Japanese Journal of Applied Physics
Vol. 46, No. 8B, 2007, pp. 5540–5542

Brief Communication

Detecting Molecular Fingerprints in Single Molecule Force Spectroscopy Using Pattern Recognition

Hendrik DIETZ and Matthias RIEF*

Physik Department E22, Technische Universität München, 85748 Garching bei München, Germany

(Received February 2, 2007; accepted March 19, 2007; published online August 23, 2007)

Single molecule force spectroscopy has given experimental access to the mechanical properties of protein molecules. Typically, less than 1% of the experimental recordings reflect true single molecule events due to abundant surface and multiple-molecule interactions. A key issue in single molecule force spectroscopy is thus to identify the characteristic mechanical "fingerprint" of a specific protein in noisy data sets. Here, we present an objective pattern recognition algorithm that is able to identify fingerprints in such noisy data sets. [DOI: 10.1143/JJAP.46.5540]

KEYWORDS: single molecule, protein mechanics, force spectroscopy, AFM, pattern recognition, GFP

One of the most fundamental and challenging problems in molecular biophysics is understanding how proteins fold into complex three dimensional structures. The folding process of proteins is generally described as diffusion in a high dimensional energy-landscape.[1] Recent advances in single molecule force spectroscopy have made it possible to explore the energy landscape of single protein molecules along well-defined reaction coordinates by applying a mechanical force.[2–7] Several aspects make mechanical experiments with single molecules a valuable tool for protein science. A large fraction of the proteins in our body have structural and thus also mechanical function.[8] Single-molecule force experiments can help to investigate the mechanical function of proteins. But, beyond physiology, force as structural control parameter also offers attractive possibilities for exploring the energy landscape of biomolecule folding.

Single molecule force spectroscopy experiments are challenging and due to abundant surface effects and interactions with multiple molecules generally less than 1% of the experimental recordings reflect true single molecule manipulation. A key issue in single-molecule force spectroscopy is to identify the characteristic mechanical "fingerprint" of a specific protein. One strategy to obtain a clear selection criterion is the use of modular proteins (polyproteins). Like in the case of the muscle protein titin [see Fig. 1(a)], stretching a chain of identical protein subunits results typically a characteristic and repetitive sawtooth pattern. Each peak reflects the sudden breakdown of a folded protein domain in the stretched chain, followed by stretching of the lengthened polyprotein. The distance between peaks is related to the number of amino acids that unravel in each unfolding event. Only curves exhibiting such sawtooth patterns are then used for data analysis. However, the vast majority of protein molecules do not exhibit modular structure. Recombinant protocols have been developed to construct artificially polyprotein chains.[9,10] Cysteine engineering can be employed to generate polyproteins with different linking geometries.[11–14] Alternatively, it may be advantageous to embed a single molecule of the protein of interest into another modular protein chain to create molecular handles that enable manipulation of the protein of interest [see Fig. 1(b)]. An example of a force-extension trace obtained with such a modular fusion protein is shown in Fig. 1(b). The protein was engineered such that a single green fluorescent protein (GFP) molecule is flanked by several modules of *Dictyostelium discoideum* filamin (DdFLN domains 1 to 5).[5] The section marked in green in the force extension trace reflects the mechanical signature arising from force-induced unfolding of the single GFP molecule, while the red sections arise from unfolding of the handle domains. The fusion-protein strategy enables control upon time and location of the desired unfolding event in force extension traces, but complicates acquisition of sufficient data. Another concern is the identification of the mechanical signature arising from stretching of the protein of interest. Here we introduce a pattern recognition algorithm for the evaluation of molecular fingerprints within noisy experimental data.

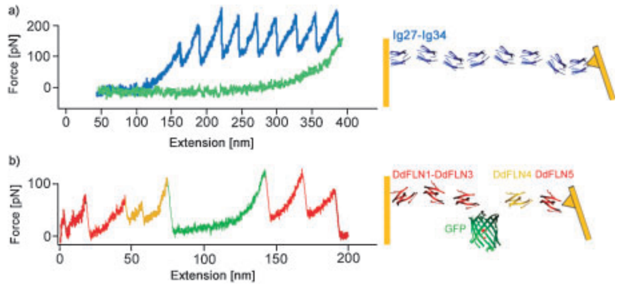

Fig. 1. (a) Typical force-extension trace obtained from stretching single titin polyprotein molecules. Each peach reflects force-induced unfolding of a single titin module in the polyprotein. (b) Typical force-extension trace measured on a chimeric protein containing a single GFP domain flanked by domains of the dictyostelium discoideum actin crosslinking protein filamin [DdFln(1–5)]. The section marked in green reflects the mechanical signature due to force-induced unfolding of GFP.[5] The section marked in yellow corresponds to unfolding of a special DdFLN domain exhibiting a stable unfolding intermediate as it has been described before.[19]

We assume a function $g(x)$ to represent a mechanical pattern of interest, while the function $f(x)$ shall be tested for sections that exhibit similarity with the pattern $g(x)$. We make use of cross correlations as functions of a displacement variable u to develop such a testing procedure.

$$K_{g,f}(u) = \int_0^b g(x) \cdot f(x+u)\,dx \tag{1}$$

The number b in eq. (1) denotes the width of the pattern $g(x)$. We seek a recognition function $C_{g,f}(u)$ which equals 1 for displacements u leading to identical matching with the

*E-mail address: mrief@ph.tum.de

pattern $g(x)$ and otherwise is always smaller than 1. Thus, if the function $f(x)$ contains exactly the pattern $g(x)$ at a certain displacement u, the condition in eq. (2) has to be fulfilled.

$$\lim_{f \to g} \frac{\int_0^b g(x) \cdot f(x+u)\,dx}{\int_0^b g(x) \cdot g(x)\,dx} = 1 \tag{2}$$

This condition motivates to introduce a displacement-dependent (u-dependent) normalization for the cross-correlation function $K_{g,f}$. The inequality of Cauchy–Schwartz can be used for such a purpose:

$$\frac{\left(\int_0^b g(x) \cdot f(x+u)\,dx\right)^2}{\int_0^b g^2(x)\,dx \cdot \int_0^b f^2(x+u)\,dx} \leq 1 \tag{3}$$

Now we define the recognition function $C_{g,f}(u)$ as the square root of the left side of inequality (3) and end up with a function fulfilling the desired condition (2).

$$C_{g,f}(u) = \frac{K_{g,f}(u)}{\sqrt{\int_0^b g^2(x)\,dx \cdot \int_0^b f^2(x+u)\,dx}} \tag{4}$$

The displacement $u_{C=\max}$ for which the function $C_{g,f}(u)$ assumes a maximum yields immediately the section of width b of the test function $f(x)$ with best matching with the pattern $g(x)$. This "best matching section" is termed $F(x)$ in the following.

$$F(x) \equiv f(x - u_{\max});\ x \in [0, b] \tag{5}$$

The value of the correlation function $C_{g,F}(u_{C=\max})$ at displacement $u_{C=\max}$ represents a measure for a *degree of coincidence*, i.e., the goodness of the matching between a pattern $g(x)$ and the best matching section $F(x)$ of a test function $f(x)$. However, this measure still does not provide an absolute measure to compare different test function. For the necessary refinement we consider the best matching sections of two different test functions, $F(x)$, and $W(x) = \eta \cdot F(x)$. In this case the correlation function $C_{g,F}(0)$ equals $C_{g,W}(0)$ for both test functions — which is undesirable. A scaling correction needs to be introduced:

$$S_{g,F} = \begin{cases} \sqrt{\dfrac{\langle F^2 \rangle}{\langle g^2 \rangle}} & \langle F^2 \rangle \leq \langle g^2 \rangle \\[2ex] \sqrt{\dfrac{\langle g^2 \rangle}{\langle F^2 \rangle}} & \langle g^2 \rangle < \langle F^2 \rangle \end{cases} \tag{6}$$

A generally valid definition of an absolute degree of coincidence Γ can now be defined by:

$$\Gamma \equiv C_{g,f}(u_{\max}) \cdot s_{g,F} \tag{7}$$

Such defined degree of coincidence can now, be used to perform statistical analysis of force-extension data sets for the appearance of a certain pattern. Γ assumes any value in the interval $0 \leq \Gamma \leq 1$, where increasing Γ reflects increasing matching between the best matching section of a test function and a given pattern. $\Gamma = 1$ is fulfilled if test function and pattern are identical. Another refinement improves substantially the resolution of the pattern recognition. This is achieved by reducing the pattern function $g(x)$ and the best matching part of the test function $F(x)$ by their corresponding average values, i.e., $\langle g(x) \rangle$ and $\langle F(x) \rangle$ and then calculate the value of the recognition function at displacement zero $C_{g-\langle g \rangle, f-\langle f \rangle}(0)$. A practical definition of an objective degree of coincidence γ is then given by

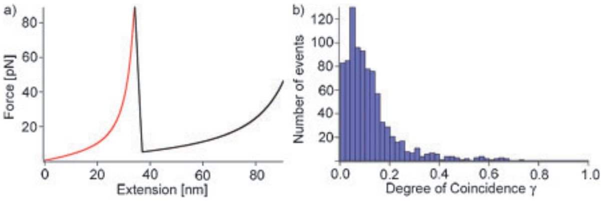

Fig. 2. (a) Simulated force-extension trace. As pattern function $g(x)$ we chose the section marked in black, corresponding to a calculated GFP unfolding pattern. (b) Distribution of γ-values as they have been calculated for an experimental data set containing 1012 force-extension traces measured on a modular protein containing a single GFP domain [DdFLN(1–5)-GFP].

$$\gamma \equiv C^2_{g-\langle g \rangle, F-\langle F \rangle}(0) \cdot s_{g,F}. \tag{8}$$

The number γ assumes positive real values ≤ 1. In the case of $\gamma = 1$ the pattern $g(x)$ matches perfectly to the section $F(x)$ of the test function $f(x)$. Squaring $C_{g-\langle g \rangle, f-\langle f \rangle}(0)$ in eq. (10) emphasizes recognition of the exact form of the pattern. Other definitions of a practical degree of coincidence may have to be considered to improve the resolution of the recognition and evaluation depending on the nature of the pattern.

The pattern recognition algorithm needs as input the pattern $g(x)$ it screens the data for. The right choice of $g(x)$ is hence an essential step. Since the unfolding pattern of GFP is already known from earlier experiments,[5] for the following test we use a simulated trace of GFP unfolding from a Monte-Carlo simulation[17] as our choice for $g(x)$. $g(x)$ corresponds to the section marked in black in Fig. 2(a). The pattern consists of a cantilever relaxation phase after unfolding and stretching of a polypeptide whose contour length has been increased by the length of the number of amino acids that make up the folded GFP barrel structure, i.e., 220 amino acids. The contour length gain equals 77 nm, that is, the contour length of 220 amino acids minus the initial distance between points of force application.[5,12]

As a set of test functions $f(x)$ we chose the complete data set of a force spectroscopy experiment containing 1012 single force extension traces. The investigated sample was DdFLN(1–5)-GFP as described above. The pattern recognition algorithm first identifies the best matching sections in each force trace and then calculates the corresponding value of the degree of coincidence γ. Figure 2(b) shows the resulting distribution of γ values obtained for this data set.

We recognize that γ values greater than 0.5 are rarely observed, while the vast majority of force traces is judged with very low γ values. Figure 3(a) shows four arbitrarily selected traces with a degree of coincidence between $[0.0, 0.01]$. The traces do not show any similarity with the given pattern. We note that they display hardly any interactions and/or only slow drift effects. Figure 3(b) (interval $[0.2, 0.21]$) shows traces with somewhat more interaction patterns with the cantilever, but we still note very poor similarity with the given pattern. Figure 3(c) (interval

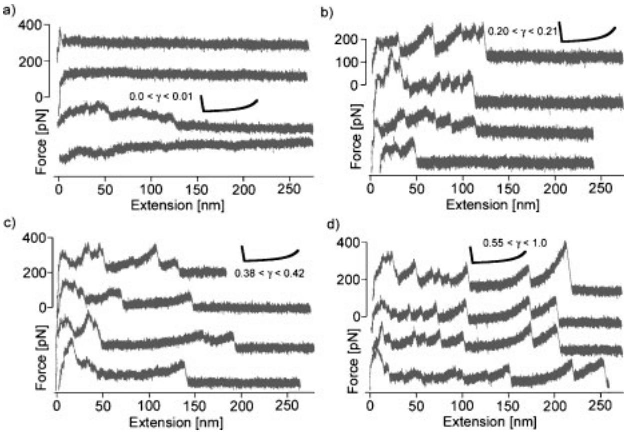

Fig. 3. Typical measured force-extension traces whose degree of coincidence with the pattern falls into four different intervals.

[0.38, 0.42]) in turn shows traces exhibiting sections with a certain similarity to the pattern, but still those structures match poorly the given pattern. Finally, Fig. 3(d) shows traces from the γ interval [0.55, 1] which exhibit sections that reproduce very well the given pattern. A manual analysis of the full data set of 1012 traces reveals that all force traces exhibiting GFP unfolding events have been assigned γ values greater than 0.55 by the pattern recognition routines.

Thus, we see that the pattern recognition enables filtering big data sets for traces exhibiting certain patterns. We also note that for each trace the calculated γ factor provides a quantitative degree of coincidence with a certain pattern.

We apply now the pattern recognition routines to compare two different data sets of similar size. One date set contains 1012 force extension traces obtained with the DdFLN(1–5)-GFP fusion protein (see above), while the other data set contains 934 force extension traces obtained on a different modular protein which does *not* contain GFP (Ig27–34 from human cardiac titin). The pattern we chose for the analysis is the idealized GFP unfolding pattern from Fig. 2(a). We thus expect that higher degrees of coincidence (γ values greater than 0.5) should only be observed in the data set obtained on the sample *containing* GFP. A superposition and close-up of the two resulting degree of coincidence distributions is shown in Fig. 4. We see that the distribution of γ values for the data set measured on the sample lacking GFP falls rapidly to zero. Only 0.5% of the total number of force curves have been assigned γ values greater than 0.2. The highest γ value is 0.46 and was assigned to only a single trace. In contrast, we see that the γ distribution calculated for the data set measured on the sample *containing* GFP falls much slower to zero. Still 6.4% of the total number of force curves have been assigned γ values greater than 0.2. Nine traces show pattern matching with a quality of γ greater than 0.5. As discussed above, such high γ values reflect very good matching with the pattern.

Thus, we see that the GFP unfolding pattern does only appear in the data set measured on a sample containing GFP, while this pattern cannot be found in the data set measured on a sample lacking GFP. We also find that the frequency of force traces with intermediate matching to this pattern is an order of magnitude higher in the data set obtained on the GFP fusion protein than in the other data set, suggesting that these more complex patterns result from multiple molecule interactions of similar mechanical properties that may be deciphered in the future.[18]

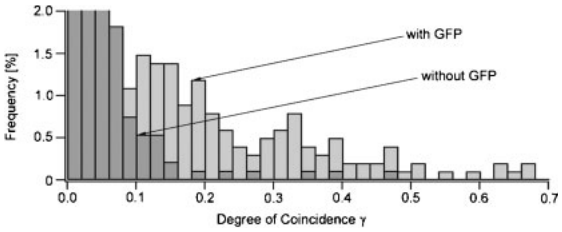

Fig. 4. Comparison of the distribution of the degree of coincidence γ as calculated for two different data sets.

It is crucial for noisy single molecule experiments to develop and apply objective criteria for the selection of data traces. We have presented such an approach based on correlation functions.

Acknowledgments

Financial support of the German Excellence Initiative via the Nanosystems Initiative Munich (NIM) is gratefully acknowledged.

1) J. N. Onuchic and P. G. Wolynes: Curr. Opin. Struct. Biol. **14** (2004) 70.
2) M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez, and H. E. Gaub: Science **276** (1997) 1109.
3) M. Carrion-Vazquez, A. F. Oberhauser, S. B. Fowler, P. E. Marszalek, S. E. Broedel, J. Clarke, and J. M. Fernandez: Proc. Natl. Acad. Sci. U.S.A. **96** (1999) 3694.
4) P. E. Marszalek, H. Lu, H. Li, M. Carrion-Vazquez, A. F. Oberhauser, K. Schulten, and J. M. Fernandez: Nature **402** (1999) 100.
5) H. Dietz and M. Rief: Proc. Natl. Acad. Sci. U.S.A. **101** (2004) 16192.
6) M. T. Alam, T. Yamada, U. Carlsson, and A. Ikai: FEBS Lett. **519** (2002) 35.
7) R. Hertadi and A. Ikai: Protein Sci. **11** (2002) 1532.
8) C. Bustamante, Y. R. Chemla, N. R. Forde, and D. Izhaky: Annu. Rev. Biochem. **73** (2004) 705.
9) H. Li, A. F. Oberhauser, S. B. Fowler, J. Clarke, and J. M. Fernandez: Proc. Natl. Acad. Sci. U.S.A. **97** (2000) 6527.
10) M. Carrion-Vazquez, A. F. Oberhauser, T. E. Fisher, P. E. Marszalek, H. Li, and J. M. Fernandez: Prog. Biophys. Mol. Biol. **74** (2000) 63.
11) H. Dietz, M. Bertz, M. Schlierf, F. Berkemeier, T. Bornschlogl, J. Junker, and M. Rief: Nat. Protocols **1** (2006) 80.
12) H. Dietz and M. Rief: Proc. Natl. Acad. Sci. U.S.A. **103** (2006) 1244.
13) H. Dietz, F. Berkemeier, M. Bertz, and M. Rief: Proc. Natl. Acad. Sci. U.S.A. **103** (2006) 12724.
14) G. Yang, C. Cecconi, W. A. Baase, I. R. Vetter, W. A. Breyer, J. A. Haack, B. W. Matthews, F. W. Dahlquist, and C. Bustamante: Proc. Natl. Acad. Sci. U.S.A. **97** (2000) 139.
15) F. Kuhner, M. Erdmann, L. Sonnenberg, A. Serr, J. Morfill, and H. E. Gaub: Langmuir **22** (2006) 11180.
16) M. Seitz, C. Friedsam, W. Jostl, T. Hugel, and H. E. Gaub: ChemPhysChem **4** (2003) 986.
17) M. Rief, J. M. Fernandez, and H. E. Gaub: Phys. Rev. Lett. **81** (1998) 4764.
18) G. E. Fantner, E. Oroudjev, G. Schitter, L. S. Golde, P. J. Thurner, M. M. Finch, P. Thurner, T. Gutsmann, D. E. Morse, H. Hansma, and P. K. Hansma: Biophys. J. **90** (2005) 1411.
19) I. Schwaiger, A. Kardinal, M. Schleicher, A. A. Noegel, and M. Rief: Nat. Struct. Mol. Biol. **11** (2004) 81.