A novel approach for highly-diverse multi-omics data fusion applied to tomato germplasm selection

M. Pividori[1], G. Stegmayer[1], A. Cernadas[2], M. Conte[2], F. Carrari[2], D.H. Milone[1]

[1] Institute for Signals, Systems and Computational Intelligence (sinc(i)), FICH-UNL, CONICET, Argentina.

[2] Institute of Biotechnology, INTA-Castelar, CONICET, Argentina.

Background:

Tomato (*Solanum lycopersicum*) is one of the major vegetable crop consumed worldwide being a valuable source of vitamins and antioxidants for the human diet. Because of the variability constraints associated with breeding programs, the phenotypic and genetic diversity in heirloom varieties emerges as a landmark to rescue desired agronomic traits for crop improvement. A germplasm collection of Andean tomato landraces materials originally cultivated by family farmers in the Cuyo region (Mendoza-ARG), was characterized based on morpho-agronomic and biochemical traits of their mature fruits. In several growing seasons, highly-diverse kinds of quantitative and qualitative measurements were obtained using GC-MS, NMR and HPLC to quantify fruit soluble and volatile metabolites; transcriptomics to assess gene expression; and tasting panels to evaluate and determine consumer preferences. The application of a classical clustering approach to integrate these kinds of heterogeneous variables for finding hidden relations would require a very complex, manual and time-consuming preprocessing to normalize each particular source of data. This should be done one-by-one, according to each particular variable and technique, being highly dependant also on the assumptions of the clustering method chosen. To the best of our knowledge, up to date there are no methods available for the integration of such highly-diverse complex data (i.e. metabolomics, transcriptomics, agronomics, tasting panels and categorical/quality assessment data) to perform an integrative analysis.

Results:

We have developed a new type of multi-modal clustering algorithm that can fuse such data but without normalization. Since direct similarities between the highly-diverse measures (variables) cannot be calculated, we propose to look at how pairs of variables influence on the clustering of the materials. If two variables, for example, a metabolite concentration and an agronomic trait, consistently produce a similar clustering of the same tomato landraces along several repetitions, then a similarity between those variables can be inferred. With this novel approach we were able to analyze and obtain clusters of diverse variables (see Fig. 1) by accumulating evidence of their variation along the complete dataset. For validation of the results, some well-known relations between the variables have been checked. For example, Guaiacol, one of the compounds with more effect on the perception of tomato scent, was clustered together with many other that share its same pathway and same precursor, such as Pinene and Limonene. Another cluster grouped several aromatic antioxidants together with Coumaric acid, which actually derives from them. Another one grouped the antioxidant capacity together with compounds that have such effect, and several other compounds for which such capacity is, at least nowadays, unknown. Thus, preliminary results revealed interesting clusters, according to agronomic traits, metabolite profiles, antioxidant properties and vitamins accumulation. In addition, we could used this novel multi-modal computational approach to explore the organoleptic properties of different landraces in order to establish further correlations between volatile content and fruit taste.

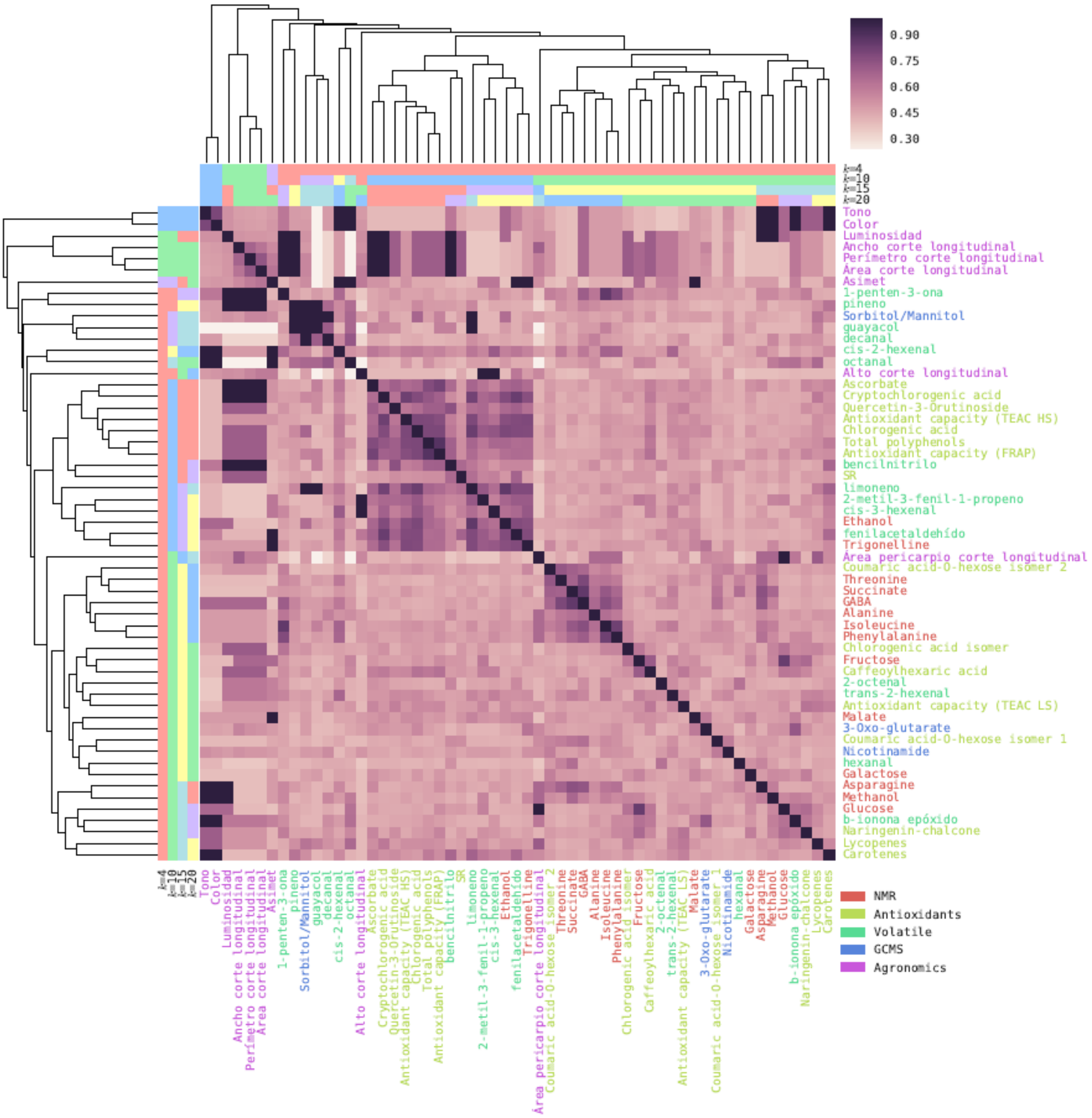

Figure 1: Highly-diverse data sources fusion with the novel multi-modal clustering algorithm.

Conclusions:

This novel approach for highly-diverse multi-modal data fusion demonstrated to have several advantages, including i) not requiring preprocessing of any of the input data to perform heterogeneous data fusion; thus enabling simple integration of categorical as well as different types of numerical data; ii) does not demand a certain (possibly different) number of replicates for each type of measure and, iii) it is specially suited for cases where highly-diverse kinds of variables (measures) have to be compared or clustered, in particular when they are not available for all the biological material under analysis. Furthermore, the new method of cluster generation and analysis based on accessions diversity and data harvested along several seasons could readily assist to infer the most probable traits to be stable inherited for germplasm selection.

Web-demo available at: http://fich.unl.edu.ar/sinc/blog/web-demo/biodatafusion/

Supported by: CONICET, MinCyT, UNL, INTA

sinc(i) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
M. Pividori, G. Stegmayer, A. Cernadas, M. Conte, F. Carrari & D. H. Milone; "A novel approach for highly-diverse multi-omics data fusionapplied to tomato germplasm selection"
4th ISCB-LA Bioinformatics Conference, 2016.