Final Project

Statistical Methods for Big Genomic Data (STAT5900F)

by

Eralda Gjika

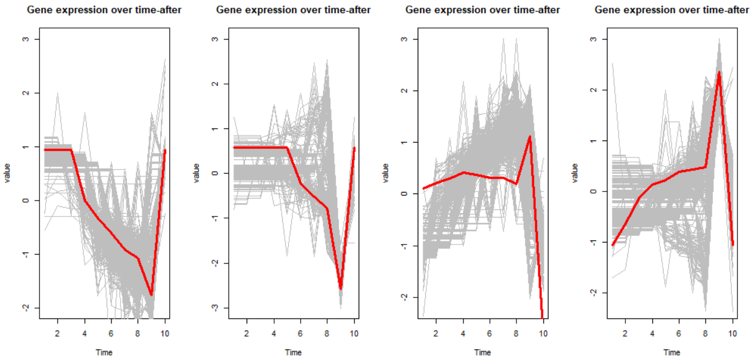
(Student ID: 101248793)

Master in Statistics (2021-2023)

TITLE

Time-Course Gene Expression Data
A Time Series Motifs Analysis for Giardia Encystation Data

(Due: December 22, 2022)



School of Mathematics and Statistics
Carleton University

Contents

1 Introduction

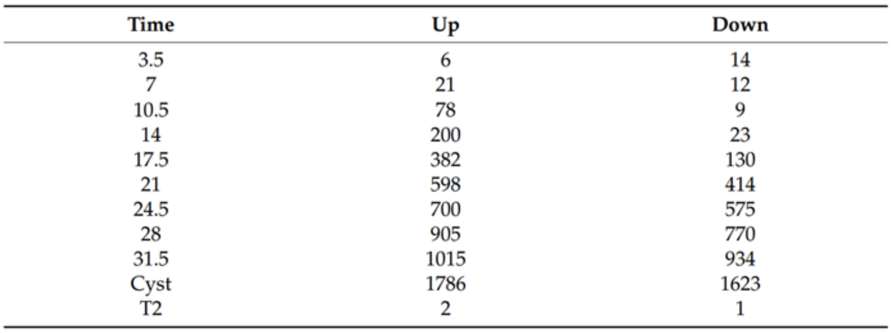
One of the oldest definition of motifs has been presented in musical patterns, literature, genomic and time series analysis as well. Motifs may be defined as repeated patterns in a sequence of observed values such as in time series. Analyzing motifs leads to a deeper understanding of the work.

In this study I will explore motifs in Giardia intestinalis data which 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. The data originaly taken from the article of Rojas-López et. al (2021) show a gradual change of gene expression along the time course of encystation, showing the most significant gene expression changes from encystation to the cyst stage. (Rojas-López et. al (2021))

The dataset consists of genes and their behaviour measured over different time points starting from trophozoites and then at: $3.5h, 7h, 10.5h, 14h, 17.5h, 21h, 24.5h, 28h, 31.5h$, *cyst* and *excysted trophozoites*. In total 12 moments of time and **5321 genes**. The authors have started their study with approximately 2 million observations and their preliminary analysis was:

They used a cut-off of 5% FDR to choose the significantly differently expressed genes, and $log2$ scale was used as a transformation. Differentially expressed genes were used to detect functional clusters at different timepoints (3.5–7 h, 10.5–14 h, 17.5–21 h, 24.5–28 h, 31.5 h–Cyst, T2) by authors.

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



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

Table 1 shows the results obtained from the authors. They showed that as time increase the number of differentially expressed genes increase as well. The highest number was achieved at cyst phase both for UP and DOWN regulated genes.

From the study presented by the authors, only six genes were found to be significantly upregulated after 3.5 hours of growth in *encystation medium*. Among these genes, the most notable upregulation was observed in the two major cyst-wall proteins, CWP-1 and CWP-2. These proteins are likely important for the formation of the cyst wall, a protective structure that is formed by some microorganisms during periods of stress or adverse conditions. The upregulation of these genes suggests that they play a crucial role in the formation of the cyst wall and may be important for the survival of the organism under the conditions of the experiment.

The expression profile of the genes in question appears to follow a cyclic pattern during the process of *encystation*, with increased expression at specific time points. This pattern of expression is unique among the "giardia cyclins," which are a group of genes involved in regulating the cell cycle of the organism. The upregulation of these genes at specific time points may be important for the proper coordination and completion of the encystation process.

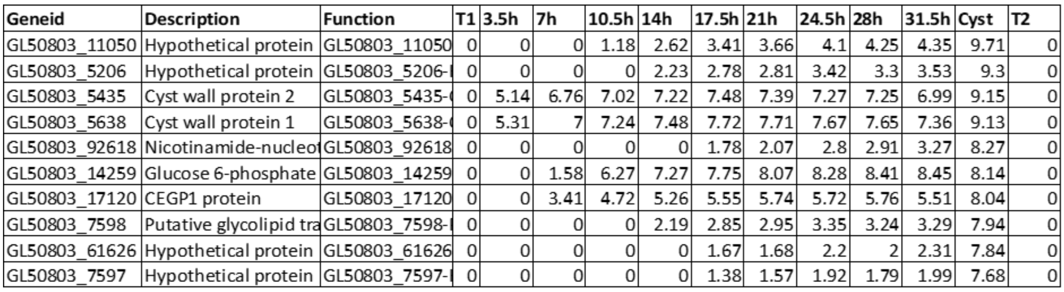
The study describes that there are significant changes in gene expression during the process of *encystation.* At 10.5 hours of encystation, there are 79 up-regulated genes and only nine down-regulated genes, with most of these genes overlapping with those that were upregulated at the 3.5-hour timepoint. At 14 hours of encystation, there are 200 up-regulated genes and 23 down-regulated genes. This suggests that there are major changes in certain cellular processes during this time, and that most of the genes needed for cyst formation peak at this stage. This information highlights the complex and dynamic nature of the encystation process and the importance of gene regulation in this process.

During the $mid - encystation\ stage$ $(17.5and21h)$, there are large changes in gene expression, with many genes being up- or down-regulated. At 17.5 hours of encystation, there are 382 up-regulated genes and 130 down-regulated genes, and at 21 hours, there are 598 up-regulated genes and 414 down-regulated genes. These changes in gene expression are associated with changes in many different metabolic pathways and protein synthesis, suggesting that the mid-encystation stage is an important transition in the Giardia life cycle. This transition likely involves significant changes in the organism's physiology and metabolism that are necessary for the completion of the encystation process and the formation of the protective cyst.

In the $Late\ Encystation$ $(24.5to31.5h)$ it appears that among the genes that are most significantly up- or down-regulated during the encystation process, there are ones that are involved in lipid metabolism and antigenic variation. Lipid metabolism refers to the processes by which an organism synthesizes, breaks down, and transports lipids, which are a type of molecule that includes fats, waxes, and sterols. Antigenic variation, on the other hand, refers to the ability of some microorganisms to change the surface proteins that are exposed on their cell surface. This allows the organism to evade the host immune system and increase its chances of survival. The up-regulation or down-regulation of these genes during encystation may be important for the proper coordination of these processes and the successful completion of the encystation process.

The largest changes in gene expression were observed in the *cyst stage*, where 3409 DEGs (from the initial number from authors) were identified, including genes involved in lipid metabolism and antigenic variation. The expression levels of these genes showed a gradual change over the encystation process, with large differences between the 31.5-hour time point and the cyst stage. These changes in gene expression likely reflect the significant physiological and metabolic changes that occur during encystation and are important for the survival of the organism in adverse conditions.

Table 2 Log2 transformed data and information. (Source: excel file)



Geneid	Description	Function	T1	3.5h	7h	10.5h	14h	17.5h	21h	24.5h	28h	31.5h	Cyst	T2
GL50803_11050	Hypothetical protein	GL50803_11050	0	0	0	1.18	2.62	3.41	3.66	4.1	4.25	4.35	9.71	0
GL50803_5206	Hypothetical protein	GL50803_5206-	0	0	0	0	2.23	2.78	2.81	3.42	3.3	3.53	9.3	0
GL50803_5435	Cyst wall protein 2	GL50803_5435-	0	5.14	6.76	7.02	7.22	7.48	7.39	7.27	7.25	6.99	9.15	0
GL50803_5638	Cyst wall protein 1	GL50803_5638-	0	5.31	7	7.24	7.48	7.72	7.71	7.67	7.65	7.36	9.13	0
GL50803_92618	Nicotinamide-nucleot	GL50803_92618	0	0	0	0	0	1.78	2.07	2.8	2.91	3.27	8.27	0
GL50803_14259	Glucose 6-phosphate	GL50803_14259	0	0	1.58	6.27	7.27	7.75	8.07	8.28	8.41	8.45	8.14	0
GL50803_17120	CEGP1 protein	GL50803_17120	0	0	3.41	4.72	5.26	5.55	5.74	5.72	5.76	5.51	8.04	0
GL50803_7598	Putative glycolipid tra	GL50803_7598-	0	0	0	0	2.19	2.85	2.95	3.35	3.24	3.29	7.94	0
GL50803_61626	Hypothetical protein	GL50803_61626	0	0	0	0	0	1.67	1.68	2.2	2	2.31	7.84	0
GL50803_7597	Hypothetical protein	GL50803_7597-	0	0	0	0	0	1.38	1.57	1.92	1.79	1.99	7.68	0

2 Methodology

Pre − processing
The authors of the study did a pre-processing analysis on the data but for a better fit to our approach we needed to further improve the way the data will be used.
The transformation used by the authors aiming to select the genes which were differentially expressed was $Log2$ transformation (shown in Table 2. What was noticed in the reduced set (obtained from the authors) was the fact that time T1 (trophozoites) which was a starting time of observations was zero for all observations. This moment of time was removed from the study since we needed to transform the data in such a way that similarity or dissimilarity measures may be applied properly. Then we also observed that there were many observations (genes) which have more than 95% of the observed moments of time values missing. After this we ended with **3908 genes**. The next step was using a standardization process by row for all observations, this will make each observation unique to its mean and standard deviation. From this moment and on every gene is considered as a time series with periodicity $p = 11$ (the 11 time points).

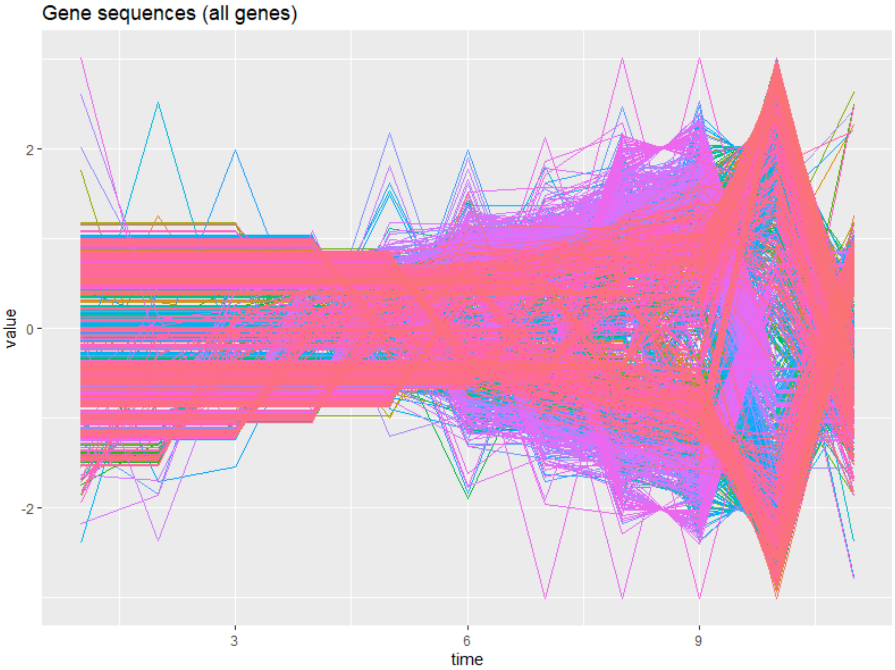


Figure 1: Gene expression over 11 moments of time (Source: E.Gjika)

Figure 1 shows the behavior of all genes over the time period of 11 moments of time. And Figure 2 shows the first 4 genes and their behavior over time. We may notice similar patterns as

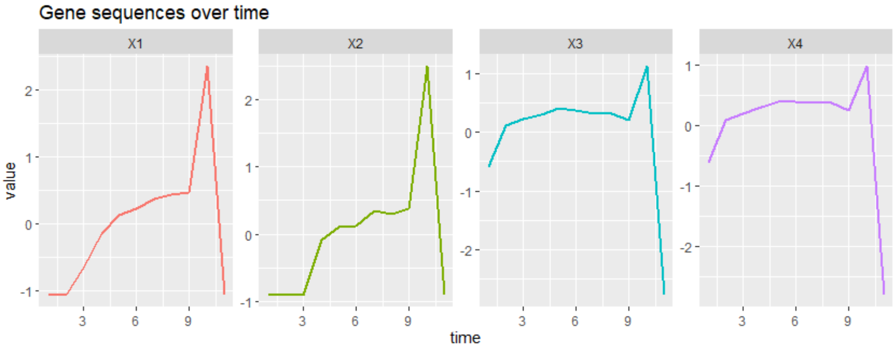


Figure 2: First four genes expression over 11 moments of time (Source: E.Gjika)

well differences especially during the last moments of time which may be effect of the conditions changed during the evolution of Gardia during the *Cyst* phase.

Similarity measures

In this study we will use similarity and dissimilarity measure to analyze genes behavior over time. 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. First we present some definitions of the mostly used similarity/dissimilarity measure used for classification methods.

Definition Let's suppose we have two vectors $X = (X_1, X_2, ..., X_n)$ and $Y = (Y_1, Y_2, ..., Y_n)$ with numerical observed values. *Manhattan distance* is defined and calculated as below:

$$Manhattan(X, Y) = \sum_{i=1}^{n} |X_i - Y_i| \tag{1}$$

Definition Let's suppose we have two vectors $X = (X_1, X_2, ..., X_n)$ and $Y = (Y_1, Y_2, ..., Y_n)$ with numerical observed values. *Minkowski distance* is defined and calculated as below:

$$Minkowski(X, Y) = (\sum_{i=1}^{n} (X_i - Y_i)^p)^{\frac{1}{p}} \tag{2}$$

Definition Let's suppose we have two vectors $X = (X_1, X_2, ..., X_n)$ and $Y = (Y_1, Y_2, ..., Y_n)$ with numerical observed values. *Euclidean distance* is defined and calculated as below:

$$Euclidean(X, Y) = \sqrt{(\sum_{i=1}^{n} (X_i - Y_i)^2)} \tag{3}$$

DTW-Dynamic Time Warping is a dynamic programming algorithm. These algorithms break the problem recursively into subproblems , 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 DTW.
To align two sequences using DTW, we construct an $n - by - m$ matrix where the (ith, jth) element

of the matrix contains the distance $d(x_i, y_j)$ between the two points x_i and y_j (i.e. $d(x_i, y_j) = (x_i$-$y_j)^2$. Each matrix element (i, j) corresponds to the alignment between the points x_i and y_j. A warping path W is a contiguous (in the sense stated below) set of matrix elements that defines a mapping between X and Y. The kth element of W is defined as $w_k = (i, j)_k$. So we have:

$$W = w_1, w_2, ..., w_k, ..., w_K max(m, n)K < m + n - 1$$

The warping path is typically subject to several constraints such as: Boundary conditions; Continuity; Monotonicity. There are exponentially many warping paths that satisfy the above conditions. However, we are only interested in the path that minimizes the warping cost:

Definition Let's suppose we have two vectors $X = (X_1, X_2, ..., X_n)$ and $Y = (Y_1, Y_2, ..., Y_m)$ with numerical observed values. $DTW - Dynamic\ Time\ Warping\ distance$ is defined as the following optimization problem:

$$D(X, Y) = min \sqrt{\sum_{i=1}^{K} w_i} \tag{4}$$

For more information you may refer to (Keogh and Ratanamahatana (2004)).

Definition 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. Since our genes were observed over regular time intervals we may consider them as unique time series of length 11.The goal of using time series analysis in our data is to find the most repeated patterns and classify genes based on their behavior to further understand their characteristics.

Definition A *distance matrix* is defined as a matrix where each entry corresponds to the distance between every pairs of observations (i, j) from the same individuals. In our case where every gene is a time series with observations measured over 11 moments of time, a distance matrix corresponds to the distance between each pair of genes. The distance may be one of the above distances or even a similarity measure when possible. So, in our case a distance matrix will be with dimensions 3908x3908.

In this study we are planning to use two different approaches to classify the genes based on their behavior during the encystation process of Giardia. The first approach is to use cluster analysis, which is a widely-used method for gene classification. This approach involves grouping the genes into different clusters based on their expression levels and other characteristics. We aim to use an ensemble model, which combines multiple algorithms to improve the accuracy and performance of the classification. In our attention is also the principal component analysis (PCA) which we believe at this step may not be as effective for time series clustering, as it does not take into account the ordering of the data and cannot capture time-evolving and time-shifted patterns. The second approach involves using a distance measure to group genes with similar behavior over time. This approach involves calculating the distance between the expression levels of different genes at different time points and grouping genes that have similar patterns of expression. This can provide insights into the relationship between the expression levels of different genes over time and

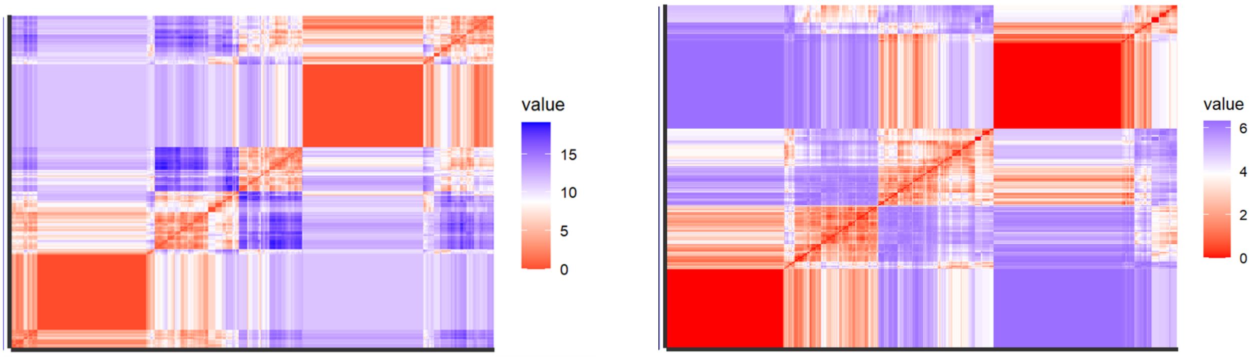


Figure 3: Manhattan (left) and Minkowski (right) distance matrix visualization for standardized observations (Source:E.Gjika)

the mechanisms underlying the encystation process. Both methodologies aim at highlighting the importance of using appropriate methods and approaches for time series data in order to accurately capture and analyze the dynamics of the encystation process. There are many studies and different approach used in clustering gene expression over time :(Hejblum et al. (2015)); (Forgetta et al. (2022))

3 Results and Discussion:

Different classification approaches were used to cluster genes based on their behavior during the encystation process. Involving use of algorithms to identify patterns and trends in the genes' expression levels and grouping them into different classes based on their behavior over the course of the experiment. This analysis provide important insights into the mechanisms underlying the encystation process and how it is coordinated at the molecular level.

Distance matrix visualization and heatmaps were used as graphical representations of data to represent the relative values of the data in a tabular format. In the context of the Giardia study described above, heatmaps have been used to visually cluster the genes that were differentially expressed during the encystation process, allowing this way to identify patterns and trends in the data. This can provide important insights into the mechanisms underlying the encystation process and the role of gene expression in this process. By clustering the genes based on their expression levels, we may be able to identify groups of genes that are co-regulated and likely function together in the encystation process. This can help to shed light on the complex and dynamic nature of the encystation process and how it is coordinated at the molecular level. As we may observe from Figure 3: both distances $Manhattan$ and $Minkowski(special\ case\ of Euclidean)$ gave almost the same visual result, by showing clearly two main clusters of genes which may display the same behavior over time.

Based on the results obtained from the authors there may be significant reasoning that a correlation between moments of time may exist. This could suggest that there is a relationship between the time points and the expression of these genes, such that changes in gene expression at certain time points are associated with changes in other variables, such as the stage of the encystation process or the organism's physiology and metabolism. So, we did a correlation plot and the results are shown in Figure 4 (left).

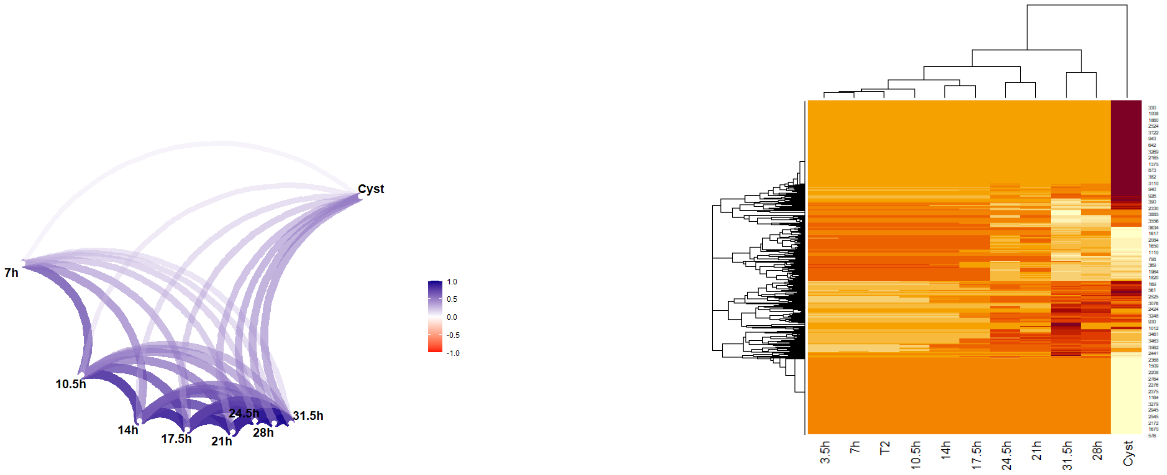


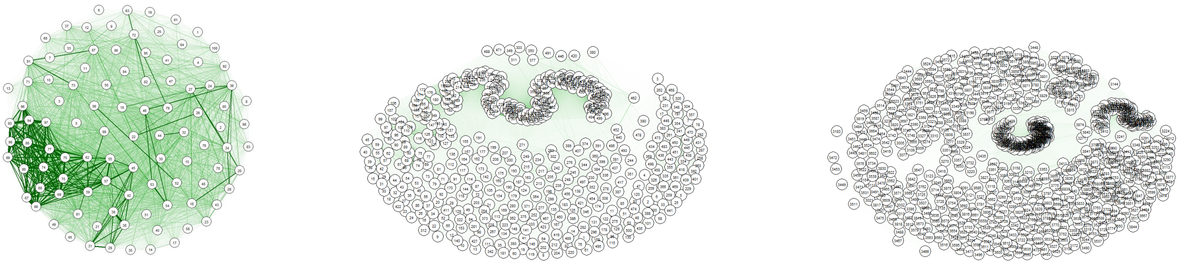
Figure 4: Time correlation graph (left) and heatmap (right) (Source:E.Gjika)

Cyst phase has a slightly low correlation with the other moments of time. this is also observed from the heat map (right) where this time point is clustered clearly away from the other moments of time. The heatmap (standardized observations) could indicate that there are four groups of genes that exhibit similar expression patterns and especially during the encystation process. The specific genes and their expression patterns in these clusters would need to be identified in order to understand the significance of this observation.

Network analysis

In the context of gene expression data, network analysis can be used to identify clusters of genes that are co-regulated and likely function together in a biological process. By constructing a network of genes based on their expression levels and other characteristics, researchers can identify groups of genes that are highly interconnected and have similar patterns of expression. This can provide important insights into the mechanisms underlying the encystation process and how it is coordinated at the molecular level. (Gao et al (2022)) Euclidean distance can be used to calculate the difference in the expression levels between pairs of genes at specific time point by constructing a graph network, where the genes are represented as nodes and the distances between them are represented as edges. The main aim here is to understand how our genes are creating networks based on distances.

We may observe from figure 5 that as the number of genes increases, the difference between them becomes less visible. This behavior could be interpreted as the number of genes included in the analysis increases, the difference in their expression levels becomes less noticeable. This could be because the gene expression levels are relatively similar, or because the distance between the genes in the graph network becomes smaller as more genes are included.



(a) $n = 100$ *genes* (b) $n = 500$ *genes* (c) $n = 900$ *genes*

Figure 5: Graph network connections (Source:E.Gjika)

3.1 Approach 1- using cluster methodologies

$Motifs\ and\ k - means$

It is true that hierarchical clustering can be less effective for visualizing clusters when the number of genes is high. This is because hierarchical clustering produces a tree-like diagram called a dendrogram, which can become difficult to interpret when there are many points in the dataset.

In contrast, k-means clustering is a clustering algorithm that aims to partition a dataset into a predetermined number of clusters by finding a set of points that can serve as cluster centers. This approach can be more effective for visualizing clusters in a high-dimensional dataset because it produces a clear, concise visualization of the clusters. Clustering gene expression data can be a powerful way to identify groups of genes that exhibit similar behavior over time. One common approach for clustering gene expression data is to use the k-means algorithm, which is a clustering method that aims to partition a dataset into a predetermined number of clusters by finding a set of points that can serve as cluster centers.

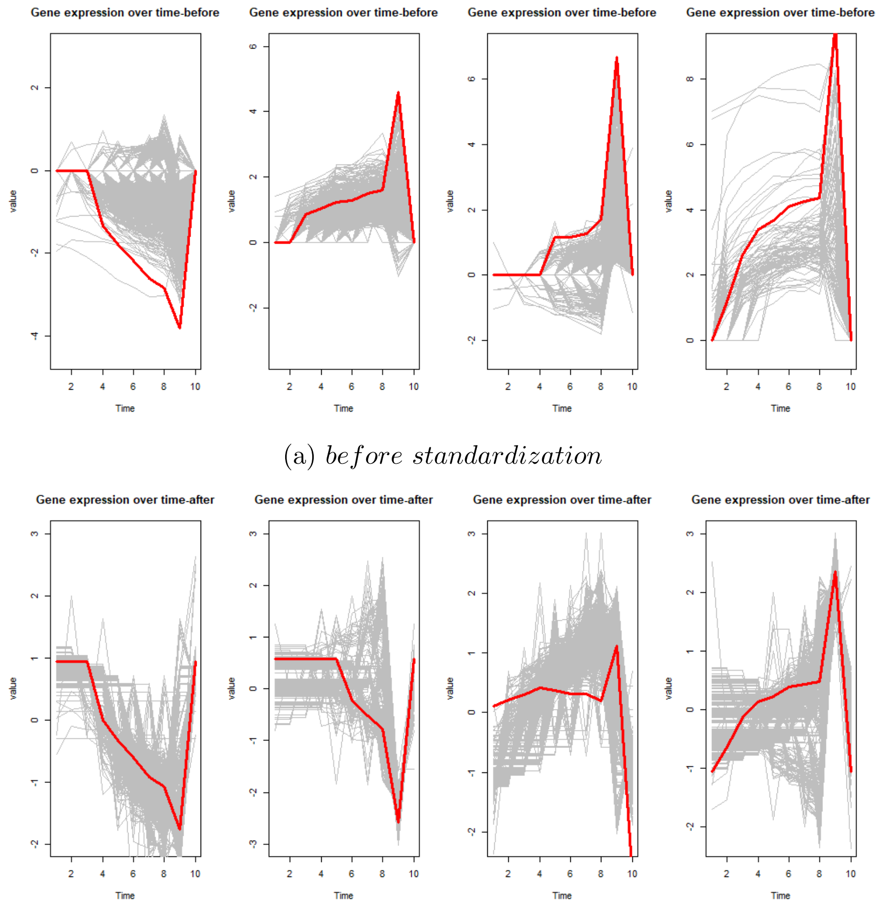
In the case of gene expression data, we can use k-means to identify groups of genes that exhibit similar patterns of expression over time. By organizing the genes into these clusters, we can gain insights into the underlying biology of the genes, as well as their potential role in the development of various diseases.

In our data we may identify four clusters of genes with similar behavior over time. Each of these clusters may represent a group of genes that are involved in a specific biological process, or that have a shared function within the cell.

By analyzing these clusters in more detail, we can gain insights into the mechanisms underlying gene expression and the potential roles of the genes within each cluster. This can be particularly useful for identifying potential targets for the development of new drugs or therapeutic approaches.

Furthermore, by comparing the gene expression patterns in different clusters, we can determine whether there are any commonalities among the genes in each cluster, which could provide valuable insights into the underlying biological processes that are at work. Ultimately, the use of k-means to cluster gene expression data can provide valuable insights into the mechanisms underlying gene expression and the potential roles of different genes within the cell.

Figure 6 shows a comparison between two situation: gene clusters before standardization and gene clusters after standardization. Standardization can improve the performance of machine learning algorithms in several ways. First, it ensures that all features are on the same scale, which can reduce the risk of bias in the model. Second, it helps to reduce the effects of outliers in the data,



(a) *before standardization*

(b) *after standardization*

Figure 6: Motifs discovered from first approach (using kmeans, k=4) (Source:E.Gjika)

which can have a detrimental effect on the performance of the model.

In the context of motif discovery, standardization can improve the accuracy and similarity of the discovered motifs by ensuring that all features are on the same scale. This can help to remove any biases that may be present in the data and can make it easier for the algorithm to identify patterns and structures in the data.

3.2 Approach 2- using time series

Second approach approach will be using a similarity measure to classify genes with similar behavior over time. The first try here was using an existing package in R *tsmp* Bischoff and Rodriguez (2020). The ‘tsmp‘ library in R is a time series analysis and modeling package. It provides tools and functions for analyzing time series data and building statistical models to forecast future values. Some of the key features of the ‘tsmp‘ library include:

- Functions for analyzing time series data, including tools for decomposing time series into trend, seasonality, and residual components - An extensive selection of time series models, including univariate and multivariate models - Support for model diagnostics and checking assumptions -

Functions for simulating and generating synthetic time series data

Using different arguments we have obtained the below motifs repeated in our time series. The challenge using this package was to create a long time series which seasonal (cyclic/periodicity) was 11 and then obtain different motifs. We were focused in three motifs.

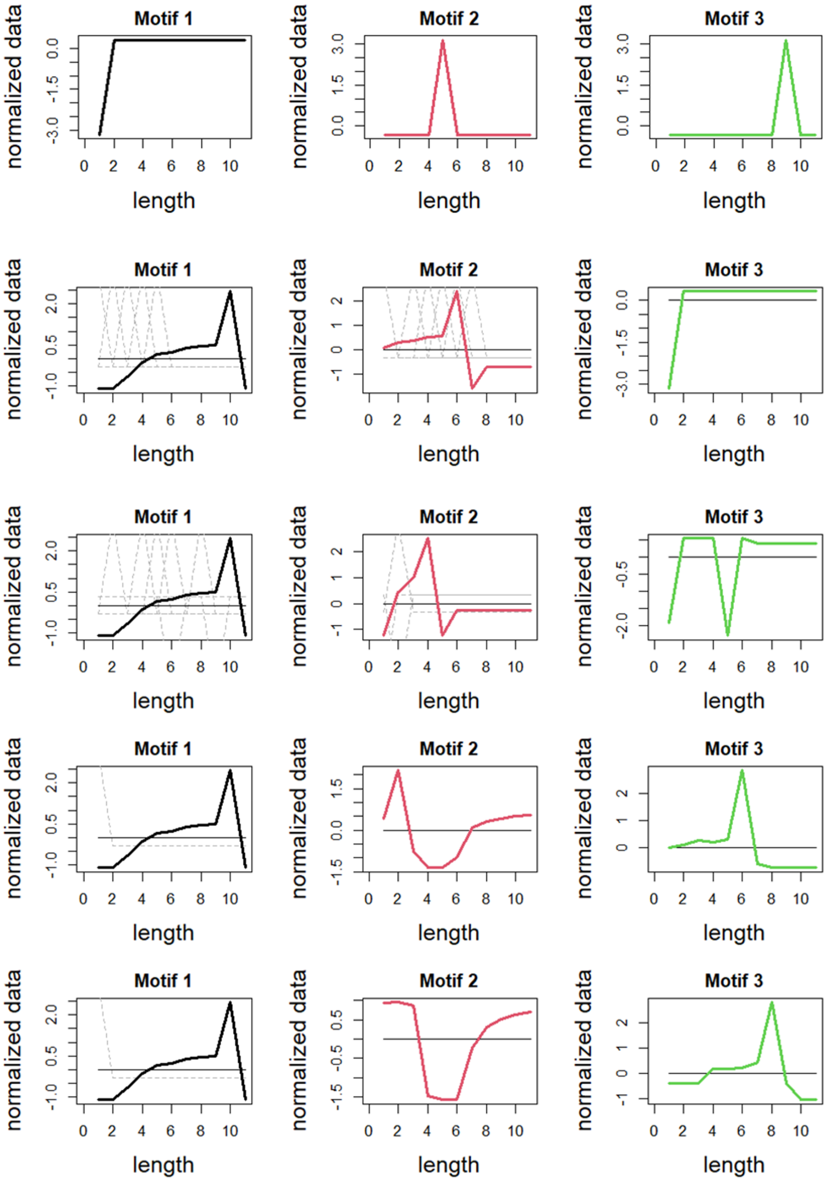


Figure 7: Motifs discovered from *tsmp*-library in R (Source:E.Gjika)

Observing carefully the images obtained for three motifs. Encystation is the process by which a microorganism, such as a protozoan or bacterium, forms a dormant, protective cyst. During encystation, the microorganism undergoes a number of physiological and biochemical changes, including changes in gene expression. Gene expression is the process by which the genetic information in DNA is used to produce proteins or RNA molecules.

From the motifs discovered we may think that there are changes in gene expression during the trophozoite stage in order to support the active metabolism and growth of the microorganism. These changes may include the activation or repression of specific genes, as well as changes in the levels of gene expression. These changes may help the microorganism to carry out essential functions, such as obtaining nutrients and reproducing, and may also help it to adapt to its environment and respond to changes in conditions. It is likely that there are extensive changes in gene expression also during the end of encystation, as the microorganism prepares to enter the dormant cyst state. These changes may involve the activation or repression of specific genes, as well as changes in the levels of gene expression. These changes may be necessary for the microorganism to survive in the dormant state and may also help protect it from environmental stresses.

Overall, changes in gene expression during encystation are an important part of the process by which microorganisms form protective cysts. These changes help the microorganism to survive in harsh environments and also help it to remain viable for long periods of time, until conditions are favorable for it to emerge from the cyst and resume its normal growth and reproduction.

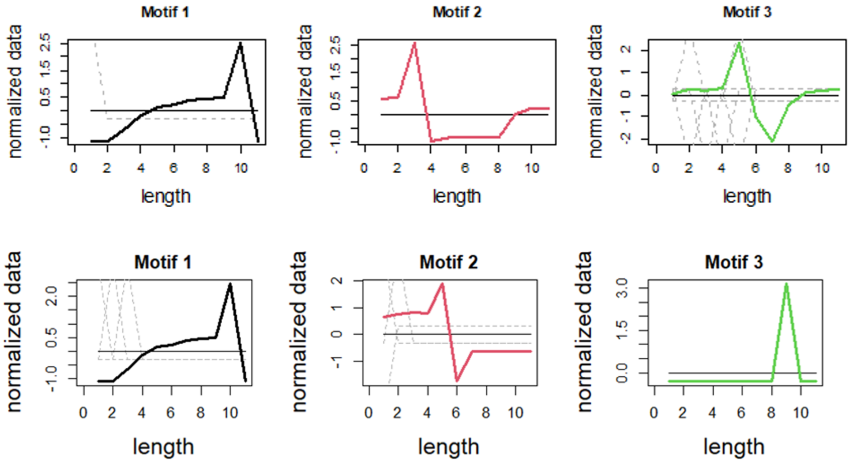
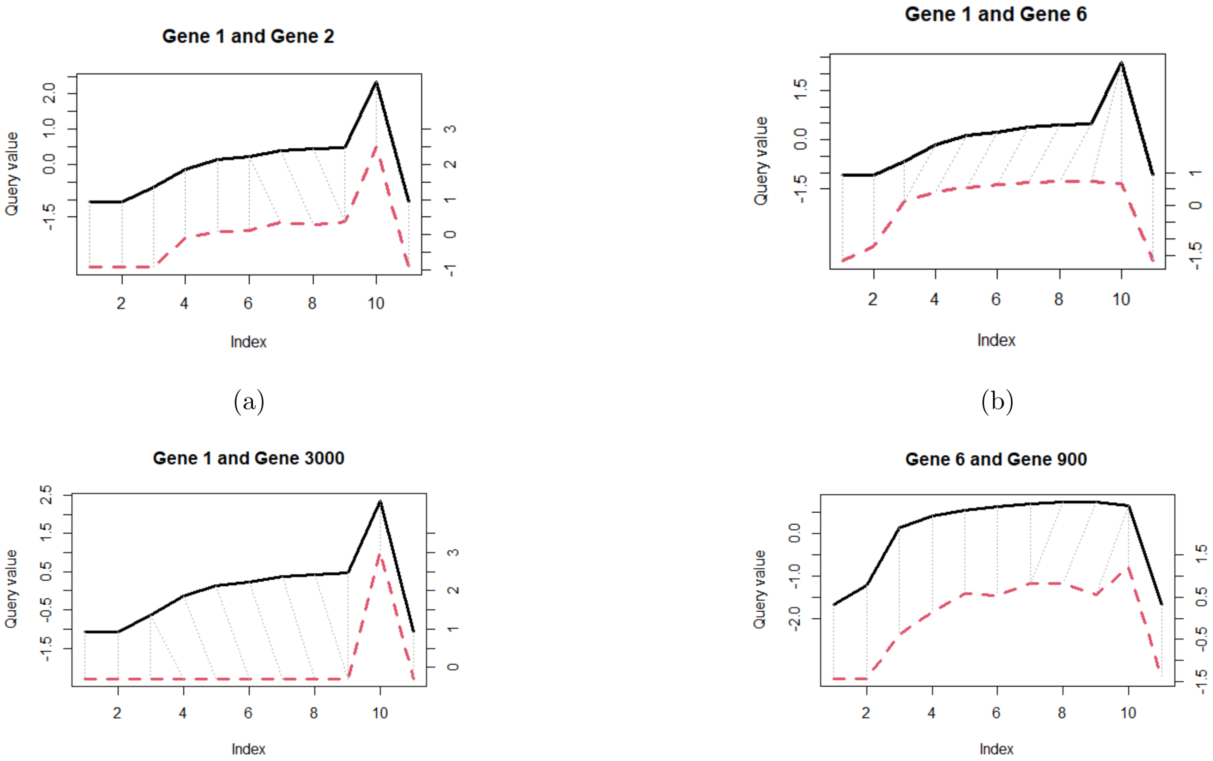


Figure 8: Motifs discovered from *tsmp*-library in R (Source:E.Gjika)

The second try using time series methodologies were considering Dynamic Time Warping (DTW) which is a dynamic programming algorithm. DTW breaks the problem down into smaller subproblems and stores the results in order to avoid recomputing them when needed.

The key advantage of DTW compared to other similarity measures is that it takes into account the fact that the two time series being compared may vary in length and speed. This allows it to compare two time series that may not be perfectly aligned in time, which can be particularly useful when dealing with real-world data.

Overall, DTW is a powerful and versatile algorithm for measuring the similarity between two time series. Its ability to account for differences in length and speed makes it well-suited to a wide range of applications, from analyzing stock prices to identifying patterns in speech signals.



(a)

(b)

(c)

(d)

Figure 9: DTW between two different gene expression of Gardia encyctation (Source:E.Gjika)

4 Conclusion:

Gene expression data is often used to study the underlying biological processes and mechanisms in an organism or tissue. Similarity in gene expression data can provide valuable insights into the relationships between genes and their roles in these processes.

There are several reasons why similarity in gene expression data is important:

1. Similarity can help identify clusters or groups of genes that have similar expression patterns, which can provide insights into the underlying biological processes and mechanisms.

2. Similarity can help identify relationships between genes, such as co-expression or regulatory interactions, which can provide additional insights into the regulation of gene expression and its role in biological processes.

3. Similarity can help identify differences in gene expression between different conditions or samples, which can provide insights into the underlying biological mechanisms involved in the response to these conditions.

To achieve the best classification using gene expression data, it is important to choose an appropriate similarity measure that is able to capture the temporal dynamics of gene expression and handle the various challenges commonly encountered in gene expression data, such as noise, missing values, and different lengths of time series.

One popular similarity measure for gene expression data is DTW (Dynamic Time Warping), which is able to handle time series with different lengths and local warping, and has been shown to have good performance in terms of accuracy and sensitivity compared to other similarity measures. However, care must be taken to choose appropriate parameters and distance measures for DTW, and to consider the computational complexity and limitations of the method. There are several advantages and disadvantages of using similarity methodologies based on time series and DTW (Dynamic Time Warping) in gene expression data instead of classical cluster analysis.

Advantages :

1. Time series and DTW can better capture the temporal dynamics of gene expression data, which is often important in understanding the underlying biological processes.
2. DTW is able to handle time series with different lengths and local warping, which is often the case in gene expression data.
3. DTW is more robust to noise and has better performance in terms of accuracy and sensitivity compared to other similarity measures.
4. DTW is able to capture the temporal relationships between genes, which can provide additional insights into the underlying biological processes.

Disadvantages :

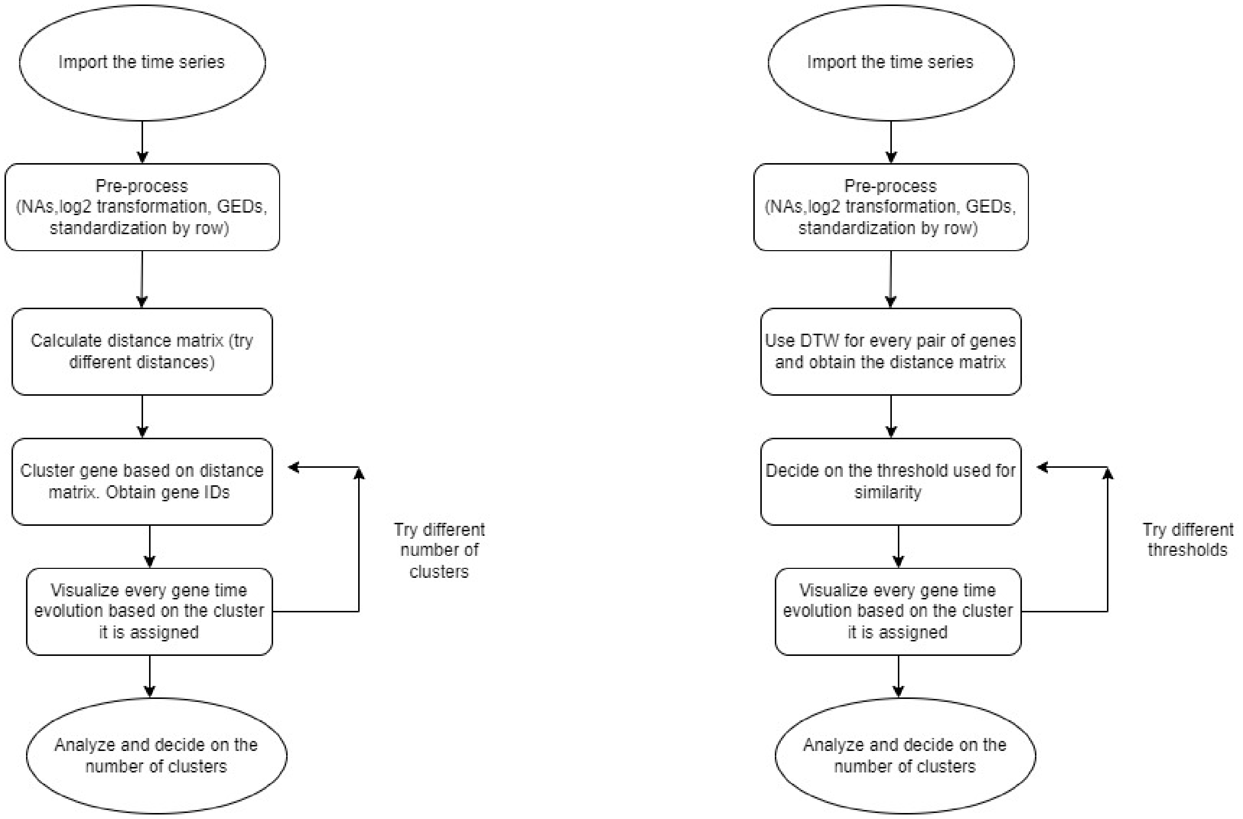
1. DTW is computationally intensive and can be slow to run on large datasets.
2. The alignment path produced by DTW may not have a clear biological interpretation.
3. There is no consensus on the best way to compute the DTW distance and its parameters, which can affect the performance and interpretation of the results.
4. The results of DTW analysis can be sensitive to the choice of distance measure and the number of clusters used for clustering.

Overall, the use of time series and DTW in gene expression data can provide valuable insights into the temporal dynamics of gene expression and help uncover biological relationships that are not readily apparent from classical cluster analysis. However, care must be taken to choose appropriate parameters and distance measures, and to consider the computational complexity and limitations of the method.

Figure 10 shows the flowchart of the two approach used in the study.

The approach involves using advanced computational methods to analyze gene expression data and identify significant motifs. This approach is unique and has the potential to offer important insights into the regulation of gene expression and the underlying mechanisms of cellular processes. By leveraging cutting-edge techniques and methodologies, I hope to shed light on important questions and issues, and to contribute to the existing body of knowledge in the field.

Furthermore, the potential benefits of the research are significant. This work has the potential to lead to new understandings and solutions to gene expression data.


Import the time series
Pre-process (NAs,log2 transformation, GEDs, standardization by row)
Calculate distance matrix (try different distances)
Cluster gene based on distance matrix. Obtain gene IDs
Visualize every gene time evolution based on the cluster it is assigned
Try different number of clusters
Analyze and decide on the number of clusters
Import the time series
Pre-process (NAs,log2 transformation, GEDs, standardization by row)
Use DTW for every pair of genes and obtain the distance matrix
Decide on the threshold used for similarity
Visualize every gene time evolution based on the cluster it is assigned
Try different thresholds
Analyze and decide on the number of clusters

(a)

(b)

Figure 10: Two approaches flowchart used in the study (left: classic cluster approach, left:DTW approach)(Source:E.Gjika)

5 Reference

Data source:
Rojas-López, L.; Krakovka, S.; Einarsson, E.; Ribacke, U.; Xu, F.; Jerlström-Hultqvist, J.; Svärd, S.G. (2021) A Detailed Gene Expression Map of Giardia Encystation. Genes 2021, 12, 1932. https://doi.org/10.3390/genes12121932 ; Zip Folder

Work reference on the topic:
Hejblum BP, Skinner J, Thiébaut R (2015) Time-Course Gene Set Analysis for Longitudinal Gene Expression Data. PLoS Comput Biol 11(6): e1004310.
doi:10.1371/journal.pcbi.1004310

Forgetta V, Li R, Darmond-Zwaig C, et al. Cohort profile: genomic data for 26 622 individuals from the Canadian Longitudinal Study on Aging (CLSA). BMJ Open 2022;12:
https://bmjopen.bmj.com/content/12/3/e059021
Gao S, Chen Y, Wu Z, Kajigaya S, Wang X, Young NS. Time-Varying Gene Expression Network Analysis Reveals Conserved Transition States in Hematopoietic Differentiation between Human and Mouse. Genes. 2022;13(10):1890. https://doi.org/10.3390/genes13101890 ;

Oh, V.-K.S.; Li, R.W. (2021) Temporal Dynamic Methods for Bulk RNA-Seq Time Series Data. Genes 2021, 12, 352. https://doi.org/ 10.3390/genes12030352

The PLOS Computational Biology Staff (2015) Correction: Time-Course Gene Set Analysis for Longitudinal Gene Expression Data. PLoS Comput Biol 11(8)
https://doi.org/10.1371/journal.pcbi.1004446

Forgetta V, Li R, Darmond-Zwaig C, et al. (2022) Cohort profile: genomic data for 26 622 individuals from the Canadian Longitudinal Study on Aging (CLSA). BMJ Open 2022;12:e059021. doi:10.1136/ bmjopen-2021-059021

Gao, S.; Chen, Y.; Wu, Z.; Kajigaya, S.; Wang, X.; Young, N.S.(2022) Time-Varying Gene Expression Network Analysis Reveals Conserved Transition States in Hematopoietic Differentiation between Human and Mouse. Genes 2022, 13, 1890.
https://doi.org/ 10.3390/genes13101890

Batista, G.E.A.P.A., Keogh, E.J., Tataw, O.M. et al. (2014) CID: an efficient complexity-invariant distance for time series. Data Min Knowl Disc 28, 634–669 (2014).
https://doi.org/10.1007/s10618-013-0312-3

R packages

Bischoff F., Pereira Rodrigues P., (2019) tsmp: An R Package for Time Series with Matrix Profile;
https://doi.org/10.48550/arXiv.1904.12626

Keogh, E., Ratanamahatana, C. Exact indexing of dynamic time warping. Knowl Inf Syst 7, 358–386 (2005). doi.org/10.1007/s10115-004-0154-9

Gillespie, C.S., Lei, G., Boys, R.J. et al. Analysing time course microarray data using Bioconductor: a case study using yeast2 Affymetrix arrays. BMC Res Notes 3, 81 (2010).
https://doi.org/10.1186/1756-0500-3-81

UCR Matrix Profile Page:

Timecourse package in R

TcGSA Cran Project

Links related to the topic:

TcGSA github ; Matrix Profile web page; tsmp package in R ; Genes journal MDPI; Image credit

Link for codes in R used in the study:

Eralda Gjika GitHub Repository Genomic Study ;