Phenotype similarities in automatically grouped T2D patients by variation-based clustering of IL-1β gene expression

Lucio José Pantazis[1*], Gustavo Daniel Frechtel[2,3*], Gloria Edith Cerrone[2,4*], Rafael García[1], Andrea Elena Iglesias Molli[2*]

[1]*Centro de Sistemas y Control, Instituto Tecnológico de Buenos Aires (ITBA), Lavardén 315 1437, Ciudad Autónoma de Buenos Aires, Argentina.*
[2]*CONICET-Universidad de Buenos Aires. Instituto de Inmunología, Genética y Metabolismo (INIGEM). Laboratorio de Diabetes y Metabolismo. Avenida Córdoba 2351, Ciudad Autónoma de Buenos Aires, Argentina.*
[3]*Universidad de Buenos Aires. Facultad de Medicina. Departamento de Medicina. Cátedra de Nutrición. Avenida Córdoba 2351, Ciudad Autónoma de Buenos Aires, Argentina.*
[4]*Universidad de Buenos Aires. Facultad de Farmacia y Bioquímica. Departamento de Microbiología, Inmunología, Biotecnología y Genética. Cátedra de Genética. Avenida Córdoba 2351, Ciudad Autónoma de Buenos Aires, Argentina.*
**Corresponding author: lpantazis@itba.edu.ar*

Article Info

Author of correspondence:
Lucio José Pantazis;
E-mail: lpantazis@itba.edu.ar;
Address:
Centro de Sistemas y Control, Instituto Tecnológico de Buenos Aires (ITBA), Lavardén 315 1437, Ciudad Autónoma de Buenos Aires, Argentina

Keywords
Shape-based clustering, Longitudinal data, Gene expression, Type 2 Diabetes, Knowledge Discovery in Databases.

Abstract

Background: Analyzing longitudinal gene expression data is extremely challenging due to limited prior information, high dimensionality, and heterogeneity. Similar difficulties arise in research of multifactorial diseases such as Type 2 Diabetes. Clustering methods can be applied to automatically group similar observations. Common clinical values within the resulting groups suggest potential associations. However, applying traditional clustering methods to gene expression over time fails to capture variations in the response. Therefore, shape-based clustering could be applied to identify patient groups by gene expression variation in a large time metabolic compensatory intervention.
Objectives: To search for clinical grouping patterns between subjects that showed similar structure in the variation of IL-1β gene expression over time.
Methods: A new approach for shape-based clustering by IL-1β expression behavior was applied to a real longitudinal database of Type 2 Diabetes patients. In order to capture correctly variations in the response, we applied traditional clustering methods to slopes between measurements.
Results: In this setting, the application of K-Medoids using the Manhattan distance yielded the best results for the corresponding database. Among the resulting groups, one of the clusters presented significant differences in many key clinical values regarding the metabolic syndrome in comparison to the rest of the data.

Conclusions: The proposed method can be used to group patients according to variation patterns in gene expression (or other applications) and thus, provide clinical insights even when there is no previous knowledge on the subject clinical profile and few timepoints for each individual.

Introduction

Type 2 Diabetes (T2D) is one of the most complex, prevalent and heterogeneous diseases whose etiology involves multiple interactions between genetic predisposing factors and environmental triggers [1]. Inflammation is a relevant component of the pathophysiological alterations that define the progression from metabolic syndrome to T2D [2]. Interleukin-1 beta (IL-1β) is a proinflammatory cytokine related to this clinical inflammation in T2D individuals, and is a well-known immune system modulator secreted by activated macrophages that can affect β-cell function and reduce insulin secretion [3]. Currently, one of the most important lines of research in diabetes is precision medicine, with the principal aim of grouping T2D individuals in different clinical subtypes defined by biomarkers. This group identification could be translated into an emerging approach to disease treatment and prevention that considers individual variability in genes, environment and lifestyle [4]. In this way, Ahlqvist et al. could recently break down T2D subjects into five distinct subgroups, with an improvement prediction of disease progression and outcome by including six variables (age at diagnosis, body mass index [BMI], glycated haemoglobin [HbA1c] , Glutamic Acid Decarboxylase Autoantibodies [GADA], estimation of insulin secretion [HOMA-B] and estimation of insulin-sensitivity [HOMA-IS]) [5]. The measurement of GADA by Ahlqvist et al. assessed the possible diagnosis of LADA (Latent Autoimmune Diabetes in Adults). The role of precision medicine in diabetes management was recognized by the American Diabetes Association (ADA) in collaboration with the European Association for the Study of Diabetes (EASD), which launched the Precision Medicine Initiative in 2018 in Diabetes [6]. The ultimate goal of precision medicine is the personalized provision of medical care, with better recognition of people at high risk for the development of T2D and its complications, and the implementation of personalized treatments at the individual level. In this sense, artificial intelligence could be used to detect clinical subtypes by matching individuals to their combinations of different biomarkers, with techniques such as large-scale prediction models. In the last decades, there have been great developments in methods for gene expression analysis, giving rise to an abundant quantity of data [7]. Since the amount of data grows faster than the ability to understand their implications, methods that allow drawing conclusions from gene expression data can be very useful to narrow this gap. The analysis of gene expression data can be very challenging due to limited prior knowledge on the observed phenomenon, heterogeneity, noise in the data and missing observations in the subject data [8]. Therefore, data mining tools that can provide potential relationships among framework. Longitudinal studies include repeated measures of a variable of interest -usually called a response- in the same subject over time, yielding multiple responses per individual noted as a response trajectory. In this work, the response variable relates to gene expression at a certain time point and the response trajectory describes the evolution of gene expression for a certain individual over time. It must be pointed out that when there are few time points and mistimed measurements, the mathematical tools that can be used are limited. For example, Fourier transformations, the standard procedure for time series, are no longer valid for few measurements. In this setting, the increase or decrease of gene expression between different measurement occasions can be studied [9]. Variables can be very useful for a clinical comprehension Clustering algorithms aim to group observations according to some measure of similarity, or conversely, to separate observations according to dissimilarity. When quantitative variables are involved, the dissimilarity can be based on distance measures. The selection of these features is closely related to the application area and the research objective. Regarding clustering algorithms, K-Means is the most popular method due to the low computational complexity of the algorithm and performance in big data. A variation of this method is the Kernel Based K-means algorithm [10]. The major disadvantage of these algorithms is the susceptibility to outliers and to the random initial group assignment. Another alternative is the K-Medoids algorithm [11]. This algorithm is more robust to outliers and initialization than K-means. Some works proposed clustering subjects according to the corresponding variation of gene expression, suggesting associations between a certain behavior in the gene expression over time with other variables [9], [12]. Many publications used this approach assuming simultaneous measurements to cluster different genes according to the increase or decrease in their expression, defining groups of co-expressed genes, or activating and repressing genes [13-18]. On most occasions, data corresponding to different subjects are not simultaneously collected, and other strategies must be used. Möller et al., applied a clustering algorithm to the transcriptional program of budding yeast, allowing mistimed measurements [19-20]. In a similar way, similarities in the variation of gene expression, can suggest associations with observable clinical features, which can be a starting point for further investigation.

Objectives

The main objective of this work was to cluster subjects in order to find relationships between patterns in Interleukin 1-beta (IL-1β) variation and clinical metabolic variables from a database of T2D patients. Also, to focus potential associations with obesity and metabolic syndrome as central clinical phenotypes. In this article, we perform a new analysis of data from a cohort of patients previously studied by our research group [21].

Materials and methods

1. Prospective controlled study database

The database used for the development of the clustering algorithm included the results of a prospective controlled study conducted in patients with newly diagnosed T2D and hyperglycemia (HbA1c > 8%), and after 6 and 12 months of treatment to achieve metabolic remission (HbA1c < 7%). The treatment was personalized: each participant received the first-line pharmacological treatment, and in all cases lifestyle changes were included through diet and physical exercise. Detailed information on this population can be found in our previously published manuscript [21]. It was the first follow-up study that evaluated IL-1β mRNA expression in hyperglycemic people with T2D after glycemic normalization treatment.

The study was conducted in a group of 30 adults (23.33% were female subjects and 76.67% male subjects) with a median age of 46 years (IQR 18.75 years) recruited from the Diabetes Care Unit. All procedures performed in the study were in accordance with the ethical standards of the institutional research committee, the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The study was approved by the Ethics Committees of the Hospital de Clínicas "José de San Martín" from Ciudad Autónoma de Buenos Aires and all the participants gave their written informed consent. An anonymized database for pre and post intervention (6 and 12 months) instances was constructed for the data mining study. All individuals informed their age and gender and anthropometric measurements (height, weight, and waist circumference), BMI and systolic and diastolic blood pressure (SBP and DBP, respectively) were determined by standardized protocols. Venous blood samples were drawn of every individual, high-density lipoprotein cholesterol (HDL-c), triglycerides (TG), fasting blood glucose (FBG) and HbA1c were measured in serum using standardized procedures [21]. Low-density lipoprotein cholesterol (LDL-C) was calculated by the Friedewald equation. Blood anticoagulated with EDTA K2 was used for mRNA extraction and IL-1β mRNA expression analysis.

2. Notation

The subjects considered in the study required the same number of repeated measures over time (noted as variable t at 0, 6 and 12 months from intervention). The different subjects were grouped according to the variation over time of the IL-1β gene expression (noted as variable r). With this notation, for example, t(i,j) and r(i,j) represent the time point and the gene expression at measurement number j for subject i, respectively.

3. Clustering methods

We used hard partitioning methods for quantitative features. Clustering applications was performed after three sequential definitions:

A) set of variables to be considered;
B) distance between different variable observations defined in item A;
C) clustering algorithm that groups observations defined in item A according to the distance function defined in item B.

A) The clustering objective was set on grouping subjects according to the increase or decrease of gene expression, therefore the algorithm considered two subjects as similar if the corresponding slopes between time points are similar. For each subject i, the variation of gene expression r between time points j−1 and j is given by the following slope value:

$$m(i,j)=\frac{(r(i,j)-r(i,j-1))}{(t(i,j)-t(i,j-1))}$$

Therefore, if each subject i has a set of J repeated measures noted r(i), the same subject has a corresponding set of slopes m(i) with J-1 values. These sets of slopes will be noted as slope vectors. Thus, the slope vectors m(i) were used instead of using the response vectors r(i) for each subject i. Hence, the automatic grouping relied on the distance between slopes.

B) Whenever it was applicable, two distance functions were considered:

• the Euclidean distance [22], that adds the squared differences of slopes and applies a square root to the results. For example, for two subjects i and k the distance is computed as:

$$dE\ (m(i),m(k))=\sqrt{(m(i,1)-m(k,1))^2+\ldots+(m(i,J-1)-m(k,J-1))^2}$$

• the Manhattan distance [23], that adds the absolute values of the slope differences. For example, for two subjects i and k the distance is calculated as follows:

$$d_M\ (m(i),m(k))=|m(i,1)-m(k,1)|+\ldots+|m(i,J-1)-m(k,J-1)|$$

C) Regarding the clustering methods, three alternatives were applied

• K-Means (based on the Euclidean distance)
• Gaussian Kernel based K-Means (based on the Euclidean distance)
• K-Medoids (based on the Manhattan distance)

These clustering methods are used to group individuals according to their corresponding set of slopes m(i). Regarding the Kernel function required for Kernel based K-Means, a Gaussian kernel function was applied [10]. It must be pointed out that the K-Means based algorithms apply exclusively the Euclidean distance, whereas the K-Medoids algorithm allows the use of other distance functions, such as the Manhattan distance. The algorithms were applied using standard commands of R software. Details on the clustering algorithms are available in Hastie et al. [11]. In the sequel we will refer to clustering algorithms applied to the individuals' set of slopes as shape-based. For example, K-Medoids applied to the slope vectors m(i) can be referred as Shape-Based K-Medoids. More details on these procedures can be found in Appendix C.

4. Statistical Inference

Once the data was grouped in clusters, statistical tests were applied to the anthropometric and metabolic variables of the

database, searching for group differences in BMI, HDL-c, TG, LDL-c, FBG, HbA1c, Waist circumference, Age and number of Metabolic Syndrome components [ncMS], according to the Adult Treatment Panel III (ATPIII) guidelines [24]. To assess the statistical significance of differences between and within groups we performed non-parametric tests due to the small sample size and unverifiable assumptions. Kruskal-Wallis test was performed to assess the differences between groups [25], and paired Wilcoxon test to assess the differences within groups [26].

Results

The database was analyzed and subjects with at least one missing response value were excluded, due to the impossibility to attain a slope set comparable with other subjects. After removal, a total of 26 individuals remained for further research. The responses were scaled prior to partitioning, thus, the mean gene expression was subtracted and the result was divided by the corresponding standard deviation [27]. Figure 1 shows the different groups resulting from the applied clustering algorithms. The partitions of the algorithms involving K-means (upper [a] and central [b] panel of Figure 1), result in groups that are likely to mix stable and highly variable gene expression trajectories. This effect can be explained by the lack of robustness of the K-means algorithm. Inspecting the results, K-Medoids clustering (lower panel [c] of Figure 1) is preferred in this application based on the following observations: subjects in Cluster 1 had an initial decrease and a posterior increase, subjects in Cluster 2 showed an initial increase and a posterior decrease, whereas subjects in Cluster 3 had a stable level of IL-1β expression throughout the study, with small increases or decreases over time. Therefore, in the following, the results of the K-Medoids will be shown since the requirements of variation similarities are met. Furthermore, although all clustering methods are subject to randomness, the K-Medoids algorithm showed such robustness that running several times the procedure yielded the same partition. For the K-Medoids algorithm, it is worth mentioning that there was a subject in Cluster 2 whose gene expression increased in both time intervals and has been classified in this group due to the initial increase, which is not present in other clusters, and therefore, the algorithm located the subject in the most similar group. This subject could be morphologically seen as an outlier, and perhaps should have been classified in a separate group. However, a single subject cluster does not allow a correct between-group comparison. Given this clustering, a subsequent analysis was performed in the remaining variables of the database. The main results are given in Table 1. We found significant differences across groups in waist circumference, BMI, HDL-c and TG; and a tendency for LDL-c; but we did not find significant differences in FBG and HbA1c (Table 1). This similarity across groups is explained by the main objective of the original design of the study in order to follow up on the T2D individuals: to attain a decrease in HbA1c levels for all the participants. Also, since the Kruskal-Wallis detects differences between groups, further inspection of the values of most variables suggest that this difference is mainly observed in subjects from Cluster 1. Table 1 shows that subjects in Cluster 1 presented a decrease in LDL-c, TG and increase in HDL-c over time, whereas these values were stable for other clusters. Also, BMI and waist circumference values for subjects in Cluster 1 were smaller compared to those of the other clusters, also suggesting healthier features for Cluster 1. In addition, the Wilcoxon paired test was applied to all variables comparing the values at the start and the end of the study. The Wilcoxon test was not performed in ncMS and Age since the values do not vary over time. The lowest p-values corresponded to Cluster 1, suggesting greater differences in key variables for subjects in this group. Even if statistical significance was not achieved, the p-value is close to 10%, which represents considerable differences in the variables, given the small number of subjects and that non-parametric tests generally provide less statistical power. In addition, the p-values for Cluster 1 are considerably lower than the values corresponding to other clusters, reinforcing the observable difference between the evolutions of people from different clusters. Although we found differences in age, none of the variables analyzed showed a significant association with age (data not shown).

5. Discussion

In the current application, the K-Medoids clustering method using the Manhattan distance applied to the slopes attained the best results concerning the main objective, which was grouping subjects according to the variation in the response of IL-1β expression and showing differential behaviour in clinical variables. The other clustering algorithms considered in our work ([11]), when applied to the slopes yielded heterogeneous groups and therefore, did not meet the desired qualities for such clustering. Similar results are shown when applied to another controlled database in Appendix B. The use of the slopes as the key features of the grouping, allows to generalize previous proposals [20]. In this new framework, any traditional clustering method can be applied to group subjects according to variations in the response. Unlike the application of clustering algorithms in the original data r(i), small distances between the slope vectors m(i) provided similar characteristics in the variation of gene expression. Therefore, the use of the slopes expands the already vast world of clustering methods since these algorithms can be applied in both settings, but yielding different results. More details in Appendix A. The clustering yielded three distinct groups, evidently differentiable when clinically and biochemically compared in Table I. There were significant differences in waist circumference and BMI between the different clusters, so it would also be necessary to analyze the contribution of obesity in the expression of IL-1β that allowed these groups to be separated. Intra-cluster analysis showed that in Cluster 1, although the proposed metabolic compensation goal was reached, the decrease in FPG and HbA1c did not reach

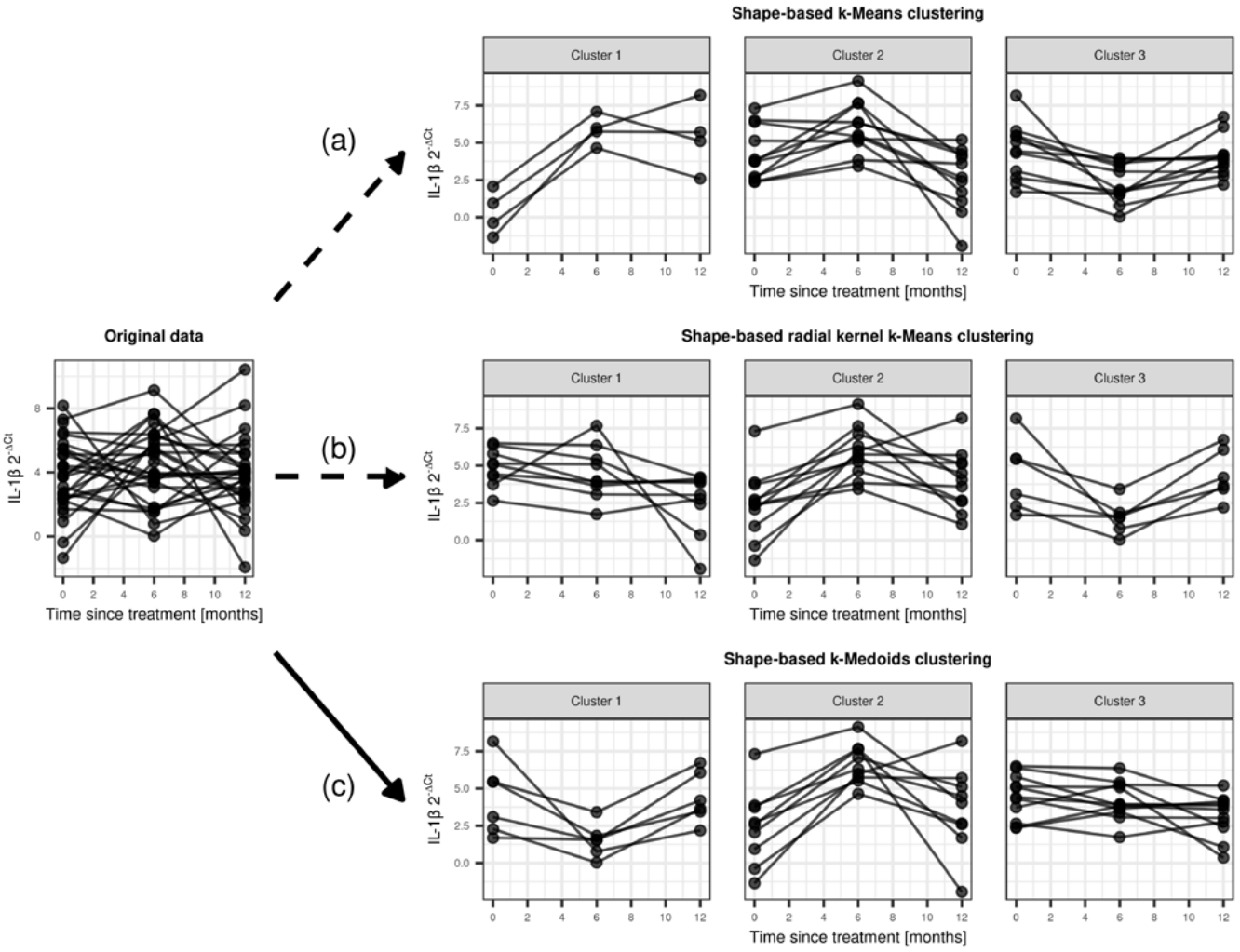

Figure 1: IL-1β (2-$^{\Delta Ct}$) expression over time, grouped according to the slopes between time points using the three clustering algorithms described in Section 3.3: (a) K-Means (Upper panel), (b) Kernel K-Means (Center panel) and (c) K-Medoids (Lower panel).

Table 1: Observed differences in quantitative variables of the dataset, separated by time measurement (at 0, 6 and 12 months). The waist circumference results at 6 months were omitted due to a low proportion of observed data. m: median; IQR: interquartile range; BMI: body mass index; HDL-c: high-density lipoprotein cholesterol; LDL-c:low-density lipoprotein cholesterol; TG: triglycerides; HbA1c: glycated haemoglobin; FBG: fasting blood glucose; ncMS: number of Metabolic Syndrome components.

Variable	Time	Cluster 1 m (IQR)	Cluster 2 m (IQR)	Cluster 3 m (IQR)	P-value (Kruskal-Wallis)
Waist circumference	**0 mo**	100 (92-102)	104 (100-109)	112 (100-117)	
(cm)	**12 mo**	100 (97-103)	106 (101-114)	111 (98-119)	*0.0149*
p-value (Wilcoxon)	*0/12 mo*	*0.8922*	*1.0000*	*0.9056*	
	0 mo	31.11 (29.22-31.42)	32.91 (31.23-38.02)	34.02 (31.60-37.92)	
BMI (kg/m²)	**6 mo**	29.68 (29.18-30.11)	33.57 (29.56-38.22)	32.50 (31.65-35.75)	*0.0106*
	12 mo	30.48 (29.91-30.85)	33.28 (30.35-37.90)	32.80 (28.19-33.85)	
p-value (Wilcoxon)	*0/12 mo*	*0.0544*	*0.0852*	*0.1358*	
	0 mo	1.10 (1.01-1.31)	1.09 (0.83-1.16)	1.01 (0.91-1.03)	
HDL-c (mmol/L)	**6 mo**	1.22 (1.00-1.47)	1.06 (0.92-1.11)	1.05 (0.94-1.14)	*0.0470*
	12 mo	1.32 (1.34-1.45)	1.11 (0.98-1.40)	1.09 (0.98-1.11)	
p-value (Wilcoxon)	*0/12 mo*	*0.0544*	*0.0852*	*0.1358*	
	0 mo	3.22 (2.97-3.30)	3.00 (2.22-3.08)	3.29 (2.90-3.75)	
LDL-c (mmol/L)	**6 mo**	3.19 (2.87-3.44)	2.56 (2.37-2.97)	3.11 (2.82-4.40)	*0.0718*
	12 mo	2.28 (2.22-2.38)	2.72 (2.57-3.09)	3.13 (2.81-3.60)	
p-value (Wilcoxon)	*0/12 mo*	*0.1250*	*0.9453*	*0.8125*	
	0 mo	1.46 (1.27-1.69)	2.06 (1.45-2.74)	1.51 (1.32-2.01)	
TG (mmol/L)	**6 mo**	1.64 (0.97-1.88)	2.42 (2.09-3.20)	2.27 (1.86-2.94)	*0.0047*
	12 mo	0.93 (0.90-0.99)	2.19 (1.67-2.72)	1.47 (1.32-2.75)	
p-value (Wilcoxon)	*0/12 mo*	*0.1250*	*1.0000*	*0.7597*	
	0 mo	8.6 (8.0-10.1)	9.5 (9.0-10.8)	8.1 (7.9-11.2)	
HbA_{1c} (%)	**6 mo**	6.2 (6.1-6.4)	6.4 (5.9-6.9)	6.7 (5.8-7.2)	*0.6652*
	12 mo	5.9 (5.7-6.1)	6.1 (5.6-6.8)	6.2 (5.9-7.0)	
p-value (Wilcoxon)	*0/12 mo*	*0.0625*	*0.0039*	*0.0029*	
	0 mo	8.69 (7.41-15.17)	8.16 (7.38-15.01)	8.77 (7.33-12.10)	
FBG (mmol/L)	**6 mo**	5.91 (5.76-6.52)	5.94 (5.27-6.97)	6.33 (5.83-7.89)	*0.8086*
	12 mo	5.99 (5.83-6.22)	6.33 (5.61-6.66)	6.49 (6.27-7.44)	
p-value (Wilcoxon)	*0/12 mo*	*0.0625*	*0.0078*	*0.0322*	
ncMS		3 (2-4)	4 (3-4)	4 (3-5)	*0.05907*
Age (Years)		60 (57-62)	42 (39-52)	46 (40-58)	*0.00423*

statistical significance. Also, a decrease trend in BMI and metabolic improvements in HDL-c values were observed. In Clusters 2 and 3, the compensation goal was reached as shown by a significant decrease in HbA1c and FBG. In Cluster 2 we also found a downward trend in BMI and HDL-c; but there were no anthropometric or lipid variations in Cluster 3. These results demonstrated that Cluster 3 showed the worst metabolic profile. In subsequent studies, it would be interesting to evaluate variables related to cardiovascular risk. Usually, non-parametric tests are less powerful (prone to discard real differences as non-significant) and the p-values can also be affected by the small sample size [25]. Consequently, the standard significance level of 5% can be too restrictive for this particular application of the statistical tests and p-values which are higher but close to 5% were considered for analysis. However, the strength of the obtained results is enforced by the large time changes considering nutrition and physical individual habits, and also by the time-varying nature of the system under study. As future work, it would be necessary to analyze a larger number of individuals to improve the individualized model and to reinforce our conclusions. Most clinical applications of gene clustering algorithms, which can be phenotype-based or gene-based, do not consider the longitudinal evolutions of gene expression. To the best of our knowledge, this approach has not yet been addressed as a clinical application in the literature. In the work of Pearson et al. [28], the consideration of longitudinal evolution was focused on phenotype follow-up, rather than gene expression and our work considered both gene expression and phenotype over time. Further investigation could profit from the use of all these perspectives to improve algorithm performance. Furthermore, works of clinical application that considered the longitudinal evolution of gene expression used supervised learning algorithms, in which the outcome variable was known and used for further predictions [29-33]. The methodology presented in this work involves unsupervised learning and can be applied when this prior knowledge is absent or limited, and new associations are required. Also, since most available gene expression data comes from countries with strong European ancestry, further research could provide data from other countries that can enrich precision medicine, based on more diverse data sources [34]. Our work used hard partitioning to automatically group individuals from the study. Many other works focused on the use of soft clustering, which is preferred for big data [35]. However, in small studies like ours, with patients undergoing large time treatments, the groups should be well-defined in order to achieve an adequate between-group comparison. A possible extension of this work could be the application of soft clustering to the slopes in longitudinal studies with a great number of subjects, allowing the determination of larger groups according to a strong association. One of the advantages of the proposed data mining procedure is that it does not require time measurements to be equal among all individuals, which is a frequent imposition for similar algorithms. However, in this study, the measurements were taken with the same protocol for every subject and do not differ with great impact in the calculations, and the algorithm easily adapts to these situations. Furthermore, the algorithm is not restricted to gene expression and performs well in other applications, or in cases in which other methods are not recommended, with few time points in which there is no prior knowledge regarding the observed phenomenon, which is a frequent issue in case studies observed in clinical investigation. Also, it is important to remark that this lack of prior knowledge allows us to search for associations between variables that are not previously thought to be linked. However, it must be pointed out that any prior knowledge regarding the application can be used to improve the algorithm, allowing the selection of specific distances between slopes. Since the presented database does not have a massive number of observations, the computational cost of K-Medoids was a drawback without major consequences. However, in other types of databases as massive databases, K-Means or Kernel-based K-means can be a better option. Another issue worth mentioning is that the proposed method is analytical and should not be used as a statistical inference tool. Any result obtained with the method should be further tested in a controlled experiment with a bigger sample size in order to attain satisfactory and pertinent inferences.

Conclusions

Our study showed that clustering individuals according to the variation in gene expression enabled us to find important clinical features that could allow the identification of differentially grouped metabolic behaviors not attained by other data analysis. With further studies, this could be translated into clinical improvement management of each individual considering the group assignment. The achieved results show that the proposed approach can significantly improve predictive performance and is effective when other established methods are not recommended due to the nature of the data, such as small sample sizes, few timepoints, heterogeneity and abrupt changes in gene expression for different timepoints. T2D is a complex and heterogeneous disease. Therefore, identifying clusters with similar clinical phenotype, will allow health professionals to evaluate increased risk, assess clinical evolutions and apply specific and personalized treatment to these groups of individuals. Precision medicine can improve the quality of life of people with T2D and help them improve glycemic control, prevent complications and provide a better quality of life.

Construction of the slope space

Selecting a simmilarity measure for a specific problem is a challenging task, due to the vast number of choices. Since this work is focused on grouping subjects according to the increase or decrease of a certain response variable, the euclidean distance between vectors may fail to recognize differences in the variation. A clear example of this problem is shown in Figure 1:

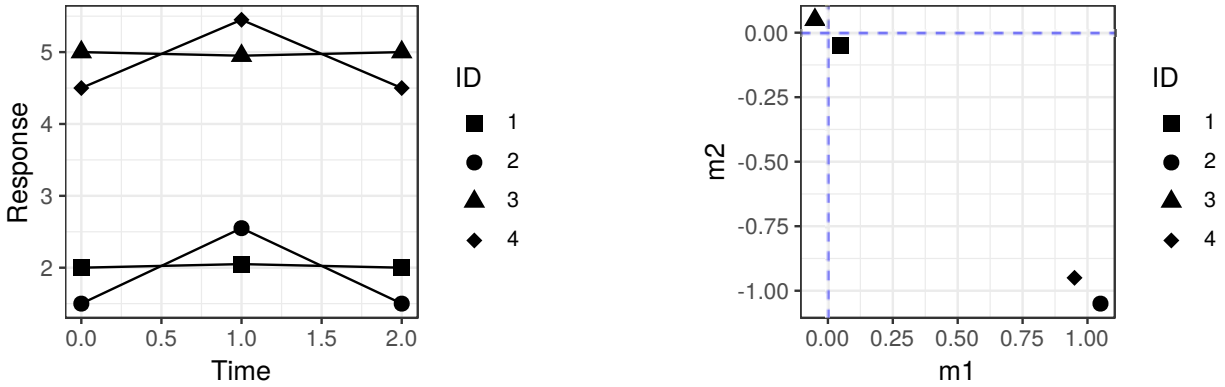

Figure 1. (a) Simulated response trajectories of 4 subjects at 3 time points. (b) Representation of these trajectories en the slope space.

In Figure 1 (a), the responses for subjects 1 and 2 are close to each other at different times, yielding a small euclidean distance between both vectors r_1 and r_2. The same feature can be observed between subjects 3 and 4. Conversely, the euclidean distance, does not consider subjects 1 and 3 as similar, as well as subjects 2 and 4. However, based on the variation of the response, the proposed distance should qualify subjects 1 and 3 as similar, the same should occur between subjects 2 and 4.

To address this issue, a natural distance is considering that to subjects are similar if the slopes between timepoints are similar. For example, according to the previous notations, for subject i, the variation of the response variable r between timepoint j and $j+1$ is given by the following slope value:

$$m_{i,j} = \frac{r_{i,j} - r_{i,j-1}}{t_{i,j} - t_{i,j-1}} \tag{1}$$

Under this construction, negative and positive slopes will correspond to decreases and increases in the response variable, respectively. On this basis, it may seem natural to group individuals according to the slope sign in the same instance. However, this classification omits the magnitude of the slope. If two subjects have slopes of different signs, but similar absolute value, these response trajectories reflect certain stability, and thus, are not qualitatively different.

Figure 1 (b) shows clearly that the slope space meets the goal of reducing the distance between subjects 1 and 3, as well as subjects 2 and 4. Furthermore, the figure shows how dividing observations by quadrants fails to assess the qualitative features of a trajectory, since subjects 1 and 3 belong to different quadrants but both correspond to stable trajectories in time and should belong to the same group. Grouping observations according to the distance in the slope space achieve this feature.

The resulting slope space yields a new set of quantitative values per individual, and the clustering methods detailed in the manuscript can be applied in this setting. For example, response trajectories $r_i = (r_{i,0}, r_{i,1}, \cdots, r_{i,J})$ and individual timepoints $t_i = (t_{i,0}, t_{i,1}, \cdots, t_{i,J})$ have $J+1$ real valued coordinates, whereas $m_i = (m_{i,1}, m_{i,2}, \cdots, m_{i,J}) \in \mathbb{R}^J$. Therefore, traditional distances for numeric vectors can be applied to the vector of slopes. For example, the distance between subject i and i' (with responses $r_{i,j}$ and

$r_{i',j}$ at times $t_{i,j}$ and $t_{i',j}$, respectively) can be expressed by the euclidean distance between the corresponding slopes $m_{i,j}$ and $m_{i',j}$:

$$||m_i - m_{i'}||_2 = \sqrt{\sum_{j=1}^{J} (m_{i,j} - m_{i',j})^2} = \sqrt{\sum_{j=1}^{J} \left(\frac{r_{i,j} - r_{i,j-1}}{t_{i,j} - t_{i,j-1}} - \frac{r_{i',j} - r_{i',j-1}}{t_{i',j} - t_{i',j-1}} \right)^2} \tag{2}$$

The K-means algorithm is based on the euclidean distance, and thus is frequently used in practice. However, other distances for quantitative vectors can be used, for example, the manhattan distance, based on the absolute value of the differences:

$$||m_i - m_{i'}||_1 = \sum_{j=1}^{J} \left| m_{i,j} - m_{i',j} \right| = \sum_{j=1}^{J} \left| \frac{r_{i,j} - r_{i,j-1}}{t_{i,j} - t_{i,j-1}} - \frac{r_{i',j} - r_{i',j-1}}{t_{i',j} - t_{i',j-1}} \right| \tag{3}$$

This distance is less susceptible to outliers, and can be used in both Hierarchal and K-Medoids algorithms, increasing the number of optional metrics beyond the Euclidean distance.

Unlike the application of these distances in the original data r_i, small distances in the slope space vectors m_i ensure similar characteristics in the response variation. Furthermore, any clustering method applied to the slope space succeeds to capture these features in a greater measure than the same clustering method applied to the trajectories. Therefore, the use of the slope space expands the already vast world of clustering methods since it can be applied in both settings, but yielding different results.

Figure 1: Appendix A

1. Growth Mixture Models

A different approach is to propose a model for the temporal evolution of the responses and a specified number K of unknown classes (the procedure is also called Latent Class Mixed Models [2],[1]). The model is usually polynomial with mixed effects for each group:

$$\begin{array}{c} r_{i,j}^{(k)} = \sum_{h=0}^{H} \alpha_h^{(k)} \cdot t_{i,j}^h + \varepsilon_{i,j} \\ \alpha_h^{(k)} = \beta_h^{(k)} + \gamma_{h,i}^{(k)} \end{array} \tag{1}$$

where each β_h is a fixed effect per group and each $\gamma_{h,i}$ is an individual random effect.

Once the model is specified, the parameters are fitted by Maximum Likelihood and a posterior group classification of each observation is performed based on the estimated parameters. This classification yields K groups that can also be seen as a clustering or partition.

This approach has been proven useful for slowly changing trends. However, the model needs to be previously specified (requiring prior information which is not always available) and the model is not always clear, specially for data with small sample sizes. Furthermore, whenever the changes in temporal trends are abrupt, the estimated coefficients are greatly affected by these variations, influencing the entire fitted trajectory.

To avoid this drawback, the trends can be modeled with polynomial splines per group, diminishing the influence of poorly estimated coefficients. However, the timepoints used as knots are not always explicit and must be determined adding a new difficulty.

2. Clustering quality measures

Given an automatic clustering, it is not always clear how to evaluate the quality of the partition. The aforementioned diversity in clustering problems and objectives lead to several different indices, but these measures can be qualified in two groups: internal and external criteria.

Internal criteria are used to evaluate desirable qualities of a clustering, such as high between-cluster variability (or separability) and low within-cluster variability (or homogeneity), without a reference grouping. On the other hand, external criteria require two partitions and do not focus on the properties of the clusterings, they assess the similarity between two different groupings. In this work, only external criteria are addressed since the aim of the experiments is set on identifying a partition given as reference.

2.0.1. *External criteria.* Measuring agreement between clustering partitions is not as simple as matching the number of objects belonging to a certain clusters, mainly because most clustering algorithms have an initial random assignment. For example, running K-means twice on the same data, can lead to the same clustering, but with different labels given as an output.

To overcome this drawback, given a database of n individuals, most external criteria focus on the $\frac{n(n-1)}{2}$ different pairings of the data observations. If two observations x_i and $x_{i'}$ are grouped in the same cluster in one partition, the other partition will agree with this result if x_i and $x_{i'}$ are also in the same cluster, regardless of the cluster labels. Also, if two observations belong to different clusters in one partition, the other partition should also be assigned to different groups.

According to the notation used in the literature [4], given two partitions $\mathcal{P}_1$ and $\mathcal{P}_2$ for a dataset of n observations, and its corresponding $n_P = \frac{n(n-1)}{2}$ pairings, the following numbers are computed:

- yy pairs of observations grouped in the same cluster in both partitions.
- yn pairs of observations grouped in the same cluster in $\mathcal{P}_1$, but not in $\mathcal{P}_2$.
- ny pairs of observations grouped in different clusters in $\mathcal{P}_1$, but not in $\mathcal{P}_2$.
- nn pairs of observations grouped in different clusters in both partitions.

We remark that these notations do not require clusters to have the same label. Furthermore, the number of clusters of each partition can be different. In all cases, the sum of these four numbers add to the number of pairings n_P.

Based on these definitions, there are several criteria that can be applied to compare different partitions. This work relies on the following criteria:

- **Precision (P):** $C_P = \dfrac{yy}{yy + ny}$
- **Recall (RC):** $C_{Rc} = \dfrac{yy}{yy + yn}$
- **Rand (RN):** $C_{Rn} = \dfrac{yy + nn}{n_P}$
- **Jaccard (J):** $C_J = \dfrac{yy}{yy + yn + ny}$
- **Czekanowski-Dice (CD):** $C_{CD} = \dfrac{2yy}{2yy + yn + ny}$
- **Folkes-Mallows (FM):** $C_{FM} = \dfrac{yy}{\sqrt{(yy + yn) \cdot (yy + ny)}}$
- **Kulczynski (K):** $C_K = \dfrac{1}{2} \cdot \left(\dfrac{yy}{yy + ny} + \dfrac{yy}{yy + yn} \right)$
- **Rogers-Tanimoto (RGT):** $C_{RGT} = \dfrac{yy + nn}{yy + nn + 2(yn + ny)}$

It is worth mentioning that all these criteria correspond a higher index with a greater agreement between partitions. Furthermore, note that in the best case scenario, both yn and ny are zero, and all these indices have a maximum value of 1.

These criteria are very useful whenever there is a reference partition and the goal is to assess the agreement of an automatic partition to the reference grouping.

3. Benchmark database

The variation-based clustering algorithms are tested in a longitudinal benchmark database, in which the subjects are naturally grouped. The main goal is to cluster automatically the trajectories according only to the variation in the response, and compare the results to the reference grouping via external criteria.

TLC Data. The Treatment of Lead-Exposed Children (TLC) trial ([3]) is a randomized study that analyses the effects of a drug named succimer in children with similar blood lead levels. These data consist of four repeated measurements of blood lead levels obtained at baseline (or week 0), week 1, week 4, and week 6 on 100 children, randomly assigned to treatment with succimer or placebo.

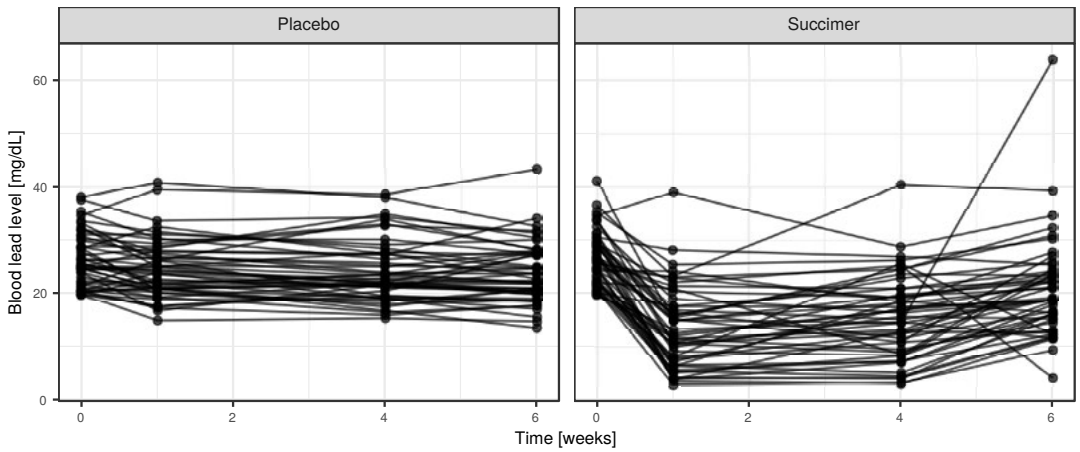

Figure 1. Response trajectories for subjects in the TLC study,

As the response trajectories in Figure 1 show, the blood lead levels are stable in the placebo group and there is a strong decrease in blood lead levels in the succimer group during the first week, but an increase in the remain of the study. Therefore, the slopes in the placebo group are expected to be close to zero.

On the other hand, the slopes in the succimer group are expected to be negative in the first instance, and relatively stable after the first week.

3.1. **Results .** Figure 2 reflects the quality indices for the different clustering methods described in the manuscript, and adding Latent Class Mixture Models applied to the TLC data. Following Figure 1, a linear spline model is considered with a knot in the timepoint corresponding to the first week. Also, Kernel-based K-Means is included with an automatic selection of parameter σ.

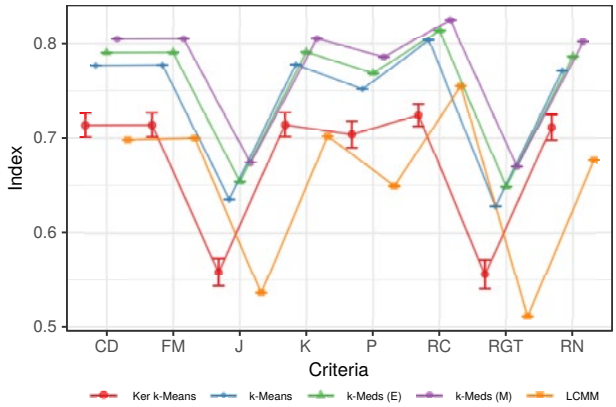

FIGURE 2. Mean indices for the TLC data

The k-Medoids algorithm using the Manhattan distance yields the best performance. Another detail worth mentioning is that even though for each method $M = 100$ repetitions were conducted, the indices remain unaffected and converge to the same partition, except for the Kernel-based k-Means. Therefore, there is a null standard error in almost all cases. The standard error for Kernel-based k-Means is considerable and can be explained by the sensitivity of this method to initialization and outliers.

REFERENCES

[1] NGP Den Teuling, SC Pauws, and ER van den Heuvel. A comparison of methods for clustering longitudinal data with slowly changing trends. *Communications in Statistics-Simulation and Computation*, pages 1–28, 2020.

[2] Daniel P Martin and Timo von Oertzen. Growth mixture models outperform simpler clustering algorithms when detecting longitudinal heterogeneity, even with small sample sizes. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(2):264–275, 2015.

[3] WJ Rogan, RL Bornschein, Jr Chisolm, AI Damokosh, DW Dockery, ME Fay, RL Jones, GG Rhoads, NB Ragan, M Salganik, et al. Safety and efficacy of succimer in toddlers with blood lead levels of 20-44 μg/dl. *Pediatric research*, 48(5):593–599, 2000.

[4] Sławomir T Wierzchoń and Mieczysław A Kłopotek. *Modern algorithms of cluster analysis*. Springer, 2018.

Figure 2: Appendix B

1. Clustering methods

The objective of clustering methods is to group observations (in this case, the response trajectories r_i) according to some measure of similarity, or conversely, separating observations according to dissimilarity. The concept of similarity (or dissimilarity) is vague and can be applied to categorical or quantitative features (and transformed features), for different similarity measures, kernels, and even stablishing degrees of association between an individual and a group, leading to an immense number of options. The selection of these features is deeply related to the application area and the research objective.

This work focuses on crisp partitioning methods for quantitative features. The advantage of grouping quantitative values methods is that they can be based on distance measures for vectors, such as the Euclidean distance.

In the sequel and unless otherwise specified, the observations are noted as $x_i \in \mathbb{R}^m$, with $1 \leq i \leq n$. The clusters corresponding to a certain partition are noted $\mathcal{C}_k$ $(1 \leq k \leq K)$, where $K \leq n$, each with n_k elements, and an average value noted $\mu_k = \frac{1}{n_k} \cdot \sum_{x_i \in \mathcal{C}_k} x_i$. The number n_k of observations in each cluster satisfy $n = \sum_{k=1}^{K} n_k$.

In this work we focus on five clustering methods with corresponding variations: K-Means, Hierarchal clustering (Single, Average and Complete methods) and K-Medoids (using Manhattan and Euclidean distance), Kernel-based k-Means (Radial Kernel) and Latent Class Mixed Models.

Details regarding the used clustering methods are available in Hastie, Tibshirani & Friedman [18] and Wierzchoń & Kłopotek [26].

1.1. **K-Means .**

K-Means is the most popular clustering method, mostly due to the low computational complexity of the algorithm and its performance in big data.

The main goal of the algorithm is to assign objects to clusters in order to minimize the within-group variance of the partition, and thus, based on the euclidean distance between vectors. However, this optimization problem is $\mathcal{NP}$-hard. Therefore, a heuristic approach is implemented iteritavely searching for a local minimum.

The algorithm requires a number of clusters K. In the first step, observations are randomly assigned to K groups. Within each group $1 \leq k \leq K$, the group mean $\mu_k \in \mathbb{R}^m$ is calculated. After this computation, each observation x_i is assinged to the group k, where μ_k is the closest group mean to x_i. The mean computation and further assingment is iterated until the resulting groups remain unchanged.

The major disadvantage to this algorithm is the lack of robustness to outliers of the calculated group mean, and to the initial group assignment. Also, the algorithm requires the number K of clusters and in practice, it can be very difficult to know in advance the number of groups.

1.2. **K-Medoids .**

An alternative to K-Means is the K-Medoids algorithm. This algorithm is more robust to outliers and initialization than K-means, since it relies on observations of the database as group centers, instead of group averages.

However, the robustness comes with a cost of increased computation complexity. Therefore, this algorithm can preform very well when the number of observations are not massive.

The algorithm starts selecting K random observations as cluster representatives, and assigns the remaining $n - K$ observations to the closest center. Once the clusters are assigned, each cluster center is updated to the observation that minimizes the within-group distance, iterating until cluster assignation does not change.

1.3. **Kernel-based K-Means .**

Kernel-based K-Means is a usual alternative when the observations are not linearly separable. In order to increase separability, a non-linear transformation Φ is applied to the data and the same K-Means algorithm described in Section 1.1 to the transformed data.

One of the most popular kernels is the Gaussian Kernel or Radial Basis Function:

$$\mathcal{K}(x_i, x_{i'}) = e^{-\frac{||x_i - x_{i'}||}{2\sigma^2}} \tag{1}$$

where σ is a parameter defined by the user.

The use of kernels flexibilizes the use of K-Means. However, this extension does not modify the previous drawbacks: the algorithm is very sensitive to outliers and initialization. Furthermore, the inclusion of parameter σ adds a new issue: the results can be qualitatively different when the parameter values are modified.

1.4. **Algorithm.** Algorithm 1 describes the pseudocode for the structure of the algorithm. Only subjects with complete responses and time variables necessary to calculate the slope are included in the clustering algorithm. For these remaining subjects, the vector of slopes is attained and a clustering algorithm is performed. It must be pointed out that any distance function or clustering method can be applied at this point in the algorithm, yielding a great versatility for this approach.

Algorithm 1 Shape-based Clustering of Longitudinal data

procedure SHAPE-BASED CLUSTERING(*Data*, *Resp*, *Time*, *J*, *ID*)
  $NetData \leftarrow$ FilterMissingData$(Data, Resp, Time, ID)$
  $IDs \leftarrow$ UniqueID$(Data, ID)$
  $NumIDs \leftarrow$ Length(IDs)
  $i \leftarrow 1$
  $IDsComp \leftarrow \emptyset$
  $m \leftarrow \emptyset$
  loop1:
  while $i \leq NumIDs$ **do**
    $IndData \leftarrow$ SelectID$(Datos, IDs(i))$
    $IndResp \leftarrow$ SelectResponse$(IndData, Resp)$
    $IndTime \leftarrow$ SelectTime$(IndData, Time)$
    if Length$(IndResp) = J + 1$ **and** Length$(IndTime) = J + 1$ **then**
      $IDsComp \leftarrow Append(IDs(i), IDsComp)$
      $m_{Aux} \leftarrow \emptyset$
      $j \leftarrow 1$
      loop2:
      while $j \leq J$ **do**
        $$m_{Aux}(j) \leftarrow \frac{IndResp(j+1) - IndResp(j)}{IndTime(j+1) - IndTime(j)}$$
        $j \leftarrow j + 1$
        go to *loop2*.
      $m \leftarrow RowBind(m, m_{Aux})$
    $i \leftarrow i + 1$.
    go to *loop1*.
  $AssignClusts \leftarrow ClusteringAlgorithm(m)$
  $Results \leftarrow RowBind(IDsComp, AsignClusts)$
  return *Results*

References

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.
[2] Sławomir T Wierzchoń and Mieczysław A Kłopotek. *Modern algorithms of cluster analysis*. Springer, 2018.

Figure 3: Appendix C

References

1. Bowman P, Flanagan SE, Hattersley AT. Future Roadmaps for Precision Medicine Applied to Diabetes: Rising to the Challenge of Heterogeneity. J Diabetes Res. 2018 Nov 27;2018:3061620. doi: 10.1155/2018/3061620. PMID: 30599002; PMCID: PMC6288579.

2. Cruz NG, Sousa LP, Sousa MO, Pietrani NT, Fernandes AP, Gomes KB. The linkage between inflammation and Type 2 diabetes mellitus. Diabetes Res Clin Pract. 2013 Feb;99(2):85-92. doi: 10.1016/j.diabres.2012.09.003. Epub 2012 Dec 14. PMID: 23245808.

3. Nackiewicz D, Dan M, He W, Kim R, Salmi A, Rütti S, Westwell-Roper C, Cunningham A, Speck M, Schuster-Klein C, Guardiola B, Maedler K, Ehses JA. TLR2/6 and TLR4-activated macrophages contribute to islet inflammation and impair beta cell insulin gene expression via IL-1 and IL-6. Diabetologia. 2014 Aug;57(8):1645-54. doi: 10.1007/s00125-014-3249-1. Epub 2014 May 12. PMID: 24816367.

4. Florez JC. Precision Medicine in Diabetes: Is It Time? Diabetes Care. 2016 Jul;39(7):1085-8. doi: 10.2337/dc16-0586. Epub 2016 Jun 11. PMID: 27289125.

5. Ahlqvist E, Storm P, Käräjämäki A, Martinell M, Dorkhan M, Carlsson A, Vikman P, Prasad RB, Aly DM, Almgren P, Wessman Y, Shaat N, Spégel P, Mulder H, Lindholm E, Melander O, Hansson O, Malmqvist U, Lernmark Å, Lahti K, Forsén T, Tuomi T, Rosengren AH, Groop L. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol. 2018 May;6(5):361-369. doi: 10.1016/S2213-8587(18)30051-2. Epub 2018 Mar 5. PMID: 29503172.

6. Nolan JJ, Kahkoska AR, Semnani-Azad Z, Hivert MF, Ji L, Mohan V, Eckel RH, Philipson LH, Rich SS, Gruber C, Franks PW. ADA/EASD Precision Medicine in Diabetes Initiative: An International Perspective and Future Vision for Precision Medicine in Diabetes. Diabetes Care. 2022 Feb 1;45(2):261-266. doi: 10.2337/dc21-2216. PMID: 35050364; PMCID: PMC8914425.

7. Baldi P, Hatfield GW. DNA Microarrays and Gene Expression: From Experiments to Data Analysisand Modeling. Cambridge University Press. 2002.

8.Altiparmak F, Ferhatosmanoglu H, Erdal S, Trost DC. Information mining over heterogeneous and high-dimensional time-series data in clinical trials databases. IEEE Trans Inf Technol Biomed. 2006 Apr;10(2):254-63. doi: 10.1109/titb.2005.859885. PMID: 16617614.

9. Bar-Joseph Z, Gitter A, Simon I. Studying and modelling dynamic biological processes using time-series gene expression data. Nat Rev Genet. 2012 Jul 18;13(8):552-64. doi: 10.1038/nrg3244. PMID: 22805708.

10. Marin D, Tang M, Ayed IB, Boykov Y. Kernel Clustering: Density Biases and Solutions. IEEE Trans Pattern Anal Mach Intell. 2019 Jan;41(1):136-147. doi: 10.1109/TPAMI.2017.2780166. Epub 2017 Dec 6. PMID: 29990278.

11. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media. 2009.

12. Ernst J, Nau GJ, Bar-Joseph Z. Clustering short time series gene expression data. Bioinformatics. 2005 Jun;21 Suppl 1:i159-68. doi: 10.1093/bioinformatics/bti1022. PMID: 15961453.

13. Chira C, Sedano J, Camara M, Prieto C, Villar JR, Corchado E. A cluster merging method for time series microarray with production values. Int J Neural Syst. 2014 Sep;24(6):1450018. doi: 10.1142/S012906571450018X. Epub 2014 Jul 24. PMID: 25081426.

14. Cinar O, Ilk O, Iyigun C. Clustering of short time-course gene expression data with dissimilar replicates. Annals of Operations Research. 2018;263(1-2):405-428.

15. Coffey N, Hinde J, Holian E. Clustering longitudinal profiles using p-splines and mixed effects models applied to time-course gene expression data. Computational Statistics & Data Analysis. 2014;71:14-29.

16. Futschik ME, Carlisle B. Noise-robust soft clustering of gene expression time-course data. J Bioinform Comput Biol. 2005 Aug;3(4):965-88. doi: 10.1142/s0219720005001375. PMID: 16078370.

17. Hestilow TJ, Huang Y. Clustering of gene expression data based on shape similarity. EURASIP J Bioinform Syst Biol. 2009;2009(1):195712. doi: 10.1155/2009/195712. Epub 2009 Apr 23. PMID: 19404484; PMCID: PMC3171421.

18. Son YS, Baek J. A modified correlation coefficient based similarity measure for clustering time-course gene expression data. Pattern Recognition Letters. 2008;29(3):232–242. doi:10.1016/j.patrec.2007.09.015.

19. Chechik G, Oh E, Rando O, Weissman J, Regev A, Koller D. Activity motifs reveal principles of timing in transcriptional control of the yeast metabolic network. Nat Biotechnol. 2008 Nov;26(11):1251-9. doi: 10.1038/nbt.1499. PMID: 18953355; PMCID: PMC2651818.

20. Möller-Levet CS, Klawonn F, Cho KH, Yin H, Wolkenhauer O. Clustering of unevenly sampled gene expression. Fuzzy sets and Systems. 2005;152(1):49-66.

21. Iglesias Molli AE, Bergonzi MF, Spalvieri MP, Linari MA, Frechtel GD, Cerrone GE. Relationship between the IL-1β serum concentration, mRNA levels and rs16944 genotype in the hyperglycemic normalization of T2D patients. Sci Rep. 2020 Jun 19;10(1):9985. doi: 10.1038/s41598-020-66751-x. PMID: 32561825; PMCID: PMC7305205.

22. Wierzchoń ST,Kłopotek MA. Modern algorithms of cluster analysis. Springer, 2018.

23. Yu J, Tian Q, Amores J, Sebe N. Toward Robust Distance Metric Analysis for Similarity Estimation. Computer Vision and Pattern Recognition. In IEEE Computer Society Conference. 2006;1:316-322.

24. Grundy SM, Cleeman JI, Daniels SR, Donato KA, Eckel RH, Franklin BA, Gordon DJ, Krauss RM, Savage PJ, Smith SC Jr, Spertus JA, Costa F; American Heart Association; National Heart, Lung, and Blood Institute. Diagnosis and management of the metabolic syndrome: an American Heart Association/National Heart, Lung, and Blood Institute Scientific Statement. Circulation. 2005 Oct 25;112(17):2735-52. doi: 10.1161/CIRCULATIONAHA.105.169404. Epub 2005 Sep 12. Erratum in: Circulation. 2005 Oct 25;112(17):e297. Erratum in: Circulation. 2005 Oct 25;112(17):e298. PMID: 16157765.

25. Ostertagova E, Ostertag O, Kováč J. Methodology and application of the Kruskal-Wallis test. Applied mechanics and materials. 2014;611115-120.

26. Wilcoxon F. Probability tables for individual comparisons by ranking methods. Biometrics. 1947 Sep;3(3):119-22. PMID: 18903631.

27. Milligan G, Cooper M. A study of standardization of variables in cluster analysis. Journal of Classification. 1988;5(2):181-204.

28. Wesolowska-Andersen A, Brorsson CA, Bizzotto R, Mari A, Tura A, Koivula R, Mahajan A, Vinuela A, Tajes JF, Sharma S, Haid M, Prehn C, Artati A, Hong MG, Musholt PB, Kurbasic A, De Masi F, Tsirigos K, Pedersen HK, Gudmundsdottir V, Thomas CE, Banasik K, Jennison C, Jones A, Kennedy G, Bell J, Thomas L, Frost G, Thomsen H, Allin K, Hansen TH, Vestergaard H, Hansen T, Rutters F, Elders P, t'Hart L, Bonnefond A, Canouil M, Brage S, Kokkola T, Heggie A, McEvoy D, Hattersley A, McDonald T, Teare H, Ridderstrale M, Walker M, Forgie I, Giordano GN, Froguel P, Pavo I, Ruetten H, Pedersen O, Dermitzakis E, Franks PW, Schwenk JM, Adamski J, Pearson E, McCarthy MI, Brunak S; IMI DIRECT Consortium. Four groups of type 2 diabetes contribute to the etiological and clinical heterogeneity in newly diagnosed individuals: An IMI DIRECT study. Cell Rep Med. 2022 Jan 4;3(1):100477. doi: 10.1016/j.xcrm.2021.100477. PMID: 35106505; PMCID: PMC8784706.

29. Baranzini SE, Mousavi P, Rio J, Caillier SJ, Stillman A, Villoslada P, Wyatt MM, Comabella M, Greller LD, Somogyi R, Montalban X, Oksenberg JR. Transcription-based prediction of response to IFNbeta using supervised computational methods. PLoS Biol. 2005 Jan;3(1):e2. doi: 10.1371/journal.pbio.0030002. Epub 2004 Dec 28. PMID: 15630474; PMCID: PMC539058.

30. Borgwardt KM, Vishwanathan SV, Kriegel HP. Class prediction from time series gene expression profiles using dynamical systems kernels. Pac Symp Biocomput. 2006:547-58. PMID: 17094268.

31. Costa IG, Schönhuth A, Hafemeister C, Schliep A. Constrained mixture estimation for analysis and robust classification of clinical time series. Bioinformatics. 2009 Jun 15;25(12):i6-14. doi: 10.1093/bioinformatics/btp222. PMID: 19478017; PMCID: PMC2687976.

32. Calvano SE, Xiao W, Richards DR, Felciano RM, Baker HV, Cho RJ, Chen RO, Brownstein BH, Cobb JP, Tschoeke SK, Miller-Graziano C, Moldawer LL, Mindrinos MN, Davis RW, Tompkins RG, Lowry SF; Inflamm and Host Response to Injury Large Scale Collab. Res. Program. A network-based analysis of systemic inflammation in humans. Nature. 2005 Oct 13;437(7061):1032-7. doi: 10.1038/nature03985. Epub 2005 Aug 31. Erratum in: Nature. 2005 Dec 1;438(7068):696. PMID: 16136080.

33. den Teuling NGP, Pauws SC, van den Heuvel ER. A comparison of methods for clustering longitudinal data with slowly changing trends. Communications in Statistics: Simulation and Computation. 2023;52(3):621-648.

34. Aiming for equitable precision medicine in diabetes. Nat Med. 2022 Nov;28(11):2223. doi: 10.1038/s41591-022-02105-6. PMID: 36333401.

35. Ben Ayed A, Ben Halima M, Alimi A. Survey on clustering methods: Towards fuzzy clustering for big data. Conference: International Conference on Computational Intelligence in Security for Information Systems. 2015;9. DOI:10.1007/978-3-319-47364-2_55