Statistical approaches for the integration of omics data

Carlos J. Peña[1], *Juan Antonio Carbonell*[2], *Sheila Zúñiga Trejos*[3]

[1]cpena@incliva.es, Biostatistics Unit, Biomedical Research Institute - INCLIVA
[2]jacarbonell@incliva.es, Biostatistics Unit, Biomedical Research Institute - INCLIVA
[3]smzuniga@incliva.es, Bioinformatics and Biostatistics Unit, Biomedical Research Institute - INCLIVA

Introduction: Combining multiple omics data sets can help provide a better understanding of diseases or complex biological processes. This combination may cover the data generated from genome, proteome, transcriptome, metabolome or epigenome but it can be further extended to other biological data.
The integration of individual omics data improve prognostics and predictive accuracy of disease phenotypes. In recent years, several tools have been developed for data integration that allow to identify subtypes of a disease based on multi-omics profiles (disease subtyping) or to predict biomarkers for diagnostics. In this work, we focus on some of the tools that perform integration of multiple omics data and we describe their methods.

Material: The Cancer Genome Atlas (TCGA) is a repository which includes genomic, epigenomic, transcriptomic, proteomic and clinical data for 32 cancers. We selected the Breast Invasive Carcinoma (BRCA) cohort and focused our analyses on characterizing the breast cancer subtypes.
For this tumor type, we downloaded TCGA-curated data sets containing gene expression, miRNA expression and protein abundance quantification as well as the patients' clinical information including the overall survival data. The individual omics data sets consist of mRNA expression, miRNA expression and protein abundance measured on 367 patients.

Methods: Comparative analysis was performed for four unsupervised multi-omics data integration methods: 1) multi-omic factor analysis [1] (implemented in the R package *MOFA2*); 2) regularized and sparse Generalized Canonical Correlation Analysis [4] (package *RGCCA* in the R software); 3) joint latent variable models implemented in the R package *iClusterPlus* [2]; and 4) Joint and Individual Variation Explained (JIVE) decomposition [3].

Future Work: We plan to evaluate the effect of the latent factors resulting from the proposed methods on overall survival along with other clinical features.

Keywords: data integration, multi-omics, disease subtyping.

[1] Argelaguet, R. et al. (2018) Multi-omics factor analysis- a framework for unsupervised integration of multi-omics data sets . *Mol. Syst. Biol*. **14**, e8124.
[2] Mo Q. et al. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc Natl Acad Sci U S A*. **110**(11):4245-50. doi: 10.1073/pnas.1208949110.
[3] O'Connell MJ, Lock EF (2016). R.JIVE for exploration of multi-source molecular data. *Bioinformatics*. **32**(18):2877-9. doi: 10.1093/bioinformatics/btw324.
[4] Tenenhaus A. et al. (2014). Variable selection for generalized canonical correlation analysis. *Biostatistics*. **15**(3):569-83. doi: 10.1093/biostatistics/kxu001.