The Endotypes* Discovery pipeline, a powerful tool in R language for patients stratification using high-dimensional omics data

EMILIEN JEMELEN[§] and EMILIE GERARD[†]
[§]emilien.jemelen@ensae.fr, ENSAE Paris, Palaiseau
[†]emilie.gerard@sanofi.com, Sanofi R&D, Chilly-Mazarin

Abstract

An Endotypes Discovery pipeline has been developed in a user-friendly format for automation of endotype discovery in high dimensional settings. The Pipeline tends to outperform classical clustering algorithms such as K-means and hierarchical clustering in various scenarii.

In the context of heterogeneous diseases, the clinical variables (phenotypes) might not be sufficient to identify which groups of patients will better respond to a given treatment. Omics data offer the opportunity for patients stratification at molecular biology level. However, identifying robust clusters reminds challenging due to the nature of omics data (few patients and thousands of features). Indeed, high dimensionality tends to make patients equidistant. Many Machine Learning algorithms have been developed throughout the years to tackle that curse of dimensionality, but always with the bias of underlying assumptions regarding the structure of the data. Our Endotypes Discovery pipeline contains several modules: generation of partitions using several clustering algorithms, removal of partitions with poor clustering quality, definition of a consensus from the retained partitions, cluster stability assessment with resampling, and cluster characterization using clinical variables. Furthermore, it was developed with optimally determined parameters to avoid random choices from the user.

Key words: patient stratification, endotypes, omics

*groups of patients based on functional/pathobiological mechanisms.