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

Introduction. 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. **Methods**. This research aims to apply several of these methods to obtain single nucleotide variants (SNV). 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 filter method, yielding a preselected set of 1974 features with the union-set criterion; Second, linear classifiers (support vector machines and fisher linear discriminant analysis) identified the relevant features to distinguish between normal subjects and disease patients, moreover, these methods produced a ranking that can give an insight of which variants most are implicated in the disease. Finally, two additional techniques (informative variable identifier and Bayesian networks were used to scrutinize the inter-feature relationships of the SNVs, in this way relationships and groups between variants were provided. **Results** and **conclusions**. It was shown a consensus between linear classifiers in which variants with higher weights coincide. With informative variable identifier, the 100 variants with higher weights were visualized to check how they are related, while with the Bayesian network, specifically with tree-augmented naive Bayes the top 30 variables with higher mutual information were visualized. 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 genetic modulators.