Coordination Finder 1.5

Basel, 16.6.2016

Contents

Thomas Gumbsch
Departement of Biosystems Science and Engineering
ETH Zürich

1. Introduction

Coordination Finder is a program developed by Thomas Gumbsch in the Borgwardt group in the Department of Biosystems Science and Engineering at ETH Zürich, and was developed in collaboration with the Schroeder group, motivated by their recent work [5]. Coordination Finder takes protein expression levels of cell lineages as input and provides methods to extract information out of the data.

Coordination Finder will help doing the following:

- Cluster time series data. It may consist of multiple, independent dimensions
- Visualize history information from mouse embryonic stem cell lineages
- Compute of changepoints in the time series
- Provide expert feedback to measure the quality of the algorithm

2. Getting started

Coordination Finder runs on both Windows and Mac OSX. It can be run from the console. The requirements are having the newest version of Python installed as well as the PyQT4 package. By default, there have one test dataset (described in Section 2.2 below) ready to explore.

2.1 Installation

Unzip the folder and navigate into it. Then run the following commands from the console:

Mac:

```
python -m pip install PyQt4
python CoordinationFinder.py
```

Windows:

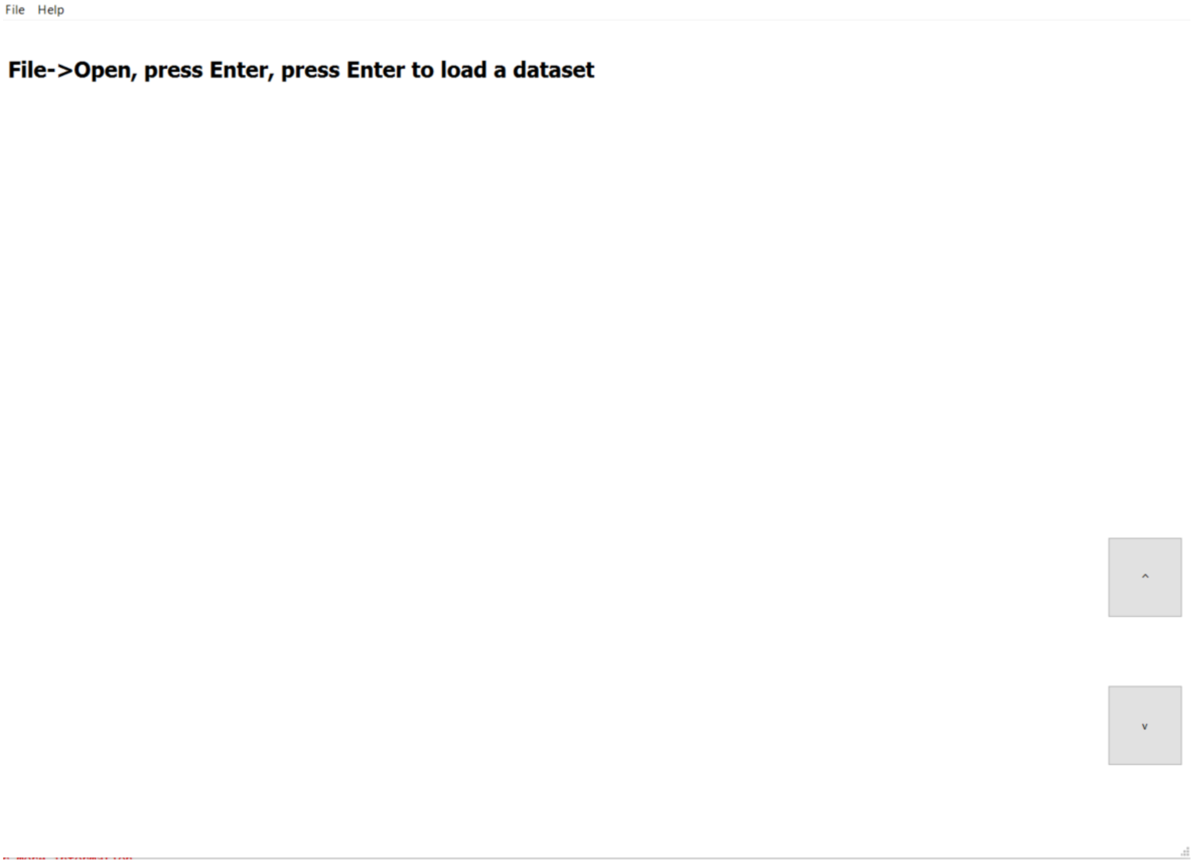
```
pip install PyQt4
start CoordinationFinder.py
```

2.2 Load *Simulation.csv*

The test dataset *Simulation.csv* consists out of two cell lineages (“trees”) with the simulated expression profiles of two Proteins: *ProteinA* and *ProteinB*. The first tree, *1nocp,* consists out of time series with no changepoint. In *1cp*, every curve has a changepoint of fixed strength at a random position.

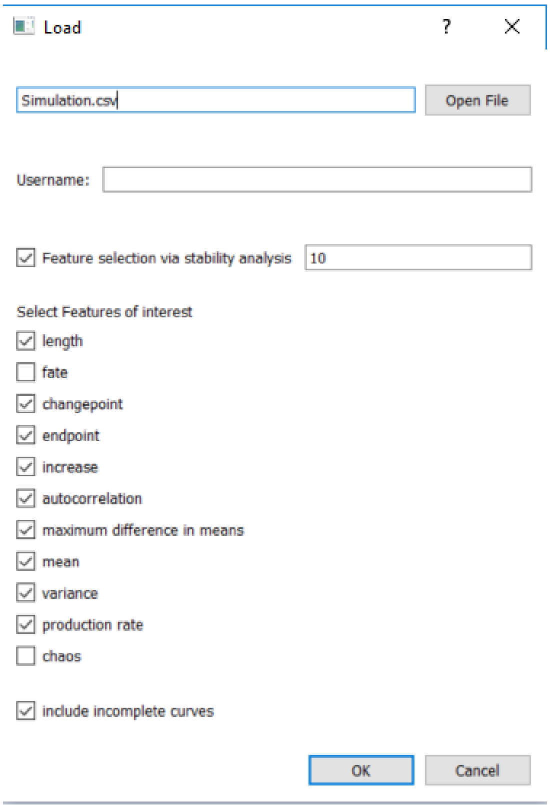
To load *Simulation.csv* it takes these five steps:

1. Follow 2.1 to start Coordination Finder. The main window will get displayed full screen and it will end up looking the following way:



File->Open, press Enter, press Enter to load a dataset

2. Go to **`File->Open`** to show the Loading screen. A recommended sample selection of what to compute will appear. For later uses, bear in mind that it can played with this configuration. For now, press okay:



Simulation.csv

Username:

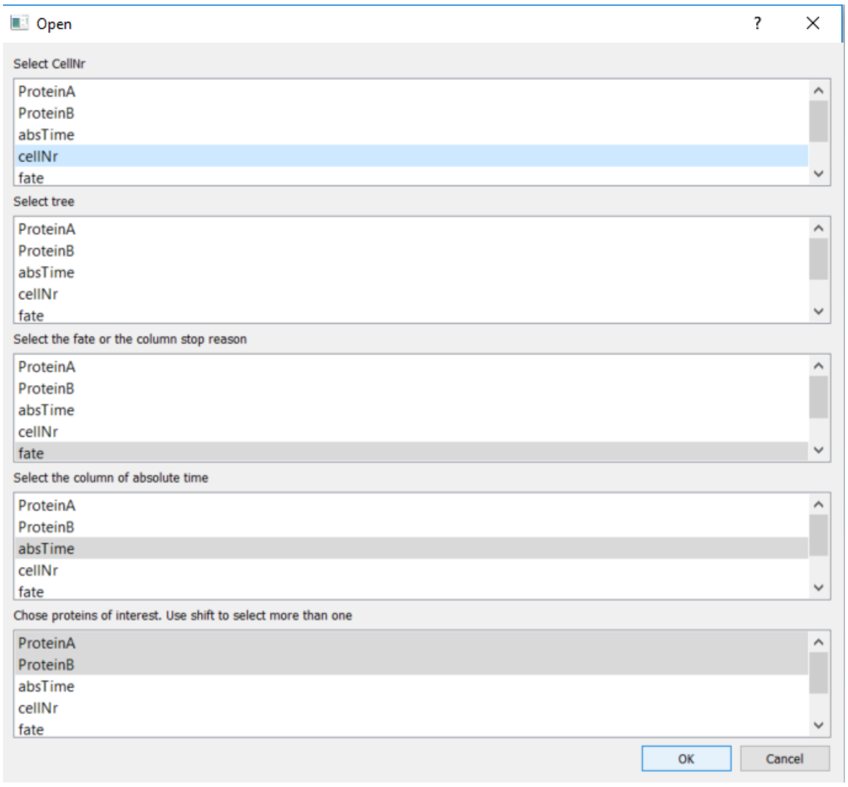
☑ Feature selection via stability analysis 10

Select Features of interest

☑ length
☐ fate
☑ changepoint
☑ endpoint
☑ increase
☑ autocorrelation
☑ maximum difference in means
☑ mean
☑ variance
☑ production rate
☐ chaos

☑ include incomplete curves

3. In the Opening-dialog, there is the option to select which columns to be evaluated. The columns for the test dataset have been preselected. For the use of different dataset, the selection has to be adjusted. Please press **OK** to start the computation:



Open

Select CellNr

ProteinA
ProteinB
absTime
cellNr
fate

Select tree

ProteinA
ProteinB
absTime
cellNr
fate

Select the fate or the column stop reason

ProteinA
ProteinB
absTime
cellNr
fate

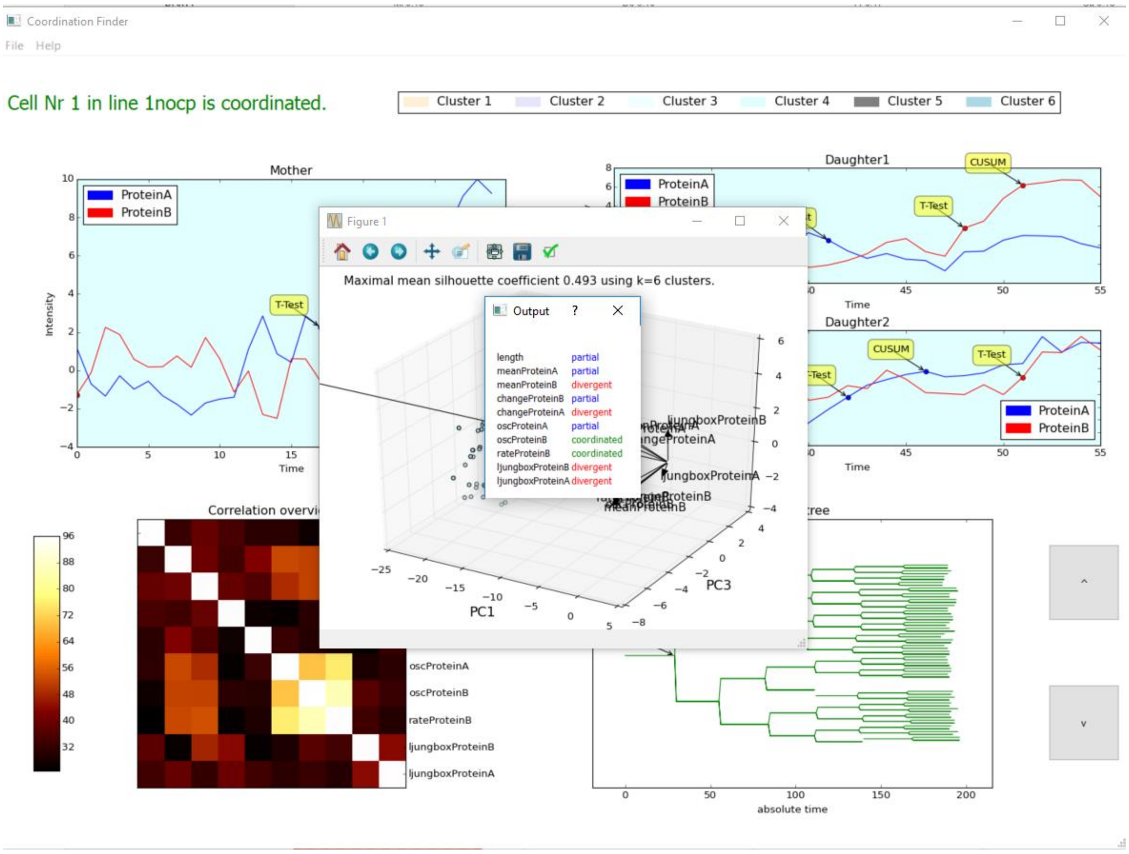
Select the column of absolute time

ProteinA
ProteinB
absTime
cellNr
fate

Chose proteins of interest. Use shift to select more than one

ProteinA
ProteinB
absTime
cellNr
fate

4. *Simulation.csv* will be evaluated. Please be patient, this may take up to 5 min. Go grab a coffee. After coming back, the screen should look like this:



Cell Nr 1 in line 1nocp is coordinated.

Maximal mean silhouette coefficient 0.493 using k=6 clusters.

```
length            partial
meanProteinA      partial
meanProteinB      divergent
changeProteinB    partial
changeProteinA    divergent
oscProteinA       partial
oscProteinB       coordinated
rateProteinB      coordinated
ljungboxProteinB  divergent
ljungboxProteinA  divergent
```

5. Well done! Enjoy exploring *Simulation.csv* and the Features of Coordination Finder. Play with the Loading options and click on buttons, graphs and menu options to get a feeling for Coordination Finder.

3. Navigating and understanding the output of Coordination Finder

The main window of Coordination displays all the results: Clustering, Changepoint analysis and Coordination. There are the following menu options:

- **`File->Open`** opens the loading dialog. The computational details as well as the username can be entered here.
- **`File->Vote`** opens a screen where to input trees to be coordinated/partially coordinated/uncoordinated and lets the user vote on where to see a changepoint.
- **`File->Analysis`** performs a meta-analysis on all collected votes from all previous and current users.
- **`File->Export`** creates a folder with the username and stores all relevant .csv files and plots.

When having the main window on display, there are two windows with additional information about the clustering and the coordination.

3.1 The username

It is recommended to enter a username after navigating to **`File->Open`**. When leaving this blank, the user will execute every command under the name of `“test_user”.` The username has an effect on these processes:

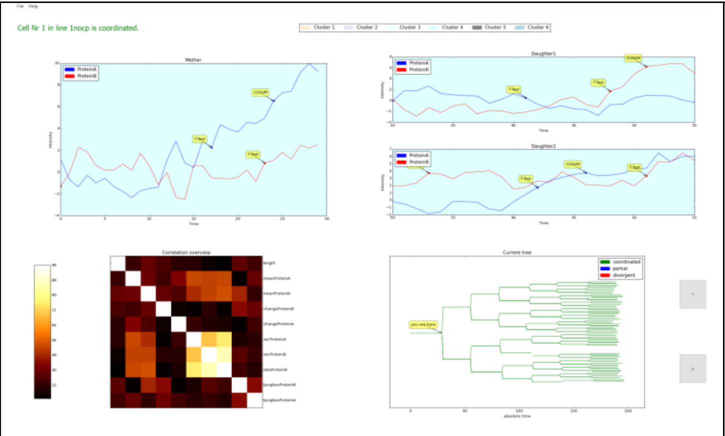
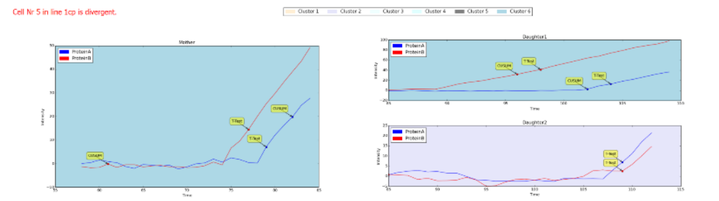
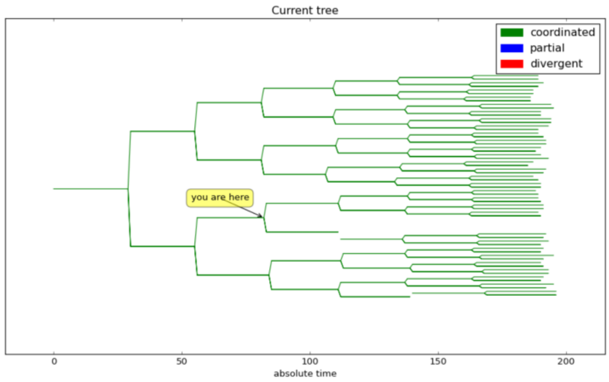
- Exporting the data.
- Saving and loading votes after closing and restarting Coordination Finder. (See section 5)

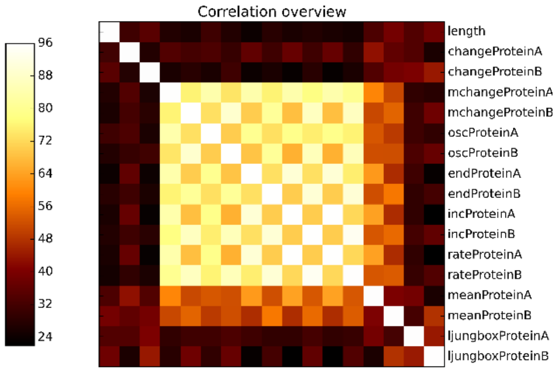
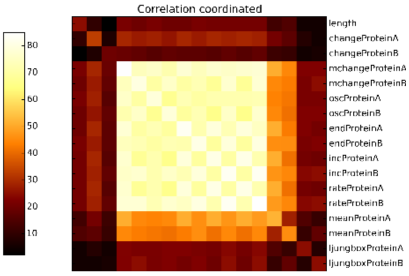
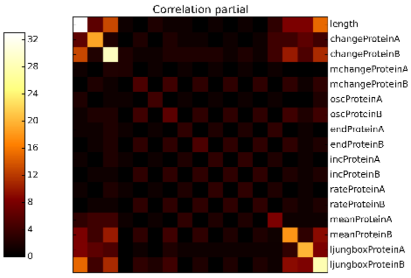
3.2 Clustering and coordination

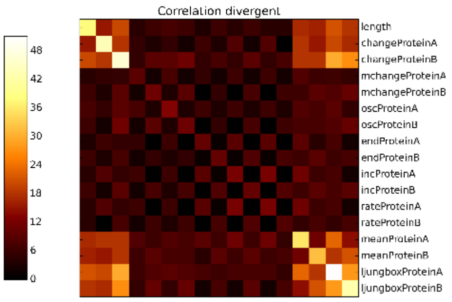
The **clustering** algorithm is K-Means. The number of clusters is determined by the respective maximal silhouette coefficient. Feature selection is performed, if selected, via a stability analysis where the feature space is perturbed by some random noise of size epsilon.

Coordination describes the relation between three data points. When all three cells belong to the same half of the population, we call them coordinated. If parent and child generation behave dissimilar, they are partially coordinated. If the sister cells belong to different halves, they are called divergent.

3.3 Explanation of the main window

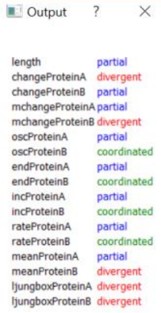
 Cell Nr 1 in line 1nocp is coordinated. File Help Cluster 1 Cluster 2 Cluster 3 Cluster 4 Cluster 5 Cluster 6 Mother Daughter1 Daughter2 ProteinA ProteinB CUSUM T-Test Intensity Time Correlation overview Current tree coordinated partial divergent you are here absolute time	This is the **main window**. The top half is a panel where to look at a single mother cell and her two daughter cells. The bottom half visualized the results from more than one subtree. Click on one of the cells, mother, daughter1 or daughter2, to move within the tree. Press ^ or v to move to the next or previous tree.
 Cell Nr 5 in line 1cp is divergent. Cluster 1 Cluster 2 Cluster 3 Cluster 4 Cluster 5 Cluster 6 Mother Daughter1 Daughter2 ProteinA ProteinB CUSUM T-Test Intensity Time	The upper half of the window consists of the **three plots of the subtree** (mother, daughter1 and daughter2). On the *x*-axis, the absolute time is on display. The *y*-axis gives the protein intensity over time for the selected proteins. If can be seen to which cluster the algorithm assigned the curves by looking at the background color of the plot. The results for the change detection (if selected) get also displayed in the graphs. Above the plot, the cell number and the tree can be read out. The color corresponds to the type of coordination.
	The bottom right plot shows the landscape of the **whole tree**. At certain time points, a line splits which means that the cell divided and two daughter cells appear. The intersections where the three viewed cells meet is labeled as *you are here* in the tree. Each subtree is colored according to the coordination of the three involved cells. Remember that coordination means that the three curves behave similarly. Partial coordination requires the two daughter cells behave similarly and the mother cell differently. Diverging subtrees are those whose two daughter cells behave differently. Moving within the displayed tree (further down one branch or up to the parent) can be done by clicking on the respective plot (see above), i.e. to

	make one of the daughters the displayed mother cell, click on the daughters. To move to the top of the tree, click on the mother cell.
˄ ˅	On the right side of the bottom half of the main window, there are two buttons for **navigation through the data**. Clicking on ˄ or ˅ makes the user jump from any part of the current tree to the uppermost complete smallest tree in the next/previous tree. Moving within the current subtree, can be accomplished by clicking on the respective plot in the top half of the main window.
	The bottom left plot **is a heat plot of the correlation between the coordination** according to different features. The numbers associated with it are the numbers of subtrees where the coordination agrees when looking pairwise at two features. By clicking on it, one can cycle through the types of coordination. For instance, entry (1,2) from “Correlation overview” displays the number of times changepoint of *ProteinA* and length output the same type of coordination.
	We look at the correlation of features filtered with respect to coordination, when clicking once on the plot “Correlation overview”, i.e. (2,2) in is the number of times, subtrees are classified as coordinated when only looking at the changepoint detection on *ProteinA*.
	We look at the correlation of features filtered with respect to partial coordination, when clicking twice on the plot “Correlation overview”, e.g. (3,2) is the number of times, subtrees are classified as partial according to the changepoint of *ProteinA* **AND** *ProteinB.*



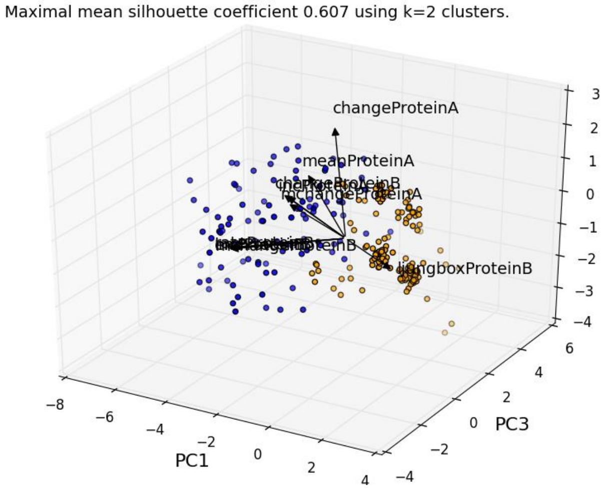
We look at the correlation of features filtered with respect to divergence, when clicking three times on the plot “Correlation overview”, i.e. (6,8) is the number of times, subtrees are classified as diverging according to the oscillation of *ProteinA* while being divergent with respect to the endpoint of *ProteinB* at the same time.

3.4 Windows for additional information



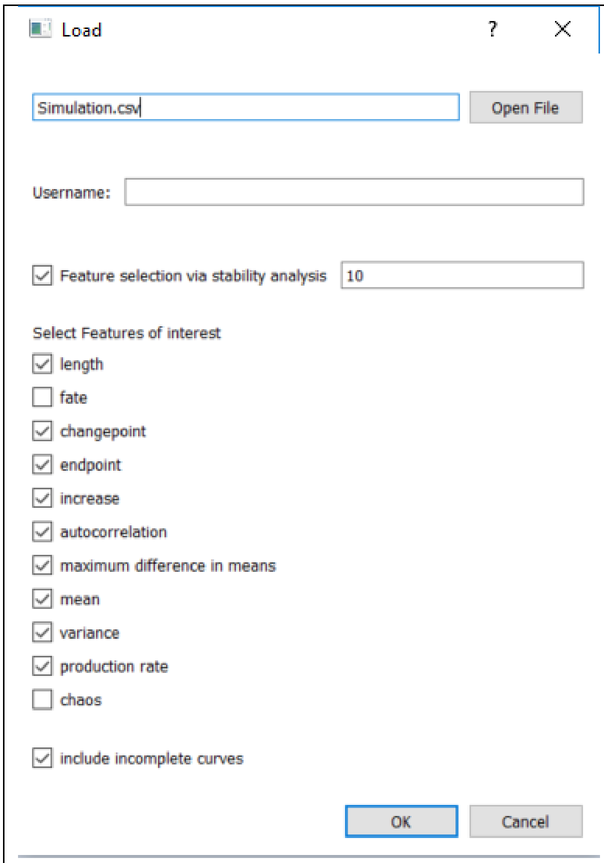
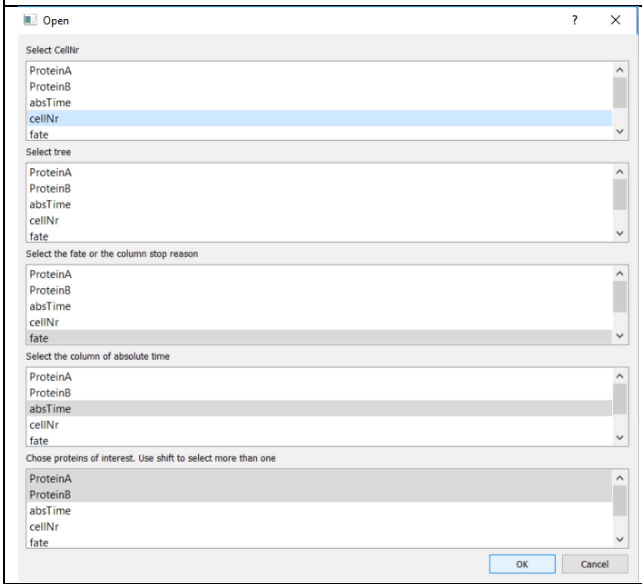
```
length          partial
changeProteinA  divergent
changeProteinB  partial
mchangeProteinA partial
mchangeProteinB divergent
oscProteinA     partial
oscProteinB     coordinated
endProteinA     partial
endProteinB     coordinated
incProteinA     partial
incProteinB     coordinated
rateProteinA    partial
rateProteinB    coordinated
meanProteinA    partial
meanProteinB    divergent
ljungboxProteinA divergent
ljungboxProteinB divergent
```

An extra window opens every time the subtree changes. It gives an overview about the coordination according to every single feature in this particular subtree. For example, on the figure left length is partially coordinated. The summary of this output is visualized in the heat plots. The raw data can be accessed after exporting it to .csv.

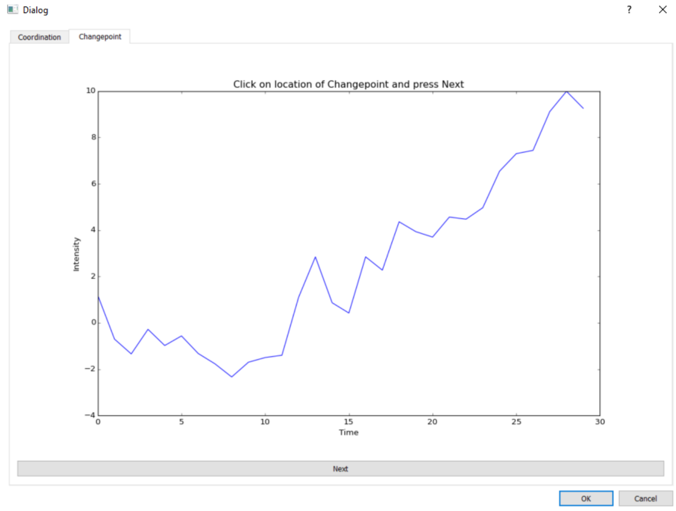


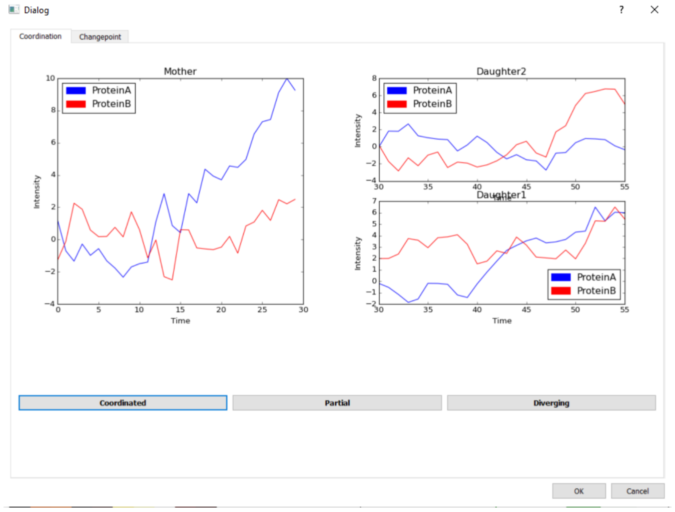
After the computation is complete, an interactive 3D plot of the clustering is displayed. Click the left mouse button and move the mouse to rotate the figure. Remember that each dot corresponds to one cell life. The three axes are the three principal component axes of the feature space. The features are then projected into this space and displayed as arrows. The bigger the arrow, the higher the influence of the specific feature. Even though the number of clusters with the minimal silhouette coefficient is selected, for real data *k=2*, mostly.

3.5 Loading datasets

 Load Simulation.csv Open File Username: ☑ Feature selection via stability analysis 10 Select Features of interest ☑ length □ fate ☑ changepoint ☑ endpoint ☑ increase ☑ autocorrelation ☑ maximum difference in means ☑ mean ☑ variance ☑ production rate □ chaos ☑ include incomplete curves OK Cancel	Use the **Load-Window** in order to select a dataset and the specifics on the data manipulation. For the input file, any data file with a .csv extension is allowed. Note that the important feature of .csv files is that column names are provided for each data column in the .csv file. It is recommended to enter a **username**. Especially for storing the progress on the voting and resume after restarting. One can then specify the features according to which to cluster the data. In general, there are two ways to select the features. One can either select every feature by hand. Or one can use a specific amount of the most stable features in terms of random perturbation of the data. Please notice that checking the box *include incomplete curves* discards the first and the last layer of cells in the trees.
 Open Select CellNr ProteinA ProteinB absTime cellNr fate Select tree ProteinA ProteinB absTime cellNr fate Select the fate or the column stop reason ProteinA ProteinB absTime cellNr fate Select the column of absolute time ProteinA ProteinB absTime cellNr fate Chose proteins of interest. Use shift to select more than one ProteinA ProteinB absTime cellNr fate OK Cancel	Continuing with **OK** in the Load-Window, a dialog opens where one selects the columns of dataset Coordination Finder accesses. Please select the right columns for the computation. In the last selection process more than one Column of interest can be named. Please hold shift when selecting several proteins. After clicking on **OK**, the dataset gets evaluated. Every feature will be computed. Depending on the type of machine, this might take time up to a couple of minutes.

3.6 Voting



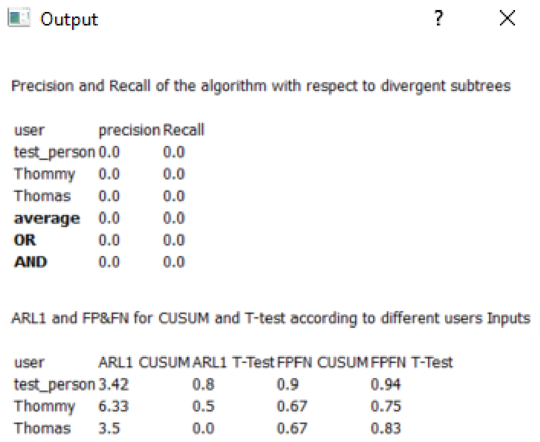


The **vote-screen** can be used for providing information about the accuracy of the algorithm.

Starting from the curve currently on display, the user can give feedback for **changepoint**s on every curve.Click on the presumed location to make a vertical line appear. Remove it by clicking again. When satisfied, press next. Then, other people’s votes as well as the results from the change detection algorithms get displayed. The screen freezes for 5s to give time to look at the results.

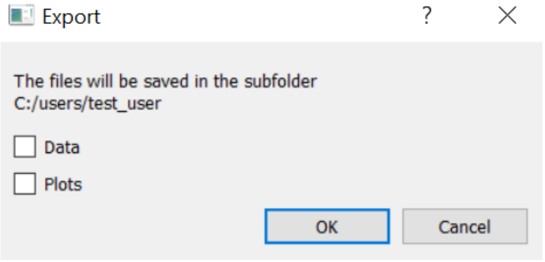
To vote on **coordination**, switch to the coordination subwindow. Starting from the tree currently on display, vote on each tree moving linearly down the list of minimal allowed trees. Remember that coordination means that the three curves behave similarly. Partial coordination requires the two daughter cells behave similarly and the mother cell differently. Diverging subtrees are those whose two daughter cells behave differently. After submitting, how other users rated the curves and what the clustering according to K-Means outputs is displayed. To save progress, press **OK**. Exiting any other way will discard the scores. Saving the results will enable to continue the voting at the stopping location of last time, even after restating and clustering newly.

Voting is only possible on the curves which can be also accessed through the main window. Moreover, the votes will be stored in separate for a different selection of Transcription factors.



```
Output                                              ?    ×

Precision and Recall of the algorithm with respect to divergent subtrees

user        precision Recall
test_person 0.0       0.0
Thommy      0.0       0.0
Thomas      0.0       0.0
average     0.0       0.0
OR          0.0       0.0
AND         0.0       0.0

ARL1 and FP&FN for CUSUM and T-test according to different users Inputs

user        ARL1 CUSUM ARL1 T-Test FPFN CUSUM FPFN T-Test
test_person 3.42       0.8         0.9        0.94
Thommy      6.33       0.5         0.67       0.75
Thomas      3.5        0.0         0.67       0.83
```

The **analysis** dialog outputs a meta-analysis of the voting of the users. In terms of **Coordination**, precision and recall with respect to the uncoordinated trees is displayed for every user. The average column averages over all collected votes (not users!). OR takes the set of all trees where every user voted at least once for uncoordinated as the true positives. AND choses the true positive trees to be the ones where every user voted for no coordination. ***CUSUM* and *T-test*** get evaluated according to statistical tests in the table below. ARL1 measures the average run length after a true changepoint before the algorithm grasps it. FPFN is the rate of false positives and false negatives combined.



Export ? ×

The files will be saved in the subfolder
C:/users/test_user

☐ Data

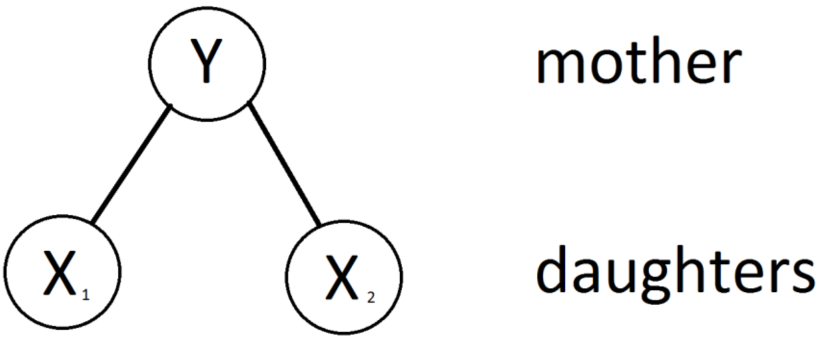
☐ Plots

The **export-dialog** enables storing the results in two different ways:

- Data: Votes, Features, Coordination output and original Data will get exported as .csv files.
- Plots: Histograms for Features, a Boxplots will be created and saved as well as all the already generated. They are saved as *.png*.

4. Methods

In Coordination Finder, one can extract information about **minimal subtrees**. These consist out of one mothercell and the daughter cell she divided into. An allowed subtree is a tree where the curves of the mothercell starts at her birth and the curves of the daughter cells end at their death or division. Excluding incomplete curves means that we discard (a) the cells which were already alive when our monitoring started, and (b) the cells which were still alive once we stopped monitoring.



1: Outline of a minimal subtree

This means, for real data, we exclude the “surface” of our trees, because we would otherwise end up with a bias towards shorter sequences.

4.1 Coordination

The main output of information happens through labeling smallest complete subtrees according to their coordination. The idea behind coordination is to spot diverging sister cells. The three types of coordination are defined the following way:

- **Coordination** is taking place when all three cells belong to the same half of the population according to a chosen feature. (Y AND X1 AND X2)
- One mother cell and her corresponding daughter cells express **partial coordination** if both daughter cells behave similarly, but different compared to the mother cell. Reference is, again, the median of the whole population. ((X1 AND X2) NAND Y)
- **Diverging** subtrees are those where sister cells that do not belong to the same half of the population according to a given feature. (X1 NAND X2)

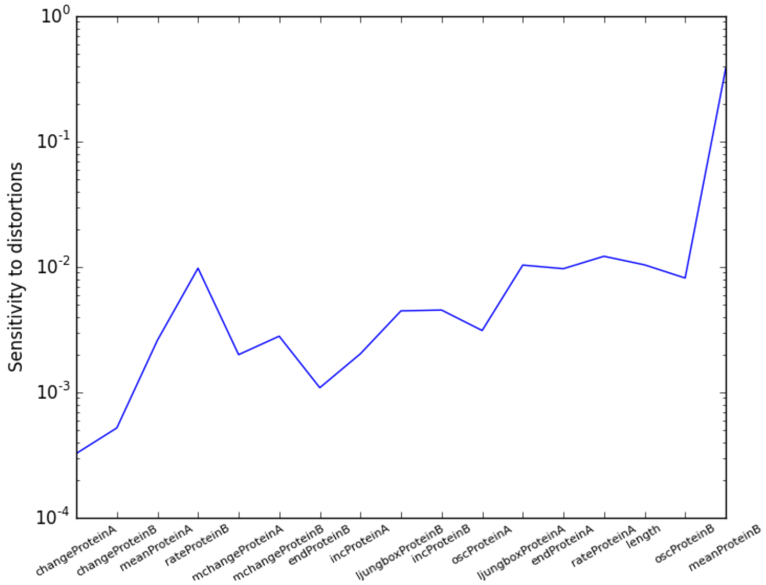
Coordination Finder uses colors for making it visually more accessible:

- **Green for coordination**
- **Blue for partial coordination**
- **Red for divergence**

Instead of taking features, we can also view coordination in terms of the Clustering output. We then do not use the median but treat cells assigned to different clusters like belonging to different parts of the population.

4.2 Clustering

The general approach is to use characteristic based clustering on global features similarly as described in [1]. We cluster with **K-Means**[4] according to the selected features. The number of clusters is determined by the corresponding maximal silhouette coefficient. Remember that if all three curves of a single tree belong to the same cluster, we call this tree coordinated. In a partially coordinated tree, the daughter cells behave differently compared to the mother cells. Uncoordinated trees have two daughters which were assigned to different clusters.



2: Stability analysis on the test data. We distort the feature space by a little bit and perform a greedy forward search on the features with the least centriod deviation after clustering.

4.3 Feature search

The **feature selection** works according to a greedy forward search algorithm proposed by [1]. First, the data is whitened. Then we distort the whitened feature space by a little epsilon *|X|*-times, each time differently. In our case, setting *X=10* is a good rule of thumb.
Next, we cluster all data sets *X* according to each feature once, i.e. if the set of all features is *Y*, we perform

|X||Y|- times K-Means. Afterwards, the centroid deviation on *X* is measured using the L2-norm. We look at each feature separately and take the distances between the centroids of the *|X|* different clustering pairwise squared and normalize to the number to clusters. As a result, we get *|Y|* numbers. We call the set of numbers *Y_1*. As the most stable first feature, we select the one with the smallest *Y_1i*, so the corresponding feature is *Y_i*.
In the next step, *Y_i* is removed from *Y*. The clustering then happens on X with two features: *Y_i* and one Feature from Y. As the second most stable feature, we take *Y_j* which has the smallest *Y_1j* after clustering according to *Y_i* and *Y_j*.
In this manner, we can sort the features according to their stability in terms of clustering after distorting of the data.

4.4 Change detection

Having a changepoint is a possible feature to select. The algorithm take tools from statistical process control ([2], [3]) and uses them to label the figures binary: Either they have a changepoint or not. We compare the performances of ***CUSUM*** and the ***students two-sample T-test*** on test sequences. We model two different time series:
In the first approach, we distribute values normally around a mean value. *ARL0* measures the average run length until a false change is detected. *ARL1* measures the average distance to a change in the mean value.
More relevant for us is the biological equivalent to this model: We model a sequence of normally distributed numbers around zero and take the cumulative sum. *FPR* measures the false positive rate for sequences whose lengths are drawn from the distribution of cell life lengths. *CWARL1* is the cumulative weighted *ARL1*: The mean of the sequence (cumulative sum) changes once by a random amount. We weight the distances between exact and measured changepoint: If the algorithm misses small changes it does not count as much as if it misses large changes. These are the results:

```
+--------+----------------+-----------------+
|  Test  |     CUSUM      |      T-Test     |
+--------+----------------+-----------------+
|  ARL0  | 94.7525773196  |       NaN       |
|  ARL1  | 3.56361584782  |  0.743204547887 |
|  FPR   |     0.185      |      0.438      |
| CWARL1 | 0.131953119035 | 0.0555763867025 |
+--------+----------------+-----------------+
Number of Runs:  1000
```

This analysis leads to the following procedure: If *CUSUM* and *T-Test* are detecting a **changepoint**, the curve is labeled to have one. If none or only one of them signal a change, the curve is taken as having no changepoint.

The model used for *CWARL1* and *FPR* is the same as for **generating the sample dataset** *Simulation.csv*. *1cp* consist only out of curves which are tested on in *CWARL1*. *1nocp* includes only curves from the simulation of *FPR*.

4.5 Other features

For the clustering with *K-Means*, Coordination Finder has a list of features for to select:

length: This feature takes the length of the cell life into account.

fate: Either division, death or lost. For the future it is planned to use this as a label for supervised learning and not as a feature.

endpoint: It refers to the absolute intensity of the protein expression at the end of cell life.

mean: This number corresponds to the average intensity of the protein throughout the cell life.

increase: Difference between start and end of the protein intensity.

rate: Increase normalized to length of Series. This value corresponds to the average production rate of the transcription factor

variance: Measured is the variance of the time series. With this feature we get an idea about the amount of oscillations happening.

Maximum difference in means: We look for the maximum difference of averages: At every step, the time series is divided into two subintervals. In each, we calculate the mean. The feature is then the maximum of the difference of means throughout one cell life. It seems to be the case that this value is roughly quadratically related to the variance.

autocorrelation: we look at autocorrelation as suggested in [1]. The Ljung-Box test for is performed. It stores the p-value based on the chi-square distribution. The Ljung-Box test is reported to have better small sample properties compared to Box-Pierce. Informally, it is the similarity between observations as a function of the time lag between them.

chaos: This feature is the Lyapunov exponent, calculated as proposed in [1]. It is a measure of how small differences evolve over time. In other words, how chaotic the sequence behaves.

5. Voting, exporting and outlook

At **`File->Vote`**, expert input of changepoint and coordination to determine the accuracy of the algorithm behind Coordination Finder can be given. The ultimate goal would be to include the votes of the users as a label in terms of supervised clustering. But that is a feature yet left to implement.

In the **voting-file**, every complete subtree is assigned to one row. A subtree is complete if the daughter cells and the mother cells curves start at birth and end with death or division. Since the subtrees which are evaluated depend on the number of proteins selected and the fact whether to include incomplete curves, the filename changes according to this selection. In terms of requirements, a minimally unique identification of the used subtrees consist out of:

```
“votes” + str(name_of_input_data) + str(Protein_column_names) +
str(bool_incomplete) + “.csv”
```

Thus it is possible for a user accessing the same data with similar input configurations as someone who came before him or her to access the votes of others. For the future it is planned to store it online and provide access for multiple user over the network.

In the file itself, for every new user, *1+3N* columns are created: One for the coordination vote on the whole subtree. *N* stands for the number of proteins taken into account. For each protein we have three curves per subtree which we could vote on. Each column name will start with the **username** of the user who created the votes. It is recommended to enter a username not only because one can reload its votes, but also connect the stored votes with the personal background of the expert.

When voting, it begins with the current subtree and cycles through the whole .csv file until it ends where it started. Please bear in mind that one needs to exit with ok to make Coordination Finder save the results permanently. If one exits in a different way, the votes will be lost.

6. References

[1]: Wang X., Smith K., Hyndman R., Data Mining and Knowledge Discovery 2006, Characteristic-Based Clustering for Time Series Data

[2]: Hawkins D.M., Qiu P., Kang C.W., Journal of quality technology 2003, The Changepoint Model for Statistical Process Control

[3]: J. Takeuchi, K. Yamanishi, IEEE Transactions on Knowledge and Data Engeneering 2006, A Unifying Framework for Detecting Outliers and Change Points from Time Series

[4]: P.J. Rousseeuw, Comput. Appl. Math. 1987, Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis.

[5]: S. Hastreiter, S. Skylaki, T. Schroeder, Nat. Cell Bio. 2015, Network plasticity of pluriotency transcription factors in embrionic stem cells.