Atlas of Science
another view on science
http://atlasofscience.org

Simultaneous integrated analysis of biological datasets: an evaluation of O2PLS

The rapid progress in high throughput technology made it possible to measure biological processes at several levels: DNA markers (genetic code), gene expression (which represents the process of reading the code of a gene), proteomics (proteins are the products of a gene and are needed for biological processes), metabolomics (molecules which play a role in many different chemic reactions in the body). By combining data from the different levels researchers aim to gain deeper understanding of biological mechanisms. State of the art methods, however, do not fully explore the joint nature of these data.

For illustration we have data from 466 participants from the Finnish DILGOM study. Here two biological aspects were measured: gene expression (6272) and metabolites (137). A straightforward approach to analyse the data is pairwise: all combinations of metabolites and gene expressions are considered at a time. However there are many pairs (more than 850k) and joint relationships (several genes related to multiple metabolites) might not be recovered. Integrative analysis of all measurements from all datasets (i.e. *simultaneous* data analysis, Fig. 1) are more likely to give an insight across the datasets and hence about the underlying biological processes. We aim to find parts of two datasets which are highly connected. To find these parts we use the O2PLS method. O2PLS constructs the joint part of the two datasets, and the remaining part consists of data-specific informative part and noise. Thus we end up with joint, metabolite-specific and gene-specific information in the data. Inferring how genes and metabolites are related, while separating the related from the unrelated part, is the aim of the paper.

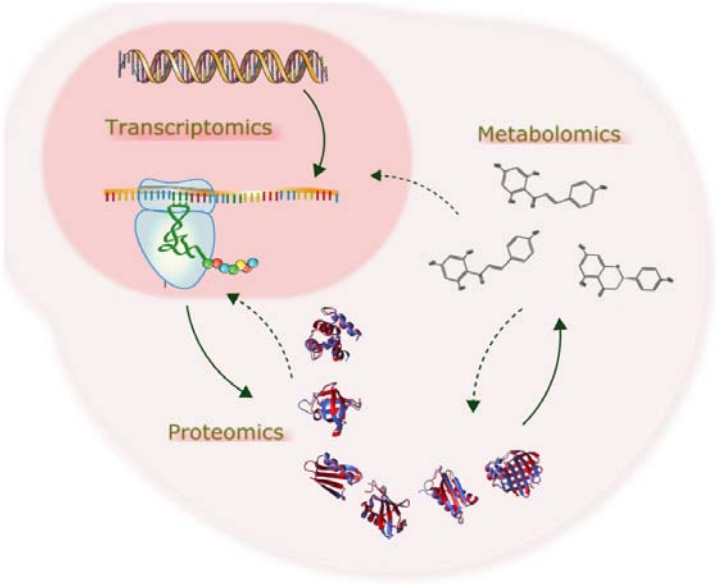

Fig. 1. Graphical description of the complex relationships between several levels of biological processes.

The parts found by O2PLS are combinations of gene expressions and metabolites. Specifically, O2PLS *simultaneously* assigns a value to each gene and metabolite indicating its importance to the joint or specific part. Large positive or negative values indicate large contribution to the corresponding part. The most important genes and metabolites in the joint part can be further investigated to understand the relationship between gene expression and metabolite concentration. The specific parts may also be interpreted by looking at the top genes and metabolites. The amount of information of each part can be quantified by its variation relative to the total variation in the corresponding dataset.

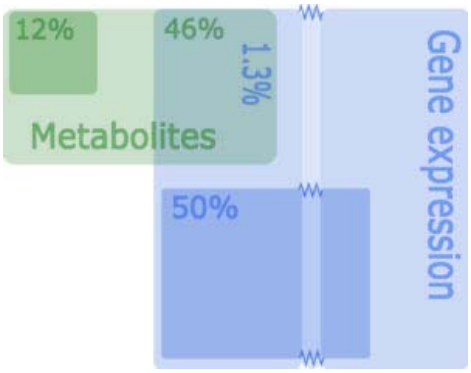

Fig. 2. Visualisation of the results of the data analysis done with O2PLS. The datasets are decomposed in an overlapping part, a data-specific part and a remaining part. The percentages indicate the variation of each part relative to the total amount of variation.

We used the measurements on gene expression and the abundance of several metabolites. Regarding the metabolites we found that 46% of the total information was in the joint part, while 12% was in the metabolite-specific part. Regarding gene-expression we found that 1,3% of the total information was in the joint part, while 50% was in the gene-specific part. The O2PLS results are visualized in Fig. 2. These results confirm the former findings based on pair-wise analysis. In addition we found interesting other genes for future research.

To conclude, O2PLS is a promising tool for summarizing information from two datasets. However the current status in biology is DNA markers, methylation, proteomics, in addition to gene expression and metabolites. These data are heterogeneous: each dataset represents a different layer of the biological mechanisms and these data are generated by different measurement techniques. Due to this heterogeneity it is highly important to model the data-specific information correctly. Ignoring this might lead to failure of recovering the joint relationships. To gain better understanding, integrated analysis should be performed of available datasets. O2PLS can be the starting point for developing such methods.

Said el Bouhaddani [1], ***Jeanine Houwing-Duistermaat*** [1,2], ***Geurt Jongbloed*** [3], ***Hae-Won Uh*** [1]
[1]*Dept of Medical statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands*
[2]*Dept of Statistics, University of Leeds, Leeds, United Kingdom*
[3]*Dept of Applied mathematics, Delft university of technology, Delft, The Netherlands*

Publication

Evaluation of O2PLS in Omics data integration.
Bouhaddani SE, Houwing-Duistermaat J, Salo P, Perola M, Jongbloed G, Uh HW
BMC Bioinformatics. 2016 Jan 20