A network-based omics integration framework: overcoming the obstacle of high dimensional data

Francesco Gadaleta[1,2], Kyrylo Bessonov[1,2], Kridsadakorn Chaichoompu[1,2], Silvia Pineda[3], and Kristel Van Steen[1,2]

[1]Systems and Modeling Unit, Montefiore Institute, University of Liege, Belgium
[2]Bioinformatics and Modeling, GIGA-R, University of Liege, Belgium
[3]Centro Nacional de Investigaciones Oncologicas, Spain
[1]{*francesco.gadaleta, kristel.vansteen*}*@ulg.ac.be*

Abstract

Genome-wide association studies can potentially unravel the mechanisms behind complex traits and common genetic diseases. Despite the valuable results produced thus far, many questions remain unanswered. For instance, which specific common variants are linked to the risk of the disease under investigation, what biological mechanism do they act through or how do they interact with environmental and other external factors?

The driving force of computational biology is the constantly growing amount of big data generated by high-throughput technologies. The amount of available data and its heterogeneity seem to play a beneficial role rather than a detrimental one in discovering new genetic insights. Each type of data, e.g. gene expression, epigenetic, methylation, RNAseq or proteomics, provided by a diverse source of information, directly contributes to complete the overall puzzling picture of genetic disorders, with its own unique local knowledge. In such a scenario, data integration, as the practice of combining evidence from different data sources, represents the most challenging activity, due to the unattainable task of merging large and heterogeneous data sets.

A practical framework that fulfils the needs of integration is provided by means of networks. Due to the manifest risks of introducing bias when integrating heterogeneous data and because of the curse of dimensionality, which is a common aspect in computational biology, preliminary procedures of variable selection are essential. We investigate two approaches that capture the multidimensional relationships between different data sets, namely conditional inference trees and linear regression methods. A strategy that is common to both methods consists in using every expression trait as a dependent variable and the remaining expression traits as covariates. The aforementioned strategy will derive variable importance scores that, in turn lead to selecting important variables. We subsequently proceed with the construction of weighted networks that integrate evidence gathered from different data sets with the purpose of

detecting pathways and interactions among genetic compounds and better describing the complex mechanisms behind the traits under study.

Keywords. omics integration, network, regression, conditional inference tree