Multi-omics data analysis using sets of variables: many needles in multiple haystacks

Renée Menezes[1]

[1]Netherlands Cancer Institute · Biostatistics Unit, Amsterdam.
E-mail for correspondence: r.menezes@nki.nl

Abstract: Understanding of molecular regulation mechanisms is important to better understand how processes in the human body occur, in particular in disease onset and development such as that of cancer. To help with this understanding, studies gather increasing amounts of molecular profiling data, such as DNA (copy number, methylation), mRNA and protein profiles. Such studies could help us better understand how often a gene dosage effect (change in copy number) occurs at the same time as a gene silencing one (hypermethylation). However, the analysis of all these data still represents a challenge. Often researchers still choose to study the association between pairs of features, for example the expression of one gene together with the copy number status at the beginning of the gene. Such pairwise analyses not only require stringent multiple testing correction (as the number of pairs of features is very large), but also cannot easily incorporate multiple omics datasets simultaneously.

We propose to use an approach that enables testing of the effect of a large number of variables on one response. By focusing on testing, no estimation of parameters is required, what makes this approach applicable to problems with $p \gg n$. We have applied this approach to studying the association between copy and gene expression in colon and breast cancer, uncovering interesting patterns that better characterize differences between these two cancer types. We have also extended the approach to explain gene expression by multiple omics data, for example by both copy number and methylation. In this way, we have uncovered genes which, when in a region with copy number gain, are silenced via methylation. We have also uncovered genes that achieve overexpression in different ways: in some samples by DNA copy gain, in others by hypomethylation. Another extension involved testing for spliceQTL, which required using multiple responses: here we tested for association between counts for all individual exons of a given gene, and the number of minor alleles of SNPs located between the start and end of the gene. By considering multiple relevant features at the same time, results yielded by this approach are more often replicated than those obtained by pairwise testing.

In conclusion, we have proposed efficient approaches to perform joint analyses of multi-omics datasets. By focusing on testing, rather than on estimation, our approaches can be used in $p \gg n$ problems without incurring very stringent multiple testing. In addition, the focus on testing helps further with separating signal from noise, a recurring problem in high-dimensional data analysis. Furthermore, by enabling the use of variable sets both as "responses" as well as "explanatory", we can study complex problems such as spliceQTL. Finally, by considering more data at once, results yielded are more often replicated than low-dimensional approaches.