bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted August 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Unfolding and identification of membrane proteins from native cell membranes

Nicola Galvanetto[1,], Sourav Maity[2], Nina Ilieva[1], Alessandro Laio[1,3], Vincent Torre[1]*

[1] International School for Advanced Studies (SISSA), via Bonomea 265, Trieste 34136, Italy. [2]Moleculaire Biofysica, Zernike Instituut, Rijksuniversiteit Groningen, Nijenborgh 4, 9747 AG Groningen, Netherlands. [3]The Abdus Salam International Centre for Theoretical Physics (ICTP), Strada Costiera 11 - 34151 Trieste, Italy.
**email: nicola.galvanetto@sissa.it*

Abstract

Is the mechanical unfolding of proteins just a technological feat applicable only to synthetic preparations or is it applicable to real biological samples? Here, we describe all the steps necessary to deal with native membranes, from the isolation of the membrane of single cells, to the characterization and identification of the embedded membrane proteins. To do so, we combined AFM-based single-molecule force spectroscopy (SMFS) with an automatic pattern classification and the cross-matching of proteomic databases (Uniprot, PDB) to identify the unfolded proteins. We applied this method to four cell types: hippocampal and dorsal root ganglia neurons, rod outer segments and disks, and we were able to classify the unfolding of 5-10% of their total content of membrane proteins. The ability to mechanically probe proteins in their native environment enables the direct mechanical phenotyping of the membrane proteins from different cell types.

Introduction

Much of what we know about the mechanics of cell membranes[1–3] and polymers[4,5] we owe it to atomic force microscopy (AFM) and to its ability to work at the nanoscale. Single-molecule force spectroscopy (SMFS) in particular uses an AFM to apply a force able to unfold directly a single molecule or a protein. The obtained force-distance (F-D) curves encode the unfolding pathway of the molecule, allowing the identification of folded and unfolded regions from the analysis of the sequence of force peaks[8].

SMFS has been mostly used to study the mechanics of purified proteins in solution or reconstituted in a lipid bilayer. However, the information that is possible to extrapolate from the F-D curves (e.g. mechanical stability[9,10], structural heterogeneity[11]) depends on the physical and chemical properties of the cell membrane[12,13], therefore it is desirable to unfold membrane proteins in their original membrane.

The obvious questions are: is the mechanical unfolding of proteins just a technological feat applicable only to synthetic preparations or is it applicable to real biological samples? If this is technically feasible, how can we identify the molecular structure of the unfolded protein among the plethora of native membrane proteins? What additional information can we get?

In the present manuscript we describe a methodology, both experimental and theoretical, to unfold and recognize membrane proteins obtained from native cell membranes (Fig. 1a). Firstly, we developed a technique to extract the membrane from single cells. Secondly, by using AFM-based SMFS we obtained hundreds of thousands of F-d curves in experiments using real biological membranes. Thirdly, we developed a filtering and clustering procedure based on pattern recognition that is able to detect clusters of similar unfolding curves among the thousands of F-d curves. Fourthly, we implemented a Bayesian meta-analysis of mass spectrometry libraries that allowed us to identify the candidate proteins. This Bayesian identification is further refined by cross-analyzing additional databases so to have very few candidates for the obtained clusters of F-d curves. We focused on native membrane proteins from hippocampal neurons, dorsal root ganglia (DRG) neurons, and the plasma and disc membrane of rod outer segments, which represent the only native sample that were approached in the past[14]. We validate the identification using the known unfolding of two proteins from rod OSs: cyclic nucleotide gated (CNG) channels[12] and rhodopsin molecules[14].

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted August 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Besides the identification, the proposed methodology generates as by-product the unfolding signature of a given cell type which could be used for phenotyping cells in screening/biomedical applications.

Results

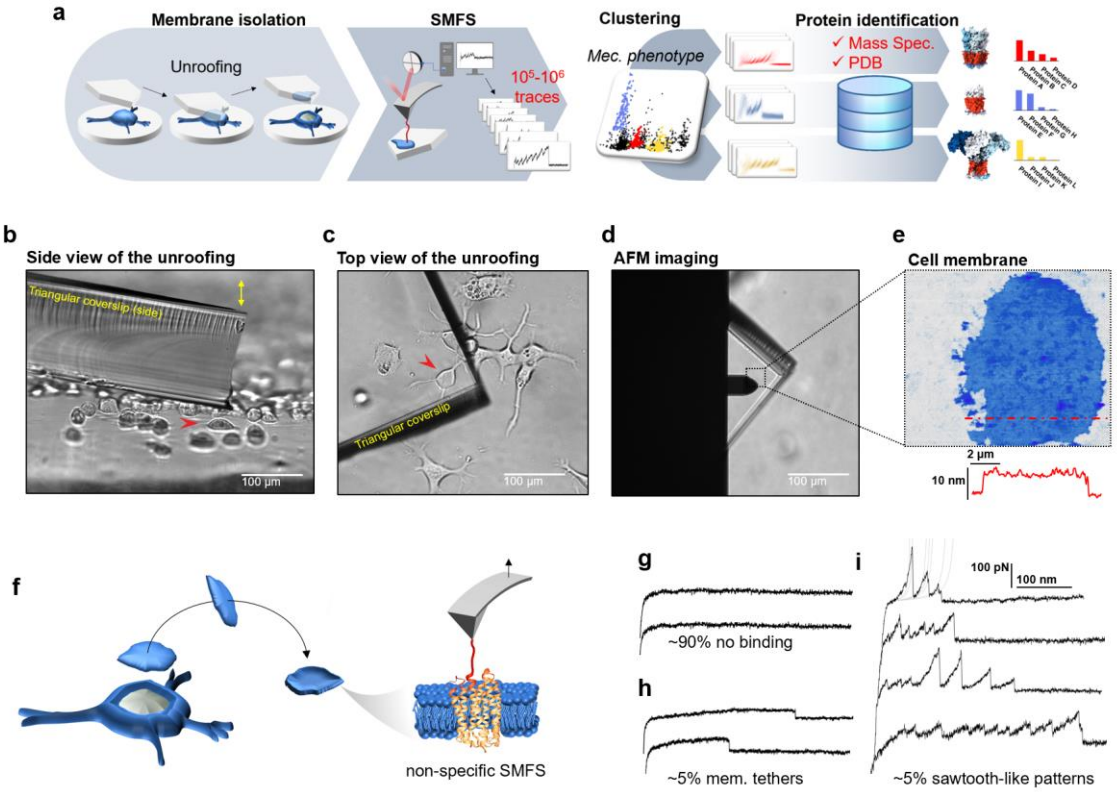

Figure 1 | Experimental method for membrane isolation and protein unfolding. a*, workflow of the method in four steps: isolation of the apical membrane of single cells; AFM-based protein unfolding of native membrane proteins; identification of the persistent patterns of unfolding and generation of the mechanical phenotype; Bayesian protein identification with mass spectrometry, Uniprot and PDB.* ***b****, side view and* ***c****, top view of the cell culture and the triangular coverslip approaching the target cell (red arrow) to be unroofed.* ***d****, positioning of the AFM tip in the region of unroofing.* ***e****, AFM topography of the isolated cell membrane with profile.* ***f****, cartoon of the process that leads to SMFS on native membranes. Examples of F-D curves of* ***g****, no binding events;* ***h****, membrane tethers that generate constant viscous force during retraction;* ***i****, sawtooth-like patterns, typical sign of the unfolding of a protein.*

Unfolding proteins from isolated cell membranes.

In order to study the unfolding of membrane proteins from their native environment, we optimized an unroofing method[15] to isolate the apical part of cell membranes. We sandwiched a single cell or neuron between two glass plates, i.e. the culture coverslip and another mounted on the AFM itself (see Fig. 1b-c, triangular coverslip). The triangular coverslip is coated with polylysine which favors membrane adhesion. When adhesion is reached, a rapid separation of the plates driven by a loaded spring permits the isolation of the apical membrane of the cell (see Fig. 1d-e, Supplementary Fig. 1). The method is reliable (n=42, ~80% success rate) with cell types grown on coverslips (epithelial cells and neurons) and the fast unroofing obtained by the introduction of the spring. For cells that do not grow in culture, like freshly isolated rods, we broke the cell with a lateral flux of medium[16].

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted August 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

After membrane isolation, we imaged the membrane with the AFM (Fig. 1f) and we verified that the isolated membrane patches have a height of 5-8 nm with rugosity in the order of 1nm. Then, we performed standard SMFS[17] with non-functionalized tips collecting 301,654 curves on the hippocampal membrane, 213,468 curves on DRG, 386,128 on rods and 221,565 on rod discs. Of the obtained curves, the ~90% shows no binding (Fig. 1 g), ~5% shows plateau ascribable to membrane tethers[18](Fig. 1 h), while the remaining >5% displays the common sawtooth-like shape that characterizes the unfolding of proteins[17,19](Fig. 1 i). Indeed, the good F-D curves are constituted by a sequence of rising concave phases followed by vertical jumps: the rising phases fit the worm-like chain (WLC) model with a persistence length of ~0.4nm indicating the stretching of an unstructured aminoacidic chain[20]. In these cases the AFM tip binds non-specifically the underlying proteins (physisorption)[8].

Architecture of membrane proteins and performance of SMFS on native membranes. The Protein Data Bank (PDB) contains 8662 entries that are also annotated in the Orientation of Proteins in Membrane (OPM)[21,22] providing the information of the position of each aminoacid relative to the cell membrane. The OPM resource provides useful statistics on the architecture of membrane proteins. We categorized all these 8662 proteins in eight different classes based on their architecture (Fig. 2 a, see Methods for details). 53% of the resolved membrane proteins are peripheral membrane proteins anchored to the membrane, of which the two thirds are located extracellularly, therefore not accessible to the AFM tip in our unroofed membrane patches (Fig.1). The intracellular peripheral membrane proteins can be unfolded only if they are tightly bounded to the membrane. The remaining 47% of these proteins are transmembrane proteins of which only the 7% have both the C- and the N-terminus in the extracellular side. Of the eight classes shown in Fig. 2 a, five (I-V) have already been investigated in purified conditions[12,14,17,23,24] and the obtained F-d curves display the usual sawtooth-like, i.e. the piece-wise WLC behavior (see also the Methods section) that is present also in our F-d curves. Class VIII is not expected to be present in our experiments as it cannot attach to a cantilever approaching from the intracellular side, while proteins of Class VI and VII can be pulled.

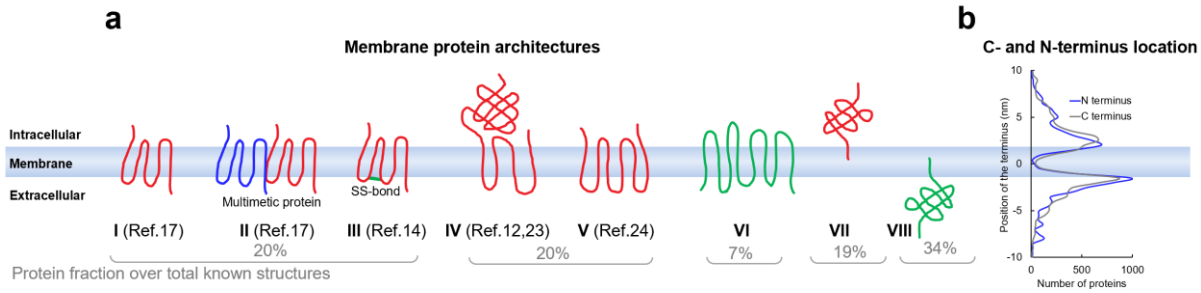

***Figure 2 | Membrane proteins architectures. a**, eight classes of membrane proteins and their fraction over all resolved proteins present in the PDB-OPM. **b**, position of the termini relative to the center cell membrane along the axis perpendicular to the membrane.*

Proteins, when pulled, generate their own characteristic pattern of unfolding[25]. By visual inspection, we observed that our F-d curves contain recurrent patterns of unfolding similar to those obtained in purified conditions when pulled from the C or N-terminus[17,23,24]. However, the attachment to either the C and N-terminus and the resulting complete unfolding of a single protein is not the only possible event that occur in our experiments. On the basis of the architectural analysis and disposition of membrane proteins, we have considered three additional cases: i) the simultaneous attachment of two or more proteins to the tip[26], ii) the incomplete unfolding of the attached protein, iii) the binding of the AFM tip to a loop of the protein instead of to a terminus end (Fig. 3 a-d).

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted August 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

i) Attachment of multiple proteins: the blind movements of the tip apex (radius of curvature 10-20 nm) leads the tip landing in random configurations on the sample so that it could bind simultaneously to multiple proteins. Since the ratio between non-empty curves over all curves is ~ 5 %, it follows that the binding probability is also close to 5%: the probability to bind 2 proteins at the same time is therefore its square (~0.2%). The attachment of multiple proteins occurs 20 times less frequently than the single attachment, and it will happen with combinations of different protein species and the resulting F-d curves will not have recurrent patterns. Furthermore, when the two chains are unfolded together the resulting spectrum is the sum of the two individual spectra: that causes deviations in the measured persistence length in the part of the curve where both chains are stretched (Supplementary Fig. 2). The simultaneous unfolding of multiple proteins is also characterized by the doubling of the peaks and evident changes in the range of the forces (Fig. 3 b and d, Supplementary Fig. 2).

ii) Incomplete unfolding of the protein: if the tip prematurely detaches from the terminus, the resulting F-d curve will display a similar but shorter pattern compared to a complete unfolding (Fig. 3 c). The fraction of curves that prematurely detaches is reported to be ~23% of the fully unfolded proteins[14], but this value could vary from protein to protein.

ii) Binding of the AFM tip to a loop: the unfolding from a loop is equivalent to the attachment of multiple proteins because the tip unfolds two chains at the same time. However, if the attachment of the cantilever tip to a loop occurs with some consistency - like to the C or the N-terminus - we will obtain a recurrent pattern with the features described in case i) (deviation of persistence length during intersection, 2 major levels of unfolding force).

We have heuristics to identify these cases which are expected to be governed by stochasticity so that the corresponding F-d curves occur *without recurrent patterns* and therefore we focused on the detection of F-d curves with clear recurrent patterns.

Finding the unfolding patterns of native membrane proteins.

The ideal methodology to find the recurrent patterns of unfolding in the data coming from native membranes is an unsupervised procedure able to filter out the stochastic events, and to identify clusters of dense patterns of any shape without setting their number *a priori*. For this purpose, we designed a pattern classification pipeline combining the density peak clustering[27] benchmarked for SMFS data[28] with a final pattern recognition method used to determine the cluster population. This pipeline can detect statistically dense patterns of unfolding within large datasets with a desktop computer (see Methods section for further details). This pipeline does not require to pre-set neither the number of clusters to be identified nor the dimension of the F-d curves and can be applied without prior knowledge of the sample composition. The method is automatic and only partially unsupervised because it considers "good" patterns only those that are in agreement with the worm-like chain model that describes the stretching of a polymer made of aminoacids with few deviations.

We found 15, 10, 8 and 5 clusters (Fig.3 i-l) in F-d curves from DRG, hippocampal neurons, rod outer segments and rod discs membranes respectively. We identified four major classes of clusters based on their unfolding behavior. Short curves with increasing forces: DRG12, H5, H8 and R3 shows repeated peaks (ΔLc 10-20 nm, distance between consecutive peaks) of increasing force that reach also 400 pN in force; these clusters resemble the unfolding behavior of tandem globular proteins[4]. Long and periodic curves: R6, H7 or DRG10 display periodic peaks of ~100 pN and with a ΔLc of 30-40 nm whose unfolding patterns are similar to the one of LacY[19]. Short curves: the majority of the identified clusters like DRG1, H3, R8 and all clusters from the rod discs have curves less than 120 nm long and with constant or descending force peaks. The F-d curves of these clusters share various features with the opsin family proteins unfolded in purified conditions[8], e.g. a conserved unfolding peak at the beginning (at contour length < 20 nm) revealing the initiation of the denaturation of the protein. We found also "unconventional" clusters such as DRG7, DRG8 and R7: DRG8 for instance shows initial high forces and variable peaks followed by a phase of more periodic low forces that recall the pattern of Fig. 2 d obtained in our model of unfolding from a loop/multiple-unfolding, while cluster R7 has a conserved flat plateau at the end of the curve of unknown origin.

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted August 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

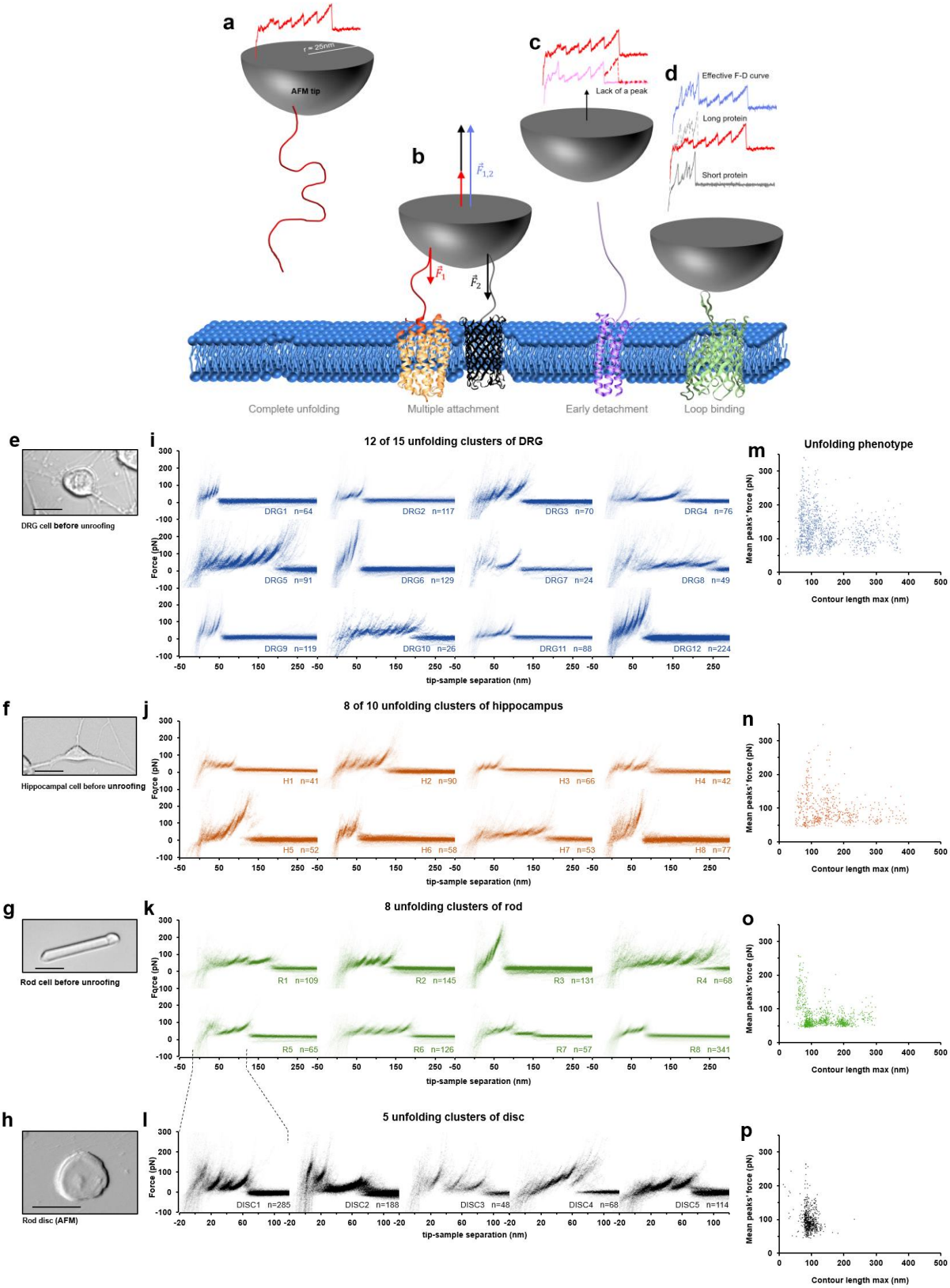

Figure 3 | Unfolding clusters in native cell membranes. *Cartoon representing* ***a****, complete unfolding of a membrane protein and its F-D curve,* ***b****, simultaneous unfolding of two proteins and the balance of the forces involved.* ***c****, incomplete unfolding of a protein,* ***d****, unfolding from a loop and prototypical F-D curve of a multiple unfolding/ unfolding from a loop (other examples in Supplementary Fig. 2). Bright field image of* ***e,*** *dorsal root ganglia neuron;* ***f***

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted August 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

hippocampal neuron; ***g*** *rot before unroofing (scale bar 15 µm).* ***h****, AFM error image of an isolated disc (scale bar 1 µm).* ***i, j, k, l****, superimposition of clustered F-D curves plotted as density maps.* ***m, n, o, p****, unfolding phenotype in the compact representation of all the clustered F-D curves in maximum contour length vs. average unfolding force space (DRG: n = 1255; hippocampus: n = 563; rod: n = 1039; disc: n = 703).*

The clustering allows also a representation of the output of the experiments in a single and compact display (Fig. 3 m-p) defining what we call the 'unfolding phenotype' of a specific cell membrane, which is peculiar of the cell type. We assign to each F-D curve different parameters related to the geometrical features that are physically relevant (maximal contour length (Lc max), average unfolding force, average ΔLc, etc.). In this way it is possible visualize the ensemble of all the clusters obtained from a specific cellular membrane and find differences in data obtained from hippocampal (Fig. 3 m) and DRG neurons (Fig. 3 n), plasma membrane of outer segments (Fig. 3 o) and discs (Fig. 3 p) so that it is possible to obtain a phenotyping of membrane proteins of a given membrane patch (Supplementary Fig. 3).

Bayesian identification of the unfolded patterns.

Having identified clusters of F-d curves from native membranes, the next question is: *which is the membrane protein whose unfolding corresponds to the identified clusters in Fig. 3?* In order to answer to this question, we developed a Bayesian method providing a limited list of candidate proteins on the basis of the information present in data from Mass Spectrometry of the sample under investigation and general proteomic databases (ProteomeXchange for mass spectrometry, Uniprot, PDB). The Bayesian identification (Fig. 4 a.) is based on two steps: firstly, the crossing of information from the cluster under investigation and the results of Mass Spectrometry analysis of the sample (hippocampal neurons, discs, etc.); secondly, a refinement of the preliminary candidates using additional information (structural and topological) present in the PDB and Uniprot databases.

The first step leverages the contour length of the last peak of the clusters (Lc_{max}; Fig. 4 a I). The SMFS-literature contains 14 examples of unfolded membrane proteins allowing a comparison between the Lc_{max} of the measured F-d curves and the real length of the same protein completely stretched (Fig.4 b). On the basis of these experiments, we extrapolated the first likelihood function of our Bayesian inference (Fig. 4 c) suggesting that, on average, the Lc_{max} corresponds to 89% of the real length of the protein. By searching for proteins with this total length in the Mass Spectrometry data from the same samples[29–31] and by using their abundance (Fig. 4 a II) we obtained a first list in which we could assign a probability to each candidate.

The refinement to the first step (Fig. 4 a III) is obtained by combining the information on the molecular structure of the proteins (Fig. 4 a IV) extracted from the PDB and Uniprot. We created a table containing all the membrane proteins present in the Mass Spec data from the sample under investigation (hippocampal neurons, rods, etc.) containing their abundance, number of amino acids, subcellular location, orientation of the N and C terminus, fraction of alpha helices and beta sheets for each protein, and the presence of Cys-Cys bonds (Fig. 4 b, Supplementary Tables). The Bayesian approach assigns to the candidate proteins a probability also based on the location of the C and N terminus, and on the fact that unfolding beta sheets typically requires forces larger than in the case of alpha helices (see Fig. 4 d). From this distribution we obtained the second likelihood function (Fig. 4 e) of our model.

A separate discussion needs to be made for proteins with disulfide bonds (i.e. covalent bonds between non-adjacent cysteines) that have a high breaking force[32], till 1 nN. As a result, the mechanical unfolding of the protein might be not sufficient to break the bonds, generating a cluster with a shorter Lc_{max}[14,32]. The effective length of the protein with disulfide bonds is therefore reduced of the length enclosed between two consecutive bonded cysteines. The crossing with the Uniprot database that contains the information of the disulfide bonds allowed us to recalculate the effective total length of the proteins in our lists. The structure of the lists is summarized in Fig. 4 f, while the complete tables are attached in the Supplementary data of the article.

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted August 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

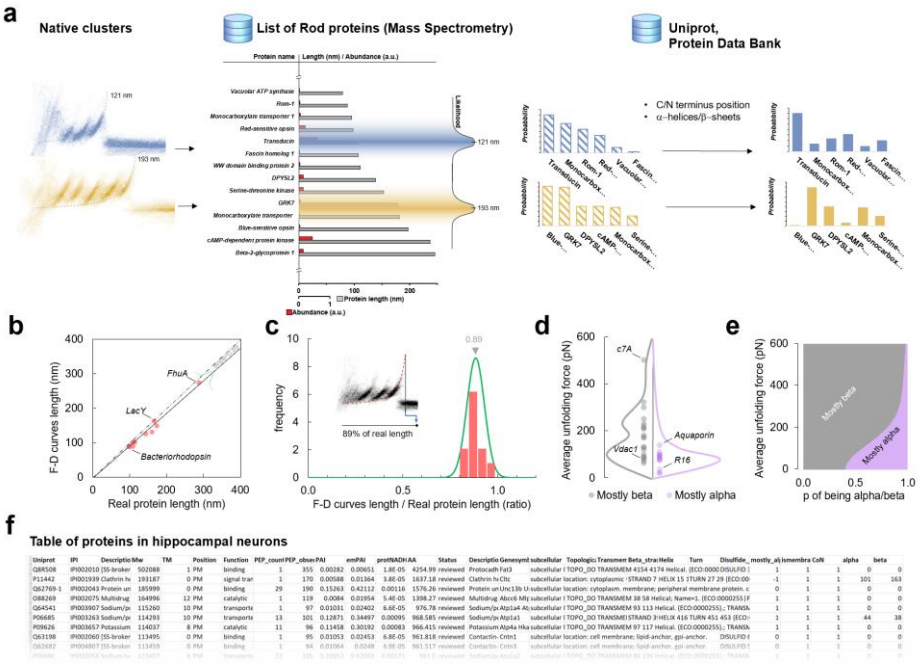

Figure 4 | Likelihoods and priors for the Bayesian identification. a*, workflow of the Bayesian steps: selection due to total length and abundance (mass spectrometry), refinement with structural and topological information (PDB and Uniprot).* ***b****, Comparison of the real length of the protein vs. the measured maximal contour length of the F-D curves in 14 SMFS experiments on membrane proteins (see Methods).* ***c****, Likelihood function of the observed maximal length of the clusters obtained from* ***b****.* ***d****, Comparison of the force necessary to unfold beta sheets and alpha helices in 22 SMFS experiments (see Methods).* ***e****, Likelihood function of the observed unfolding forces obtained from* ***d****.* ***f****,*

Following the Bayesian inference, we developed a software to estimate the probability of the candidate proteins for all the unfolding clusters found in hippocampal neurons, rod membranes and discs (Fig. 5 a-c). Starting from no information on the nature of these unfolding events, the software provides a list of known proteins which are the candidates of the molecules unfolded in the clusters of Fig.3. The software not only provides the candidates but assign to each known protein a probability based on the Bayesian inference (Fig.4). Therefore, by simply crossing and exploiting the large information available in various databases, we identified a restricted number of molecular candidates for the identified unfolding clusters (Fig. 5). The more accurate assignations happen when a protein has a very high abundance (e.g. rhodopsin in discs and rods) or when there are few proteins of the same mass (length) of the identified protein.

To verify this analysis, we looked for an orthogonal validation of the proposed method, based on the results of two membrane proteins unfolded in native membranes, i.e. the cyclic nucleotide gated channel yet unfolded in semi-purified conditions[12] and hypothesized in previous experiments in the plasma membrane of rod outer segments[33], and the rhodopsin unfolded in discs[14,33]. The CNG got a probability equal to 29% for the longest cluster in the rod plasma membrane (Fig 2 k, R4) mostly due to a combination of the correct Lc window and high abundance. The pattern shows the occurrence of the 5 major unfolding barriers. We engineered a chimera of the CNG with N2B on the C-terminus that we overexpressed in the hybrid conditions explained in ref. [12]. These experiments generated an unfolding cluster with the same unfolding barrier shifted of ~ 85 nm, i.e. the length of the N2B, which confirmed also the fact that we were unfolding from the C- terminus (Supplementary Fig. 4 a-f).

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted August 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

With rhodopsin we reproduced the experiments performed in discs in ref. [14,33]. In discs we obtained 5 unfolding clusters of which 2 matches the unfolding pattern of Tanuj et al. (Supplementary Fig. 4 g-l): the identity of these clusters was demonstrated with the enzymatic digestion that caused a truncation in the C-III loop of the rhodopsin molecule. The experiments performed after enzymatic digestion showed a 40-fold reduction of long curves confirming the identity of rhodopsin

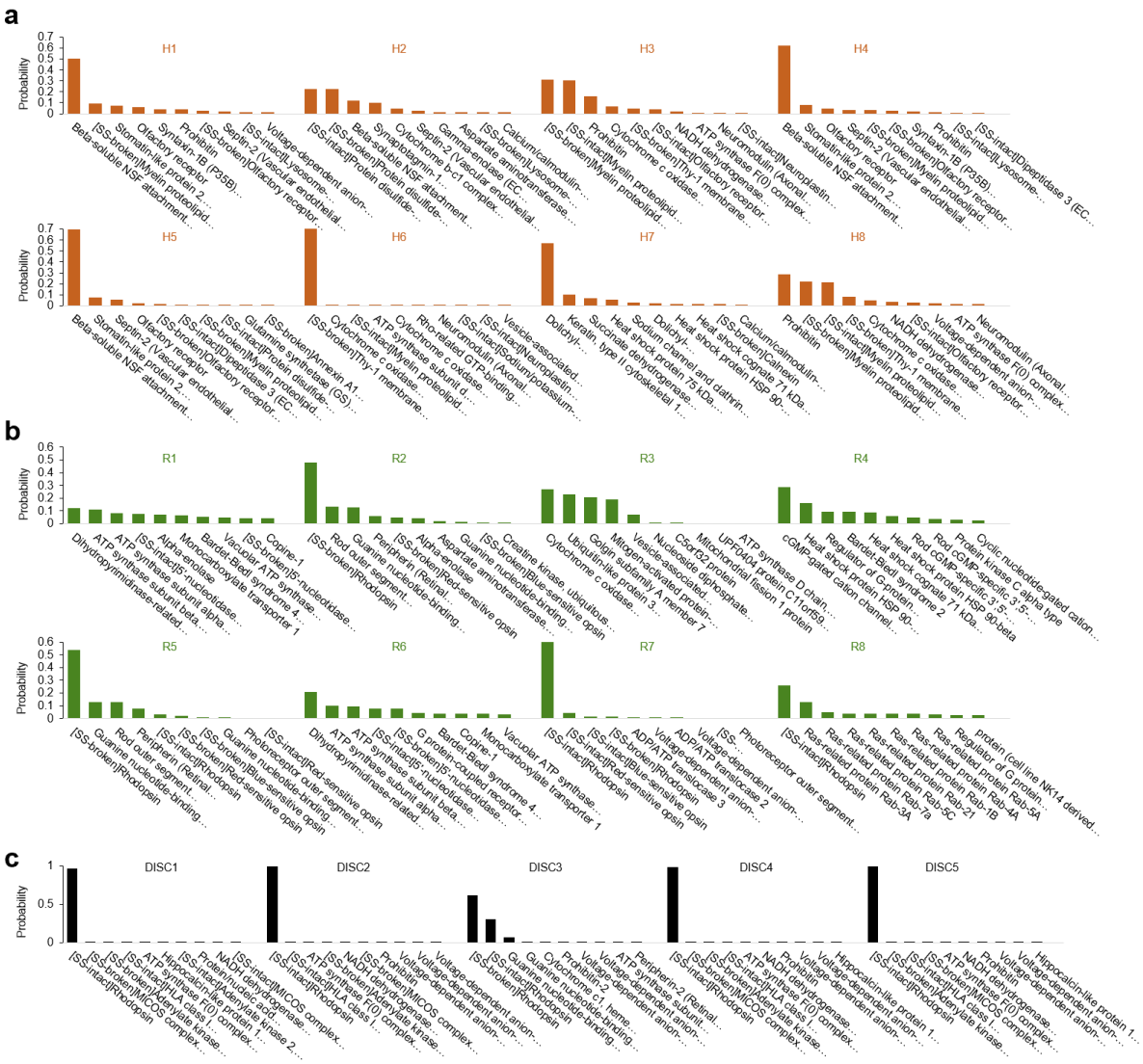

Figure 5 | Bayesian identification of the unfolding clusters. *Most probable candidates for the unfolding clusters found in* ***a****, hippocampal neurons;* ***b****, rods;* ***c****, rod discs. The label 'SS-broken' and 'SS-intact' refers to state of the disulfide bond after the unfolding.*

Discussion

The method here illustrated describes all the necessary steps to obtain F-d curves from biological membranes of many cell types that grow in culture, and provides an automatic way to obtain clusters of F-d curves representing the unfolding of the membrane proteins present in the sample. We describe also a Bayesian approach able to provide a list of known proteins as candidates to be the unfolded protein. The Bayesian approach depends on the information present in Mass Spectrometry data and on the PDB and Uniprot databases. Therefore, the list of candidate proteins is expected to be refined as these databases will become richer and

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted August 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

more complete, and the quality of Mass Spectrometry data will be improved. Let us discuss, now, the advantages and the weaknesses of the proposed method.

The possibility to perform SMFS experiments in natural samples obtained from native cells provides a clear breakthrough in the field of protein unfolding by avoiding purification and reconstitution, but it has implications in particular in complementary fields. An example is the possibility to characterize molecules coming from a very limited amount of native material (membranes isolated from 1 to 10 cells). The unfolding phenotype is a univocal tool to characterize the sample under investigation (see Fig.3) and this approach could be extended to characterize membrane proteins in neurons/cells in healthy and sick conditions. Indeed, it is remarkable that the distribution of the detected proteins in our SMFS experiments (solid lines in Fig.6) is similar to that obtained in the Mass Spec experiments of thousands of cells (broken lines). This is also an *ex post* confirmation of the goodness of using the Mass Spec data in the Bayesian inference.

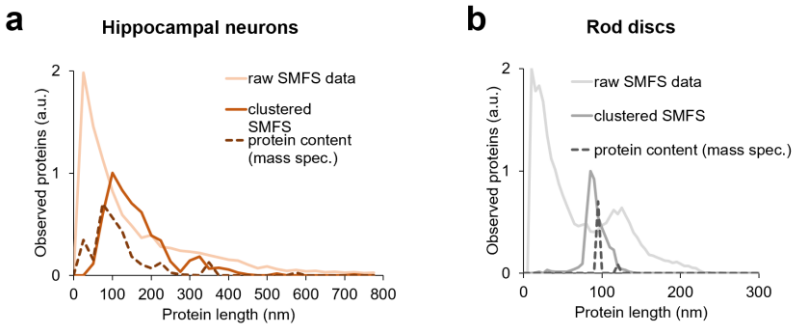

Figure 6 | Comparison of mass spectrometry protein detection vs. SMFS data. a*, hippocampal neurons.* ***b****, rod discs.*

In our experiments we collected a limited number of F-d curves – some hundreds of thousands – and by increasing their number by 10- or 100-fold, we expect to improve the total number of detected clusters – as those in Fig.3 – possibly close to 100. As the total number of different membrane proteins from a native sample is on the order of hundreds, we would be able to detect and characterize a significant fraction of the total membrane proteins present in the sample. Improvements of the proposed method, primarily by increasing its throughput, could potentially provide a new screening method with clinical applications: indeed, the characterization of the changes of the unfolding phenotype caused by a disease will provide a better understanding of the malfunction of membrane proteins. Moreover, the proposed method is able to explore the variety of proteins present in a sample with an accuracy almost similar to that obtained by Mass Spec, but using a much simpler apparatus.

The proposed method has inherent limitations: indeed, the molecular identity of the unfolded proteins is guessed by a Bayesian estimator, which can be improved, but cannot be firmly established as in experiments with purified proteins. A possible way to obtain a better and more reliable identification of the proteins in the membrane would be to couple the SMFS analysis of the native sample with a high-resolution AFM imaging of the same native samples: in this way it will be possible also to “see” all the proteins present in the sample. Unfortunately, the necessary molecular resolution is nowadays difficult to be achieved even in purified conditions, therefore we don’t expect this to be possible in few years.
The proposed method for clustering F-d curves is automatic but it is not fully unsupervised indeed Block 3 - in which we evaluate the quality of the F-d curve - assumes that a good F-d curve is piece-wise close to WLC. Block 5 of clustering method requires also a refinement which is done by the experimenter. The development of an unsupervised and fully automatic clustering method is under way.
Another major limitation of the proposed method – in its present form – is the possibility to merge in the same cluster the unfolding of proteins with a different molecular identity: indeed,

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted August 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

from the Mass Spec data it's evident that different proteins have the same – or approximately the same – molecular weight and total unfolded length Lc, in particular for Lc between 50 nm and 200 nm. In order to overcome this limitation, it will be desirable to couple to SMFS some chemical information on the unfolded protein. In our opinion, this will be a desirable achievement, which will make a substantial improvement to the method here proposed.

Acknowledgments

We thank dr. Kosaku Shinoda for support in the emPAI determination.

References

1. Al-Rekabi, Z. & Contera, S. Multifrequency AFM reveals lipid membrane mechanical properties and the effect of cholesterol in modulating viscoelasticity. *PNAS* **115**, 2658–2663 (2018).
2. Casuso, I. *et al.* Characterization of the motion of membrane proteins using high-speed atomic force microscopy. *Nature Nanotechnology* **7**, 525–529 (2012).
3. García-Sáez, A. J., Chiantia, S. & Schwille, P. Effect of Line Tension on the Lateral Organization of Lipid Membranes. *J. Biol. Chem.* **282**, 33537–33544 (2007).
4. Carrion-Vazquez, M. *et al.* Mechanical and chemical unfolding of a single protein: A comparison. *PNAS* **96**, 3694–3699 (1999).
5. Sarkar, A., Caamano, S. & Fernandez, J. M. The Elasticity of Individual Titin PEVK Exons Measured by Single Molecule Atomic Force Microscopy. *J. Biol. Chem.* **280**, 6261–6264 (2005).
6. Scheuring, S. & Sturgis, J. N. Chromatic Adaptation of Photosynthetic Membranes. *Science* **309**, 484–487 (2005).
7. Baumgartner, W. *et al.* Cadherin interaction probed by atomic force microscopy. *Proceedings of the National Academy of Sciences* **97**, 4005–4010 (2000).
8. Engel, A. & Gaub, H. E. Structure and Mechanics of Membrane Proteins. *Annual Review of Biochemistry* **77**, 127–148 (2008).
9. Otten, M. *et al.* From genes to protein mechanics on a chip. *Nat Meth* **11**, 1127–1130 (2014).
10. Thoma, J., Burmann, B. M., Hiller, S. & Müller, D. J. Impact of holdase chaperones Skp and SurA on the folding of β-barrel outer-membrane proteins. *Nat Struct Mol Biol* **22**, 795–802 (2015).
11. Hinczewski, M., Hyeon, C. & Thirumalai, D. Directly measuring single-molecule heterogeneity using force spectroscopy. *PNAS* **113**, E3852–E3861 (2016).
12. Maity, S. *et al.* Conformational rearrangements in the transmembrane domain of CNGA1 channels revealed by single-molecule force spectroscopy. *Nature Communications* **6**, 7093 (2015).
13. Thoma, J. *et al.* Protein-enriched outer membrane vesicles as a native platform for outer membrane protein studies. *Communications Biology* **1**, 23 (2018).
14. Tanuj Sapra, K. *et al.* Detecting Molecular Interactions that Stabilize Native Bovine Rhodopsin. *Journal of Molecular Biology* **358**, 255–269 (2006).
15. Galvanetto, N. Single-cell unroofing: probing topology and nanomechanics of native membranes. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1860**, 2532–2538 (2018).
16. Clarke, M., Schatten, G., Mazia, D. & Spudich, J. A. Visualization of actin fibers associated with the cell membrane in amoebae of Dictyostelium discoideum. *PNAS* **72**, 1758–1762 (1975).

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted August 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

17. Oesterhelt, F. *et al.* Unfolding Pathways of Individual Bacteriorhodopsins. *Science* **288**, 143–146 (2000).
18. Chu, C., Celik, E., Rico, F. & Moy, V. T. Elongated Membrane Tethers, Individually Anchored by High Affinity α4β1/VCAM-1 Complexes, Are the Quantal Units of Monocyte Arrests. *PLoS ONE* **8**, e64187 (2013).
19. Serdiuk, T. *et al.* YidC assists the stepwise and stochastic folding of membrane proteins. *Nat Chem Biol* **12**, 911–917 (2016).
20. Li, H. *et al.* Reverse engineering of the giant muscle protein titin. *Nature* **418**, 998–1002 (2002).
21. Lomize, M. A., Pogozheva, I. D., Joo, H., Mosberg, H. I. & Lomize, A. L. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Research* **40**, D370–D376 (2012).
22. Lomize, A. L., Pogozheva, I. D. & Mosberg, H. I. Anisotropic Solvent Model of the Lipid Bilayer. 2. Energetics of Insertion of Small Molecules, Peptides, and Proteins in Membranes. *Journal of Chemical Information and Modeling* **51**, 930–946 (2011).
23. Thoma, J., Bosshart, P., Pfreundschuh, M. & Müller, D. J. Out but Not In: The Large Transmembrane β-Barrel Protein FhuA Unfolds but Cannot Refold via β-Hairpins. *Structure* **20**, 2185–2190 (2012).
24. Sapra, K. T. *et al.* One β Hairpin after the Other: Exploring Mechanical Unfolding Pathways of the Transmembrane β-Barrel Protein OmpG. *Angewandte Chemie International Edition* **48**, 8306–8308 (2009).
25. Marsico, A., Labudde, D., Sapra, T., Muller, D. J. & Schroeder, M. A novel pattern recognition algorithm to classify membrane protein unfolding pathways with high-throughput single-molecule force spectroscopy. *Bioinformatics* **23**, e231–e236 (2007).
26. Walder, R. *et al.* Rapid Characterization of a Mechanically Labile α-Helical Protein Enabled by Efficient Site-Specific Bioconjugation. *Journal of the American Chemical Society* **139**, 9867–9875 (2017).
27. Rodriguez, A. & Laio, A. Clustering by fast search and find of density peaks. *Science* **344**, 1492–1496 (2014).
28. Ilieva, N. Computational analysis of single molecule force spectroscopy experiments on native membranes. *PhD thesis*
29. Chen, P. *et al.* Proteomic analysis of rat hippocampal plasma membrane: characterization of potential neuronal-specific plasma membrane proteins. *Journal of Neurochemistry* **98**, 1126–1140 (2006).
30. Kwok, M. C. M., Holopainen, J. M., Molday, L. L., Foster, L. J. & Molday, R. S. Proteomics of Photoreceptor Outer Segments Identifies a Subset of SNARE and Rab Proteins Implicated in Membrane Vesicle Trafficking and Fusion. *Molecular & Cellular Proteomics* **7**, 1053–1066 (2008).
31. Panfoli, I. *et al.* Proteomic Analysis of the Retinal Rod Outer Segment Disks. *Journal of Proteome Research* **7**, 2654–2669 (2008).
32. Ainavarapu, S. R. K. *et al.* Contour Length and Refolding Rate of a Small Protein Controlled by Engineered Disulfide Bonds. *Biophysical Journal* **92**, 225–233 (2007).
33. Maity, S., Ilieva, N., Laio, A., Torre, V. & Mazzolini, M. New views on phototransduction from atomic force microscopy and single molecule force spectroscopy on native rods. *Scientific Reports* **7**, (2017).