traps (whether they are on or not). The cyan-shaded region shows chamber locations (such as specific regions to get beads or perform experiments) and the current calibration (conversion between PSD output and force). The blue-shaded region shows the position of the moveable trap. The violet-shaded region is the most important area, since it shows live force-time and force-extension readouts. Of the other open windows, the red one is the protocol VI, which is what is ran once a tether is found. The orange-shaded VI handles the offset, which measures the background as a function of trap position for later subtraction. The yellow-shaded windows a text document to record what data files are which.

The following discussions will include descriptions of the VIs that drive the instrument. But first, a small introduction to the LabVIEW language is in order. LabVIEW is a visual programming language, where instead of writing lines of text, wires (which carry values) and nodes (which store or act on values) are arranged in a *block diagram*. User input and results display is handled through the front panel, the forward-facing interface to the program, which show the values of all of the variables in the program. Transitioning to coding in LabVIEW from a background knowledge of conventional text-based coding can be difficult, as some basic concepts in text coding look very different in LabVIEW. In contrast to text code, which is read top-to-bottom, LabVIEW block diagrams are (usually) arranged in a left-to-right fashion. Some concepts in text-based coding are shown as their LabVIEW counterparts in Figure 2.2.

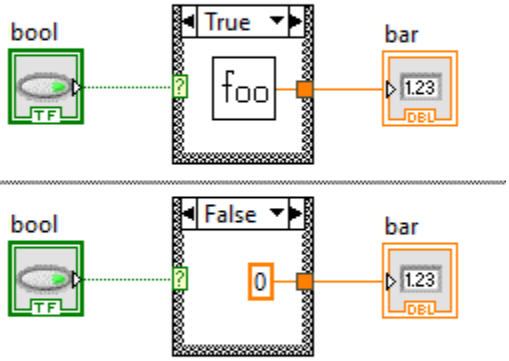
a

```
c=a+b
```



b

```
if(bool)
   bar = foo
else
   bar = 0
```



c

```
for i in 0:9
   bar(i) = foo(i)
```

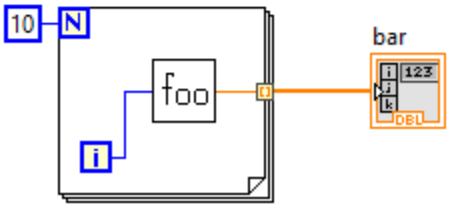


Figure A2.3: Some basic concepts in programming expressed in pseudocode and in LabVIEW code
On the left are some basic code snippets, and on the right is the corresponding code in LabVIEW. a) To add two numbers a and b, they need to be wired to the two input terminals (left side) of the "add" function (triangle with +), and the resulting sum will be output to the terminal on the right. To store it in the variable c, a wire is drawn to connect the sum to c. b) An if statement is represented by a box with multiple pages of code (both of which are shown), one page is run if the query (bool, wired into the green ? symbol) is true (top diagram) and the other if false (bottom diagram). Other VIs can be embedded in block diagrams; in this case foo a VI. c) A for loop is represented by a box, with the number of iterations signaled by the number wired to the N in the upper-left. Here foo is a function that takes in the loop iteration i as an argument and outputs a number. The thickening of the wire coming out of the for loop represents the outputs of foo being composed into an array (to be stored in bar).

The Main VI's front panel should include whatever one needs to see while performing the experiment at a glance, in one window without scrolling. An image of the Main VI is in Figure A2.2. It has a force-time graph, force-extension graph (useful for checking for tethering, and if the right tether was formed), trap position, trap shutter state (for turning traps on and off), calibration values (trap conversion and spring constant), and chamber

positions (top dispenser, bottom dispenser, experiment), and has buttons to initiate collecting data or setting parameters such as bead size. When actively tweezing, most of the time I am looking at the force-extension graph, to check if a tether forms and at what length.

The main runtime loop of the Main VI is shown in Figure 2.4. Improvements made include incorporating a force-extension graph, a more useful graph than force-time since it shows the stretching properties of the tether, which allows differentiating proper tethers from double tethers and other unwanted tethers. Also, an automation of the *fishing* process, the act of forming a tether by moving the traps together, rubbing the beads together, and separating the traps while looking for an increase in force (the sign that a tether has formed), removes some of the busy work of performing experiments. If fishing needs to be done manually, trap movement can be done by moving the mouse, too.

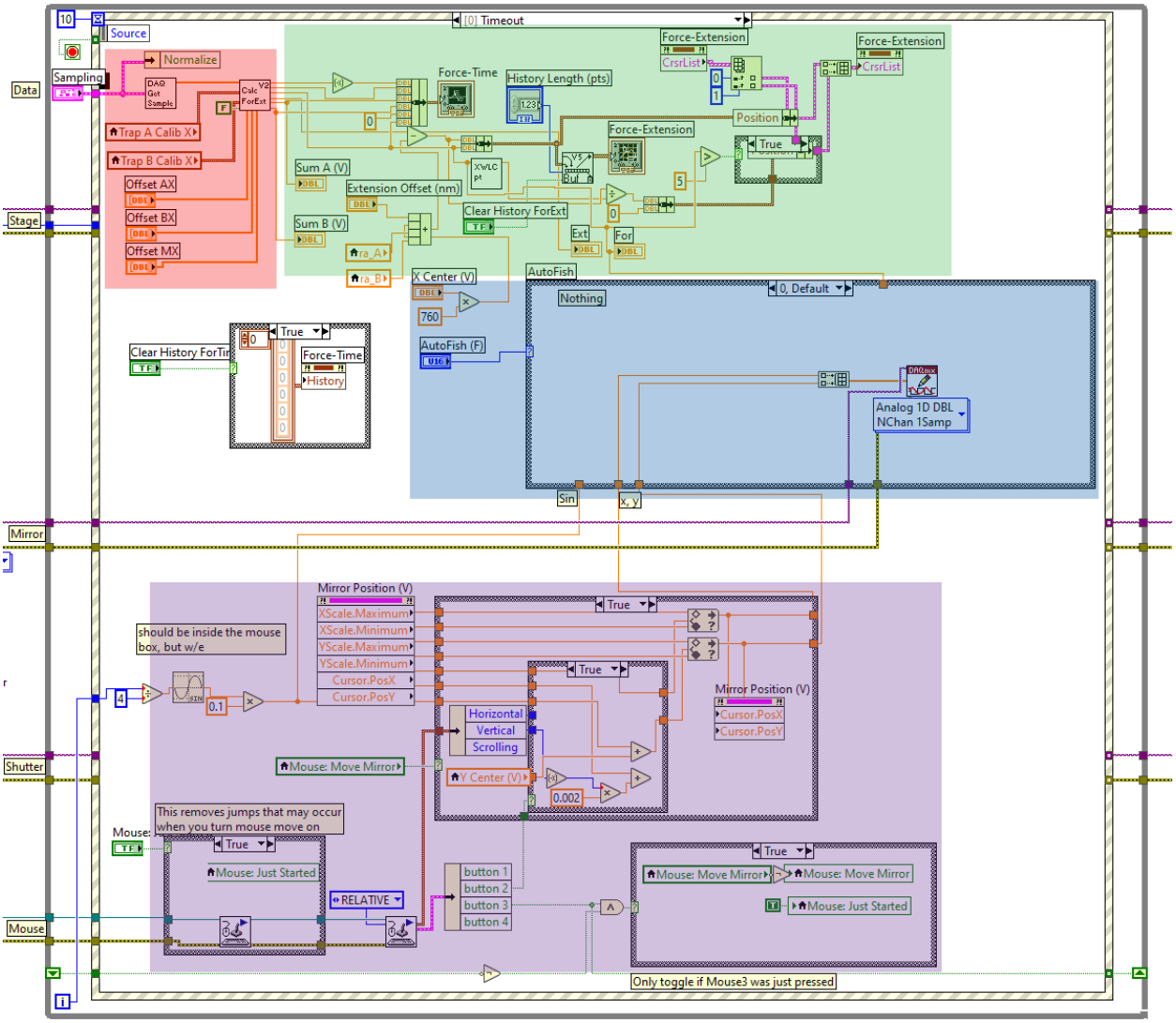


Figure A2.4. The Main VI block diagram
Pictured is the main loop of the Main VI, with colored overlays added to highlight

regions. The red-shaded code queries the PSD and calculates the current force and extension. This data is sent to the green section, where it gets plotted as force-time and force-extension graphs. The blue-shaded section automates the *fishing* process, the act of forming a tether by rubbing the beads together. The purple-shaded section lets the mouse move the trap if Mouse3 is clicked.

To further streamline the use of the instrument, I implemented keyboard shortcuts to common functions of the VI. The user's left hand rests on the normal typing home row, while the user's right uses the trackball in order to move the chamber. Being able to issue commands without having to look at the screen (as you would if the commands were buttons to click) is important since the user's focus is usually on the bright field monitor (see Figure A2.1). Toggling the traps on and off, communicating with the motion controller to go to the saved locations, initiating fishing, and taking data are all accessible via keys situated near the left hand home row. Making the tweezer easier to use and moving the rote tasks to automation can relieve mental drain during tweezing and decrease the time it takes to grab beads, increasing throughput for what is an unavoidably draining process. The user is only in charge of catching beads, which could potentially be automated via image detection, as is done in the Lumicks C-trap instruments; doing so would remove all user interaction from performing an optical tweezers experiment.

Once a tether is formed, a separate VI is run to handle the experimental protocol, so the movement of the traps can be done in a specific manner. For phage experiments, however, most data is taken using a *semipassive* protocol, where the traps are held stepwise constant, only moving to keep the increasing force within a specified range, usually between 7 and 12pN for "low force" experiments and between 30 and 35pN for "high force" experiments. An example of this protocol is shown in Figure A2.5a, and the LabVIEW code in Figure A2.5b. The other major protocol for motors is *force feedback*, where the traps are moved to keep the force constant. The advantage of using semipassive is that it removes noise that could be introduced due to the movement of the traps. However, the advantage of force feedback is that the constant force means that the movement of the motor can be inferred directly from the movement of the traps, as opposed to in semipassive where changes in extension also comes from changes in the stretching of the polymer. In semipassive mode, the polymer physics is removed by converting to contour length space. The DNA stretching response is approximated by a worm-like chain, and those model parameters (persistence length, stretch modulus) are used to convert the *extension* (the distance from one bead to the other) to *contour* (the length of the tether if it was completely extended). To obtain the worm-like chain parameters of the tether, either set values can be used (I use 50nm and 900pN for the persistence length and stretch modulus, respectively), or if possible, a pulling curve can be done for each tether to obtain parameters specific to that tether, as the calculated values can vary. For φ29 experiments, the motor is never paused for enough time to take one. Fortunately, we can use the mirror as an internal control for length to correct for errors such as deviations from the assumed worm-like chain parameters[20].

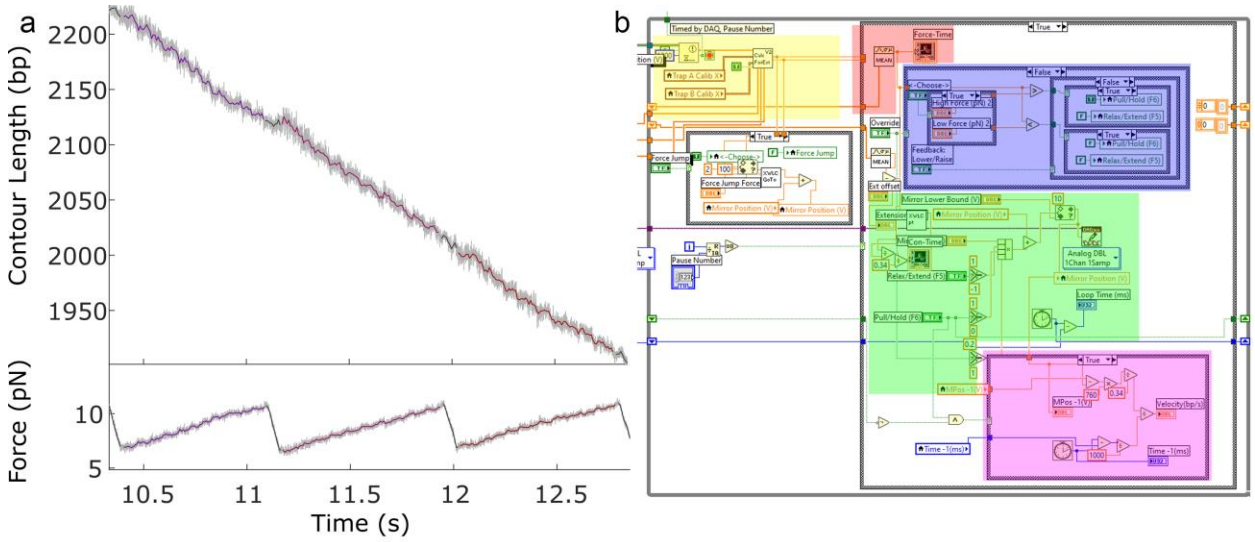


Figure A2.5 Semipassive mode
a) A trace taken in 7-12pN semipassive mode is displayed at 2.5kHz (grey) and filtered to 250Hz (colored). In the force channel, we see the increase in force as the motor packages DNA while the traps are held constant. Once the force goes above the upper force range (12pN), the traps move to relax the tension down to 7pN (black regions of traces), and the process continues. The data when the mirror moves is not considered for data analysis (black data). b) The loop of the smipassive VI. Important regions are highlighted: The current data is read and converted to force and extension in the yellow box. The data is smoothed and displayed in the red box. The blue box handles the semipassive protocol, checking the current force and adjusting it to keep it within the semipassive range. The code in the green box is where the actual movement of the traps happens. The code in the pink box calculates and displays the current velocity of the trace.

Appendix 3: Continuous elution gel electrophoresis

For generation of the chimeric DNA-hybrid-dsRNA substrate, it is imperative to be able to purify the genomic DNA fragment while keeping the terminal gp3 protein intact. So, a continuous elution gel electrophoresis apparatus was devised, to allow for direct elution of DNA fragments out of an agarose gel. The eluted product is very dilute, but can be concentrated via spin concentrators. In addition to DNA-gp3 fragments, this purifcation process was deemed more gentle than say gel extraction, so the later steps in the process that have a long product are purified via this method. The custom part of the apparatus is the gel tray, a device that allows for the creation of a chamber in the middle of the gel, from which samples can be eluted. An annotated picture of the gel tray is shown in Figure A3.1a.

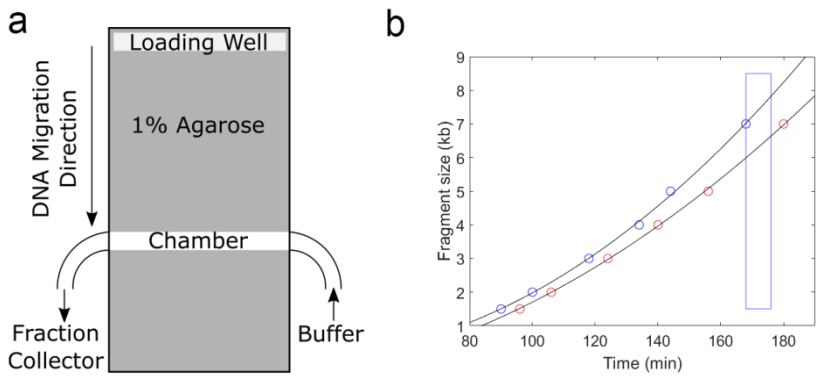


Figure A3.1 Continuous elution agarose gel electrophoresis apparatus
a) The gel tray for the apparatus has a chamber in the middle from which samples can be taken. So, as the DNA migrates from the loading well through the gel and into the chamber, it gets collected in a fraction collector. The chamber is constantly exchanging liquid while collection is occurring. See Appendix 3 for more info. b) Elution calibration timeline of DNA in a 1% agarose gel. Eluting a DNA ladder (GeneRuler 1kb plus) allows us to create a calibration curve for elution times in this device. A blue and red circle marks the start and end, respectively, of the elution of the fragments (1.5kb, 2kb, 3kb, 4kb, 5kb, 7kb) from 80 to 190 minutes. The start times and end times are fit to quadratics. The blue box shows the elution time of the xylene cyanol dye, which can be used as a visual indicator.

The 1% agarose gel is cast with a well on one and and an insert mid-way through that creates a cavity in the middle of the gel. This will be the area from which samples are taken. After the gel sets, the insert is removed and the chamber is closed off. The sample is loaded and run like normal, until just before the band of interest is anticipated to enter the chamber. The timeline in Figure 3.1c is useful for determining when to start collection. Once collection starts, a peristaltic pump adds new buffer into the chamber while it also collects from the chamber into a fraction collector. Collection is done at 1mL/min. After collection, a sample from each fraction is run on a gel to find which contain the band of interest, and these are pooled and concentrated with spin concentrators (Amicon, 10k MWCO).

Appendix 4: Hidden Markov models

A Markov chain is a model for a system transitioning between a finite number of states in a memoryless fashion. The simplest case is a two-state hopper: a system which can either be in state A or B, with a chance at every time point to change state (Figure A4.1a). A molecular system can be modeled by a Markov chain; an example in optical tweezers would be a DNA hairpin opening and closing (Figure A4.1b). However, we do not directly observe the underlying state of the system; we see some external observable (in this case, the extension) which may not directly correlate with the state. Hence, the state of the Markov chain is *hidden* from us, and we must use analysis methods to use the experimental data to infer the underlying state of the chain. In the DNA hairpin case, the two states are well-separated in space, so inferring the state from

the data is not difficult, but there are cases of lower signal-to-noise ratio (SNR) or redundant states (two different states that have similar extensions) that require more complex analysis to solve.

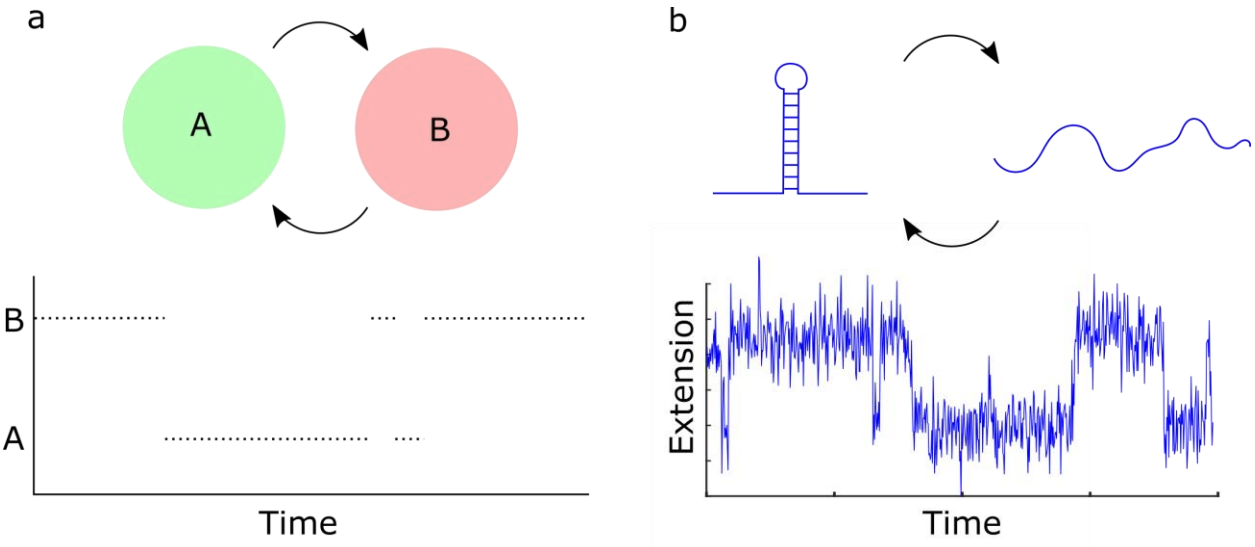


Figure A4.1: A two-state hopper
a) A Markov model with two states, A and B is shown above, with a potential chain resulting from the model shown below, depicting transitions between states A and B. b) The cooperative folding and unfolding of a DNA hairpin can be modeled as a Markovian process, with model (above) and simulated data (below). When the hairpin folds, the distance between the two ends of the DNA is shorter, leading to a decrease in the extension. The data is noisy because in this theoretical experiment, the extension is measured by optical tweezers, which has noise from the fluctuations of the beads and DNA tether. From this noisy data, the underlying folding state of the DNA hairpin can be inferred.

A Markov model is defined by the number of states *n* it has, the initial state distribution π, and the transition matrix a_{ij} between these states (the chance for a molecule in state 1 to either stay in 1 or transit to 2, 3, etc.). Given an experimental chain, we can extract the transition matrix by simply counting transitions: a_{ij} = (transitions from state *i* to state *j*)/(time spent in state *i*). To expand this to be a hidden Markov chain, we need to also add the observed trace $x(t)$, the position of these states $\vec{\mu}$ and the noise of the system *σ* (the noise can be made state-dependent without much additional complexity, but in this case we will assume the noise to not be state-dependent), so the system in state *i* will be observed as a number drawn from a normal distribution with mean μ_i and the standard deviation *σ*. A generic hidden Markov model (HMM) is depicted in Figure S2. The problem then is to take the data and derive the model that best describes that data. The solving of an HMM involves starting with a model guess, which is then refined to increase the fit between the model and data. The algorithm behind solving HMMs numerically will not be covered in detailed here, but is contained this citation[50]. The solution is based on maximum likelihood estimation, which places the chain in the most likely state given the input model. The algorithm takes advantage of the Markovian

nature of the chain, specifically the memoryless nature of it, meaning that the evolution of the chain only depends on adjacent timepoints (not larger time windows), which cuts down on the analytical complexity of the problem.

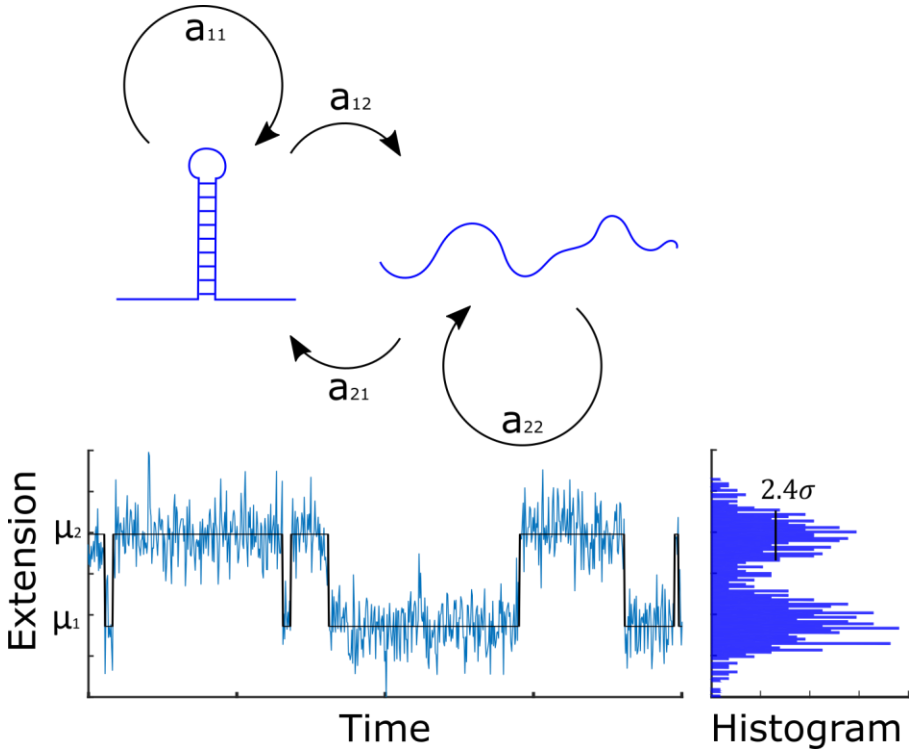


Figure S4.2: Hidden markov model for a DNA hairpin
Shown is a Markov model for a DNA hairpin labeled with the model parameters. Above: the model states and the elements of the transition matrix. Below: Sample trace and the mean and noise of the two states to characterize how the underlying state is observed in an experiment.

I will cover the method for generating a guess for an HMM, though. The initial state can be inferred from the initial point, i.e. $\vec{\pi} = NormalDistribution(\vec{\mu}, x(t_0), \sigma)$. This depends on the guess of the state vector, which can be obtained from a residence time histogram, and the noise, which can be estimated based on the data. Making a guess for the noise can be done by subtracting a slow-moving average of the data to center the values around zero, and then taking the standard deviation of the result. The transition matrix guess can be simplified to one parameter, the chance to transition, where all possible transitions have equal probability, and the chance to not transition is then one minus the number of transitions times this probability. From these guesses, fitting an experimental trace to an HMM is possible.

For a Markov chain with known state positions $\vec{\mu}$, such as in the dwellfinder (see Chapter 4), the chain has a simple transition matrix, where the only nonzero elements are the major diagonal (a_{ii}, the chance to not change state) and the first superdiagonal (a_{ii+1}, the chance to transit to the next state) since we assume that the system is simply moving from state 1 to 2 to 3, etc. in order. A graphical representation of this model is in

Figure A4.3a. If the state positions are unknown, we can still perform stepfinding by imposing a grid of states (with the distance between grid points is much smaller than the expected step size), and the transition matrix is derived from a step size distribution $\vec{s}$ (e.g., s(2) = chance to make a step of twice the grid size, s(17) is the chance to make a step of size 17 times the grid size, s(0) = chance to not step). For simplicity, let's assume the trace is monotonic (but the algorithm can be easily adapted to include negative steps), then the transition matrix is upper-diagonal, with each row being the step size distribution shifted to place the first element of $\vec{s}$ at the diagonal. Equivalently, $a_{ij} = s(j - i)$. The step size distribution after optimization is then the step size distribution of the system. Because of the gridding, there are a lot of states in this model, potentially too many to consider them all at a time. So, to speed computation time and memory requirements, only a small neighborhood of points around each position is considered, since we know that the trace won't make a super-long jump. A graphical representation of this model is in Figure A4.3b.

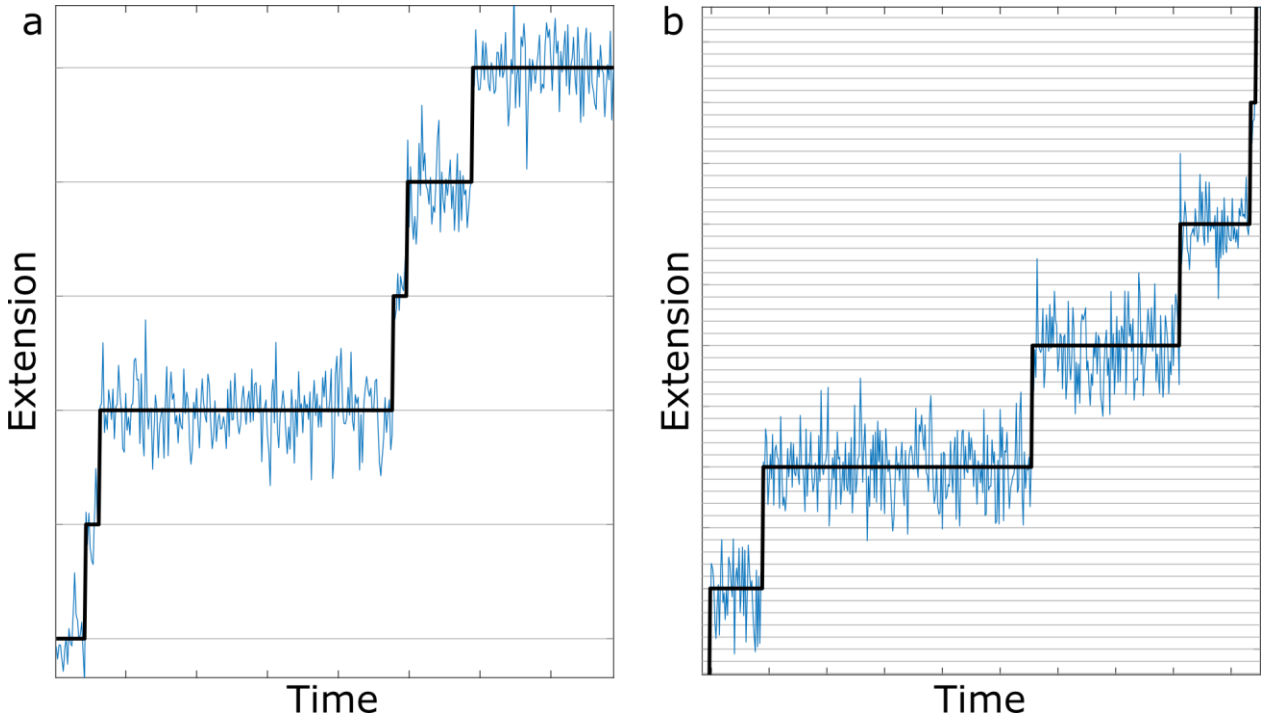


Figure A4.3: Comparison of a hidden Markov model dwellfinder and stepfinder Pictured is a simulated trace (blue), which needs to be fit to the HMM states (grey horizontal lines) by a path (thick black line). a) When the state locations are known, the chain just progresses from state to state in order (in this case, increasing in magnitude). b) When the state locations are not known, a grid of states is used instead, and the number of states traversed at each transition is variable.