Data and text mining

Automatic classification of single-molecule force spectroscopy traces from heterogeneous samples

Nina I. Ilieva [1,], Nicola Galvanetto [1,5]*, Michele Allegra [1,3] , Marco Brucale [4], and Alessandro Laio [1,2]*

[1] Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, 34136, Italy
[2] The Abdus Salam International Centre for Theoretical Physics (ICTP), Trieste, 34151, Italy
[3] Institut de Neurosciences de la Timone UMR 7289, Aix Marseille Université, CNRS, Marseille, 13005, France
[4] Consiglio Nazionale delle Ricerche, Istituto per lo Studio dei Materiali Nanostrutturati (CNR-ISMN), Bologna, Italy
[5] Present Address: University of Zurich, Winterthurerstrasse 190, 8057, Zurich, Switzerland

* To whom correspondence should be addressed.

Abstract

Motivation: Single-molecule force spectroscopy (SMFS) experiments pose the challenge of analyzing protein unfolding data (traces) coming from preparations with heterogeneous composition (e.g. where different proteins are present in the sample). An automatic procedure able to distinguish the unfolding patterns of the proteins is needed. Here, we introduce a data analysis pipeline able to recognize in such datasets traces with recurrent patterns (clusters).

Results: We illustrate the performance of our method on two prototypical datasets: $\sim$ 50,000 traces from a sample containing tandem GB1 and $\sim$ 400,000 traces from a native rod membrane. Despite a daunting signal-to-noise ratio in the data, we are able to identify several unfolding clusters. This work demonstrates how an automatic pattern classification can extract relevant information from SMFS traces from heterogeneous samples without prior knowledge of the sample composition.

Availability: https://github.com/ninailieva/SMFS_clustering

Contact: laio@sissa.it; nicola.galvanetto@sissa.it

1 Introduction

Atomic force microscopy (AFM)-based single molecule force spectroscopy (SMFS) is a powerful tool for studying proteins at the single molecule level. In a typical AFM-SMFS experiment, the protein is bonded on one side to a surface, and attaches on the other side to the AFM tip (Engel and Gaub, 2008). As the tip retracts from the surface, the protein gets stretched and unfolded. The resulting force and extension values are stored in the form of a force-extension (F-x) curve. In a single experimental session, thousands of F-x curves are generated. Therefore, a dataset can easily contain more than 10^5 curves. F-x curves, also called traces, are direct representations of the protein unfolding pathway and can be used to fingerprint specific proteins (Rief *et al.*, 1997; Oesterhelt *et al.*, 2000; Maity *et al.*, 2015).

The vast majority of protein-related AFM-SMFS studies published so far were performed on either membrane proteins or soluble globular proteins. Membrane proteins can be purified and reconstituted in synthetic lipids. This procedure yields SMFS datasets that are very homogeneous, as they include traces with the same F-x profile and length. Recently, however, an improvement in extracting membrane proteins directly from their original membrane has been made (Galvanetto, 2018). Datasets coming from native membranes are highly heterogeneous: the majority of traces do not represent meaningful unfolding events, and if they do, they likely represent the unfolding of different proteins, since a native membrane hosts hundreds of different proteins. Soluble globular proteins are usually engineered in tandem constructs so as to have multiple copies of the same domain in the same aminoacid chain. These datasets are usually heterogeneous, too. Even though they contain one single protein type, the proteins in the sample can be hooked at different positions generating traces of different length.

Available data analysis tools (Kuhn *et al.*, 2005; Marsico *et al.*, 2007; Bosshart *et al.*, 2012; Galvanetto *et al.*, 2018), which work reasonably well for AFM-SMFS traces coming from experiments performed with purified membrane proteins, perform poorly when applied to sets of highly heterogeneous traces from multiprotein samples. The most important stumbling block is possibly trace selection, because complete unfolding

is observed in less than 1% of the traces (Bosshart *et al.*, 2012). In heterogeneous samples, the selection cannot be based on the protein contour length like in homogeneous samples.

In this work, we introduce a procedure which allows the classification of highly heterogeneous SMFS datasets. The main idea is to detect sets of traces with recurrent F-x patterns, emerging in a vast population of traces corresponding to statistically isolated events. To find these patterns we use density-peak clustering (Rodriguez and Laio, 2014), an approach which allows detecting the maxima in multidimensional probability distributions using as input only the distance between each pair of data points (*the traces*, in the case of this work). We estimate the distance between pairs of traces using a modified version of the metric introduced by Marsico *et al.*. The procedure is fully automatic. In addition, it allows the processing of large amount of data in a reasonable computational time. It takes $\sim$ 30 minutes to process 10^5 traces on a workstation with 16 CPUs.

We first show that our method can discriminate a single set of meaningful traces, corresponding to the unfolding of a protein, from a set of traces containing noise. We then show that the approach can discriminate between different types of meaningful traces, corresponding to the unfolding of different proteins. To this aim we first analyze a dataset containing subsets of traces corresponding to the unfolding of four different proteins. Next, we test our method on data generated in 20 experiments on a tandem globular construct of GB1 proteins, and finally on a highly-heterogeneous dataset of $\sim$ 400,000 traces from experiments performed in the native plasma membrane of the rod outer segment under physiological conditions.

2 The Algorithm

2.1 Overview and workflow.

The workflow of the algorithm we developed consists of the four major blocks depicted in Figure 1. The first block, "Cutting & filtering", aims at removing the physically irrelevant parts of the original trace, clearing the space to meaningful unfolding events. The filtering consists in discriminating spurious traces (Figure 1). In the second block, a quality score is computed for each trace based on the features of its contour length (L_c) histogram. Depending on the score, a trace selection is performed, significantly reducing the total number of traces. For each pair of the remaining traces, a similarity distance, almost identical to the one used by Marsico *et al.*, is computed and density-peak clustering is applied to classify the emerging recurrent F-x patterns into separate clusters (Figure 1). In what follows, we provide a detailed description of each block of the algorithm.

2.2 Cutting and filtering.

Initially, each F-x curve is processed in order to remove all irrelevant parts of the signal. Typically, a F-x curve contains a *contact* and a *non-contact* part. The contact part comes from the interactions between the AFM tip and the membrane surface. This part starts with highly negative forces due to the upwards bending of the cantilever in the beginning of the retraction cycle. The non-contact part (or the tail) is noise coming from the free motion of the cantilever that is no longer in contact with the surface. This part is used to set the baseline of zero force. We remove the negative contact part and the tail of each trace as follows. For each trace, we find the first point at extension larger than 0 nm, followed by $n_{cont} = 20$ (all the parameters used in the algorithm are listed in Table S1) consecutive force measures having positive values. We mark this point as the *starting point* because this is where the positive contact part begins. We exclude from the analysis the signal preceding the starting point. In order to identify the non-contact part, we start from the end of the trace and move backwards until the force exceeds the range compatible with the free motion of the cantilever. In detail, we estimate the standard deviation of the force σ_{NOISE} in 10

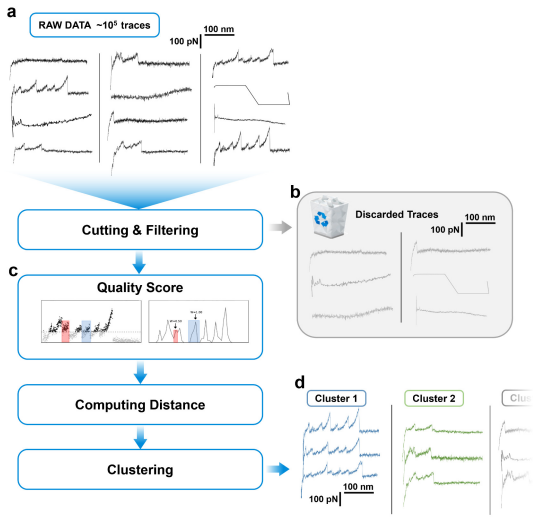

Fig. 1. Block scheme of the algorithm.

manually selected tails. In general, this value depends on the instrument and on the experimental condition (e.g. it is $\sigma_{NOISE} = 5.67$ pN for our dataset 'Rod'). We then perform a linear fit to the last 8 nm of each trace, extending the window stepwise in the backward direction by 2 nm (hence considering the last 10 nm, 12 nm, 14 nm, etc.; see Figure S1) . At each step we compute the standard deviation from the fit and check whether it exceeds the cutting threshold, $\sigma_{cut} = 4\sigma_{NOISE}$. If not, we continue, otherwise we stop and cut the trace there. We assume that at this point the last force peak has been reached and the non-contact part has ended. In our procedure, the position of the last force peak determines the trace length.

At this point we store the traces on a regular grid with width $\Delta x_{interp} = 1$ nm, and we perform filtering, which aims at selecting only traces which are likely to correspond to the unfolding of a protein. In Figure S2 we show some examples of traces discarded by our procedure. A trace is discarded if it does not contain a detectable contact point, e.g. if the starting point of the contact part is blurred; if the points occupy an anomalously wide force range; if after the tail removal, the trace is shorter than a minimal length $L_{min} = 50$ nm and if the trace contains abnormal deflection points (with values larger than $x_{abn} = 5000$ nm or/and $F_{abn} = 5000$ pN), namely if the entire signal is shifted upwards with respect to the zero force baseline. The non-contact part in relevant traces is normally flat with fluctuations of the force compatible with σ_{NOISE}. By detecting the position of the last force peak, we obtain the total length of the non-contact part. This allows to compute its standard deviation from a horizontal zero-force line. If this deviation exceeds the threshold $\sigma_{tail} = 2\sigma_{NOISE}$, the trace is considered spurious and is discarded.

2.3 Quality score.

We then compute a quality score which we use for further selecting the meaningful F-x curves. The score we propose quantifies how well the experimental data satisfy the worm-like chain (WLC) model, which is the standard model used for the analysis of F-x curves of linear (bio)polymers (Rief *et al.*, 1997; Oesterhelt *et al.*, 2000; Maity *et al.*, 2015, 2017). This choice excludes possibly meaningful traces corresponding to an unfolding not described by this model. In future applications one can consider removing or modifying this filter. According to the WLC model, a F-x curve represents the unfolding of different domains, each unfolding curve (each "tooth" in the pattern) being described by

$$F(x) = \frac{k_B T}{l_p}\left(\frac{1}{4}\left(1-\frac{x}{L_c}\right)^{-2} + \frac{x}{L_c} - \frac{1}{4}\right) \qquad (1)$$

where F is force, x is extension, k_B is Boltzmann's constant, T is temperature, l_p is the persistence length and L_c is the contour length of the domain. Usually, the WLC model is *assumed* and one retrieves the L_c corresponding to different domains by means of an L_c histogram (Bosshart *et al.*, 2012; Galvanetto *et al.*, 2018). The WLC equation is inverted to find an L_c value for any x, and the resulting L_c values are plotted in a histogram. Ideally, the L_c histogram should consist of narrow peaks centered at the L_c values corresponding to the contour lengths of each domain. Thus, the L_c values corresponding to the maxima of the histogram are taken to be the contour length values for each domain. L_c histograms of meaningful traces are characterized by the presence of a few maxima, well separated by deep minima. We will define a score that quantifies how well the data agree with this expectation.

For each point in the F-x curve we compute L_c value by solving Equation (1). We use a persistence length $l_p = 0.4$ nm, which is considered appropriate for membrane proteins (Oesterhelt *et al.*, 2000). The L_c is computed in this manner in the force range from $F_{min}^{WLC} = 30$ to $F_{max}^{WLC} = 500$ pN which is the range of validity of the model (Bosshart *et al.*, 2012).

A critical parameter for our algorithm is the bin width used for computing the histogram. Since the traces are unavoidably affected by noise, if the bin width is too small (say smaller than 5 nm), the histograms are characterized by the presence of spurious peaks, which do not correspond to genuine force peaks. On the opposite, if the bin width is too large (say more than 10 nm), peaks corresponding to the unfolding of different domains get merged. In Figure S3 b, d, f we show such histograms for three traces from the CNG dataset. We have chosen a bin size of 8 nm corresponding to approximately 20 a.a., which is close to the typical length of a single transmembrane helix in membrane proteins (Lehninger, 2000). In fact, in Supplementary Information we show that the results of a clustering analysis are only mildly affected if one changes this to 7 nm or 9 nm, and only the precise size of the clusters varies and only the precise size of the clusters is affected by this parameter.

Once we have the histogram, we find all maxima and minima. A maximum is considered meaningful if it is generated by more than 5 points and it includes more than 1% of the force measures of a trace Next, we compute score W quantifying the consistency of each maximum with the WLC model. We assume that a high quality peak has its two surrounding minima falling under $\frac{1}{2}$ of the peak height. We define $f_{left} = P_{left}/P_{max}, f_{right} = P_{right}/P_{max}$ where P_{max}, P_{left} and P_{right} are the probability densities of the maximum, of the minimum at its left and of the minimum at its right. The ideal trace will yield $f = \frac{1}{2}(f_{left} + f_{right}) \sim 0$. We embed this requirement by estimating the score of the peak as $W = \exp\left(-2f^2\right)$. If, for example, $P_{left} = 1$, $P_{right} = 2$ and $P_{max} = 16$ one gets $W = 0.98$. If instead $P_{left} = 13$ and $P_{right} = 14$, the peak has a low quality and one gets $W = 0.24$. In Figure S3 we provide a few examples of F-x curves, their L_c histograms and the W-score for some peaks.

Subsequently, we assign the corresponding peaks scores to all points in each trace. A score is assigned to a point in two steps: we assign the peak's score to all points contributing to the peak. If a point has a force smaller than 30 pN it is not assigned to any peak, since the signal-to-noise ratio is too small for small forces. We therefore assign to it the same score of the first successive point whose force is larger than 30 pN. We apply this criterion only for points that are within 75 nm from the last point assigned to the peak (Figure S4). We selected this value by visual inspection of the traces, estimating the maximum widths of force peaks.

Finally, we sum up the scores for all points and we obtain the global score or the quality score of the trace, S_w. The higher the global score, the higher the overall quality of that trace. The ratio between the global score S_w and the trace length L is used to select high quality traces for subsequent analysis. If this ratio is smaller than 0.5, the trace is discarded (Figure S2). This is the same as saying: if more than half of the trace is inconsistent with the WLC model, it is a low quality trace and we are not interested in analyzing it. On the contrary, if more than half of the trace is consistent with the WLC model, it is possibly a meaningful trace. In Supplementary Information we show that the final results we obtain are quite robust with respect to small variations in the value of this parameter. Therefore, the score/length threshold is not a critical parameter of our approach.

2.4 Compute distances.

The final goal of our procedure is finding in an automatic manner meaningful F-x curves bearing a specific unfolding pattern and grouping them into clusters based on their similarity to each other. We use the distance introduced by Marsico *et al.*, which is based on dynamic programming alignment score, to evaluate the similarity between two traces. The distance between traces a and b, denoted by d_{ab} is:

$$d_{ab} = 1 - \frac{S_D(N_a, N_b)}{N_{max}} \tag{2}$$

where $S_D(N_a, N_b)$ is the global alignment score, N_a is the length of trace a, N_b is the length of trace b, and $N_{max} = max(N_a, N_b)$. In our method, we modified the original scoring function used to evaluate match/mismatch score as follows:

$$M(i,j) = \begin{cases} 1 - \frac{|F_a(i) - F_b(j)|}{F_{scoring}} & \text{if } |F_a(i) - F_b(j)| < F_{scoring} \\ -\frac{|F_a(i) - F_b(j)|}{F_{scoring}} & \text{otherwise} \end{cases} \tag{3}$$

where $F_a(i)$ and $F_b(j)$ are the forces in points i and j in traces a and b, and $F_{scoring} = 4\sigma_{NOISE}$. The difference between this scoring function and the one used by Marsico *et al.* is that in the latter the force difference was divided by the average of the maximum force values ΔF_{max} in the two traces, and not by a single value $F_{scoring}$, equal for all the traces. If this choice is made, when two widely different traces both have high ΔF_{max}, their difference may be weighted less than the difference between two similar traces with low ΔF_{max}. In other words, the distance magnitude depends on ΔF_{max}, leading to low distance values for traces with high ΔF_{max}. Note that in Marsico *et al.* this problem was absent, since the ΔF_{max} values were approximately uniform for all traces in the dataset. We are using the same gap penalties δ_1 and δ_2 like Marsico *et al.*.

In order to make the method computationally more efficient, we compute the distance only between traces that differ by no more than 2 peaks in their L_c histograms or by no more than 20 % in terms of their trace length difference.

2.5 Density peak clustering.

To group the traces in clusters, we use the density peak clustering (DPC) algorithm (Rodriguez and Laio, 2014). In the datasets we are analyzing meaningful clusters correspond to small subsets of the traces, while most of the traces are statistically isolated events. In such a situation, partitioning algorithms like K-means are not fully appropriate because they classify in clusters all the traces, including isolated ones. DPC is suitable because it excludes automatically the outliers, which by definition do not belong to a density peak. The algorithm can be summarized in the following steps:

1. First, one computes the densities, representing the density of data points in the neighborhood of each point. The densities are computed with the k-nearest neighbor (k-NN) density estimator (Altman, 1992), as the ratio between k and the volume occupied by the k nearest neighbors:
$$\tilde{\rho}_i = \frac{k}{\omega_d r_{k,i}^d} \tag{4}$$
where d is the intrinsic dimension (ID) of the dataset (Facco *et al.*, 2017), ω_d is the volume of the d-sphere with unitary radius and $r_{k,i}$

is the distance of point i from its k-th nearest neighbor. In DPC, the cluster membership of a data point is determined uniquely by the rank of its density, and not by the exact value of the density. Therefore, without loss of generality, instead of estimating the density by Equation (4), we estimate it as: $\rho_i = -\log r_{k,i}$. It is easy to verify that the rank of ρ_i is equal to the rank of $\tilde{\rho}_i$, as the two are related by a simple monotonic transformation. Using this definition of the density we are not required to estimate the intrinsic dimension of the dataset. In addition, we multiply ρ_i by the score-length ratio of trace i. By doing so we assign bigger weight to the traces which satisfy better the WLC model.

2. One then finds the minimum distance between point i and any other point with higher density, $\delta_i = min_{j:\rho_j>\rho_i}(d_{ij})$, where d_{ij} is the distance between points i and j. This will be used to idenitify local maxima of ρ_i.
3. Next, one finds the cluster centers, identified as density peaks, namely points with high values of ρ_i and δ_i. To identify the peaks, following Rodriguez and Laio, we compute for each point $\gamma_i = \rho_i \delta_i$. Points with high values of γ_i are good cluster center candidates. One then sorts in descending order all points by the value of γ_i. The first point is a cluster center. The second point is also a cluster center, unless it is at a distance smaller than r_{cut} from the first center, where r_{cut} is a free parameter of the approach (see below). One then considers the third point, which is a cluster center, unless it is at a distance smaller than r_{cut} from the two points with higher γ. This test is then performed for all the points, finding in this manner all the cluster centers, which, by construction, will be further than r_{cut} from all the points with a higher γ.
4. Subsequently, all points that are not centers are assigned to the same cluster of the nearest point with higher density (Rodriguez and Laio, 2014).

In the standard implementation of DPC the distance between a cluster member and the cluster center can be arbitrarily large, if the density peak has an elongated shape. This is not appropriate for the analysis of SMFS traces, where the similarity between all the traces belonging to a cluster is essential. We therefore assume all the clusters to be spherical, and consider meaningful the assignation to a cluster of a trace only if its distance from the cluster center is smaller than r_{cut} and smaller than its distance to any other cluster center. The traces satisfying this condition will be called *core traces*. To determine an appropriate value for the parameter r_{cut} we performed a careful visual inspection on sets of traces characterized by an increasing distance from a high-quality trace. We concluded that at distances larger than 0.3 we can no longer be confident that two traces are likely to correspond to the same protein. We therefore fix $r_{cut} = 0.3$ and determine the size of the cluster core accordingly. In the following, to simplify the description of the results, we discard from the analysis clusters whose core has less than 10 members.

In Table S1 the values of all parameters used in this approach are listed. Furthermore, we explored the effects of changing the values of all relevant parameters. We report the results of these tests in Supplementary Information.

3 Results

3.1 Benchmark on the dataset 'Oocyte'

Dataset Oocyte contains 4,128 traces, 101 of which were selected based on their contour length and visual inspection and attributed to the unfolding of the membrane protein CNGA1 (Maity *et al.*, 2015) thanks to molecular tags (see Materials and Methods for details). After filtering the traces with our procedure, their number was reduced to 440, which is approximately 11% of the total amount of traces. 52% of the previously selected CNG traces passed the filters. The traces were divided into 4 clusters. All selected CNG traces were found in cluster 1 and therefore, cluster 1 is the CNG cluster. With the data that are available, we cannot relate the remaining clusters to proteins or further investigate their molecular origin.

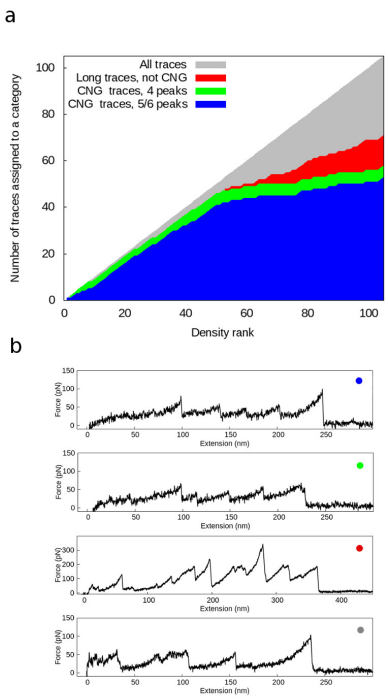

Fig. 2. Graphical representation of the CNG cluster content. a. The cluster members are ranked by density in descending order. The blue area shows the manually selected CNG traces with 5 or 6 force peaks; the green area - manually selected CNG traces with 4 peaks; the red area - traces with contour length greater than 350 nm; the gray area corresponds to all traces assigned to the cluster. b. Representative traces for the different groups in panel a. The painted dot in the top right corner indicates the group affiliation.

We then analyzed more in detail the content of the CNG cluster. In Figure 2a we plot the cluster members ranked by their density in a descending order. The highest density traces are the CNG traces with 5 to 6 peaks (the blue area) followed by the CNG traces with 4 peaks (the green area). In Figure 2b we represent each group with a single F-x curve. When we looked more closely to the highest density region in Figure 2a, we noticed a gray area representing high density traces that haven't been included in the selection. We looked at these traces and found out that they are very similar to the cluster center of the CNG cluster (Figure S5). Therefore, these traces can be considered CNG traces which escaped manual selection. Remarkably, our procedure was able to detect previously unknown CNG traces and group them together in the right cluster. These results demonstrate that our method can discriminate a set of meaningful traces, corresponding to the unfolding of a protein, from a set of traces containing noise.

3.2 Benchmark on the dataset 'Mixed'

Dataset Mixed contains four groups of selected F-x curves, representing the unfolding of four different proteins (see Materials and Methods for details). Each group is depicted in Figure S6a. Since all the traces of this dataset are manually selected, 83% of the traces passed the filters. In Figure S6b we show the distribution of the traces in the two-dimensional space of two representative descriptors, commonly used to discriminate among F-x curves: the maximum contour length (approximately the total length of the curve from the contact to the detachment of the polymer) and the average force of the peaks. This representation does not allow

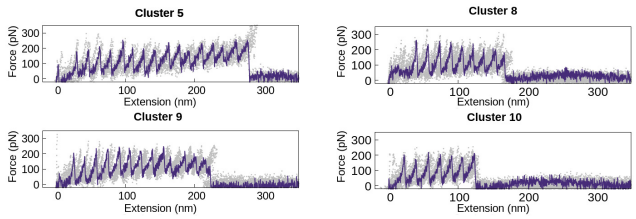

Fig. 3. Most representative cluster centers (blue lines) and closest 10 neighbours (grey points) of the dataset 'GB1'.

distinguishing the four groups: indeed, the traces belonging to groups 2, 3 and 4 occupy approximately the same region.

Figure S6c shows that with our procedure the vast majority of traces belonging to each group was correctly assigned to a separate cluster. Our approach proved capable of clustering together three out of the four groups which are not discriminated by standard descriptors. More in detail, the total number of clusters we obtained was 5. Cluster 1 contains only traces representing the unfolding of the tandem globular polyprotein Alpha3D+4xNUG2 (group 2). Cluster 2 contains only traces representing the unfolding of CNG (group 1). Group 3 (see Materials and Methods) is split in three clusters: 3, 4 and 5, one of which contains 50% of the traces. Visual inspection of the traces belonging to these three different clusters reveal some marginal differences, which justify their assignation do different groups. The traces belonging to group 4 do not belong to the core of any of these clusters, indicating that they are not similar enough, according to the criteria implemented in the code. Indeed, if the r_{cut} parameter value is changed from 0.3 to 0.4, all the four groups of traces are found, and assigned to a single cluster in a one-to-one correspondence. Indeed, with a larger r_{cut} the three clusters in which group 3 was split are merged in a single one, and the traces belonging to group 4 are similar enough to form their own cluster.

For this dataset we compared the performance of the method with spectral clustering (Von Luxburg, 2007) using the Scikit-learn implementation (Pedregosa *et al.*, 2011). The number of clusters in this approach is chosen by visual inspection of the eigenvalue spectrum, which, in the case of the Mixed dataset, has a first gap after the third eigenvalue, and another gap after the fifth eigenvalue. Accordingly, we performed the analysis retaining three and five eigenvectors. We quantified the consistency of a clustering partition c and the ground truth partition gt by estimating two different mutual information measures, $\text{NMI}_{gt} = \frac{H_c + H_{gt} - H_{c,gt}}{H_{gt}}$, and $\text{NMI}_{symm} = \frac{2(H_c + H_{gt} - H_{c,gt})}{H_c + H_{gt}}$, where H_c and H_{gt} are the entropy of the clustering and of the ground truth partition and $H_{c,gt}$ is the cross entropy. NMI_{symm} is a symmetric measure of consistency, and is equal to 1 only if two partitions are fully equivalent, except for a permutation of the labels. NMI_{gt} is a measure of consistency in which the ground truth category prevails, and is equal to 1 also if a ground truth cluster is split in two or more clusters. For spectral clustering, we find NMI_{symm}=0.77 (using three eigenvectors) and 0.84 (using five eigenvectors). NMI_{symm} is equal to 0.79 for our approach. NMI_{gt} for spectral clustering is 0.69 (three eigenvectors) and 0.87 (five eigenvectors), indicating that the clusters contain traces with different ground truth classification. Instead, with our approach we find NMI_{gt}=1, indicating that the clusters found for this dataset are pure, namely they contain only traces with the same ground truth classification. These results demonstrate that our algorithm is able to distinguish different unfolding patterns arising in the same dataset and to properly assign the corresponding F-x curves to different clusters without knowing *a priori* the protein composition. Remarkably, at variance with other methods, the number of clusters is determined automatically.

3.3 Analysis of the dataset 'GB1'

Dataset GB1 consists of 48,769 F-x curves generated in 20 SMFS experiments of a polyprotein tandem construct of GB1 domains (Aioanei *et al.*, 2011) (see Materials and Methods). The tandems of 8 GB1 domains were designed with a cysteine in the N-terminus so that the two tandems could dimerize in a single 16 domain protein. The expected unfolding patterns should have 16 unfolding peaks spaced by 19 nm plus the detachment peak. However, not all the proteins are expected to dimerize, therefore a significant amount of traces with 8 peaks plus the detachment peak should be present in the dataset. Moreover, as opposed to membrane proteins—which are used to generate unfolding patterns of the same length of the stretched amino-acid chain—tandem globular proteins may be hooked at a random spot along the chain, resulting in shorter trace recordings. In summary, the dataset should contain mostly 17 and 9 peaks traces plus a fraction of shorter traces.

After filtering the initial 48,769 F-x curves, only 2,701 traces survived ($\sim 6\%$). More than 90% of the F-x curves in the GB1 dataset showed no binding events. We found 24 clusters. In Figure 3 and Figure S7 we show all clusters which contain more than 10 core traces. Clusters 5, 8, 9, and 10 in Figure 3 are the clusters having the highest quality scores (e.g. showing the best agreement with the WLC model). These traces clearly bear the prototypical patterns with 19 nm-spaced peaks—the length of the GB1 domain. Clusters 5 and 8 represent the full unfolding of a dimeric and a monomeric tandem respectively. The other clusters, shown in the Supplementary Information host much shorter traces, likely corresponding to (rare) events in which the AFM tip hooks at a random position, rather than at the N-terminus. The results in this dataset reveal the performance of our clustering method on F-x curves with different number of peaks. Namely, it tends to separate them in different clusters while preserving each cluster's homegeniety.

3.4 Analysis of the dataset 'Rod'

Dataset Rod, together with dataset GB1, motivates our work since it poses a challenge to the traditional methods for AFM-SMFS data analysis. The analysis of this dataset can be considered blind, since the traces are collected in experiments performed in the plasma membrane of the rod outer segment (rod OS) under native conditions, and no preprocessing or selection was performed on the traces before our analysis. It contains 386,912 F-x curves. For this specific data set we restrict the analysis to traces with a quality score $W > 0.65$ and with at least three peaks in the L_c histogram. After filtering and score-based trace selection, the total number of traces is reduced to 7,843, $\sim 2\%$ of the initial amount of data (also in this case the majority of the traces shows no binding event).

With our approach we found 27 clusters. Six of them are shown in Figure 4b. In Figure S8 the cluster centers of the other clusters with more than 10 members are shown. In order to visualize the distribution of the most abundant traces in the relevant clusters, we used also here the maximum contour length and the mean peaks' force as descriptors (Figure 4a). In dataset 'Rod', by using these descriptors it is impossible to discriminate the different clusters because they form an almost continuous distribution.

Given that two of the dominant proteins in the plasma membrane of the rod OS are rhodopsin and the CNG channel, one might expect to find a rhodopsin cluster and a CNG cluster. The contour length of rhodopsin with the intact S-S bond is $\sim$ 95 nm and $\sim$ 120 nm when fully stretched (Sapra *et al.*, 2006) with three–four major peaks equally spaced by $\sim$ 20 nm. With our procedure we found that cluster 1—the most abundant one (Figure 4b), —might correspond to the unfolding of rhodopsin, although in this case the molecule must have been pulled from the opposite terminus compared to the work of Sapra *et al.*. The wild type CNGA1 channel of the Xenopus has a slightly different sequence than the bovine one overexpressed in the

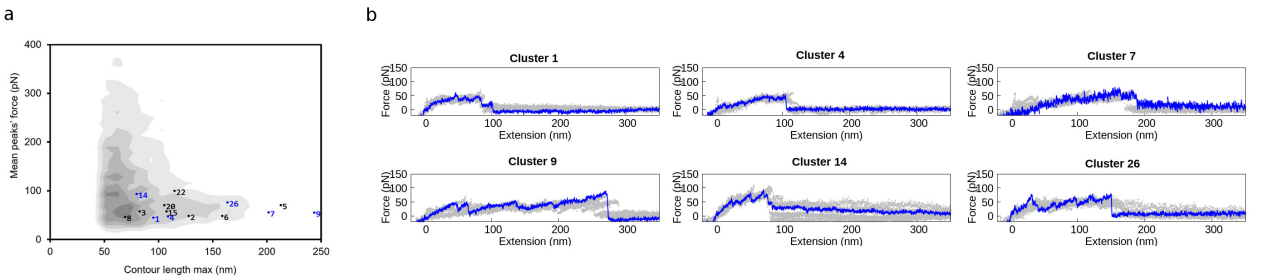

Fig. 4. Results of the cluster analysis in dataset Rod. a. Isoline density plot of the 7,843 traces that survived the quality filter (each trace was codified as one single point in this two-dimensional representation, the grey levels indicate the region of this space with higher density of points). The blue points indicate the position of the representative trace shown in panel b while the black points represent the traces in Figure S8. b. Representative trace (solid lines) of 6 clusters with other 5 traces belonging to the same cluster core (grey points) of Rod Dataset.

oocytes (Maity *et al.*, 2015). We expect the contour length of the fully-stretched CNG channel in the rod OS to be around 260 nm. The most likely candidate for the CNG cluster is cluster 9 which includes traces with L_c values up to 250 nm (Figure 4b) and with a periodicity of the peaks similar to the one described by Maity *et al.*.

4 Discussion

There is an increasing need for the development of automatic analysis tools for large scale AFM experiments (Müller *et al.*, 2019; Minelli *et al.*, 2017). Our approach is designed to face a specific challenge: analyzing the huge amount of data obtained by AFM-SMFS experiments in highly heterogeneous samples, for example sets of traces harvested in native membrane patches. Our method does not require any previous knowledge on the sample composition and the proteins contour length. In previous approaches (Kuhn *et al.*, 2005; Marsico *et al.*, 2007; Bosshart *et al.*, 2012; Galvanetto *et al.*, 2018), the most significant filtering step is based on the expected contour length of the protein under investigation. Such an approach requires knowledge of the sample composition. It reduces tremendously the number of analyzed traces but is not suitable for data obtained from native cell membranes.

An important step in our procedure is trace selection based on a quality score which measures the consistency of each trace with the WLC model. By filtering traces according to this criterion one can miss some meaningful patterns that deviate from the WLC. If a model characterizing the force-extension features of these patterns was available, one could retain also traces consistent to that model. Dedicated experiments in controlled conditions like the ones performed by Takahashi *et al.* may offer a route to study these deviations. We also remark that by using a bin width of 8 nm for computing the L_c histograms, we are allowing for significant deviations of the persistence length (Sarkar *et al.*, 2005), such that anomalous patterns may still survive the filtering. In other words, traces not perfectly adhering to the WLC model but still exhibiting the sawtooth pattern with well-defined peaks will survive the filter.

Following Marsico *et al.*, we use dynamic programming alignment to measure how similar to each other two traces are. In order to group similar traces into clusters, we use density-peak clustering (Rodriguez and Laio, 2014). The major advantages of this approach are that it does not require knowledge on the number of clusters in advance, and is able to distinguish "density peaks" formed by sets of similar traces from the background noise, formed by traces associated with isolated unfolding events.

We benchmarked our method on a data set containing a manually selected sample of CNG traces and $\sim$ 40 times more unevaluated traces. Our method successfully detected the CNG traces and grouped them in a separate cluster. Furthermore, the method proved to be able to distinguish between groups of traces corresponding to the unfolding of four different proteins, and to automatically find the different patterns of GB1s among $\sim$ 50,000 traces generated in a real-world experiment. Finally, we analyzed a dataset containing $\sim$ 400,000 traces of unidentified molecular origin from experiments in the plasma membrane of the rod outer segment. Our program turned out to be efficient taking only $\sim$ 30 minutes to process the entire data set revealing several unknown unfolding patterns calling for further molecular identification.

It is important to underline that the method is mainly aimed at finding statistically meaningful sets of similar traces which are likely to correspond to the unfolding of the same protein. After a meaningful pattern has been found, it is useful to adopt more conventional methods based on fingerprinting (like in Fodis, (Galvanetto *et al.*, 2018)) to enrich the clusters with other traces, that can be initially discarded due to the filtering procedure.

This algorithm has still two parameters that affect in different ways the overall results: bin size of L_c and r_{cut}. The bin size of L_c should be bigger than the noise of the instrument, and smaller than the expected feature of the traces. If it is chosen in this range, it affects only slightly the size of the clusters but it does not affect the identification of their centers as discussed in Supplementary Information. On the other hand, r_{cut} plays a role mostly in the 'front-end' side of the algorithm: it works as a threshold that allows the user to decide how precisely the clustering should operate depending on the quality of the data and the scope of the analysis. Higher r_{cut} (e.g. > 0.5) will generate less clusters but more populated, suitable when the data are very noisy (e.g. with native samples). Lower r_{cut} (e.g. 0.3) will still find the bigger clusters found with high r_{cut}, but it will also generate more smaller clusters (which might be suitable for experiments with purified proteins).

We should also underline that the method is not designed to distinguish different unfolding pathways of the same protein. The filtering and the clustering procedure are by far too coarse for this scope. After the clusters have been found, one can further investigate them by one of the approaches in refs. (Kuhn *et al.*, 2005; Marsico *et al.*, 2007; Bosshart *et al.*, 2012; Galvanetto *et al.*, 2018), which are much more appropriate for this scope.

5 Materials and Methods

5.1 Experimental data

Dataset Oocyte: The first data set contains 101 traces ascribed to the unfolding of the CNGA1 channel and 4,027 other traces generated in the same experiments. CNGA1 channels were expressed in *Xenopus laevis* oocytes with sample preparation, experimental procedure and selection described in Maity *et al.*. SMFS experiments were performed in the oocytes membrane with the AFM (NanoWizard 3, JPK). The cantilever was calibrated before the start of each experiment; its spring constant was $\sim$ 0.08 N/m. The AFM tip was pushed into the surface and a force of 1 nN was applied for 0.5 s to enhance the non-specific binding to the proteins. The tip was retracted from the surface at pulling speed 500 nm/s. The selection of the CNG traces was based on two criteria: the contour length of the curves and their force pattern: according to the interpretation of the

experimental data made by Maity *et al.*, the last peak in the CNG traces has a L_c value larger than 220 nm and all CNG traces share a common unfolding fingerprint. The unfolding fingerprint consists of a peak at L_c around 100 nm corresponding to the unfolding of the cyclic nucleotide-binding (CNB) domain attached to the C-terminus; 3 or 4 force peaks between L_c 120 nm and 250 nm corresponding to the unfolding of the six transmembrane helices and the detachment peak. The 101 CNG traces include traces that satisfy these criteria and some other traces that miss a peak in the middle or the last peak assuming different unfolding pathways as suggested in Maity *et al.*.

Dataset 'Mixed': The second data set contains a mixture of four manually selected groups of F-x curves corresponding to the unfolding of different proteins. Group number 1 includes the 101 manually selected CNGA1 traces included in dataset Oocyte. Group number 2 includes 38 F-x curves representing the unfolding of a tandem globular polyprotein (Alpha3D + 4xNUG2) (Heenan and Perkins, 2018b). These experiments were performed by Marc-Andre LeBlanc and are available on Dryad (Heenan and Perkins, 2018a). Group number 3 includes 131 traces from dataset Rod representing the unfolding of an unknown protein. To build this group we selected a template trace by visual inspection, and used the tool 'Fingerprint ROI' in the software Fodis (Galvanetto *et al.*, 2018) to mark the sawtooth pattern of the template, and find traces similar to that template. Finally, group number 4 includes 43 traces from dataset Rod representing the unfolding of another unknown protein found following the same protocol used for group number 3.

Dataset 'GB1': The data set consists of 48,769 F-x curves generated in 20 SMFS experiments performed on a (GB1) 8 synthetic tandem polyprotein terminated with a cysteine residue as described in ref. (Aioanei *et al.*, 2011). Of the 48,769 F-x curves, approximately only the 10% shows binding events, the 90% are flat curves. The extended details of the protein engineering and purification can be found in the Supplementary Information of ref. (Aioanei *et al.*, 2011). Briefly, (GB1)8 gene was obtained by iterative cloning of the sequence on the basis of the identity of the sticky ends generated by BamHI and BglII restriction enzymes. Then, (GB1)8 polyprotein was overexpressed in DH5 strain and purified by Ni2+-affinity chromatography. The purified polyprotein sample was at a final concentration of 0.34 mg/ml and was stored at -20 °C in PBS buffer with 0.2% (v/v) sodium azide. The experiments were performed with a Bruker Multimode on a template-stripped gold substrate. All the measurements were carried out in Tris/HCl (10 mM, pH 7.5) buffer.

Dataset 'Rod': The fourth data set comes from pulling SMFS experiments performed in the native plasma membrane of the rod outer segment (OS) of *Xenopus laevis* retinas. A detailed experimental protocol is provided in ref. (Maity *et al.*, 2017). As a consequence the dataset is highly heterogeneous and poses a challenge to traditional analysis tools. The entire dataset contains 386,912 F-x curves. Briefly, the AFM (NanoWizard 3, JPK) was used with cantilever with spring constant 0.08 N/m. The cantilever was calibrated before each experiment. The AFM tip was pushed into the sample surface with 1 nN force and held for 0.5 s. It was then retracted at constant speed 500 nm/s.

References

Aioanei, D. *et al.* (2011). Single-molecule-level evidence for the osmophobic effect. *Angewandte Chemie International Edition*, **50**(19), 4394–4397.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, **46**(3), 175–185.

Bosshart, P. D. *et al.* (2012). Reference-free alignment and sorting of single-molecule force spectroscopy data. *Biophysical journal*, **102**(9), 2202–2211.

Engel, A. and Gaub, H. E. (2008). Structure and mechanics of membrane proteins. *Annu. Rev. Biochem.*, **77**, 127–148.

Facco, E. *et al.* (2017). Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, **7**(1), 12140.

Galvanetto, N. (2018). Single-cell unroofing: probing topology and nanomechanics of native membranes. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, **1860**(12), 2532–2538.

Galvanetto, N. *et al.* (2018). Fodis: software for protein unfolding analysis. *Biophysical journal*, **114**(6), 1264–1266.

Heenan, P. R. and Perkins, T. T. (2018a). Data from: Feather: automated analysis of force spectroscopy unbinding and unfolding data via a bayesian algorithm. `https://datadryad.org/resource/doi:10.5061/dryad.1615c2p`. Accessed: 2010-09-30.

Heenan, P. R. and Perkins, T. T. (2018b). Feather: Automated analysis of force spectroscopy unbinding and unfolding data via a bayesian algorithm. *Biophysical journal*, **115**(5), 757–762.

Kuhn, M. *et al.* (2005). Automated alignment and pattern recognition of single-molecule force spectroscopy data. *Journal of microscopy*, **218**(2), 125–132.

Lehninger, A. L. (2000). *Lehninger principles of biochemistry*. Worth Publishers: New York.

Maity, S. *et al.* (2015). Conformational rearrangements in the transmembrane domain of cnga1 channels revealed by single-molecule force spectroscopy. *Nature communications*, **6**, 7093.

Maity, S. *et al.* (2017). New views on phototransduction from atomic force microscopy and single molecule force spectroscopy on native rods. *Scientific reports*, **7**(1), 12000.

Marsico, A. *et al.* (2007). A novel pattern recognition algorithm to classify membrane protein unfolding pathways with high-throughput single-molecule force spectroscopy. *Bioinformatics*, **23**(2), e231–e236.

Minelli, E. *et al.* (2017). A fully-automated neural network analysis of afm force-distance curves for cancer tissue diagnosis. *Applied Physics Letters*, **111**(14), 143701.

Müller, P. *et al.* (2019). nanite: using machine learning to assess the quality of atomic force microscopy-enabled nano-indentation data. *BMC bioinformatics*, **20**(1), 1–9.

Oesterhelt, F. *et al.* (2000). Unfolding pathways of individual bacteriorhodopsins. *Science*, **288**(5463), 143–146.

Pedregosa, F. *et al.* (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Rief, M. *et al.* (1997). Reversible unfolding of individual titin immunoglobulin domains by afm. *Science*, **276**(5315), 1109–1112.

Rodriguez, A. and Laio, A. (2014). Clustering by fast search and find of density peaks. *Science*, **344**(6191), 1492–1496.

Sapra, K. T. *et al.* (2006). Detecting molecular interactions that stabilize native bovine rhodopsin. *Journal of molecular biology*, **358**(1), 255–269.

Sarkar, A. *et al.* (2005). The elasticity of individual titin pevk exons measured by single molecule atomic force microscopy. *Journal of Biological Chemistry*, **280**(8), 6261–6264.

Takahashi, H. *et al.* (2018). α-helix unwinding as force buffer in spectrins. *ACS nano*, **12**(3), 2719–2727.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, **17**(4), 395–416.