POTATO: Automated pipeline for batch analysis of optical tweezers data

Stefan Buck [†1], Lukas Pekarek[†1], Neva Caliskan* [1,2]

[†] Authors contributed equally to this work.

* Corresponding author

[1] Helmholtz Institute for RNA-based Infection Research (HIRI), Würzburg, Germany

[2] Medical Faculty, Julius-Maximilians University Würzburg, Würzburg, Germany

ABSTRACT

Optical tweezers is a single-molecule technique that allows probing of intra- and intermolecular interactions that govern complex biological processes involving molecular motors, protein–nucleic acid interactions and protein/RNA folding. Recent developments in instrumentation eased and accelerated optical tweezers data acquisition, but analysis of the data remains challenging. Here, to enable high-throughput data analysis, we developed an automated python-based analysis pipeline called POTATO (Practical Optical Tweezers Analysis TOol). POTATO automatically processes the high-frequency raw data generated by force-ramp experiments and identifies (un)folding events using predefined parameters. After segmentation of the force-distance trajectories at the identified (un)folding events, sections of the curve can be fitted independently to worm-like chain and freely-jointed chain models, and the work applied on the molecule can be calculated by numerical integration. Furthermore, the tool allows plotting of constant force data and fitting of the Gaussian distance distribution over time. All these features are wrapped in a user-friendly graphical interface (https://github.com/REMI-HIRI/POTATO), which allows researchers without programming knowledge to perform sophisticated data analysis.

SIGNIFICANCE

Studying (un)folding of biopolymer structures with optical tweezers under different conditions generates very large datasets for statistical data analysis. Recent technical improvements accelerated data acquisition by coupling modern instruments with microfluidic systems, at the same time creating the need for a high-throughput, and unbiased data analysis. We developed Practical Optical Tweezers Analysis TOol (POTATO); an open-source python-based tool that can process data gathered by any OT force-ramp experiment in an automated fashion. POTATO is principally designed for data preprocessing, identification of (un)folding events and the fitting of the force-distance curves. In addition, all parameters for preprocessing, statistical analysis and fitting of the curves can be adapted to suit the dataset under analysis in an easy-to-use graphical user interface.

INTRODUCTION

Arthur Ashkin received the Nobel Prize in 2018 for his research on trapping dielectric particles with laser light in optical tweezers (OT) (1). Optical tweezers enable probing of structural dynamics of individual molecules by monitoring internal forces and short-lived intermediate states in real-time (2-5). This technique has been widely used to study structures of nucleic acids and dynamics of RNA/protein folding (6-10). In addition, OT can also be used to probe the molecular interactions between small molecules, proteins, and nucleic acids (11-13). Recently, the combination of optical tweezers with confocal microscopy enabled simultaneous measurements of force and fluorescence that provided unprecedented insights into molecular mechanisms such as timing and order of events during transcription or translation (12,14-16). Basically, in a typical OT experiment, a biopolymer, such as a protein, DNA, or RNA molecule, is tethered between two dielectric beads via labeled handles. The beads are then trapped by focused laser beams, the so-called optical traps. Following this several modes of operation are possible. In force-ramp mode the beads are precisely displaced in a monotonous manner, which applies increasing forces onto the biopolymer (**Fig. 1A**). Since trapped beads behave as if they were attached to mechanical springs, the applied force can be calculated from the measured displacement of the beads out of the trap focus according to Hooke's law (**Fig. 1B**) (17). This mode is commonly used to determine the elastic properties of the molecule and/or to determine the rupture forces at which transitions in folding and unfolding occur.

On the other hand, a constant-force operation mode allows tracking the molecule of interest in real time as it transitions between different conformational states, yielding kinetic parameters of folding-unfolding of molecules or progressive movements of molecular motors (5). Accordingly, optical tweezers experiments also allow precise calculation of the work done

on the system of interest (18,19). Previously, OT instruments were self-built by researchers and thus application required substantial physics and engineering background. Furthermore, such experiments were highly time demanding and labor intensive because a large amount of data need to be collected for a quantitative analysis. Recently, commercial instruments became available on the market. Another breakthrough was the integration of OT instruments with microfluidic systems, which accelerated both experimental setup and data acquisition (14,15). Nowadays, high-frequency data acquisition allows the generation of large data sets in a relatively short time. Subsequent data analysis, however, still requires custom written scripts to perform data preprocessing, identification of (un)folding events or different folding states, mathematical modeling, and statistical analysis. There are few algorithms developed for the analysis of single-molecule force spectroscopy data, which can perform alignment and pattern recognition functions (20-23). Such algorithms are mostly tailored for atomic force spectroscopy data analysis, thus are not directly applicable for optical tweezers data (20-25). In addition, device manufacturers would provide basic solutions for the analysis of force spectroscopy data, yet processing of the data still require bioinformatics and statistics skills, therefore remain to be a major bottleneck.

Here, we present an automated python-based pipeline for the analysis of optical tweezers force-ramp and constant-force data (POTATO). Using statistical analysis of the time-derivative of force and distance data, both unfolding as well as refolding steps are deduced automatically, and values such as (un)folding force and step length are derived. These values are then directly employed for fitting of force-distance (FD) curves. Additionally, we provide a basic constant-force analysis function. In order to allow the users to modify the analysis parameters to suit their needs, we integrated an easy-to-use graphical user interface (GUI) in POTATO. Since the pipeline allows automated processing of multiple raw data files, our tool reduces the analysis time substantially and the automated analysis ensures reproducibility and eliminates inconsistencies of manual analysis (26). Next, applicability of the tool is demonstrated on an artificially generated dataset, which covers a broad range of possible parameter combinations for force-ramp data, and also on real experimental data (27,28). Finally, we also evaluated the performance of POTATO on a published dataset independently generated using a self-built optical tweezers system (29). Our results indicate that POTATO exhibits a robust performance in identifying (un)folding events with high accuracy, precision, and recall.

MATERIALS AND METHODS

Algorithm implementation

The algorithm is written in python 3. We designed a graphical user interface and wrapped the code into a windows stand-alone executable with *pyinstaller* to open this tool to a broader audience without a bioinformatics background. The code is freely available on GitHub (https://github.com/REMI-HIRI/POTATO) and the architecture of the python files and GUI is further explained in the Supporting Material.

Artificial data generation

Artificial force spectroscopy data were generated using a custom-written python script (Supporting Material). The fully folded part of FD curves was modeled using an equation for extensible worm-like chain (WLC) models (**Eq. 4**). The partially unfolded region was modeled using a combination of WLC and freely-jointed chain (FJC) models (**Eq. 5 and 6**). For a more detailed description, see the Supporting Material.

Optical trapping system

Optical tweezers experiments were performed using a C-Trap® instrument (LUMICKS, NL). This device offers two laser traps combined with a 5-channel laminar-flow microfluidics system and a confocal microscope. Experiments were conducted as described in (27,28,30).

RESULTS AND DISCUSSION

Data preprocessing

Raw data (**Fig. 1B**) from various input file formats (.h5 or .csv files containing force and distance information) can be loaded and preprocessed before marking the (un)folding events (Supporting Material). Depending on the data collection frequency downsampling can be performed, which accelerates the analysis and saves storage space. Downsampling is especially crucial when data are collected at high frequencies. The instrument we used automatically collects data in the high-frequency mode (78,000 Hz) and the raw data need to be downsampled for ease of analysis. On the other hand, self-built systems allow collecting the data at lower frequencies. In principle, if the data frequency is sufficiently high to detect the molecule while transitioning from folded to unfolded states and vice versa, POTATO can perform the analysis. Therefore, downsampling rate should be defined by the user empirically.

We also note that data sets of very low data gathering frequency may not be suitable for direct analysis by POTATO. In that case, further preprocessing steps can be implemented (see data augmentation in Supporting Material). At the next step, a low pass Butterworth filter is employed to reduce the noise out of the signal (**Eq. 1**) (31). This filter allows efficient noise removal while keeping the actual (un)folding events intact and is therefore commonly used (**Fig. 1C**). The algorithm then trims the data at a minimum force threshold set by the user (**Table S1**). Similar to downsampling, also the noise filtering can be disabled in the GUI if the loaded data is already preprocessed.

(1) Butterworth filter:

$$G^2(\omega) = \frac{{G_0}^2}{1 + \left(\frac{\omega}{\omega_c}\right)^{2n}}$$

G is gain, ω is frequency, ω_c is cut-off frequency, and *n* is filter degree.

Force-ramp data analysis

For the identification of (un)folding events, we employed a derivative-based approach, which has been previously demonstrated to allow efficient step recognition (23). There are also other algorithms available that are based on probabilistic approaches, such as FEATHER (22). However, it must be noted that these tools are mostly developed for the analysis of atomic force microscopy (AFM) generated data (20-25). Here, we aimed to combine step recognition with downstream data fitting and determination of work, based solely on recorded force and distance values. Furthermore, we aimed to keep the pipeline intuitive and adjustable to user requirements. Although this tool was initially developed for the analysis of LUMICKS FD data in H5 format, in principle POTATO can be employed to analyze any dataset format independent of the type of optical tweezers instrument.

Statistical analysis

In force-ramp trajectories, an unfolding event is characterized by a simultaneous drop in force and a quick increase in distance as the secondary structure of the polymer undergoes a sudden transition from the folded to the unfolded state (**Fig. 1C**). Refolding events have opposite characteristics, in which the distance decreases and the force increases upon refolding. When flipped, the refolding data cannot be distinguished from the unfolding data and the processing, therefore step identification can be performed in an identical manner. Ultimately, these (un)folding events can be identified as a local maximum in the derivative of the distance and a local minimum in the derivative of the force (**Eq. 2**). When plotted, the

numerical derivative data of both distance and force show two populations of values. The first is a normal-like distribution representing the measurement noise, while outliers from the normal distribution represent the second population – the actual (un)folding events. To distinguish real (un)folding events from background noise, we calculate the moving median and the standard deviation (SD). These are then used to separate the normally distributed data from the extreme values outside a given z-score (i.e. number of standard deviations = 3 by default) (**Fig. 1D**). This should include 99.73% of the normally distributed data points. As the initially calculated SD is affected by the outliers, a second SD is calculated from the data points inside the threshold, and the data are sorted again. The cycle is repeated until the difference between initial and secondary SD is < x (with x-default = 5%). After the force- and distance derivatives are sorted, our algorithm finds the local extrema of the derivatives, representing the saddle points of the (un)folding events in the FD curve. Then, it finds the adjacent crossing points of the derivative with the moving median, representing the start or end of the corresponding unfolding events.

(2) Numerical approximation of the derivatives:

$$\frac{dF}{dt} = \frac{F(t+dt) - F(t)}{dt} \approx \lim_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{\Delta t} = \frac{F(x + step\ d) - F(x)}{step\ d}$$

$$\frac{dD}{dt} = \frac{D(t+dt) - D(t)}{dt} \approx \lim_{\Delta t \to 0} \frac{D(t+\Delta t) - D(t)}{\Delta t} = \frac{D(x + step\ d) - D(x)}{step\ d}$$

F is force, D is distance, *t* is time, *x* is position, and step *d* is a change in position.

Data fitting

Once the respective (un)folding steps are identified, this information is employed for data fitting. Data fitting is performed on the untrimmed data to model the trajectories more precisely. For the characterization of the mechanical properties of the (bio)polymer under tension, the extensible worm-like chain (WLC) model is commonly used relating the applied force and molecular extension (**Eq. 3**) (32). For that, the FD curve is split into multiple parts. The fully folded part (until the first detectable unfolding step) is fitted with an WLC (32) to calculate the persistence length (dsL_P) of the tethered molecule, while the contour length (dsL_C) is fixed. In addition, baseline and offsets in both force and distance are included in the model to compensate for the experimental variability in the FD curves.

The partially and fully unfolded parts of the FD curves are subsequently fitted using a combined model comprising WLC (describing the folded double-stranded handles) and freely

jointed chain (FJC) (**Eq. 4, 5**), or another worm-like chain (WLC) model (representing the unfolded single-stranded parts) (**Eq. 6**) **(Fig. 1E**) (32,33). To mathematically fit the models, we applied model polymer stretching functions from the free python package *pylake* (LUMICKS).

(3) Extensible worm-like chain model (WLC):

$$x_{WLC} = L_C \left[1 - \frac{1}{2} \left(\frac{k_B T}{(F - F_{off\ set}) \cdot L_P} \right)^{1/2} + \frac{(F - F_{off\ set})}{K_0} \right] - \mathrm{d_{offset}}$$

X is an extension, L_C is contour length, *F* is force, L_P is persistence length, k_B is Boltzmann constant, *T* is thermodynamic temperature, K_0 is stretch modulus, f_{offset} is force offset and d_{offset} is distance offset.

(4) Freely jointed chain (FJC):

$$x_{FJC} = L_C \left[\coth\left(\frac{2F \cdot L_P}{k_B T} \right) - \frac{k_B T}{2F \cdot L_P} \right] \left(1 + \frac{F}{K_0} \right)$$

(5) WLC + FJC:

$$x_{total} = x_{ds} + x_{ss} = x_{WLC} + x_{FJC}$$

(6) WLC + WLC:

$$x_{total} = x_{ds} + x_{ss} = x_{WLC1} + x_{WLC2}$$

Work calculations

Unfolding and refolding force-distance trajectories also yield crucial information on the thermodynamic properties of the molecule under study. Accordingly, the work applied by the optical tweezers instrument onto the system can be calculated from the area under the FD curve (AUC), here using composite Simpson's rule (**Eq. 7**). First, we determine the work applied to the whole construct, including the handles (**Fig. 2A**). The total work on the construct is the sum of the AUC of the folded model until the starting point of the step (W_{ds}) and work performed during the step transition (W_{step}), represented by the rectangular area of the step length times force average ((F_{start} + F_{end}) / 2) (**Fig. 2A**). In order to extract the amount of work applied only to the structure of interest ($W_{structure}$, **Fig. 2C**), the work applied to the handles, represented by the AUC of the combined model (W_{ss}), is subtracted from the sum of the work on the whole construct (**Eq. 8, Fig 2B-C**). It shall be noted that the work derived from these calculations equals the Gibbs free energy of the studied structure provided the system is in thermodynamic equilibrium. However, if the (un)folding trajectories do not coincide, it indicates that the molecule is out of equilibrium. In non-equilibrium scenario, Gibbs free energy can be

extracted from the work values (5,18,19,29,34-36) (**Fig. S3**). It should be noted that while POTATO performs work calculations, the estimations of free energy values have to be derived by the user separately.

(7) Numerical integration using composite Simpson's rule:

$$\int_a^b f(x)\,dx \approx \frac{h}{3}\sum_{j=1}^{n/2}\left[f\left(x_{2j-2}\right) + 4f\left(x_{2j-1}\right) + f\left(x_{2j}\right)\right]$$

where $x_j = a + jh$ *for* $j=0, 1, \ldots, n-1$ with $h=(b-a)/n$; $x_0 = a$ and $x_n = b$.

(8) Non-equilibrium work calculation:

$$W_{structure} = W_{ds} + W_{step} - W_{ss}$$

$W_{structure}$ is work needed to unfold the structure of interest. W_{ds} is numerical integration of the fully folded model, W_{ss} is numerical integration of the unfolded model, and W_{step} is numerical integration of the step region between the two models.

Constant-force data analysis

In addition to force-ramp experiments, the algorithm we provide can also analyze constant-force data (**Fig. S1** in the Supporting Material). In this way, the dynamics of the structure at a given force can be investigated. This way the equilibrium force at which the chance of the structure to be folded or unfolded are equal can be derived.

The constant-force analysis accepts the same input formats as the force-ramp batch analysis, and data preprocessing is performed similarly by downsampling and filtering of the data without trimming. First, it is necessary to display the constant-force data in order to optimize the preprocessing parameters and the plot's axis (**Fig. S1B)**. At this step, two plots are generated for visualization. In the first plot, distance is plotted against time. Here, the difference in distance corresponds to the change in the contour length of the tethered molecule. The second plot is a histogram of the distance distribution (**Fig. S1C)**. From this histogram, the number of different folding states can be deduced. Afterward, the histogram is fitted with multiple Gaussian functions. According to the position distribution histograms, the user can interactively provide initial estimates for various parameters including the number, localization, width (standard deviation, z-score), and amplitude of the fits. After the optimization, the model parameters are exported together with the percentage of each folding state as a table in csv format (comma separated values).

Artificial data sets to test the limits of detection

To test the limits of (un)folding events detectable by the POTATO pipeline, an artificial dataset was generated (Supporting Material). In this data set, some curves can show a negative step-length that would not be observed in real unfolding events. We considered these steps as non-identifiable and used them as negative controls. The phenomenon of negative steps can mainly be observed for small contour length changes (ΔL_C) between the models, combined with high force drop (ΔF) values. To test the performance of the algorithm, we defined identifiable steps as events with a drop in force and a simultaneous increase in distance (Supporting Material). To evaluate if a specific parameter combination results in an identifiable curve, **Eq. 9** with x = 0 was solved for all sets of parameters. Each time two parameters were fixed, and the third parameter was optimized.

(9) Minimal step calculation:

$$x = \mathrm{WLC}_{ss}(\mathrm{step}_{\mathrm{end}}) + \mathrm{WLC}_{ds}(\mathrm{step}_{\mathrm{end}}) - \mathrm{WLC}_{ds}(\mathrm{step}_{\mathrm{start}})$$

Where WLC corresponds to expression from **eq. 3**, "ss" refers to the model corresponding to single-strand values, while "ds" describes the double-stranded region.

A hyperplane showing the interface of theoretically identifiable and non-identifiable steps was generated from these optimized values (**Fig. 3A**). This allowed us to classify the generated dataset based on a combination of parameters: One with curves where POTATO is expected to find an unfolding step (x > 0) and the other one where POTATO should not identify the steps (x ≤ 0). After analyzing the artificial dataset (comprising 2520 curves) with different z-scores, the expected results, based on the input parameters when the data were generated, were compared to the steps identified by POTATO. For the default z-score of 3, the expected parameters were then plotted into the 3D plot and colored based on the identification by POTATO (**Fig. 3A**). For an unfolding force of 25 pN, the ΔF and ΔL_C values are shown in a 2D plot, making it easier to identify and compare single unfolding events analyzed with different z-scores. It can be seen that all identified steps at this specific unfolding force are above the theoretical threshold and that more unfolding events are identified at z-score 2.5 than at z-score 3 (**Fig. 3B**). Accordingly, the effect of the z-score on the derivative of force (**Fig. 3C)** and distance (**Fig. 3D**) can be investigated for an individual force-distance trajectory. In the representative trajectory, the local maximum in the derivatives of distance is above the z-score

threshold for both cases. In the derivative of force, the local minimum at the same position is only detected for the lower z-score **(Fig. 3C-D)**.

Next, we calculated performance measures such as accuracy, precision, sensitivity, specificity, and F1-score to validate the performance of POTATO. For a z-score of 3.2, a precision score of 0.974 indicates that most of the positive classified steps were actual steps, and even for a z-score of 2.5, the precision was still above 0.944 (**Table S2**). As expected, higher precision comes with the trade-off to miss certain positive events (recall 0.870 - 0.939), and the optimal z-score has to be chosen depending on the application. For smaller unfolding events that are difficult to detect, lower z-scores should be employed, as for distinct unfolding events the z-score can be set to higher values. This way number of false-positive events detected can be minimized. Since the present dataset was generated using artificial parameter combinations, those might not be found in actual OT measurements. Therefore, it is important to keep in mind that we were exploring the limits of the tool by using these strict parameter constraints. Performance measures would also vary depending on where a specific dataset is located in the parameter space, and which z-scores were employed.

Furthermore, we investigated how accurately POTATO estimates step parameters (F_U, ΔL_C, ΔF). For that, we compared the expected and measured values of these parameters for all curves analyzed (**Fig. 4**). We then calculated the linear regression of the true positive values to estimate possible biases of POTATO estimated F_U and ΔL_C values. Our analysis shows that in the case of F_U (**Fig. 4A**), the values determined by POTATO are in perfect agreement with the expected values (slope of the linear regression = 0.9912). For ΔL_C (**Fig. 4B**), the comparison shows a broader distribution of the measured values with an overall trend suggesting a minor overestimation (slope of the linear regression = 1.0282) of around 3%. Lastly, in the case of ΔF (**Fig. 4C**), the trend shows a slight underestimation of the measured values (slope of the linear regression = 0.8517), resulting in a bias of 12-15%. Taken together, our performance measures analysis suggests that the presented tool successfully identifies most (un)folding events correctly with only few false classifications (false positives/false negatives). Accordingly, in most of the cases, performance measures were above 0.9 (**Table S2**). Moreover, we show that POTATO can precisely estimate the parameter values describing the (un)folding events (F_U, ΔL_C, ΔF, **Fig. 4**). Overall, the performance measures and the accuracy of the estimates show that POTATO represents a reliable tool for optical tweezer data analysis.

Applicability of POTATO on real experimental data

Next, we employed POTATO to test its performance on real experimental data generated from force-distance measurements of the programmed ribosomal frameshifting (PRF) element of the Encephalomyocarditis virus (EMCV) and SARS-CoV-2 (27,28). We compared the POTATO results with manually annotated steps of a subset of our dataset. The results obtained with manual step identification and data fitting were in good agreement with the automated analysis using the pipeline (**Fig. S2A**). Harnessing POTATO in the data processing allowed us to speed up the analysis significantly compared to previous manual analysis. Furthermore, we saw that POTATO is not only suitable for curves with a single (un)folding event like in the artificial dataset, but we successfully fit force-distance curves with as many as five unfolding steps and we were able to identify even short-lived intermediate states of the unfolding process (**Fig. S2B** and **C**). In addition to the contour length change obtained by curve fitting, also the Gibb's free energy is an important variable to conclude on the nature of the (un)folded structure as the Gibb's free energy is dependent on the base pairing of the RNA. We were able to use the work calculated by the POTATO, to estimate the Gibb's free energy of the structures and thereby distinguish between different secondary structures (27). Here to demonstrate the energy calculation, we used a stem-loop mRNA of 30 nucleotides in length (**Fig. S3) (28)**. First, we use mfold (37) to predict the secondary structure and its Gibb's free energy (**Fig. S3A**). Then, we plot the unfolding as well as refolding work distributions calculated by POTATO (**Fig. S3B**). We then employed the results of POTATO analysis to estimate the Gibb's free energies by applying (i) Crooks fluctuation theorem, and (ii) Jarzynski equality with bias correction (**Fig. S3C**) as described in (18,34-36).

To evaluate the performance of POTATO on other published datasets generated using a self-built optical tweezers instrument we analyzed the SARS-CoV-2 pseudoknot RNA force-distance data by Neupane et al (29). Since the dataset provided had a lower data frequency resulting in less than 250 datapoints per FD curve, we first had to artificially augment the datapoints (see Supporting Material). Despite that, we could still successfully assign the steps and reproduce the unfolding force distribution (**Fig S2**) as well as the contour length estimate (**Table S3**). We were also able to detect the refolding steps force distribution and detected steps as low as 6 pN (**Fig. S2**). In conclusion, regardless of the system used, we demonstrate that the pipeline output matched well with manual data analysis on real-experiment datasets and that POTATO performed analysis of FD trajectories with multiple steps or even short-live intermediates in a reliable way. Therefore, POTATO represents a versatile tool for high-throughput OT data analysis for many upcoming studies.

Limitations

Processing automation comes with trade-offs (38,39). First, the statistical analysis applied in the pipeline might be prone to false-positive event discoveries due to external causes, such as vibration that might induce step-like events in the force-distance profile of gathered data. We split the force-distance data and analyze the derivatives of force and distance separately to minimize this effect. Only the events found by both approaches are considered real (un)folding events. Therefore, the robustness of the analysis is increased.

Second, the pipeline output strongly depends on parameters and threshold values that are applied throughout the analysis. The default values were set empirically to suit our needs. Therefore, it might require optimization to fit specific needs and reach an analysis output consistent with the manual data analysis. User input is required despite the user-friendly GUI environment, and an understanding of the analysis workflow is necessary to adjust the parameters rationally.

The current algorithm does not annotate the repeated folding and unfolding of a structure during force-ramp measurements and identifies this oscillation as independent steps. Nevertheless, this mainly occurs at slow loading rates and does not affect the contour length estimates. To overcome any unexpected issues with the automated analysis, POTATO also includes a tab that allows full manual analysis of the force-ram data files. This should help to eliminate bias caused by omission of certain files from the analysis during the automated analysis.

SUMMARY

Here we present a publicly available pipeline for batch analysis of optical tweezers data. Our pipeline allows OT raw or preprocessed data processing from force-ramp or equilibrium measurements (constant force/position). These are widely employed experimental approaches in the OT field, applied to nucleic acid structure probing, protein folding, RNA-protein interactions, or even to analyze events as complex as translation. Here, by wrapping our algorithm in a standalone application and designing an intuitive graphical user interface, we aim to open the data analysis to a broader audience without the need for a bioinformatics background. The user can adjust all parameters directly in the GUI without diving into the code to tailor the pipeline to their exact needs. With the parameters optimized for the here presented datasets, POTATO showed high precision and accuracy in the identification of (un)folding events. Moreover, compared to manual data analysis, the pipeline is faster and, most importantly, consistent throughout the analysis, thus yielding reproducible results.

SUPPORTING DATA

Supporting Data can be accessed in the GitHub repository (https://github.com/REMI-HIRI/POTATO)

AUTHOR CONTRIBUTIONS

NC, LP, and SB designed the pipeline. LP and SB wrote the python scripts. LP generated the artificial data. SB analyzed the artificial data. LP and SB performed the optical tweezers experiments. LP analyzed experimental data. LP and SB prepared the figures with input from NC. NC, LP, and SB wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

ACKNOWLEDGMENTS

We thank Vojtech Vrba for helpful python discussions. We thank Dr. Anke Sparmann for critically reviewing the manuscript. The work in our laboratory is supported by the Helmholtz Association and grants from the European Research Council (ERC) Grant Nr. 948636.

SUPPORTING CITATIONS

References **(40-44)** appear in the Supporting Material.

REFERENCES

1. Ashkin, A., J. M. Dziedzic, . . . S. Chu. 1986. Observation of a single-beam gradient force optical trap for dielectric particles. *Opt. Lett.* 11(5):288-290, doi: 10.1364/OL.11.000288.
2. Moffitt, J. R., Y. R. Chemla, . . . C. Bustamante. 2008. Recent advances in optical tweezers. *Annu Rev Biochem*. 77:205-228, doi: 10.1146/annurev.biochem.77.043007.090225.
3. Choudhary, D., A. Mossa, . . . C. Cecconi. 2019. Bio-Molecular Applications of Recent Developments in Optical Tweezers. *Biomolecules*. 9(1):23, doi: 10.3390/biom9010023.
4. Hashemi Shabestari, M., A. E. C. Meijering, . . . E. J. G. Peterman. 2017. Chapter Four - Recent Advances in Biological Single-Molecule Applications of Optical Tweezers and Fluorescence Microscopy. In Methods Enzymol. M. Spies, and Y. R. Chemla, editors. Academic Press, pp. 85-119.
5. Bustamante, C. J., Y. R. Chemla, . . . M. D. Wang. 2021. Optical tweezers in single-molecule biophysics. *Nature Reviews Methods Primers*. 1(1):25, doi: 10.1038/s43586-021-00021-6.
6. Chen, Y.-T., K.-C. Chang, . . . J.-D. Wen. 2017. Coordination among tertiary base pairs results in an efficient frameshift-stimulating RNA pseudoknot. *Nucleic Acids Res*. 45(10):6011-6022, doi: 10.1093/nar/gkx134.
7. Mukhortava, A., M. Pöge, . . . M. Schlierf. 2019. Structural heterogeneity of attC integron recombination sites revealed by optical tweezers. *Nucleic Acids Res*. 47(4):1861-1870, doi: 10.1093/nar/gky1258.
8. Stephenson, W., G. Wan, . . . P. T. Li. 2014. Nanomanipulation of single RNA molecules by optical tweezers. *J Vis Exp*.(90), doi: 10.3791/51542.
9. Zhong, Z., L. Yang, . . . G. Chen. 2016. Mechanical unfolding kinetics of the SRV-1 gag-pro mRNA pseudoknot: possible implications for -1 ribosomal frameshifting stimulation. *Sci Rep*. 6:39549, doi: 10.1038/srep39549.
10. Jiao, J., A. A. Rebane, . . . Y. Zhang. 2017. Single-Molecule Protein Folding Experiments Using High-Precision Optical Tweezers. *Methods Mol Biol*. 1486:357-390, doi: 10.1007/978-1-4939-6421-5_14.
11. Ritchie, D. B., J. Soong, . . . M. T. Woodside. 2014. Anti-frameshifting ligand reduces the conformational plasticity of the SARS virus pseudoknot. *J Am Chem Soc*. 136(6):2196-2199, doi: 10.1021/ja410344b.
12. Desai, V. P., F. Frank, . . . C. Bustamante. 2019. Co-temporal Force and Fluorescence Measurements Reveal a Ribosomal Gear Shift Mechanism of Translation Regulation by Structured mRNAs. *Molecular Cell*. 75(5):1007-1019.e1005, doi: https://doi.org/10.1016/j.molcel.2019.07.024.
13. Liu, T., A. Kaplan, . . . C. J. Bustamante. 2014. Direct measurement of the mechanical work during translocation by the ribosome. *Elife*. 3:e03406-e03406, doi: 10.7554/eLife.03406.
14. Eriksson, E., J. Enger, . . . D. Hanstorp. 2007. A microfluidic system in combination with optical tweezers for analyzing rapid and reversible cytological alterations in single cells upon environmental changes. *Lab on a Chip*. 7(1):71-76, doi: 10.1039/B613650H, (10.1039/B613650H).
15. Gross, P., G. Farge, . . . G. J. Wuite. 2010. Combining optical tweezers, single-molecule fluorescence microscopy, and microfluidics for studies of DNA-protein interactions. *Methods Enzymol*. 475:427-453, doi: 10.1016/s0076-6879(10)75017-5.
16. Whitley, K. D., M. J. Comstock, . . . Y. R. Chemla. 2017. 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.
17. Rocha, M. S. 2009. Optical tweezers for undergraduates: Theoretical analysis and experiments. *American Journal of Physics*. 77(8):704-712, doi: 10.1119/1.3138698.
18. McCauley, M. J., I. Rouzina, . . . M. C. Williams. 2020. Significant Differences in RNA Structure Destabilization by HIV-1 GagDp6 and NCp7 Proteins. *Viruses*. 12(5):484, doi: 10.3390/v12050484.
19. McCauley, M. J., I. Rouzina, . . . M. C. Williams. 2015. Targeted binding of nucleocapsid protein transforms the folding landscape of HIV-1 TAR RNA. *Proc Natl Acad Sci U S A*. 112(44):13555-13560, doi: 10.1073/pnas.1510100112.
20. Kuhn, M., H. Janovjak, . . . D. J. Müller. 2005. Automated alignment and pattern recognition of single-molecule force spectroscopy data. *J Microsc*. 218(Pt 2):125-132, doi: 10.1111/j.1365-2818.2005.01478.x.

21. Bosshart, P. D., P. L. T. M. Frederix, . . . A. Engel. 2012. Reference-free alignment and sorting of single-molecule force spectroscopy data. *Biophysical journal*. 102(9):2202-2211, doi: 10.1016/j.bpj.2012.03.027.
22. Heenan, P. R., and T. T. Perkins. 2018. FEATHER: Automated Analysis of Force Spectroscopy Unbinding and Unfolding Data via a Bayesian Algorithm. *Biophysical Journal*. 115(5):757-762, doi: https://doi.org/10.1016/j.bpj.2018.07.031.
23. Andreopoulos, B., and D. Labudde. 2011. Efficient unfolding pattern recognition in single molecule force spectroscopy data. *Algorithms for Molecular Biology*. 6(1):16, doi: 10.1186/1748-7188-6-16.
24. Gergely, C., B. Senger, . . . J. Hemmerlé. 2001. Semi-automatized processing of AFM force-spectroscopy data. *Ultramicroscopy*. 87(1-2):67-78, doi: 10.1016/s0304-3991(00)00063-2.
25. Roduit, C., B. Saha, . . . S. Kasas. 2012. OpenFovea: open-source AFM data processing software. *Nature Methods*. 9(8):774-775, doi: 10.1038/nmeth.2112.
26. Muhs, K. S., W. Karwowski, . . . D. Kern. 2018. Temporal variability in human performance: A systematic literature review. *International Journal of Industrial Ergonomics*. 64:31-50, doi: https://doi.org/10.1016/j.ergon.2017.10.002.
27. Hill, C. H., L. Pekarek, . . . I. Brierley. 2021. Structural and molecular basis for Cardiovirus 2A protein as a viral gene expression switch. *Nature Communications*. 12(1):7166, doi: 10.1038/s41467-021-27400-7.
28. Zimmer, M. M., A. Kibe, . . . N. Caliskan. 2021. The short isoform of the host antiviral protein ZAP acts as an inhibitor of SARS-CoV-2 programmed ribosomal frameshifting. *Nature Communications*. 12(1):7193, doi: 10.1038/s41467-021-27431-0.
29. Neupane, K., M. Zhao, . . . M. T. Woodside. 2021. Structural dynamics of single SARS-CoV-2 pseudoknot molecules reveal topologically distinct conformers. *Nature Communications*. 12(1):4749, doi: 10.1038/s41467-021-25085-6.
30. Pekarek, L., S. Buck, . . . N. Caliskan. 2022. Optical Tweezers to Study RNA-Protein Interactions in Translation Regulation. *JoVE*.(180):e62589, doi: doi:10.3791/62589.
31. Butterworth, S. 1930. On the Theory of Filter Amplifiers. *Experimental Wireless and the Wireless Engineer*. 7:536-541.
32. Odijk, T. 1995. Stiff Chains and Filaments under Tension. *Macromolecules*. 28(20):7016-7018, doi: DOI 10.1021/ma00124a044.
33. Smith, S. B., Y. Cui, . . . C. Bustamante. 1996. Overstretching B-DNA: The Elastic Response of Individual Double-Stranded and Single-Stranded DNA Molecules. *Science*. 271(5250):795, doi: 10.1126/science.271.5250.795.
34. Gore, J., F. Ritort, . . . C. Bustamante. 2003. Bias and error in estimates of equilibrium free-energy differences from nonequilibrium measurements. *Proceedings of the National Academy of Sciences*. 100(22):12564, doi: 10.1073/pnas.1635159100.
35. Liphardt, J., S. Dumont, . . . C. Bustamante. 2002. Equilibrium Information from Nonequilibrium Measurements in an Experimental Test of Jarzynski's Equality. *Science*. 296(5574):1832, doi: 10.1126/science.1071152.
36. Collin, D., F. Ritort, . . . C. Bustamante. 2005. Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies. *Nature*. 437(7056):231-234, doi: 10.1038/nature04061.
37. Zuker, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*. 31(13):3406-3415, doi: 10.1093/nar/gkg595.
38. Alberdi, E., L. Strigini, . . . P. Ayton (2009). Why Are People's Decisions Sometimes Worse with Computer Support? In B. Buth, G. Rabe, and T. Seyfarth, eds. Computer Safety, Reliability, and Security. Springer Berlin Heidelberg.
39. Cummings, M. L., F. Gao, . . . K. M. Thornburg. 2016. Boredom in the Workplace: A New Look at an Old Problem. *Hum Factors*. 58(2):279-300, doi: 10.1177/0018720815609503.
40. Harris, C. R., K. J. Millman, . . . T. E. Oliphant. 2020. Array programming with NumPy. *Nature*. 585(7825):357-362, doi: 10.1038/s41586-020-2649-2.
41. Collette, A. 2013. Python and HDF5. O'Reilly.
42. McKinney, W. 2011. pandas: a Foundational Python Library for Data Analysis and Statistics. *Python High Performance Science Computer*.
43. Virtanen, P., R. Gommers, . . . Y. Vázquez-Baeza. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*. 17(3):261-272, doi: 10.1038/s41592-019-0686-2.
44. Hunter, J. D. 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*. 9(3):90-95, doi: 10.1109/MCSE.2007.55.

FIGURE LEGENDS

FIGURE 1: Schematic of the pipeline. **(A)** Diagram illustrating the optical tweezers experiments. RNA is hybridized to ssDNA handles and immobilized on beads. These are used to exert a pulling force on the RNA with a focused laser beam. In force-ramp operation mode, the force is gradually increased until the structure in the middle is unfolded (bottom). Release of the force allows the structure to refold (top). *RAW data files* **(B)** are downsampled, the noise is filtered using a Butterworth signal filter, and the data are trimmed at a minimum force threshold to yield the *Trimmed filtered data* **(C)**. Then the time derivative is calculated numerically to yield the *Derivative data* **(D)**; histogram of the derivative value distribution (right) shows two populations - normal-like distribution represents the experimental noise, while the other population of outliers represents the (un)folding steps. The derivative data are then statistically analyzed – the standard deviation and moving median are calculated. Peaks in derivative data that exceed median (white line) ± z-score (grey region) are classified as (un)folding events. The beginning and end of each event are derived. The coordinates of the events are then used to define the region for fitting, yielding the *Fitted steps* **(E)**. Finally, the output data files are exported according to the selected settings. The FD curve shown here was simulated (see Supporting Material).

FIGURE 2: Work determination of a simple hairpin. (**A-C)** FD curve obtained during force-ramp experiment of a short stem-loop of 30 nucleotides. Inlets: the optical tweezers construct stretched between the beads with grey regions indicating to what parts of the construct the calculated work relates. (**A**) Marked region (grey) corresponding to the work necessary for stretching of the whole construct including the structure of interest. (**B**) Marked region (grey) corresponding to the work necessary for stretching of the handles and the unfolded single-stranded RNA. (**C**) Marked region (grey) corresponding to the work necessary for stretching of the RNA structure of interest. See the subsequent analysis in Supplementary Figure S3.

FIGURE 3: Testing the limits of POTATO. For each combination of the parameters unfolding force (F_U), force drop (ΔF), and contour length change (L_C), two parameters were fixed, and the third one was optimized so that the **eq. 9** (Supporting Material) evaluates to zero. **(A)** A hyperplane was generated from the optimized values that separate the resolvable space above the hyperplane (parameter combinations that result in identifiable steps) from the unresolvable space below the hyperplane (parameter combinations that result in unidentifiable steps). Each analyzed curve is plotted in blue if its step was identified by POTATO or in grey if it was not recognized. **(B)** Slices of the 3D plot at F_U = 25 pN were analyzed with different z-scores. The black line corresponds to the theoretical limit of resolvable/unresolvable parameter combinations. The black dots represent curves with identified steps, whereas the grey dots represent curves where POTATO could not identify the step. The derivatives of force **(C)** and distance **(D)** of the curve that is marked with a red arrow in **(B)** are displayed at different z-scores.

FIGURE 4: Evaluation of the performance of POTATO. The parameters used for the generation of the dataset compared to the parameters identified by POTATO are plotted against each other. All three parameters used for the data generation are evaluated with a z-score of 3. The values of the true positive steps (black) and the values of the false-positive steps (grey) are visualized for **(A)** the unfolding Force (F_U), **(B)** the contour length change (ΔL_C), and **(C)** the force drop (ΔF). A dashed line represents the theoretical perfect correlation between measured and expected value.