Proceedings of the LXII SIGA Annual Congress
Verona, Italy – 25/28 September, 2018
ISBN **978-88-904570-8-1**

Poster Communication Abstract – 6.30

QUALITY CONTROL ON THE MEDICAGO MICROARRAY DATABASE TO IMPROVE PREDICTIONS OF GENE FUNCTION

WANG C.*, PAVESI G.*, SAIA S.**, MIZZI L.*, MORANDINI P.*

*) Dept of Biosciences, University of Milan, Via Celoria 26, 20133 Milano (Italy)
**) Council for Agricultural Research and Economics (CREA), Research Centre for Cereal and Industrial Crops (CREA-CI), S.S. 11, Km 2,5, 13100 Vercelli (Italy)

Medicago sativa, correlation analysis, coexpression

Microarray-based technologies permit the simultaneous measurement of expression levels across an entire genome. Their widespread application has led to the production, in several species, of hundreds to thousands datasets obtained from different experiments, tissues and treatments. Thus, this huge amount of data can be mined for example to predict mutual functional relationships among genes, by using 'guilt by association'-like approaches. There is however little control over the quality of the data stored in the various microarray databases, weakening their predictive power and hampering integration among them and with data from other sources (e.g. RNAseq). We devised a strategy for data quality control based on logical and statistical relationships among parameters and conditions applied to the Medicago microarray database, that contains around 700 different measurements based on the Affymetrix technology. Out of 715 different hybridizations, we first removed a group of 24 that were duplicated and a further group of 6 lacking data. We then identified and deleted 'problematic' samples or hybridizations using two indicators: a) sum of expression levels over the entire transcriptome probed and b) Pearson correlation coefficient, to detect replicate experiments with too large variation in expression values. Using these criteria, around 10% of hybridizations resulted to be inconsistent or containing poor quality data. The resulting reduced dataset showed a substantial improvement in the consistency of the predictions using genes of known function (ribosomal, photosynthetic…etc.) as tests, thus allowing the reduction of both false positives (e.g. genes that, using the complete database, were erroneously predicted to be highly correlated with other genes or processes) and false negatives (genes which, using the entire database, showed poor or inconsistent correlation, but improved substantially after polishing). Examples of identified false positives are many transcription factors of the bHLH and WRKY families, while examples of false negatives are genes in metabolic pathways (e.g. saponins and Krebs cycle). Some of the false negatives were validated by measuring the enrichment in Gene Ontology terms of the best correlators before and after polishing.

Since the approach is of general validity (i.e. it is not limited by the species under examination), we are extending the analysis to other plant microarray databases.