bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 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 The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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], Zhongjie Ye[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 a method providing all the necessary steps to deal with native membranes, from the isolation of the plasma membrane of single cells, to the characterization and identification of the embedded membrane proteins. We combined single-molecule force spectroscopy with an automatic pattern classification of the obtained Force-distance curves, and we provide a Bayesian identification of the unfolded proteins. The Bayesian identification is based on the cross-matching of Mass Spectrometry datasets with proteomic databases (Uniprot, PDB). We applied this method to four cell types where we classified the unfolding of 5-10% of their total content of membrane proteins. The ability to mechanically probe membrane proteins directly in their native membrane enables the phenotyping of different cell types with almost single-cell level of resolution.

Introduction

Much of what we know about the mechanics of cell membranes[1–3] and polymers[4,5] comes from 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. 1 a). 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

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 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.

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[15].

Besides the identification, the proposed methodology generates as by-product the phenotyping of the membrane proteins content of specific cells that may become relevant in biomedical applications.

Results

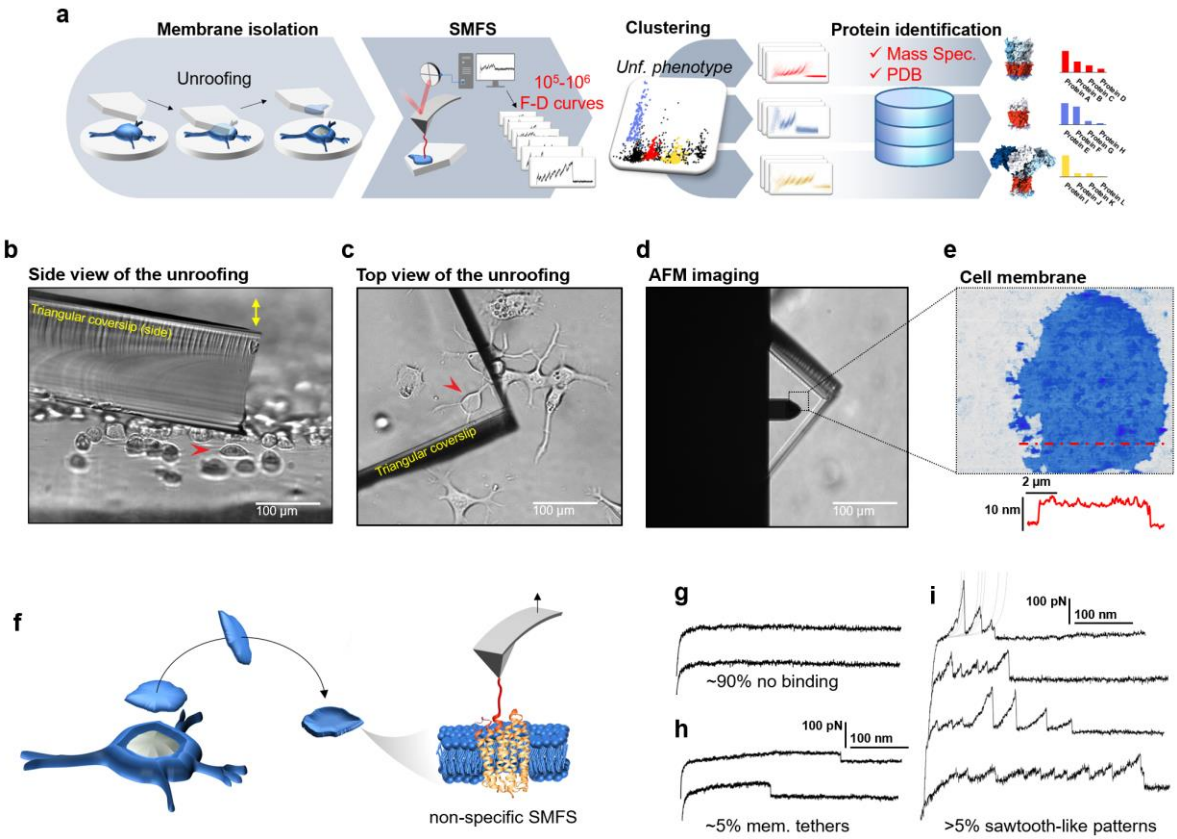


Figure 1 | Experimental method for single-cell membrane isolation and protein unfolding. a*, workflow of the method in four steps: isolation of the apical membrane of single cells; AFM-based protein unfolding of native membrane proteins; identification of the persistent patterns of unfolding and generation 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[16] to isolate the apical part of cell membranes containing primarily membrane proteins with negligible contamination of cytoplasmic proteins (Supplementary Fig. 1). 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

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 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.

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). With non-adherent cells, like freshly isolated rods, we isolated the membrane with a lateral flux of medium[17] (see Methods).

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 1 nm. Then, we performed standard SMFS[18] 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[19](Fig. 1 h), while the remaining >5% displays the common sawtooth-like shape that characterizes the unfolding of proteins[18,20](Fig. 1 i). Indeed, F-D curves representing unfolding events are constituted by a sequence of rising concave phases followed by vertical jumps, where 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[21]. In these cases the AFM tip binds non-specifically the underlying proteins through physisorption[8].

The architecture of membrane proteins and their unfolding.

The Protein Data Bank (PDB) contains 8662 entries that are also annotated in the Orientation of Proteins in Membrane (OPM)[22,23] providing the information of the position of each aminoacid relative to the cell membrane. The OPM is therefore a useful resource from which we extrapolated 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 (class VIII of Fig. 2 a), thus not accessible by the AFM tip in our experiments (Fig.1). The intracellular peripheral membrane proteins (class VII) can be unfolded only if they tightly bound 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 (class VI). Of the eight classes shown in Fig. 2 a, five (I-V) have already been investigated in purified conditions[12,14,18,24,25] and the obtained F-d curves display the usual sawtooth-like, i.e. the piece-wise WLC behavior 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 by a cantilever approaching from the cytoplasmic side.

Analysis of SMFS data from native cell membranes

Membrane proteins, when pulled, generate their own characteristic pattern of unfolding which is used for their selection[26,27]. Visual inspection shows that the obtained F-d curves contain recurrent patterns of unfolding similar to those obtained in purified conditions when pulled from either the C or N-terminus[18,24,25]. 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, we have considered three additional cases: i) the simultaneous attachment of two or more proteins to the tip[28], ii) the incomplete unfolding of the attached protein[14], iii) the binding of the AFM tip to a loop of the protein instead of to a terminus end (Fig. 2 c-f).

<u>i) Attachment of multiple proteins</u> (Fig. 2 d): 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 (~0.25%). 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 a recurrent

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 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.

pattern. Furthermore, when the two chains are unfolded together, the resulting spectrum is the sum of the two individual spectra: that causes deviations of 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 and persistence length (Fig. 2 d and g, Supplementary Fig. 2).

ii) Incomplete unfolding of the protein (Fig. 2 e): 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.

iii) Binding of the AFM tip to a loop (Fig. 2 f): 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 with 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 all these cases (see also Supplementary Fig. 2), and in particular case i) and ii) 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.

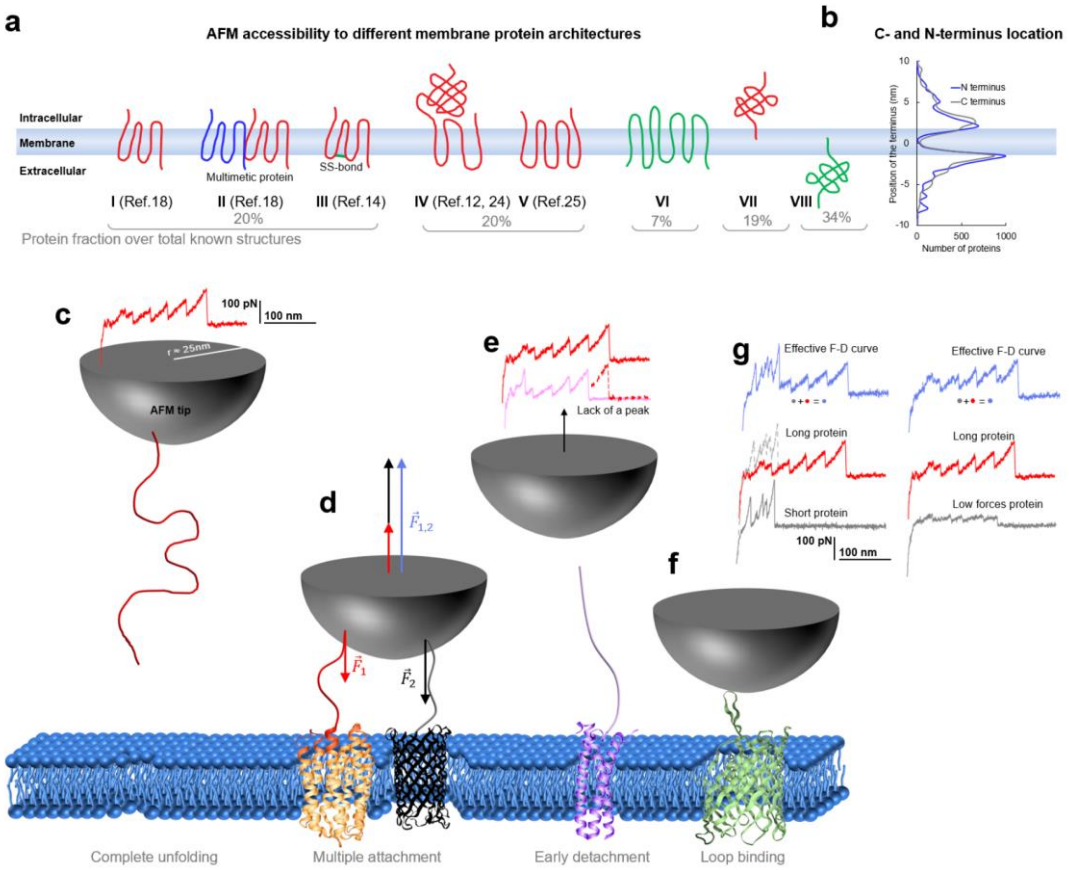


***Figure 2 | Membrane proteins architectures. a**, eight classes of membrane proteins and their fraction over all resolved proteins present in the PDB-OPM. **b**, position of the termini relative to the center cell membrane along the axis perpendicular to the membrane. Cartoon representing **c**, complete unfolding of a membrane protein and its F-D curve, **d**, simultaneous unfolding of two proteins and the balance of the forces involved. **e**, incomplete unfolding of a protein, **f**, unfolding from a cytoplasmic loop. **g**, prototypical F-D curves of a multiple unfolding/ unfolding from a loop.*

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 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.

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[29] benchmarked for SMFS data {Ref. to Nina's thesis} 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). 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.

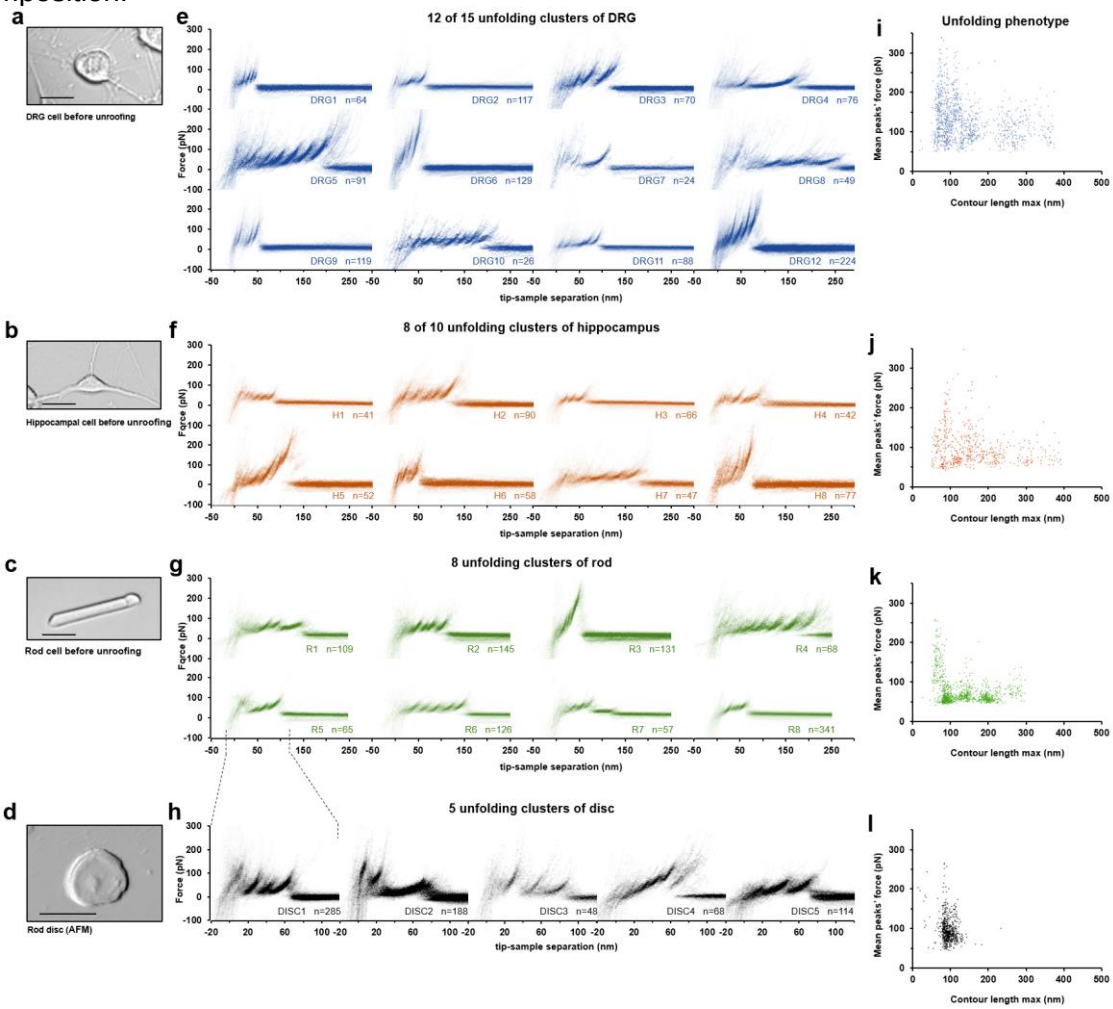


Figure 3 | Unfolding clusters in native cell membranes. *Bright field image of* ***a,*** *dorsal root ganglia neuron;* ***b*** *hippocampal neuron;* ***c*** *rot before unroofing (scale bar 15 µm).* ***d****, AFM error image of an isolated disc (scale bar 1 µm).* ***e, f, g, h****, superimposition of clustered F-D curves plotted as density maps.* ***i, j, k, l****, 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).*

We found 15, 10, 8 and 5 clusters (Fig.3 e-h) of 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 reaches also 400 pN in force; these clusters resemble the

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 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 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 what seen in the LacY[20]. 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) associated to the initiation of the denaturation of the protein. We found also "unconventional" clusters such as DRG7, DRG8 and R7: DRG8 exhibiting initial high forces and with variable peaks followed by more periodic low forces; while cluster R7 has a conserved flat plateau at the end of the curve of unknown origin. This last class displays features in common with the hypothesized unfolding from a loop or from multiple proteins.

The clustering allows also a representation of the output of the experiments in a single and compact display (Fig. 3 i-l) defining what we call the 'unfolding phenotype' of a specific cell membrane, which is peculiar of the cell type. We assigned 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 to phenotype the membrane protein landscape across cell types by visualizing the ensemble of all the clusters (see Supplementary Fig. 4).

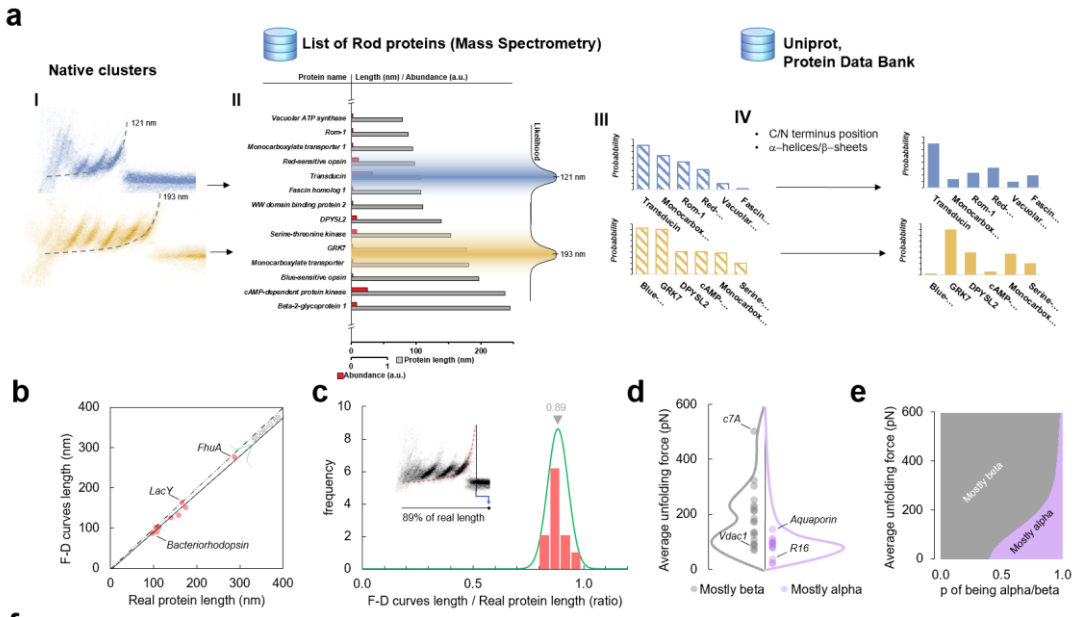


***Figure 4 | Likelihoods and priors for the Bayesian identification. a**, workflow of the Bayesian steps: selection due to total length and abundance (mass spectrometry), refinement with structural and topological information (PDB and Uniprot). **b**, Comparison of the real length of the protein vs. the measured maximal contour length of the F-D curves in 14 SMFS experiments on membrane proteins. **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. **e**, Likelihood function of the observed unfolding forces obtained from **d**. **f**, example of table entries resulting from the combination of mass spectrometry, Uniprot and PDB (Supplementary data).*

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 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.

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 the data from Mass Spectrometry of the sample under investigation and other proteomic databases (Uniprot, PDB). The Bayesian identification (Fig. 4 a.) is based on two steps: firstly, the crossing of information between 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) indicating that, on average, the Lc_{max} corresponds to 89% of the real length of the protein (R^2=0.98). By searching for proteins with this total length in the Mass Spectrometry data from the same samples[30–32] 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.) reporting their abundance, number of amino acids, subcellular location, orientation of the N- and C-terminus, topology, fraction of α-helices and β-sheets, and presence of SS-bonds (Fig. 4 f, 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 β-sheets typically requires larger forces than in the case of α-helices (see Fig. 4 d). Indeed, from this force distribution we obtained the second likelihood function (Fig. 4 e) of our model.

There are proteins for which it is available a precise annotation of their topology (usually the most abundant proteins), in these cases we can be more precise assigning them an effective contour length (Lc_{max}) based on the real structure, and also identifying whether they are unfolded from the C- or the N-terminus.

Disulfide bonds (i.e. covalent bonds between non-adjacent cysteines) are known to have a high breaking force[33], till 1 nN. As a result, the mechanical unfolding of the protein with SS-bonds is usually not sufficient to break the bonds, generating a cluster with a shorter Lc_{max}[14,33]. 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 framework of the lists is shown in Fig. 4 f, while the tables with all the information can be found in the Supplementary data of the article.

Following the Bayesian inference, we developed a method 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.

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 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.

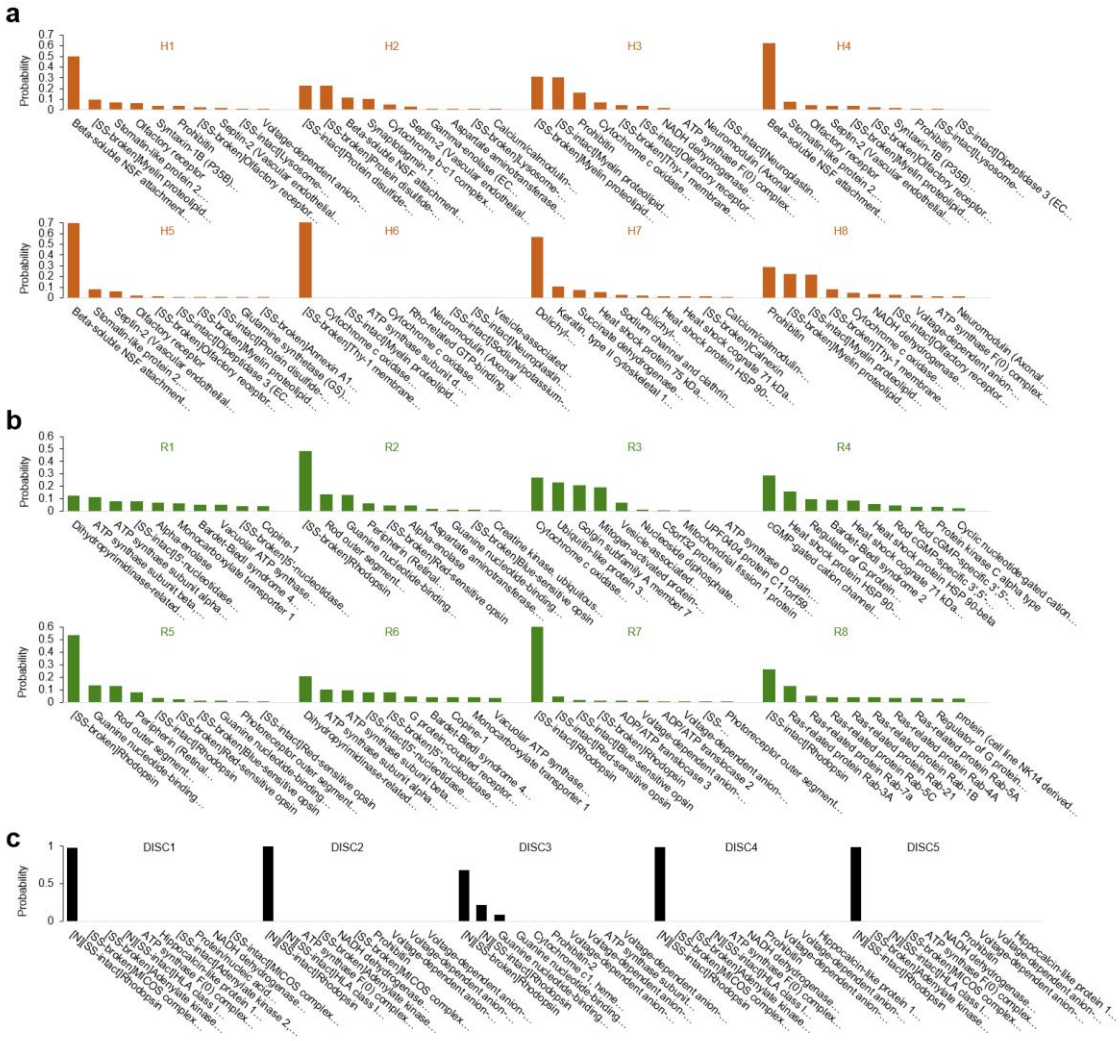


Figure 5 | Bayesian identification of the unfolding clusters. *Most probable candidates for the unfolding clusters found in* ***a****, hippocampal neurons;* ***b****, rods;* ***c****, rod discs.*

If the available data of Mass Spec from disks is complete—or almost complete—the proposed method will identify the proteins corresponding to the identified clusters with a probability close to 1.

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[34], and the rhodopsin unfolded in discs[14,34]. The unfolding pattern from the C-terminus of the CNGA1 in semi-purified conditions[12] displays 5 major unfolding barriers starting from 100 nm and with a periodicity of ~30 nm, which are features similar to those observed in cluster R4. The CNGA1 is a highly abundant protein in the rod membrane, and indeed the Bayesian identification assign a probability of 29% for cluster R4 mostly due to a combination of the correct Lc window and its high abundance. 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. 5 a-f).

With rhodopsin we reproduced the experiments performed in discs in ref. [14,34]. In discs we obtained 5 unfolding clusters of which DISC1 and DISC3 match the rhodopsin unfolding patterns of Tanuj et al. (see Supplementary Fig. 5 g-l), while DISC2, DISC4 and DISC5—

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 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.

according to our identification—represent alternative unfolding pathways for rhodopsin. The identity of these clusters was demonstrated by enzymatic digestion with endoproteinase Glu-C 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 the F-D curves with a length comparable with rhodopsin, confirming the molecular origin of our unfolding clusters.

Discussion

The method here illustrated describes all the necessary steps to obtain F-d curves from biological membranes of cell types that grow in adhesion, 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 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 bypassing purification and reconstitution. In addition, the comparison of F-d curves obtained from the same protein in its natural environment and after purification will provide new insights on the role of the physico-chemical environment of the mechanical properties of proteins, a very important issue not yet properly investigated.

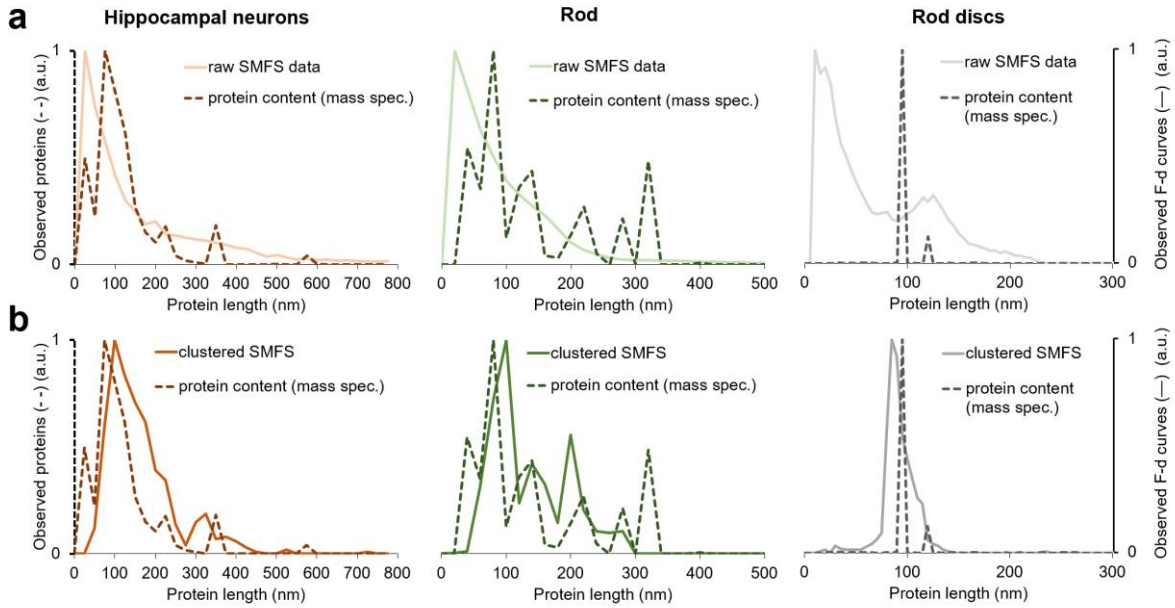


Figure 6 | Comparison of mass spectrometry protein detection vs. SMFS data. *Number of proteins / number of F-d curves observed for each length interval.* ***a****, normalized distribution of membrane proteins observed by mass spectrometry (broken lines) and distribution of the raw F-d curves.* ***b****, distribution of the clustered F-d curves of Fig. 3 approximate the distribution of the observed membrane proteins by mass spec. To the distributions of the F-d curves was applied the length correction of Fig. 4 c.*

One of the most relevant follow-ups of the method here proposed 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 Supplementary Fig. 4) and this approach could be

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 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.

extended to characterize membrane proteins in 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 using millions of cells (broken lines). This is also an *ex post* confirmation of the goodness of using the mass spectrometry data in the Bayesian inference, and that our clustering method selects correctly the protein unfolding events.

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 expect 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 some 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 is to couple the SMFS analysis of the native sample with a high-resolution AFM imaging of the same samples or, alternatively, we envision the use of the AFM cantilever as a mass sensor[35,36] of the unfolded proteins that could permit to exclude F-d curves where there is a mismatch between mass and length. However, in both cases, current technology is at least one order of magnitude away from the resolution needed.

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, from the Mass Spec data it is clear that different proteins have the same—or approximately the same—molecular weight and the total unfolded length Lc. this issue is rather significant for short proteins, I,e, those with values of Lc between 50 nm and 200 nm. In order to overcome this limitation, it will be desirable to couple SMFS with 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 calculation. We thank Prof. Anna Menini who provided the mTMEM16A-GFP and mTMEM16A-GFP plasmids. We thank Prof. Guidalberto Manfioletti who provided the peGFP-N1 plasmid.

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 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.

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. Kawamura, S., Colozo, A. T., Müller, D. J. & Park, P. S.-H. Conservation of Molecular Interactions Stabilizing Bovine and Mouse Rhodopsin. *Biochemistry* **49**, 10412–10420 (2010).
16. Galvanetto, N. Single-cell unroofing: probing topology and nanomechanics of native membranes. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1860**, 2532–2538 (2018).
17. 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).
18. Oesterhelt, F. *et al.* Unfolding Pathways of Individual Bacteriorhodopsins. *Science* **288**, 143–146 (2000).
19. 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).
20. Serdiuk, T. *et al.* YidC assists the stepwise and stochastic folding of membrane proteins. *Nat Chem Biol* **12**, 911–917 (2016).
21. Li, H. *et al.* Reverse engineering of the giant muscle protein titin. *Nature* **418**, 998–1002 (2002).
22. 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).

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 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.

23. 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).
24. 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).
25. 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).
26. 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).
27. Spoerri, P. M. *et al.* Structural Properties of the Human Protease-Activated Receptor 1 Changing by a Strong Antagonist. *Structure* **26**, 829-838.e4 (2018).
28. 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).
29. Rodriguez, A. & Laio, A. Clustering by fast search and find of density peaks. *Science* **344**, 1492–1496 (2014).
30. 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).
31. 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).
32. Panfoli, I. *et al.* Proteomic Analysis of the Retinal Rod Outer Segment Disks. *Journal of Proteome Research* **7**, 2654–2669 (2008).
33. 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).
34. 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).
35. Boisen, A. & Thundat, T. Design & fabrication of cantilever array biosensors. *Materials Today* **12**, 32–38 (2009).
36. Martínez-Martín, D. *et al.* Inertial picobalance reveals fast mass fluctuations in mammalian cells. *Nature* **550**, 500–505 (2017).
37. Mazzolini, M. *et al.* The phototransduction machinery in the rod outer segment has a strong efficacy gradient. *PNAS* **112**, E2715–E2724 (2015).
38. Butt, H.-J., Jaschke, M. & Ducker, W. Measuring surface forces in aqueous electrolyte solution with the atomic force microscope. *Bioelectrochemistry and Bioenergetics* **38**, 191–201 (1995).
39. Thoma, J. *et al.* Maltoporin LamB Unfolds β Hairpins along Mechanical Stress-Dependent Unfolding Pathways. *Structure* **25**, 1139-1144.e2 (2017).
40. Müller, D. J. *et al.* Stability of Bacteriorhodopsin α-Helices and Loops Analyzed by Single-Molecule Force Spectroscopy. *Biophysical Journal* **83**, 3578–3588 (2002).

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 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.

Methods

All experimental procedures were in accordance with the guidelines of the Italian Animal Welfare Act, and their use was approved by the SISSA Ethics Committee board and the National Ministry of Health (Permit Number: 630-III/14) in accordance with the European Union guidelines for animal care (d.1.116/92; 86/609/C.E.).

Cell preparation and culture.

Hippocampal and DRG neurons.

Hippocampal and DRG neurons were obtained from Wistar rats (P2-P3) as described in ref. [16]. In short, the animals were anesthetized with CO_2 and sacrificed by decapitation. The dissociated cells were plated at a concentration of 4×10^4 cells/ml onto glass round coverslips (170 µm in thickness) coated with 0.5 mg/ml poly-D-lysine (Sigma-Aldrich, St. Louis, MO, USA) for 1 h at 37°C and washed 3 times in deionized water. It is fundamental to obtain an optimal adhesion of the cells to prevent detachment in the next step (isolation of the cell membrane). The medium used for hippocampal neurons is in Minimum Essential Medium (MEM) with GlutaMAX supplemented with 10% Fetal Bovine Serum (FBS, all from Invitrogen, Life Technologies, Gaithersburg, MD, USA), 0.6% D-glucose, 15 mM Hepes, 0.1 mg/ml apo-transferrin, 30 µg/ml insulin, 0.1 µg/ml D-biotin, 1 µM vitamin B12 (all from Sigma-Aldrich), and 2.5 µg/ml gentamycin (Invitrogen). The medium used for DRG neurons is Neurobasal medium (Gibco, Invitrogen, Milan, Italy) supplemented with 10% Fetal Bovine Serum (FBS, from Invitrogen, Life Technologies, Gaithersburg, MD, USA).

Rods.

Rod cells were obtained from adult male Xenopus laevis as described in ref. [37]. Under infrared illumination, the eyes of dark-adapted frogs after anesthesia with MS-222 were surgically extracted. Eyes were preserved in the Ringer solution (110 NaCl, 2.5 KCl, 1 CaCl2, 1.6 MgCl2, 3 Hepes-NaOH, 0.01 EDTA, and 10 glucose in mM; pH 7.8 buffered with NaOH), and hemisected under a dissecting microscope. The extracted retina was maintained in the Ringer solution.

NG108-15.

Mouse neuroblastoma NG108-15 cells were obtained from Sigma-Aldrich. The cells were grown in Dulbecco's Modified Eagle Medium (DMEM, ThermoFisher) plus 10% Fetal bovine serum (FBS, Gibco), 100 U/ml Penicillin and 100 U/ml Streptomycin. The cells were cultured into a humidified incubator (5% CO2, 37 °C).

Cell transfection.

NG108-15 cells were transiently transfected with 300 ng of each cDNA expression plasmids by using Lipofectamine 2000 Transfection Reagent (ThermoFisher) according to its handbook. Briefly, mTMEM16A-GFP plasmids (with GFP at their C-terminal) expression vector peGFP-N1 plasmid and the Lipo2000 were diluted into Opti-MEM Reduced Serum Medium (Gibco), respectively. 5 mins later, we added the diluted DNA to the diluted Lipo2000 to make the plasmid DNA-lipid complexes. After incubating 30 min, we plated the cells on the 12 mm round coverslips coated with 1x Poly-L-Ornithine (Sigma-Aldrich) in 12 well plate, and in the meanwhile, we added DNA-lipid complexes to the cells. We performed membrane isolation about 48 hours after transfection.

Isolation of cell membranes.

Single-cell unroofing (for cell types that grow in adhesion).

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 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.

The apical membrane of Hippocampal neurons, DRGs and NG108-15 were isolated with an optimized version of the unroofing method[16]. Briefly, additional empty glass coverslips (24 mm in diameter, 170 µm in thickness) were plasma cleaned for 15 seconds and broken in 4 quarters (with the use of the hands) in order to obtain optically sharp edges, as described in [16]. The coverslip quarters were immersed in 0.5 mg/ml poly-D-lysine for 30 minutes, and then they were immersed in deionized water for 10 seconds before use. A petri dish was filled with Ringer solution (2 ml), where the glass quarter was placed tilted of 7-15 degrees in the middle of it, supported by a 10 x 10 x 1 mm glass slice and Vaseline. The cover of the petri dish was then fixed on the stage of the AFM-inverted microscope setup (JPK Nanowizard 3 on an Olympus IX71).

The cell culture was then mounted on a 3D printed coverslip holder connected to the head stage of the AFM. The AFM head was put on top of the stage in measurement position. The cell culture was immersed into the solution and a target cell was identified and aligned with the underlying corner of the glass quarter. The cell culture was moved towards the corner of the underlying glass with the motors of the AFM until the target cell was squeezed and it doubled its area. At this point the cell is kept squeezed for 3 minutes, then a loaded spring under the AFM is released to abruptly separate the corner from the cell culture, and break the target cell membrane. The glass quarter with the isolated cell membrane was laid down and fixed on the petri dish. The medium was replaced with Ringer's solution without exposing the cell membrane to the air.

Membrane isolation of non-adherent cells.

Cells that do nott grow in adhesion usually do not establish a tight binding with the substrate on top of which they are deposited. For these cells (e.g. rod cells), instead of unroofing, it is more reliable to break the cells with a lateral flux of medium[17].

Isolated and intact rods were obtained by mechanical dissociation of the Xenopus retina in an absorption buffer (150 mM KCl, 25 mM $MgCl_2$, and 10 mM Trizma base; pH 7.5); they were then deposited on cleaved mica as described in ref.[34]. Incubated rods were maintained for 30–45 minutes over the mica in order to be adsorbed by its negatively charged surface. In the meanwhile, the position of the rods in the field of view of the microscope was annotated. The absorption buffer was substituted by a solution containing (in mM): 150 KCl, 10 Tris-HCl, (pH 7.5) and then a lateral flux of medium was applied to the rods until all the cell bodies were removed.

Isolation of rod discs.

Purification techniques with multiple centrifugations are usually required to isolate membrane-only organelles like rod discs or outer membrane vesicles[13]. Rod discs were obtained starting from the extracted retina as described in ref. [34]. Briefly, discs were separated with two series of centrifugations of the sample overlaid on a 15-40% continuous gradient of OptiPrep (Nycomed, Oslo, Norway). 40 µl of the sample were diluted with 40 µl of absorption buffer, and incubated on freshly cleaved mica for 40 minutes. After 40 minutes, the incubation medium was removed and substituted with the solution used in the AFM experiments (150mM KCl, 10mM Tris-HCl, pH 7.5).

AFM imaging and Single-Molecule Force Spectroscopy (SMFS).

AFM experiments was performed using an automated AFM (JPK Nanowizard 3) with 50 µm long cantilevers (AppNano HYDRA2R-NGG, nominal spring constant = 0.84 N/m). We calibrated the AFM cantilevers in the experimental medium before each experiment using the equipartition theorem[38]. The AFM experiments of Hippocampal neurons and DRGs were performed with Ringer's solution (NaCl 145 mM, KCl 3 mM, CaCl2 1.5 mM, MgCl2 1 mM, HEPES 10 mM, pH 7.4); Rod membrane and discs experiments were performed with 150mM KCl, 10mM Tris-HCl, pH 7.5. All experiments were performed at 24 Celsius.

AFM imaging.

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 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.

The position of the cells before unroofing was annotated in the monitor of the computer in order to start the AFM imaging where the cells was in contact with the substrate (cell membrane is not visible in bright-field). The membrane obtained with single-cell unroofing (hippocampal neurons and DRG) can easily be found in proximity of the glass corner (~80% success rate). In the case of the rod membrane (non-adherent cells), usually different positions need to be scanned before finding a patch of membrane. Rod discs can be identified only via AFM imaging. We performed imaging both in contact mode (setpoint ~0.4 nN) and intermittent-contact mode (lowest possible), but the intermittent-contact mode if preferable because it does not damage the border of the patches of membrane.

AFM-based SMFS (protein unfolding)

we performed automated SMFS on top of the imaged membranes by setting grid positions for the approaching and retraction cycles of the cantilever. All experiments were performed with a retraction speed of 500 nm/s. The membrane proteins present in the sample were attached non-specifically to the cantilever tip by applying a constant force of ~1 nN for 1 second between the AFM tip and the cytoplasmic side of the membrane. This method proved to work with different membrane proteins[14,24,39], and to allow a higher throughput compared to methods that involve a specific attachment between the tip and the protein[18,40–42]..

Automatic classification of SMFS data.

The selection of the F-d curves that represent the unfolding of membrane proteins is usually based on the search for a specific pattern of unfolding in the SMFS data, after a filtering based on the length of the protein under investigation[26,27]. In the case of a native preparation (like ours) that contains a mixture of unknown proteins a) the filtering based on the distance cannot be applied and b) the number of specific patterns to be found is unknown. In order to find recurrent patterns of unfolding in a SMFS dataset we developed an algorithm that consists of five major blocks (Supplementary Fig. 3 a). In the first block, the parts of the F-d curves not related to the unfolding process are removed, and a coarse filtering aimed at the detection of spurious traces is performed. In the second block, a quality score based on the consistency of the experimental data with the worm-like chain (WLC) model is computed and assigned to each trace. This score is used to select physically meaningful traces for further analysis. In the third block, distances between pairs of traces are computed to assess their similarity. The distances are used in the fourth block for density peak clustering. The fifth and final block consists in the refinement and possibly in the merging of some of these clusters. In what follows we provide a detailed overview of each block.

Block 1: filtering.

The standard F-d curve preprocessing was applied to all the data within 'Fodis'[43]. The zero of the force of the curve was determined averaging the non-contact part (baseline after the final peak) and subtracted to all the points of the curve. The piezo position was transformed in tip-sample-separation considering the contribution of the bending of the tip to the extension of the polymer. Given that the F-d curves are subject to noise (due to thermal fluctuations, coming from the instrument, etc.), we smooth the original signals through interpolation on a grid with width δ_{interp}= 1 nm.

A curve is discarded if it does not contain a:

- detectable contact point (i.e. a transition from negative forces to positive forces in respect to the baseline set at zero force);
- if the points occupy force ranges over 5000 pN;

Some of the F-d curves show deviations from the horizontal zero-force line in the non-contact part (wavy final part due to imperfect detachment of the polymer or other noise from the environment). We detect and discard these traces by computing the standard deviation of the tails from the zero-force line. If it exceeds two times σ_{NOISE} (average standard deviation of the baseline of the batch of curves) the trace is discarded.

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 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.

Block 2: Quality score.

The quality score is used for refine selection of traces with high information content vs. noisy traces. It is based on the description provided by the worm-like chain (WLC) model, which is the standard model in the analysis of SMFS data[33]. The WLC model implies the equation:

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

where F is force, x is extension, k_B is Boltzmann's constant, T is temperature, l_p is persistence length and L_c is contour length. Each unfolding curve in the trace is fitted with the WLC equation and a L_c value, corresponding to the length of the unfolded protein domain is obtained. The L_c values are computed by solving equation (1) for each x and F. An appropriate value for the persistence length l_p for membrane proteins is 0.4 nm as reported in ref. [33]. The WLC model is applicable in the force range 30-500 pN[44].

Once we compute the L_c values, we can build a L_c histogram. Normally, the L_c histogram describing a successful unfolding experiment is characterized by the presence of a few maxima separated by deep minima. We implement these features in the definition of our quality score to distinguish meaningful F-d curves.

An important parameter is the bin width of the L_c histogram. If the bin width is too small the histogram is noisy; if the bin width is too large, peaks corresponding to the unfolding of different domains might be merged. We use bin width 8 nm which is an efficient value for evaluating the goodness of a curve and it allows to consider also curves that deviate from the WLC model (l_p =0.4 nm) but that contain information. Furthermore, the choice of such large bin width is based on visual inspection of the histograms of proteins with known structure. Once the L_c histogram is built, we detect all maxima and minima. A maximum is meaningful if it is generated by more than 5 points and it includes more than 1 % of the force measures of a trace.

For each maximum in the L_c histogram, we compute a score W quantifying the consistency of the peak with the WLC model. A high-quality peak is clearly separated from other peaks of the histogram, therefore it should be surrounded by two minima. We define $f_{left} = \frac{P_{left}}{P_{max}}, f_{right} = \frac{P_{right}}{P_{max}}$ where P_{max} , P_{left} and P_{right} are the probability densities of the maximum, of the left and the right minima. Ideally, $f \sim \frac{1}{2}\left(f_{left} + f_{right}\right)$ should go to 0. We define the peak score as $W = exp(-2f^2)$. According to this definition, if $P_{left} = 1, P_{right} = 2$ and P_{max}=16, W=0.98. Whilst if $P_{left} = 13, P_{right} = 14$, the peak doesn't fit well with the WLC model and W=0.24.

Once a score is computed for each relevant peak in the L_c histogram, that score is assigned to all points in the corresponding trace. This is accomplished in two steps: first, the peak score is assigned to all points in the histogram belonging to that peak. Second, to all points with force values below 30 pN, for which an L_c values cannot be computed due to the model's limitations. To these points, we assign the score of the first successive point with force larger than 30 pN. This criterion applies only to points within 75 nm from the last point assigned to the peak. The peak width value is selected by visual inspection of traces, evaluating the maximum width of their force peaks.

The quality score of a trace, S_w, is the sum of the scores for all points in the trace. The higher the global score, the higher the trace quality. We use the ratio between the quality score and the trace length to select high quality traces. If this ratio is below 0.5, we discard the trace. We assume that if more than half of the trace is inconsistent with the WLC model, it is a low-quality trace and as such we exclude it from the analysis. While if more than half of the trace is in good agreement with the WLC model, it is possibly a meaningful trace.

We point out that the goal of blocks 1–4 is only to find dense recurrent patterns in the SMFS data: in block 5 we reevaluate the F-d curves to form the selections shown in Fig. 3 of the main text.

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 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.

Block 3: Computing distances

In block 3 we quantify the similarity between the traces in order to find the recurrent pattern of unfolding within the data. To accomplish this goal, we use a modified version of the distance introduced by Marsico et al.[26]. This distance is defined using the dynamic programming alignment score computed for a pair of traces. For two traces, *a* and *b*, the distance d_{ab} is simply:

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

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} is the maximum length between the two. We have modified the match/mismatch scoring function used by Marsico et al as follows:

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

where $F_a(i)$ and $F_b(j)$ are the forces in points *i* and *j* in traces *a* and *b*, and $F_{scoring} = 4\sigma_{NOISE}$. In the work done by Marsico et al, $F_{scoring}$ is replaced by ΔF_{max}, which is the average of the maximum force values in the two traces. When two widely different traces have high ΔF_{max} their distance will be lower with respect to two traces with low ΔF_{max} but overall higher level of similarity. Namely, the distance magnitude depends on the ΔF_{max} value and traces with high ΔF_{max} have by definition lower distance values. It is important to note that this problem did not occur in Marsico's work since the ΔF_{max} values were uniformly distributed for all traces.

In order to gain computational efficiency, the distance is computed only for traces which differ by no more than 2 peaks in the L_c histograms or by no more than 20 % in their trace length difference.

Block 4: Density peak clustering.

The density peak clustering (DPC) algorithm[29] is used for clustering. This choice is appropriate given that a fraction of traces in the analyzed datasets correspond to statistically isolated events and DPC automatically excludes the outliers. DPC can be summarized in the following steps:

1. We compute the density of data points in the neighborhood of each point using the *k*-nearest neighbor (*k*-NN) density estimator[45]. The density is 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[46], 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 it is the density rank which is relevant for the final cluster assignation. Therefore, without loss of generality, we compute the density using the following equation:

$$\rho_i = -\log r_{k,i} \tag{5}$$

$\tilde{\rho}_i$ and ρ_i are related by a simple monotonic transformation and thus, have the same rank. By using equation (4) we don't have to compute the intrinsic dimension of the dataset. In order to assign bigger weight to high quality traces, we multiply ρ_i by the score-length ratio of trace *i*.

2. Next, we find the minimum distance between point *i* and any other point with higher density, denoted as δ_i:

$$\delta_i = \min_{j:\rho_j > \rho_i} d_{ij} \tag{6}$$

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 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.

where d_{ij} is the distance between points *i* and *j*. δ_i is used to identify the local density maxima.

3. We identify the cluster centers as density peaks, e.g. points with high values of both ρ_i and δ_i. For each point we compute the quantity $\gamma_i = \rho_i \delta_i$. Points with high values of γ_i are good cluster center candidates. We sort all points by the value of γ_i in descending order. The first point is a cluster center. The second point is a cluster center unless its distance from the first point is smaller than $r_{cut} = 0.3$ (which represents the distance below which on average two traces are considered as the same pattern). Regarding the third point, it is a cluster center if it is at a distance smaller than r_{cut} from the preceding two points. Following the same logic, all the points are assessed and all cluster centers are identified.
4. All points that are not cluster centers are assigned to the same cluster of the nearest point with higher density.

Block 5: Refinement and merging

The previous blocks, from 1 to 4, were optimized for finding the centers of dense patterns of unfolding in the SMFS data, but not for finding the borders of the clusters. To solve this issue, i.e. finding the F-d curves that are similar to each pattern of unfolding, we used the conventional definition of similarity (degree of superposition of F-d curves in the Force/tip-sample-separation plane) automated in the Fodis software in the tool 'fingerprint_roi'[43].

In brief, we superimposed each cluster center with its two closest neighbors creating the effective 'area of similarity' (AoS) for each cluster. The AoS is defined as the area generated by all the points of the three curves above 30 pN and before the last peak (see Supplementary Fig. 3 b), each point forming a square of 5 nm x 5 pN. Then, the SMFS curves are preliminary filtered based on their length with their final peak falling between 0.7 x *L* and 1.3 x *L* (with *L* length of the cluster center). Each of the remaining F-d curves is compared with the AoS, and the number of its points that fall within the AoS is annotated: this number constitutes the similarity score. As depicted in Supplementary Fig. 3 c, the plot of the scores in descending order interestingly forms a line with two different slopes. The change of the slope empirically defines a threshold that reflects the limit of similarity for each cluster. If two clusters share more than 40% of the traces above the threshold, they are considered the same cluster, thus merged (all the merges are reported in Supplementary Fig. 3 d).

Bayesian identification of F-D curves.

Bayesian inference is widely used in modern science[47,48] because it allows to univocally determine the level of uncertainty of a hypothesis[49]. We used the same framework to determine the molecular identity of the unfolding clusters. In the most general terms, we observed the unfolding cluster C_X, and we want to find the probability that the unfolding of a certain protein $Prot_A$ corresponds to the unfolding cluster C_X, i.e. we want to find the posterior probability $P(Prot_A|C_X)$. In the form of the Bayes theorem:

$$P(Prot_A|C_X) = \frac{P(C_X|Prot_A)P(Prot_A)}{P(C_X)} \tag{7}$$

where $P(Prot_A)$ is the prior, i.e. the probability of $Prot_A$ to be in the sample; $P(C_X|Prot_A)$ is the likelihood, i.e. the probability to find a cluster with the features of C_X coming from the unfolding of $Prot_A$; and $P(C_X)$ is the normalizing factor. In the case of a classical experiment with a single purified protein, $P(Prot_A|C_X)$ is assumed to be equal to 1, but this is not the case for a native environment where there are $Prot_B$, $Prot_C$, etc.

The observables of an unfolding cluster for which we determined the likelihood functions are the contour length of the last detectable peak $Lc_{max,Cx}$ (~ length of the F-d curve), and the average unfolding force of the detected peaks $\bar{F}_{Cx}$, but the method is modular therefore it could incorporate also other observables. The equation (7) becomes:

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 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.

$$P\left(Prot_A \middle| Lc_{max,Cx}, \bar{F}_{Cx}\right) = \frac{P\left(Lc_{max,Cx}|Lc_{Prot_A}\right) P\left(\bar{F}_{Cx}|\bar{F}_{Prot_A}\right) P(Prot_A)}{N} \quad (8)$$

where $N = \sum_i (P\left(Lc_{max,Cx}|Lc_{Prot_i}\right) P\left(\bar{F}_{Cx}|\bar{F}_{Prot_i}\right) P(Prot_i))$ is the normalizing factor that takes into consideration all the proteins $Prot_i$ present in the sample. In the next paragraph we will describe the determination of the numerator of equation (8).

Determination of prior probability $P(Prot_A)$

The most crucial part of the method is the determination of the list of proteins present in the sample, together with all their properties (length, abundance, secondary structure, topology, etc.). To do so we combined the Mass Spectrometry results of the cells under investigation[30–32] with other structural and topological information available in Uniprot and PDB. The crossing of the databases is done thanks to the unique Uniprot identifier. The complete list of proteins of Hippocampal neurons, Rod outer segments and Rod discs with the information necessary for the Bayesian inference are shown in the Supplementary Data of the article. In case the data of the species of interest are not available, cross species proteomic analysis demonstrated that the majority of proteins are conserved in terms of relative abundance[50,51].

$P(Prot_A)$ is the probability of finding $Prot_A$ and not $Prot_B$, $Prot_C$, etc., which corresponds to the normalized relative abundance of $Prot_A$ in the sample—a parameter that is usually calculated in mass spectrometry analysis. Indeed, *in silico* calculation of abundances gives rather trustworthy values:

1. the most accurate option is the emPAI[52];
2. if the emPAI is not available, the second best option is the spectra counting for each peptide (PSM)[53];
3. if the PSM is not available, the sequence coverage can be used as loose estimation[54].

We used the emPAI for hippocampal neurons and rods; for the discs, the emPAI does not give accurate values because of the extreme concentration of Rhodopsin, therefore we used the abundances obtained with other quantitative methods[55].

We demonstrated in Supplementary Fig 1 that the isolated paches of membrane contain the membrane proteins of the original cells but not the cytoplasmic proteins, therefore we created an additional binary variable $ismembrane$ for each protein equal to 0 if the protein is not a membrane protein, 1 otherwise. This information is extracted from the annotation in the Uniprot database. The final prior probability is:

$$P(Prot_A) = abundance_A \times ismembrane_A \quad (9)$$

Determination of the Likelihood function $P\left(Lc_{max,Cx}|Lc_{Prot_A}\right)$

The F-d curves encode a reliable structural information, that is the total length of the unfolded protein[18]. We revised 14 published unfolding clusters of membrane proteins[12,18,20,24,25,39,41,42,56–60] that allowed us to create the likelihood function for the observable $Lc_{max,Cx}$ as shown in Fig. 4 b–c. This likelihood is a Gaussian centered at $0.89Lc_{Prot_A}$ with a standard deviation of $0.05Lc_{Prot_A}$.

Determination of the Likelihood function $P\left(\bar{F}_{Cx}|\bar{F}_{Prot_A}\right)$

The force necessary to unfold a protein domain depends on the stability of the domain itself. α-helices and β-sheets are unfolded at different force levels as shown in Fig. 4 d. We revised the unfolding forces of 32 proteins and we used as $P\left(\bar{F}_{Cx}|\bar{F}_{Prot_A}\right)$ the smoothed trend line of the distribution (Fig. 4 e of the main text).

All the data and the Matlab functions for the Bayesian inference are present in the Supplementary Data of the article.

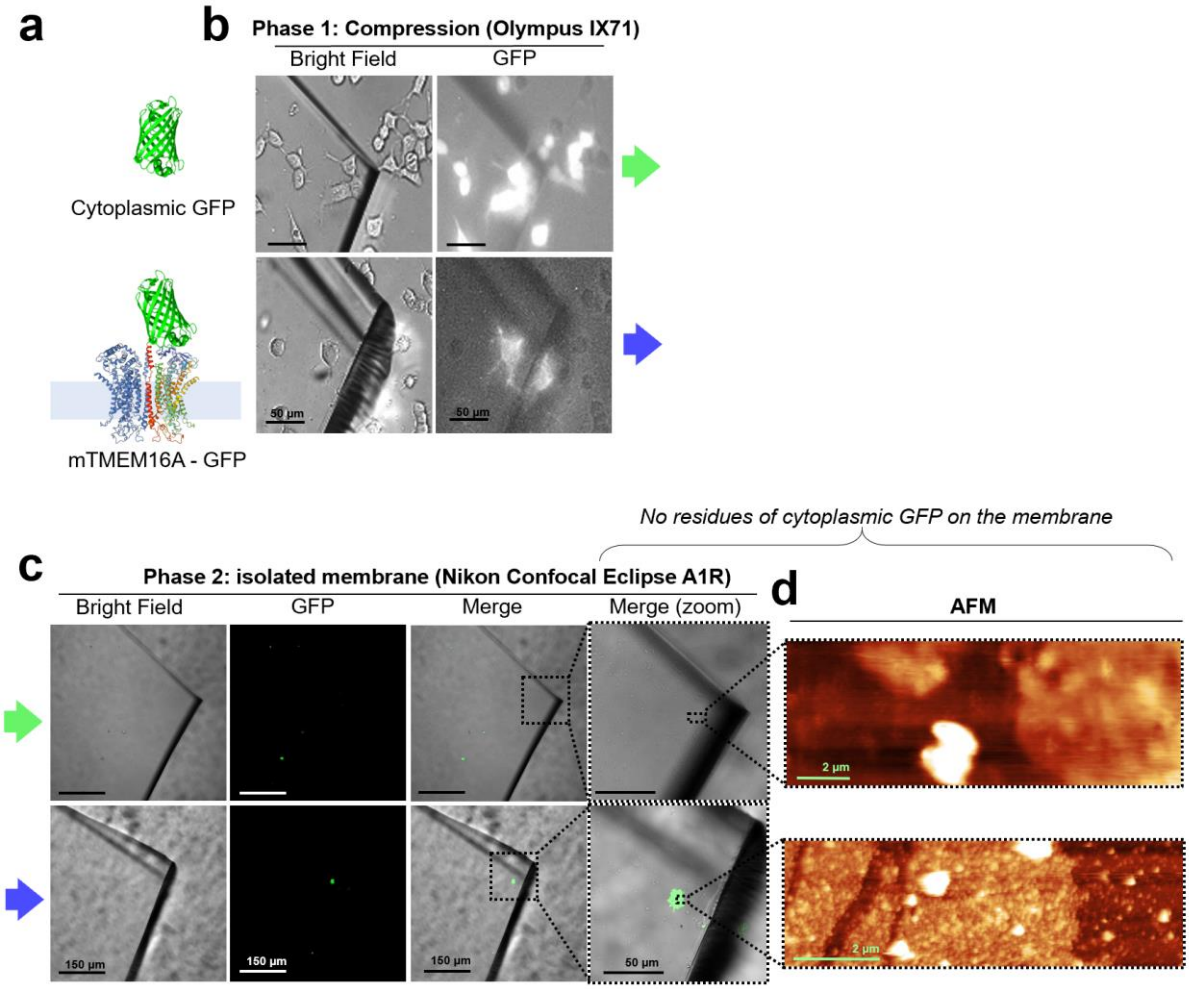
bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

41. Cisneros, D. A., Oesterhelt, D. & Müller, D. J. Probing Origins of Molecular Interactions Stabilizing the Membrane Proteins Halorhodopsin and Bacteriorhodopsin. *Structure* **13**, 235–242 (2005).
42. Kedrov, A., Ziegler, C., Janovjak, H., Kühlbrandt, W. & Müller, D. J. Controlled Unfolding and Refolding of a Single Sodium-proton Antiporter using Atomic Force Microscopy. *Journal of Molecular Biology* **340**, 1143–1152 (2004).
43. Galvanetto, N. *Fodis: a Software for Single Molecule Force Spectroscopy*. (2018).
44. Petrosyan, R. Improved approximations for some polymer extension models. *Rheol Acta* 1–6 (2016). doi:10.1007/s00397-016-0977-9
45. Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician* **46**, 175–185 (1992).
46. Facco, E., d'Errico, M., Rodriguez, A. & Laio, A. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports* **7**, (2017).
47. Wang, Q. *et al.* A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data. *Nature Neuroscience* **22**, 691–699 (2019).
48. Heenan, P. R. & Perkins, T. T. FEATHER: Automated Analysis of Force Spectroscopy Unbinding and Unfolding Data via a Bayesian Algorithm. *Biophysical Journal* **115**, 757–762 (2018).
49. Jaynes, E. T. *Bayesian Methods: General Background*. (1986).
50. Bayram, H. L. *et al.* Cross-species proteomics in analysis of mammalian sperm proteins. *Journal of Proteomics* **135**, 38–50 (2016).
51. Wright, J. C., Beynon, R. J. & Hubbard, S. J. Cross Species Proteomics. in *Proteome Bioinformatics* (eds. Hubbard, S. J. & Jones, A. R.) **604**, 123–135 (Humana Press, 2010).
52. Ishihama, Y. *et al.* Exponentially Modified Protein Abundance Index (emPAI) for Estimation of Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein. *Mol Cell Proteomics* **4**, 1265–1272 (2005).
53. Liu, H., Sadygov, R. G. & Yates, J. R. A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics. *Anal. Chem.* **76**, 4193–4201 (2004).
54. Florens, L. *et al.* A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419**, 520 (2002).
55. Milo, R. & Phillips, R. *Cell biology by the numbers*. (2016).
56. Kawamura, S. *et al.* Kinetic, Energetic, and Mechanical Differences between Dark-State Rhodopsin and Opsin. *Structure* **21**, 426–437 (2013).
57. Möller, C. *et al.* Determining molecular forces that stabilize human aquaporin-1. *Journal of Structural Biology* **142**, 369–378 (2003).
58. Ge, L. *et al.* Molecular Plasticity of the Human Voltage-Dependent Anion Channel Embedded Into a Membrane. *Structure* **24**, 585–594 (2016).
59. Bosshart, P. D. *et al.* The Transmembrane Protein KpOmpA Anchoring the Outer Membrane of Klebsiella pneumoniae Unfolds and Refolds in Response to Tensile Load. *Structure* **20**, 121–127 (2012).
60. Klyszejko, A. L. *et al.* Folding and Assembly of Proteorhodopsin. *Journal of Molecular Biology* **376**, 35–41 (2008).

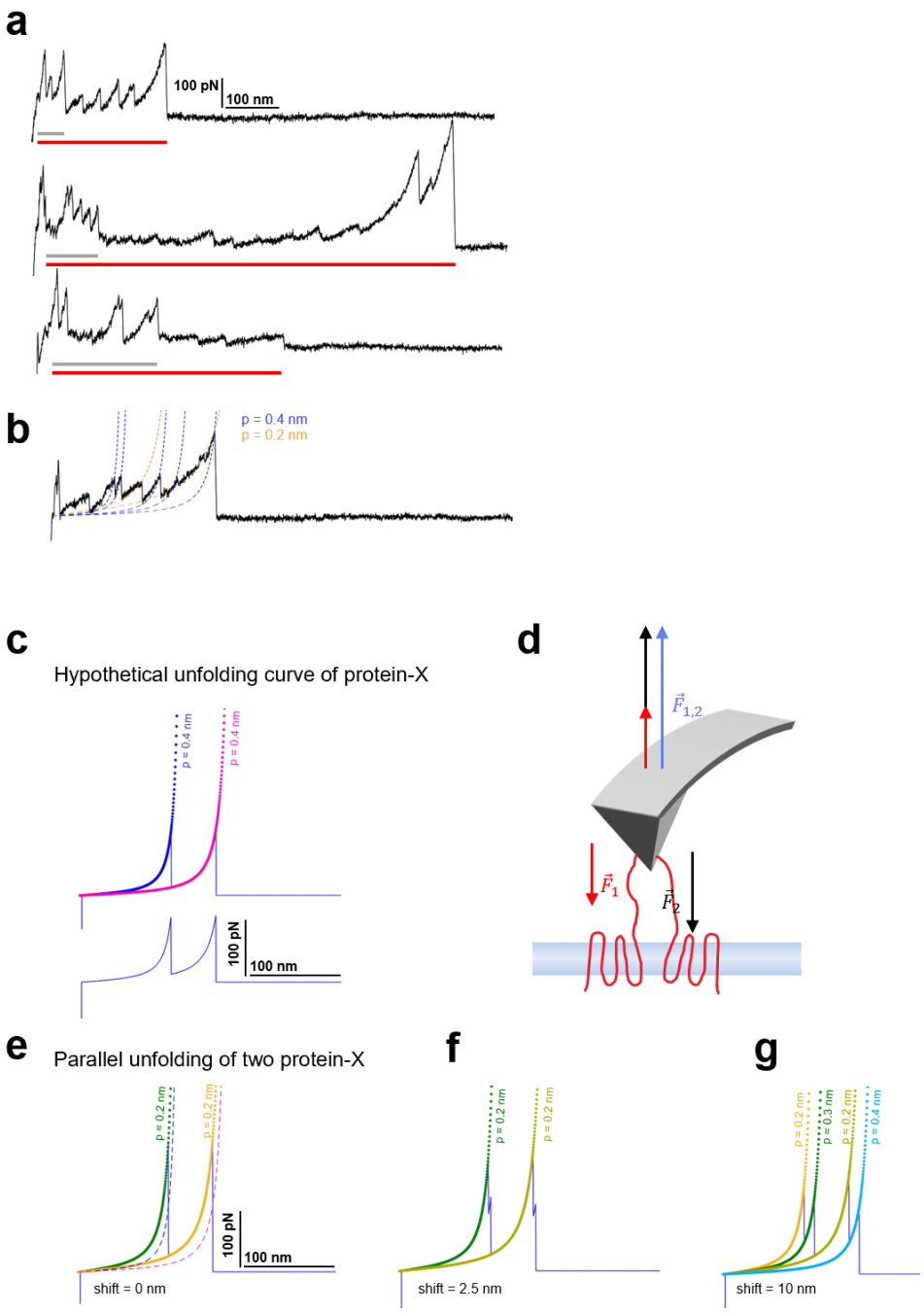
bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Figures



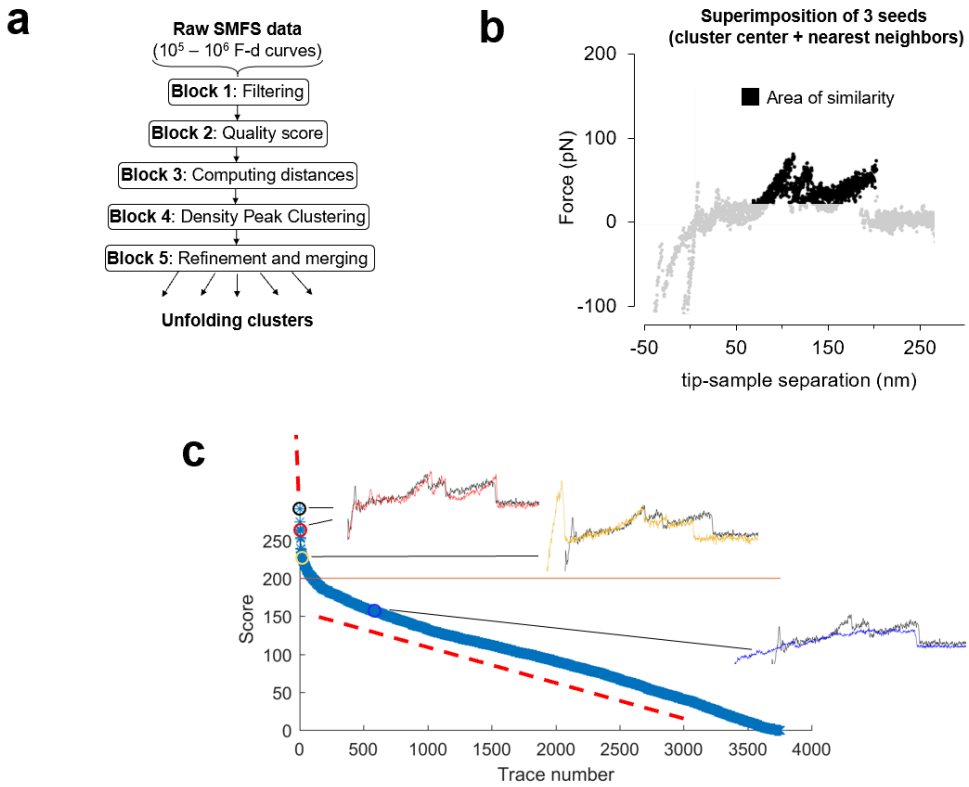
Supplementary Figure 1
Membrane proteins remain in membrane after unroofing, cytoplasmic proteins don't.
a, (top) cytoplasmic GFP and (bottom) mTMEM16A-GFP overexpressed in NG108-15 cells. **b**, bright field and fluorescence images taken during the compression on the inverted microscope-AFM system. **c**, images of the coverslip quarter only after the unroofing process taken with the confocal microscope; the rough surface under the glass is due to the Vaseline layer used to fix the coverslip quarter. **d**, AFM images of the area in **c** showing the presence of membrane parches in both cases.

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 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.



Supplementary Figure 2
Candidates of multiple unfolding and origin of persistence length deviation.
a, observed F-d curves with the features of multiple unfolding events shown in Fig 3 g (red: long protein, gray: short protein). **b**, F-d curve with intra-deviations of persistence length. **c**, hypothetical unfolding curve of protein-X (peak 1: Lc=100 nm, F=150pN; peak 2: Lc=150 nm, F=150pN) fitted with the WLC model with standard persistence length p=0.4nm. **d**, the force applied by the AFM tip balances the unfolding forces of the two proteins during the retraction. **e**, the effective F-d curve recorded during parallel unfolding of two protein-X corresponds to the sum of a single unfolding curve **c** and is best fitted with p=0.2nm. **f**, relative shift of 2.5 nm and **g**, 10 nm still result in deviations of the measured persistence length and display the doubling of the peaks.

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 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.

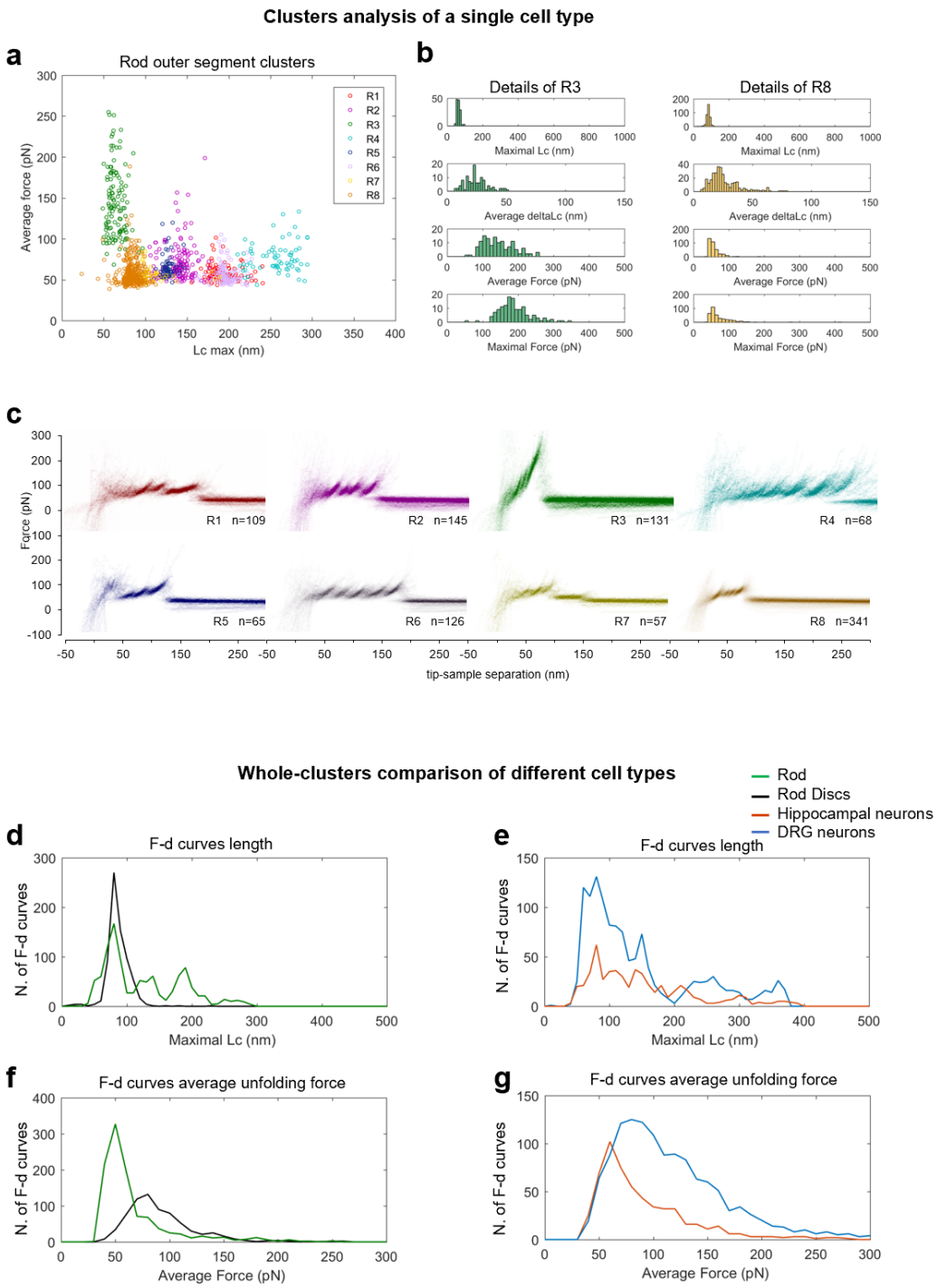


d

DRG	C. merged	Hippo.	C. merged	Rod	C. merged	Disc	C. merged
drg1	1	h1	1	r1	2	disc1	1
drg2	1	h2	1	r2	2	disc2	4
drg3	2	h3	7	r3	2	disc3	1
drg4	1	h4	2	r4	2	disc4	3
drg5	1	h5	1	r5	2	disc5	1
drg6	3	h6	1	r6	1		
drg7	1	h7	1	r7	1		
drg8	1	h8	1	r8	1		
drg9	1	h9	4				
drg10	1	h10	1				
drg11	1						
drg12	1						
drg13	1						
drg14	1						
drg15	1						

Supplementary Figure 3
Clustering.
a, block scheme of the clustering method. **b**, area of similarity (AoS) for cluster R1 used for block 5. **c**, plot of the scores in descending order. **d**, table showing the number of clusters that was merged to form the final selection of Fig. 3.

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 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.

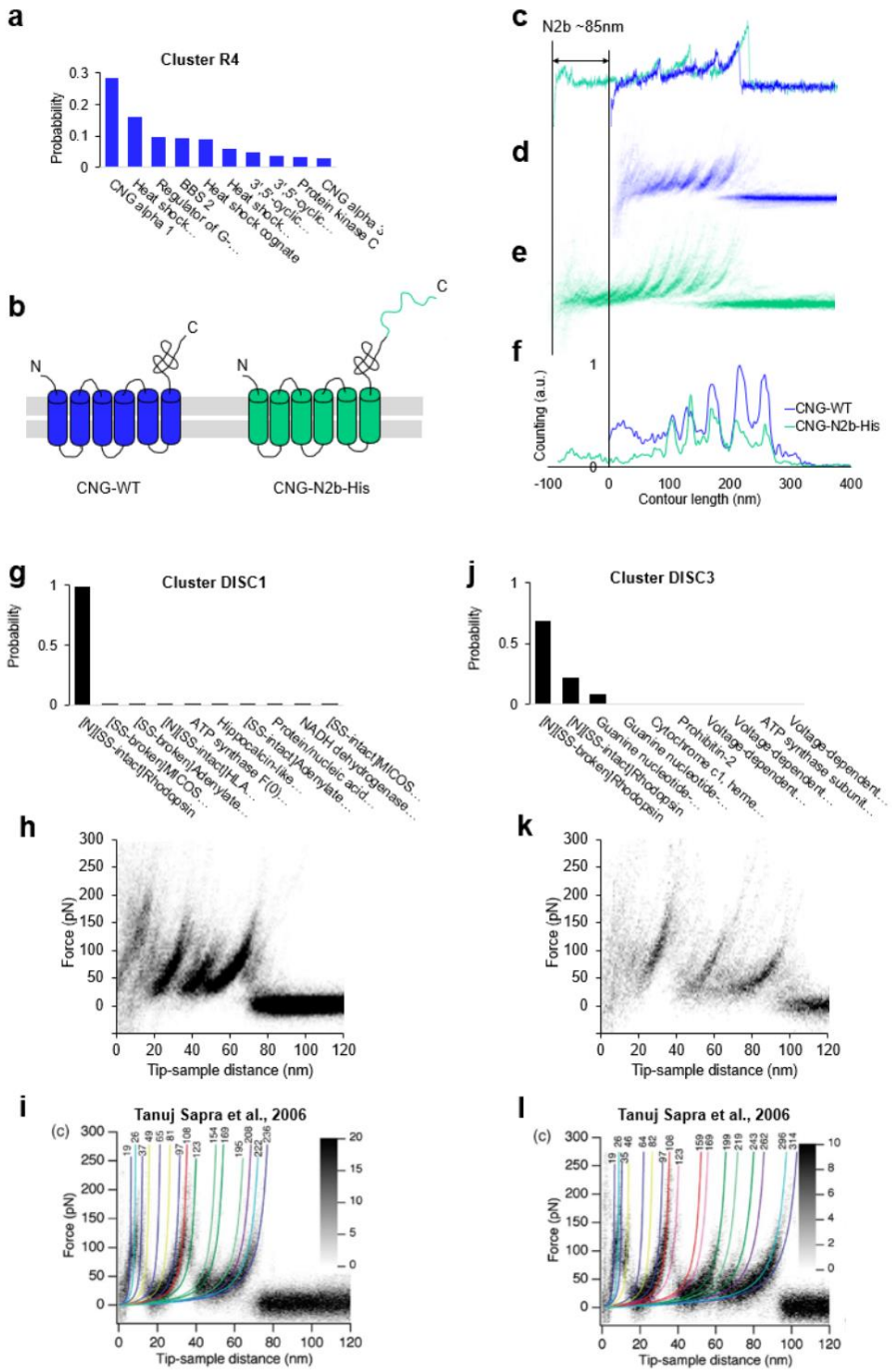


Supplementary Figure 4

Alternative visualizations for clusters analysis.

a, plot of all the F-d curves belonging to the clusters of the rod outer segment plotted with different colors in the 'average unfolding force' vs 'maximal contour length' space. **b**, distribution of representative observables for clusters R3 and R8. **c**, density plots of clusters in **a**. Comparison of the maximal contour length profiles (**d** and **e**) and of the average unfolding forces (**f** and **g**) of the four cell types investigated.

bioRxiv preprint doi: https://doi.org/10.1101/732933; this version posted September 26, 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.



Supplementary Figure 5

Orthogonal validations of the Bayesian identification.

a, identification probability for cluster R4. **b**, scheme of the wild type CNG and of the engineered CNG linked to the N2B (unstructured amino acid chain). **c**, representative F-d curves obtained in rod outer segments (blue) and after overexpression of the CNG-N2b-His (green). **d,** density plot of R4 (n=68) and **e,** density plot of the cluster obtained after overexpression of CNG-N2b-His (n=64). **f**, comparison of the global histograms of **d** and **e** in the contour length space. **g** and **j**, identification probabilities of cluster DISC1 and DISC3 and their density plots (**h** and **k**). **i**, unfolding of rhodopsin identified by Tanuj Sapra et al. (2006) with intact SS-bond and **l**, with broken SS-bond (figure adapted from Tanuj Sapra et al. (2006)).