Published in final edited form as:

Ultramicroscopy. 2014 January ; 136: . doi:10.1016/j.ultramic.2013.07.020.

Improving single molecule force spectroscopy through automated real-time data collection and quantification of experimental conditions

Zackary N. Scholl¹ and Piotr E. Marszalek²

¹Program in Computational Biology and Bioinformatics, Duke University, Durham, North Carolina, 27708, USA

²Department of Mechanical Engineering and Materials Science, Center for Biologically Inspired Materials and Material Systems, and Duke University, Durham, North Carolina, 27708, USA

Abstract

The benefits of single molecule force spectroscopy (SMFS) clearly outweigh the challenges which include small sample sizes, tedious data collection and introduction of human bias during the subjective data selection. These difficulties can be partially eliminated through automation of the experimental data collection process for atomic force microscopy (AFM). Automation can be accomplished using an algorithm that triages usable force-extension recordings quickly with positive and negative selection. We implemented an algorithm based on the windowed fast Fourier transform of force-extension traces that identifies peaks using force-extension regimes to correctly identify usable recordings from proteins composed of repeated domains. This algorithm excels as a real-time diagnostic because it involves <30 ms computational time, has high sensitivity and specificity, and efficiently detects weak unfolding events. We used the statistics provided by the automated procedure to clearly demonstrate the properties of molecular adhesion and how these properties change with differences in the cantilever tip and protein functional groups and protein age.

Keywords

Atomic force; single molecules; single molecule experiments; single molecule microscopy; proteins; protein adsorption of surfaces

1. Introduction

The advent of atomic force microscopy (AFM) [1, 2] has been a major proponent for single molecule force spectroscopy (SMFS) which enables mechanical manipulation and precise characterization of biopolymers through single molecule stretching experiments (reviewed in [3]) or through measuring receptor-ligand rupturing strengths (reviewed in [4]). This characterization has produced novel insights about mechanics and protein folding [5–19], conformational transitions in polymers [20–26], protein-ligand interactions [4, 27–37], and bond mechanochemistry [38–40]. Though this field is promising, SMFS is hindered by the

*Correspondence should be addressed to ZNS (zns@duke.edu) or PEM (pemar@duke.edu).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

slow and tedious data collection which is necessary in AFM experiments to ensure accurate and reproducible data.

Here, we focus on SMFS stretching experiments which often use synthetic proteins composed of the protein of interest flanked by a repeating “self-reporting” protein. An example of a self-reporting protein is the I27 domain of the human muscle protein titin which has a force peak at ~200 pN and a characteristic worm-like chain (WLC) contour length increment of about 28 nm due the elongation of 91 amino acids (~32 nm) minus the initial folded length of the protein (~4 nm) [5, 7]. The presence of this positive control helps to distinguish recordings that have the protein of interest from the nonspecific recordings. The positive control also unequivocally determines single molecule recordings, allowing detection of a protein of interest in a single recording.

Even with a positive control, the recording of a single molecule with the protein of interest is a relatively rare event; often at least 50,000 recordings are needed before enough unfolding events are selected to create a converged histogram of the characteristic contour length increment and unfolding forces. Normally the researcher saves only a few hundred recordings during data acquisition in order to save space and analysis time. Dynamic force spectroscopy [31, 33, 41] experiments must be conducted at several pulling speeds which increases the required amount of data by a factor of four. Each recording is done at a speed of 5–5000 nm/s so a given recording may take 0.2–30 seconds; this means that the full mechanical characterization of a protein can require 12–1400 hours of operating the AFM (not including gene engineering, protein purification, and sample preparation).

Automation has become more recurrent in science fields, enabling better reproducibility and allowing higher throughput of scientific data [42, 43]. A notable example is the robot developed by Oliver et al. that generates hypotheses and performs functional genomic assays which are simple - but laborious – experiments [44]. Many drawbacks of AFM can be partially removed using an accurate automated procedure for capturing and selecting data. Currently, most AFM operation is done by trained researchers except for a few recent advances that allow for the automation of calibration of cantilever spring constants [45] and saving most recordings using a force threshold [46]. This last method works well for many proteins. However, proteins with weak unfolding events (<20 pN) would require setting the threshold near the baseline which would capture nearly all empty or nonspecific recordings. Since only 1% of data is usable it is not advisable to save the other 99%, which would require additional storage and backup resources and make future reviewing cumbersome.

How, then, can one save only the usable force-extension recordings from AFM experiments in real-time? Accurate analysis of force curves can be performed through algorithms that use methods like force thresholds [46], fuzzy logic [47], peak detection [48], transformation into contour length space [49], pattern recognition [50], convolution functions [51] or correlation functions [52] or combinations of these methods [53, 54]. Our method introduces an alternative approach that focuses on real-time data selection by improving the positive selection of usable recordings and the negative selection of unwanted nonspecific recordings. A given AFM experiment may yield only 0.5–1% of usable recordings, 20–30% nonspecific recordings, while the remainder are often blank recordings. A successful real-time algorithm should be able to accept only the small fraction of the recordings that are usable (positive selection) and efficiently reject all other recordings (negative selection) to save space and time.

With these characteristics in mind, we developed an algorithm for accurate automation of real-time data collection for AFM stretching experiments of polyproteins. This automation helped to increase the throughput of our laboratory because it allows parallel

experimentation and increased duration of experiments. Also, this procedure instantly provides new information about the statistics of usable and nonspecific recordings which can be exploited to describe surface and tip chemistry effects on molecules for AFM. We use this information to test the hypothesis that Au-coated cantilever tips or substrates and proteins with N-terminal added cysteine can capture recordings with greater efficiency and efficacy. This work serves as an important step toward creating a fully-automatic AFM system that operates similarly to other spectroscopy machines.

2. Materials and methods

2.1. Procedure for automated real-time data collection

Force extension (FE) traces are mostly either empty (background) or contain nonspecific adhesion to multiple molecules or molecules not of interest. We developed a robust method for determining peaks to distinguish force-extension recordings containing the protein of interest from the nonspecific recordings. This method uses two steps, 2.1.1.) using a windowed fast Fourier transform across the recordings, and then 2.1.2.) using preset force-extension regimes to detect usable peaks and assess whether the recording is usable (Figure 1). This method is efficient at capturing force peaks because of the discontinuity that occurs when the force drops during an unfolding event. The force drop creates a discontinuity in the signal causing the Fourier sums to overshoot the original function (known as the “Gibbs phenomenon”). The overshooting of the Fourier sums at discontinuities provides a signal containing peaks at each unfolding event, and allows for accurate identification of unfolding events, even in the presence of noise.

2.1.1. Windowed fast Fourier transform—The algorithm to produce a windowed fast Fourier transform of the data is as follows (scripts and test sets of proteins are available at <http://smfs.pratt.duke.edu/downloads.html>):

1. Designate a window size that is smaller than the smallest unfolding contour length increment. For this study a window size of 5 nm was used.
2. Allocate an empty matrix to hold the FFT coefficients for each point of the windowed recordings. The number of rows should be equal to the number of windows to take the FFT and the number of columns should be equal to the half of the window length.
3. Set location to the beginning of the FE vector.
4. Draw a window at the current location of the FE vector.
5. Calculate the magnitude of the FFT of the windowed region at the current location. The windowed region need not be a power-of-two as the small window size can be efficiently calculated using a FFT algorithm that takes advantage of small prime factors.
6. Save the magnitude of the FFT coefficients to a new row in the matrix.
7. Step to next point to take the next window and go back to step 4 until the end of the FE vector is reached.
8. The WFFT is calculated from the sum of the odd coefficients, by adding the odd columns of the matrix. The corresponding sum is the vector containing the WFFT.

2.1.2. Determining force-extension peaks—The determination of the force-extension peaks is straightforward. First, all peaks are detected by comparing each element of the WFFT to its neighboring values. A local maximum is declared if it is larger than both its

neighbors. The peaks are then sorted in descending order according to their heights. Starting from the highest peak, any subsequent peak separated by a distance smaller than the minimum FE regime extension increment (Figure 1B) is ignored. Additionally, any peaks less than the minimum force specified in the FE regime are ignored. To determine the type of unfolding event for each local maximum, the user sets a 2D space for the possible forces and extensions of a given molecule (Figure 1B). The force of peaks and distances between peaks from the WFFT of the FE recording are analyzed and then each peak is assigned specific types or declared nonspecific if no type is found.

The presence of a usable recording is based using previously published heuristics. For polyI27 proteins, there must be at least three I27 peaks (not including the detachment peak) for it to be considered a usable recording. For consensus Ankyrin repeat proteins, there must be at least five Ankyrin events. For the SNase constructs there must be at least three peaks of either SNase or I27 to register as a usable recording. These more stringent heuristics maximize the positive selection of true recordings while improving the negative selection by minimizing false positives.

2.2. Controlled stage movement for automated real-time data collection

AFM instruments are generally equipped with X, Y, and Z piezoelectric stages. The X-stage and Y-stage can be used to scan the surface of the substrate. We use the X-Y piezoelectric stages to raster scan wide areas (50 μm x 50 μm) at small 500 nm intervals. The Z-stage is critical for performing force spectroscopy experiments as it is used to set the sample a small distance away from the cantilever, called the Z-set point (usually set to ~50% greater than the fully stretched length of a single molecule unless cyclic measurements are to be performed). A measurement is performed by moving the sample towards the cantilever tip and then moving back a distance prescribed by setting a Z-scan size. The force exerted on the substrate surface from the tip during this measurement can be controlled through the Z-scan size and the initial Z-set point. The trained researcher makes sure that the Z-set point and the Z-scan size can capture the entire unfolding of a protein and makes sure it does not exert too much force on the substrate.

Often, the sample substrate or the AFM head is not perfectly flat and the initial Z-set point becomes too close or too far from the sample as the cantilever moves across the surface. This can be compensated by controlling the Z-stage when the range of forces against the substrate exceeds or is less than expected. This is implemented by performing a test during the raster scan on each X and Y move. The Z-scan size is set to be constant and the Z-set point is either moved away from the substrate if the force is above a high force level or moved closer to the substrate if the force is below a low force level. For the experiments in this study we used a low force level of 300 pN and a high force level of 600 pN which corresponds to a distance of about 80 nm for our sample and cantilever's spring constant.

2.3. Cloning and protein purification

For this study we used a polyprotein composed of seven I27 domains connected with short two amino acid linkers which we refer to as I27₇. The gene for I27₇ was obtained from modifying the 8th terminal of the poly(I27) pRSETa vector, a kind gift from Jane Clarke [55], with a C-terminal 6xHis-tag. We also use a construct containing alternating I27 domains and Staphylococcal nuclease [56], (I27-SNase)₃-I27, and a construct containing consensus ankyrin repeats flanked by I27 domains (Kim *et al.* to be submitted), I27₃-NI₁₀C-I27₃, that were created previously in our laboratory. The Cys-I27₇ construct was made by mutating the Gly residue after the N-terminal 6xHis-tag into a Cys, which should leave an unburied thiol group for reaction with the Au tip or surface. The mutation was made using 5' phosphorylated primers with the desired mutation and verified by sequencing

(Supplementary Figure S1). All engineered plasmids were transformed into *Escherichia coli* C41(DE3)pLysS cells, and expression was induced using isopropyl β -D-thiogalactopyranoside and purified on a Ni-NTA column (Qiagen, Valencia, CA) and stored in elution buffer (PBS pH 7.4 with 250 mM Imidazole) at 4°C.

2.4. Assay of thiol accessibility in I27₇ and Cys-I27₇

Each I27 molecule could potentially contribute two thiol bonds from cysteines, Cys63 and Cys47. However, the crystal structure of I27 [57] shows that only Cys47 is exposed to the solvent but that the sulfur atom is partially buried in a pocket made by the loop region between stand C and strand D (Supplementary Figure S2). To test whether any of these cysteines are reactive, we used attachment to 5' thiol modified DNA handles as an assay. We followed the protocol for attaching 5' thiol modified DNA handles for the I27₇ protein and the Cys-I27₇ protein (Supporting material) and visualized using AFM imaging. [58] The number of full length DNA handles with zero, one, or more than one protein attached was determined for each set of images. Whole DNA handles were determined based on their length (0.35 nm = 1 bp) and the presence of a protein attachment was determined by a 1–3 nm bump on the end of the DNA molecule. DNA handles attached to a Cys-I27₇ protein in 10% of the handles while the other 90% of the DNA handles attached to no Cys-I27₇ proteins (n=281, Supplementary Figure S4A). For the unmodified I27₇ protein we found that 0% of the DNA handles attached to a protein (n=302) (Supplementary Figure S4B). This data indicates that only the Cys-I27₇ protein construct is able to form a single thiol bond while the unmodified I27₇ protein construct is incapable of forming thiol bonds. Therefore we can test whether the thiol bond capable Cys-I27₇ protein construct can produce more full length recordings due to the attachment to the gold-covered substrate or a gold covered AFM tip.

2.5. AFM spectroscopy

All AFM measurements were obtained using a custom-built AFM instrument [14]. Automation routines to control the AFM as described in 2.2 were implemented in Labview 7.0 (National Instruments, Austin, Texas). Cantilever spring constants were calibrated in the buffer solution using the energy equipartition theorem [29]. Force-extensions measurements were all performed at a speed of 500 nm/s, in a phosphate buffered saline solution at room temperature. Glass substrates were prepared by sonicating glass for 5 minutes, boiling in piranha solution (25% sulfuric acid, 75% hydrogen peroxide v/v), and then sonicating for another 5 minutes and drying with filtered air. In all experiments the purified protein was diluted to 9 μ g/ml, which corresponds to about 121 nM for I27₇ (a 74.5 kDa protein), and applied to a freshly cleaned glass slide or recently evaporated gold and incubated for an hour. Measurements were performed using Au-coated OBL-10 (OBL) cantilevers or Silicon nitride MLCT cantilevers (Bruker, Camarillo, CA).

The usable FE recordings were determined from the automatically collected dataset based off previously published heuristics to determine the number of unfolding events [59]. There are four main criteria: 1) nonspecific forces must be absent, 2) The end of the approach and retraction curves should overlay, 3) the detachment force is not counted as a protein force rupture, and 4) peaks must fit a family of worm-like chain (WLC) models [60] given below:

$$F(x, p, L) = \frac{k_b T}{p} \left[\frac{1}{4(1 - \frac{x}{L})^2} + \frac{x}{L} - \frac{1}{4} \right]$$

A WLC family was fit by manually fitting a WLC to the final desorption peak, and then extended to preceding peaks using a contour length spacing (using $\Delta L = 28$ nm and

persistence length of $p=0.3$ nm for I27 domains). Normally another criterion imposes a constraint to select only recordings that have at least three peaks (not including detachment force peak) from the self-reporting protein. However, for this study we relax this criterion to select recordings that have as little as one peak (not including detachment force peak), as long as they have the correct WLC contour length spacing and additionally have a force rupture within an acceptable range for I27 (~ 200 pN). Peaks were counted as an I27 unfolding force rupture if these criteria were validated. Examples of recordings are shown in Supplementary Figure 11.

3. Results

3.1. Fidelity of algorithm to select usable recordings

The fidelity of our algorithm to select usable recordings (described in 2.1) was gauged using appropriate test sets of data. The sensitivity, or true positive rate ($\#$ of true positives/ $(\#$ of true positives + $\#$ of false negatives)), was tested against a set of manually selected recordings of various protein constructs. The specificity ($\#$ of true negatives/ $(\#$ of true negatives + $\#$ of false positives)) of the negative selection was tested against a set of all nonspecific recordings from several AFM experiments. The robustness of these characteristics was tested using simulated levels of noise added to the force-extension recordings in these test sets.

3.1.1. Effect of different unfolding characteristics—To test the fidelity of the procedure we performed automated selection using several different constructs that would sample a large space in the possible force-extension characteristics of proteins. The proteins used in this study are: 1) I27 polypeptide, I27₇, 2) alternating domains of Staphylococcal nuclease and I27, (I27-SNase)₃-I27 (SNase), and 3) consensus ankyrin repeats flanked by I27 domains I27₃-NI₁₀C-I27₃ (NI10C) or I27₃-NI₆C-I27₃ (NI6C). Sets of the FE recordings of these proteins were compiled from previously published works or collected from new experiments and are available online. The FE recordings of SNase, NI10C, and NI6C contained the unfolding events of the specific protein alone or contained the unfolding events of the protein followed by unfolding events of I27 domains. The proportion of each set is shown in Table 1. The specificity of each system was determined separately using recordings that show nonspecific recordings (examples shown in Supplementary Figures S9 and S10). The results are shown in Table 1.

All the protein constructs tested here showed 100% sensitivity, with the exception of the detection of only 3 domains of the polyI27. These protein constructs demonstrate a wide range of the force-extension regimes of proteins (Figure 1B). For instance, Ank repeats are on the extreme of the force-extension regime with very small unfolding forces and very short extension increments. I27 is on the other end of the spectrum with a high unfolding force and a moderately high length increment. SNase fills the gap in between these two. All these proteins had near perfect sensitivity meaning that all usable recordings were successfully detected and that this algorithm is applicable to a vast range of force-extension regimes. Representative examples of the correctly identified recordings for these proteins are shown in Supplementary Figures S5–S8.

A real-time processing algorithm for AFM FE traces must also discern the true recordings from nonspecific recordings. The metric to evaluate this negative selection is the specificity. The specificity for each FE regime was determined based on a test set of nonspecific recordings. For all constructs the specificity is $>90\%$ which means that fewer than 10 nonspecific recordings are selected for every 100 nonspecific recordings. Examples of accurate negative selection are shown in Supplementary Figure S9. The negative selection fails for $<10\%$ of the cases, and examples of these issues are shown in Supplementary

Figure S10. In these false positive results there are complicated nonspecific FE patterns that are suggestive of the FE regime for I27. These could be eliminated by using a more stringent heuristic, such as using more domains for selection, at the risk of decreasing the sensitivity.

3.1.2. Effect of noise on the fidelity of algorithm—Noise is inherent in AFM experiments and can be alleviated using lower loading rates and decreasing recording bandwidths for unfolding, or by sophisticated techniques like Lock-in amplifiers [61]. Here we used a test set of proteins with various levels of synthetic noise (as determined by RMS) added to the original signal. An example is shown in Figure 2 and the results are tabulated in Table 2. The baseline RMS noise in all recordings is between 3 pN and 10 pN.

The FE of I27 proteins are resistant to large amounts of noise because of their high unfolding force and large extension increment. The sensitivity is unchanged upon adding up to 12 pN of noise and the detection of recordings with only three I27 peaks becomes less sensitive upon adding 24 pN. The I27 force regime is also very specific, as adding up to 24 pN had no effect on the negative selection.

The FE of NI6C and NI10C proteins are also well defined in their force regimes so that they remain perfectly sensitive for positive selection up to 24 pN. However, adding more noise increases the chance of small nonspecific peaks that resemble the Ank FE regime and the specificity decreases drastically between adding 6 pN and 24 pN of noise. The FE of SNase is also well defined, although there is a slight drop in sensitivity when adding 12 pN of noise. There is also a drastic drop in sensitivity for SNase when increasing the noise level from 12 pN to 24 pN, because the added noise is approaching the FE regime of SNase. As a result, more nonspecific events become present.

3.2. Quantifying experimental conditions using statistics from automation

Using the automation procedure we were able to obtain statistics about the frequency of force extension curves. As an example of the power of such statistics, we quantified a common experimental practice of using single thiol group at the N-terminus and Au-covered cantilever or substrate to record force-extension traces with higher efficiency (more usable unfolding events) or with higher efficacy (more full length molecules unfolded). The application of the bond that forms between a thiol group of the Cysteine residue and Au has been an important advance for biomolecular assembly [62]. It has been exploited for specific adhesion of a molecule to a substrate in protein biology [63], and biological membranes [64]. It has been shown that the specific adhesion of a molecule to a cantilever tip can be attained through functional modification of the cantilever tip [34, 65, 66]. However it is unknown whether the specific adhesion between the protein and the tip will increase the probability of picking up full length molecules.

To test experimental conditions we performed AFM measurements on six systems using the Cys-I27₇ protein construct and the I27₇ protein construct: 1) I27₇ with MLCT on glass, 2) I27₇ with OBL on glass, 3) Cys-I27₇ with MLCT on glass, 4) Cys-I27₇ with OBL on glass, 5) I27₇ with MLCT on gold, or 6) Cys-I27₇ with MLCT on gold. These systems were measured over several prepared samples, with an average of 6,630 recordings per sample for an average total of 60,749 recordings for each system (~8 experiments for each). Each system was used with old (>3 month) and new (< 2 week old) proteins.

3.3. Analysis of efficacy of pulling full-length I27₇ proteins

An AFM system with high efficacy would have a high proportion of single molecule recordings that unfold the full length molecule. Such a system will have a better chance of unfolding the protein of interest and would be very useful for researchers' data collection.

Efficacy results for the six systems are shown in Figure 3. The distributions for Cys-I27₇ OBL-Glass, I27₇ MLCT-Gold and Cys-I27₇ MLCT-Gold had a great proportion of molecules with more unfolding events than either I27₇ MLCT-Glass, I27₇ OBL-Glass, or Cys-I27₇ MLCT-Glass ($p < 10^{-5}$ with Bonferroni correction for all tests). The Cys-I27₇ MLCT-Gold shows higher proportion of full length molecules than I27₇ MLCT-Gold, however this observation is not statistically significant ($p = 0.14$ with Bonferroni correction).

It is possible that the Cys-I27₇ could react with itself to form an I27₇-Cys-Cys-I27₇ protein. This would result in a higher number of longer molecules when using an unreactive substrate and tip as compared to using I27₇ with the same experimental conditions. Comparing the distributions for Cys-I27₇ MLCT-Glass and I27₇ MLCT-Glass we found that the null hypothesis that the distributions come from the same underlying distribution could not be rejected ($p=0.73$). Thus the intra-molecular connections formed between Cys-I27₇ did not influence the results.

Overall, these results are consistent with our hypothesis that an unburied thiol group at the end of the molecule would increase the proportion of longer molecules picked up by an Au-covered tip or on a gold coated surface. Surprisingly, the I27₇ MLCT-Gold system was also highly efficacious. This could be due to the inert reactivity of gold. Glass may be stickier and better able to hold down the protein than gold. However, combining the Cys-I27₇ protein with the Gold substrate and MLCT cantilever seemed to have the best overall results, combining the assets of the gold substrate and the connection between the free thiol and the gold surface.

3.4. Analysis of efficiency of pulling I27₇ proteins

An AFM system with high efficiency is able to pick up more single molecule recordings, though not necessarily full length recordings. Efficiency for the six systems was determined by counting the average number of positive recordings that were obtained during a single experiment. A single experiment consisted of a freshly prepared sample on newly cleaned glass and an average of 6,630 total force-extension recordings. Since all samples were diluted to the same concentration and placed on the same substrate we expect that the density of molecules between each experiment should be equal.

We compiled the results from different tests on the six different experimental setups and did ANOVA assuming that components contributing to the variation were either protein type (I27₇ or Cys-I27₇), substrate type (gold or glass), cantilever type (OBL or MLCT), protein age (old or new), or their interactions. The only statistically significant source of variation was the protein age ($p=0.008$ with Bonferroni correction) shown in Figure 4.

4. Discussion

We show that there are feasible heuristics and software capabilities to automate the SMFS experimentation for stretching experiments of polypeptides by correctly identifying usable recordings from nonspecific recordings. This step is useful for AFM operators who spend tens to hundreds of hours manually operating AFM machines. Unfortunately, the simple method of saving all recordings is not feasible here because of the huge amounts of hard drive space necessary for data storage and preservation. Previous methods for AFM force-extension analysis prioritize accurately identifying each individual peak but this is too computationally intensive to be used in a real-time setting. Thus, we developed an algorithm that performs selection of the usable recordings from the unusable nonspecific recordings in a timely manner that is efficient for real-time selection, which is applicable to weak and strong unfolding events in proteins composed of repeated domains.

The fidelity of our algorithm was analyzed and determined to have 100% sensitivity in constructs with Ankyrin repeat proteins, SNase proteins, and I27 (except for detecting recordings of only three I27 domains in which sensitivity is >90%). These three types of proteins cover a large area of the possible force-extension regime for unfolding of proteins. We believe that this shows that this algorithm is pliable and can readily be applied to a great variety of unfolding events. This sensitivity also remains very high when simulated noise is added. Thus, this algorithm is very robust and should perform similarly across different types of AFM setups, buffer conditions, or cantilevers.

In addition to providing high sensitivity of selecting true recordings, this algorithm also has high specificity. A high specificity is necessary for a real-time algorithm to triage the <1% of the correct recordings from the 20–30% of the recordings that have nonspecific events. In all cases tested here, this algorithm shows greater than 90% specificity. This specificity could be greater using more stringent heuristics, but at the risk of losing sensitivity. The sensitivity remains robust at modest levels of added noise but too much noise can greatly hinder the ability to select against nonspecific recordings.

This algorithm is limited in its ability to correctly identify all peaks in a single FE recording. Since it uses the heuristics of a minimally usable recording to define a positive event (i.e. at least three I27 peaks for an I27 poly protein) it does not rely on the ability to correctly identify each and every peak in a recording. Post-processing methods could be used to more accurately identify all relevant peaks, to model the unfolding characteristics of the protein, and to perform alignments. But since these methods require more time, they may be best left to a postprocessing algorithm.

There are many benefits of using an automated procedure for real-time AFM data collection. Automation provides a means to save hundreds of hours of measurement that would otherwise be necessary for a human researcher to supply. Also, the experiments can be performed in parallel if multiple atomic force microscope setups exist in the laboratory. The automated system can also provide information about the statistics of the frequency of usable and nonspecific recordings. Here we exploited these statistics to test a hypothesis about the reactive effects of thiol-group modified molecules with Au-covered cantilevers or Au-coated substrates. We find that a free cysteine residue on the terminal end of a protein can increase the proportion of full length molecule recordings. This same technique can be applied to investigate efficiency and efficacy of other chemical functionalization methods such as the HaloTag7 system [67], the Ni-NTA-6xHis system [68], or the Streptavidin-biotin system [69].

We also envision application of this algorithm in conjunction with protein-based force probes [70] to automate force spectroscopy measurements of interacting protein systems such as ligand-receptors pairs. For such systems, detecting the rupture of single-pair events and rejecting recordings of nonspecific events is quite challenging because force extension data typically consists of a single force peak. Addition of our force probes to the interacting molecules (at the DNA level) will introduce a positive control and this will enable automation of these SMFS measurements as well.

5. Conclusion

Automation is a simple procedure and provides an objective framework to the subjective process of data collection and selection from force-extension recordings. We hope researchers are encouraged to automate their systems to expedite the production of high quality single molecule force spectroscopy data because there is still a great deal of molecular information that can be probed with this powerful instrumentation. We hope to further develop these automation techniques to adapt it to other forms of single molecule

measurements like AFM-FRET [71] or AFM-TIRF [72]. This development is the first step on the way to creating a fully automatic AFM-based force spectrometer that would operate similarly to traditional spectrometers such as CD, UV, VIS, IR or NMR in that it would generate and save useful force spectra with minimal User supervision.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The AFM Labview data acquisition and analysis program that was modified for automatic data collection in this work was originally developed by Dr. J.M. Fernandez. Data for SNase and NI10C was obtained by Dr. Minkyu Kim, data for NI6C was obtained by Dr. Whasil Lee. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1106401 to ZNS and by the National Science Foundation grant MCB-1052208 and the National Institutes of Health grant GM079563 to PEM. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. Binnig G, Quate CF, Gerber C. Atomic force microscope. *Phys Rev Lett*. 1986; 56:930–933. [PubMed: 10033323]
2. Rugar D, Hansma P. Atomic Force Microscopy. *Physics Today*. 1990; 43:23–30.
3. Marszalek PE, Dufrene YF. Stretching single polysaccharides and proteins using atomic force microscopy. *Chem Soc Rev*. 2012; 41:3523–3534. [PubMed: 22331199]
4. Hinterdorfer P, Dufrene YF. Detection and localization of single molecular recognition events using atomic force microscopy. *Nat Meth*. 2006; 3:347–355.
5. Rief M, Gautel M, Oesterhelt F, Fernandez JM, Gaub HE. Reversible Unfolding of Individual Titin Immunoglobulin Domains by AFM. *Science*. 1997; 276:1109–1112. [PubMed: 9148804]
6. Oberhauser AF, Marszalek PE, Erickson HP, Fernandez JM. The molecular elasticity of the extracellular matrix protein tenascin. *Nature*. 1998; 393:181–185. [PubMed: 9603523]
7. Carrion-Vazquez M, Oberhauser AF, Fowler SB, Marszalek PE, Broedel SE, Clarke J, Fernandez JM. Mechanical and chemical unfolding of a single protein: a comparison. *Proc Natl Acad Sci U S A*. 1999; 96:3694–3699. [PubMed: 10097099]
8. Viani MB, Schaffer TE, Chand A, Rief M, Gaub HE, Hansma PK. Small cantilevers for force spectroscopy of single molecules. *J Appl Phys*. 1999; 86:2258–2262.
9. Oberhauser AF, Hansma PK, Carrion-Vazquez M, Fernandez JM. Stepwise unfolding of titin under force-clamp atomic force microscopy. *Proc Natl Acad Sci U S A*. 2001; 98:468–472. [PubMed: 11149943]
10. Williams PM, Fowler SB, Best RB, Toca-Herrera JL, Scott KA, Steward A, Clarke J. Hidden complexity in the mechanical properties of titin. *Nature*. 2003; 422:446–449. [PubMed: 12660787]
11. Law R, Carl P, Harper S, Dalhaimer P, Speicher DW, Discher DE. Cooperativity in Forced Unfolding of Tandem Spectrin Repeats. *Biophys J*. 2003; 84:533–544. [PubMed: 12524305]
12. Fernandez JM, Li H. Force-Clamp Spectroscopy Monitors the Folding Trajectory of a Single Protein. *Science*. 2004; 303:1674–1678. [PubMed: 15017000]
13. Gutsmann T, Fantner GE, Kindt JH, Venturoni M, Danielsen S, Hansma PK. Force spectroscopy of collagen fibers to investigate their mechanical properties and structural organization. *Biophys J*. 2004; 86:3186–3193. [PubMed: 15111431]
14. Lee G, Abdi K, Jiang Y, Michael P, Bennett V, Marszalek PE. Nanospring behaviour of ankyrin repeats. *Nature*. 2006; 440:246–249. [PubMed: 16415852]
15. Bornschlgl T, Rief M. Single Molecule Unzipping of Coiled Coils: Sequence Resolved Stability Profiles. *Physical Review Letters*. 2006; 96:118102. [PubMed: 16605876]
16. Dietz H, Berkemeier F, Bertz M, Rief M. Anisotropic deformation response of single protein molecules. *PNAS*. 2006; 103:12724–12728. [PubMed: 16908850]

17. Kessler M, Gottschalk KE, Janovjak H, Muller DJ, Gaub HE. Bacteriorhodopsin folds into the membrane against an external force. *Journal of Molecular Biology*. 2006; 357:644–654. [PubMed: 16434052]
18. Junker JP, Rief M. Single-molecule force spectroscopy distinguishes target binding modes of calmodulin. *Proc Natl Acad Sci U S A*. 2009; 106:14361–14366. [PubMed: 19667195]
19. Lee W, Zeng X, Zhou HX, Bennett V, Yang W, Marszalek PE. Full Reconstruction of a Vectorial Protein Folding Pathway by Atomic Force Microscopy and Molecular Dynamics Simulations. *Journal of Biological Chemistry*. 2010; 285:38167–38172. [PubMed: 20870713]
20. Kikuchi H, Yokoyama N, Kajiyama T. Direct Measurements of Stretching Force-Chain Ends Elongation Relationships of a Single Polystyrene Chain in Dilute Solution. *Chemistry Letters*. 1997; 26:1107–1108.
21. Rief M, Oesterhelt F, Heymann B, Gaub HE. Single Molecule Force Spectroscopy on Polysaccharides by Atomic Force Microscopy. *Science*. 1997; 275:1295–1297. [PubMed: 9036852]
22. Marszalek PE, Oberhauser AF, Pang YP, Fernandez JM. Polysaccharide elasticity governed by chair-boat transitions of the glucopyranose ring. *Nature*. 1998; 396:661–664. [PubMed: 9872313]
23. Rief M, Clausen-Schaumann H, Gaub HE. Sequence-dependent mechanics of single DNA molecules. *Nat Struct Biol*. 1999; 6:346–349. [PubMed: 10201403]
24. Li HB, Rief M, Oesterhelt F, Gaub HE, Zhang X, Shen JC. Single-molecule force spectroscopy on polysaccharides by AFM - nanomechanical fingerprint of alpha-(1,4)-linked polysaccharides. *Chem Phys Lett*. 1999; 305:197–201.
25. Ho D, Zimmermann JL, Dehmelt FA, Steinbach U, Erdmann M, Severin P, Falter K, Gaub HE. Force-driven separation of short double-stranded DNA. *Biophys J*. 2009; 97:3158–3167. [PubMed: 20006953]
26. Rakshit S, Sivasankar S. Cross-linking of a charged polysaccharide using polyions as electrostatic staples. *Soft Matter*. 2011; 7:2348–2351.
27. Florin E, Moy V, Gaub H. Adhesion forces between individual ligand-receptor pairs. *Science*. 1994; 264:415–417. [PubMed: 8153628]
28. Moy VT, Florin EL, Gaub HE. Intermolecular Forces and Energies between Ligands and Receptors. *Science*. 1994; 266:257–259. [PubMed: 7939660]
29. Florin EL, Rief M, Lehmann H, Ludwig M, Dornmair C, Moy VT, Gaub HE. Sensing Specific Molecular-Interactions with the Atomic-Force Microscope. *Biosensors & Bioelectronics*. 1995; 10:895–901.
30. Fritz J, Katopodis AG, Kolbinger F, Anselmetti D. Force-mediated kinetics of single P-selectin/ligand complexes observed by atomic force microscopy. *Proceedings of the National Academy of Sciences*. 1998; 95:12283–12288.
31. Evans E. Looking inside molecular bonds at biological interfaces with dynamic force spectroscopy. *Biophysical Chemistry*. 1999; 82:83–97. [PubMed: 10631793]
32. Harada Y, Kuroda M, Ishida A. Specific and Quantized Antigen–Antibody Interaction Measured by Atomic Force Microscopy. *Langmuir*. 1999; 16:708–715.
33. Merkel R, Nassoy P, Leung A, Ritchie K, Evans E. Energy landscapes of receptor-ligand bonds explored with dynamic force spectroscopy. *Nature*. 1999; 397:50–53. [PubMed: 9892352]
34. Zhang, X.; Rico, F.; Xu, AJ.; Moy, VT. Atomic Force Microscopy of Protein–Protein Interactions *Handbook of Single-Molecule Biophysics*. Hinterdorfer, P.; Oijen, A., editors. Springer US; 2009. p. 555-570.
35. Zhu R, Howorka S, Proll J, Kienberger F, Preiner J, Hesse J, Ebner A, Pastushenko VP, Gruber HJ, Hinterdorfer P. Nanomechanical recognition measurements of individual DNA molecules reveal epigenetic methylation patterns. *Nat Nano*. 2010; 5:788–791.
36. Kim M, Wang CC, Benedetti F, Marszalek PE. A nanoscale force probe for gauging intermolecular interactions. *Angew Chem Int Ed Engl*. 2012; 51:1903–1906. [PubMed: 22253141]
37. Kaur P, Qiang F, Fuhrmann A, Ros R, Kutner LO, Schneeweis LA, Navoa R, Steger K, Xie L, Yonan C, Abraham R, Grace MJ, Lindsay S. Antibody-unfolding and metastable-state binding in force spectroscopy and recognition imaging. *Biophys J*. 2011; 100:243–250. [PubMed: 21190677]

38. Grandbois M, Beyer M, Rief M, Clausen-Schaumann H, Gaub HE. How strong is a covalent bond? *Science*. 1999; 283:1727–1730. [PubMed: 10073936]
39. Szoszkiewicz R, Ainarapu SRK, Wiita AP, Perez-Jimenez R, Sanchez-Ruiz JM, Fernandez JM. Dwell time analysis of a single-molecule mechanochemical reaction. *Langmuir*. 2008; 24:1356–1364. [PubMed: 17999545]
40. Wiita AP, Perez-Jimenez R, Walther KA, Grater F, Berne BJ, Holmgren A, Sanchez-Ruiz JM, Fernandez JM. Probing the chemistry of thioredoxin catalysis with force. *Nature*. 2007; 450:124–127. [PubMed: 17972886]
41. Strunz T, Oroszlan K, Schäfer R, Güntherodt HJ. Dynamic force spectroscopy of single DNA molecules. *Proceedings of the National Academy of Sciences*. 1999; 96:11277–11282.
42. King RD, Rowland J, Oliver SG, Young M, Aubrey W, Byrne E, Liakata M, Markham M, Pir P, Soldatova LN, Sparkes A, Whelan KE, Clare A. The automation of science. *Science*. 2009; 324:85–89. [PubMed: 19342587]
43. Muggleton SH. 2020 computing: exceeding human limits. *Nature*. 2006; 440:409–410. [PubMed: 16554781]
44. King RD, Whelan KE, Jones FM, Reiser PG, Bryant CH, Muggleton SH, Kell DB, Oliver SG. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*. 2004; 427:247–252. [PubMed: 14724639]
45. Materassi D, Baschieri P, Tiribilli B, Zuccheri G, Samori B. An open source/real-time atomic force microscope architecture to perform customizable force spectroscopy experiments. *Rev Sci Instrum*. 2009; 80
46. Jens S, Reiner W, Mirko L, Joao N, Harald J, Ulrich G, Gerd H, Torsten J, Daniel JM. Fully automated single-molecule force spectroscopy for screening applications. *Nanotechnology*. 2008; 19:384020. [PubMed: 21832579]
47. Kasas S, Riederer BM, Catsicas S, Cappella B, Dietler G. Fuzzy logic algorithm to extract specific interaction forces from atomic force microscopy data. *Rev Sci Instrum*. 2000; 71:2082–2086.
48. Odorico M, Teulon JM, Berthoumieu O, Chen SwW, Parot P, Pellequer JL. An integrated methodology for data processing in dynamic force spectroscopy of ligand–receptor binding. *Ultramicroscopy*. 2007; 107:887–894. [PubMed: 17644254]
49. Puchner EM, Franzen G, Gautel M, Gaub HE. Comparing proteins by their unfolding pattern. *Biophys J*. 2008; 95:426–434. [PubMed: 18550806]
50. Kuhn M, Janovjak H, Hubain M, Muller D. Automated alignment and pattern recognition of single-molecule force spectroscopy data. *J Microsc*. 2005; 218:125–132. [PubMed: 15857374]
51. Sandal M, Benedetti F, Brucale M, Gomez-Casado A, Samori B. Hooke: an open software platform for force spectroscopy. *Bioinformatics (Oxford, England)*. 2009; 25:1428–1430.
52. Dietz H, Rief M. Detecting Molecular Fingerprints in Single Molecule Force Spectroscopy Using Pattern Recognition. *Japanese Journal of Applied Physics*. 2007; 46:5540–5542.
53. Andreopoulos B, Labudde D. Efficient unfolding pattern recognition in single molecule force spectroscopy data. *Algorithms for Molecular Biology*. 2011; 6:16. [PubMed: 21645400]
54. Bosshart, Patrick D.; Frederix, Patrick LTM.; Engel, A. Reference-Free Alignment and Sorting of Single-Molecule Force Spectroscopy Data. *Biophysical Journal*. 2012; 102:2202–2211. [PubMed: 22824285]
55. Steward A, Toca-Herrera JL, Clarke J. Versatile cloning system for construction of multimeric proteins for use in atomic force microscopy. *Protein Sci*. 2002; 11:2179–2183. [PubMed: 12192073]
56. Wang CC, Tsong TY, Hsu YH, Marszalek PE. Inhibitor binding increases the mechanical stability of staphylococcal nuclease. *Biophys J*. 2011; 100:1094–1099. [PubMed: 21320455]
57. Stacklies W, Vega MC, Wilmanns M, Grater F. Mechanical network in titin immunoglobulin from force distribution analysis. *PLoS Comput Biol*. 2009; 5:e1000306. [PubMed: 19282960]
58. Cecconi C, Shank EA, Marqusee S, Bustamante C. DNA molecular handles for single-molecule protein-folding studies by optical tweezers. *Methods Mol Biol*. 2011; 749:255–271. [PubMed: 21674378]
59. Rounsevell R, Forman JR, Clarke J. Atomic force microscopy: mechanical unfolding of proteins. *Methods*. 2004; 34:100–111. [PubMed: 15283919]

60. Marko JF, Siggia ED. Stretching DNA. *Macromolecules*. 1995;8759–8770.
61. Anczykowski B, Gotsmann B, Fuchs H, Cleveland JP, Elings VB. How to measure energy dissipation in dynamic mode atomic force microscopy. *Applied Surface Science*. 1999; 140:376–382.
62. Nuzzo RG, Allara DL. Adsorption of bifunctional organic disulfides on gold surfaces. *Journal of the American Chemical Society*. 1983; 105:4481–4483.
63. Nakanishi K, Sakiyama T, Kumada Y, Imamura K, Imanaka H. Recent Advances in Controlled Immobilization of Proteins onto the Surface of the Solid Substrate and Its Possible Application to Proteomics. *Current Proteomics*. 2008; 5:161–175.
64. Mrksich M, Chen CS, Xia Y, Dike LE, Ingber DE, Whitesides GM. Controlling cell attachment on contoured surfaces with self-assembled monolayers of alkanethiolates on gold. *Proc Natl Acad Sci U S A*. 1996; 93:10775–10778. [PubMed: 8855256]
65. Schmitt L, Ludwig M, Gaub HE, Tampé R. A Metal-Chelating Microscopy Tip as a New Toolbox for Single-Molecule Experiments by Atomic Force Microscopy. *Biophysical Journal*. 2000; 78:3275–3285. [PubMed: 10828003]
66. Kamruzzahan ASM, Ebner A, Wildling L, Kienberger F, Riener CK, Hahn CD, Pollheimer PD, Winklehner P, Hölzl M, Lackner B, Schörkl DM, Hinterdorfer P, Gruber HJ. Antibody Linking to Atomic Force Microscope Tips via Disulfide Bond Formation. *Bioconjugate Chemistry*. 2006; 17:1473–1481. [PubMed: 17105226]
67. Taniguchi Y, Kawakami M. Application of HaloTag protein to covalent immobilization of recombinant proteins for single molecule force spectroscopy. *Langmuir*. 2010; 26:10433–10436. [PubMed: 20527958]
68. Kienberger F, Kada G, Gruber HJ, Pastushenko VP, Riener C, Trieb M, Knaus HG, Schindler H, Hinterdorfer P. Recognition Force Spectroscopy Studies of the NTA-His6 Bond. *Single Molecules*. 2000; 1:59–65.
69. Wong J, Chilkoti A, Moy VT. Direct force measurements of the streptavidin-biotin interaction. *Biomol Eng*. 1999; 16:45–55. [PubMed: 10796984]
70. Kim M, Wang CC, Benedetti F, Marszalek PE. A Nanoscale Force Probe for Gauging Intermolecular Interactions. *Angewandte Chemie*. 2012; 124:1939–1942.
71. He Y, Lu M, Cao J, Lu HP. Manipulating protein conformations by single-molecule AFM-FRET nanoscopy. *ACS Nano*. 2012; 6:1221–1229. [PubMed: 22276737]
72. Sarkar A, Robertson RB, Fernandez JM. Simultaneous atomic force microscope and fluorescence measurements of protein unfolding using a calibrated evanescent wave. *Proc Natl Acad Sci U S A*. 2004; 101:12882–12886. [PubMed: 15326308]
73. Shlyakhtenko LS, Gall AA, Filonov A, Cerovac Z, Lushnikov A, Lyubchenko YL. Silatrane-based surface chemistry for immobilization of DNA, protein-DNA complexes and other biological materials. *Ultramicroscopy*. 2003; 97:279–287. [PubMed: 12801681]

Highlights

- An algorithm for automating selection in AFM spectroscopy experiments is proposed
- Algorithm succeeds in being fast, specific, and sensitive enough for use in real-time selection of usable recordings
- The algorithm is found to be robust, especially when used with proteins composed of repeated domains, and can efficiently detect weak unfolding events
- Real-time selection provides information about statistics of efficacy and efficiency of AFM spectroscopy
- Statistics were used to pinpoint the optimal set of experimental conditions

FIG. 1. Example of algorithm correctly identifying the unfolding events in a recording of I27₃-NI₁₀C-I27₃ (NI10C) which contains both small, tightly spaced Ankyrin unfolding events, and high-force I27 unfolding events. (A) The windowed fast Fourier transform (WFFT) of the force-extension (FE) curve is calculated as described in 2.1.1. The local maxima (red triangles) are then determined by comparing each element of the WFFT with its neighbors. (B) The user supplies the force-extension regimes for necessary protein unfolding events. (C) The peaks are assigned to be specific proteins based on distance between neighboring peaks and their force threshold according to the force-regimes (here only regimes for Ank and I27 are used).

FIG. 2.

Representative example of adding noise to a recording of NI10C. The detected peaks are shown with triangles and then amount of noise added is labeled (RMS of the noise). The detection of the small Ank unfolding events become less reliable as the noise is increased, but remain fairly stable until 12 pN of noise is added. The identification of I27 events is very reliable up to 24 pN, although many nonspecific events become present at this high level of noise.

FIG. 3.

The probability for detecting 1–7 force peaks is plotted, normalized against the total number of molecules in which force peaks were detected.

FIG. 4.

The efficiency of the pulling experiments was determined by the percentage of all recordings that resulted in at least one force rupture event divided by the total number of recordings. The only statistically significant contribution to the variance in the efficiency was the age of proteins, shown here ($p = 0.008$ with Bonferroni correction).

Table 1

Protein	Sensitivity %	Specificity % (n=158)
I27 (3 domains, n=20)	90.5	96.2
I27 (4+ domains, n=46) Supplementary Figure S5	100.0	96.2
SNase (n=39, 8 w/I27) Supplementary Figure S6	100.0	90.5
NI10C/NI6C (n=41, 25 w/I27) Supplementary Figure S7,S8	100.0	90.5

Table 2

Protein	Sensitivity %	Specificity % (n=158)	Noise added (pN)
3I27/4+I27	90.5/100.0	96.2	+3
3I27/4+I27	90.5/100.0	96.2	+6
3I27/4+I27	90.5/100.0	96.2	+12
3I27/4+I27	80.6/100.0	96.2	+24
NI6C/NI10C	100.0	74.1	+3
NI6C/NI10C	100.0	71.4	+6
NI6C/NI10C	100.0	57.7	+12
NI6C/NI10C	100.0	37.4	+24
SNase	100.0	89.2	+3
SNase	100.0	89.2	+6
SNase	97.4	79.8	+12
SNase	97.4	58.9	+24