Consensus clustering from heterogeneous measures of *S. Lycopersicum*

Milton Pividori[1,2], Georgina Stegmayer[1,2], Fernando Carrari[3] , Diego Milone[2]

[1]CIDISI (CONICET), Universidad Tecnológica Nacional, Facultad Regional Santa Fe, Santa Fe, Argentina

[2]sinc(i) (CONICET), Universidad Nacional del Litoral, Facultad de Ingeniería y Ciencias Hídricas, Santa Fe, Argentina

[3]INTA-Institute of Biotechnology, Castelar, Argentina

Background

Clustering methods are key tools that aid in understanding the structure of biological datasets. However, the correct choice of a clustering algorithm requires the user to have previous knowledge about the data distributions assumed by the algorithm and also about how its parameter setting affects final results. Moreover, biological data are often composed of heterogeneous measures of the same objects, like gene expressions, metabolite profiles and other phenotypic measures. In the last years, *consensus clustering* has emerged as an attempt to solve these problems by combining a set of different clustering solutions (or partitions of the data) into a single consolidated one [1]. Plain clustering algorithms require the integration of all diverse data sources, which involves complex preprocessing of the data (like normalization). Instead, the consensus clustering method provides a simplified and high-quality approach to consider each measure type separately in a first step. Then, by using a consensus function, it combines the multiple solutions into a consensus partition, maximizing the information used from all data sources and providing better results. Several works have studied this new clustering approach and proposed different consensus functions [2–4].

Proposal

In this work, a new method based on *groups of solutions* for consensus clustering is proposed. It consists in two steps. In the first one, for each data source, clustering solutions are generated using the k-means algorithm (KM) and varying k, the number of clusters. For each k value, KM is run several times with random initializations, producing a group of solutions and thus extracting all the information from each KM configuration. This means that solutions within a group have the same number of clusters, and that all the groups have the same size. In the second step, this heterogeneous information is combined into a single clustering solution by using a supra-consensus function [1], which tries to use as much information from each source as possible. The proposed approach, named group consensus KM (gcKM), provides an extensive analysis for each data source: a wide range of different KM solutions over the same data is generated, obtaining different *points of view* for each source of data. The method also avoids the requirement of data preprocessing (normalization) and obtains better results than plain KM (pKM).

Materials and methods

Metabolite profiles, antioxidant capacity, sensory panels and volatile profiles measured in fruits and leaves from 8 different Solanum lycopersicum (tomato) accessions collected along Andean valleys of Argentina [5] were used as a source of data. Data were generated in three independent replicates, giving a total of 24 objects.

For the gcKM method, all the different data sources were analyzed separately by running KM over each one of them. The number of clusters was set in the range from 2 to 10, obtaining 9 groups of solutions. As the group size was set to 20, a total of 180 clustering solutions for each data source was generated. After that, to obtain a final result, the gcKM method combined all the clustering solutions from each data source, thus obtaining a consensus solution. Note that, at this step, the consensus function does not access the raw data, just the clustering solutions given by the groups of KM partitions.

The proposed method (gcKM) was compared with a classical approach (pKM), which consisted in running KM over the complete normalized dataset. To assess the quality of both approaches, two aspects were evaluated: (1) if accessions replicates are clustered together when solutions with 8 clusters (the same number of tomato accessions) are obtained; and (2) the amount of information used from each source of data by the final solutions of pKM and gcKM. To obtain the second measure, a set of representative partitions is first derived from each data source. Once these representatives are obtained, the mutual information between them and each final result is computed. This measure was obtained by calculating the Average Normalized Mutual Information (ANMI) [1]: $\bar{\Upsilon}(\Lambda, \Pi^{'}) = 1/M \sum_{i=1}^{M} \Upsilon(\Pi^{'}, \Pi_i)$, where M is the number of groups of solutions for each data source, and Υ is the Normalized Mutual Information (NMI) between the final partition $\Pi^{'}$ and each representative partition Π_i.

Results

For the first quality assessment, gcKM always clustered the 3 replicates of each accession together, in contrast with pKM, which failed in 70 % of the cases. The results regarding the amount of information used from each data source in the final clustering (for both gcKM and pKM) are shown in Figure 1. The top of the bars indicate the ANMI mean over 100 repetitions, together with a 95% confidence interval. At the x-axis, the number of clusters used in the final solution is shown. For example, final solutions with 4 clusters ($k = 4$) obtained by gcKM reaches an average ANMI of 0.77 when compared against the sensory panels (S. Pan.) source, whereas pKM obtains 0.57. That is, the amount of information retained by gcKM from this source is higher than the amount obtained by pKM.

It is important for the analysis to take into account the relationship between solutions with the same number of clusters over all sources of data. If a clustering solution uses more information from one data source, this may imply using less from another. Thus, it is interesting to assess, in addition to the highest ANMI values, how balanced is the ANMI obtained for gcKM and pKM solutions among all data sources. Although pKM obtains a higher ANMI under certain configurations and sources, gcKM always uses more information in at least 4 out of the 6 data sources. The pKM method tends to take into account only one or two sources of data for the final solution, whereas consensus solutions obtained by gcKM not only reach the highest ANMI values overall, but also avoid heavily preferring some data sources over another ones.

Finally, an example of a gcKM solution with $k = 4$ is shown in Figure 2. Equally shaped marks represent the 3 replicates of each accession, while each cluster is indicated with a different color (red, blue, green and purple). To be able to depict the accessions graphically, two PCA components were obtained from the complete normalized dataset. It can be seen that these clusters group tomato accessions exclusively by their metabolite complements and there is a link to the genetic groups they have been assigned to by using molecular markers [6]. For example, accessions 572/C526 and 552/C169 have been classified in two distinct genetic groups (G3 and G4), however they both were collected in the same Andean region (Catamarca). The blue cluster groups two accessions (557/C237 and 3806/ALGR) which also belong to different genetic groups, but they have similar fruit shapes.

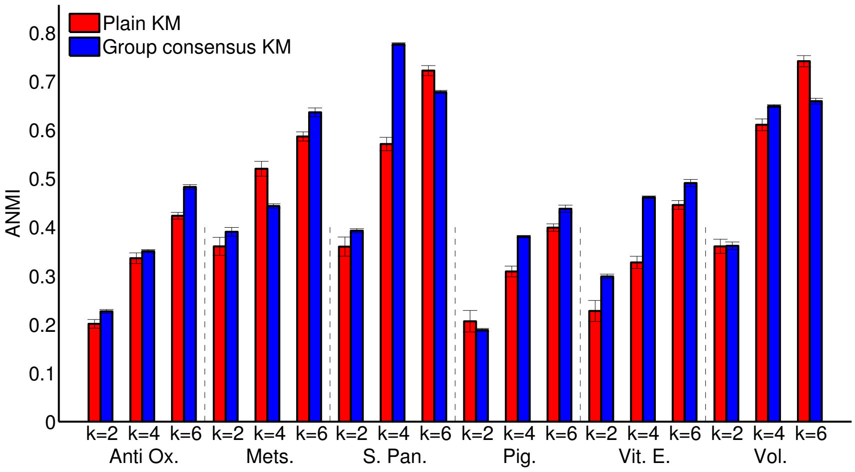

Figure 1: Mutual Information of the final solution with each data source for $k = 2, 4, 6$.

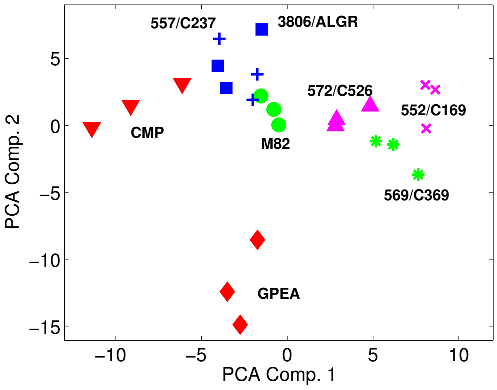

Figure 2: Example of consensus clustering solution with 4 clusters.

Conclusions

Group consensus clustering over heterogeneous biological data using groups of solutions represents an important step to make analysis easier, whilst effectively improves the quality of the solutions obtained. The gcKM method, in contrast with classical approaches like pKM, not only obtains solutions that maximize the information taken from all sources of data, no matter how heterogeneous they are, but also avoids complex preprocessing, like normalization, that are required when all diverse data sources are integrated. By producing several solutions with different configurations to finally combine them into a single consensus partition, the method also reduces the requirement for a user to know how the clustering algorithms work and which is the best set of parameters to analyze a dataset. The results obtained and the fact that tomato accessions are indistinguishable in genetic terms, suggest that it will be difficult to breed tomatoes with enhanced nutritional values if dedicated programs are not pointed to the mechanism(s) behind the low heritability values of these kinds of trait.

References

1. Strehl A, Ghosh J, Cardie C: **Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions**. *Journal of Machine Learning Research* 2002, **3**:583–617.

2. Iam-On N, Boongoen T, Garrett S, Price C: **A Link-Based Approach to the Cluster Ensemble Problem**. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 2011, **33**(12):2396–2409.

3. Yan D, Chen A, Jordan MI: **Cluster Forests**. *Computational Statistics & Data Analysis* 2013, **66**(0):178 – 192.

4. Iam-on N, Boongoen T, Garrett S: **LCE: a link-based cluster ensemble method for improved gene expression data analysis**. *Bioinformatics* 2010, **26**(12):1513–1519, [http://bioinformatics.oxfordjournals.org/content/26/12/1513.abstract].

5. Peralta I, Makuch M, García S, Occhiuto O, Asprelli P, Lorello I, Togno L: **Catálogo de poblaciones criollas de pimiento, tomate y zapallo colectadas en valles andinos de la Argentina**. *Ediciones INTA. Ed. INCA, Mendoza* 2008, pp 128, ISBN: 978-987-521-335-7.

6. Asprelli P, Occhiuto P, Makuch M, Lorello I, Togno L, García Lampasona S, Peralta I: **Recolección de germoplasma criollo de especies cultivadas y su distribución en regiones andinas de Argentina**. *Horticultura Argentina; Lugar: La Consulta* 2011, **30**(71):30 – 45.