Ultramicroscopy 136 (2014) 7–14

Contents lists available at ScienceDirect

Ultramicroscopy

journal homepage: www.elsevier.com/locate/ultramic

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

Zackary N. Scholl [a,*], Piotr E. Marszalek [b,*]

[a] Program in Computational Biology and Bioinformatics, Duke University, Durham, NC 27708, USA
[b] Department of Mechanical Engineering and Materials Science, Center for Biologically Inspired Materials and Material Systems, and Duke University, Durham, NC 27708, USA

ARTICLE INFO

Article history:
Received 28 January 2013
Received in revised form
17 July 2013
Accepted 25 July 2013
Available online 7 August 2013

Keywords:
Atomic force
Single molecules
Single molecule experiments
Single molecule microscopy
Proteins
Protein adsorption of surfaces

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.

Published by Elsevier B.V.

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 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 s; this means that the full mechanical characterization of a protein can require 12–1400 h 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

* Corresponding authors.
E-mail addresses: zns@duke.edu (Z.N. Scholl), pemar@duke.edu (P.E. Marszalek).

0304-3991/$ - see front matter Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.ultramic.2013.07.020

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 (Fig. 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.

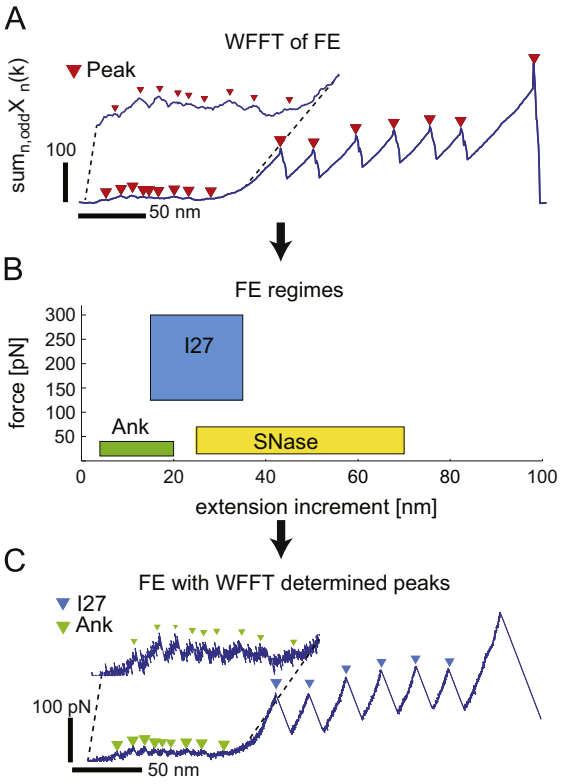

Fig. 1. Example of algorithm correctly identifying the unfolding events in a recording of $I27_3$-$NI_{10}C$-$I27_3$ (NI10C) which contains both small, tightly spaced Ankyrin unfolding events, and high-force I27 unfolding events. (A) The windowed fast Fourier transform (WFFT) of the force–extension (FE) 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). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article)

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