An Embedding Approach for Biomarker Identification in Hypertrophic Cardiomyopathy

Arash Kazemi-Díaz[1], Luis Bote-Curiel[1], María Sabater-Molina[2], Juan Ramón Gimeno-Blanes[3], Salvador Sala-Pla [4], Francisco Javier Gimeno-Blanes[5], Sergio Muñoz-Romero [1], José Luis Rojo-Álvarez[1]

[1] Teoría de la Señal y Comunicaciones y Sistemas Telemáticos y Computación. Universidad Rey Juan Carlos, Spain.
[2] Laboratorio de Cardiogenética. Instituto Murciano de Investigación Biosanitaria, Spain.
[3] Unidad de Cardiopatías Familiares. Hospital Clínico Universitario Virgen de la Arrixaca, Spain.
[4] Departamento de Fisiología. Universidad Miguel Hernández, Spain.
[5] Teoría de la Señal y Comunicaciones. Universidad Miguel Hernández, Spain.

Abstract

Hypertrophic Cardiomyopathy (HCM) consists of a thickening of the cardiac muscle, causing fatigue, changes in the cardioelectric system, arrhythmias, and even sudden deaths. Variants in gene MYBPC3 are a well-known cause of this illness. Our objective was to find variants in other genes that can cause this pathology. For that purpose, genetic data from a group of patients is analyzed using embedding methods, which allow a lower dimensional representation, which is very helpful for visualization, diagnosis, and personalized therapy. Our results, applying different methods –Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP), Orthonormalized Partial Least Squares (OPLS) and Supervised Autoencoders– on genetic data showed a very good separability in the embedded space, allowing us to identify 10 variants that cause that separability. These results may be useful for identifying new HCM cases and implementing new Machine Learning models in those embedded spaces.

1. Introduction

Hypertrophic Cardiomyopathy (HCM) is an inherited cardiac disease mainly characterized by a thickening in the cardiac muscle. This leads to changes in the electric system of the heart and can cause fatigue, arrhythmias, and, in some cases, sudden cardiac deaths. Epidemiological studies in the last years estimate a prevalence of 1 in 500 people in the general population [1]. This disease is associated with mutations in genes encoding proteins of the cardiac sarcomere, Z-disc, and calcium-controlling proteins. Some of them are well known as the ones that affect the Myosin Binding Protein C (MYBPC3) [2]. Nevertheless, many other variants may be also a cause of this disease.

To identify some possible causes, very large datasets containing genetic information are usually used. This makes it very difficult to find patterns in data and relations between variants. Intending to solve this problem, Machine Learning has broken into medicine [3], giving some very useful tools that allow researchers to find intrinsic data relations that is impossible to visualize for humans.

In this work, we present some embeddings-based methods, which project the original data onto lower dimension spaces creating new latent variables that capture the information of the original variables. This allows the visualization of the high-dimensional data distribution in 2 or 3 dimensions showing the differences between HCM patients and control samples, leading to the identification of some additional biomarkers that can be causes of these differences. In that embedded space, we can also find how similar or different the patients and controls are depending on how far or close they are from each other.

2. Materials and Methods

2.1. Data

This study used genetic data obtained with the Next Generation Sequencing technique. The available data corresponded to 62 HCM patients, who belong to 62 families. There are 46 men and 16 women and the mean age is 46.08 ± 16.78 years. On the other hand, we have 73 control subjects. All the data is provided by the Hospital Clínico

Universitario Virgen de la Arrixaca (HCUVA, Spain). The data was provided in Variant Call Format, and it contained information on Single Nucleotide Polymorphisms (SNPs). The data from the subjects were genes related to different cardiomyopathies. The genes selected for the study were those sequenced in all patients and controls to avoid biases in the comparisons as much as possible.

The first step of our study consisted of preprocessing the data, eliminating all the information that was not relevant for the study, and codifying the SNPs in a suitable form for applying the corresponding models. We hypothesized that if a mutation is relevant, it will appear in all the patients. Therefore, we selected only the positions mutated in all the HCM patients. There was a total of 57 variants that fulfilled this condition. Then, we constructed a matrix, namely $\mathbf{X} \in \mathbb{R}^{n \times m}$, where $n = 135$, was the number of samples, and $m = 57$ was the number of variables. The variables considered contained the chromosome, the position, and the alteration in that position (letter of the nucleotide). The elements, x_{ij}, of that matrix were: 0 if the sample had the reference nucleotide at that position; 1 if it had the $j-th$ alteration in heterozygosis; And 2 if it had the $j-th$ alteration in homozygosis. We also create the output vector, $\mathbf{y} \in \mathbb{R}^n$, whose elements were $+1$ for HCM patients and -1 for control.

2.2. Embedding methods

Although this dataset was not as complex as the initial data, it was still impossible to find any pattern in the data or visualize it. So we wondered if an embedded space of latent variables and low dimensions could properly represent our data. To check this hypothesis, we propose the following methods.

The first embedding method used for this purpose is the **unsupervised methods**. These models are those in which the algorithm is trained without using labeled data [4]. Here, we propose 3 methods. Firstly, Principal Component Analysis (PCA) [5]. It is a multivariate technique that extracts the most important information from a dataset and represents it as a set of new orthogonal variables called Principal Components (PC). Mathematically, PCA depends on the matrix's Singular Value Decomposition (SVD). Secondly, t-distributed Stochastic Neighbor Embedding (t-SNE) [6] is an algorithm that embeds high dimensional points in low dimensions respecting the similarities between data on a multidimensional distribution sense. The embedding is a nonlinear map created respecting the statistical proximity among points in higher dimensions. And thirdly, Uniform Manifold Approximation and Projection (UMAP) [7] is a manifold learning technique for dimensionality reduction, which is based on Riemann geometry and algebraic topology.

The second embedding method used is the **supervised** algorithms. Those are the models in which the algorithm is trained, taking into account the labels of the data [4]. The methods proposed in this class were 3. Firstly, we used Supervised Autoencoders [8]. An autoencoder is a Neural Network where the outputs are set to the inputs, with 2 symmetric parts, an encoder and a decoder. In a supervised autoencoder, we add a supervised loss between both blocks on the representation layer. Secondly, we used the supervised variation of UMAP. And thirdly, we also proposed Orthonormalized Partial Least Squares (OPLS) [9], a method that removes variation from the data matrix that is not correlated with the output. This algorithm can be seen as a preprocessing method to remove systematic orthogonal variation from the dataset. The limitation of this method with respect to the others is that we can only embed the data onto a space that that have as much dimensions as classes (in this case, 2).

3. Experiments and Results

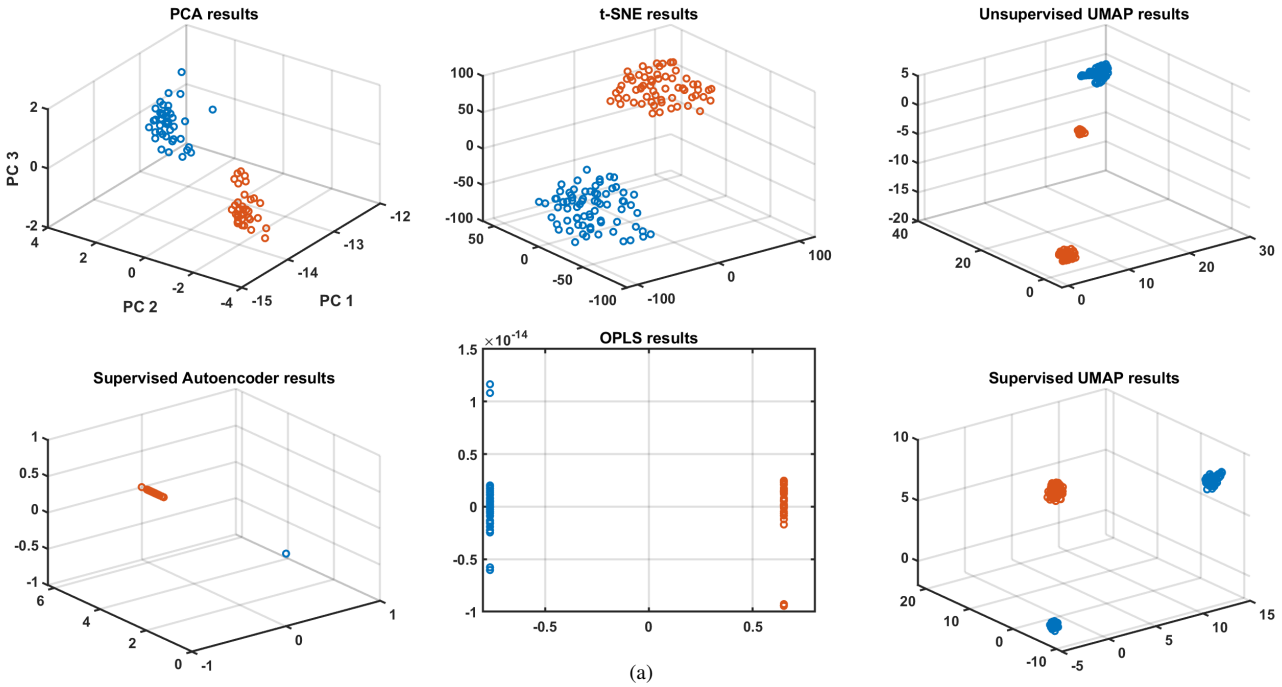
Let the dataset used for the experiments be $\{\mathbf{X}, \mathbf{y}\}$, which was defined before. The idea of these experiments is to represent the 57-dimensional dataset that we have in 3 dimensions, using the previously described techniques to visualize it and see if any separability exists between HCM patients and controls.

The results are shown in Figure 1a. All the methods show perfect separability between the controls and the HCM patients, although the embedded subspaces created are different. This means that some variants exist that make the difference between the affected patients and the controls, which means that they can be involved in the development of the disease.

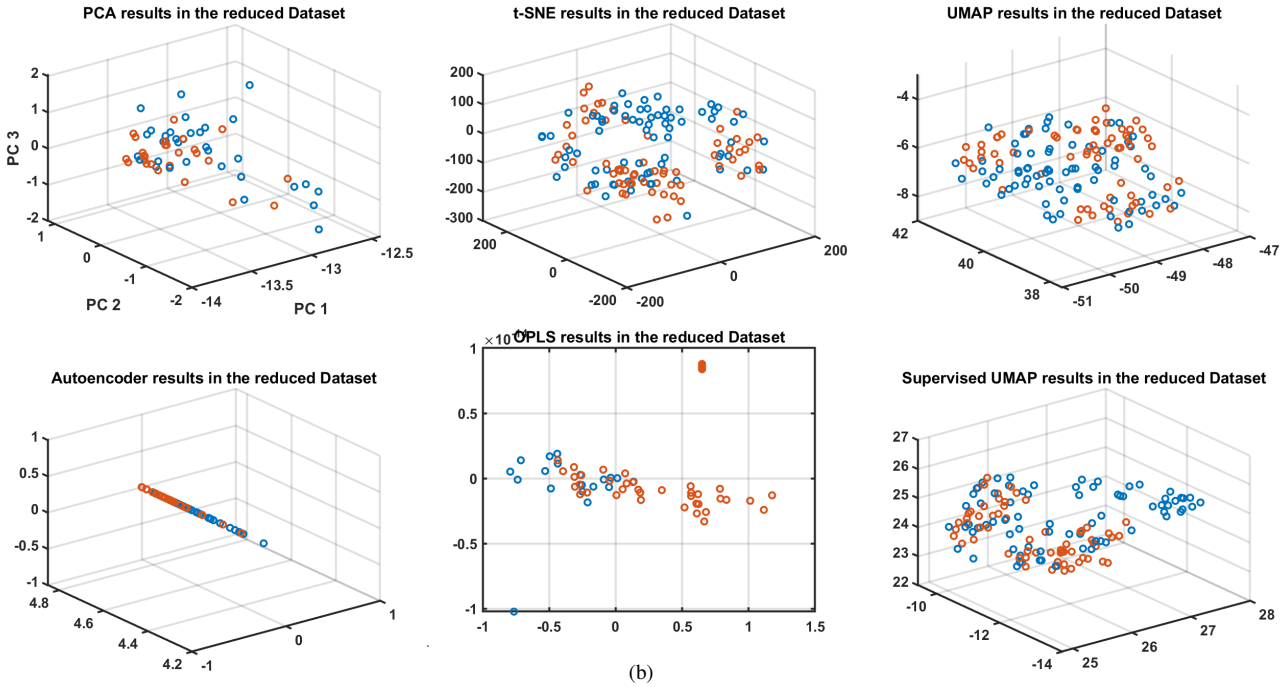
Therefore, the next step is identifying the variables that can cause this separability. It might be obvious that the separability will occur due to the variants appearing in all the patients but not in any control. In this step, we identified 10 variants that belong to 9 genes, shown in Table

Gene	Chromosme	Position	SNP
CASQ2	1	116,311,198	C
RYR2	1	237,730,124	G
SOS1	2	39,224,351	T
TTN / TTN-AS1	2	179,623,939	C
TTN /RP11-88L24.4	2	179,643,886	G
CACNA1D	3	53,529,140	C
ANK2	4	114,267,023	A
MYBPC3	11	47,364,762	G
ABCC9	12	22,047,151	T
HCN4	15	73,616,635	C

Table 1: Variants present in all the HCM patients but not in any control.



(a)



(b)

Figure 1: Results of the implementation of all the models to the matrix (a) with all the variables, and (b) without the variants that do not appear in any control. The orange points correspond to HCM patients and the blue points correspond to control subjects.

1. Each row of the table shows the next information for one of those variables: the name of the gene, the chromosome and position of the SNP (according to the reference GRCh37/hg19), and the SNP, the nucleotide present in that position. Note that all the patients also presented the same alteration at that position.

Once we identified the variants, we applied the same methods again, but to a matrix without the variables shown in Table 1, to see how the data is distributed. These results are shown in Figure 1b. All the methods presented a loose separability now in the embedded space. As expected, there were no big differences between the controls and the patients in the rest of the variables.

To measure the separability in the embedded space, we constructed a classifier using the Support Vector Machines method [10], which is a Kernel method in which the hyperplane is created by maximizing the margins, i.e., the distance to each class. In this case, we choose the *Radial Basis Kernel* and tune the hyperparameters (C and γ) using a 5-fold cross-validation. For measuring the quality of the results, we used 3 metrics, accuracy (ACC), sensitivity (SENS), and specificity (SPE), defined as usual.

In our case, the mean results are the following after doing 30 realizations and choosing different training and test sets. First, for PCA, $ACC = 0.64$, $SENS = 0.52$, and $SPE = 0.76$. Second, for tSNE $ACC = 0.68$, $SENS = 0.82$, and $SPE = 0.58$. Third, for UMAP, $ACC = 0.68$, $SENS = 0.81$, and $SPE = 0.59$. Fourth, for supervised Autoencoders, $ACC = 0.63$, $SENS = 0.51$, and $SPE = 0.73$. Fifth, in the case of OPLS, $ACC = 0.66$, $SENS = 0.61$, and $SPE = 0.71$. Finally, for Supervised UMAP $ACC = 0.65$, $SENS = 0.73$, and $SPE = 0.59$. The ACC is low and medium-low in all cases, and regarding the SENS and SPE, we can see that the classifiers can struggle when finding and identifying positive or negative values depending on the methods.

4. Conclusion

This study explored the differences between controls and HCM patients embedding the original data onto lower-dimensional latent spaces. By doing this, we have identified 10 variants in 9 different genes that are causes of that difference, and we have been able to quantify how far or close is the HCM-affected subjects from non-affected ones.

Although this information is not conclusive enough to determine whether those variants are a cause of HCM or not, it may help clinicians in the task of identifying new HCM cases. On the other hand, working on those embedded spaces may be very useful in future works for implementing new classifiers, expecting improved predictions concerning the original spaces.

Acknowledgements

This work was funded by the European Union Next Generation EU, in the context of the 2022 Recovery, transformation, and Resilience Plan, project budget 30G1ININ22. It was also supported by the Ministry of Economy and Competitiveness, grant IPT-2012-1126- 300000, AEI/10.13039/5011000110033-PID2019-106623RB, AEI/10.13039/ 5011000110033 PID2019-104356RB, AEI/10.13 039/ 5011000110033-PID2022-140786NB-C31, 2022 -REGING-95982, and 2022-REGING-92049.

References

[1] de Oliveira Antunes M, Luis Scudeler T. Hypertrophic cardiomyopathy. IJC Heart Vasculature 2020;27.

[2] Sabater Molina M, et al. A novel founder mutation in MYBPC3: Phenotypic comparison with the most prevalent MYBPC3 mutation in spain. Revista española de cardiología 2017;70(2):105–114.

[3] Sidey-Gibbons JA, et al. Machine learning in medicine: a practical introduction. BMC medical research methodology 2019;19:1–18.

[4] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nature Reviews Genetics 2015; Vol:16:321–332.

[5] Abdi H, Williams LJ. Principal component analysis. Wiley interdisciplinary reviews computational statistics 2010; 2(4):433–459.

[6] Van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of machine learning research 2008;9(11).

[7] McInnes L, et al. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. arXiv preprint arXiv180203426 2018;.

[8] Le L, Patterson A, White M. Supervised Autoencoders: Improving generalization performance with unsupervised regularizers. Advances in neural information processing systems 2018;31.

[9] Muñoz-Romero S, et al. Sparse and kernel OPLS feature extraction based on eigenvalue problem solving. Pattern Recognition 2015;48(5):1797–1811.

[10] Salcedo-Sanz S, et al. Support vector machines in engineering: an overview. Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery 2014;4(3):234–267.