Unsupervised multiple kernel learning to integrate various metagenomic sources

Jérôme J. Mariette, Nathalie Villa-Vialaneix

Unsupervised multiple kernel learning to integrate various metagenomic sources

Jérôme Mariette and Nathalie Villa-Vialaneix

MIAT, Université de Toulouse, INRA, 31326 Castanet-Tolosan, France

1 Abstract

In metagenomic analysis, the integration of various sources of information is a difficult task since produced datasets are often of heterogeneous types. These datasets can be composed of species counts, interaction networks or phylogenetic information. The combinations of all these types of data have been shown relevant to provide a better comparison between communities. However, standard integration methods (like PLS) can take advantage of external information but do not allow to analyse heterogenous multi-omics datasets in a generic way.

We propose to use similarity functions, called kernels, to integrate multiple datasets of various types into a single exploratory analysis. Kernels can be computed for various data types, such as numerical vectors, phylogenetic trees but also any diversity indexes. They can also be combined into a single meta-kernel. In this work, we provide several solutions to learn either a consensual meta-kernel or a meta-kernel that preserves the original topology of the datasets. This kernel is subsequently used in kernel PCA to provide a fast and accurate visualisation of similarities between samples, in a non linear space and from the multiple source point of view. A generic procedure is also proposed to improve the interpretability of the kernel PCA in regards with the original data. We applied our framework to the multiple metagenomic datasets collected during the *TARA* Oceans expedition. We demonstrated that our method is able to retrieve previous findings in a single analysis as well as to provide a new image of the sample structures when a larger number of datasets from different sources are included in the analysis.

Proposed methods are available in the R package **mixKernel**, released on CRAN. It is fully compatible with the **mixOmics** package and a tutorial describing the approach can be found on **mixOmics** web site `http://mixomics.org/mixkernel/`.