

Optimal Subset of Differentially Expressed Genes for Atrial Fibrillation Identification and Anatomical Origin Localization

Meri Gjika*, Agnese Sbröllini, Vincenzo Lucio Caputo, Sara Paratico, Carlo Pappone, Emanuela Teresina Locati, Laura Burattini

Università Politecnica delle Marche, Ancona, Italy

Introduction: Atrial fibrillation (AF) is a common sustained cardiac arrhythmia, associated with complex and heterogeneous atrial remodeling processes. Transcriptomic analyses allow the identification of differentially expressed genes (DEGs) associated with AF; however, high-dimensional genetic data often contain redundancy that can obscure meaningful information and reduce the generalizability of classification algorithms. In this study, we aimed to identify the optimal DEGs subset size capable of discriminating AF from SR and to localize AF anatomical origin in terms of left atrium (LA) or right atrium (RA).

Methods: Gene expression profiles from 26 paired atrial biopsy specimens from the LA and the RA were obtained from the GSE79768 dataset. Significant DEGs were identified and systematically reduced to evaluate the impact of feature space size on a multiclass classifier's (with class being LA-AF, LA-SR, RA-AF, RA-SR) performance, using the averaged area under the receiver operating curve ($\overline{AUC}$) as the primary metric.

Results: The results demonstrate that the highest classification performance, with an $\overline{AUC}$ of 87.80%, was achieved using a subset size of only 7% of the total DEGs. This optimal subset included 91 genes, consisting of 70 up-regulated and 21 downregulated transcripts. Notably, increasing the number of DEGs beyond this 7% threshold led to a progressive decline in $\overline{AUC}$, suggesting that larger datasets introduce noise and redundancy. Statistical assessment revealed that SR-related classes achieved the highest accuracy (LA-SR: 93.34%; RA-SR: 91.67%). Among AF classes, the LA-AF group outperformed the RA-AF group in all parameters (AUC of 87.97% vs. 78.21%). This difference may suggest a homogeneous molecular remodeling pattern in the left atrium compared to the higher transcriptomic heterogeneity of the right atrium.

Conclusion: The results indicate that a small subset of DEGs (7%) of the most significant genes can achieve optimal classification performance. Future studies will investigate integrating transcriptomic data, including other molecular layers.