UCL Université catholique de Louvain

École polytechnique de Louvain (EPL)

epl

Force-Distance Curves Analysis in Single-Molecule Force Spectroscopy

Dissertation presented by
Bruno DEGOMME

for obtaining the Master's degree in
Mathematical Engineering

Supervisor
Christophe DE VLEESCHOUWER

Readers
David ALSTEENS, Laurent JACQUES

Academic year 2017-2018

Abstract

Single-Molecule Force Spectroscopy is an application of Atomic Force Microscopy, and is used to mechanically unfold proteins, producing Force-Distance (FD) curves. These FD curves can then be analyzed and fitted to identify unfolding pathways and intermediate states.

In this thesis we analyze a set of a hundred FD curves of the LmrP integral membrane protein. The main factor complicating this analysis is that, for experimental reasons, there is a different and unknown offset on the distance of each FD curve.

We start out by fitting each FD curve with a Worm-Like Chain model, then cluster and align them to identify a main unfolding pathway and two tentative intermediate states.

Acknowledgments

This Master Thesis wouldn't have been possible without the invaluable advice of my thesis advisor, Prof. Christophe de Vleeschouwer (ICTM, UCLouvain). Discussing ideas with him nearly each week was a true motivation booster. Thank you for your great availability and earnest curiosity.

Many thanks also go to Prof. David Alsteens (LIBS, UCLouvain) for coming to us with this interesting problem, starting a fruitful interdisciplinary collaboration. I would also like to thank you and Dr. Melanie Köhler for taking the time to answer our many questions. We learned a lot in the process.

Contents

1. Introduction

Cell membrane proteins play essential roles in many cellular processes like photosynthesis, cell-cell adhesion and transport of ions [4]. The way they fold and unfold can vary, and is linked to their ability to play those roles. Their misfolding can cause human diseases like *cystic fibrosis* and *retinis pigmentosa* [2].

A protein doesn't unfold in one go, but transitions between multiple intermediate states. At each one of these intermediate states, part of the protein is still folded while part of it is already unfolded. The unfolding pathway of a protein is the succession of its intermediate states, from the completely folded to the completely unfolded protein[9].

1.1. Single-Molecule Force Spectroscopy

Single-Molecule Force Spectroscopy (SMFS) is an application of Atomic Force Microscopy (AFM), used to determine the unfolding pathway of proteins. Simplifying *heavily*, a very fine AFM tip located at the end of a cantilever is approached to a protein, and sticks to one of its ends. The cantilever is then pulled up, stretching and unfolding the protein. The force exerted on the tip is determined based on the deflection of the cantilever, and recorded as the protein stretches and unfolds. This force can then be plotted against the distance, yielding a Force-Distance (FD) curve. See figure 1 for an illustration of this process.

Typically, this curve will look like a succession of peaks. First, the force exerted on the tip will grow as we stretch the loose end of the protein (the rest of the protein remains completely folded at this point). This goes on until the pulling force is high enough to induce the unfolding of another protein section. The amino-acids that were part of this section then join the loose end, leading to a sharp drop in the pulling force. We thus reach a new intermediate state. This process of stretching and unfolding sections goes on until the protein leaves the surface on which it lays, at which point the pulling force drops to zero. The sections unfold sequentially, starting from the tip [6].

The distance between two successive peaks is strongly dependent on the length of the protein section that unfolded at the end of the first one. If the amino-acid structure of the protein is known, then we can use these lengths to determine which amino-acids are present in each protein section. Hence we can characterize the whole unfolding pathway using the FD curve obtained through SMFS.

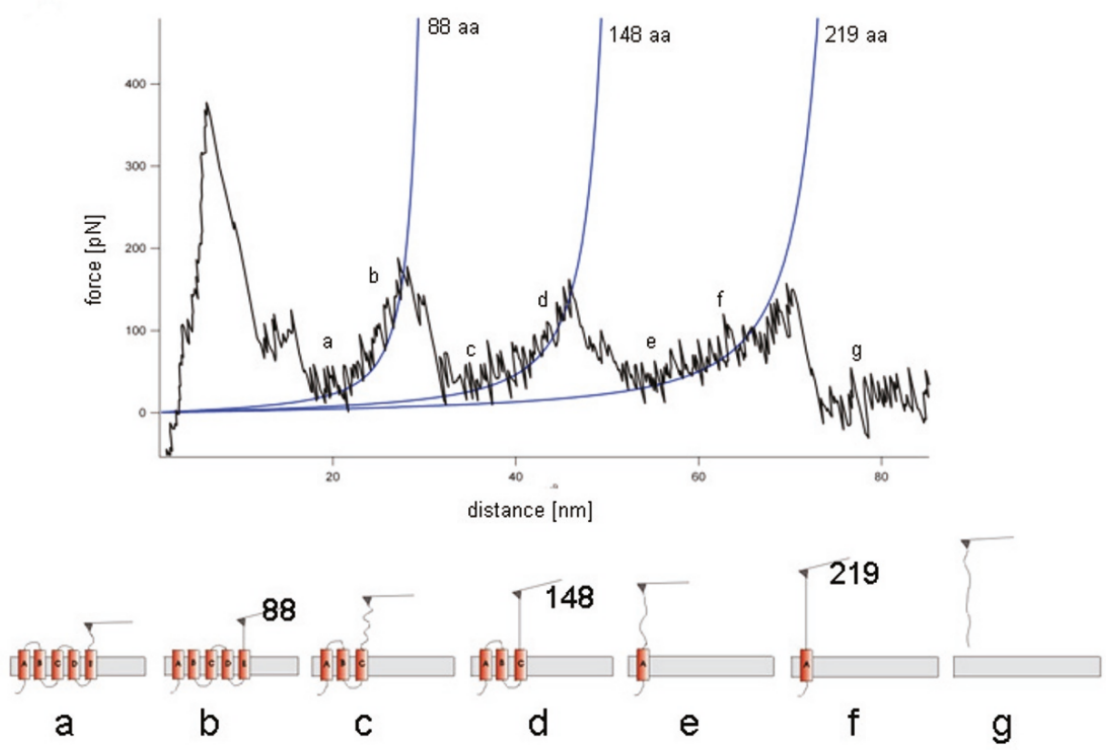


Figure 1: **FD curve and illustration of the unfolding process**. This is the typical FD curve we get when unfolding bacteriorhodopsin. The peaks (top) can be correlated to the length of the corresponding unfolding sections (bottom). Source : [4]

In state-of-the art protocols, bare AFM tips are used and consequently they are pressed with high force (around 500-1000 pN) onto the membrane protein to ensure attachment before unfolding it. While such high forces are mandatory to enable adsorption of an exposed part of the polypeptide to the tip, it also likely induces some conformational change within the receptor. The group of Prof. David Alsteens (LIBS, UCLouvain) has therefore improved the method with the aim of reducing the force applied and increasing the overall throughput efficiency. To this end, they use functionalized AFM tips, i.e. NTA-terminated tips that target with high-affinity the terminal His-tagged end of membrane protein. Preliminary results have been obtained to unfold LmrP from the C-terminal end. NTA-tips will allow to reduce the applied force below the threshold of intramolecular interactions ($\simeq$ 50-100 pN), reducing membrane protein perturbation. Preliminary results also suggested that it increases efficiency by a factor of 10.

1.2. Problem

In this thesis we will focus on analyzing a specific dataset of a 100 FD curves of the LmrP integral membrane protein. We will attempt to identify the most common unfolding pathway in this data, as described by the successive protein section lengths. We will also check if there are any secondary unfolding pathways, or intermediate states that aren't part of the main unfolding pathway but that are shared by multiple proteins.

Two factors make this problem hard. First, there is substantial noise on the recorded force, mostly due to the thermal motion of the cantilever [4].

Second, the origin of a FD curve is defined so that the mean recorded force is zero once the protein has left the surface, and the distance is zero when the tip reaches the surface on which the proteins lays (detected by a sharp increase in the recorded force).

On the distance side of things, this definition is problematic : each FD curve will end up shifted by a (presumably small) random amount. That is because the end of the protein attached to the tip isn't always attached to exactly the same position on the tip, and the other end of the protein is rarely attached to the surface right under the tip. Hence, our FD curves do not share the same origin, which prevents us from easily comparing them to find commonalities. We will need to find some way to align them before we can compare them.

1.3. Analysis

Our analysis will proceed as follows : we start out by defining a clear terminology for concepts used in this dissertation (2). We preprocess the raw FD curves, removing the first and last part of the FD curve which are irrelevant to our analysis (A). We explain a commonly accepted model for the behavior of unfolding proteins (3), and try different techniques to fit this model to each of our FD curves (4,5). We also try to use this model to align all FD curves to the same origin (6), but end up abandoning this approach because of its lack of robustness.

This model gives us a low-dimensional description of each FD curve in term of the lengths of its successive unfolding sections (or equivalently, of its intermediate states). We use a clustering technique on these descriptions to identify the main intermediate states and unfolding pathway (7). We check for the existence of a common secondary pathway but fail to find it. We use the main unfolding pathway as a template to align all FD curves to the same origin (8). And use these aligned FD curves to find two tentative secondary intermediate states (9). We conclude by discussing the robustness of our results(10).

2. Terminology

Before we attempt to fit such a model to real world data, lets define and disambiguate some terms to avoid confusion. We call

- a single record of a distance and corresponding force a **FD point**
- the sequence of FD points of an unfolding protein obtained by SMFS a **FD curve**.
- the point at which a section of the protein unfolds an **unfolding point**. It is characterized by a distance X_u^i.
- a sequence of FD points corresponding to the stretching of a protein section, bounded by two unfolding points (or zero and the first unfolding point), a **peak** of the FD curve.
- a single curve that follows the WLC model and is fitted to one of the peaks of an FD curve a **WLC curve** for this peak. It is characterized by a contour length L_c^i. A WLC curve defines an intermediate state.
- the union of sequence of unfolding points and corresponding sequence of WLC curves for all the peaks of a FD curve, a **WLC profile** for that curve. A WLC profile is characterized by vectors $P_{Lc} = [L_c^1, L_c^2, ..., L_c^n]$ and $P_{Xu} = [X_u^1, X_u^2, ..., X_u^n]$. A WLC profile defines an unfolding pathway.

3. The Worm-Like Chain Model

One of the most common models describing the behavior of semi-flexible polymers upon stretching is the Worm-Like Chain (WLC). "The WLC model assumes that the polymer is inextensible, has a linear elastic bending energy and is subjected to thermal fluctuations" [5], and is described by the following function, which we call a WLC curve.

$$\begin{aligned} WLCc(x) &= -\frac{k_B T}{l_p}\left(\frac{1}{4\left(1-\frac{x}{L_c}\right)^2}-\frac{1}{4}+\frac{x}{L_c}\right) \text{ when } 0 \leq x < L_c \\ &= 0 \text{ when } x \geq L_c \end{aligned}$$

where k_B is the Boltzmann constant, T is the temperature, l_p is the persistence length of the polymer (the distance over which two segments remain directionally correlated), L_c is the contour length (the length of the fully extended chain), x is the recorded distance and $F(x)$ the recorded force as a function of the distance.

We will also use the notation $WLCc_{L_c}(x)$ to mean the WLC curve of parameter L_c evaluated at point x.

This model, used with persistence length $l_p = 0.36$ nm, appropriately describes the behavior of the loose end of the protein upon stretching [7]. But it doesn't take into account the unfolding of successive protein sections. These unfoldings should be seen as discrete events, happening at multiple unfolding points X_u^i, and increasing the contour length L_c of the loose end of the protein by the contour length of the unfolding section.

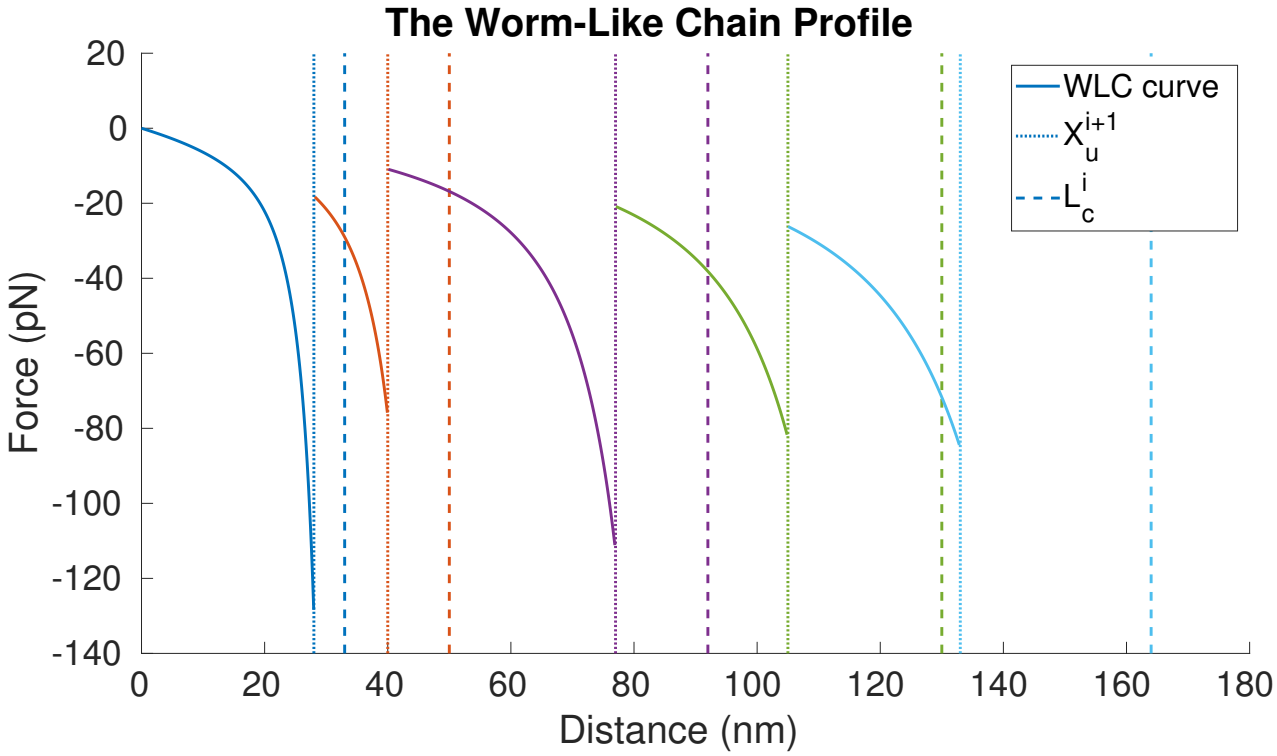
Therefore the model for the unfolding of a whole protein with n folded sections can be described by the following function, which we call a WLC profile

$$\begin{aligned} WLCp(x) &= -\frac{k_B T}{l_p}\left(\frac{1}{4\left(1-\frac{x}{L_c^i}\right)^2}-\frac{1}{4}+\frac{x}{L_c^i}\right) \text{ when } X_u^i \leq x < X_u^{i+1} \\ &= 0 \text{ when } x \geq X_u^{n+1} \end{aligned}$$

Where $(X_u^i)_{i=1..n}$ are the points at which the section i unfolds ($X_u^1 = 0$, the first section always starts out unfolded), X_u^{n+1} is the point at which the protein leaves the surface and $(L_c^i)_{i=1..n}$ are the contour lengths of the loose end of the protein after the i^{th} unfolding episode. The contour length of the i^{th} section is thus given by L_c^1 for the first section and $L_c^{i+1} - L_c^i$ for all the other sections.

A WLC profile is thus fully described by a list of contour lengths $P_{Lc} = [L_c^1, L_c^2, ..., L_c^n]$ and of unfolding points $P_{Xu} = [X_u^1, X_u^2, ..., X_u^n]$.

Let's look at what a WLC profile looks for concrete parameter values, choosing $P_{Lc} = [33, 50, 92, 130, 164]$ nm and $P_{Xu} = [28, 40, 77, 105, 133]$ nm. We get



In this model, we consider unfoldings of sections as discrete events, but this is an over-simplification. In practice the successive points of a FD curve do not switch instantaneously from one WLC curve to the next. An unfolding event will often appear as a succession of increasing points, connecting the last FD points of one peak to the first FD points of the following peak. These points will have to be treated as outliers to our model.

4. First Fitting Algorithm : Minima Fit

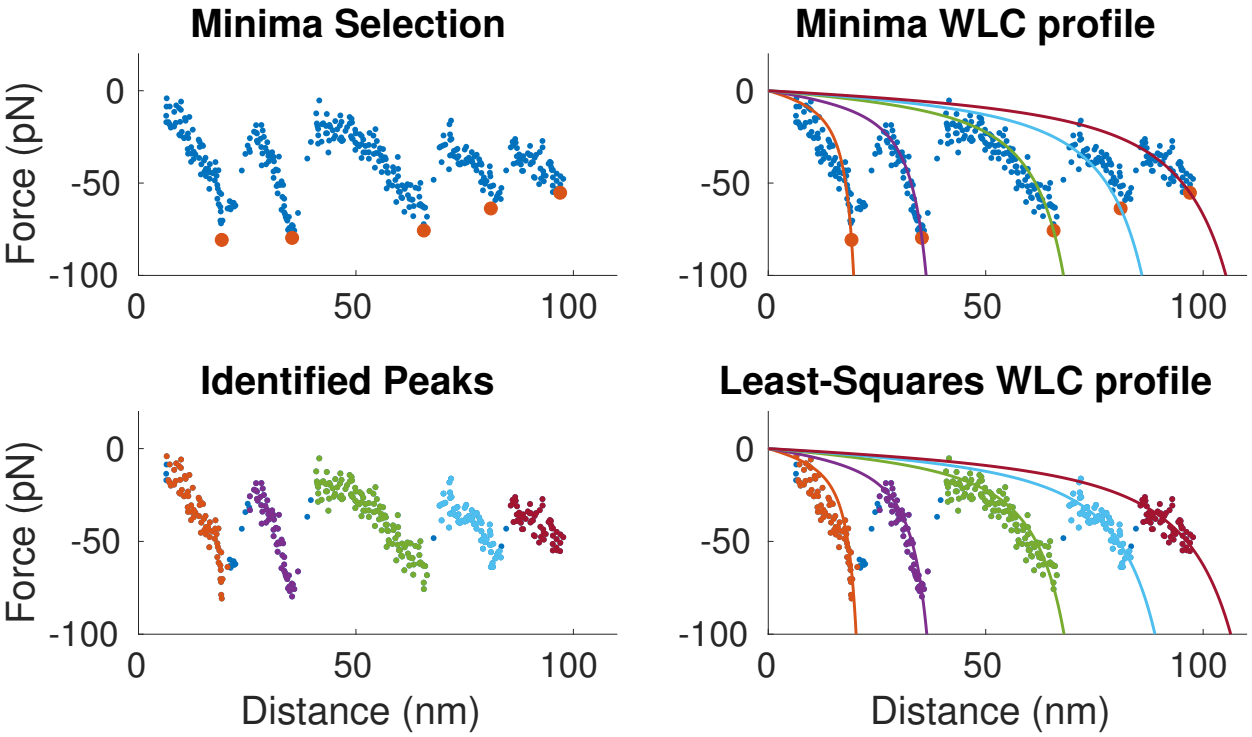
Our first objective is to fit a WLC profile robustly to a FD curve. To do so we propose the following algorithm. The idea is to use local minima of the FD curve as proxies for unfolding events, giving us a first WLC profile. We use this profile to identify peaks which we in turn use to update our estimate of the profile.

4.1. Algorithm

- **Find the minima of the FD curve**
  that correspond to an unfolding event. And since there is noise on the measured force, there are many minima that do not. We therefore need to select points that are minima on a large enough interval. At the same time, if we take too large a comparison interval, we might miss some good minima. We resolve this dilemma by taking a relatively small comparison interval but merging close WLC curves. This will be explained in section 4.2.
- **For each minima, find the L_c corresponding to the WLC curve interpolating this minima**
  To accomplish the second step of this algorithm we need to factor L_c out of the WLC model. To do so we find the only real and positive root of the following polynomial : $4\frac{Fl_p}{k_bT}L_c^3 + 2x(3 - 4\frac{Fl_p}{k_bT})L_c^2 - x^2(9 - 4\frac{Fl_p}{k_bT})L_c + 4x^3$, with x and F the distance and force of the minima respectively. See B for the derivation of this polynomial.
- **For each WLC curve, select the points close to it and find the updated WLC curve that minimizes the Mean-Square Error (MSE) on these points, using a Least-Squares fit. Iterate**
  The points close to our WLC curve are likely to be the points of the corresponding peak. We then reestimate the WLC curve based on all points of the peak. Two parameters need to be taken into account when selecting the points "close" to the WLC curve : the actual (vertical) distance to the curve, and temporal consistency. We do not want to select points that are too far from the curve because we suspect they do not belong to the same unfolding event, and we need to select a set of (temporally) adjacent points. This will be explained in section 4.3.

We can illustrate the different steps of this algorithm with the following plot showing successively the selected minima - the WLC profile fitted to these minima - the peaks identified based on this WLC profile - the WLC profile providing a Least-Squares fit of these peaks.

Peaks and their corresponding WLC curve have the same color, and the dark blue points in the dark blue points in the two last plots are the outliers (the points that haven't been selected as inliers to any peak).



Notice that the positions of the WLC curves are mostly determined by the position of the minima. This is because the WLC curves are very steep at the minima, and so a slight modification of L_c leads to a high increase in the quadratic error term associated with the minima, while not necessarily decreasing the total square error over the other FD points as much. Hence the Least-Square fit step of our algorithm often doesn't update each estimated L_c by a lot. Still, in some cases this step does provide an improved estimation (as for the fourth peak in the above illustration).

4.2. Merging WLC Curves

There is no optimal comparison interval that allows us to detect all minima corresponding to proper peak endings while never taking in minima that don't. To avoid missing peaks, we thus set a small comparison interval, and find some extraneous minima. To compensate for that we add a merging step to the algorithm after each Least-Squares fitting step.

This step will merge two WLC curves if their L_c are too close according to some threshold, keeping only the second WLC curve. That is because a minima that is not at the end of a peak always leads to an estimation of the L_c of that peak that is close to but below the correct value.

4.3. Selecting Inliers

After finding a minima and interpolating it with a WLC curve, we must select the points of the corresponding peak. These points must be recorded consecutively, to respect temporal consistency, and come after the previous peak.

Since the noise is on the measured force, the basic idea is to select a force threshold and select as inliers all FD points that are at a vertical distance of the WLC curve that is below this threshold. But since we also need to respect temporal consistency, we actually choose as inliers the set of points between the first and last FD points that are inside the threshold.

4.4. Parameter Tuning

This algorithm thus uses multiple parameters that have to be tuned to give good fit. The parameters are the following :

- **interval_length** : the length of the interval on which a FD point has to be the smallest to be selected as minima.
- **merge_thresh** : the distance between successive L_c below which they are merged into one.
- **force_thresh** : the threshold used to select the first and last inliers to a peak.
- **min_inliers** : the minimum number of inliers a peak needs to have to be considered valid.

After some experimentation, the value of these parameters is set to 10 nm, 10 nm, 10 pN and 5 respectively.

4.5. Discussion

On the positive side, the way this algorithm functions is intuitive, and it has a short execution time (less than 1 second per FD curve).

On the other hand, this algorithm is not very elegant, because it has 4 parameters that need to be tuned to give a good fit. The tuning of the parameters is mostly done by visual aid, and their value ends up being somewhat arbitrary. Performance is not great, as we miss many of the smaller peaks. It is also not quite clear whether this approach could be generalized to different kinds of protein, as their might not always be a set of parameter values that leads to a good fitting of the peaks.

We therefore try to design another algorithm, with the goal of being able to fit even small peaks and of having a cleaner, more generalizable approach.

5. Second Fitting Algorithm : Exhaustive Fit

We now develop a second algorithm to fit a WLC profile to a FD curve. We design this algorithm so that we don't need to find any minima to determine where the peaks are. We also avoid the merging step, hence reducing the number of parameters our algorithm needs, and increasing its robustness. We end up with an algorithm that fits even very small peaks whereas our first algorithm only fitted large peaks.

5.1. Algorithm

First we realize that if we can fit the first peak of a FD curve, then we can fit all of them because we can continuously fit the first peak and then remove it from the points under consideration.

To fit a first peak, we can work exhaustively : since a WLC curve has only one parameter, L_c, we can try out a range of possible values for L_c and keep the one providing the best fit. More precisely, for each L_c we compute the points that would be part of the first peak if its corresponding WLC curve was described by L_c (the *inliers* to this WLC curve) and compute the MSE of these inliers w.r.t. the WLC curve. Provided that the inliers are properly computed, the WLC having the lowest MSE will be a good fit to the first peak of the FD curve.

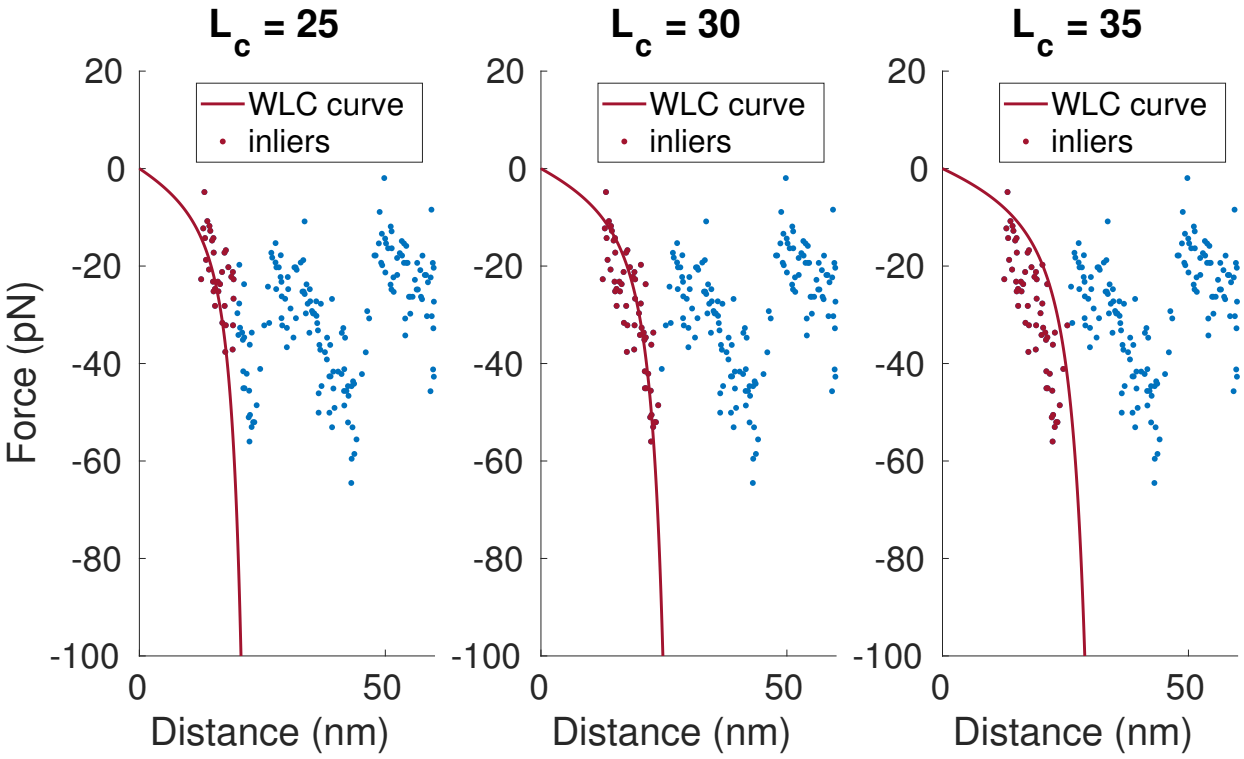
For this algorithm to work, we need to cleverly design the distance function. This is the function that takes as input L_c and outputs the corresponding inliers. When L_c corresponds to the proper WLC curve, the returned inliers should be all the points of the corresponding peak. By definition the proper WLC curve is a good fit for its peak, hence it will have a low MSE. On the other hand, when L_c is too low or too high then the returned inliers should be such that we get a high MSE. We define our distance function as follows

Distance Function
Input: L_c and the FD curve $(f_i, x_i)_{i=1}^n$
Output: $(f_i, d_i)_{i=1}^k$, where (f_k, d_k) is the last point satisfying the condition $f_i < WLCc_{L_c+m}(d_i)$.

Where the margin m is set to 3 nm.

This distance function satisfies the mentioned constraints. The proper WLC curve will get the points of its peak as inliers. When L_c is too low, we will select some FD points after the WLC curve which, because of the steepness of the WLC curve, will lead to a high increase in the MSE. When L_c is too high, the points of the peak will necessarily be selected as inliers and lead to a high MSE.

For example, when attempting to fit the first peak of the following FD curve, trying $L_c = 25$, $L_c = 30$ and $L_c = 35$, we will find the following inliers, and we will get an MSE of 204, 71 and 201 respectively, leading us to find the proper value of $L_c = 30$.

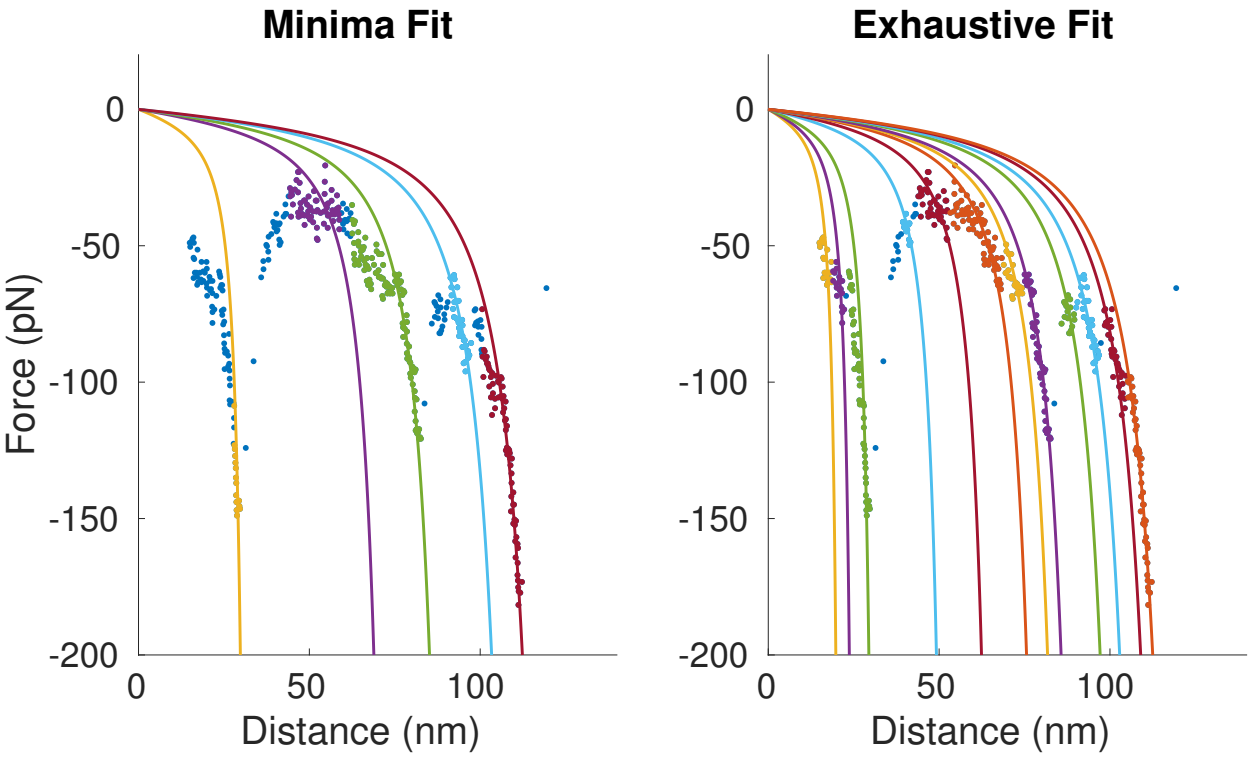


As is, the algorithm will properly fit all the peaks present in the FD curve, but it will also fit the outliers (the FD points connecting the unfolding point of a peak to the start of the next peak). Characteristically these outliers will be fitted in small groups, and therefore discarding all peaks containing fewer than $min\_inliers = 10$ cleans up our WLC profile. But this may also remove some very small real peaks, so we could consider using another method to get rid of outliers.

The attentive reader will have noticed that this algorithm is a kind of modified version of RANSAC, where instead of trying a random selection of models we allow ourselves to try them all. We can afford this due to the low dimension of the model. We will use an other modified version of RANSAC in section 7.3.

5.2. Comparison with Minima Fit

Here follows an illustration between the way both our algorithms fit a FD curve. As you can see, the exhaustive fit is able to identify much smaller peaks, and peaks that aren't obvious to the naked eye. This is true for all our FD curves.



This comes with the additional benefit that this method has only two parameters : *margin* and *min_inliers*, compared to the 4 parameters needed for the minima fit, and that this approach has more generality.

6. Finding the Offset

As stated in the introduction, the distance 0 of each FD curve is initially set to be the point at which the tip reaches the surface on which the protein lays (detected by a sharp increase in the recorded force). But since both ends of the protein aren't exactly attached to the end of the tip and to the surface right under the tip, all FD curves end up shifted by an unknown amount, preventing us from easily comparing them.

When fitting a WLC profile to a FD curve, we implicitly assumed that the distance 0 (current origin) of this FD curve corresponds to the point where the protein is completely folded (proper origin). This assumption was unwarranted, and we call the difference between the current and the proper origin the offset. If we could find the proper origin of each FD curve (or equivalently, its offset), then comparing them would be much easier because we could directly compare the values of each of their WLC profiles to identify common peaks.

Note that if the offset is small then the WLC profile we find will be close to the proper WLC profile plus the offset. We can thus still find the successive section lengths (except the first one) by computing $L_c^{i+1} - L_c^i$, because the offsets cancel out.

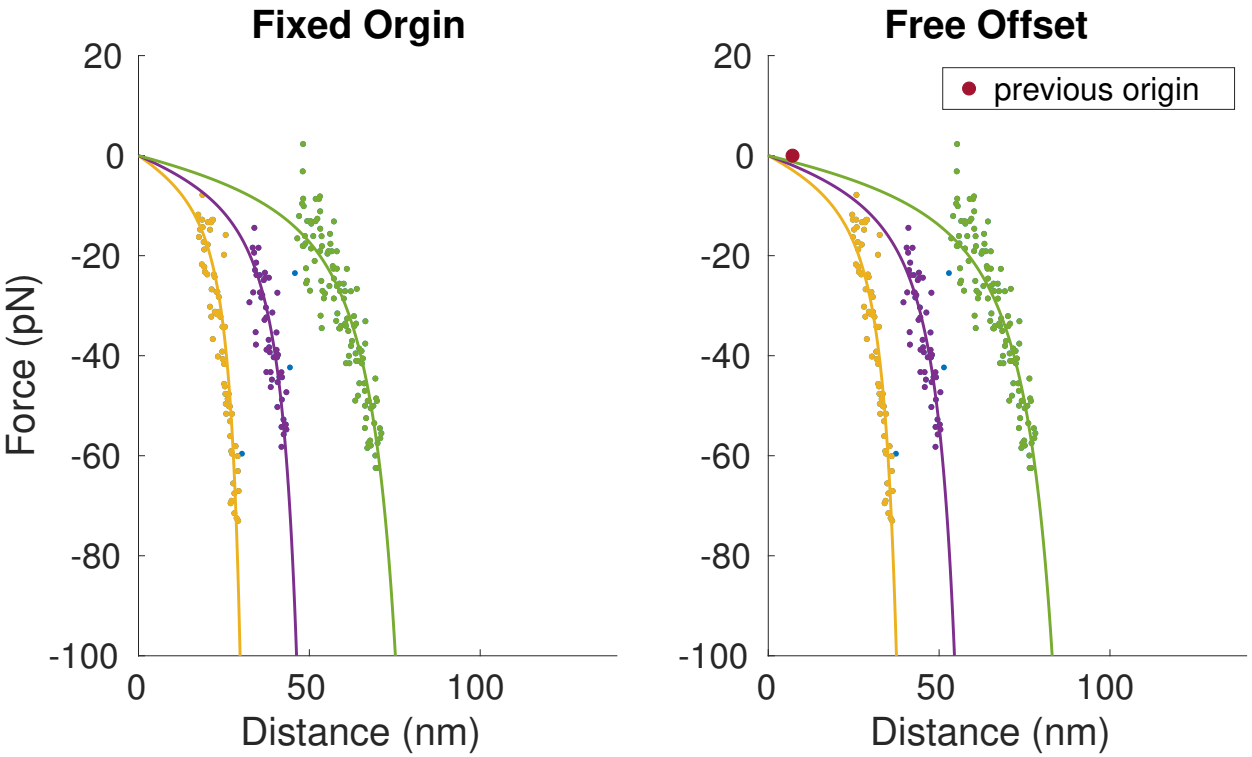
6.1. Least-Squares Optimal Offset

Both for the minima and the exhaustive fit, we tried to find the contour lengths L_c^i of the WLC model that leads to the the best fit of a FD curve. But nothing stops us from also finding the offset that leads to the best fit of a FD curve.

After running our exhaustive fit algorithm on a FD curve we get a WLC profile and the points that are present in each peak. To find out the offset between the current and proper origin, we can do a Least-Squares fit step of the different peaks, leaving the profile P_{Lc} and an offset *delta* as free variables. Hence, we will find the offset that minimizes the MSE of the WLC profile over all the FD points selected as inliers to a peak.

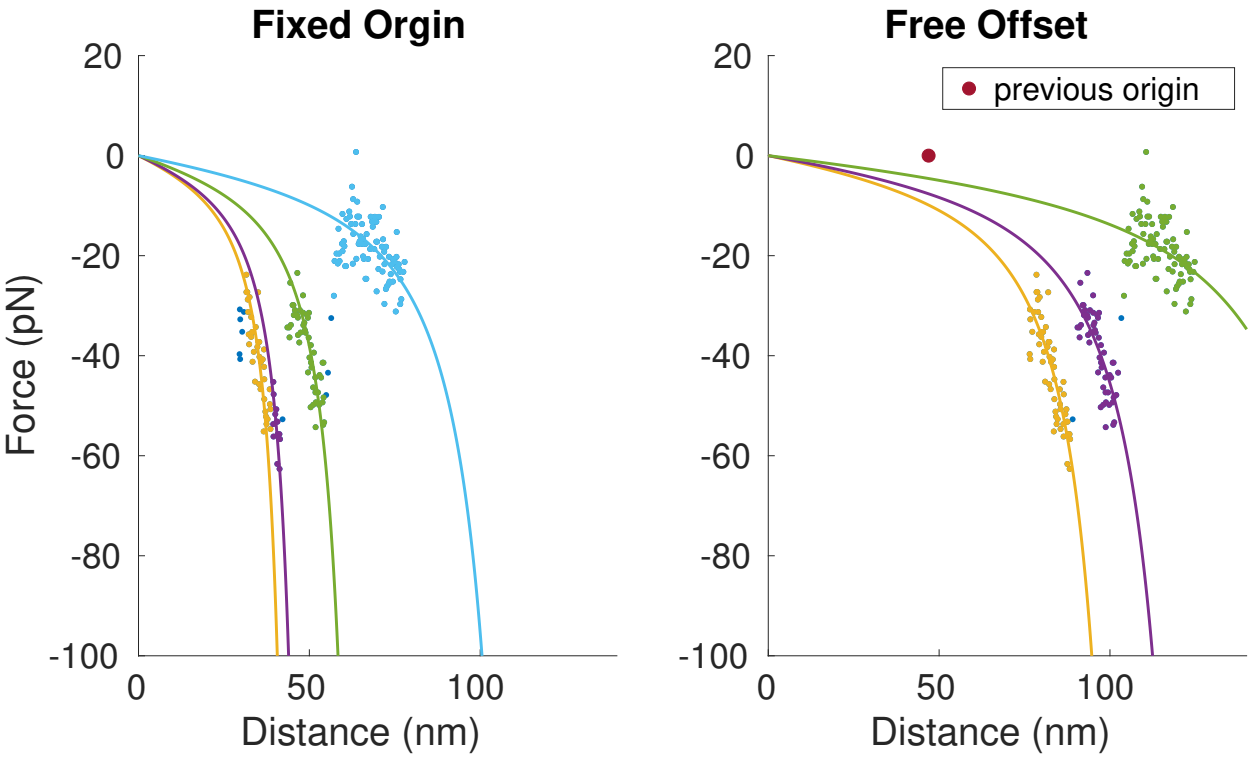
If the offset is large, then our first estimation of the composition of the different peaks might have been incorrect. Therefore, we iterate the two steps a few times (using the exhaustive fit algorithm to re-estimate the position of the peaks of the shifted FD curve). We will nearly always converge quickly.

The first results are encouraging : most FD curves end up slightly shifted and their WLC profile then provides a better overall fit. For example



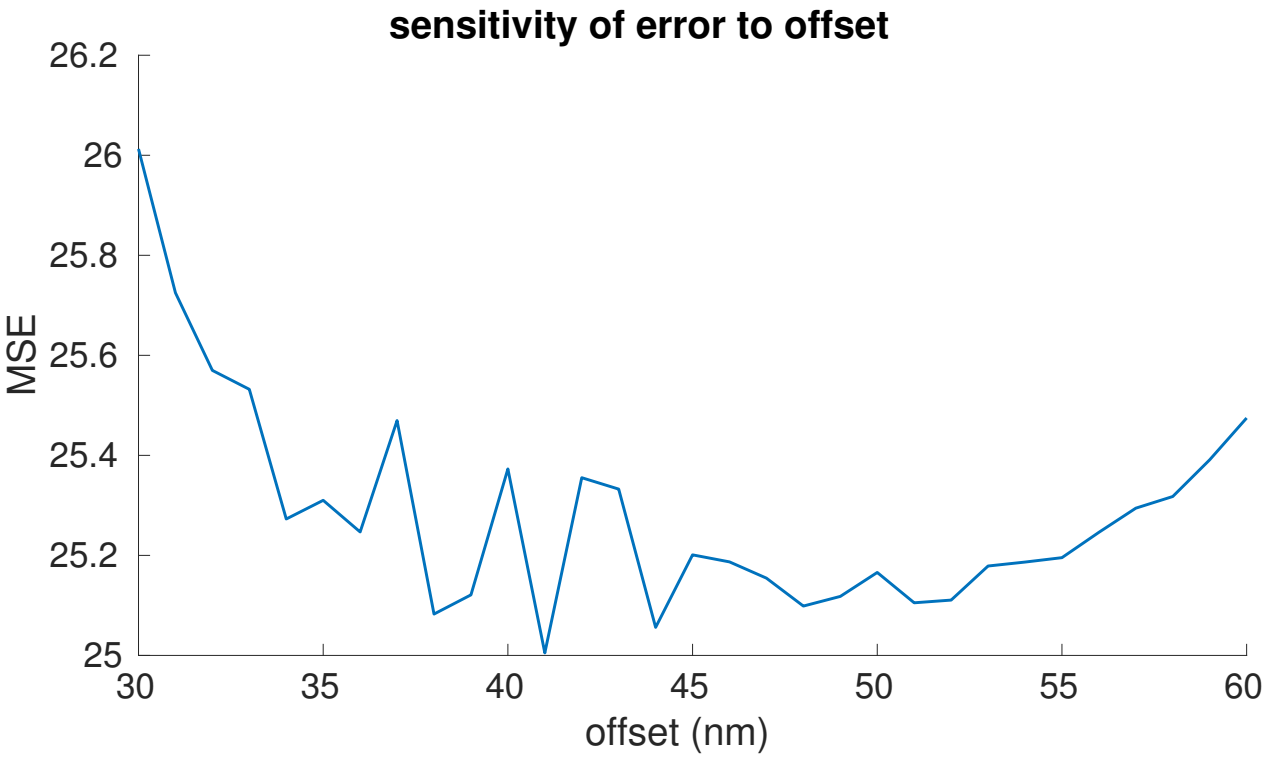
Where the optimal offset is found to be 7.10 nm.

But for some FD curves we find very large offsets, which doesn't mesh well with our expectations. For example



Where the optimal offset is found to be 46.92 nm.

We must therefore question the robustness of our method. Since our method is based on finding the shift minimizing the MSE of the FD curve w.r.t the WLC profile, let's look at how this MSE is impacted by the value of the offset, for the specific FD curve above.



Here we can see that the MSE isn't very sensitive to the offset : for an offset of 60 nm, we find a MSE of 25.47 whereas for the an offset of 41 nm we have 25.01 : only a 1.01% difference in MSE for a 19 nm difference in offset!

The MSE is also highly influenced by the selection of a few points as inliers or not, which explains its volatility around the optimal value. And because we used fixed peaks at our Least-Squares fit step, we did not find the optimal offset of 41 nm, but a local optima of 47 nm.

Furthermore the WLC model is already an approximation of the behavior of an unfolding protein. Peaks don't necessarily all correspond perfectly to this model. Using a method that is very sensitive to the FD curve strongly corresponding to the WLC model is thus not a robust approach

We can still use this method as an aid to estimate the offset, but we should remain skeptical of the results. This is why we end up using another method to align the different FD curves, which will be presented in section 8.

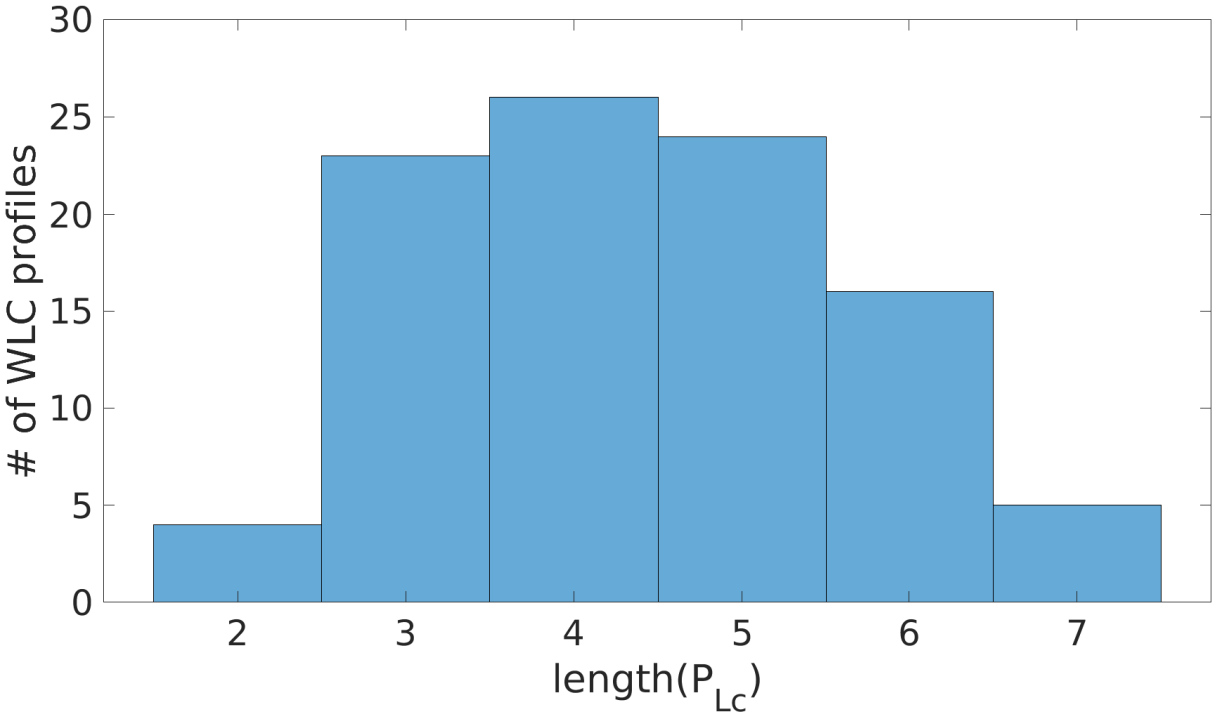
7. Clustering of WLC Profiles

At this point we have thus found a way to reduce a FD curve to a WLC profile characterized by vectors P_{Lc} and P_{Xu}. But we know there is a different offset on each of these FD curves, and thus on their WLC profile. But this shouldn't stop us from trying to compare the different WLC profiles to find main unfolding pathways (common successions of WLC curves). We just need our comparison method to be translationally invariant, i.e. insensitive to shifts on the profile.

The first step of our analysis will be to detect clusters in the contour lengths of our profiles $(P_{Lc}^j)_{j=1}^N$, obtained with the minima fit method. We use the minima fit method here because it only fits the main peaks of a FD curve, and these are the peaks we expect to be part of a main unfolding pathway. This also simplifies our analysis, because the WLC profiles will have smaller numbers of peaks. The exhaustive fit method will be used later on, when aligning the WLC profiles (8).

7.1. Simple Classification

We first classify our WLC profiles according to their number of peaks (which is equal to $length(P_{Lc})$). The number of elements in each class is distributed as follows.

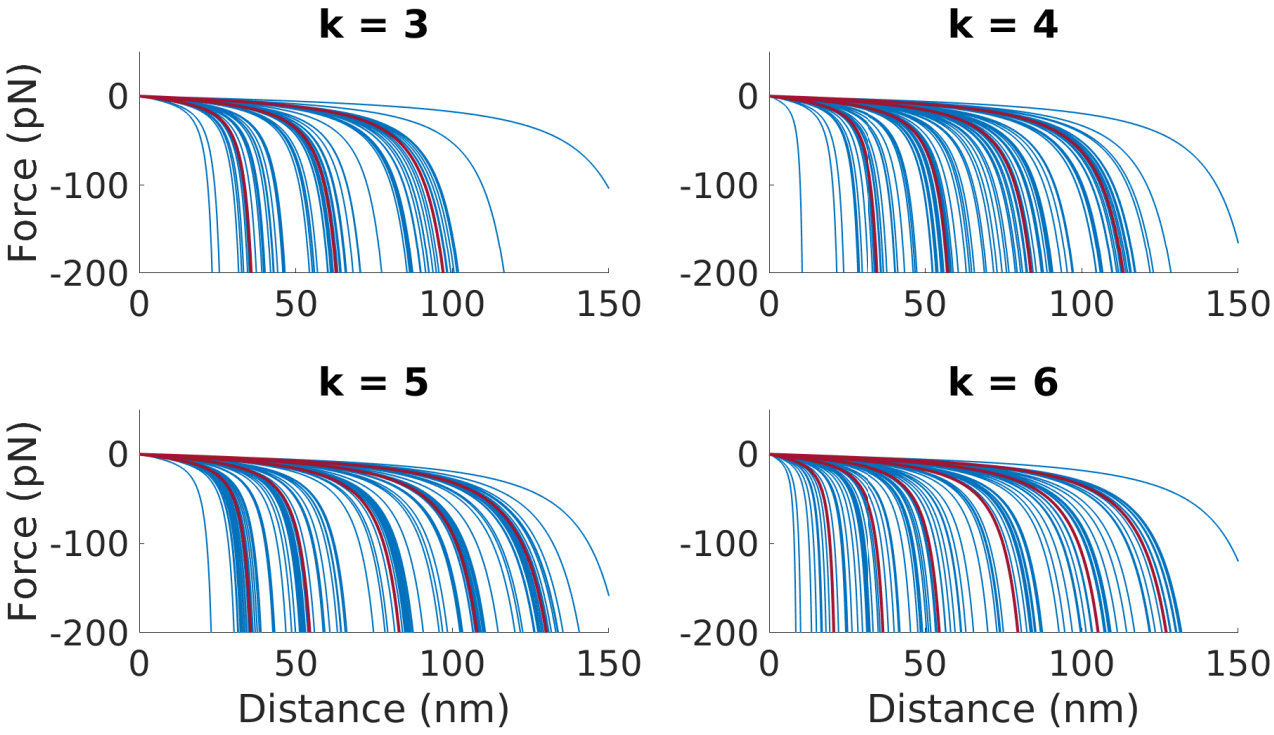


Meaning most WLC profiles have between 3 and 6 peaks.

Do these classes suffice to properly split the WLC profiles? They are if the different profiles of a class can be aligned through proper shifting, implying that they encode the same unfolding pattern. To check this hypothesis, we compute the element-wise mean profile of each class and see how the different elements of the class align to it.

More precisely, we assume each class k contains N_k WLC profiles ($\sum_{k=2}^{k=7} N_k = N$). For each class we compute the template WLC profile, which is the element-wise mean WLC profile, i.e. $P^k_{Lc\mu} = \underset{j \in N_k}{\text{mean}}(P^j_{Lc})$. We then align all profiles in the class to the template, i.e. for each profile P^j_{Lc} we find the offset δ_j minimizing the MSE w.r.t. the template. We find $\delta_j = \underset{\delta_j}{\text{argmin}}(\|P^k_{Lc\mu} - (P^j_{Lc} + \delta_j)\|^2/k) = mean(P^k_{Lc\mu} - P^j_{Lc})$.

For each class, we can then plot the template profile $P^k_{Lc\mu}$ (in red), and the shifted profiles of the class (in blue). We get the following :



Clearly, not all the different WLC profiles of each class align well the template. But for $k = 5$, we can already see that many WLC profiles do align well (because the WLC curves are denser around the template). There is thus probably a large fraction of the WLC profiles that are very similar to each other, and correspond to a canonical unfolding pathway, while the other WLC profiles are outliers.

For each class we can try to find the main cluster. Finding a cluster in a set of data that contains outliers is a typical use-case for RANSAC, which we explain in the following section.

7.2. RANSAC

RANSAC is "*an iterative method to estimate parameters of a mathematical model from a set of observed data that contains outliers*" [1]. This definition fits our problem well, since we want estimate the parameter ($P^k_{Lc\mu}$) of a mathematical model (WLC profile) from a set of observed data ($(P^j_{Lc})_{j=1}^{N_k}$, the set of profiles that are in class k) that contains outliers.

Before applying a slightly modified version of this algorithm to our problem, let's look at how it works in general. The standard RANSAC framework requires that we define 3 things : a fitting function, a distance function and a selection criterion.

- The fitting function takes as input a set of data points and outputs a model that fits those points.
- The distance function takes as input a set of data points and a model and outputs the set of inliers, i.e. the subset of data points that correspond to the model.
- The selection criterion defines the quality of a model. It usually favors a combination of two things : a high number of inliers to the model, and a low MSE on these inliers (w.r.t. the model). Those criteria are often in tension, which is resolved by using a hybrid criterion, requiring the best model to have a minimum number of inliers and to have the lowest MSE among all models satisfying this first criterion.

Once these are defined, the algorithm can be run as follows.

for iter = 0 **to** iter = n **do**
  Select a random subset of data points
  Fit a model to those points using the fitting function
  Select the inliers given by the distance function on this model
  if this is the best model yet according to the selection criteria **then**
    Save the model as current best
  end if
end for
return the best model found

The framework requires us to determine how many iterations we want to do, and to define a distribution function for randomly selecting data points. Usually a uniform distribution is used, but this can be changed to improve performance depending on the problem.

7.3. Aligned Clustering

For the specific problem at hand, we define the following.

The fitting function is, straightforwardly

Fitting Function
Input: A set of profiles $P^i_{Lc}, i = 1..n$
Output: $P_{Lc\mu}$ the (element-wise) mean of the set

Our distance function is slightly more tricky to design. We want it to select as inliers the profiles that, if they were properly shifted, would align best with the template. And the quality of an alignment is properly described by the MSE, since this measure strongly penalizes any strong deviation on one peak of the profile.

Therefore, we take as inliers all WLC profiles P^j_{Lc} that, once shifted to $P_{Lc\mu}$ to minimize the MSE $\|P^k_{Lc\mu} - (P^j_{Lc} + \delta_j)\|^2/k$, are such that $\|P^k_{Lc\mu} - (P^j_{Lc} + \delta_j)\|^2/k < thresh$.

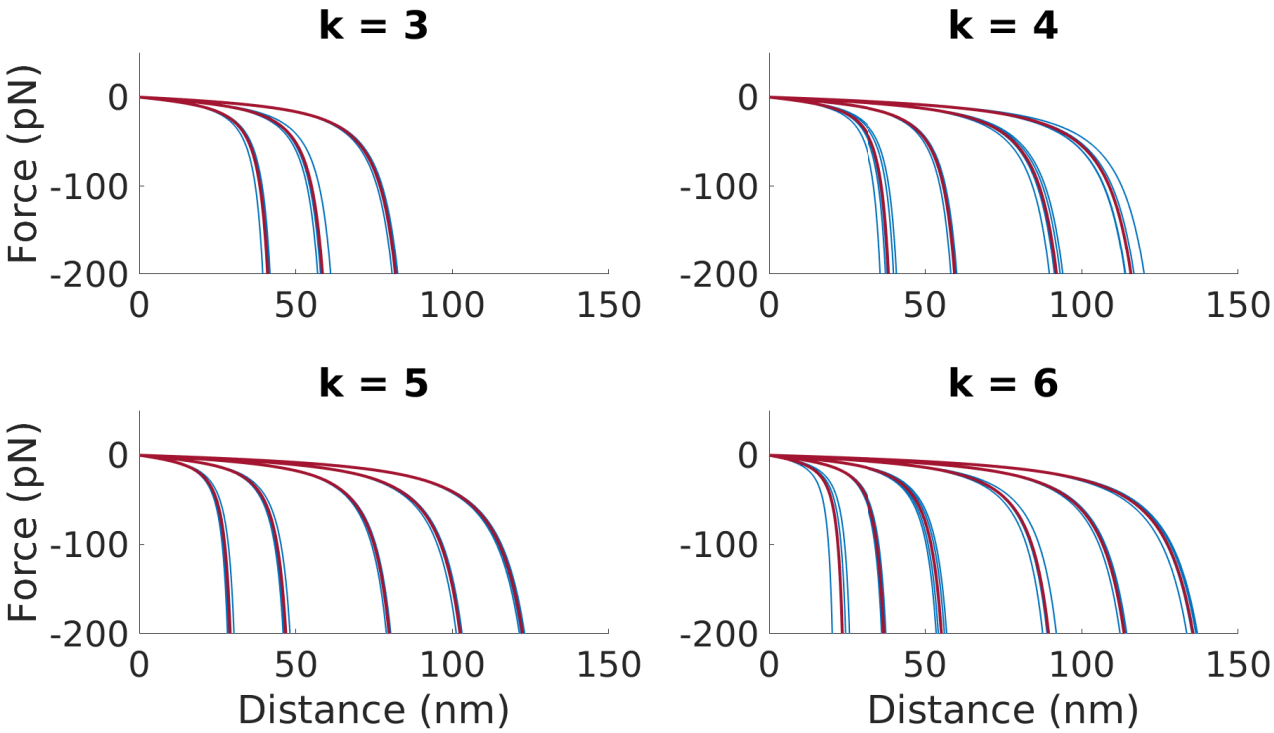
Distance Function
Input: A template profile $P_{Lc\mu}$, the set of class profiles $(P^j_{Lc})^{n_k}_{j=1}$, a threshold $thresh$
Output: $(P^{j2}_{Lc} | j2 \in \{1..n_k\}$ and $\|P^k_{Lc\mu} - (P^{j2}_{Lc} + \delta_{j2})\|^2/k < thresh))$

Our selection criterion is the following : we require the best model to have a certain number of inliers $min\_inliers$, and among these ones we select the one that has the smallest MSE.

In this specific problem, points and models are of the same type (WLC profiles). Hence instead of randomly drawing a sample of WLC profiles, using the fitting function to find a template WLC profile and then finding the inliers to this template, we can simply pick one of our already existing WLC profiles as template. And since there aren't that many WLC profiles (100 in total), we don't pick a WLC profile randomly but try out each WLC profile exhaustively. Our algorithm is thus not strictly speaking a RANSAC algorithm, because we don't use any random drawing of points, and the result is entirely deterministic.

After finding the inliers to our template, we use the fitting function to update our estimation of the template.

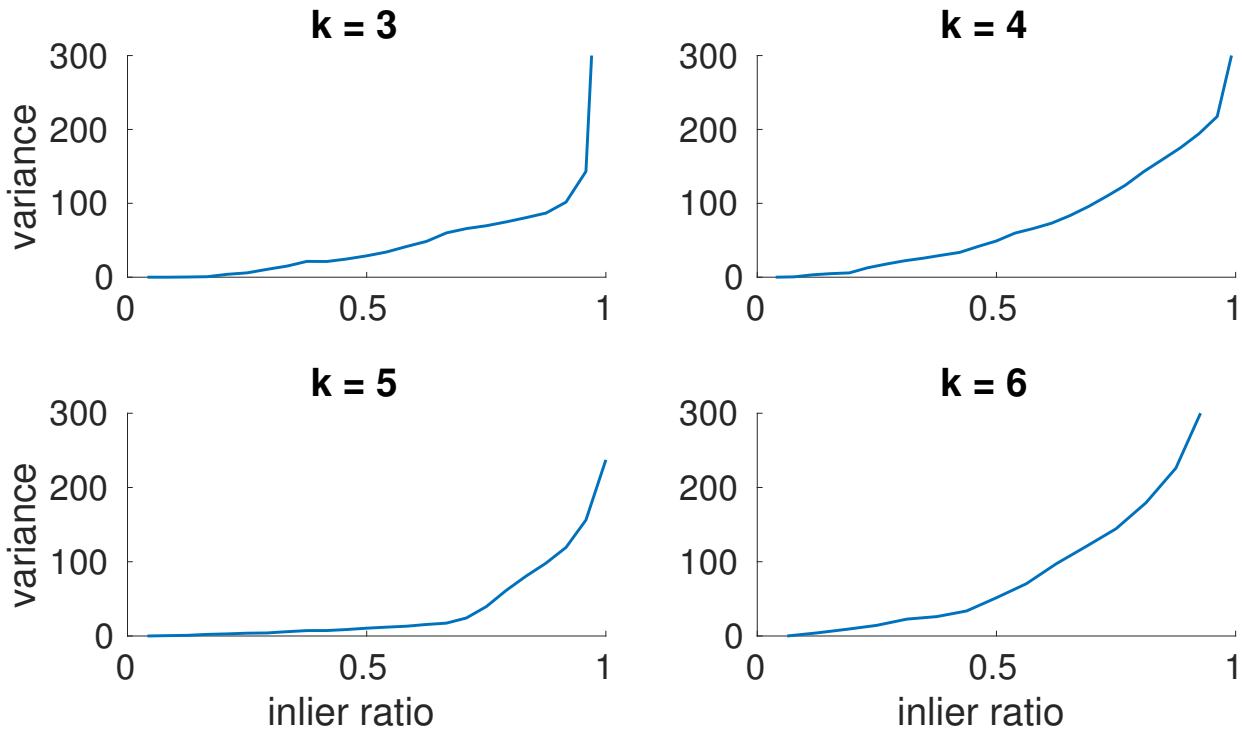
Choosing $thresh = 50/k$ and $min\_inliers = 0.2N_k$, we get the following clusters. The profiles of the cluster are in blue, whereas the template is in red



The clusters now appear a lot more sharply, as we might have expected. But our threshold was chosen quite low, and we end up with very few WLC profiles in each of our clusters. How should we set this threshold, if we don't know *a priori* how many inliers there are in each cluster?

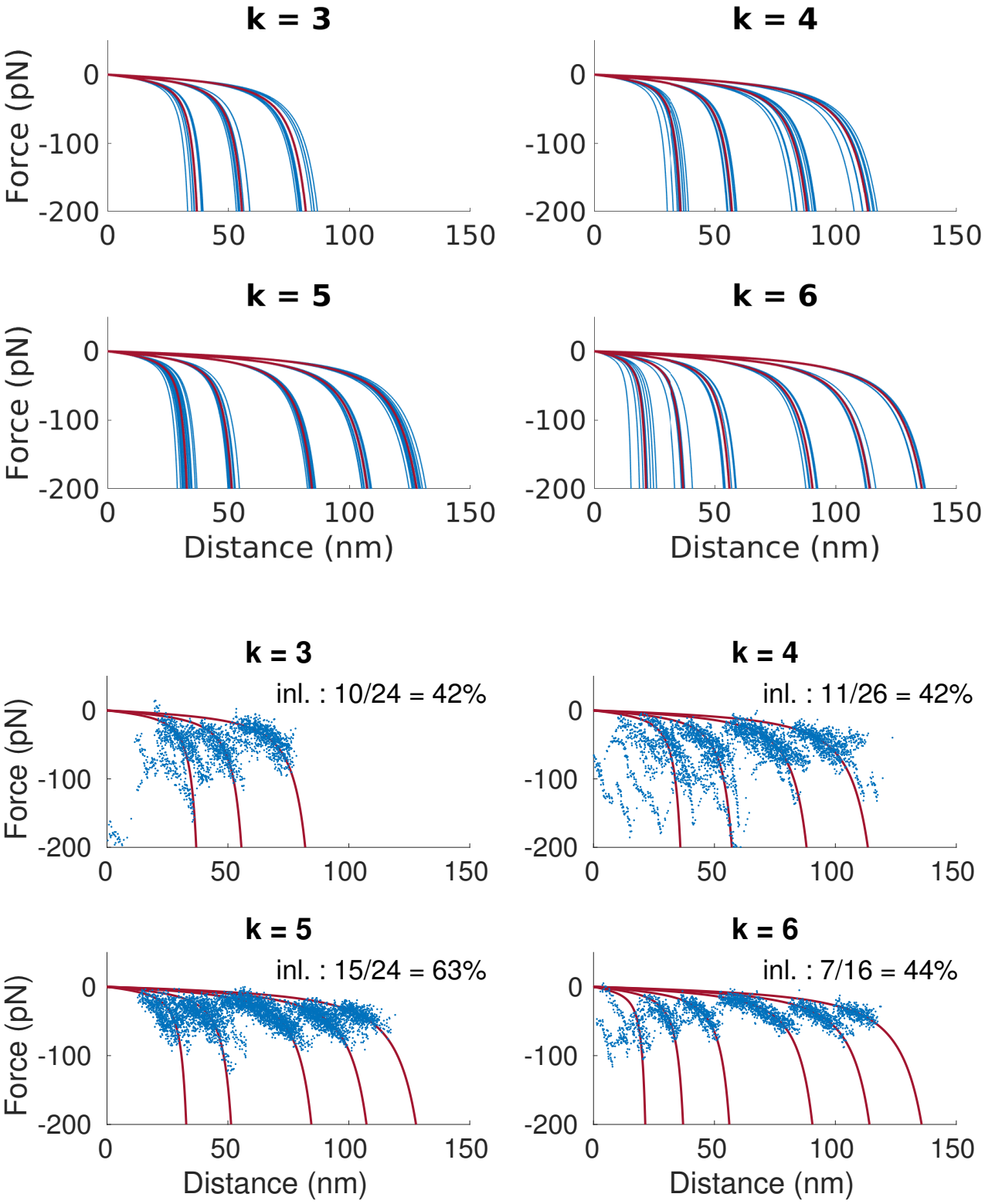
For this we get inspiration from [3] and modify our method. Instead of setting an arbitrary error threshold for choosing inliers in the distance function, we use the following condition : we set a required number of inliers $n\_inliers$ we want our model to have, and among all such models choose the one with the lowest MSE. More precisely, for each template we will select as inliers the closest $n\_inliers$ WLC profiles (after alignment, in the min MSE sense) and compute the MSE of these inliers. We return the best such template and its inliers.

We can now analyze how the variance of the inliers of the best model evolves when increasing $n\_inliers$. Presumably, the variance should increase slowly as long as $n\_inliers$ is lower than the true number of inliers, but after that point we should see a sharp increase in variance, since we would be adding true outliers. Let's take a look at how the variance increases with the required $n\_inliers$ (normalized here to the required inlier ratio).



The predicted inflection point in the evolution of the variance is only clearly present for $k = 5$, around 63% of inliers. For the other groups this is not so clear. Nevertheless we must choose a ratio, and we set it at 42%, 42% and 44% respectively (we choose ratios corresponding to $n\_inliers$ being an integer).

For these specific ratios, we thus find cluster that contain many WLC profiles, while (hopefully) not taking in any true outliers. We also plot the FD curves that correspond to each these WLC profiles, shifted by the same amount. This allows us to check if the alignment of these FD curves seems proper, regardless of the quality of our fitting method.


k = 3
k = 4
k = 5
k = 6
Force (pN)
Distance (nm)
0
-100
-200
0
50
100
150
k = 3
inl. : 10/24 = 42%
k = 4
inl. : 11/26 = 42%
k = 5
inl. : 15/24 = 63%
k = 6
inl. : 7/16 = 44%

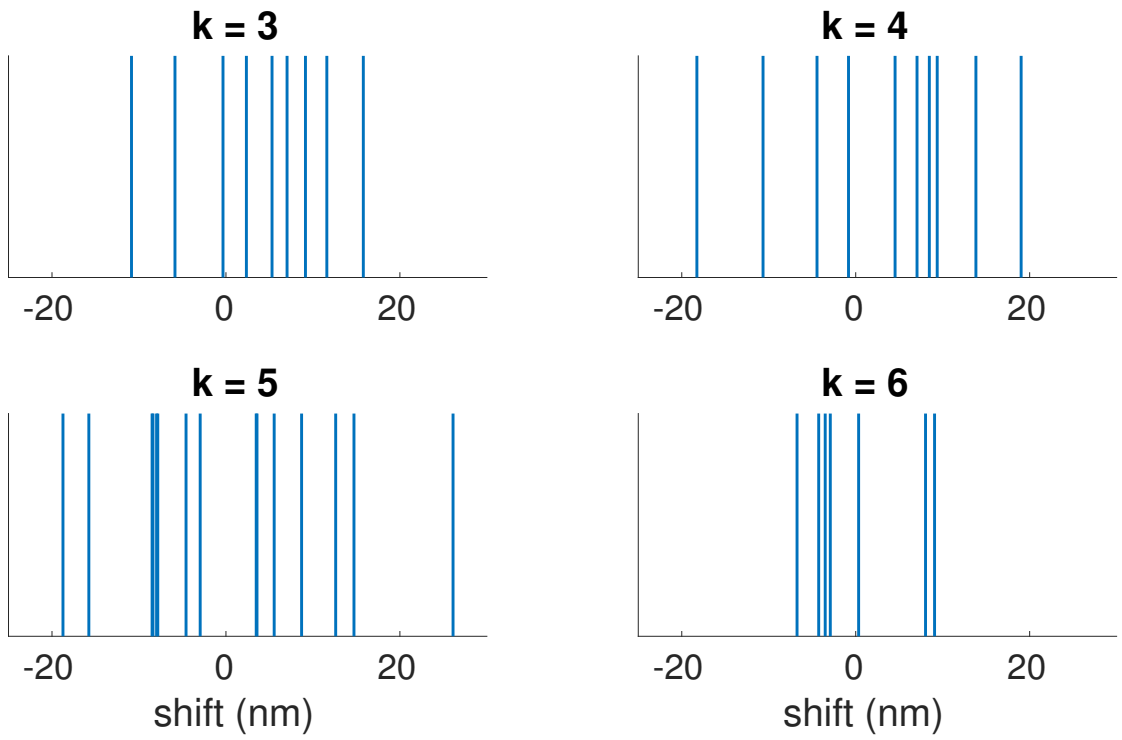
The different FD curves do align well, especially for $k = 5$ and $k = 6$. Some outliers might have been taken into $k = 4$.

We notice a clear similarity between all those clusters : it appears that the cluster for $k = 5$ is the same as the one for $k = 3$ and $k = 4$ with one and two extra peaks at the end respectively, and a slight shift. This nicely fits with the hypothesis that some proteins detach from the surface before being totally unfolded (thus loosing the last peaks). $k = 6$'s cluster is also very similar to $k = 5$'s but with one extra peak at the start. But we can see that the WLC curves are not very dense around this first peak. We deduce from this that the first peak is probably due to contact forces between the tip of the microscope and the surface, and not any section unfolding. These kind of peaks are supposed to be removed when preprocessing the FD curves (see A) but they can't always be because they are, *a priori*, hard to distinguish from peaks due to protein sections unfolding.

So we actually found one main cluster, that in its canonical form has 5 peaks, with $P_{Lc\mu} = [34.5, 54.6, 92.4, 118.0, 140.6]$ nm.

7.4. Shift Distribution

In our final clusters, each FD curve had to be shifted a certain amount to align properly with the template. We can take a look at how the shifts of the FD curves are distributed

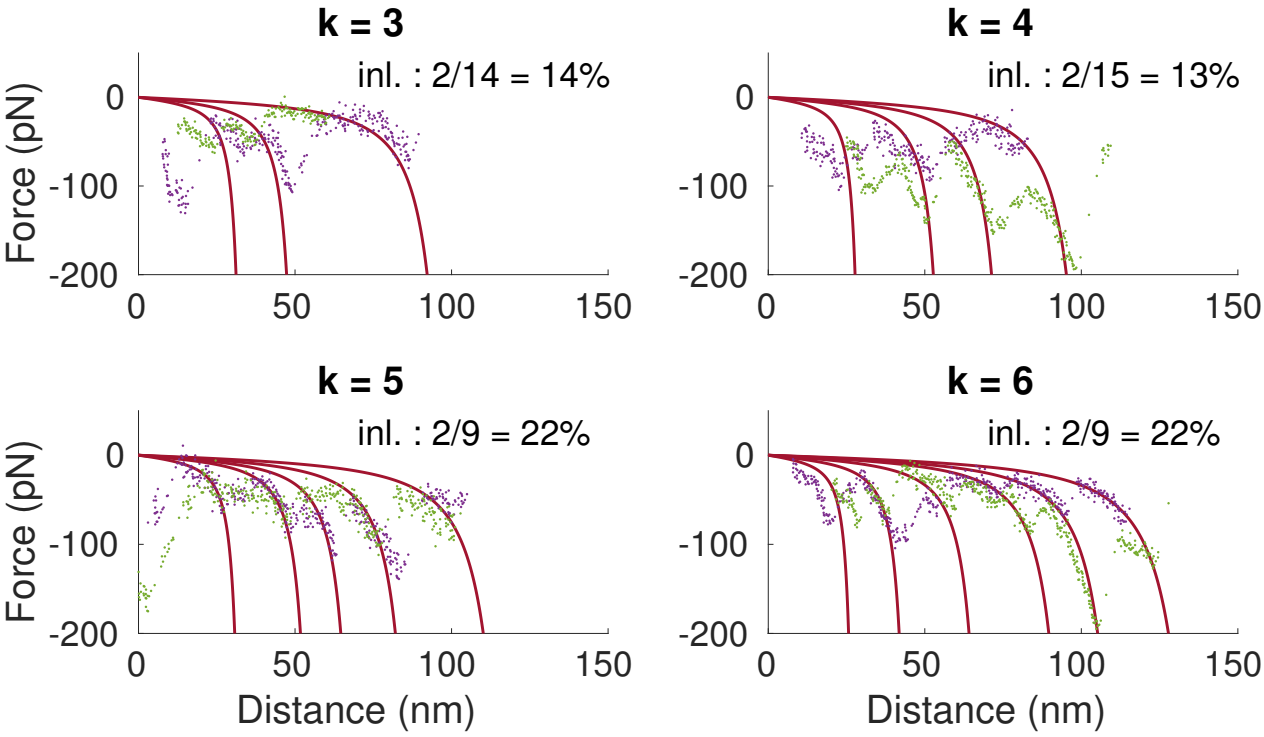


We can see that they are distributed on a wide range, between -25 and $+30$ nm. This confirms that there is a high uncertainty on the proper origin to our FD curves.

We could recompute the WLC profile of each shifted FD curve and then reiterate the clustering step, so as to find an updated estimation of the template WLC profile.

7.5. Secondary Clusters

We could imagine their being a second cluster of WLC profiles among the outliers to the first cluster. This would correspond to a second unfolding pathway. To check for this possibility, we can simply apply our RANSAC algorithm again to the outliers of the first clusters.



For each cluster we are requiring only 2 inliers, hence we are finding the 2 remaining FD curves that are closest to each other. But as we can see, even these FD curves aren't very similar, their peaks aren't properly aligned. Thus we don't expect to find any meaningful secondary cluster. This might be due to the fact that there are very few FD curves left at this point. Having more FD curves could then allow us to find a secondary cluster.

But actually, we don't really expect to find a second major unfolding pathway. Intermediate states of the main unfolding pathway are stable intermediates, and we expect most FD curves to have to have some of them. But their might be secondary intermediate states that are shared by multiple FD curves even though these FD curves have different unfolding pathways. So we can still hope to find secondary intermediate states. But to find those, we must first align all FD curves.

8. Alignment of WLC Profiles

We have now found the template WLC profile corresponding to the most common unfolding pathway. Around 50% of our FD curves follow this unfolding pathway. These FD curves can easily be aligned to the template by MSE alignment of their WLC profile. But what about the other 50% of FD curves? Can we somehow align them to the same template? This would help us analyze other interesting features, like the presence of common secondary peaks.

Even though these other FD curves don't have all the peaks of the main unfolding pathway, we can assume that they have some : since these peaks correspond to stable intermediates states, they should be present more often than not. A suitable generalization of the MSE alignment method can help us exploit this assumption.

8.1. Exhaustive Alignment

How should we go about aligning the remaining FD curves to the template, using their WLC profiles?

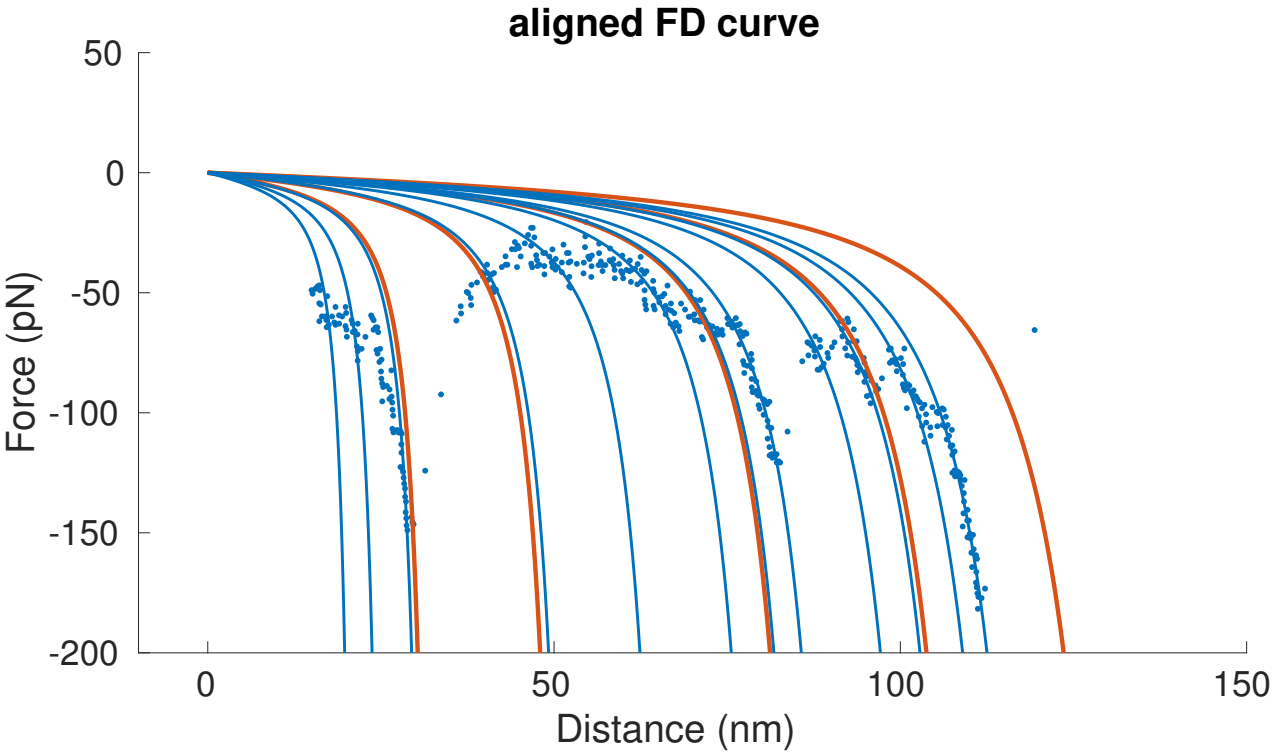
Simply using the shift minimizing the MSE on the peaks, as we did before, will not work for two reasons. First, we are now trying to align WLC profiles that aren't necessarily of the same size as the template profile. Second, we know each profile can contain WLC curves that aren't present in the template, and the other way around, and if we take these extra WLC curves in consideration they will mess up our estimation of the shift.

To avoid taking these extra WLC curves in consideration, we develop the idea of pairing. A WLC curve from the template is considered *paired* to a WLC curve from the FD curve if they are at a distance lower than a certain threshold (in practice, 1.5 nm). To align the profiles based on WLC curves that are likely to be common, we then use two criteria : to maximize the number of paired WLC curves and to minimize the MSE on those WLC curves respectively.

Our algorithm works exhaustively : for a large range of shifts (between -40 and $+40$ nm), we compute the profile of the shifted curve (with an exhaustive fit) and compute the number of paired WLC curves and the MSE on those WLC curves. Among all those shifts, select the one corresponding to the highest number of paired WLC curves and lowest MSE.

We recompute the profile of the FD curve for each shift instead of simply shifting the WLC profile because this increases accuracy. Recall that shifting a FD curve does not simply lead to an equivalent shift of the corresponding WLC profile.

Here follows the alignment obtained for one of our FD curves whose WLC profile wasn't part of the main cluster found in section 7. WLC curves one through four from the template have been paired with WLC curves from the FD curve, and the optimal shift has been found to be -13.9 nm.



8.2. Discussion

The shift found this way should be seen as a good guess of the proper shift, but it is likely that some FD curves will be incorrectly aligned. This can happen for two reasons.

First, if we can only pair one WLC curve from a FD curve to the template then we cannot say anything about the proper alignment of this FD curve. That is because getting one paired WLC curve can be done using many different shifts (while always getting an equal MSE of 0). This happens for 3 out of our 100 FD curves.

Second, if we can only pair two WLC curves from a FD curve with the template, then this alignment is very fragile. We may have 2 WLC curves that just happen by random luck to be at a distance that is close to the one between 2 WLC curves of the template. Worse, if for example, a FD curve gets paired to WLC curves 1 and 4 from the template (distance between curves 1 and 4 : 83.5 nm), then it could almost as well be paired to WLC curves 2 and 5 (distance between curves 2 and 5 : 86 nm). The

tiebreaker, taking the shift providing the lowest MSE on the paired peaks, should help us choose the right alignment. But it will not necessarily always work, especially if we keep in mind that the template WLC profile was obtained through clustering, and is itself approximative. This could lead to systematic misalignments.

And this problem is not unique to these two pairs of WLC curves. You can see this in the following table giving the distance between each pair of WLC curves from the template. We can see there that the distance between WLC curves 1-2 is close to the one between 3-4 and 4-5, the distance between 1-3 close to the distance between 2-4 and the distance between 1-4 close to the distance between 2-5. Even the triplet 1-3-4 is similar to the triplet 2-4-5. Whenever using any of those pairs or triplet for alignment, we run a higher risk of misalignment.

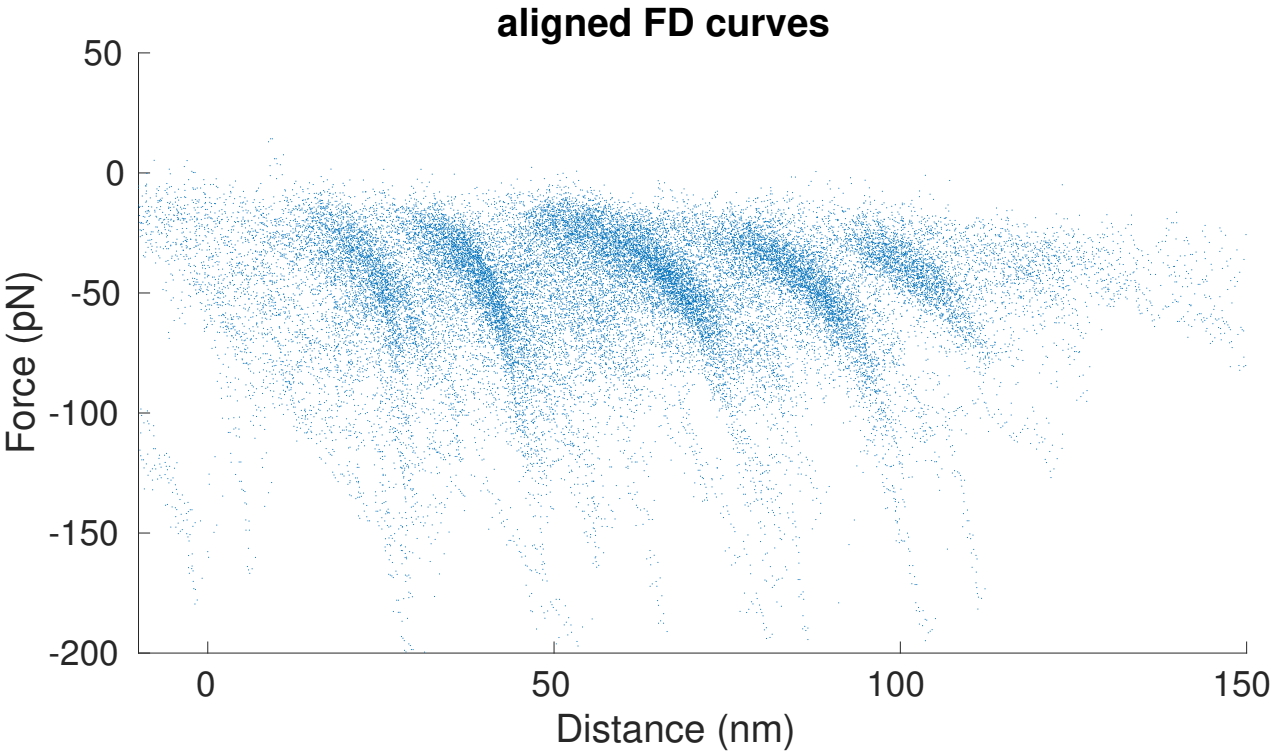
WLC curves	1	2	3	4	5
1	0	20.12	57.89	83.56	106.10
2		0	37.76	63.44	85.98
3			0	25.68	48.22
4				0	22.54
5					0

Still, lacking a better alternative we use this method to align all FD curves to the template. We can then analyze the set of aligned FD curves, while keeping in mind that the misalignment of some of them may impact further results.

9. Searching for Intermediate States

Our entire set of FD curves is now tentatively aligned to the template. Recall that the template profile corresponds to the main unfolding pathway, and that its WLC curves correspond to the main intermediate states. We would like to know whether there might be secondary WLC curves shared by multiple FD curves. These WLC curves would then correspond to common secondary intermediate states.

If there are any secondary WLC curves, then the corresponding peaks across all FD curves should be very similar (potentially varying in their unfolding point, but not in their overall position). If we plot a superposition of all aligned FD curves, we could thus hope to see a secondary peak appear, and know there was a corresponding secondary intermediate state.

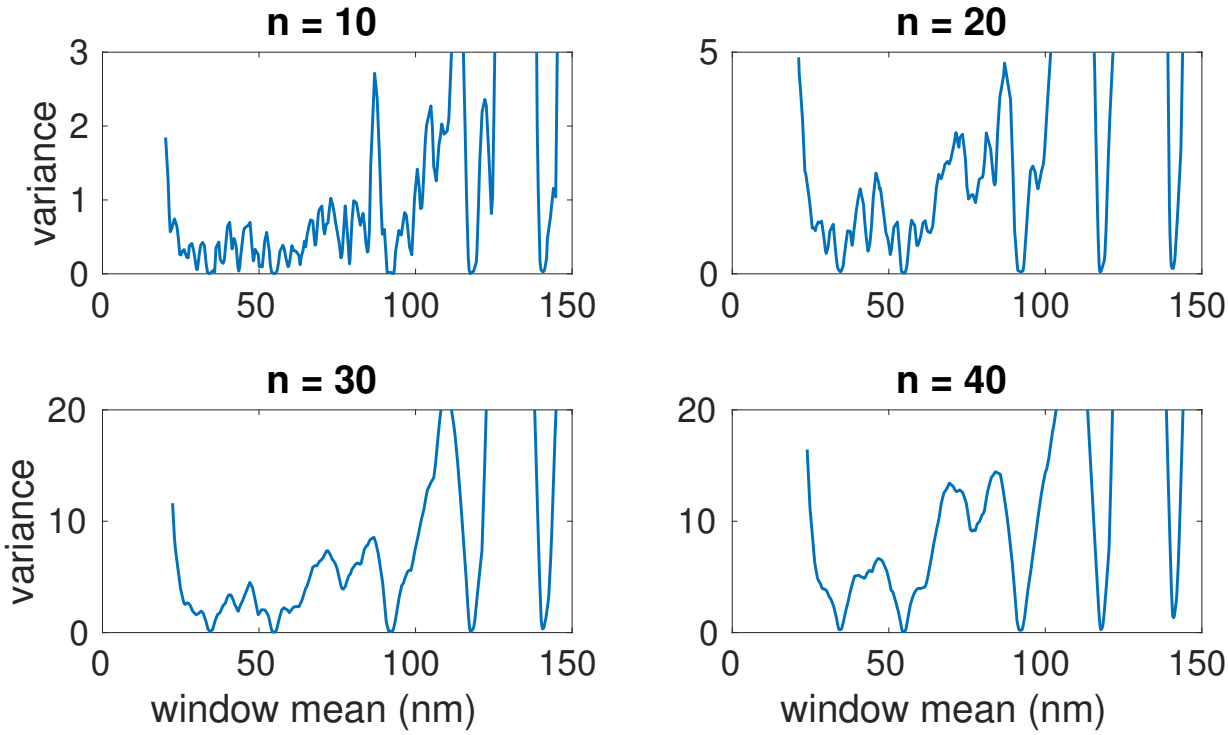


We don't see any clear peaks except for the 5 main ones. This doesn't prove that there are none, but if there are any they will be small and not present across many FD curves. To find those, we will need to be a bit more subtle.

9.1. Clustering of WLC Curves

We try to find WLC curves that are very similar across different FD curves (other than the 5 main WLC curves). To find those clusters of similar WLC curves, we proceed as follows.

We assume a specific cluster size n. For a large range of L_c values, we then compute the closest n WLC curves across all FD curves (max. 1 WLC curve per FD curve), and compute the mean and variance of this set. We then plot the obtained variances against the obtained means. Assuming we used a proper cluster size, clusters of FD curves should appear on this graph as dips in the variance. Let's look at how the variance evolves with the mean L_c for different values of n.



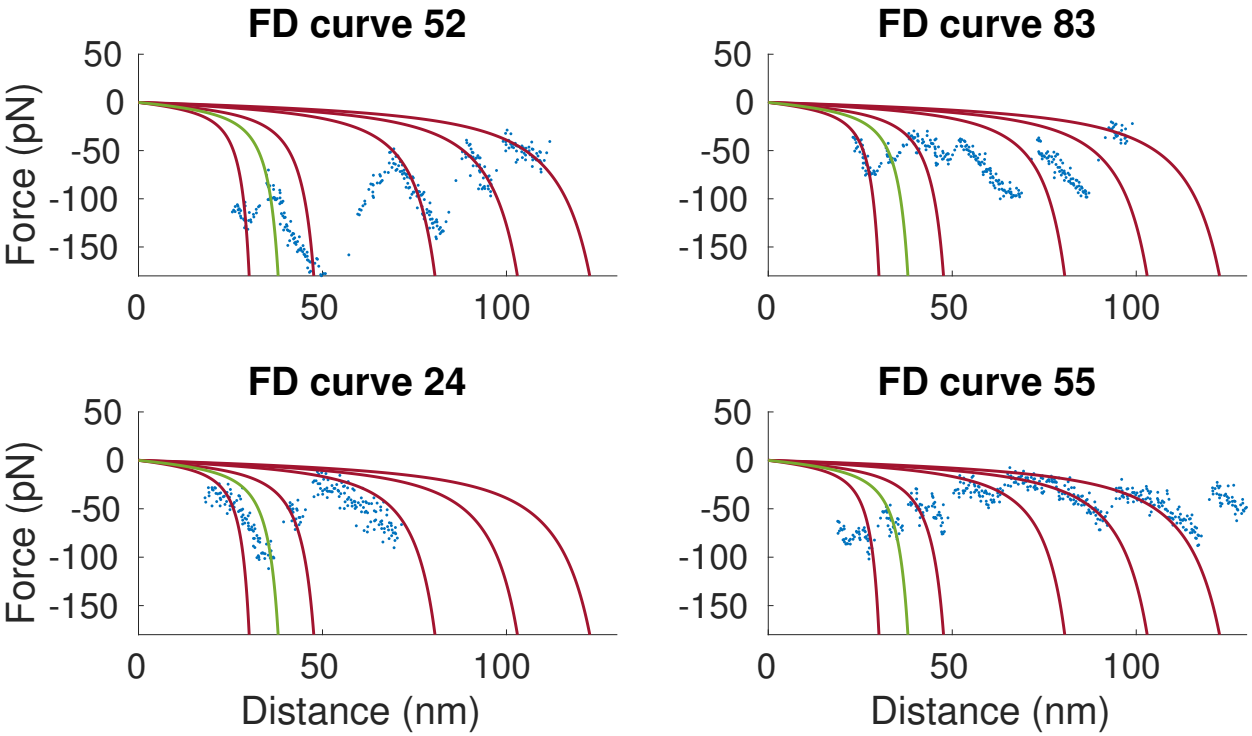
We observe the following things

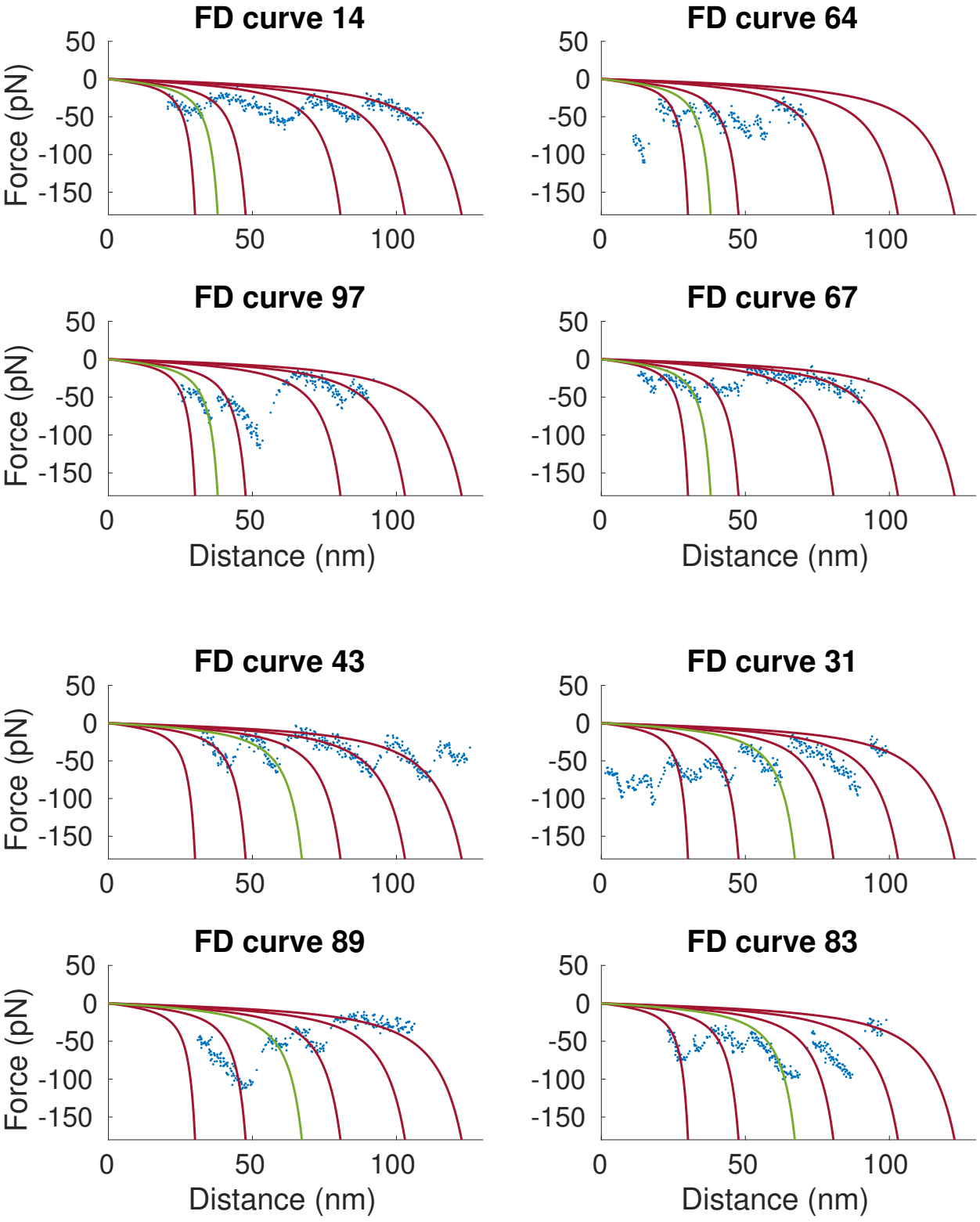
- Overall the 5 main WLC curves appear as consistent dips in each of these graphs, as expected.
- For $n = 10$ and $n = 20$, we can see many small dips in the variance. It is likely that many of those dips are just noise.
- We find two points that are local minima of variance across the different window sizes : one around 43.5 nm and one around 77 nm. Both of these could suggest

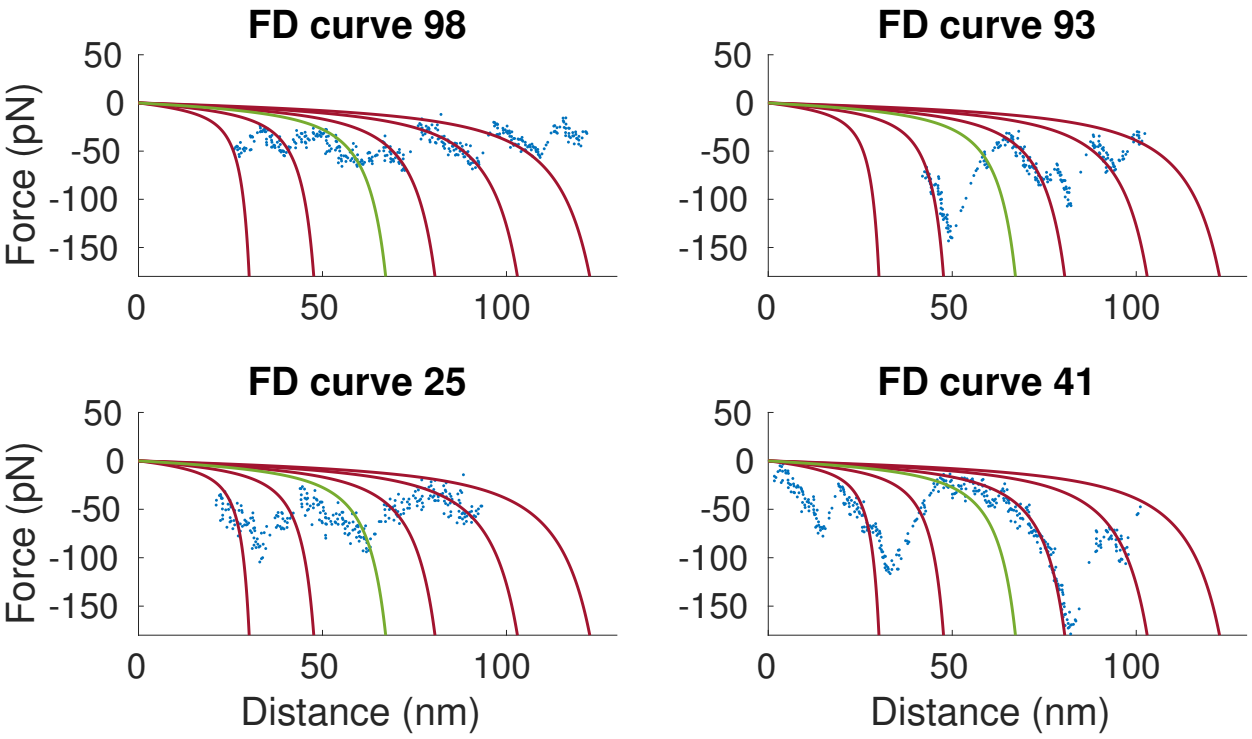
the presence of secondary WLC curves.

We plot the 8 aligned FD curves having the closest WLC curve to 43.5 and 77 nm respectively, to check whether they are proper secondary WLC curves. We want to verify if these WLC curves properly fit their corresponding peak, and if the FD curves seem properly aligned.

The main WLC profile is drawn in red and the tentative secondary WLC curve is drawn in green.




FD curve 14
FD curve 64
FD curve 97
FD curve 67
FD curve 43
FD curve 31
FD curve 89
FD curve 83
Force (pN)
Distance (nm)
50
0
-50
-100
-150
0
50
100



We suspect some of these FD curves to not have been properly aligned, as for example FD curve 43 for which we could have paired the WLC curves differently, shifting the whole profile 20 nm to the left.

Nevertheless, we come to the tentative conclusion that there might be true secondary WLC curves at 43.5 and 77 nm, with a higher confidence for 43.5 than for 77 nm. That is because most FD curves appear properly aligned, and the secondary WLC curves seem to properly fit their peaks. And even if one or a few of the FD curves were shown to be improperly aligned, this wouldn't necessarily invalidate our conclusion.

But we should remain cautious, since we have made many simplifying assumptions to get to this point, when fitting the FD curves, to find the template and when aligning the FD curves. There is thus a very real possibility that we overfitted the data.

10. Conclusion

The objective of this thesis was to analyze the unfolding pathway of the LmrP integral membrane protein. To do so, we had access to a 100 FD curves obtained through Single-Molecule Force Spectroscopy.

The analysis proceeded as follows. First, we preprocessed the FD curves to only keep the peaks corresponding to unfolding events (A). We then explained the Worm-Like Chain model (3), and designed two algorithms to fit it to our FD curves. Our first algorithm, Minima Fit (4), based on the identification of minima, was a bit unsatisfactory. It required the tuning of many parameters, and only fitted the main peaks of a FD curve.

Our second algorithm, Exhaustive Fit (5), based on an exhaustive comparison of possible fitting curves, was an improvement. The number of parameters was much reduced, and it managed to properly fit even very small peaks. We then tried using a Least-Squares step on the peaks identified through Exhaustive Fit to find the proper origin of each FD curve (6) . This approach would have allowed us to align all FD curves and easily compare them, but it proved too fragile to be used in practice.

We then used a translationally invariant clustering technique to find a most common WLC profile (or equivalently, main unfolding pathway) (7). We found that about 50% of our FD curves follow this template WLC profile. We also attempted to align all our FD curves to this template (8).

We then looked for the presence of clusters in the set of all aligned WLC curves (9). We found two such clusters, which correspond to two tentative secondary intermediate states.

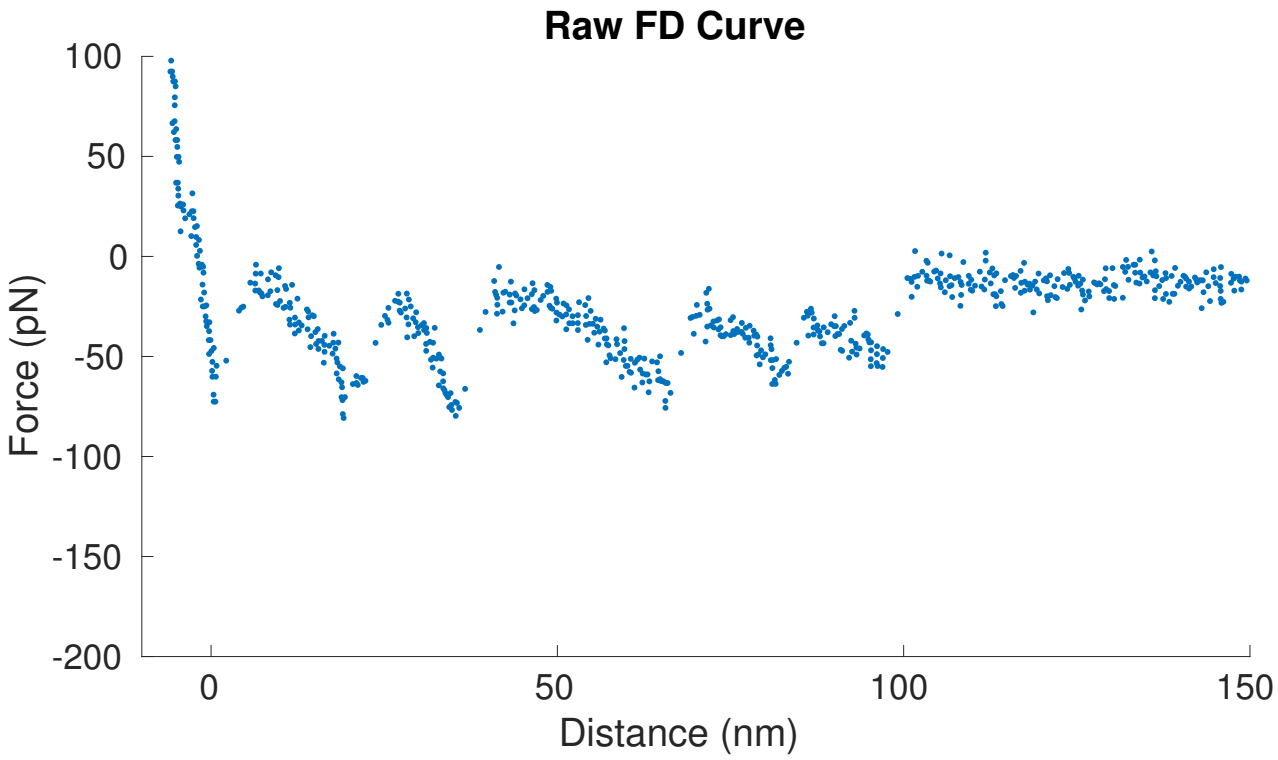
We had to make many assumptions and approximations to get to this result, which is why these secondary intermediate states are only tentative. The fitting step does not give a perfect evaluation of the WLC profile of a FD curve, especially considering the existence of an offset. The template WLC profile found through clustering is also approximative. And when aligning the FD curves to the template, there is a high likelihood of misalignment.

Clearly aligning the different FD curves is the hardest step, and the one most likely to introduce errors in our analysis. A possible technique that we didn't consider here is to use a global sequence alignment approach directly on the FD curves [4]. Finding a more trustworthy alignment technique or improving the experimental precision so as to not introduce an offset on each FD curve would make the analysis of a protein's unfolding pathway and intermediate states much easier.

Appendix

A. FD Curve Preprocessing

A typical FD curve obtained experimentally contains points that are not useful to our analysis, and can safely be removed. This can be illustrated by the following curve.

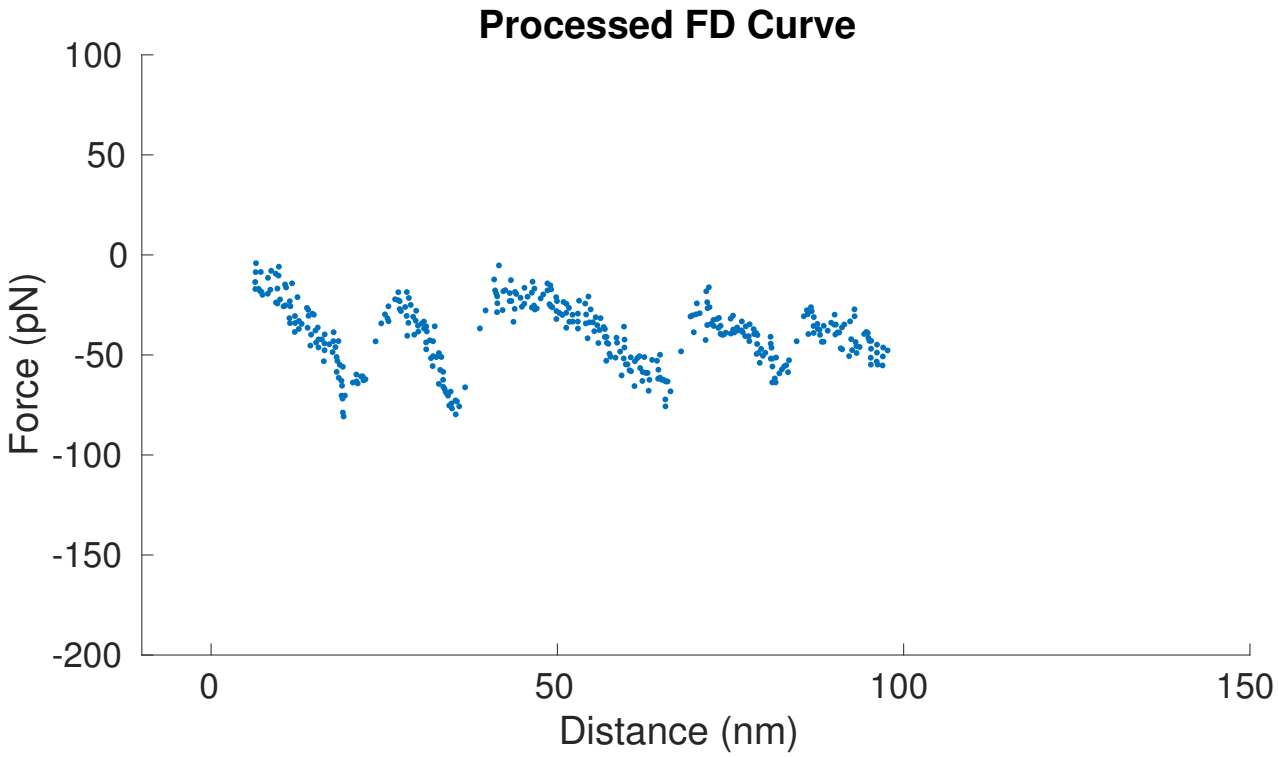


We point out the few things :

- the points at the start of the WLC are not useful to us, as they result from the retraction of the tip from the surface and not the unfolding of a section of protein. The first dip of the force below 0 pN is due to contact forces.
- the points at the end of the FD curve are not useful to us, as they result from the protein not sticking to the surface anymore and not the unfolding of a section of protein.
- in between successive peaks there are FD points going from the unfolding point of the first peak to the start of the next peak. This corresponds to the unfolding of a protein section not actually being instantaneous (as we assumed in our simple WLC model). We will treat them as outliers, but won't remove them in the preprocessing step.

To simplify the problem of fitting a WLC profile to FD curves, we will remove the starting and ending phase of each one of them before attempting a fit. We will do so in a heuristic and ad-hoc fashion, but one that works pretty well : for the starting phase, we remove all points as long they are decreasing, then all points as long as they are increasing. For the ending phase, we compute the mean and standard deviation of the force on the last 100 points. We then successively remove all points starting from the end, until there are at least 5 successive points situated at more than 2.5 standard deviations from the mean.

We get the following corresponding preprocessed FD curve. We are, as expected, only left with the FD points useful for our analysis, the peaks.



B. Factoring-out of L_c

Suppose we want to find the WLC curve that interpolates a FD point (x, F) (with $0 < x$ and $F < 0$). We know that L_c is real and positive and

$$
\begin{aligned}
F &= \frac{k_B T}{l_p}\left(\frac{1}{4\left(1-\frac{x}{L_c}\right)^2} - \frac{1}{4} + \frac{x}{L_c}\right) \\
1 &= \left(\frac{1}{4} - \frac{F l_p}{k_B T} - \frac{x}{L_c}\right) 4(1-\frac{x}{L_c})^2 \\
L_c^3 &= \left((1 - 4\frac{F l_p}{k_B T})L_c - 4x\right)(L_c - x)^2 \text{ multiplying by } L_c^3 \text{ on each side} \\
0 &= 4\frac{F l_p}{k_b T}L_c^3 + 2x(3 - 4\frac{F l_p}{k_b T})L_c^2 - x^2(9 - 4\frac{F l_p}{k_b T})L_c + 4x^3
\end{aligned}
$$

Finding the only real and positive root of this polynomial will give us L_c which defines the WLC curve.

References

[1] Random sample consensus. `https://en.wikipedia.org/wiki/Random_sample_consensus`. Accessed: 2018-05-02.

[2] B. Adreopoulos and D. Labudde. Efficient unfolding pattern recognition in single molecule force spectroscopy data. *Algorithms for Molecular Biology*, June 2011.

[3] R. Litman, S. Korman, A. Bronstein, and S. Avidan. Inverting ransac: Global model detection via inlier rate estimation. *2015 IEEE Conference on Computer Vision and Pattern Recognition*, June 2015.

[4] A. Marsico, D. Labudde, T. Sapre, and D. Müller. A novel pattern recognition algorithm to classify membrane protein unfolding pathways with high-throughput single-molecule force spectroscopy. *Bioinformatics*, 23:e231–e236, January 2007.

[5] J. Milstein and J. Meiners. *Worm-Like Chain (WLC) Model*, pages 2757–2760. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[6] D. Müller, M. Kessler, F. Oesterhelt, C. Möller, D. Oesterhelt, and H. Gaub. Stability of bacteriorhodopsin alpha-helices and loops analyzed by single-molecule force spectroscopy. *Biophysical Journal*, 83:3578–3588, December 2002.

[7] M. Rief, M. Gautel, F. Oesterhelt, JM. Fernandez, and HE. Gaub. Reversible unfolding of individual titin immunoglobulin domains by afm. *Science*, 276:1109–1112, May 1997.

[8] G. Tsaousis, K. Tsirigos, X. Adrianou, T. Liakopoulos, P. Bagos, and S. Hamodrakas. Extopodb: a database of experimentally derived topological models of transmembrane proteins. *Bioinformatics*, 26:2490–2492, October 2010.

[9] H. Yu, M. Siewny, D. Edwards, A. Sanders, and T. Perkins. Hidden dynamics in the unfolding of individual bacteriorhodopsin proteins. *Science*, 355:945–950, March 2017.

Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve **www.uclouvain.be/epl**