Multivariate methods for genomic data integration and visualization

Ferran Reverter[1], José Fernández-Real[2], Esteban Vegas[1], Francesc Carmona[1], Jacques Amar[3], Remy Burcelin[3], Eduardo García Fuentes[4], Matteo Serino[3], Francisco Tinahones[4], and Alex Sánchez-Pla[1]

[1] *University of Barcelona, Spain*
[2] *Girona Biomedical Research Institute, Spain*
[3] *Institut de Medecine Moleculaire de Rangueil, Toulouse, France*
[4] *Hospital Interuniversitario Virgen de Victoria, Malaga, Spain*

Abstract

As the developments in high throughput technologies have become more common and more accessible it is becoming usual –and affordable– to take different simultaneous approaches to study the same problem. In practice this means that different sets of data of different types (expression, proteins, metabolites...) may be available or generated for the same study, highlighting the need for methods and tools to use them in a combined way.

In recent years there have been developed many methods that integrate the analysis of different types of data. Corresponding to a certain tradition in bioinformatics many methodologies are rooted in machine learning tools such as bayesian networks, support vector machines or graph-based methods. In contrast with the high number of applications from these fields, another that seems to have contributed less to genomic data integration is multivariate statistics, which has however a long tradition in being used to combine and visualize multidimensional data. In this work we discuss the application of multivariate statistical approaches to integrate bio-molecular information by combining several multivariate statistical approaches such as principal components analysis, simple and multiple correspondence analysis and canonical correlation analysis and its variants. The techniques are applied to a real unpublished data set consisting of four different data types: DGGE bands, expression microarrays, high-throughput sequence data and clinical variables.

We show how these statistical techniques can be used to perform reduction dimension and then visualize data of one type useful to explain those from other types. Whereas this is more or less straightforward when we deal with two types of data it turns to be more complicated when the goal is to visualize simultaneously more than two types. Comparison between the approaches shows that the information they provide is complementary suggesting their combined use yields more information than simply using one of them.

Keywords

Data integration, Genomic data, Visualization.

References

Hamid, J.S., Hu, P., Roslin, N.M., Ling, V., Greenwood, C.M.T., and Beyene, J. (2009). Data integration in genetics and genomics: methods and challenges. *SAGE-Hindawi Access to Research. Human Genomics and Proteomics.* doi:10.4061/2009/869093.

Serino, M., García Fuentes, E., Quipo-Ortuño, M., Moreno, J.M., Liche, E., Amar, J., Sanchez-Pla, A., Tinahones, F., Burcelin, R., and Fernández-Real, J.M.L. (2009). A specific gut microbiota genomic profile defines insulin sensitivity. Submitted.

de Tayrac, M., Lê, S., Aubry, M., Mosser, J., and Husson, F. (2009). Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: multiple factor analysis approach. *BMC Genomics.* Jan 20, 10–32.