BIOINFORMATICS

Vol. 23 ECCB 2006, pages e231–e236
doi:10.1093/bioinformatics/btl293

A novel pattern recognition algorithm to classify membrane protein unfolding pathways with high-throughput single-molecule force spectroscopy

Annalisa Marsico*, Dirk Labudde, Tanuj Sapra, Daniel J. Muller and Michael Schroeder

Biotec, TU Dresden, Germany

ABSTRACT

Motivation: Misfolding of membrane proteins plays an important role in many human diseases such as *retinitis pigmentosa*, hereditary deafness and *diabetes insipidus*. Little is known about membrane proteins as there are only very few high-resolution structures. Single-molecule force spectroscopy is a novel technique, which measures the force necessary to pull a protein out of a membrane. Such force curves contain valuable information on the protein structure, conformation, and inter- and intra-molecular forces. High-throughput force spectroscopy experiments generate hundreds of force curves including spurious ones and good curves, which correspond to different unfolding pathways. Manual analysis of these data is a bottleneck and source of inconsistent and subjective annotation.

Results: We propose a novel algorithm for the identification of spurious curves and curves representing different unfolding pathways. Our algorithm proceeds in three stages: first, we reduce noise in the curves by applying dimension reduction; second, we align the curves with dynamic programming and compute pairwise distances and third, we cluster the curves based on these distances. We apply our method to a hand-curated dataset of 135 force curves of bacteriorhodopsin mutant P50A. Our algorithm achieves a success rate of 81% distinguishing spurious from good curves and a success rate of 76% classifying unfolding pathways. As a result, we discuss five different unfolding pathways of bacteriorhodopsin including three main unfolding events and several minor ones. Finally, we link folding barriers to the degree of conservation of residues. Overall, the algorithm tackles the force spectroscopy bottleneck and leads to more consistent and reproducible results paving the way for high-throughput analysis of structural features of membrane proteins.

Contact: annalisa.marsico@biotec.tu-dresden.de

1 INTRODUCTION

Integral membrane proteins play essential roles in cellular processes, including photosynthesis, transport of ions and small molecules, maintenance of osmotic balance, cell–cell adhesion, signal transduction and light harvesting. They account for ~20–30% of the open reading frames of a typical genome. Despite the central importance of transmembrane proteins, the number of high-resolution structures remains small due to the practical difficulties in crystallizing them (Bowie, 2005). Many human disease-linked point mutations occur in transmembrane proteins: human rhodopsin and its mutants causing *retinitis pigmentosa* (Rader *et al*., 2004; Sanders and Myers, 2004), human aquaporin and its mutants causing diabetes (Tamarappoo *et al*., 1999). These mutations cause structural instabilities in a transmembrane protein, leading it to unfold or to fold in an alternative conformation (Filipek *et al*., 2003; Mirzadegan *et al*., 2003). Protein folding is described by multidimensional energy landscapes or folding funnels and this is the result of complex inter- and intra-molecular interactions (Onuchic and Wolynes, 2004). Atomic Force Microscopy (AFM) is mostly known for its imaging capabilities, but it also provides a novel tool for detecting and locating forces on a single-molecule level, like the inter- and intra-molecular interactions that stabilize protein structures, forces mediating receptor–ligand bonds or controlling antibody–antigen binding (Janshoff *et al*., 2000; Kedrov *et al*., 2005). Single-molecule force spectroscopy experiments allow measuring the stability of membrane proteins and also probing the energy landscapes (Janovjak *et al*., 2004). In Figure 1A we show a schematic representation of the force spectroscopy instrumentation. Molecules with complex three-dimensional (3D) structures, such as proteins, can be unfolded in a controlled way. Titin and bacteriorhodopsin are examples of proteins that have been intensively studied (Rief *et al*., 1997; Oesterhelt *et al*., 2000; Janovjak *et al*., 2004; Sapra *et al*., 2006). When transmembrane proteins are unfolded in force spectroscopy experiments, during continuous stretching of the molecule, the applied forces are measured by the deflection of the cantilever and plotted against extension (tip-sample separation), yielding a characteristic force–distance curve for the specific molecule under investigation (Fig. 1). The force–distance curve is the result of subsequent events of molecular interactions (Zhuang and Rief, 2003; Oesterhelt *et al*., 2000). From the analysis of single-molecule force spectra it is possible to associate the peaks to single potential barriers stabilizing segments within membrane proteins. These segments can be represented by transmembrane helices, polypeptide loops or fragments and are established by collective interactions of several amino acids (Kessler *et al*., 2005). For a given molecule under study, the force–distance curves exhibit certain patterns, which contain information about strength and location of molecular forces established inside the molecule, about stable intermediates and interaction pathways, and the probability with which they occur. In contrast to soluble proteins, investigated by single-molecule force spectroscopy, the folded part of a membrane protein is anchored within the membrane and the sequence of the unfolding peaks follows the amino acid sequence of the protein (Muller *et al*., 2002). For each peak the number of already unfolded amino acids can be determined from the length of the unfolded part of the polypeptide chain, obtained from a fit to a hyperbolic function, the worm-like chain model (WAC), of the given peak (Rief *et al*., 1997). Consequently, with the peaks and the predicted secondary structure, it is possible

*To whom correspondence should be addressed.

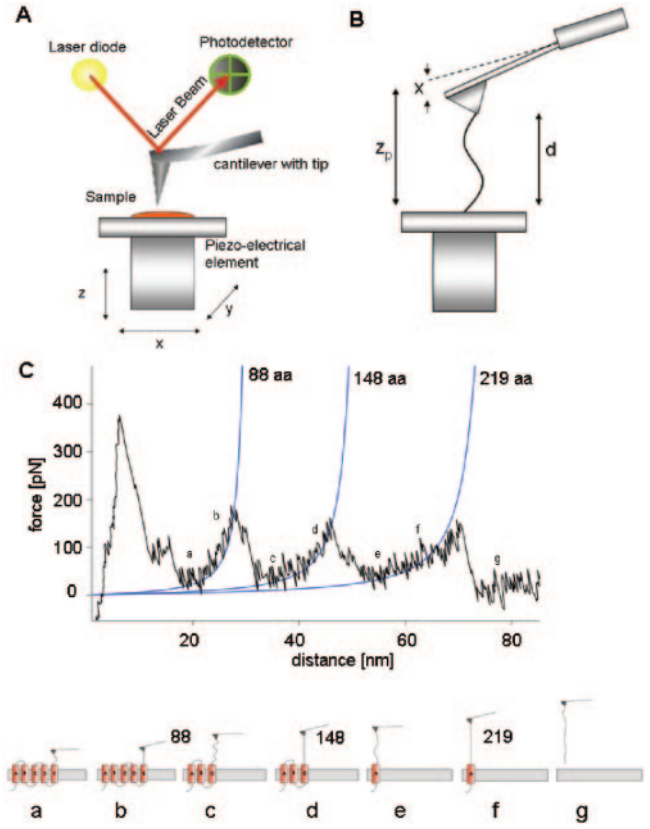

Fig. 1. (**A**) Schematic representation of atomic force microscopy. The sample is mounted on a piezo-electrical element and scanned under a sharp tip attached to the cantilever. The voltage difference of the photodetector is proportional to the deflection of the cantilever. (**B**) Unfolding of a transmembrane protein. A single molecule is kept between the tip and the sample while the tip-sample separation is continuously increased. (**C**) Typical spectrum obtained from an experiment of unfolding of bacteriorhodopsin with the main peaks fitted by a hyperbolic function (worm-like chain model) and correlated to the unfolding of secondary structure elements.

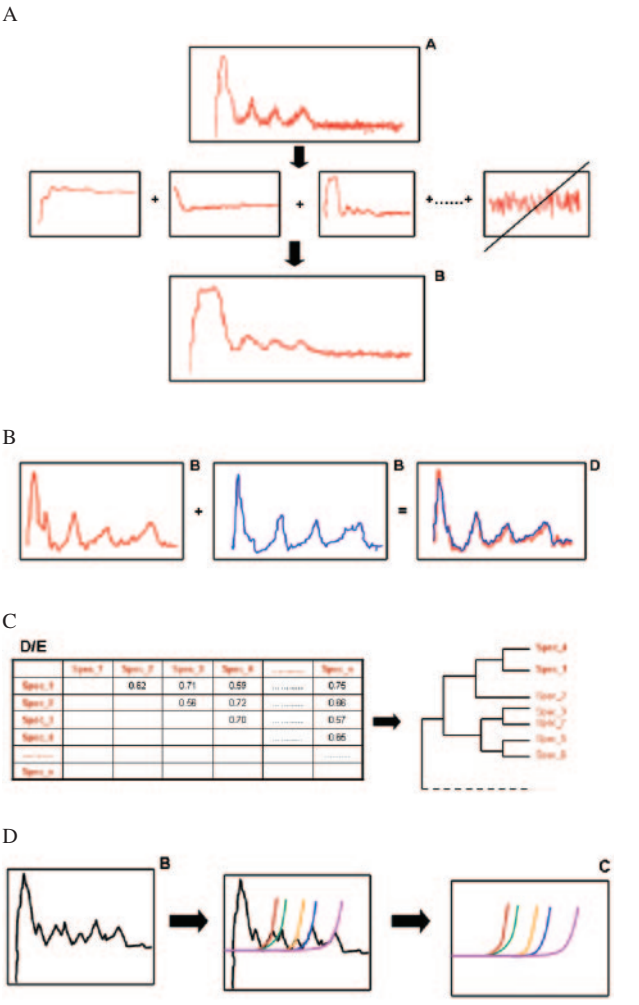

Fig. 2. Algorithm for noise reduction, alignment and clustering of force curves. (**A**) Noise reduction: Dimension reduction with singular value decomposition is applied. (**B**) Curve alignment: The two curves are aligned with dynamic programming. (**C**) Clustering: The pairwise distances obtained with curve alignment are clustered with hierarchical clustering and average linkage. (**D**) Peak detection: Peaks are detected with the worm-like chain model for evaluation of the clustering.

to associate the peaks with structural domains (Fig. 1C) (Muller *et al*., 2002).

To obtain statistically relevant results, many single molecules of the same species need to be studied and a large number of force curves collected and analysed. To discriminate force curves showing specific and unspecific interactions and different unfolding pathways, classification and pattern recognition algorithms have to be applied, as the manual annotation is slow and subject to human mistakes. Even if some off-line software packages have been developed to analyse single-molecule force spectroscopy data (Kasas *et al*., 2000; Kuhn *et al*., 2005), there is an increasing demand for data analysis techniques and suitable pattern recognition algorithms that offer fully automated processing of force spectroscopy datasets on the basis of scientific criteria. Here, we develop an algorithm for high-throughput classification and statistical analysis of force spectra and apply it to a benchmark dataset of 135 force curves obtained for a bacteriorhodopsin mutant P50A. We interpret the results and discuss main and minor unfolding events of bacteriorhodopsin.

2 MATERIALS AND METHODS

We briefly describe how the experimental curves are obtained and then we illustrate the method developed to automatically analyse them. To identify different unfolding pathways with single-molecule force spectroscopy we devise a method, which proceeds in three stages as shown in Figure 2: First, noise is removed from the curves; second, two force curves are aligned and their pairwise distance is computed; third, the curves are clustered hierarchically. The resulting clusters correspond to spurious curves and to different unfolding pathways.

2.1 Experimental setup

Bacteriorhodopsin mutant P50A is a kind gift of Prof. James Bowie (UCLA, USA) and is prepared as described (Faham *et al*., 2004). The protein is attached non-specifically to silicon nitride cantilever by applying a contact force of 1 nN between the AFM stylus and the membrane surface. Single-molecule AFM imaging and force spectroscopy is performed as described

(Muller *et al*., 2002; Oesterhelt *et al*., 2000). First, membrane patches are imaged using contact mode AFM (Muller *et al*., 1999). For force measurements, the AFM stylus is approached to the membrane protein surface while applying a constant force of <1 nN. After a contact time of 500 ms −1 s, the stylus is retracted from membrane surface at a constant velocity.

2.2. Data preparation: filtering bad curves and determining zero-force baseline and contact point

Before applying the alignment algorithm a careful pre-processing of experimental data is needed. The quality of measured force curves varies from an unfolding event to the other. Not every force curve contains unfolding signals, some of them exhibit an overall length indicative of partial or multiple unfolding events, others show peaks due to non-specific interactions (corrupted curves) or a slope in the baseline that makes further analysis difficult. The quality of input data is very important for an automated procedure since corrupted data often lead to incorrect results and make it difficult to give a reliable interpretation. We automatically detect these corrupted curves and remove them and, for the subsequent analysis, we filter those curves that satisfy the following criteria:

- presence of at least one peak (point whose force magnitude is higher than twice the standard deviation of the noise in the final part of the spectrum) in the force–distance curve;
- position of the last peak in a suitable distance range, depending on the length of the protein under study. This criterion ensures that all the analysed curves correspond to complete single unfolding events (Muller *et al*., 2002).

According to a classical analysis, a linear fit of the non-contact part of each spectrum allows one to determine the zero-force baseline and evaluate the standard deviation of the noise. The non-contact part of a spectrum corresponds to the final part of the signal, pure noise due to the free motion of the cantilever, which is no more in contact with the membrane surface. The first intersection of the baseline with the spectrum determines the contact point (the point at which the deflected cantilever is in contact with the sample surface) (Kuhn *et al*., 2005).

We also detect the last peak of the curve and remove unspecific noise after this peak. We achieve this by a 5 nm linear fit on the end of the curve, by calculating its standard deviation and by walking forwards until the standard deviation increases by a factor of 1.5.

2.3 Noise reduction: singular value decomposition

In force spectroscopy measurements, the dominant source of noise is the thermal motion of the cantilever, but other sources (like electronic, optical or vibrational noise of the instrument) can affect the sensitivity of the measurements, causing the real relevant peaks in the force–distance curves to be difficult to detect and distinguish from random noise. In a typical force spectroscopy experiment, the standard deviation of the noise in the final part of the spectrum is usually between 10 and 40 pN and strongly depends on the spring constant of the cantilever (Janovjak, 2005). In order to reduce the noise we apply dimension reduction with singular value decomposition (SVD) to the spectra.

Singular value decomposition decomposes a matrix X into two orthogonal matrices U and V and a diagonal matrix S as follows: $X = \mathrm{USV}^{\mathrm{T}}$ (Goldberg, 1992). The diagonal matrix S contains singular values $s_1,s_2,\ldots,s_n$ in decreasing order on its diagonal. By setting $s_k,s_{k+1},\ldots,s_n$ to zero we can reduce the matrix X's dimensionality from n to k and thus reduce noise from the original signal. Figure 2A shows a schematic decomposition of the curves. For our purposes, the matrix X consists of a spectrum per row.

For the 135 force curves of bacteriorhodopsin we remove 15 dimensions with singular value decomposition, which represent 2% of the original signal. The standard deviation is 8 pN afterwards, which is considerably smaller than 14 pN obtained from applying a moving average for noise reduction.

2.4 Curve alignment with dynamic programming

In the second stage (Fig. 2B) of our method the curves are aligned using global sequence alignment with dynamic programming(see e.g. Eddy, 2004):

Given two sequences seq_1 and seq_2
Let $M(i,j)$ be the match score of $seq_1(i)$, $seq_2(j)$
Let g be a negative gap penalty
For $i = 1$ to n: $A[i,0] = i*g$
For j=1 to m: $A[0,j] = j*g$
For $i = 1$ to n
  For $j = 1$ to m
    $A[i,j] = \max(A[i\text{-}1,j]+g,\ A[i,j\text{-}1]+g,\ A[i\text{-}1,j\text{-}1]+M(i,j))$
Output A[n,m] as final score

The key reason for using a sequence alignment technique is the meaningful definition of matches/mismatches, insertions and deletions. Matches and mismatches reward/penalize more or less fitting parts of the force curves. Insertions and deletions are important, as peaks in the curves may vary by up to six residues and as peaks may be missing completely between two curves. Here, the match/mismatch score $M(i,j)$ is defined as follows:

$$M(i,j) = \begin{cases} 1 - \dfrac{|seq_1(i) - seq_2(j)|}{avg(max\_force\_value)} & if\ |seq_1(i) - seq_2(j)| \\ & \leq 2\sigma_{noise} \\ -\dfrac{|seq_1(i) - seq_2(j)|}{avg(max\_force\_value)} & else \end{cases}$$

The scoring function is proportional to the absolute value of the force magnitude difference, normalized by the average of the maximum force values from the two spectra. We define a match between two force values when the absolute value of their difference is not greater than twice the standard deviation of the noise, otherwise we define a mismatch. The standard deviation of the noise strongly depends on the cantilever used in the experiment, so the value of this parameter changes according to the experimental data under study. The gap penalty g used in the global alignment procedure is position-dependent and its value can be changed in the algorithm according to the particular dataset.

In our case the gap penalty g is set to a value of 0.002 for force values in the first 10 nm of the spectra and to a value of 0.8 in rest of the trace. An unspecific attachment of the cantilever to the C-terminus of the protein (24 amino acids long = 10nm) needs to be compensated by a low gap penalty for these first 10 nm. The similarity score sim (seq_1,seq_2) of two curves seq_1 and seq_2 is $A(n,m)$, i.e. the final alignment score.

As an example of our approach consider Figure 3. To achieve the optimal alignment, the algorithm introduces a gap within the first peak. As a result, the rest of the curves matches apart from one missing peak.

2.5 Defining unfolding classes with hierarchical clustering

Finally, we cluster the curves using hierarchical clustering with average linkage (Fig. 2C). We define the distance of two sequences as 1-sim(seq_1,seq_2). Only curve pairs with a Z-score of sim(seq_1,seq_2) better than 0.65 are considered for clustering, as a lower value indicates an outlier, which belongs to the class of spurious curves.

2.6 Peak detection

For the evaluation of our approach peaks have to be detected. The standard approach for fitting force peaks revealed from stretching a polypeptide is the WLC. Consider Figure 2D. Unfolded proteins behave in an approximation like random coils whose elasticity is described by the WLC with a persistence length $l_p = 4$ Å (Rief *et al*., 1997; Muller *et al*., 2002). The gradual, nonlinear increase in the extension traces can be fitted using the

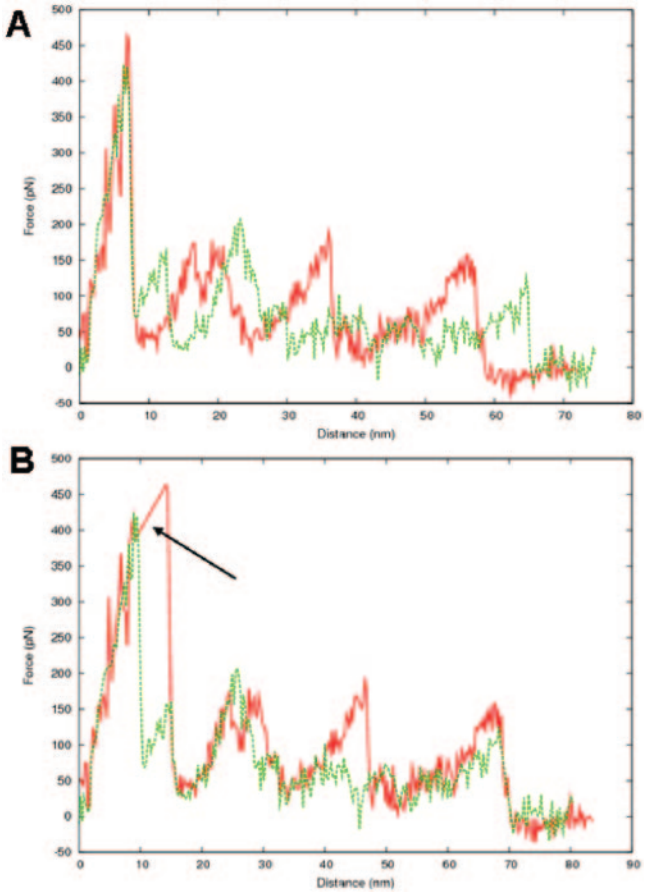

Fig. 3. Aligning two force curves with dynamic programming. (**A**) The curves are not aligned. (**B**) After introducing a gap in the first peak, the overall alignment of peaks is clear with one peak missing.

WLC model with only one free parameter: the contour length L of the stretched portion of the molecule. The equation below describes the increasing slope of the force-distance trace at low forces (few hundreds pN) with good agreement:

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

where k_b is the Boltzman's constant and T is the temperature.

The WLC fit of a peak thus provides the contour length L of the unfolded portion of the protein, that is the position of the corresponding barrier against unfolding. From the knowledge of the attachment point of the protein to the cantilever tip, the position of an unfolding barrier with respect to the amino acid sequence of the backbone can thus be counted backwards from the terminus.

For the evaluation of our algorithm, we use the WLC model and detect peaks of curves manually, so that each curve is represented by a sequence of contour lengths $seq = (L_1, \ldots, L_n)$ for n detected peaks.

3 EVALUATION

3.1 Bacteriorhodopsin

In order to assess the reliability of our method, we test it on a dataset of 135 force curves collected from unfolding experiments of the P50A bacteriorhodopsin mutant. Bacteriorhodopsin is a compact 27 kDA light-driven proton pump in *Halobacterium salinarum*, converting the energy of green light into an electrochemical proton gradient. Bacteriorhodopsin represents one of the most extensively studied membrane proteins. Its structural analysis has revealed the photoactive retinal embedded in seven closely packed transmembrane α-helices, which build a common structural motif along a large class of related G-protein coupled receptors. As bacteriorhodopsin was already intensively studied by single-molecule force spectroscopy experiments, we test our algorithm on a new experimental dataset consisting of force–distance traces from a bacteriorhodopsin mutant where Pro50 in helix B is mutated to Ala. Prolines in the helices of bacteriorhodopsin are studied as they are linked to kinks in the helices.

3.2 Spurious curves and peak detection

According to the manual annotation, there are 61 good curves among the total dataset of 135 curves. Our algorithm identifies spurious curves during the pre-processing step and as outliers with low similarity to other curves during the clustering phase. Our algorithm achieves a success rate of 81% in comparison to the manual annotation in classifying bad and good curves.

Overall, there are three main peaks (at 88, 148 and 219 amino acids) present in all curves and five minor ones [94 (18%), 105 (18%), 158 (34%), 175 (56%) and 232 amino acids (16%)] in the 61 good curves of the manual annotation.

3.3 Unfolding pathways

Our alignment algorithm generates a hierarchical clustering tree where each cluster corresponds to a class of unfolding events. We compare it against a manual annotation of the curves based on a manual peak detection with the WLC described above.

A hierarchical clustering based on curve alignment reflects the similarity of clustered spectra. Clusters in the tree correspond to classes of unfolding events at different levels of granularity. We find that all the spectra in the dataset share three main peaks (88 ± 5, 148 ± 5 and 219 ± 5 amino acids) that indicate the unfolding of two transmembrane helices and their connecting loop in a single step. Furthermore, we can identify subclusters in the hierarchical tree relating to different unfolding events showing the presence of side peaks besides the main unfolding pathway, indicating that helices not always unfold pairwise but exhibit more unfolding intermediates. This result agrees with previous studies that analyse individual unfolding pathways of bacteriorhodopsin (Oesterhelt *et al*., 2000).

Figure 4 shows an example of an unfolding pathway in the bacteriorhodopsin mutant P50A. The figure shows three spectra which are very close to each other in the hierarchical tree of the curve alignment and share the same number of peaks as the manual annotation. This unfolding pathway corresponds to helix E and helix D unfolding pairwise in a single step, then helix C with loop B-C, then helix B unfolds and finally helix A together with the N-terminus.

The full tree with all the unfolding events consists of 12 levels starting from the root (dendrogram not shown). We progressively cut the tree at different levels and, at each step, we identify subclusters that can be possibly associated to different unfolding pathways (Fig. 5). Each table in Figure 5 corresponds to the analysis of subclusters at different levels of cutting, and, for each peak position identified in the manual annotation, we show the percentage of curves sharing a given peak. We discuss also the correlation

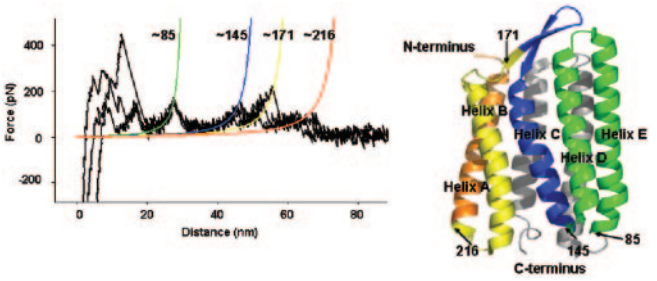

Fig. 4. Three curves clustered by the algorithm share the same number of peaks according to the manual annotation. They all share a side peak at about 171 amino acids. On the right, we show the mapping of the detected unfolding barriers onto the three-dimensional structure (PDB ID 1PXR) of the protein. According to the positions of the detected peaks, helix E and helix D unfold pairwise in a single step, then helix C with loop B-C, then helix B unfolds and finally helix A together with the N-terminus.

between subclusters and different unfolding pathways. We underline in red the peaks that possibly identify a given unfolding pathway.

3.4 Curve alignment versus manual annotation: 76% success rate

The manual annotation consists of the number of manually detected peaks and their positions. We calculate the percentage of peak positions which match within each subcluster of the algorithm's hierarchical clustering tree. This calculation is based on a sum-of-pairs score inside each subcluster, where we do not split the dendrograms at any distance, but we consider subclusters as succession of partitions in the hierarchical tree. On the basis of the manually annotated dataset we achieve a precision of 75.6% for the curve alignment method. This value indicates that the algorithm is consistent with a classification of the spectra in different unfolding pathways.

4 DISCUSSION

Single-molecule force spectroscopy is a convenient and promising tool to measure interaction forces inside and between molecules. But suitable algorithms to process the data still have to be developed. The solution we propose greatly simplifies and accelerates the data processing step in specific force measurements, compared with a manual selection and annotation. With our automated approach, the recognition of unfolding events is no more considered subjective as for manual recognition but has the advantage to be reproducible and quantitative. The hierarchical tree, as output from the described procedure can be helpful in the interpretation of the experimental data, in discriminating different possible unfolding pathways and calculating their probability of occurrence. A similar dynamic programming approach was used successfully in spectral alignment of mass spectrometry data for peptides masses identification, where two spectra are represented as sets of peaks and the problem consists of finding the best similarity (or the minimum edit distance) between them, allowing insertions and deletions of peaks (Pevzner *et al*., 2000). Unfortunately, unlike mass spectra, force spectra do not consist of limited sets of peaks and the problem of peak detection is not easy to solve. To overcome this problem, we presented a solution based on a curve alignment procedure and compared it with manually annotated data. Overall,

sub-cluster 1

Peak-position	% of curves
88 aa	100% = (12/12)
94 aa	33% = (5/12)
105 aa	8% = (1/12)
148 aa	100% = (12/12)
158 aa	33% = (5/12)
175 aa	58% = (7/12)
219 aa	100% = (12/12)
232 aa	0% = (0/12)

sub-cluster 2

Peak-position	% of curves
88 aa	100% = (10/10)
94 aa	30% = (3/10)
105 aa	30% = (3/10)
148 aa	100% = (10/10)
158 aa	20% = (2/10)
175 aa	70% = (7/10)
219 aa	100% = (10/10)
232 aa	70% = (7/10)

sub-cluster 3

Peak-position	% of curves
88 aa	100% = (6/6)
94 aa	67% = (4/6)
105 aa	0% = (0/6)
148 aa	100% = (6/6)
158 aa	20% = (1/6)
175 aa	100% = (6/6)
219 aa	100% = (6/6)
232 aa	20% = (1/6)

sub-cluster 4

Peak-position	% of curves
88 aa	100% = (10/10)
94 aa	10% = (1/10)
105 aa	30% = (3/10)
148 aa	100% = (10/10)
158 aa	70% = (7/10)
175 aa	50% = (5/10)
219 aa	100% = (10/10)
232 aa	10% = (1/10)

sub-cluster 5

Peak-position	% of curves
88 aa	100% = (12/12)
94 aa	0% = (0/12)
105 aa	0% = (0/12)
148 aa	100% = (12/12)
158 aa	9% = (1/12)
175 aa	41% = (5/12)
219 aa	100% = (12/12)
232 aa	0% = (0/12)

Fig. 5. Five different unfolding pathways. Subcluster 1 is obtained by cutting the hierarchical tree before the third level (starting form the root). Of the curves, 58% share a common peak at about 175 amino acids, 33% at 94 and 33% at 158. We find that all the curves showing a peak at 158 amino acids inside subcluster 1 also share the peak at 175 amino acids. We can possibly associate subcluster 1 to the following unfolding pathway: 88, 148, 158, 175, 219 (unfolding of helices E and D in a single step, then helix C, then loop B-C, then helix B and then helix A and *N*-terminus in a single step). Subcluster 2 is obtained by cutting the tree before the sixth level. The most relevant percentages correspond to peak positions 175 amino acids, shared from the 70% of the curves and 232 amino acids (70% of the curves also in this case). We also find that most of the curves (5/7) share a peak at 232 amino acids also share a peak at 175 amino acids. This suggests an association between subcluster 2 and a possible unfolding pathway characterised from the presence of the following peaks: 88, 148, 175, 219, 232 (unfolding of helix E and D in a single step then helix D, then helix C with loop B-C, then helix B and then helix A and then N-terminus alone). Subcluster 3 is obtained by cutting the hierarchical tree before the seventh level. In this case all the curves in the subcluster share a common peak at 175 amino acids (100%) and 67% of the curves share a peak at 94 amino acids. We suggest that this subcluster can be highly associated to the following unfolding pathway: 88, 94, 148, 175, 219 (unfolding of helix E in a two-step process, then helix D, then helix C with the loop B-C, then helix B and then helix A with the N-terminus). Subcluster-4 and subcluster 5 are detected by cutting the tree before the ninth level. Subcluster-4 can be associate to two different unfolding pathways: 88, 148, 158, 219 and 88, 148, 175, 219. Subcluster-5 can be associated to the main unfolding pathway: 88, 148, 219 (pairwise unfolding of helices E and D, then pairwise unfolding of helices C and B and finally helix A and N-terminus in a single step).

our approach leads to good results (success rate of 81% identifying spurious curves and a success rate of 76% classifying unfolding events). The method has also the advantage to be configurable based on requirements e.g. our current match/mismatch function penalizes the absolute difference of force, which does not align

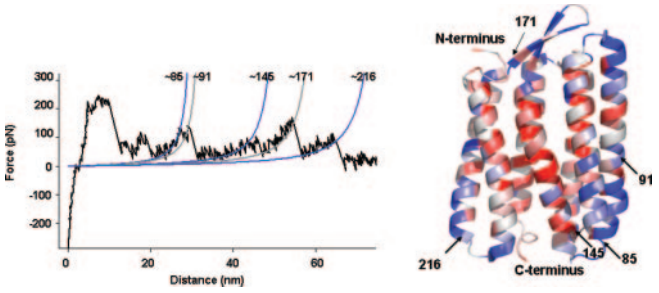

Fig. 6. Example of a spectrum showing the mapping of main peaks (85, 145 and 216 amino acids) and side peaks (91 and 171 amino acids) positions on the structure of the bacteriorhodopsin mutant P50A. The residues in the structure are coloured on the basis of their conservation score in a multiple sequence alignment. The colour scale ranges from red (highly conserved residues) to blue (weakly conserved residues). The positions of peaks in the force curves are located in regions of low conservation. The highly conserved residues fall inside two detected unfolding barriers.

curves of similar behaviour of different absolute force values. To detect these, the scoring function could be adapted to include correlation of the curves.

To summarize, we propose a novel algorithm for high-throughput classification of single-molecule force spectroscopy data. Current, semi-automated approaches based on peak detection with the worm-like chain model are manual and therefore slow and possibly inconsistent. Our approach based on sequence alignment and clustering addresses this bottleneck and leads to reproducible results.

As the next step, we will link unfolding events to sequence and structure features such as conservation, residue–residue contacts and protein topology. As an example, consider Figure 6, which shows three main and two minor unfolding events in combination with the degree of conservation of the residues. Unfolding barriers are highly conserved, while the peaks are associated with low conservation residues. We hope that a bioinformatics approach to high-throughput atomic force microscopy can shed new light onto the structure and function of membrane proteins.

We think that a bioinformatics analysis of mostly conserved residues, residue–residue contacts and protein topology prediction can give a helpful feedback in the interpretation of measured and assigned molecular interactions that determine a specific unfolding pathway in a given molecule.

ACKNOWLEDGEMENTS

Thanks to Andreas Henschel for conservation colouring in PyMol, Christof Winter and Harald Janovjak for interesting discussions, Gihan Dawelbait, Andreas Doms and Thomas Wächter for critical reading of the manuscript. ZHI, TU Dresden's supercomputing facility is kindly acknowledged for computing support and EFRE projects for funding.

Conflict of Interest: none declared.

REFERENCES

Bowie,J.U. (2005) Solving the membrane protein folding problem. *Nature*, **438**, 581–589.

Eddy,S.R. (2004) What is dynamic programming? *Nat. Biotechnol.*, **22**, 909–910.

Faham,S. *et al.* (2004) Side-chain contributions to membrane protein structure and stability. *J. Mol. Biol.*, **335**, 297–305.

Filipek,S. *et al.* (2003) The crystallographic model of rhodopsin and its use in studies of other g protein-coupled receptors. *Annu. Rev. Biophys. Biomol. Struct.*, **32**, 275–397.

Goldberg,J.L. (1992) Matrix theory with applications. McGraw-Hill, pp. 395–401.

Janovjak,H. (2005) Exploring the mechanical stability and visco-elasticity of membrane proteins by single-molecule force measurements. Ph. D. thesis, Technischen Universitat Dresden, Dresden.

Janovjak,H. *et al.* (2004) Probing the energy landscape of the membrane protein bacteriorhodopsin. *Structure*, **12**, 871–879.

Janshoff,H. *et al.* (2000) Force spectroscopy of molecular systems-single molecule spectroscopy of polymers and biomolecules. *Angew. Chem. Int. Ed. Engl.*, **39**, 3212–3237.

Kasas,S. *et al.* (2000) Fuzzy logic algorithm to extract specific interaction forces from atomic force microscopy data. *Rev. Scientific Instruments*, **71**, 2082–2086.

Kedrov,A. *et al.* (2005) Locating ligand binding and activation of a single antiporter. *EMBO Rep.*, **6**, 668–674.

Kessler,M. *et al.* (2005) Bacteriorhodopsin folds into the membrane against an external force. *J. Mol. Biol.*, **357**, 644–654.

Kuhn,M. *et al.* (2005) Automated alignment and pattern recognition of single-molecule force spectroscopy data. *J. Miscrosc.*, **218(Pt2)**, 125–132.

Mirzadegan,T. *et al.* (2003) Sequence analysis of g-protein coupled receptors: similarities to rhodopsin. *Biochemistry*, **42**, 2759–2767.

Muller,D.J. *et al.* (1999) Surface structures of native bacteriorhodopsin depend on the molecular packing arrangement in the membrane. *J. Biol. Chem.*, **274**, 34825–34831.

Muller,D.J. *et al.* (2002) Stability of bacteriorhodopsin alpha-helices and loops analyzed by single-molecule force spectroscopy data. *Biophys. J.*, **83**, 3578–3588.

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

Onuchic,J.N. and Wolynes,P.G. (2004) Theory of protein folding. *Curr. Opin. Struct. Biol.*, **14**, 70–75.

Pevzner,P. *et al.* (2000) Mutation-tolerant protein identification by mass spectrometry. *J. Comput. Biol.*, **7**, 777–787, 70–75.

Rader,A.J. *et al.* (2004) Identification of core amino acids stabilizing rhodopsin. *Proc. Natl Acad. Sci. USA*, **101**, 7246–7251, 70–75.

Rief,M. *et al.* (1997) Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science*, **276**, 1109–1112, 70–75.

Sanders,C.R. and Myers,J.K. (2004) Disease-Related Misassembly of Membrane Proteins. *Annu. Rev. Biophys. Biomol. Struct.*, **33**, 25–51.

Sapra,K.T. *et al.* (2006) Characterizing molecular interactions in different bacteriorhodopsin assemblies by single-molecule force spectroscopy. *J. Mol. Biol.*, **355**, 640–650, 70–75.

Tamarappoo,B.K. *et al.* (1999) Misfolding of mutant aquaporin-2 water channels in nephrogenic diabetes insipidus. *J. Biol. Chem.*, **274**, 34825–34831.

Zhuang,X. and Rief,M. (2003) Misfolding of mutant aquaporin-2 water channels in nephrogenic diabetes insipidus. *Curr. Opin. Struct. Biol.*, **13**, 88–97.