Carleton University

TIME-COURSE GENE EXPRESSION DATA A TIME SERIES MOTIFS ANALYSIS FOR GIARDIA ENCYSTATION DATA

ERALDA GJIKA

Keyword:

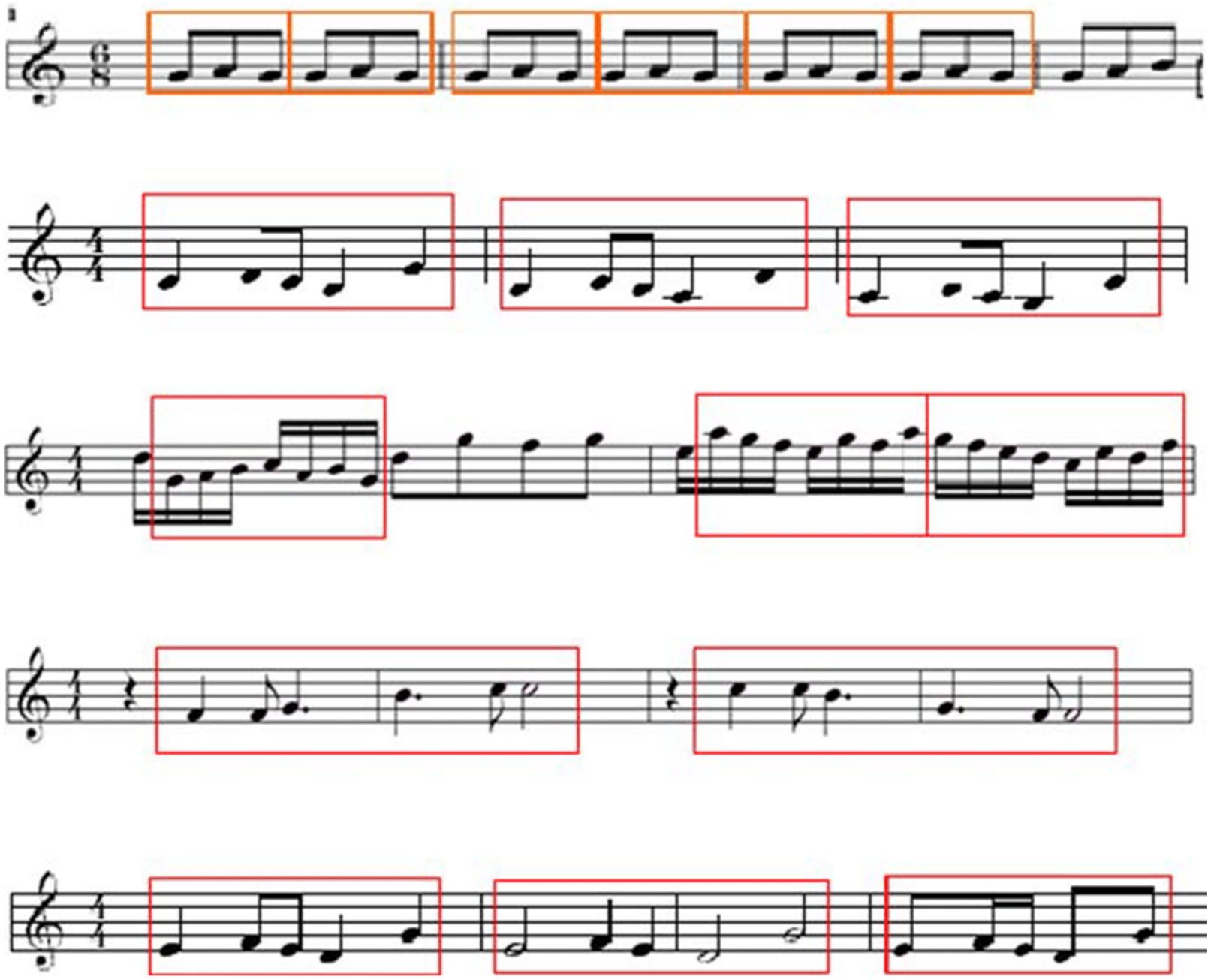
MOTIFS

Where?

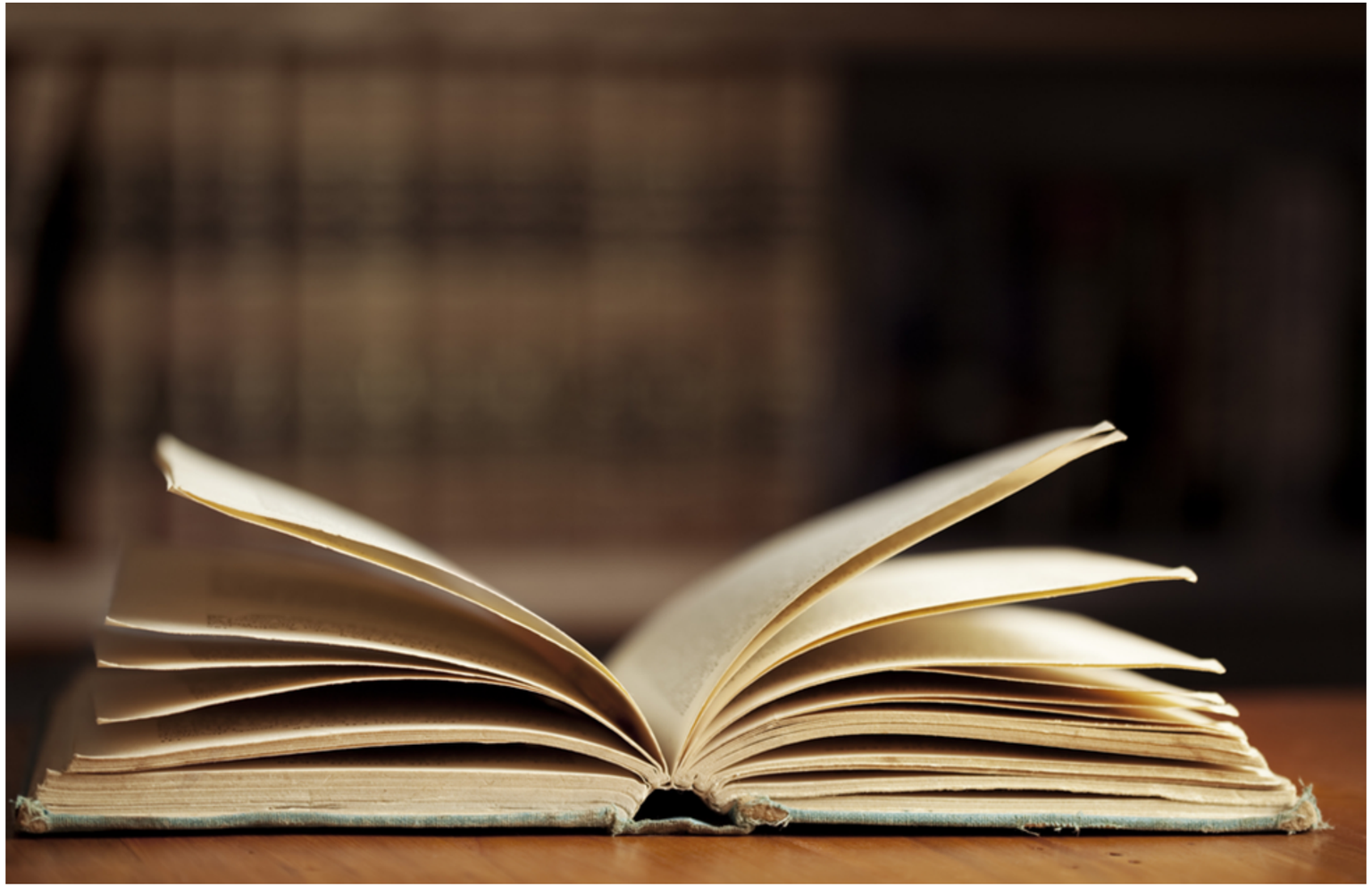
What?

Why?

MUSICAL PATTERNS



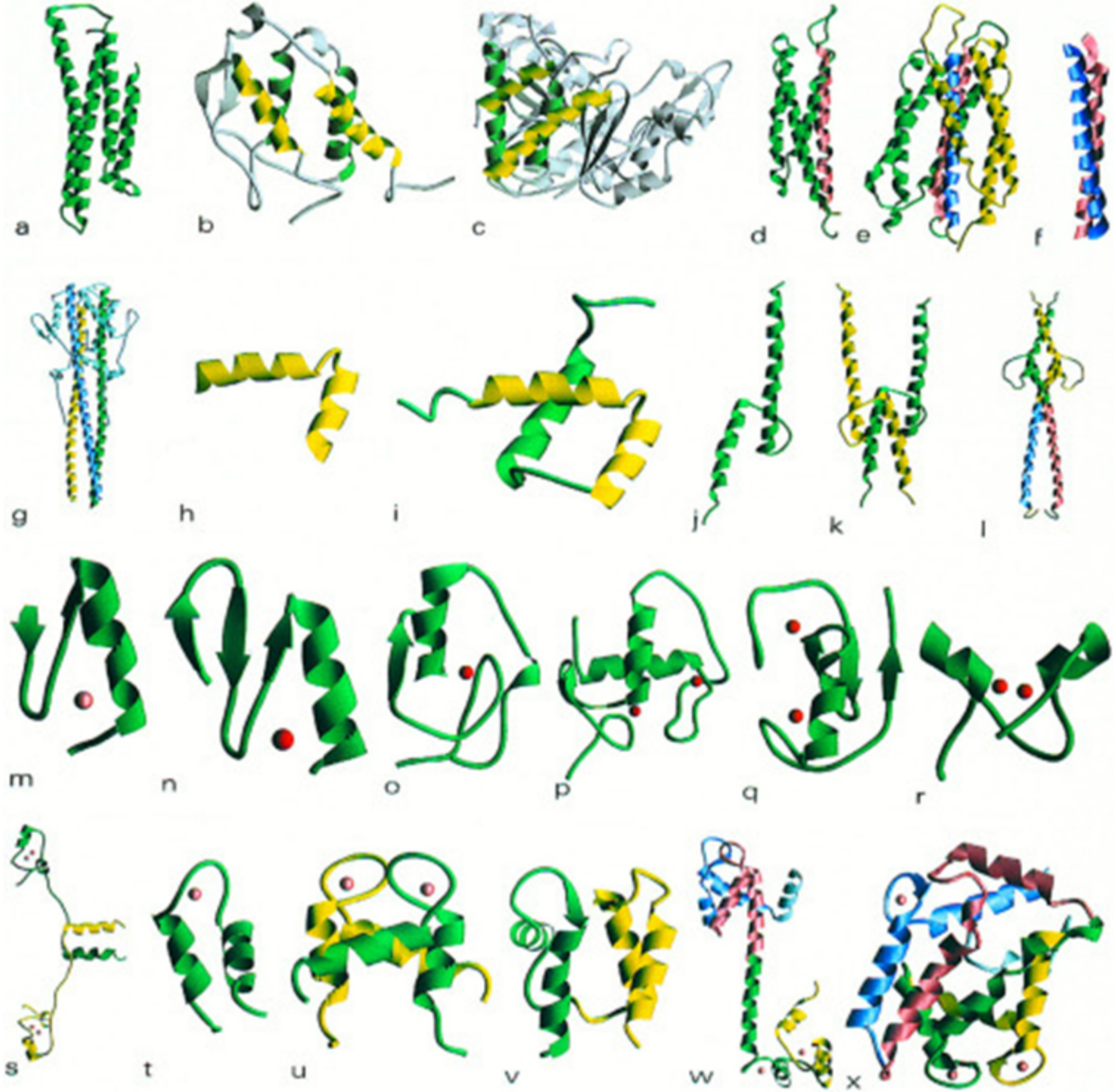
REFERENCE: HTTPS://LINK.SPRINGER.COM/ARTICLE/10.1007/S11042-009-0303-Y



Literature

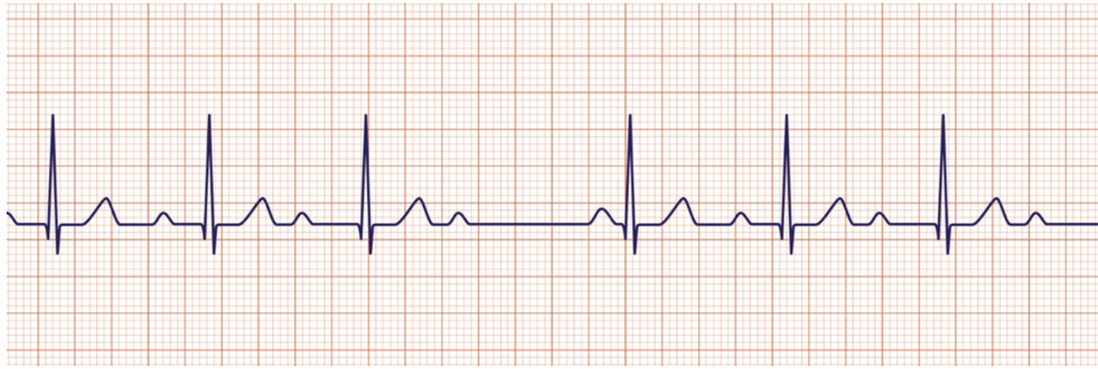
A MOTIF IS A RECURRING ELEMENT, IDEA OR CONCEPT THAT HAS A SYMBOLIC VALUE IN A TEXT.

ANALYZING MOTIFS IN A LITERARY WORK LEADS TO A DEEPER UNDERSTANDING OF A WORK.



a-Helical Protein Assembly Motif

Electrocardiogram



REFERENCE: HTTPS://WWW.JBC.ORG/ARTICLE/S0021-9258%2819%2967344-8/FULLTEXT

Time series Motifs

We may use the repeated patterns to predict the behavior!

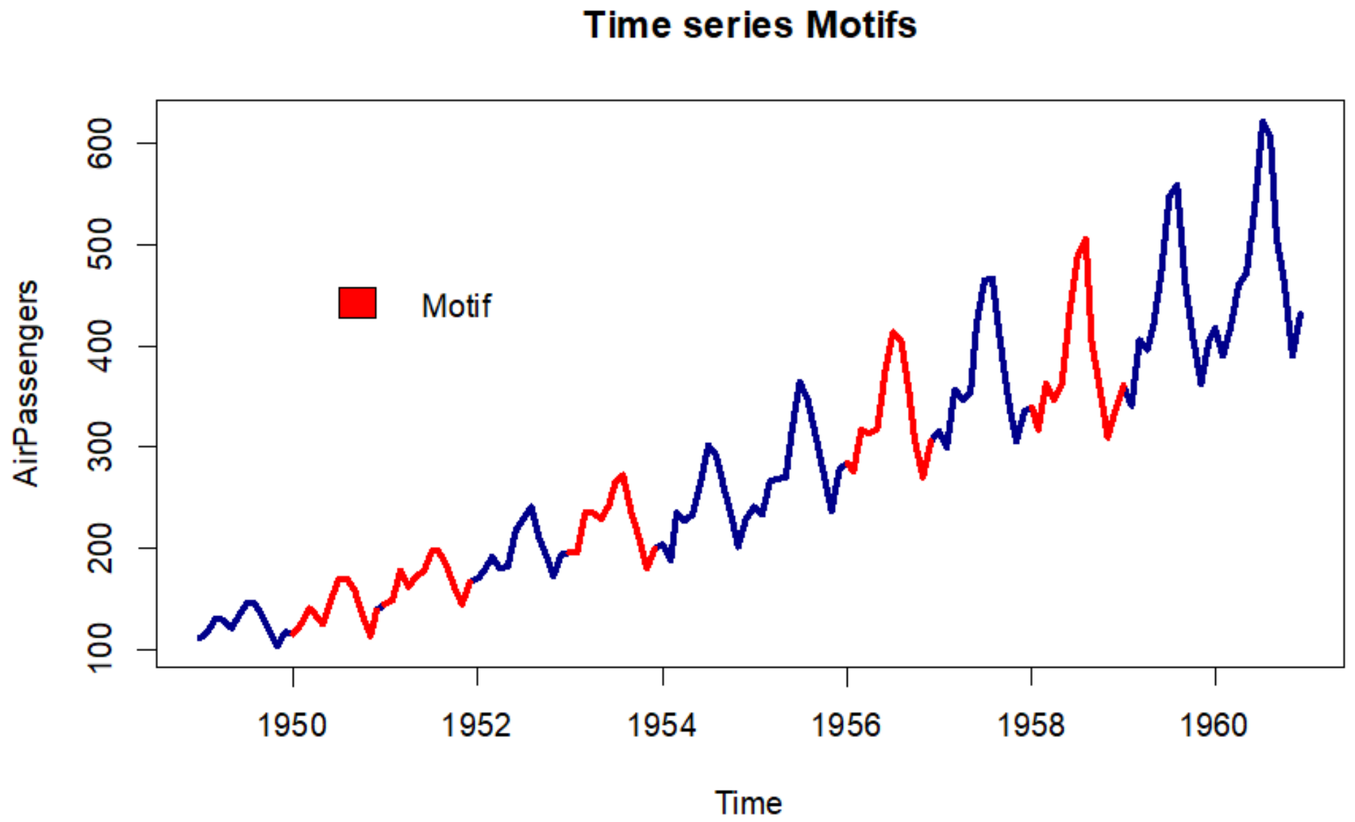


Figure 1

genes

MDPI

Article

A Detailed Gene Expression Map of *Giardia* Encystation

Laura Rojas-López [1], Sascha Krakovka [1], Elin Einarsson [1], Ulf Ribacke [2], Feifei Xu [1], Jon Jerlström-Hultqvist [1] and Staffan G. Svärd [1,3,*]

1 Department of Cell and Molecular Biology, Uppsala University, SE-751 24 Uppsala, Sweden; laura.rojas@icm.uu.se (L.R.-L.); sascha.krakovka@icm.uu.se (S.K.); elineinarsson86@gmail.com (E.E.); feifei.xu@icm.uu.se (F.X.); jon.jersltromhultqvist@icm.uu.se (J.J.-H.)
2 Department of Microbiology, Tumor and Cell Biology (MTC), Karolinska Institutet, SE-171 65 Stockholm, Sweden; ulf.ribacke@ki.se
3 SciLifeLab, Uppsala University, SE-751 24 Uppsala, Sweden
* Correspondence: Staffan.svard@icm.uu.se

Abstract: *Giardia intestinalis* is an intestinal protozoan parasite that causes diarrheal infections worldwide. A key process to sustain its chain of transmission is the formation of infectious cysts in the encystation process. We combined deep RNAseq of a broad range of encystation timepoints to produce a high-resolution gene expression map of *Giardia* encystation. This detailed transcriptomic map of encystation confirmed a gradual change of gene expression along the time course of encystation, showing the most significant gene expression changes during late encystation. Few genes are differentially expressed early in encystation, but the major cyst wall proteins CWP-1 and -2 are

Source: https://doi.org/10.3390/genes12121932

mRNA Sequencing Data

- **Giardia** encystation can be divided into an **early** and **late** phase.
- The **encystation** process, differentiation of **trophozoites** to **cysts,** is triggered by environmental changes.

- Gene expression was analyzed every 3.5 h throughout the process, starting with **trophozoites (T1)**, 3.5, 7, 10.5, 14, 17.5, 21, 24.5, 28, and 31.5 h post-induction of **encystation** (p.i.) and **mature cysts** after three days water treatment.

Pre processing done by authors:

- Started with approximately **two million** reads in three samples.
- A cut-off of **5% FDR** was used to choose the significantly differently expressed genes
- **log2** scale was used as a transformation
- in total: **5320** genes were considered

Table 1. Number of DEGs at each timepoint during encystation and excystation of Giardia (Source:Author)

Time	Up	Down
3.5	6	14
7	21	12
10.5	78	9
14	200	23
17.5	382	130
21	598	414
24.5	700	575
28	905	770
31.5	1015	934
Cyst	1786	1623
T2	2	1

Data

Table 2

Geneid <chr>	Description <chr>	T1 <dbl>	3.5h <dbl>	7h <dbl>	10.5h <dbl>	14h <dbl>	17.5h <dbl>	21h <dbl>	24.5h <dbl>	28h <dbl>	31.5h <dbl>
GL50803_11050	Hypothetical protein	0	0.00	0.00	1.18	2.62	3.41	3.66	4.10	4.25	4.35
GL50803_5206	Hypothetical protein	0	0.00	0.00	0.00	2.23	2.78	2.81	3.42	3.30	3.53
GL50803_5435	Cyst wall protein 2	0	5.14	6.76	7.02	7.22	7.48	7.39	7.27	7.25	6.99
GL50803_5638	Cyst wall protein 1	0	5.31	7.00	7.24	7.48	7.72	7.71	7.67	7.65	7.36
GL50803_92618	Nicotinamide-nucleotide adenylyltransferase	0	0.00	0.00	0.00	0.00	1.78	2.07	2.80	2.91	3.27
GL50803_14259	Glucose 6-phosphate N-acetyltransferase	0	0.00	1.58	6.27	7.27	7.75	8.07	8.28	8.41	8.45

Actions!

Before: 5320 **After**: 3908

More processing:

- Observations with more than **95% missing values** were removed from the study.
- **Standardization** by row was performed (this will make each observation unique to its mean and sd)
- **T1 (trophozoites) time** was removed from consideration (it was zero for all genes)

Gene expression over time

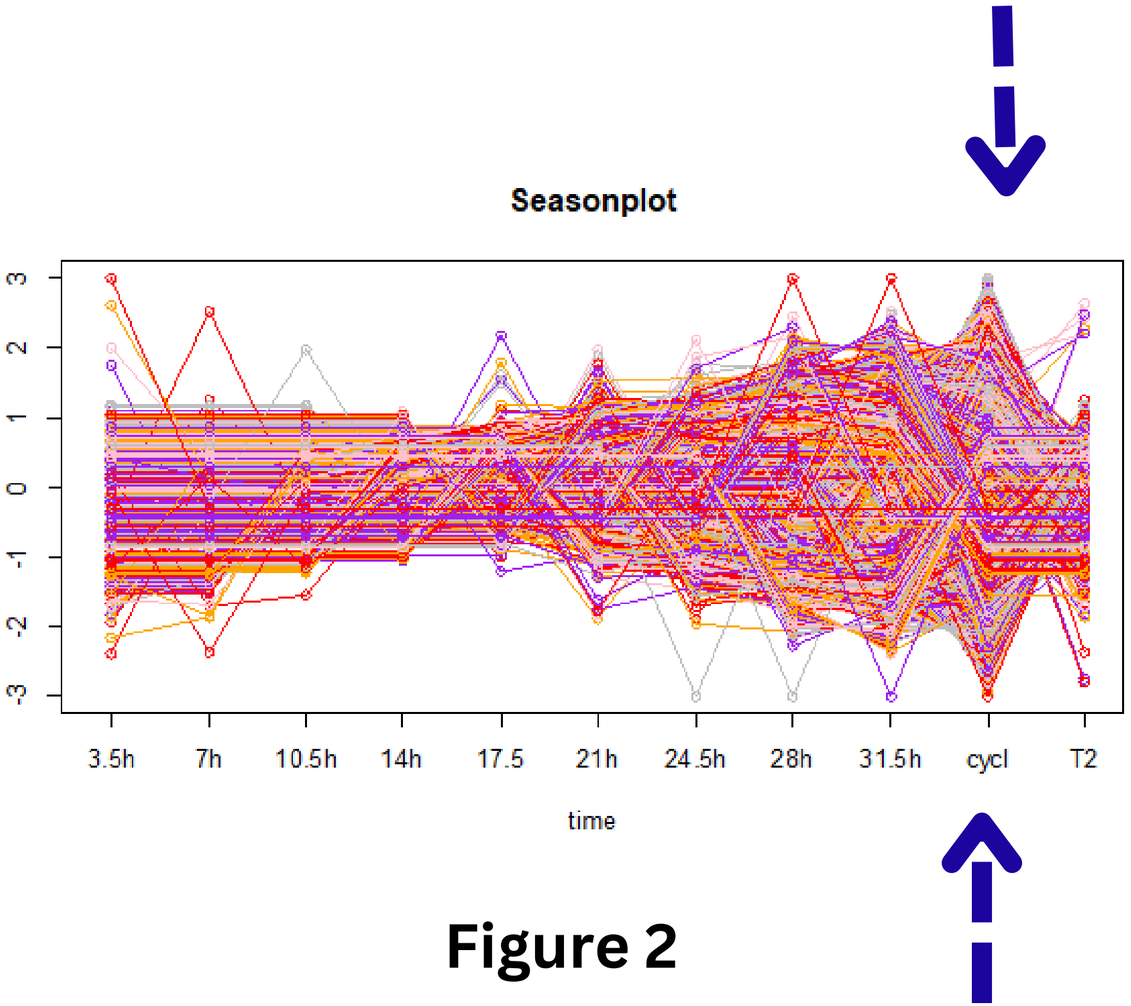


Figure 2

We observe changes over time, especially during the last period.

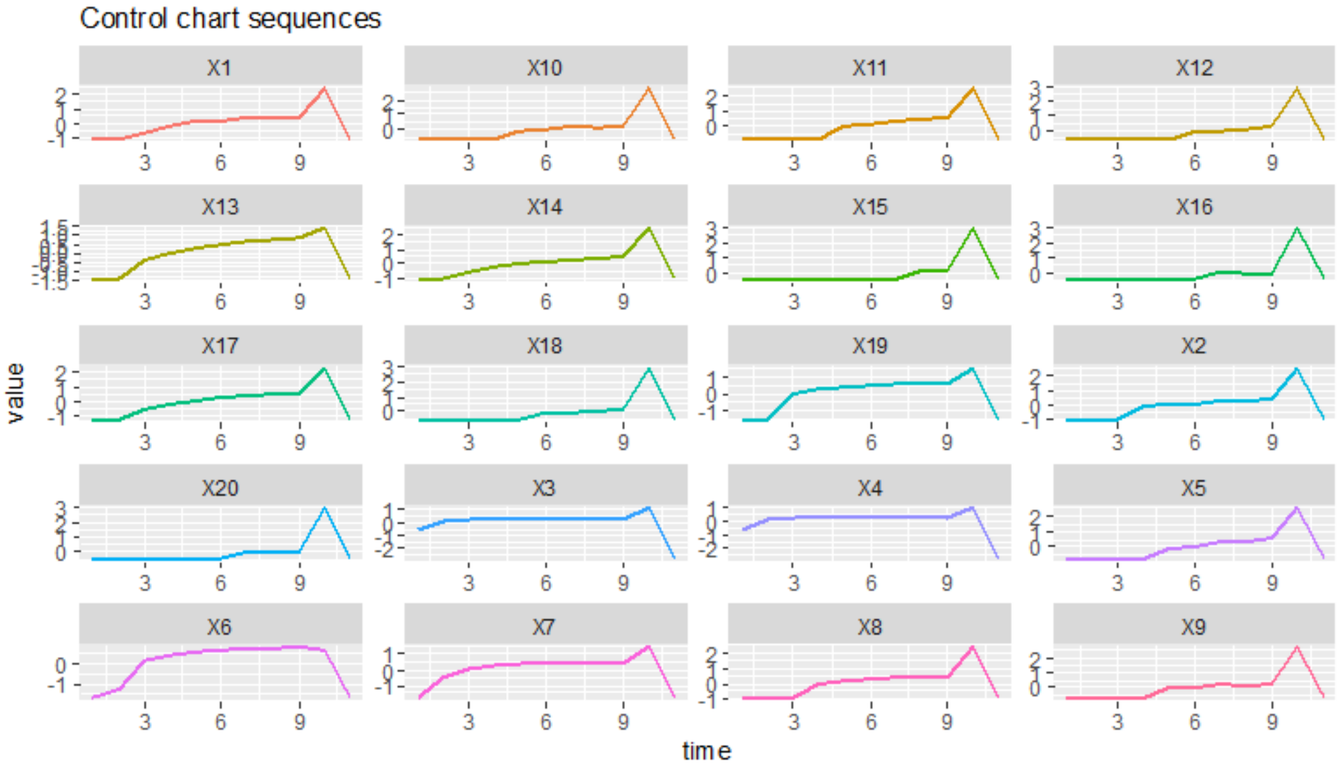


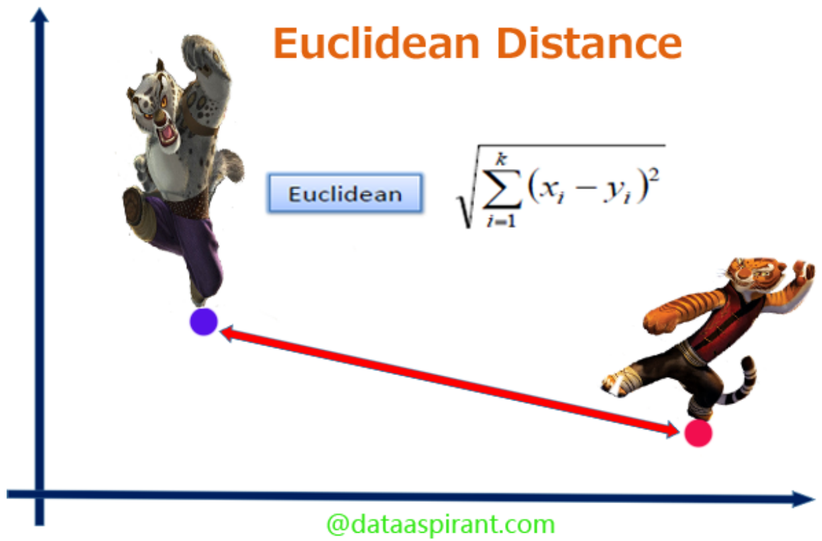
Figure 3

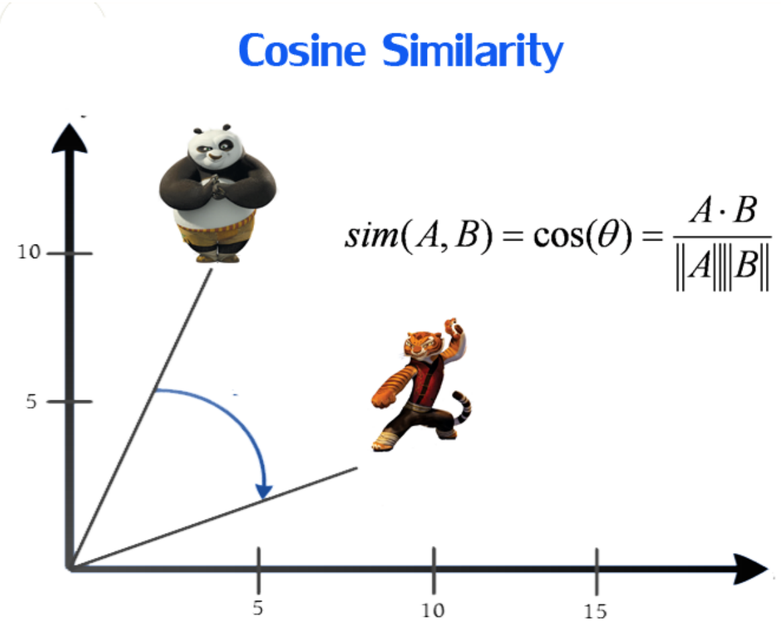
(Dis) Similarity measures

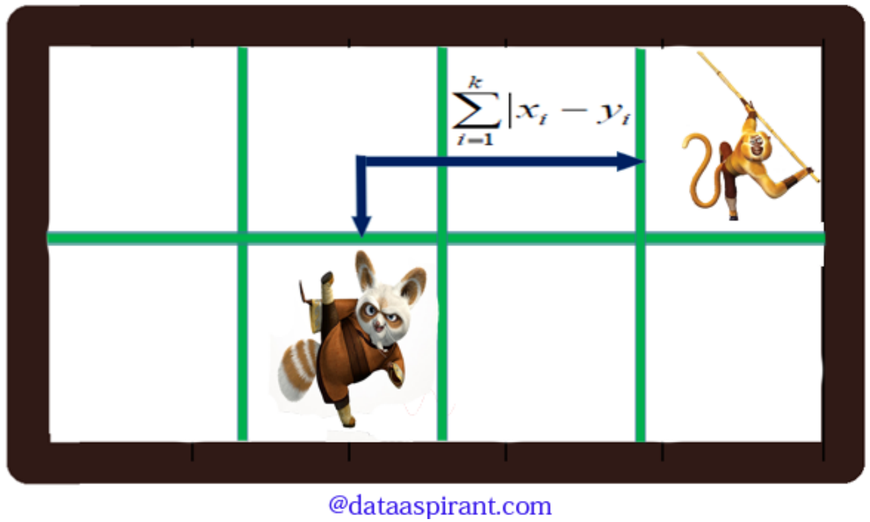
- The **similarity measure** is a way of measuring how data samples are related or **closed** to each other.

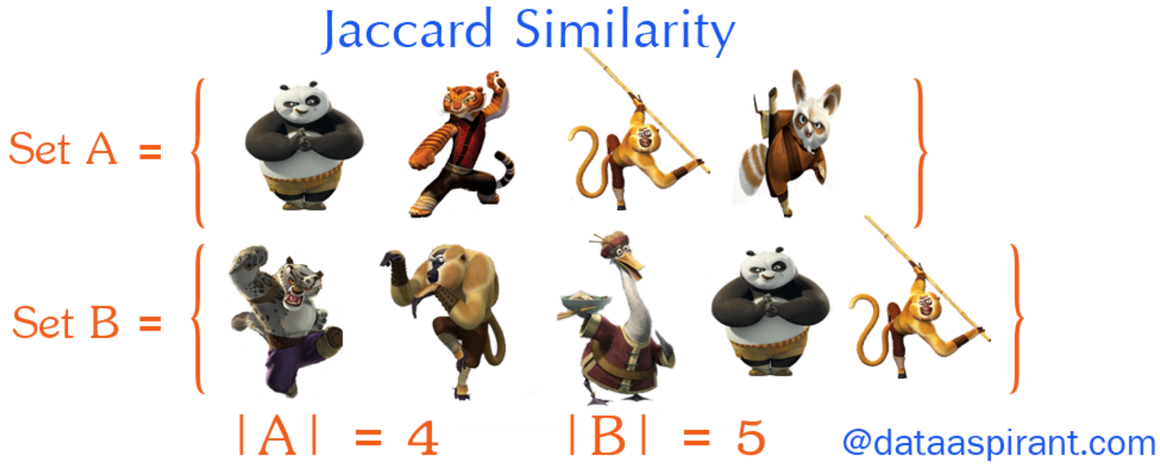
- On the other hand, the **dissimilarity measure** is used to show how much the data objects are **distinct.**

(dis)Similarity measure







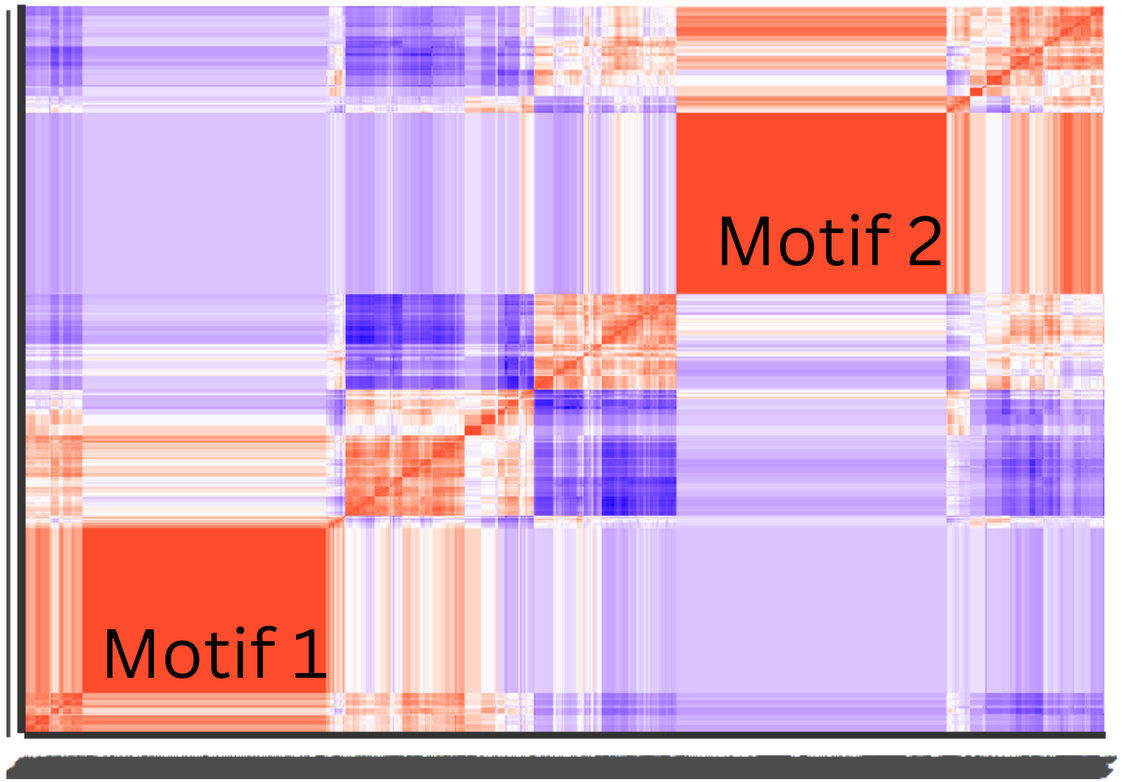


https://dataaspirant.com/five-most-popular-similarity-measures-implementation-in-python/

After standardization

Both distances have clearly created **two main clusters (red)** (corresponding to two main motifs) and other clusters with observations approximately close, and other which are significantly different among others.

Manhattan



Minkowski

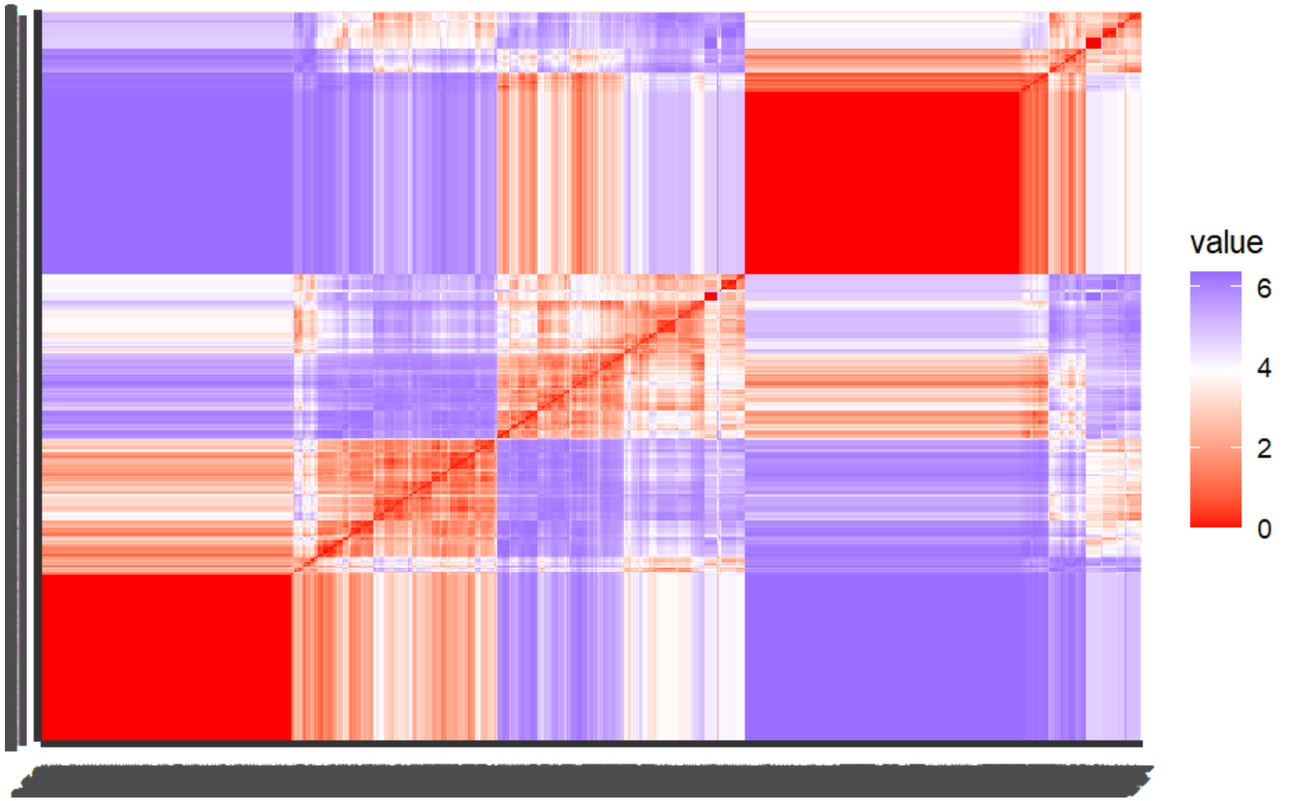


Figure 4

Univariate Time Series

- A univariate time series is defined as a series of observations of the same variable collected over time.

- Most often, the measurements are made at **regular time intervals.**

- Each gene was considered a time series with 11 observations.

- **Goal:** find most repeated patterns (motifs) and classify genes based on their behavior to further understand their characteristics.

tsmp -library in R

Three motifs (of length 11) were suggested by this algorithm.
The patterns are not clear so other tests were performed to clarify more in detail the behavior!

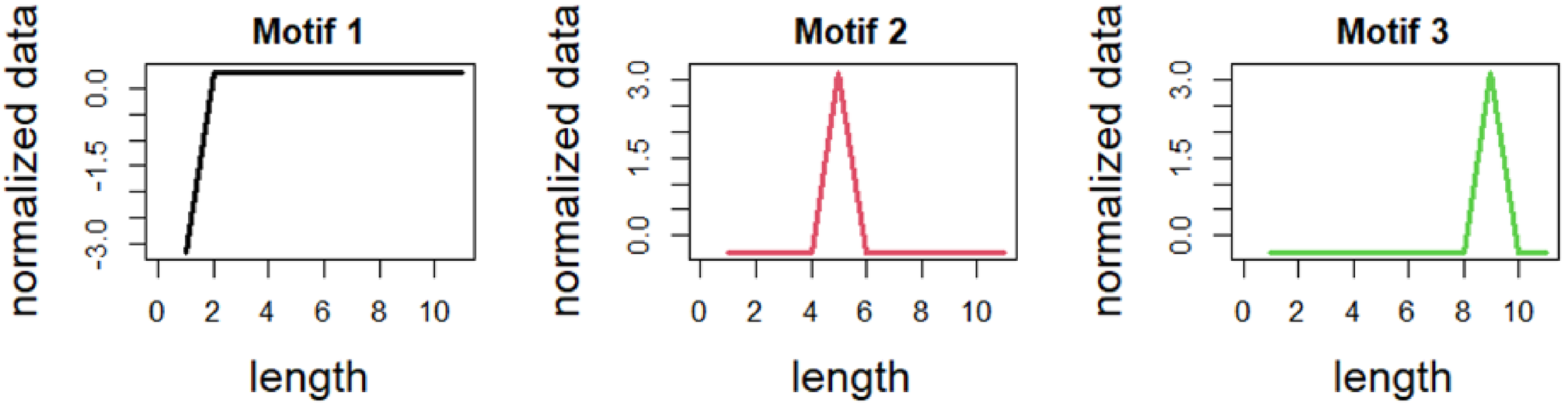


Figure 5

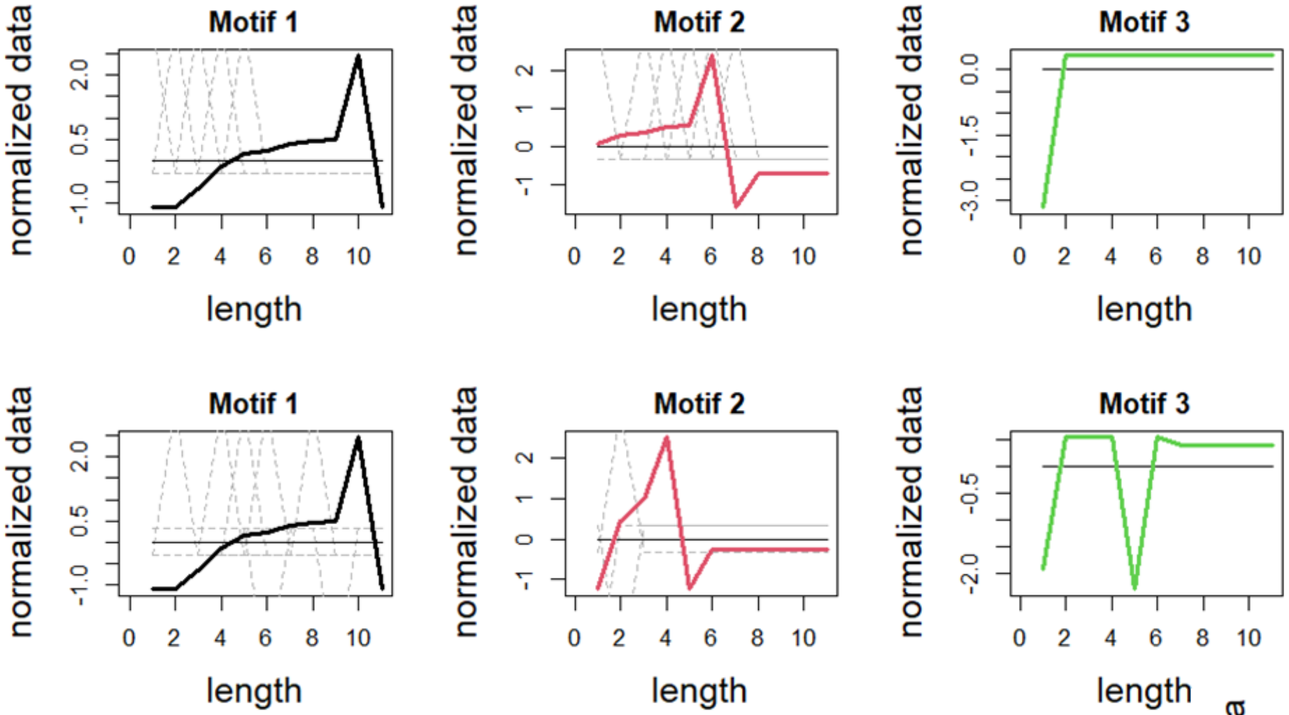


Figure 6.1

What may be observed is the fact that there are **3 motifs** or more visualized in most of the test.

Different motifs were suggested based on different coefficients considered when performing the **motifs discovery algorithm**.

length=11

Figure 6.2

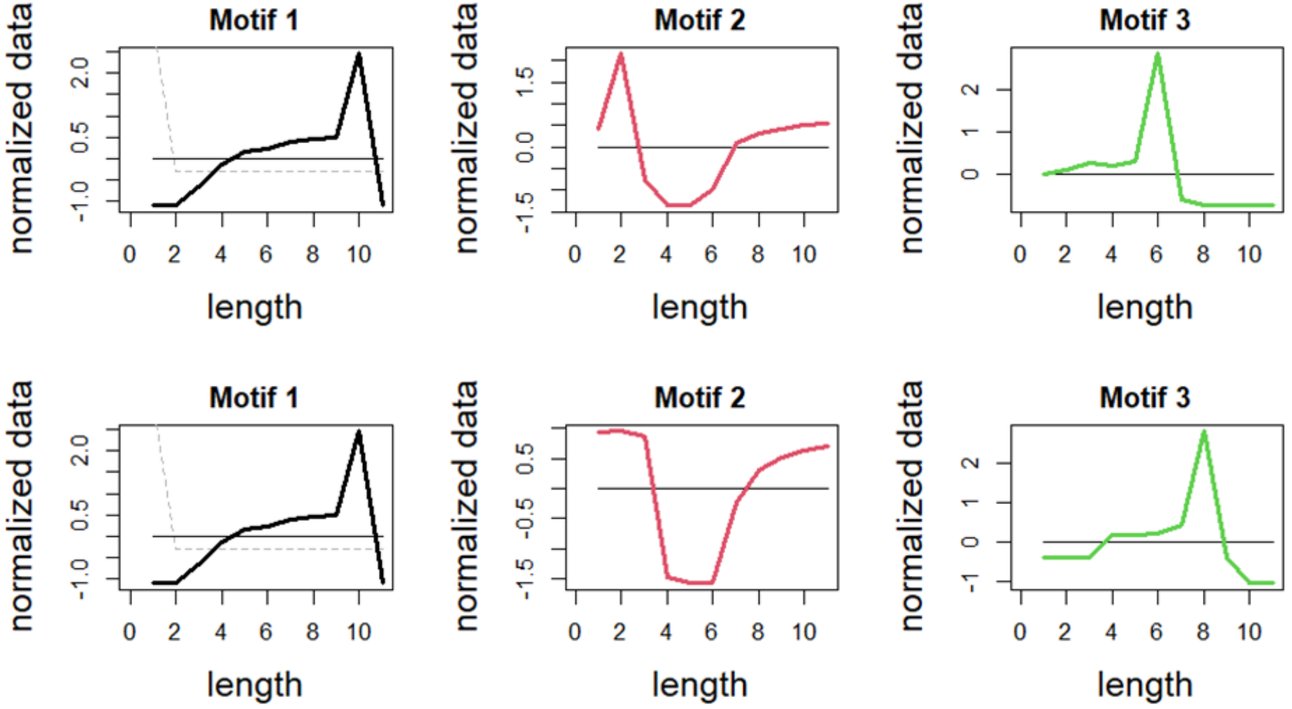
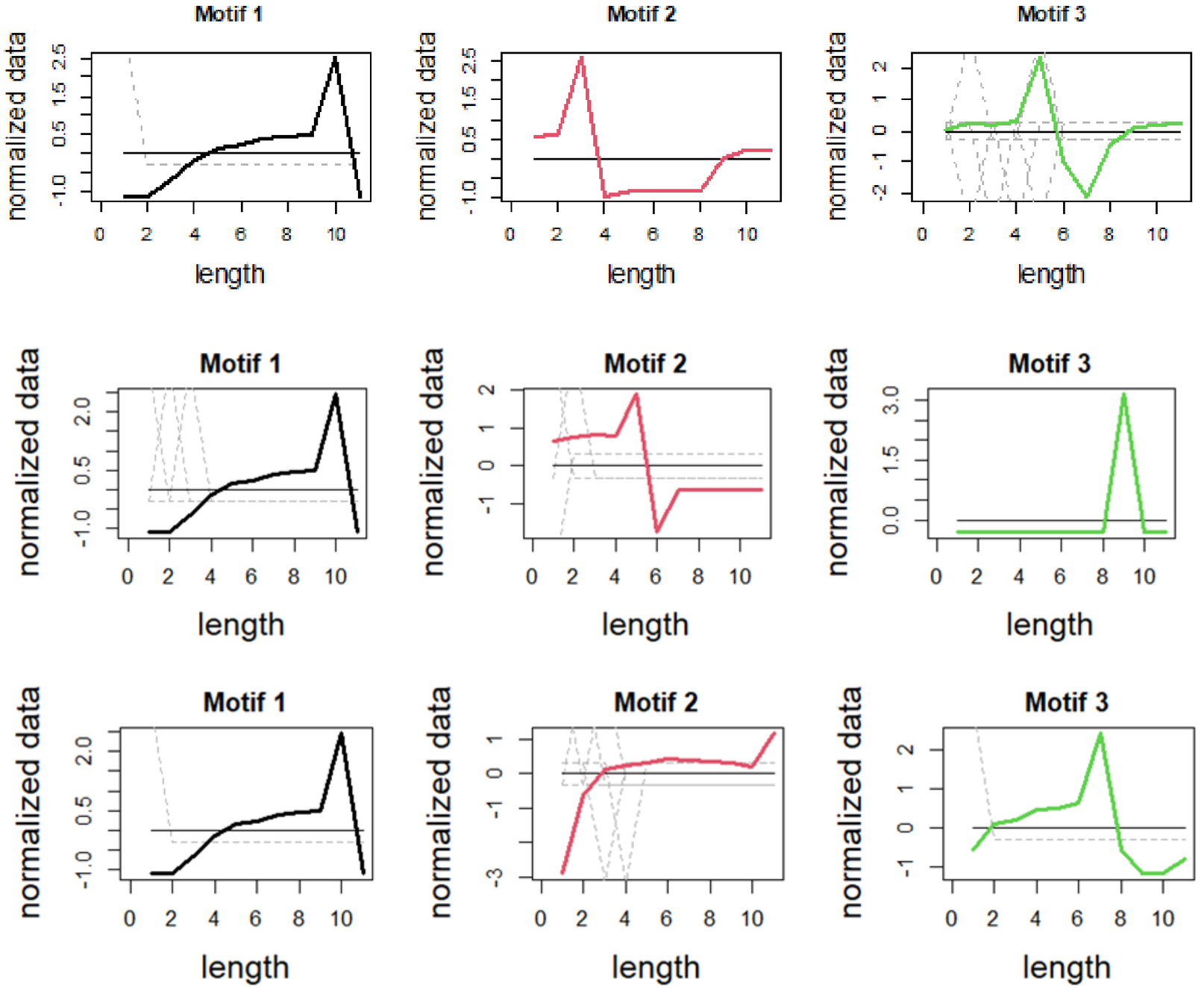


Figure 6.3


Motif 1
Motif 2
Motif 3
normalized data
length
Motif 1
Motif 2
Motif 3
normalized data
length
Motif 1
Motif 2
Motif 3
normalized data
length

Correlation graph

Correlation between **moments of time** is shown.
Cyst phase which was also observed from heatmap clearly differ from other moments of time.

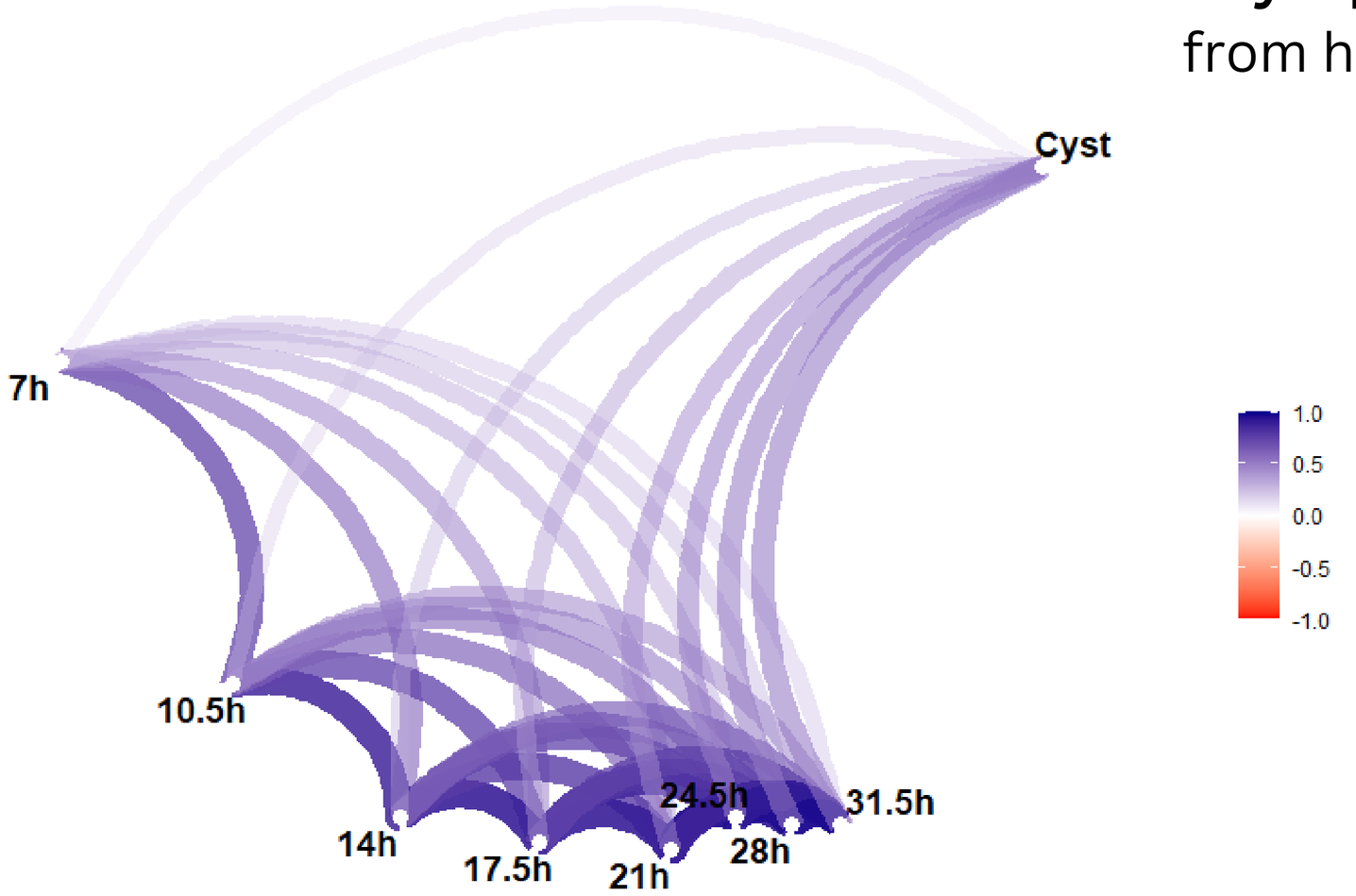


Figure 7

Cluster analysis

4 clusters clearly observed and also moment of time **10 (Cycl) has high impact** in gene expression evolution**.**

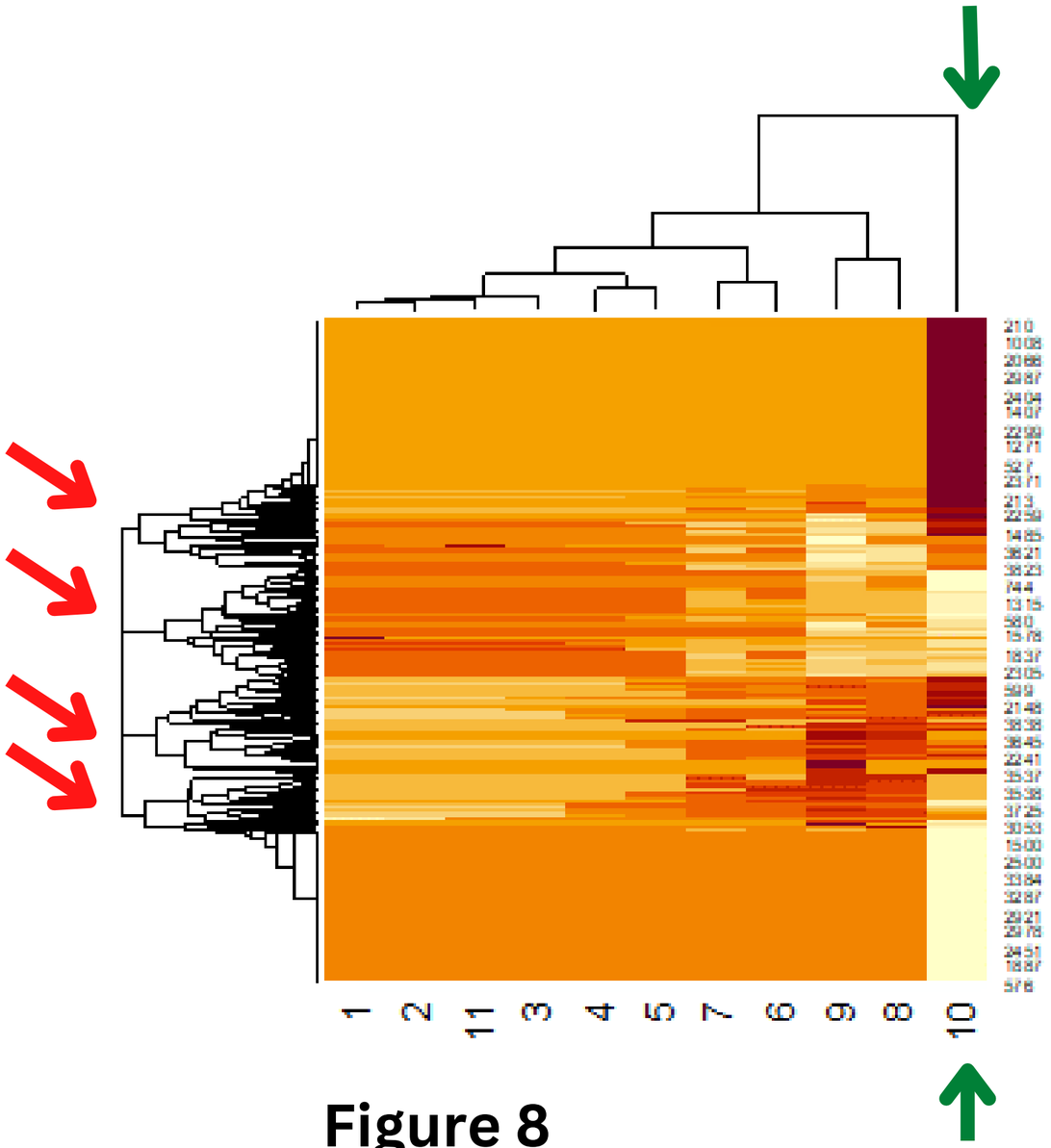


Figure 8

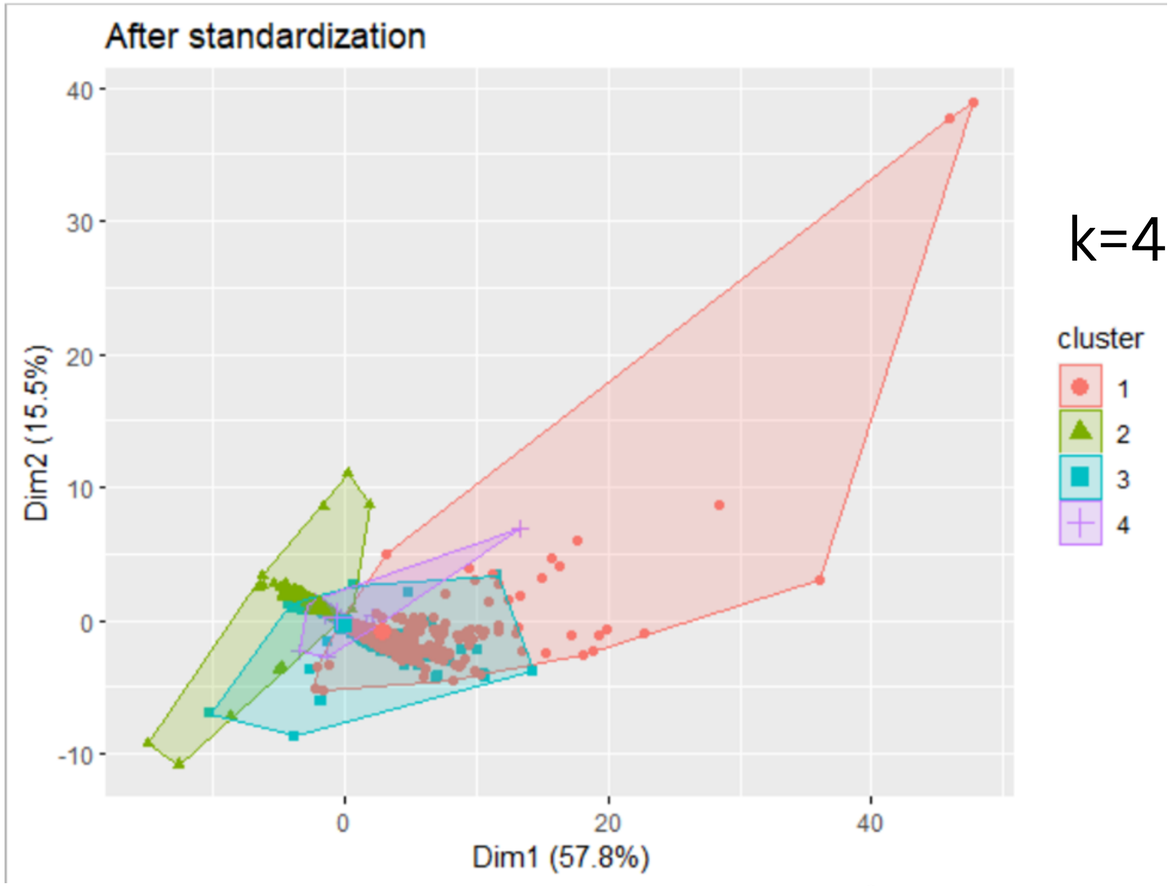
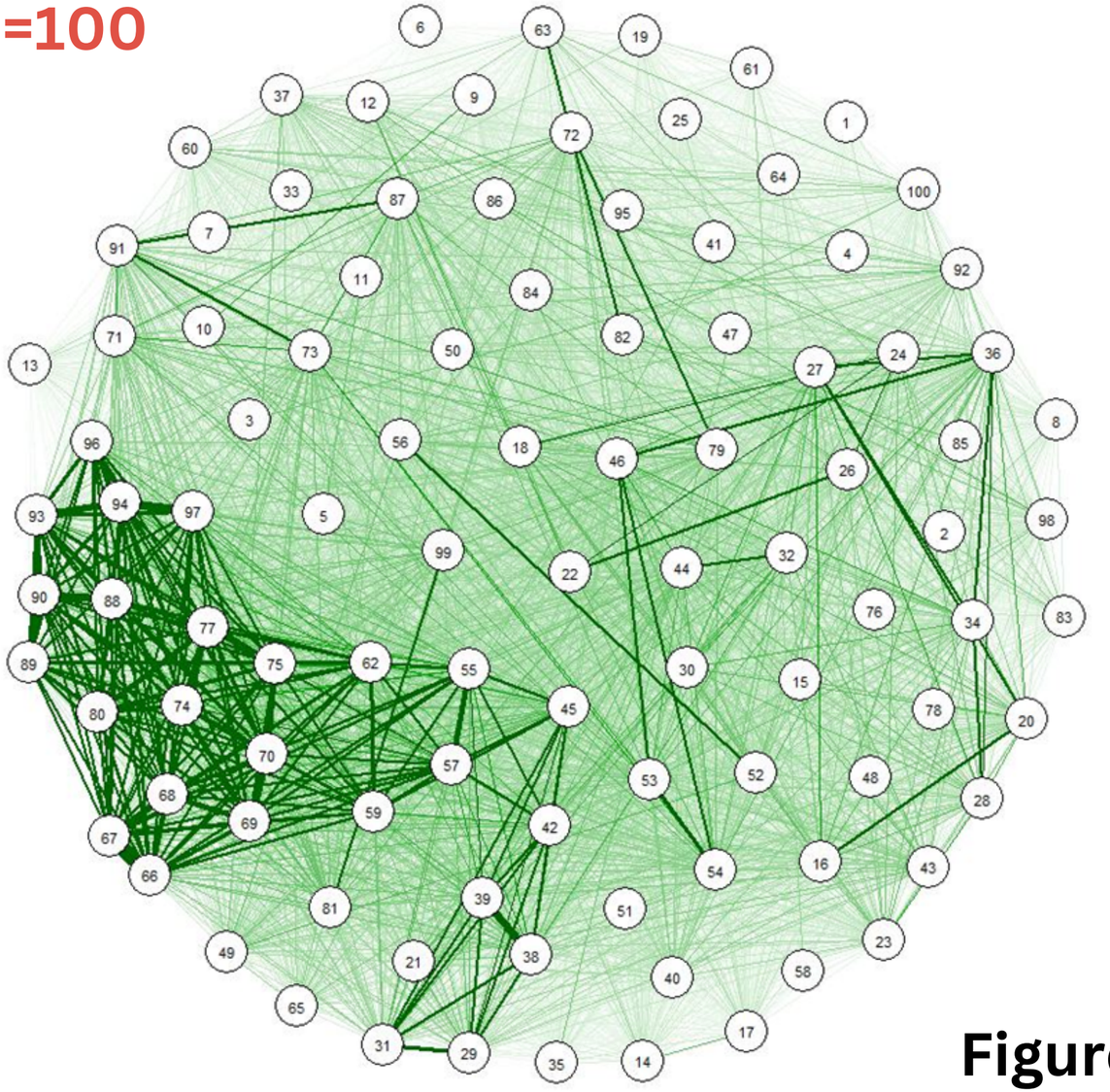


Figure 9

Graph network

Considering also the **euclidean distance** between genes a graph network was created for a better consideration of the collaboration between genes.

n=100



n=500

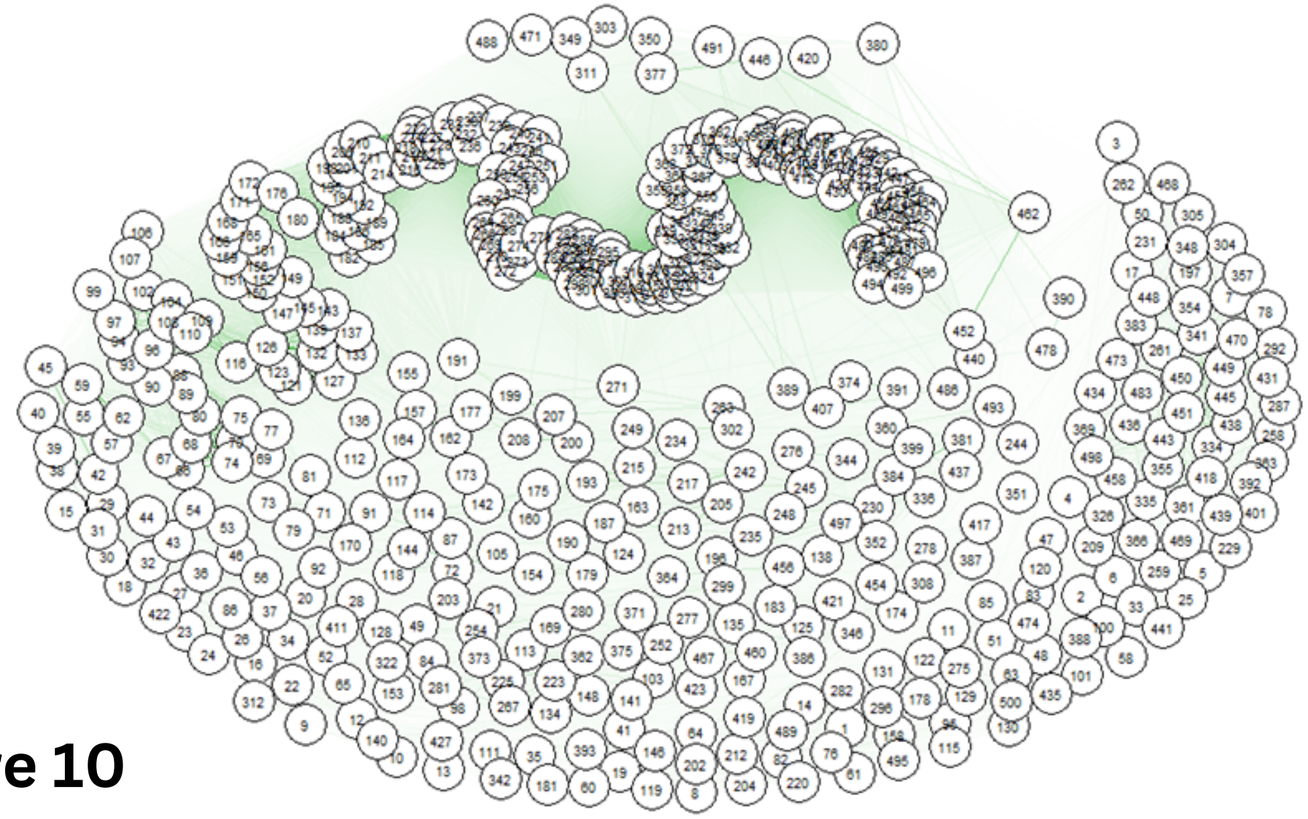


Figure 10

Graph network

Considering also the **euclidean distance** between genes a graph network was created for a better consideration of the collaboration between genes.

n=900

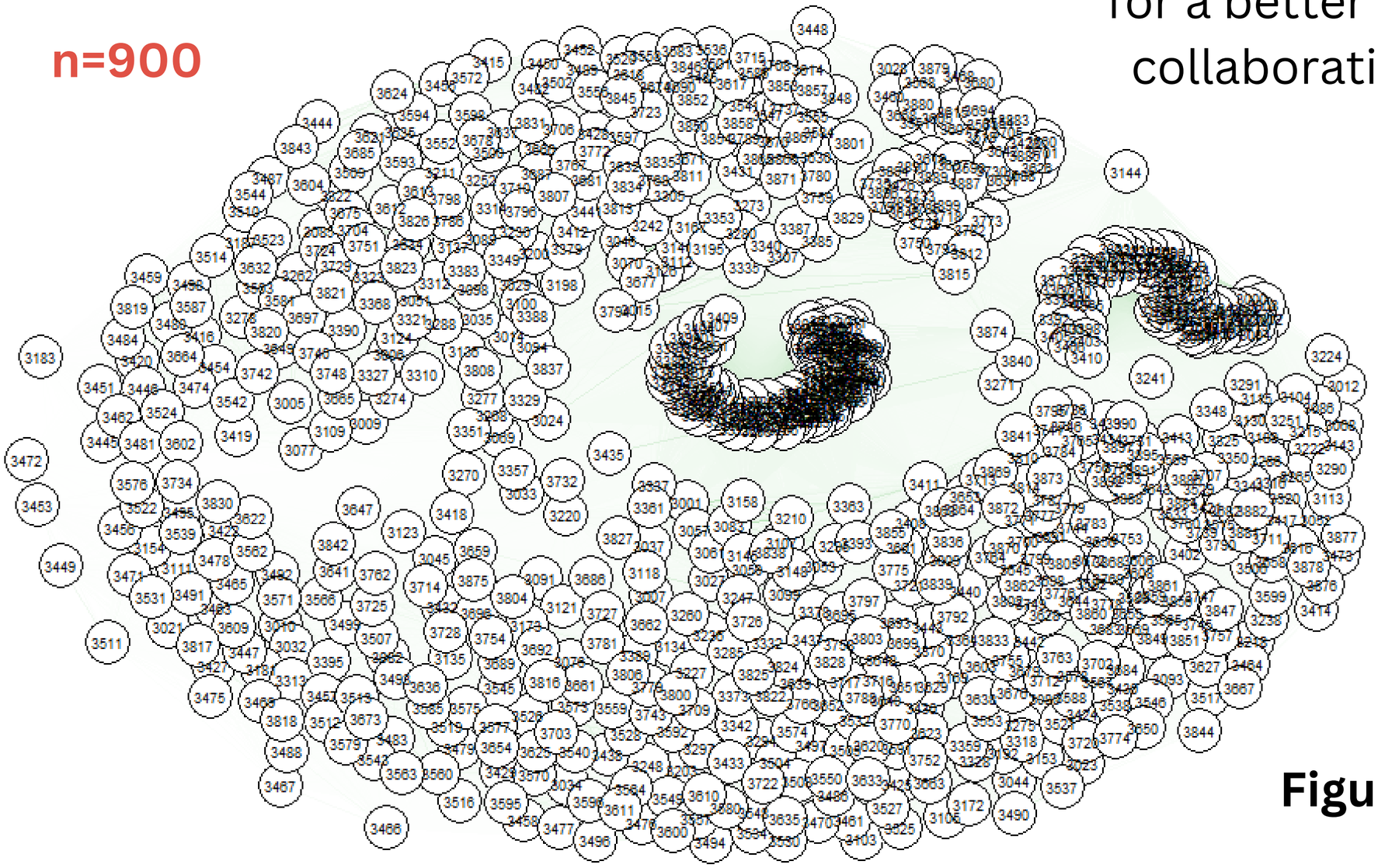


Figure 11

Cluster Analysis with kmeans- (k=4)

Euclidean distance, after standardization.

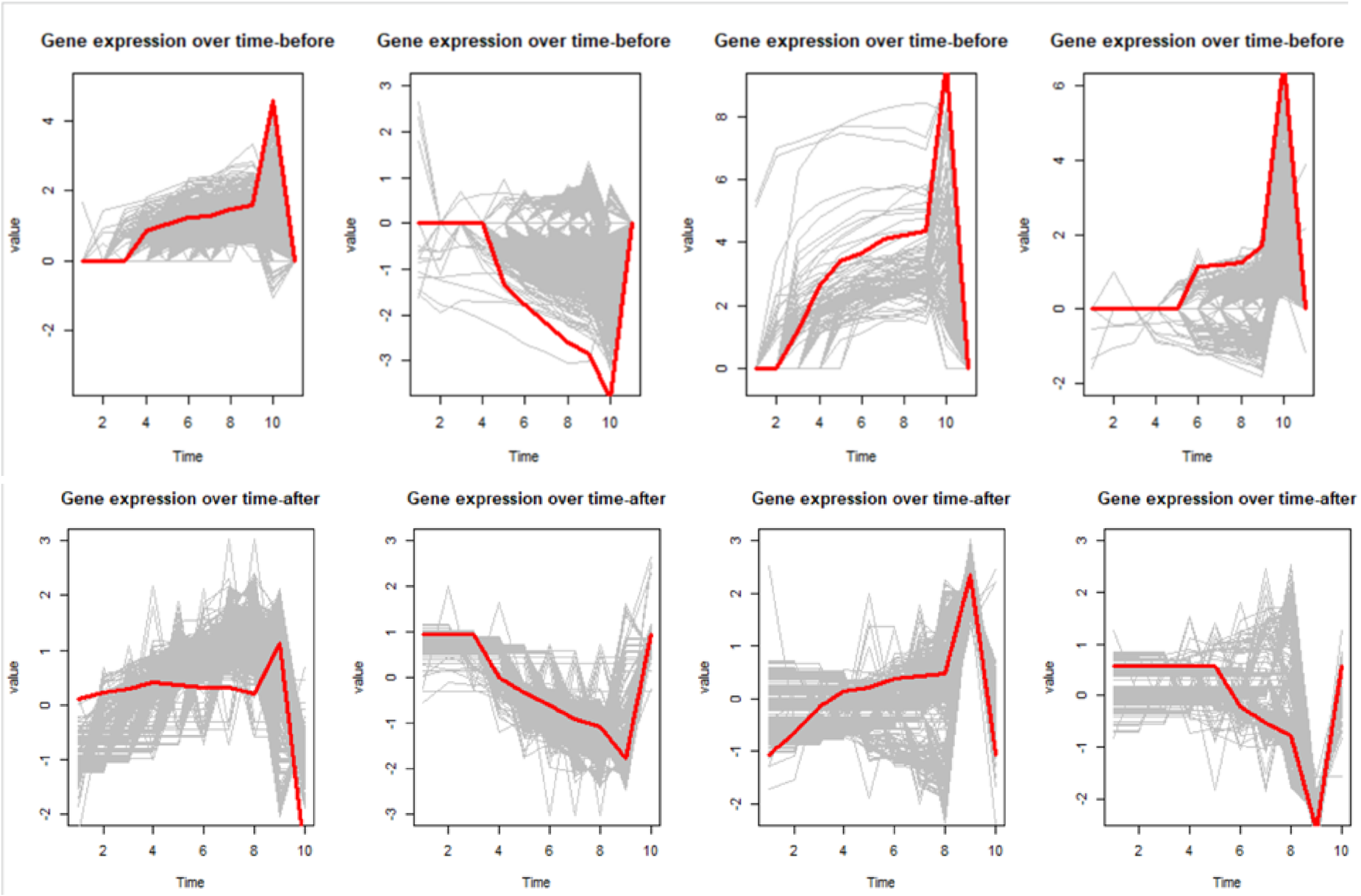
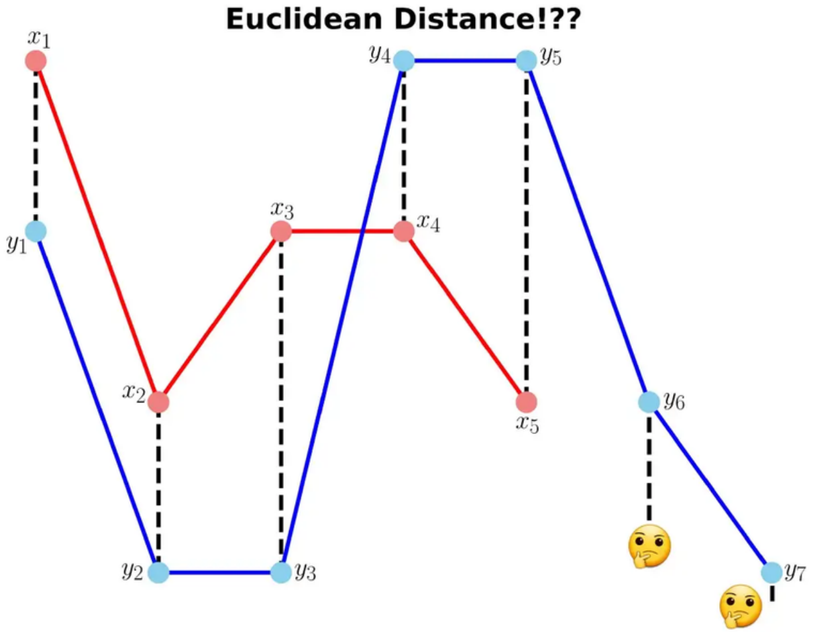
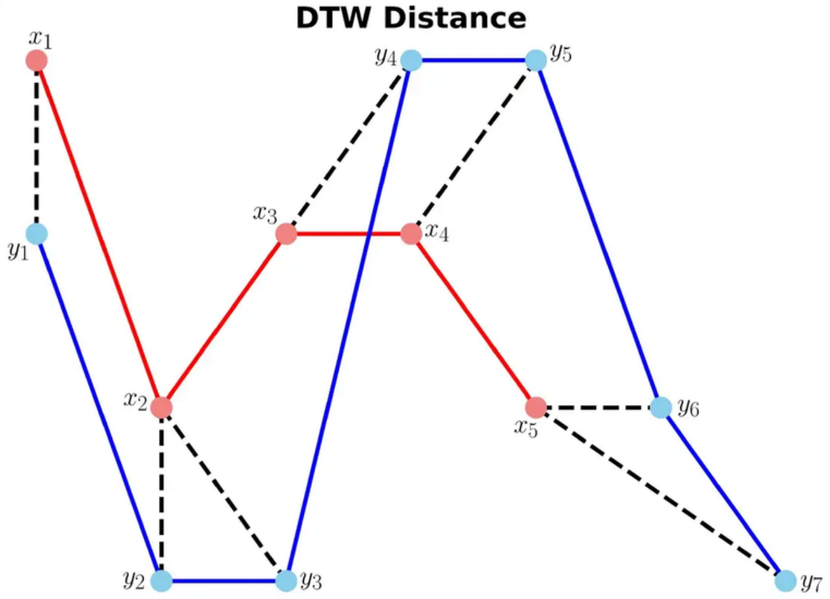


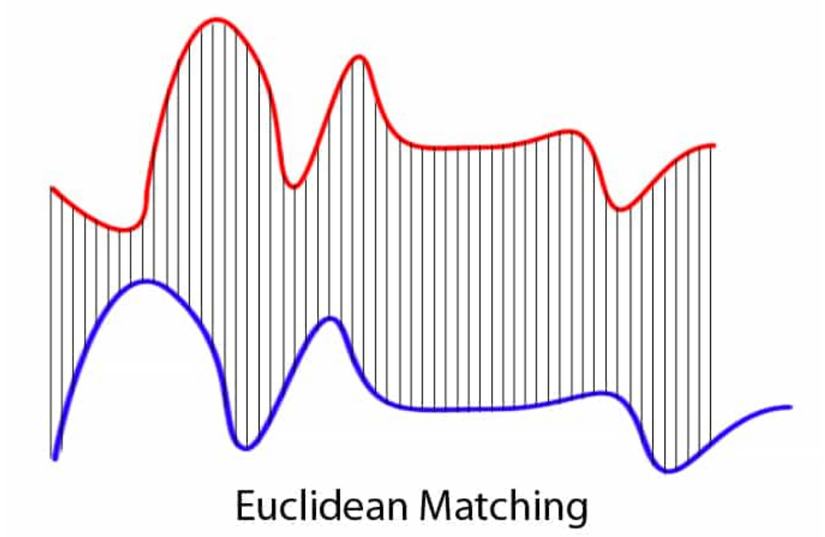
Figure 12

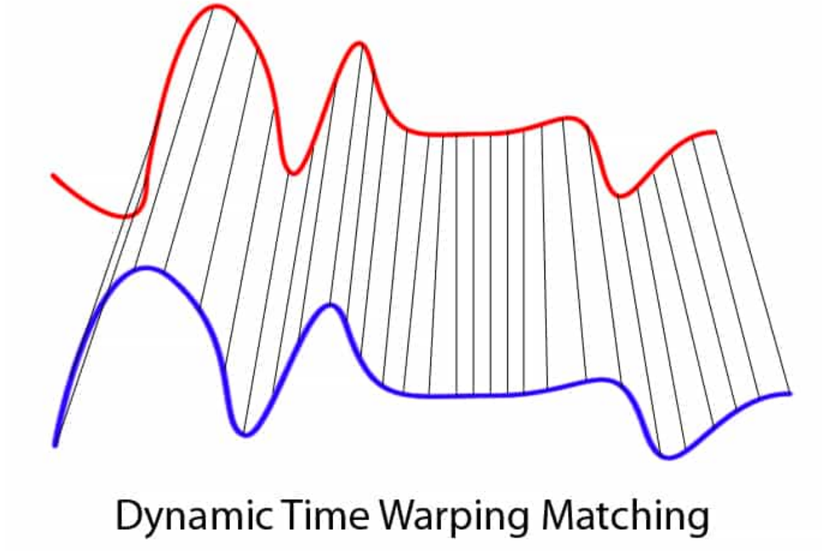
DTW-Dynamic Time Warping





Source: Wiki Commons





https://towardsdatascience.com/an-illustrative-introduction-to-dynamic-time-warping-36aa98513b98

Figure 13

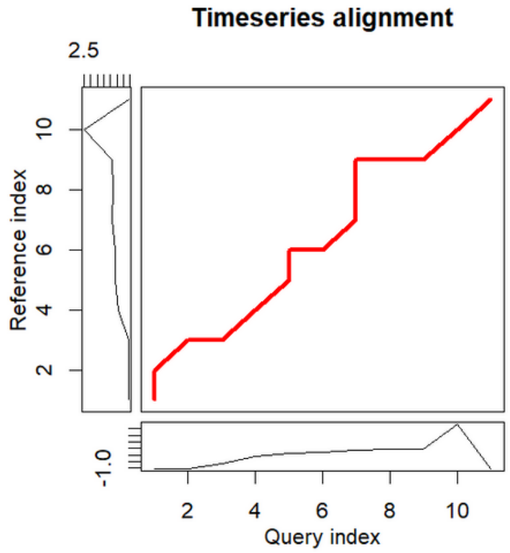
Advantages of **DTW**:

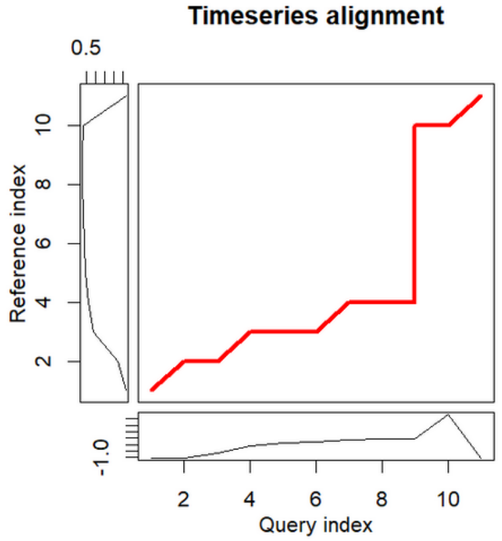
DTW is a **dynamic programming algorithm**. These algorithms break the problem recursively into subproblems (if applicable), store the results, and later use those results when needed, instead of recomputing them.

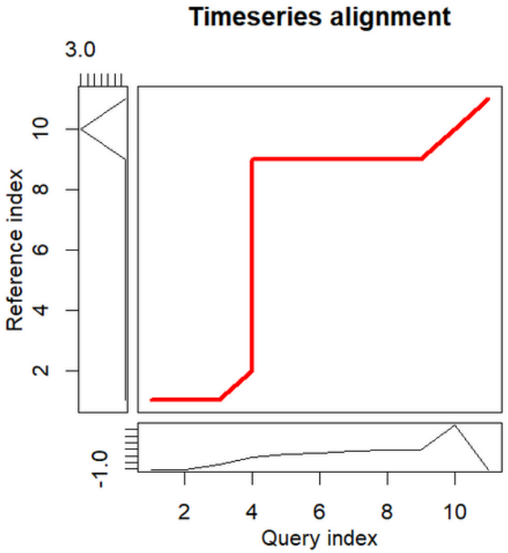
DTW takes into consideration the fact that the two time series being compared may **vary in length and speed**.

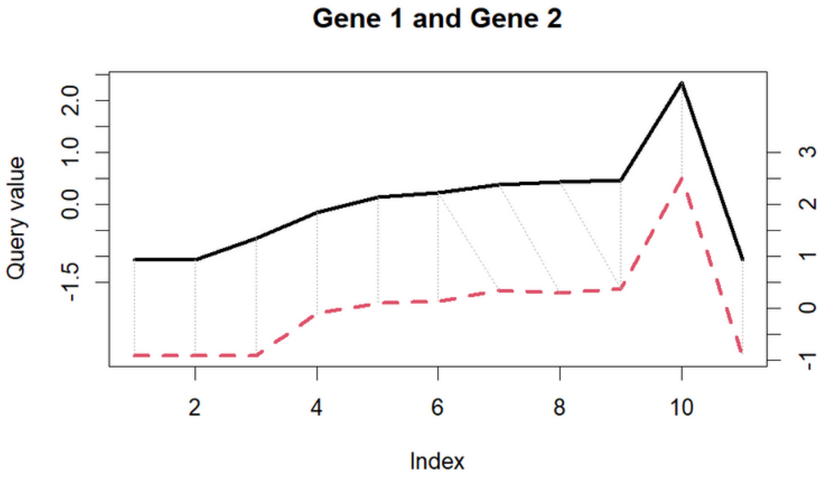
There are packages in **R** and **Python** for this similarity measure.

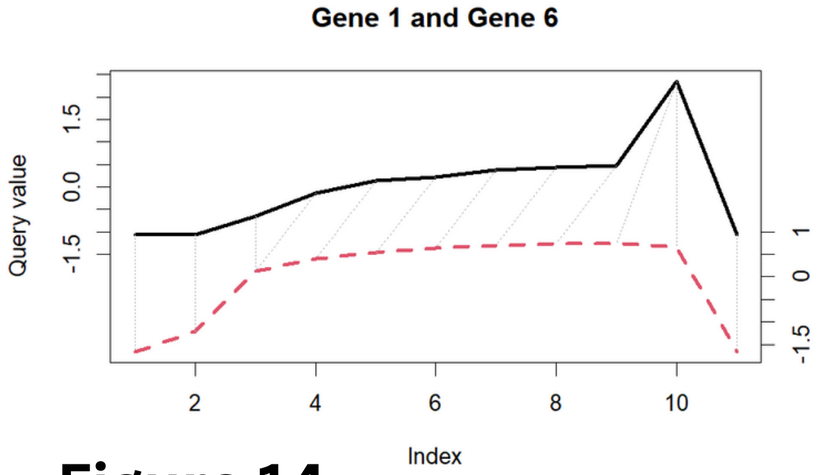
DTW library application for sequences of two genes











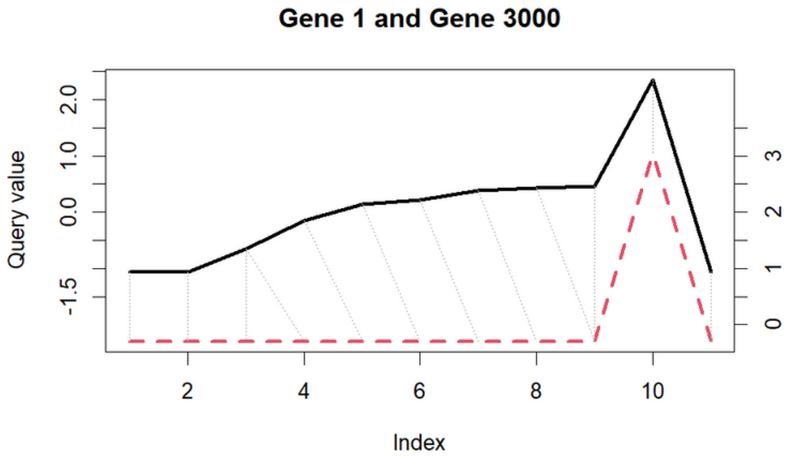


Figure 14

Conclusion and future work

Time series analysis may be a very good tool to analyze and classify gene behavior evolution based on time.

Advantages:

- Observations may be recorded on regular or irregular time period.
- You may use existing similarity measure or write your own modification and apply it. (i.e. apply DTW in k-means)

Disadvantage:

- Not many libraries in R and most of the work should be done by coding your own approach.
- Distance matrix calculation and visualizations may need high compilation time!

Next challenge! Why not create a package in R which may deal with gene expression over time data!

THANK YOU FOR WATCHING

Eralda Gjika

https://www.linkedin.com/in/eralda-dhamo-gjika-71879128/