
Fodis

A Software for Single Molecule Force Spectroscopy

User Guide

Nicola Galvanetto

TRIESTE, ITALY

JUNE 2017

Contents

Preface	3
Installation	5
1 Basics	7
1.1 Software Framework	7
1.2 Open data	8
1.2.1 Import traces	8
1.2.2 Open samples	9
1.3 Data saving	10
1.4 Single trace representation - Plot data	11
1.4.1 Traces	11
1.4.2 Traces-Lc	12
1.4.3 Dynamic persistence length	12
1.4.4 Contour length (Lc,Fc)	13
1.4.5 Contour length histogram	13
1.4.6 Delta Lc histograms	13
1.5 Global trace representation	14
1.5.1 Global contour length histogram	14
1.5.2 Global histogram max	14
1.5.3 Global contour length - Force plot	15
1.5.4 Global delta Lc histograms (and plots)	15
1.5.5 Global peaks (or Global Matrix)	16
1.5.6 Global persistence length	16
1.6 Graphical features	17
1.6.1 Density plots	17

1.6.2	Parameters Box	17
1.6.3	Export plot	19
1.7	Selected & Valid Traces	19
1.8	Trace filtering: basic tool	20
1.9	Align to <code>zero</code>	22
1.10	Manual Alignment	22
2	Advanced Tools	25
2.1	Path Analysis	25
2.1.1	Path Plot	26
2.1.2	Combined Path Plot	30
2.2	Fingerprint ROI	30
2.3	Automatic Alignment	32
2.4	Selection through correlation	35
	Bibliography	39

Preface

Fodis (for “Force-distance software”) is a program for the analysis of Single Molecule Force Spectroscopy data. It is mainly intended for automated analysis of force-distance curves of Protein Unfolding Events obtained during Atomic Force Spectroscopy experiments. However, it can be used for visualization and organization of any force-distance curve. Fodis provides a large number of force-curve visualization option, including function for basic statistical analysis. Since the developers are active AFM users, the program also contains some specific data processing methods intended for protein unfolding studies. Fodis is Free and Open Source software, covered by CC Public License. It was born in *Matlab* and we encourage third-parties to participate in developing functions and modules. Being a free software, it provides the source code to developers and users, which makes easier both the verification of its data processing algorithms and any further program improvements. Fodis was written in *Matlab 2016b*.

Installation

Fodis can be downloaded from <https://github.com/nicolagalvanetto/Fodis>. Users need to

- unzip *Fodis* folder and move it to the Matlab **Current Folder**
- right clik on the folder and select **Add to path folder and subfolders**
- type **Fodis** in the Command Window and **Enter**

Non-Matlab users working on Windows, Mac or Linux need to launch the installer of their own version and follow the instructions. Please read the **README.txt** file before installation.

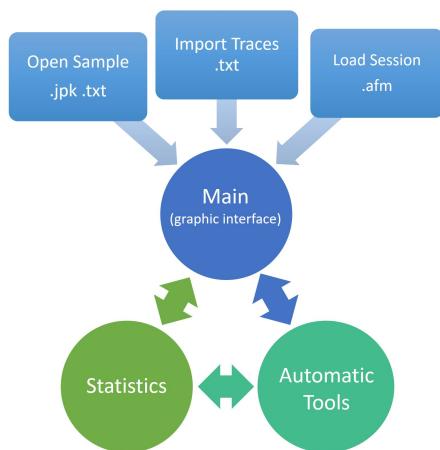
For any problem related to the software, please use GitHub issue section <https://github.com/nicolagalvanetto/Fodis/issues>.

Chapter 1

Basics

In this chapter we present the basic working rules of the software and how to use it.

1.1 Software Framework



The general architecture of Fodis and the relationship among the various *blocks* is shown in this diagram. *Main* is the hub, everything can be managed into that. *Main* is linked to different input blocks, which are designed to facilitate expansion to any desired input file formats.

1.2 Open data

There are three modalities to import data and start working with Fodis.

File	Tools	Filter	Align	Path A
Open samples				Ctrl+O
Load session				Ctrl+L
Save session				Ctrl+S
Import traces				Ctrl+G
Export traces				Ctrl+K
Clear all				Ctrl+8
Remove duplicates				Ctrl+0

- **Import traces** uploads matrices of numbers. This is the most universal import option in Fodis (see example below).
- **Open samples** uploads folders containing multiple files of single traces. The format allowed is listed in the GUI menu.
- **Load session** uploads .afm files previously generated by Fodis through **Save session**. These files contain not only the saved dataset, but also the specific configuration and parameters set in the previous working sessions.

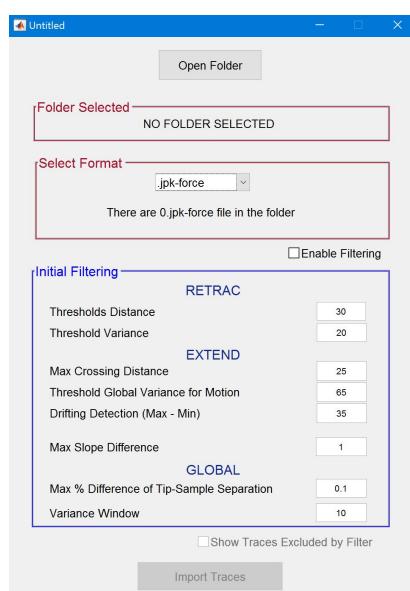
1.2.1 Import traces

Force 1	tss 1	Force 2	tss 2	Force 3	tss 3
N	0	N	0	N	0
-6.18289e-11	-2.09318e-08	-6.846298e-13	+9.80119e-09	+9.20008e-12	
-6.69519e-11	-2.83687e-08	-1.40984e-11	-8.75944e-09	5.62841e-12	
-1.10949e-10	-2.76247e-08	-1.67578e-11	-8.48523e-09	-2.10615e-13	
-4.42698e-11	-2.81514e-08	-2.99126e-12	-8.08778e-09	9.39727e-13	
-4.19844e-11	-2.77138e-08	-2.81514e-12	-8.10166e-09	1.48811e-13	
-7.05283e-11	-2.73118e-08	-2.60602e-12	-8.20984e-09	-1.88353e-12	
-4.50939e-11	-2.74242e-08	-5.93442e-12	-7.63895e-09	-5.39043e-12	
-1.64628e-11	-2.75198e-08	-1.04414e-11	-7.34291e-09	2.69829e-11	
-4.36695e-11	-2.69098e-08	9.21356e-12	-7.32972e-09	3.84139e-12	
-3.95089e-11	-2.64891e-08	-1.46133e-12	-6.97394e-09	6.49191e-12	
-2.25884e-11	-2.67271e-08	-5.54563e-13	-6.729e-09	1.96454e-11	
-3.25583e-11	-2.63701e-08	-2.35804e-12	-6.51598e-09	7.8957e-12	
-3.50057e-11	-2.61013e-08	-3.30533e-12	-6.21103e-09	7.48038e-12	
-1.42995e-11	-2.61013e-08	-6.43045e-12	-6.93213e-09	2.05564e-11	
-1.19046e-11	-2.60994e-08	-6.43045e-12	-6.93213e-09	-5.66290e-11	
-2.19373e-11	-2.55334e-08	-1.04769e-11	-5.39704e-09	2.35925e-12	
-1.15546e-12	-2.55356e-08	-1.73032e-11	-5.07372e-09	1.28047e-11	
-2.36624e-13	-2.53064e-08	3.46724e-13	-5.03688e-09	2.15476e-11	
-2.24945e-13	-2.51186e-08	-4.46734e-12	-5.03688e-09	6.41915e-11	
-4.16734e-13	-2.48228e-08	-1.17071e-12	-4.46816e-09	3.93916e-11	
1.72054e-13	-2.45898e-08	6.74303e-12	-3.82025e-09	2.10162e-11	
4.83275e-12	-2.42905e-08	8.49396e-12	-4.11238e-09	2.11776e-11	
-8.08149e-12	-2.40239e-08	-2.00928e-12	-5.20611e-11	-5.32441e-11	

Every .txt file can be imported as long as it follows these simple rules: The .txt file must be organized in columns. The first column contains the force values of the first trace in Newton, the second column contains *tip-sample-separation* values (or polymer extension) in meters, and so on. Columns are space-separated.

N.B. By convention, in Fodis, keep repulsive forces *negative*, attractive forces *positive*. (*Please check additional datasets for more details*).

1.2.2 Open samples



Open sample user interface is designed to help the user to upload JPK or Bruker - exported traces. First select the desired format on the **Select format** menu and then **Open Folder**.

- **JPK .txt** : JPK Data Processing software allows to process curves, transforming default input parameter (*Head Height*) into *tss*, and at the same time extract force-tss curves as .txt files. Those files can be directly imported in Fodis. (For those who have trouble in exporting .txt files from JPK Data Processing, Fodis can import directly .jpk-force files, already processed to tss, but we encourage the use of .txt for celerity). An example of JPK-extracted .txt files can be found in *additional datasets*.

- **Bruker .txt** : Nanoscope Analysis allows to process curves, transforming default input parameter (*Distance*) into *tss* (Nanoscope command **Baseline Correction x2**), and at the same time extract force-tss curves as .txt files. Resulting .txt files can be imported in Fodis. An example of Bruker-extracted .txt files can be found in *additional datasets*.

During this process, it is possible to enable coarse filtering for the parameters reported in **Initial Filtering** box.

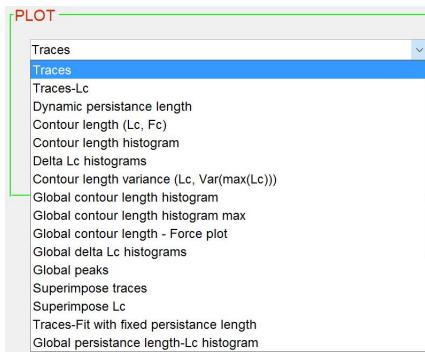
1.3 Data saving

Any time, during a working session it is possible to save the data in two different modalities:

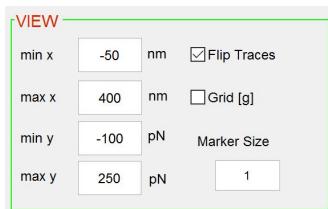
- **Export traces** generates a .txt of the same format presented in the **Import traces** section. It is highly recommended to use this format for portability reasons.
- **Save session** generates a .afm file that can be uploaded only by Fodis through the **Load session**. These files contain not only the current dataset, but also the specific configuration and parameters set during the working session.

1.4 Single trace representation - Plot data

In this section we present all possible trace representations of single events implemented in Fodis.

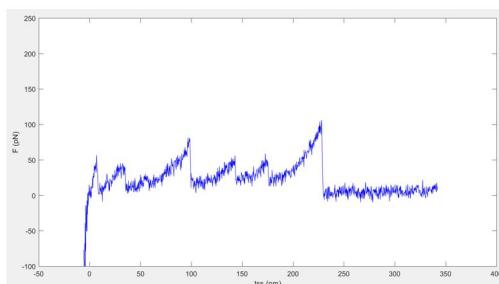


The plot-box at the top-left corner controls the graph displayed in the central panel. There are two major groups among these graphical representations: single trace representations and global representations of all selected traces.

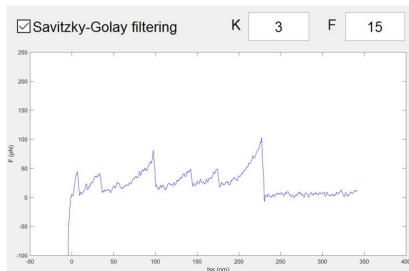


In the view-box, all the parameters necessary to control plotting range, marker size and grid are present. To perform **Export plot** click on **tool** in the menu bar: this creates a Matlab figure containing all the objects present in the Fodis graphic panel.

1.4.1 Traces

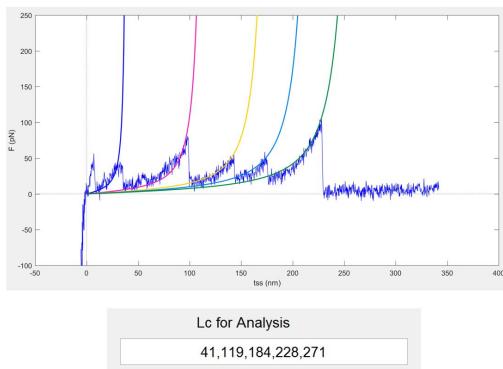


Typical *Force-tip sample separation* curve: exact representation of input numbers. The slide bar at the lower part of the panel allows to move across different loaded traces.



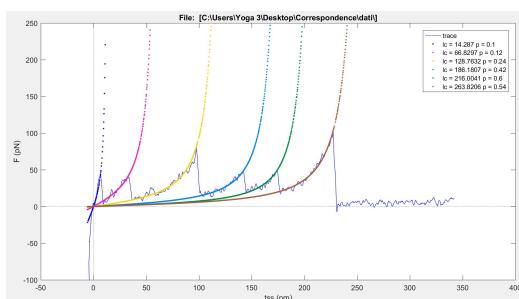
Trace smoothing performed through a discontinuity-preserving filter (Savitzky-Golay algorithm). K is the polynomial order of the interpolant curve while F is the number of points used for interpolation (F must be an odd number).

1.4.2 Traces-Lc



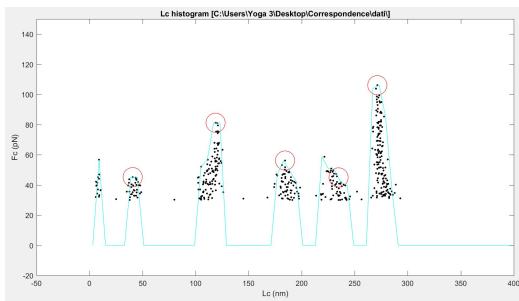
Traces-Lc finds relevant peaks and fits them with the WLC equation. *Peak detection is enhanced on smoothed traces.* The numbers reported in Lc for Analysis are all detected contour lengths (in nm) of plotted rainbow curves. In case a peak is not detected, type the number in the edit box and then enter. That new Lc will be saved for all the following analyses.

1.4.3 Dynamic persistence length



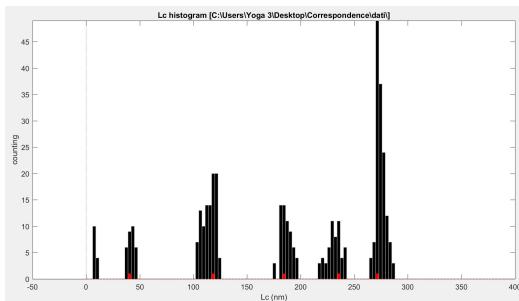
The persistence length, which is typically kept fixed ($p=0.4\text{nm}$ for a.a. chains), in dynamic mode is relaxed to find the best p -Lc couple for each peak. The final parameters are plotted in the legend.

1.4.4 Contour length (Lc,Fc)



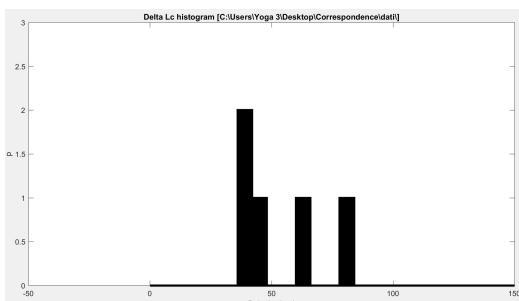
This representation transforms every point above a threshold force (in this case 30 pN, tunable in Parameters box **Min F**) from F-tss space to F-Lc space according to the transformation presented by Puncher and colleagues [1, 2].

1.4.5 Contour length histogram



This is the histogram of Contour length (Lc,Fc) representation. It is also called Barrier position histogram [1]. The red bars represent the detected peaks. Bin width can be tuned in the Parameters box.

1.4.6 Delta Lc histograms

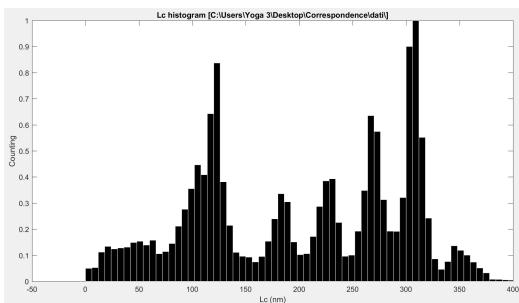


Delta Lc histograms shows the value of the difference in Lc between two consecutive peaks. The first peak's Delta-Lc, by definition, is equal to the first Lc.

1.5 Global trace representation

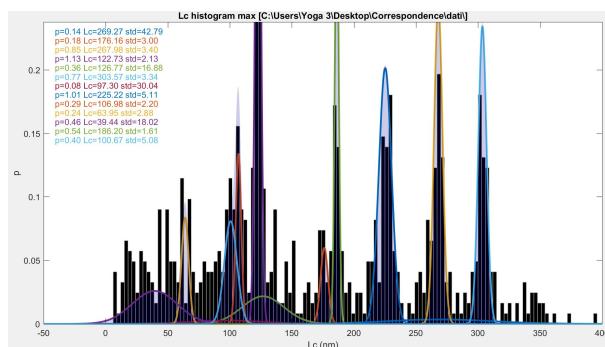
Single molecule force spectroscopy experiments generate a large amount of data. Single trace representations are useful to establish the quality and features of single events, but only a statistical evaluation of Lc distribution or other specific features allows to perform a quantitative analysis. For this reason, Fodis includes a variety of global representations for the population of traces.

1.5.1 Global contour length histogram



The Global contour length histogram is the sum of all contour length histograms, normalized on the highest bin. Bin size can be tuned in the Parameters box.

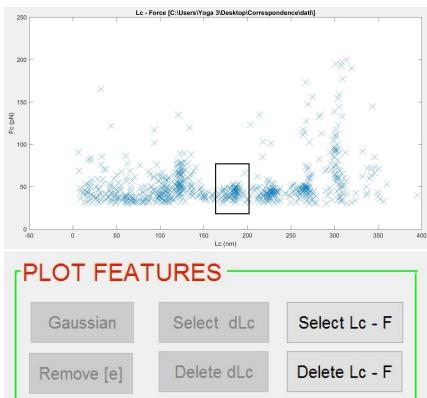
1.5.2 Global histogram max



The Global histogram of Maxima (also called *Contour length histogram* by Kawamura and colleagues [3]) counts the detected peaks of all se-

lected traces (*i.e.* red bars shown in **Contour length histograms** section). The resulting distribution can be fitted with the Gaussian Mixture Model (GMM) checking **Tools > Automatic multi Gaussian** option on the menu bar. This representation is useful to determine the probability of a certain unfolding event happening along the Lc coordinate. (Bin size can be tuned in the Parameters box).

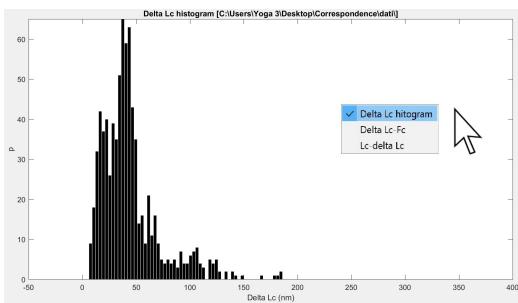
1.5.3 Global contour length - Force plot



The *Lc - Force plot* shows the contour length *vs* Force of all detected peaks. This plot can be used also to detect the detachment force in binding experiments (*e.g.* with antibodies).

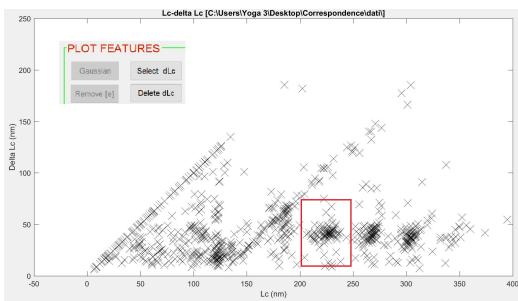
If the user is interested on the force distribution of a particular subgroup of peaks (*i.e.* black rectangle), using **Select Lc-F** button it is possible to select a region (draw a rectangle and double click inside de area). A force distribution of the selected points will be automatically generated.

1.5.4 Global delta Lc histograms (and plots)



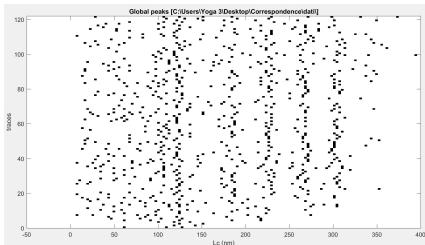
Delta Lc is defined as the Lc difference between two consecutive peaks. On the left, histogram of all Delta Lc values of the population.

N.B. Right clicking on the graphs allows the user to change the view to **Delta Lc-Fc** or **Lc-Delta Lc**.



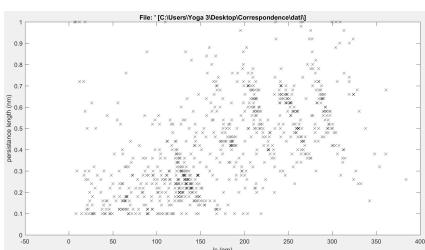
The Lc-Delta Lc plot shows the scatter plot of all detected peaks. If the user wants the Delta Lc distribution of a particular subgroup of peaks (i.e. red rectangle), using Select dLc button it is possible to select a region (draw a rectangle and double click inside de area). A Delta Lc distribution of the selected points will be automatically generated.

1.5.5 Global peaks (or Global Matrix)



Automatically detected peaks are used to form a string of ones and zeros for each trace (zeros where there are no peaks, ones where there are a peak). It is a sort of binarized coding for the traces. Each string is therefore plotted one above the other in **Global peaks** with white pixels for zeros and black pixels for ones.

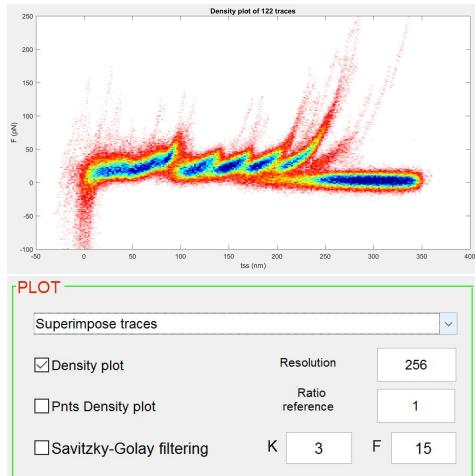
1.5.6 Global persistence length



All the peaks of the population are fitted with two free parameters - persistence length (p) and Lc as described in **Dynamic persistence length-** and plotted. This allows to find correlations between the peaks positions and their curvature(i.e. p).

1.6 Graphical features

1.6.1 Density plots



A heat map is plotted when the **Density plot** check box is selected. The image is constituted by pixels (**Resolution**) and the color of the pixel is proportional to the number of points contained in the pixel area. **Ratio reference** changes the heat map color limits. The **Points Density plot** highlights each point with the color of the corresponding pixel. Density plot is available for:

- Superimpose Traces
- Superimpose Lc
- Global Lc - Delta Lc

1.6.2 Parameters Box

Many parameters can be set to tune both graphical views and threshold limits (parameters not described here will be discussed later).

- **Min F**: minimum value of force valid for peaks detection or Lc evaluation;
- **Max F**: maximum value of force valid for peaks detection or Lc evaluation;
- **Min peak proximity**: minimum interval for peaks detection in **Global histogram max**;
- **Threshold N points**: parameter related to peaks detection. The lower N, the easier the peaks to be found;

PARAMETERS BOX

LIMITS		BIN SIZE		PERSISTENCE LENGTH	
30	Min F (pN)	3	Lc bin size (nm)	0.4	Persistence Length p (nm)
500	Max F (pN)	10	Fc bin size (pN)	0.05	p step (free p) (nm)
6	Min peak proximity (nm)	2	Delta Lc bin size (nm)	0.2	Min p (free p) (nm)
2	Threshold N points			1	Max p (free p) (nm)
1000	Max Lc (nm)				
0	Min Lc for delta Lc (nm)				
500	Max Lc for delta Lc (nm)				

Lc for Analysis

Update Settings

- **Max Lc:** maximum value of Lc taken into account;
- **Min Lc for Delta Lc:** minimum value of Lc taken into account for Delta Lc plots;
- **Max Lc for Delta Lc:** maximum value of Lc taken into account for Delta Lc plots;
- **Lc bin size:** bin width of Lc histograms or Global Lc representations;
- **Fc bin size:** bin width of Force histograms or Global Force representations;
- **Delta Lc bin size:** bin width of Delta Lc histograms or Global Delta Lc representations;
- **Persistence Length p:** fixed value of persistence length for Lc evaluation;
- **p step (free p):** minimum step when evaluating free persistence length (dynamic persistence length);

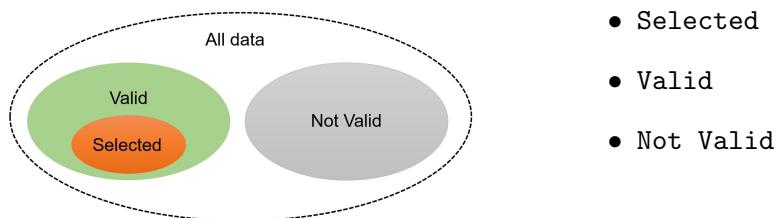
- **Min p (free p)**: minimum value for dynamic persistence length algorithm;
- **Max p (free p)**: maximum value for dynamic persistence length algorithm;
- **Lc for Analysis**: detected Lc values of a current trace (they can be edited);
- **Update Settings**: after parameters changes, click this button to update displayed results.

1.6.3 Export plot

Every plot can be extracted and saved individually in Matlab format .fig, or any other file format for figures, from the menu bar **Tools > Export plot**.

1.7 Selected & Valid Traces

Loaded traces have a *visibility label*. There are only three possible labels:



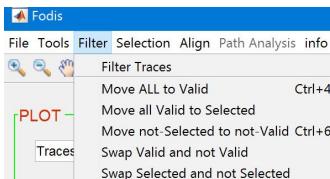
By default, loaded traces are all **Selected**. Global representations act on **Selected** (that is a sub-group of **Valid**). All **Valid** traces are displayed moving the slide bar. **Not Valid** traces are not shown. Filtering procedures described in the next section and chapter, filter out traces labeling them as **Not Valid**.

DETAILS	
Valid Traces	101/122
Selected Traces	85/101

Details Box shows the numbers:
85 Selected,
101 Valid,
122 Valid plus Not Valid, i.e. all loaded data.

SELECT TRACES	
<input type="checkbox"/> Remove [r]	
Offset (nm)	0
Selected Traces	17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36

Select traces Box allows to edit selected traces. Removed traces remain Valid but are not Selected any more.



In the Filter menu bar the user can manage the *visibility labels* of loaded data:

- Move ALL to valid: all Not Valid are labeled Valid;
- Move all Valid to Selected: all Valid are labeled Selected;
- Move not-Selected to not-Valid: all Valid but not Selected are set Not Valid;
- Swap Valid and not Valid: all Valid, therefore also Selected, are labeled Not Valid, and vice versa;
- Swap Selected and not Selected: all Selected are labeled Valid, and vice versa; Not Valid remains unchanged.

1.8 Trace filtering: basic tool

SMFS experiments typically generate a large amount of data. In this section we present a basic filtering tool that allows the user to select traces

upon two characteristics: points position in F-tss space and peaks position. Traces that do not satisfy desired requirements will be labeled as **Not Valid**.



The top panel displays single traces. On this panel, the user can draw two kinds of rectangular ROIs: mandatory regions (green) and forbidden regions (red). **Green** ROIs define Force and tss intervals in which at least one point of the trace must fall into, otherwise the trace is filtered out. **Red** ROIs define Force and tss intervals in which no point of the trace can fall into, otherwise the trace is filtered out. The ROIs are equal for all the traces. To draw a ROI, the user needs to click on **Draw Rectangle ROI** and then move on the panel and click&drag; to erase a ROI, just click **Erase ROI**. Multiple ROIs can be drawn.

The bottom panel displays the single contour length histograms in red and the detected peaks in black. Color limits are set by the four edit boxes on the right. The **Blue** section defines the area in which at least one peak of the trace must fall, otherwise the trace is filtered out. The **Pink** section defines the area in which the last peak of the trace must fall, i.e. no peaks

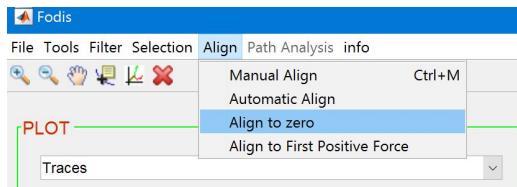
on higher Lc values, otherwise the trace is filtered out.

N.B. We suggest to perform this filtering procedure with traces smoothed with Savitzky-Golay algorithm in order to enhance peaks detection.

Once all the desired limits have been set, **Compute Filter** finds the traces to be removed and **Update Valid Traces** updates the **Not Valid** group.

1.9 Align to zero

Loaded traces can be all aligned to a zero position. There are two **zero** definitions:



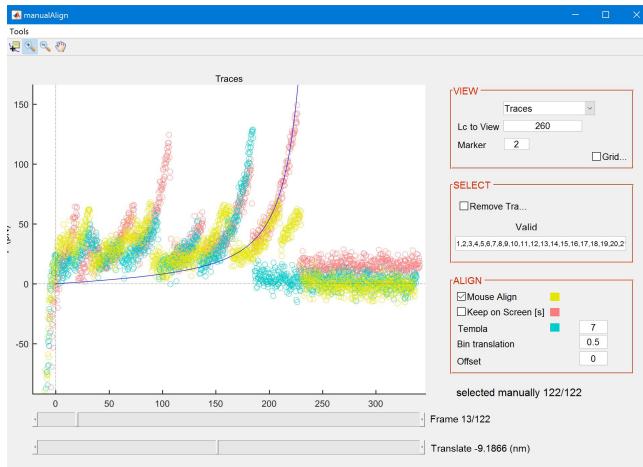
- Align to zero
- Align to first positive force

Align to zero finds the average value of tss corresponding to negative forces between -200pN and -400pN, and translates each trace accordingly. Sometimes this procedure may not work because the contact phases are not homogeneous: in this case Align to first positive force sets to zero the tss corresponding to the first positive value of Force of each trace.

1.10 Manual Alignment

Manual Alignment tool (accessible from the menu bar **Align > Manual alignment**) allows the user to slightly refine the relative shift among valid traces, typically caused by misalignment of the base line. The graphical panel shows three traces of different colors:

- Yellow - chosen trace, its number is visible in the **Frame** text box;
- Green - template trace, editable in the **Template** edit box;
- Pink - trace(s) saved on screen, **Keep on screen** check box.



Traces can be moved only horizontally, both with the Translate slide bar and clicking&dragging on the graphic panel. Only chosen trace (yellow trace) can be translated. To move across valid traces, use Frame slide bar. **Select** box contains all the information related to Selected traces, the chosen trace can be removed editing the selected edit box or by clicking on **Remove trace** check box. **View** box allows to decide the trace representation and to set an Lc of reference.

Chapter 2

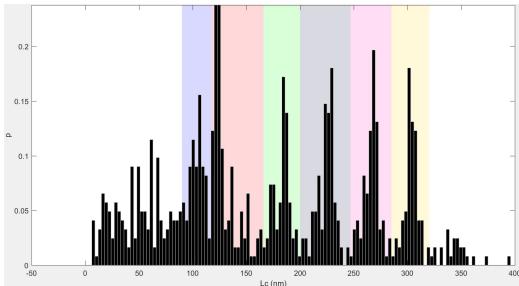
Advanced Tools

In this chapter we describe the tools implemented in Fodis. These tools were developed in order to solve some common issues in SMFS experiments, like clustering and alignment.

2.1 Path Analysis

Given a population of traces that represent the unfolding of a certain protein, an important issue in SMFS is to determine *how* this unfolding happens. Global representations presented in Chapter 1, for instance, are very useful to generate statistics in an *aggregated* fashion. Here, with Path Analysis we intend to present three representations that, in spite of being *global*, show unfolding pathways in a *disaggregated* fashion.

2.1.1 Path Plot



The idea is to generate a graphical representation of the different unfolding pathways of a set of traces, and cluster them. Starting from Global Histogram of Maxima (GHM), we can divide the Lc coordinates in intervals (different colors).



The extremities of each interval can be set in the **Path Interval** edit box. Intervals cannot be overlapping.

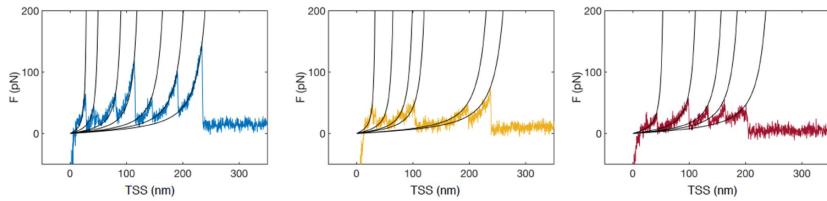
There are two different ways to construct intervals:

1. **min1-max1,min2-max2,...** in nm;
2. **N** where N is an integer: divide Lc coordinates in N equal intervals.

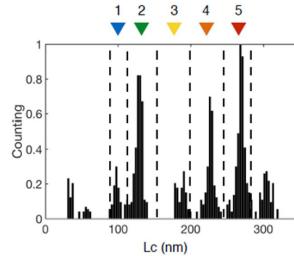
Once the intervals have been chosen, the user can update the group division by clicking on **Update Grouping**. Then checking **Enable Group Division** it is possible to move across different resulting groups with the menu **Show Group Nr.**.

Path Plot construction

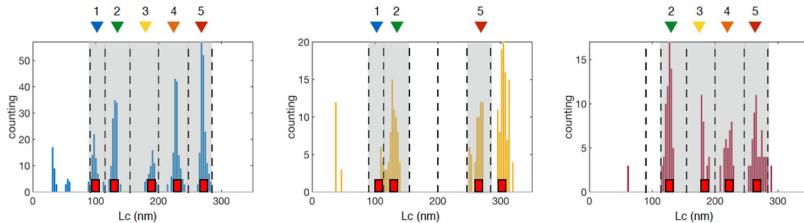
Group division, and resulting *Path Plot* generation, is based on the following 5 steps.



Traces are fitted with the WLC model, transformed into Lc histograms and grouped in the Global Histogram (or GHM).



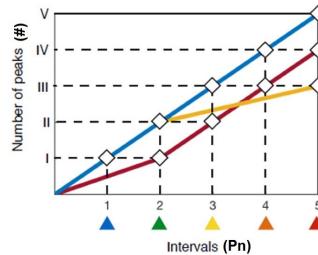
The Global Histogram is divided into intervals according to with the main ensembles of occurring events, to partition the Lc coordinates in regions with distinct maxima.



Following the aforementioned rule, in this explanatory panel, we selected 5 intervals 90-118,118-155,155-200,200-247,247-285nm.

1 1 1 1 1	1 1 0 0 1	0 1 1 1 1
Pn = 1 2 3 4 5	Pn = 1 2 5	Pn = 2 3 4 5
# = 1 2 3 4 5	# = 1 2 3	# = 1 2 3 4

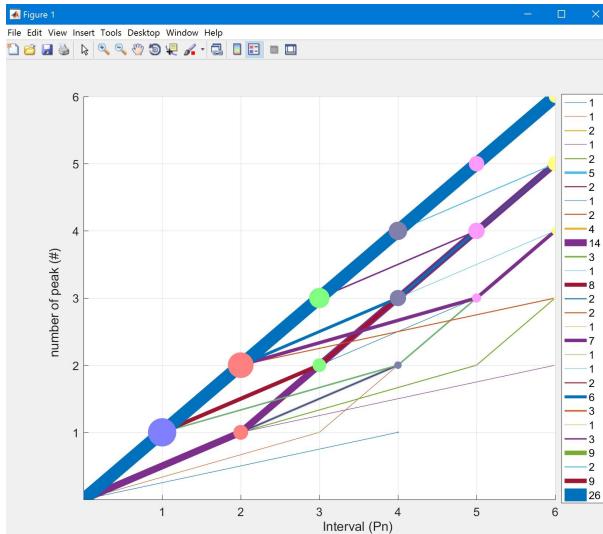
On the base of this division, each trace is coded in a binary string of 5 digits: 0 is assigned if no force peak is detected within the interval, 1 is assigned if at least one event is detected. From each string, we created two additional sequences: $\#$ and P_n . $\#$ is the sequence referred to the order of appearance of the force peak along the trace (in a trace with 2 peaks, the 1st peak has $\#=1$ and the 2nd peak has $\#=2$). P_n refers to the interval position occupied by a peak (a peak that falls within the 3rd interval has $P_n=3$).



Traces so coded are finally plotted as broken lines into an orthogonal $\#$ - P_n space, the line-width is proportional to the number of traces that follow the same path. This method provides a representation to distinguish different unfolding behaviors/clusters of a given set of traces, based on the number and position of occurrence of unfolding events.

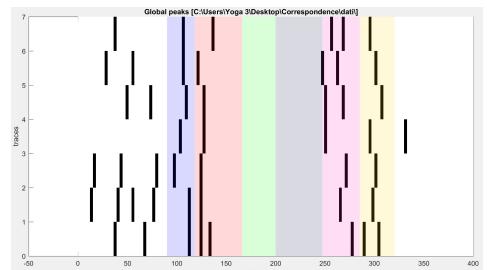
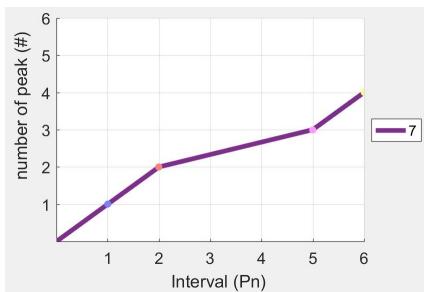
Path Plot of a larger population

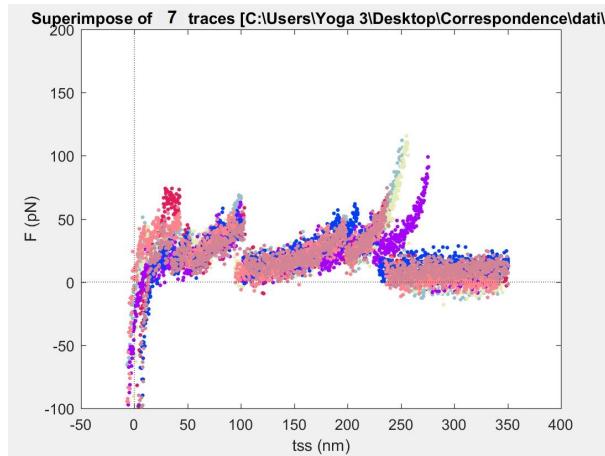
The previous example shows how the Path Plot is constructed, now we present how it looks when generated on a larger dataset (CNG dataset).



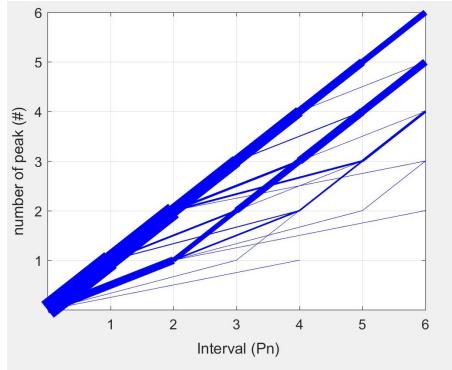
Each line width is proportional to the number of traces that cover that specific path. The size of the point at a node is proportional to the number of traces that pass through that node.

By Enabling Group Division we can select only one group, the 18th in this case. We can see the peaks distribution and relative Path Plot of 7 traces that has no peak in intervals 3 and 4, indeed, in the *Path Plot* representation, it does not have any node in the 3rd and 4th intervals.





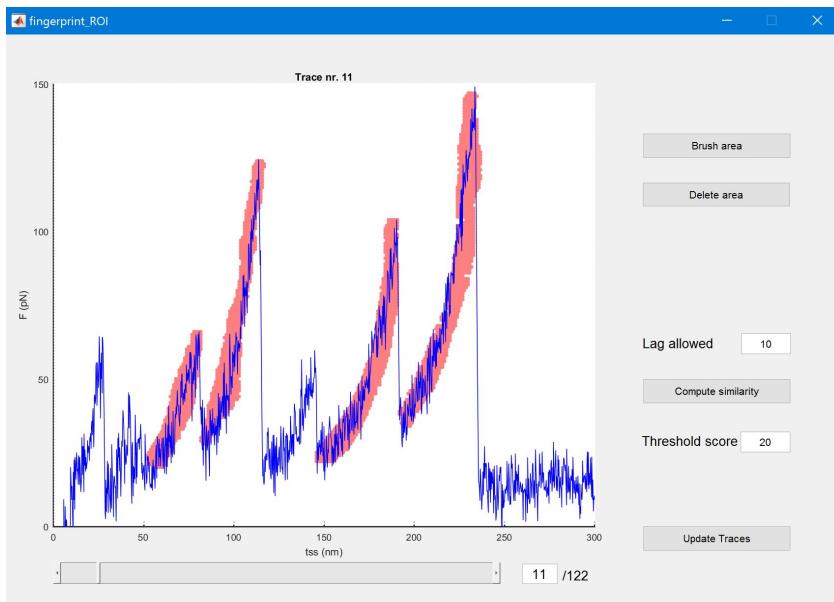
2.1.2 Combined Path Plot



Similar to Path plot, *Combined Path Plot* displays #-Pn space but with no color distinction. Each segment-width is proportional to the absolute number of traces that travel over that particular segment path.

2.2 Fingerprint ROI

Fingerprint ROI is a tool intended to find specific pattern over the population of traces. It is accessible from the menu bar **Selection > Fingerprint**



ROI.

This graphic interface shows the plot panel and the slide bar. The general idea behind this tool is to evaluate how many points of a trace fall into the drawn area (pink area). By clicking the **Brush area** button, move the mouse cursor on the graph and finally click&drag to draw a continuous ROI. The user can draw as many ROIs as needed just by clicking the button **Brush area** for any new ROI. The **Delete area** button deletes all ROIs previously drawn. Once all desired ROIs have been drawn, the evaluation can start. **Lag allowed** edit box sets how many nm each trace is shifted from its starting position. The maximum value of intersection between ROIs and a trace is saved and displayed with **Compute similarity**. The **Threshold score** edit box sets the threshold value of the score: the **Update Traces** button filters out all the traces that score lower than the threshold value and then they are available in the main Fodis interface.

2.3 Automatic Alignment

The Automatic Alignment (menu bar `Align > Automatic Alignment`) is a tool intended to solve the critical and highly operator-dependent issue of *traces alignment*. This tool is based on the work of Bosshart and colleagues [2], with some specific improvements developed to extend its applicability to more general SMFS datasets.

In particular, they proposed an algorithm consisting of 4 steps. Starting from the `contour length histogram` of every trace, they:

- subdivided the traces into groups of homogeneous traces (i.e. traces with the same number of peaks);
- recursively aligned traces into the same group with the maximum-correlation principle, building an average contour-length reference for each group;
- formed a *global reference*;
- aligned all the traces of the dataset to the *global reference*.

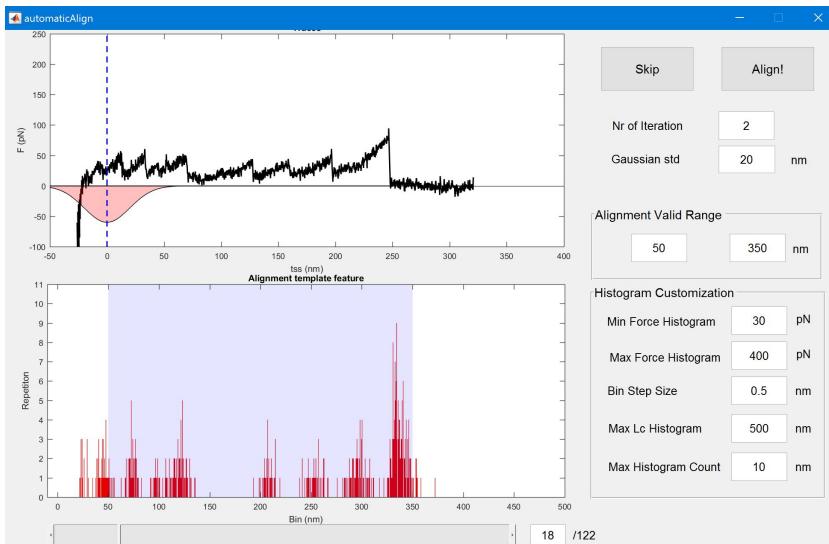
To this remarkable computational procedure, we added two features that help to manage a larger variety of traces that do not display regular occurrence of peaks.

1. In addition to the contour-length histogram, we assign to every trace a zero-point, that is the point of tip-sample contact. Given the correlation curve of two traces, we then multiply the correlation curve with a Gaussian curve centered at the point in which the zero-points of the two traces match with each other (red Gaussian curve). The idea is to apply a “potential well” to reduce the maximum displacement of the two zeros.
2. Group division proposed by Bosshart and colleagues works only if all the traces with the same number of peaks have the peaks in the same position, but this is not generally true for F-D curves of the same protein. Therefore, we used a group division following the method described in the *Path Plot* section. In this way, we imposed two constraints for a given trace to be part of a given group: to have

a specific number of peaks and to have those peaks in a specific position.

N.B. Group division is done as shown in *Path Plot* section, with custom or equally-spaced intervals.

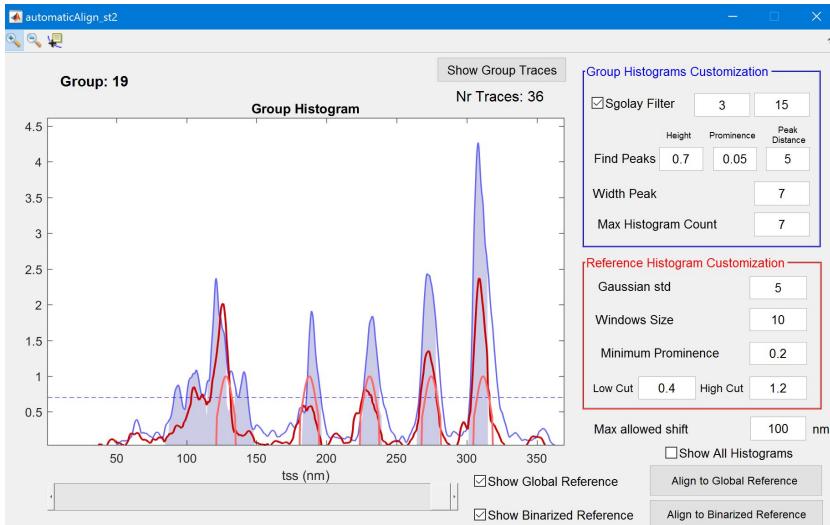
The user performing Automatic Alignment (menu bar **Align > Automatic Alignment**) in *Fodis* is guided across the aforementioned steps with the help of two Graphic Interfaces.



Graphic interface 1 for *global reference* construction

The first interface helps the user to set all relevant parameters for the formation of the *global reference*. The **Histogram Customization** box is crucial to determine the shape of the histograms to be aligned. Default parameters are generally adequate, but in case of some specific requirements, the user is allowed to change them. The **Alignment Valid range** edit boxes set the limits over which histograms are set to zero. **Gaussian std.** sets the standard deviation of the Gaussian "potential" and **Nr. of Iterations** sets how many iterations the algorithm executes: iterations are time consuming, but only one iteration generally gives unstable results. Use the button **Align!** to execute the algorithm, **Skip** if first alignment

has already been performed thus to go directly to the second Alignment interface.



Graphic interface 2 for *global reference* tuning

The scope of the second alignment step is to prepare the desired *global reference* (i.e. the reference histogram onto which every single histogram is finally aligned). Default parameters should be adequate, but for specific requirements, it is possible to change some settings.

The tuning process consists of:

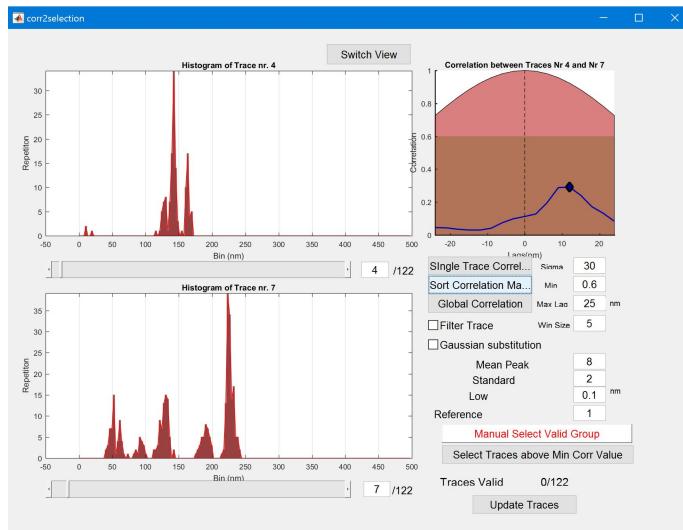
- tuning shape of the blue profiles (i.e. *group references*, resulting histograms from alignment *intra-group*). This can be done in the **Group Histograms Customization** box moving through different groups with the slide bar.
- **N.B.** Group division is done as shown in the *Path Plot* section, with custom or equally-spaced intervals;
- tuning shape of the red profile (i.e. *global reference*) and pink profile (i.e. *binarized reference*). This can be done in **Reference Histogram Customization** box.

The *global reference* is the weighted average of *group references*. The

binarized reference is based on the *global reference*. It contains peaks found in the *global reference*, but they are leveled and normalized. Align to Global-Custom Reference to finally re-align all the traces to the reference.

2.4 Selection through correlation

A common problem in SMFS is finding and isolating *relevant* traces among all collected data. The *Basic Filtering Tool* described in Chapter 1 allows the operator to filter out traces longer or shorter than the expected ones; the *Fingerprint ROI* tool, described in this Chapter, looks at a specific pattern drawn by the user and can list the traces accordingly. All these tools need the intervention of the user to choose some specific interval or shape, in contrast with the purpose of an unsupervised filtering procedure.

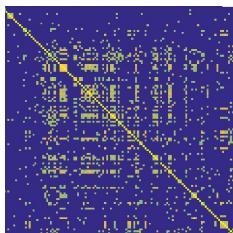


The *Selection through correlation* (**Selection > ***) calculates the cross correlation between every couple of loaded (Valid) traces forming a symmetric *similarity matrix*. This matrix can then be sorted (ordered) with the MATLAB function `symamd` (symmetric approximate minimum permu-

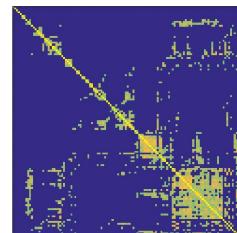
tation algorithm).

The slide bars change the visualized contour length histograms. The blue curve in the top-right panel is the calculated cross correlation of the two visualized histograms of *trace 1* and *trace 2*. The diamond points the maximum value, the red area is the Gaussian curve that weights the correlation (as described in *Automatic Alignment*).

- **Single trace correlation** button displays the values of correlation relative to *trace 1* (the one shown in the upper panel);
- **Sort Correlation Matrix** applies the symmetric approximate minimum permutation algorithm to the *similarity matrix* (i.e. it changes the order of the traces to find clusters of similar traces);
- **Global Correlation** shows the *similarity matrix*;
- **Sigma** determines the width of the Gaussian curve that weights the correlation curve;
- **Min** is the threshold value for the *similarity matrix*;
- **Max Lag** is the maximum lag allowed when calculating the correlation between two traces.



Similarity matrix



Sorted similarity matrix

At this point the user has two options: he can decide to select a specific cluster in the sorted *similarity matrix*, or he can select traces similar to the reference *trace 1* (the trace visualized in the upper panel).

Manual Select Valid Group

Select Traces above Min Corr Value

With the first button, the sorted *similarity matrix* pops up and, with the cursor, the user can select a cluster of traces (click-drag&drop). The selected group of traces will be visible in the main interface only after clicking **Update traces**. If the selection does not satisfy the user, he can **Filter > Move ALL to Valid** and redo all the process.

The second button allows to select the traces above the **Min** threshold, that are similar to the reference trace in the upper panel. The selected traces are previously visible clicking **Single trace correlation** button.

A further refinement of the calculation of the correlation is possible tuning the contour length histograms with the options:

- **Filter trace** applies a smoothing filter to the contour length histogram of every trace;
- **Gaussian substitution** substitutes a sum of Gaussian curves to each contour length histograms. It detects the peaks with the parameters on the right.

Bibliography

- [1] Puchner, Elias M., et al. "Comparing proteins by their unfolding pattern." *Biophysical journal* 95.1 (2008): 426-434.
- [2] Bosshart, Patrick D., Patrick LTM Frederix, and Andreas Engel. "Reference-free alignment and sorting of single-molecule force spectroscopy data." *Biophysical journal* 102.9 (2012): 2202-2211.
- [3] Kawamura, Shiho, et al. "Kinetic, energetic, and mechanical differences between dark-state rhodopsin and opsin." *Structure* 21.3 (2013): 426-437.