Biomarker discovery through integrated models of patient metabolic and genetic fingerprints

Effrosyni Karakitsou[1,2,3]; Carles Foguet[1,2]; Silvia Marin[1,2], Pedro de Atauri[1,2]; Jean-Baptiste Cazier[3]; Marta Cascante[1,2]

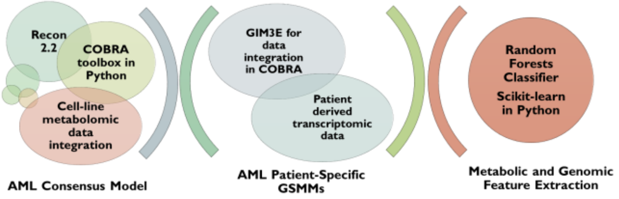
1 Department of Biochemistry and Molecular Biomedicine, Faculty of Biology, Universitat de Barcelona, Av Diagonal 643, 08028, Barcelona, Spain
2 Institute of Biomedicine of University of Barcelona (IBUB) and Associated unit to CSIC, Av Diagonal 643, 08028, Barcelona, Spain
3 Centre for Computational Biology, University of Birmingham, B15 2TT, Birmingham, UK

Introduction

The classical approach in cancer research focuses on the identification of specific genomic features, such as single mutations or the role specific genes play in the development and progression of different types of cancer. However, the need to treat cancer as a multi-factorial disease and therefore take under consideration the entirety of events and the different levels of regulation that occur is becoming increasingly evident. In addition, accounting for the specific genetic and phenotypic fingerprint of each patient will pave the way towards an accurate stratification of patients, better treatment efficacy and elimination of potential drug side-effects. Through multi-omic data integration different types of medical and biological datasets can be integrated, so that a more accurate interpretation can be achieved. Genome-Scale Metabolic Models (GSMMs) provide an excellent platform to integrate various omics, whereas statistical modeling and the application of machine learning provides the means to handle the ever-increasing amount of biomedical data and their inherent complexity.

Methods

Metabolomic data from AML cell-lines (THP1, HL60) were used to build an Acute Myeloid Leukemia (AML) consensus GSMM.
Transcriptomic data from AML patients (TCGA, NEJM 2013) were integrated to build patient-specific GSMMs.
sparse Canonical Correlation Analysis (sCCA) was applied to perform feature extraction and reduce number of parameters.
The flux distributions calculated from the personalized GSMMs together with the transcriptomics and the meta-data of the patients were used to build a Random Forest (RF) classifier for AML subtypes and identify genomic and metabolic biomarkers in respect to different risk categories.


Recon 2.2
COBRA toolbox in Python
Cell-line metabolomic data integration
AML Consensus Model
GIM3E for data integration in COBRA
Patient derived transcriptomic data
AML Patient-Specific GSMMs
Random Forests Classifier
Scikit-learn in Python
Metabolic and Genomic Feature Extraction

Results

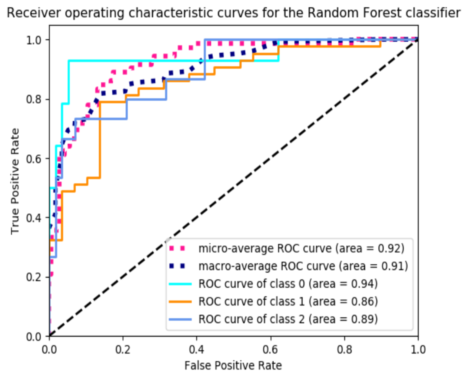
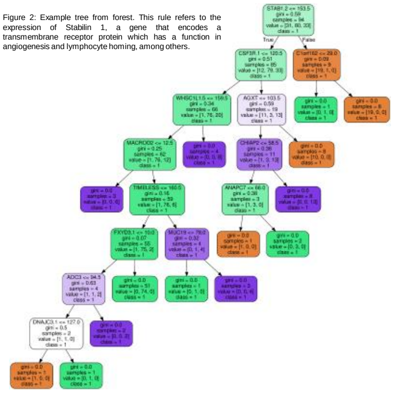


Figure 1: Receiver Operating Characteristic (ROC) curves were calculated appropriately for the multi-class Random Forest classifier. The micro -average ROC curve depicts the overall performance of the model across all classes. The macro-average ROC curve illustrates the overall performance of the model for each individual class. The corresponding Area Under the Curve (AUC) is presented in the legend.

Figure 2: Example tree from forest. This rule refers to the expression of Stabilin 1, a gene that encodes a transmembrane receptor protein which has a function in angiogenesis and lymphocyte homing, among others.



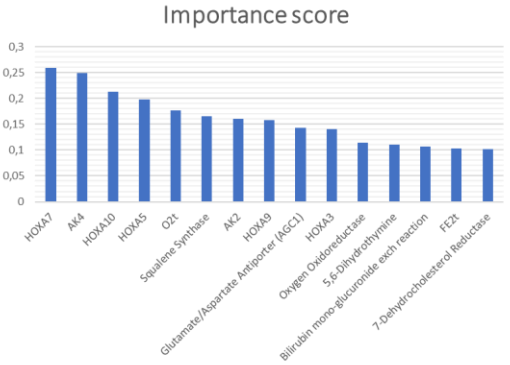


Figure 3: Importance scores as calculated using the Random Forest model for genomic and metabolic features. These features hold the highest importance regarding their significance in the classification of patients into different risk prognosis groups.

- The micro-average performance of the RF classifier was 0.92, whereas the macro-average performance was 0. 91 (Figure 1).
- The average number of nodes of the trees in the RF was 29 and the average depth of the trees was 6 (Figure 2).
- The age factor was not identified by the algorithm as an important classification feature for AML patients in regards to the disease risk category.
- The genes with the highest importance for the grouping of patients were five members of the HOXA family and two adenylate kinase isoforms, whose role in AML disease prognosis has already been reported (Figure 3).
- Haem catabolism and its end product bilirubin, as well as the sterol branch of cholesterol metabolism were distinctly represented in our results as valuable new putative prognostic biomarkers in AML (Figure 3).

Conclusions

○ GSMMs can serve as a comprehensive repository for the integration of transcriptomic data facilitating their biological interpretation
○ The RF algorithm handled efficiently the complexity of the data set, the multi-class classification application and overcame the challenge of dealing with a significantly smaller number of samples (=180) compared to the larger number of features (=49,362).
○ Our workflow of combining patient-specific metabolic models with machine learning has provided new prognostic biomarkers to explore for AML.

Financial Support

This work was supported by:
-European Commission (PhenoMeNal [654241],HaemMetabolome[675790])
-MINECO-European Commission FEDER funds – "Una manera de hacer Europa" (SAF2017-89673-R)
- Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) – Generalitat de Catalunya (2017SGR1033)
Carles Foguet acknowledges the support received through "Becas de la Caixa para estudios de doctorado en universidades españolas" funded by the "La Caixa" foundation
Marta Cascante acknowledges the support received through the prize "ICREA Academia" for excellence in research, funded by ICREA foundation – Generalitat de Catalunya.