High-Dimensional Feature Characterization of Single Nucleotide Variants in Hypertrophic Cardiomyopathy

Dafne Lozano[1], Luis Bote[1], Concha Bielza[2], Pedro Larrañaga[2], María Sabater[3], Juan Ramón Gimeno[4], Sergio Muñoz[1], Francisco Javier Gimeno[5], José Luis Rojo[1]

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

Abstract

Hypertrophic cardiomyopathy is a genetic disorder that affects the structure of the heart muscle, which can lead to sudden cardiac arrest. The genetic characterization of biomarkers remains an open area, and machine learning techniques are being proposed for its detection. This research aims to apply several of these methods to obtain single nucleotide variants (SNVs). We followed a three-stage approach: First, the initial set of 118142 *SNV features were filtered with the union of Manhattan threshold from biostatistics together with the Chi-squared test and with a logistic regression based univariate filtering method, yielding a preselected set of* 1974 *features; second, linear classifiers (support vector machine and Fisher linear discriminant analysis) identified and ranked the relevant features to distinguish between normal subjects and diseased patients. Finally, two additional techniques (informative variable identifier and Bayesian networks) were used to scrutinize the inter-feature relationships of the SNVs. The results showed a consensus between linear classifiers in which variants with higher weights coincide. The* 100 *variants with higher weights were visualized to analyze their relationships. To validate the result, the top-ranked variants were checked in the literature. Most of them were directly implicated with the disease or participated in cardiac remodeling, meaning that these variants can be considered modulators of the disease.*

1. Introduction

Heart disease is the leading cause of death worldwide. Despite the advances in the knowledge of its treatment, it continues to cause the death of approximately 17.9 million people yearly. Hypertrophic cardiomyopathy (HCM) is one of the most common hereditary heart diseases. HCM is characterized by left ventricular hypertrophy and an undilated left ventricle with an altered ejection fraction and it has a genetic origin. However, their complex features suggest a complex genetic substrate [1].

On the other hand, in recent years, machine learning (ML) methods have been applied broadly to health problems that can be addressed from a data-driven perspective. So far, different works have been developed on ML and cardiac diseases. For example, they have been applied for diagnosing heart diseases by interpreting electrocardiogram signals [2]. Studies using genomic data from heart disease often focus on parametric statistical methods.

In this context, it is important to conduct research on ML techniques to process and analyze vast amounts of genetic information to identify patterns and determine the variants that are associated with the disease. Therefore, the first stage of our research consists of the consensus of univariate filtering methods to obtain the most significant single nucleotide variants (SNVs) associated with the disease. Then, with linear classifiers, the variants associated with the disease can be detected and ranked. Hence support vector machines (SVMs) and Fisher linear discriminant analysis (FLDA) provide the variants with higher weights. Once the ranking is established, the top 100 variants are selected to visualize their interactions and how the variables are related using informative variable identifier (IVI) and Bayesian networks (BNs). Finally, the top features are checked in the literature, and it is found that they are variants of genes associated directly with the disease or produce cardiac remodeling.

2. Materials and Methods

In this section, the description of the dataset as well as an introduction to the filtering methods and ML methods, are provided.

2.1. The Dataset

In this study, genetic data from 62 patients with HCM with an average age of 46 years and 73 control subjects were obtained using Next Generation Sequencing. The data was provided by the Hospital Clínico Universitario Virgen de la Arrixaca (HCUVA) in Spain in variant call format (VCF) and contained information on SNVs. The data sequenced for the HCM patients contained approximately 500 genes, while the control data provided by a partner company was composed of patients without heart disease. However, the genes selected for the study were those sequenced in all patients and controls. The preprocessing of the raw data consisted of selecting just the genetic information containing SNV; insertions or deletions were discarded and numerically codified the genotype depending on whether the variant corresponds to monozygotic or heterozygotic. Finally, the genetic information is composed of 118142 SNVs.

2.2. Filter Methods

Since we are starting from a large data set of 118142 SNVs, we have used the union result of three filter methods to identify and prioritize relevant features, which consists of the preselection of variables based on the p-value.

The first method is the chi-squared (χ^2) test [3] for each SNV, which is a measure of the difference between the observed frequency (O) of the SNV in each class and the expected frequency (E) under the null hypothesis of independence: $\chi^2 = \sum_i \sum_j \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$. The resulting χ^2 statistic is then compared to the distribution of χ^2 values with one degree of freedom under the null hypothesis to obtain a p-value.

Logistic regression is another used method for feature selection by calculating the corresponding p-value. It is a statistical model used to predict a binary outcome (the presence or absence of a disease) based on predictor variables (SNVs in our case) [4]. The coefficients in the logistic regression model can be used to assess the importance of each predictor variable, using the statistical $z = \frac{\beta}{SE(\beta)}$, where β is the estimated coefficient, and $SE(\beta)$ is the standard error of the coefficient. The resulting z statistic is compared to the standard normal distribution to obtain a two-sided p-value. Hence, SNVs with lower p-values are considered to be more associated with the outcome.

In both previous methods, we need to control the false positive rate (FPR) [5], which is the expected number of false positives from all the hypothesis tests performed and we want to guarantee that all hypothesis tests' percentage of false positives is 5% or less. For this purpose, the Bonferroni correction is used. However, it is highly criticized in these studies for not taking into account that the tests are not independent since the tests are correlated by linkage disequilibrium (LD). Many modifications to the Bonferroni method have been proposed, including methods that take LD, false discovery rate, and false positive rate into account. Essentially, the currently accepted genome-wide significance threshold is 5×10^{-8}, which can be considered the Manhattan threshold since it is usually accompanied by the Manhattan plot. This value can be obtained by calculating the $-log(p\text{-value})$ and considering the significance of the SNV that exceeds the threshold [6].

2.3. Machine Learning Methods

In this section, we introduce the methods used in this study. The first two, SVMs and FLDA are used to obtain the weights of all the variables to identify and rank the relevant features, while IVI and BNs are used as methods to scrutinize the inter-feature relationships.

Firstly, SVMs [7] goal is to establish a decision boundary between two classes that allow for label prediction using one or more feature vectors. So the objective is to maximize the margin $1/\|\mathbf{w}\|^{\mathbf{2}}$ between classes, where $\mathbf{w}$ refers to the vector of weights. We can use the kernel method to create non-linear and higher-dimensional models. The kernel function maps the input to a higher-dimensional space, $K(\mathbf{x},\mathbf{y}) =< \mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{y}) >$, where K is the kernel function, the n dimensional inputs are $\mathbf{x}$ and $\mathbf{y}$, and $\mathbf{f}$ maps the input to m dimensional space from n and $< \mathbf{x},\mathbf{y} >$ denotes the dot product. Therefore, with this model, we can obtain the different weights $\mathbf{w}$ of each variant.

The second method to obtain the vector of weights is FLDA which is a technique for finding a linear combination of features that can distinguish between two or more groups [8]. The goal is to maximize the distance between the projected data classes while minimizing the relative variances of the points around their means. Hence, obtaining the vector $\mathbf{w}$ such that $g(\mathbf{x}) = \mathbf{w^T x} + \boldsymbol{w_0} = 0$, where $\boldsymbol{w}_0$ is the bias term. This is achieved by introducing Fisher's discriminant ratio and by differentiating and equating the criterion to zero.

The next step is to analyze the inter-feature relationships of the SNVs, which can be done using IVI and BNs. IVI method is a technique that can categorize each feature as informative or noisy, find connections between informative features, and create a ranking in order of importance [9]. The IVI algorithm is based on the initial assumption that the weights learned by a linear classifier method can summarize important relationships between features, so the input variable space is converted into a weights space by a low-cost weight generator. To identify redundant features, the IVI method calculates the similarity between different variables using the Pearson correlation coefficient, $\rho_{l,z} = \mathrm{corr}\,(w_l, w_z)$, where corr indicates the normalized correlation coefficient, w_l indicates the weights cor-

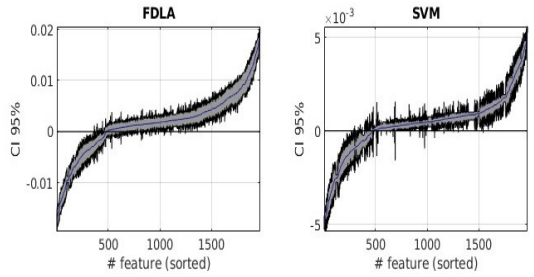

Figure 1. CI of sorted variables in FLDA and SVM.

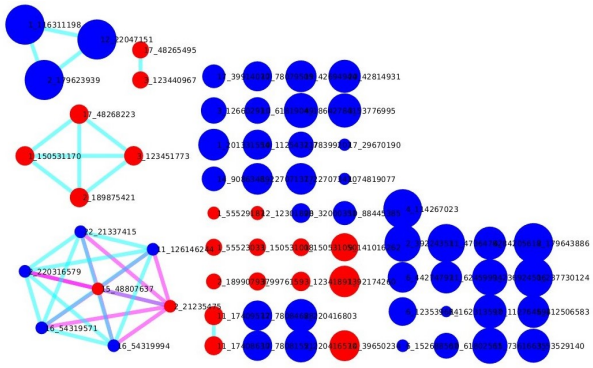

Figure 2. IVI representation of 68 top-ranked variants.

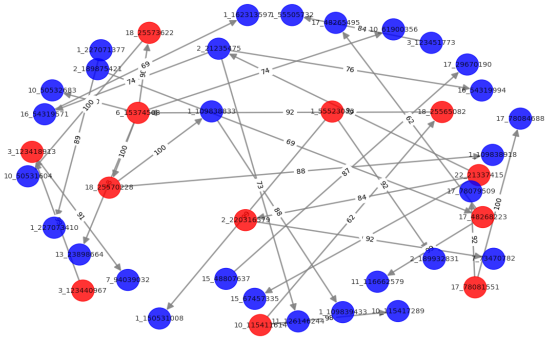

Figure 3. BN structure learned of the variables associated with the disease (class 1).

responding to the l-th variable, while w_z indicates those corresponding to the z-th variable.

Finally, BNs are also used to obtain the feature relationships. Specifically, we learn the BN using the hill climbing algorithm with bootstrap resampling. This approach allows us to extract the most significant relationships between variants, emphasizing those that appear more frequently in the resampling process. Notably, our specific interest lies in uncovering the relationships between variants in individuals with the disease, further enhancing our understanding of the intricate genetic mechanism associated with the condition.

3. Experiments and Results

The three univariate filter methods were applied separately. As a result, 1767, 207, and 764 SNVs were considered significant for the χ^2, logistic regression, and Manhattan threshold methods, respectively. Due to the discrepancy in the number of significant SNVs as well as the possibility of losing potential SNVs related to the disease, the selection consists of the union of the three results, resulting in 1974 SNVs.

After obtaining the preselected set of features, both linear classifiers were applied, the way of proceeding consisted of choosing the most appropriate parameter values by means of a 10-fold cross-validation. The performance of the methods is appropriate since the mean error probability during 10-fold is 0.03 and 0 for FLDA and SVM respectively. In other words, the separation produced by the weights, especially in SVM was complete between the two classes. A 95% confidence interval (CI) was obtained after performing bootstrap resampling. In both methods, the number of variables that do not overlap zero is similar, 1570 for FLDA and 1588 for SVM, as shown in Figure 1.

To scrutinize the inter-feature relationships of the SNVs, IVI was applied to the dataset to obtain the relationship between genes and the corresponding informative and redundant variables. The algorithm was executed 150 times, and the variables considered relevant are those that appear in every iteration. Moreover, a narrow interval was established for the redundant variables. However, for an appropriate visualization process, just the 100 top-ranked SNVs were selected since they were strongly related to the disease according to previous ML methods. As a result, the algorithm considers significant 68 variants. In Figure 2 the relationships between those variants are shown. At the left of the image, we can see the variants that are related between them forming groups, while at the right of the image, the variants that are positively and negatively related to the disease by themselves can be observed.

On the other hand, for obtaining the inter-feature relationships with BNs, hill climbing algorithm is used to learn the structure while performing bootstrap resampling. In this way, we can focus on the strongest relationships in the learning process which are the ones that appear at least 60 out of 100 times as shown in Figure 3. Moreover, the red nodes represent a subset of variants that exhibit strong relationships, appearing at least 80 times in the resampling process.

It is important to note that the majority of SNVs have not been extensively studied clinically. This means that we often lack comprehensive knowledge about their functional impact and whether they are associated with malignancy

or other disease conditions. By identifying the genes associated with the highest-ranking variants, we can focus our attention on specific genomic regions probably linked to HCM. As a result, a literature review of the 5 top-ranked genes obtained from the classifiers has been carried out: RAF1 is frequently reported to produce an increased kinase activity, causing HCM [10]. Another gene obtained is CACNA1C which has been linked to the mixed phenotype of HCM, congenital heart disease, and long QT syndrome [11]. Moreover, variations in NRAP genes have been found in elderly patients with HCM, affecting cardiac muscle organization [12]. It has been discovered that mutations in MYPN can disrupt muscle structure and signaling, leading to HCM [13]. Therefore, the genes obtained with higher weights have been discovered to be directly related to the target disease, or are involved in changes in the cardiac function. These results indicate that the variants in these genes can be modulators of expression without being the main genes causing the disease, that is, HCM can be caused by the known causal genes, and others can act as modulators changing the clinical outcome.

4. Conclusions

The results have shown the reliability of ML methods used to obtain the weights of each SNV due to the results achieved, including a low probability of error. Additionally, the methods that allowed us to obtain the inter-feature relationship provided a comprehensive view of how the SNVs are related, specifically if they are joined to form groups of greater relevance. Moreover, identifying the top-ranked genes as being related directly to the disease or as participating in cardiac remodeling gives evidence of the presence of genetic modulators in the disease.

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. TED2021-131310B-I00 project, funded by the Ministry of Science and Innovation. 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, AEI/10. 13039/501100011033-PID2022- 139977NB-I00, 2022 - REGING-95982, and 2022- REGING-92049 and by the grant to the ELLIS Unit Madrid by the Autonomous Region of Madrid. Furthermore, special thanks to the collaborating company Health in Code for providing the data belonging to the controls.

References

[1] Szczesna-Cordary D, Morimoto S, Gomes AV, Moore JR. Cardiomyopathies: Classification, clinical characterization, and functional phenotypes. Biochemistry Research International 2012;2012:870942.

[2] Ayano YM, Schwenker F, Dufera BD, Debelee TG. Interpretable machine learning techniques in ecg-based heart disease classification: A systematic review. Diagnostics 2023;13(1).

[3] Fisher RA. On the interpretation of χ^2 from contingency tables, and the calculation of p. Journal of the Royal Statistical Society 1922;85(1):87–94.

[4] Tolles J, Meurer WJ. Logistic regression: Relating patient characteristics to outcomes. JAMA 2016;316(5):533–534.

[5] Goeman JJ, Solari A. Multiple hypothesis testing in genomics. Statistics in Medicine 2014;33(11):1946–1978.

[6] Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. Nature Reviews Genetics 2014;15(5).

[7] Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genomics Proteomics 2018; 15(1):41–51.

[8] Theodoridis S. Chapter 7 - classification: A tour of the classics. In Theodoridis S (ed.), Machine Learning (Second Edition). Academic Press, 2020; 301–350.

[9] Muñoz-Romero S, Gorostiaga A, Soguero-Ruiz C, Mora-Jiménez I, Rojo-Álvarez JL. Informative variable identifier: Expanding interpretability in feature selection. Pattern Recognition 2020;98:107077.

[10] Pandit B, Sarkozy A, Pennacchio LA, Carta C, Oishi K, Martinelli S, Pogna EA, Schackwitz W, Ustaszewska A, Landstrom A, Bos JM, Ommen SR, Esposito G, Lepri F, Faul C, Mundel P, López Siguero JP, Tenconi R, Selicorni A, Rossi C, Mazzanti L, Torrente I, Marino B, Digilio MC, Zampino G, Ackerman MJ, Dallapiccola B, Tartaglia M, Gelb BD. Gain-of-function raf1 mutations cause noonan and leopard syndromes with hypertrophic cardiomyopathy. Nature Genetics 2007;39(8):1007–1012.

[11] Gakenheimer-Smith L, Meyers L, Lundahl D, Menon SC, Bunch TJ, Sawyer BL, Tristani-Firouzi M, Etheridge SP. Expanding the phenotype of CACNA1C mutation disorders. Mol Genet Genomic Med 2021;9(6):e1673.

[12] Ankit S, Koranchery R, Rajendran R, Mohanan KS, Shenthar J, Dhandapany PS. Next generation sequencing reveals nrap as a candidate gene for hypertrophic cardiomyopathy in elderly patients. bioRxiv 2019;.

[13] Purevjav E, Arimura T, Augustin S, Huby AC, Takagi K, Nunoda S, Kearney DL, Taylor MD, Terasaki F, Bos JM, Ommen SR, Shibata H, Takahashi M, Itoh-Satoh M, McKenna WJ, Murphy RT, Labeit S, Yamanaka Y, Machida N, Park JE, Alexander PMA, Weintraub RG, Kitaura Y, Ackerman MJ, Kimura A, Towbin JA. Molecular basis for clinical heterogeneity in inherited cardiomyopathies due to myopalladin mutations. Hum Mol Genet 2012;21(9):2039–2053.