Analysis and detection of correlations in High-throughput transcriptional assays

Andrés Rabinovich[1], Maximiliano Beckel[1], Ariel Chernomoretz[1,2]
[1]Integrative Systems Biology Group, Fundación Instituto Leloir, Capital Federal, Argentina.
[2]Department of Physics, FCEyN, UBA/IFIBA

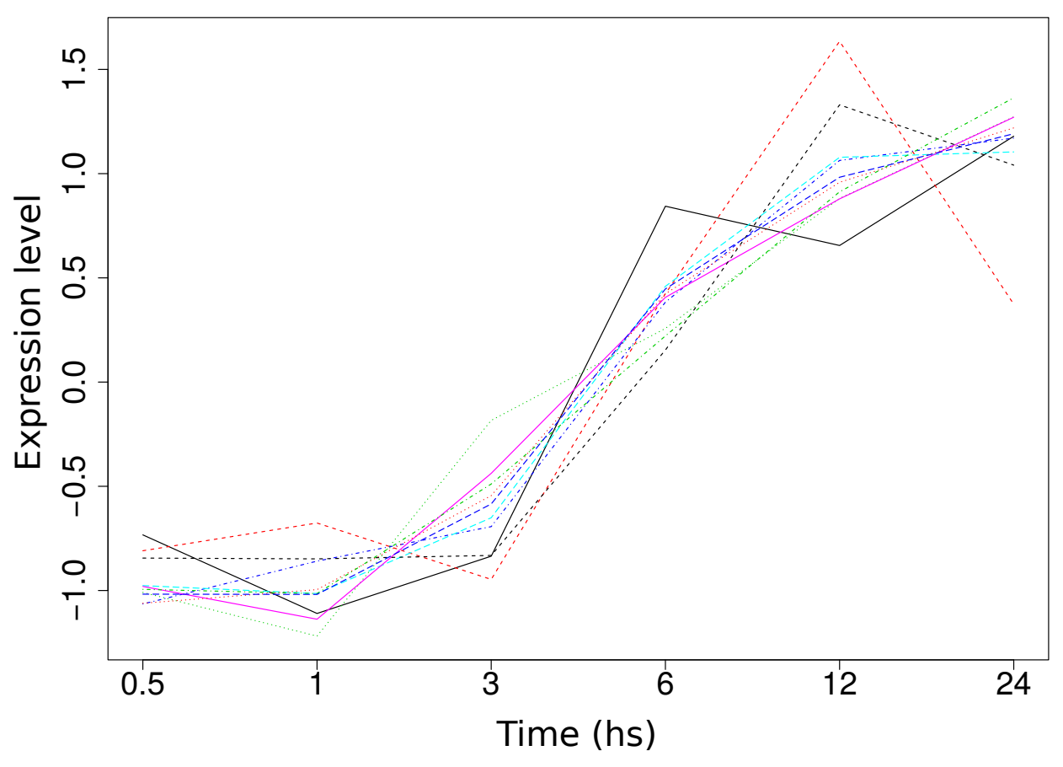
Abstract

Clustering techniques are used to unveil common patterns of gene expression across different samples. Genes with similar expression profiles could have related functions or could be regulated by common mechanisms. Thus, detection and analysis of gene correlations can reveal unknown meaningful biology. Different clustering methods produce descriptions at different resolutions. Therefore an external biological knowledge space (GO) must be incorporated in order to quantify the information obtained with each method. The aim of this study is to obtain biologically homogeneus groups of genes with highly correlated expression profiles for several stress treatments on *Arabidopsis thaliana*. This was accomplished by constructing a transcriptional network and using the information contained on local neighborhoods.

Similarity and clustering

Central to clustering analysis is the concept of similarity, that quantifies the strength of the relationship between two elements of a set.
A sensible way of defining similarity between genes A and B in gene-expression space is the Pearson correlation coeficient, with a distance given by:

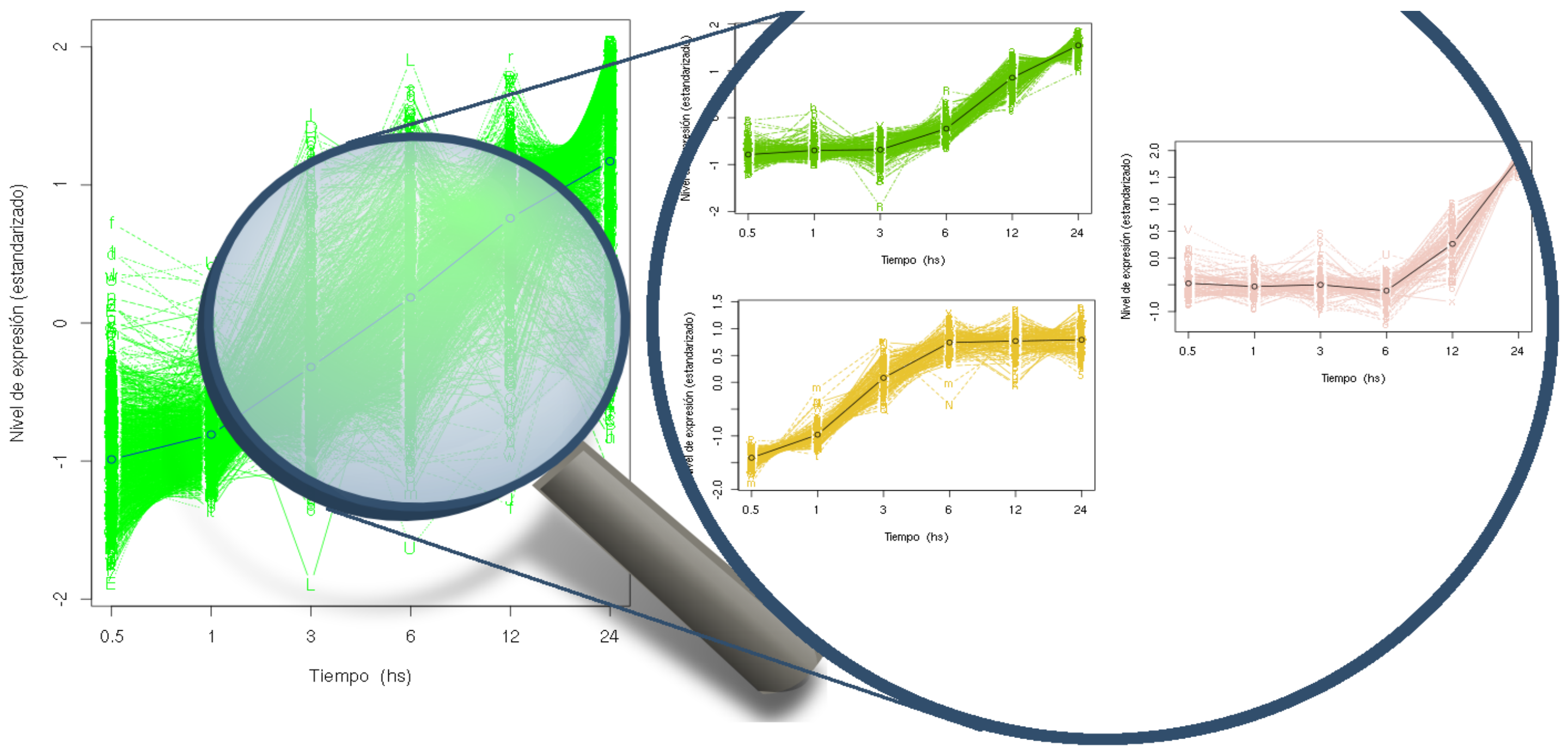
$$d_{ccp}(A,B) = 1 - cor(A,B)$$



Once a distance is chosen, different clustering procedures can be used. Each method will produce a different description based on the resolution of the method. Therefore, clustering is considered an ill-posed problem.
In order to find out if an optimal scale on a biological sence exists, external information (GO) must be incorporated.

Clusters in gene-expression space

We analyzed a microarray dataset for *A. thaliana* [1] using two well known clustering algorithms: k-means and dynamic tree cut with two different granularities (DS1: medium granularity, DS4: large granularity)[2].



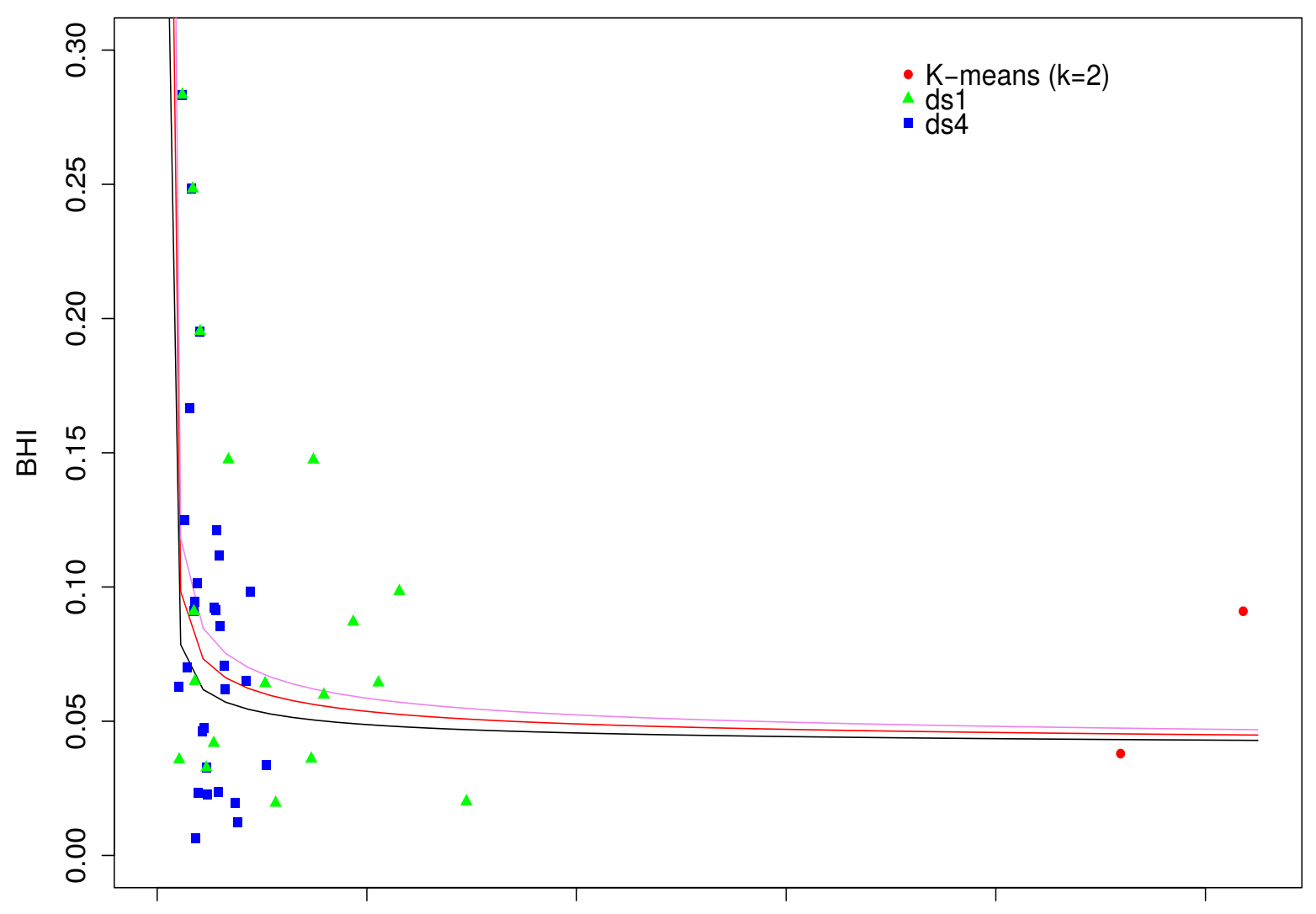
For the various treatments, k-means found few large clusters. Biological interpretation of such large structures is almost impossible.
Dynamic tree cut, on the other hand, finds several clusters of varying size, effectively finding substructures inside the structures found by k-means.
Clusters found with these methods have a high mean correlation, between 0.75 and 0.95.

Biological homogeneity index

It is a way of quantifing the biological congruency of the clusters[3]:

$$BHI_j = \frac{1}{n_j(n_j-1)} \sum_{x \neq y \in D_j} I(C(x) = C(y))$$

In order to improve the biological homogeneity of the clusters, a mixed metric can be defined to simultaneously mine for correlations in gene expression and gene ontology spaces.



References
[2] P. Langfelder et al. Dynamic Tree Cut : in-depth description , tests and applications (2007) 1.
[1] Kilian et al. The AtGenExpress global stress expression data set: Protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. Plant Journal 50 (2007) 347.
[3]S. Datta. Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. BMC bioinformatics 7 (2006) 397.
[4] Ashburner. Gene ontology: tool for the unification of biology. Nat Genet 25 (2000)
[5] C. Pesquita et al. Semantic similarity in biomedical ontologies. PLoS Computational Biology 5 (2009).
[6] N. Cristianini et al. On kernel target alignment. Studies in Fuzziness and Soft Computing 194 (2006) 205.
[7] Beissbarth,T. and Speed,T.P. GOstat: find statistically over-represented Gene Ontologies within a group of genesBioinformatics, 20, 1464–1465 (2004)

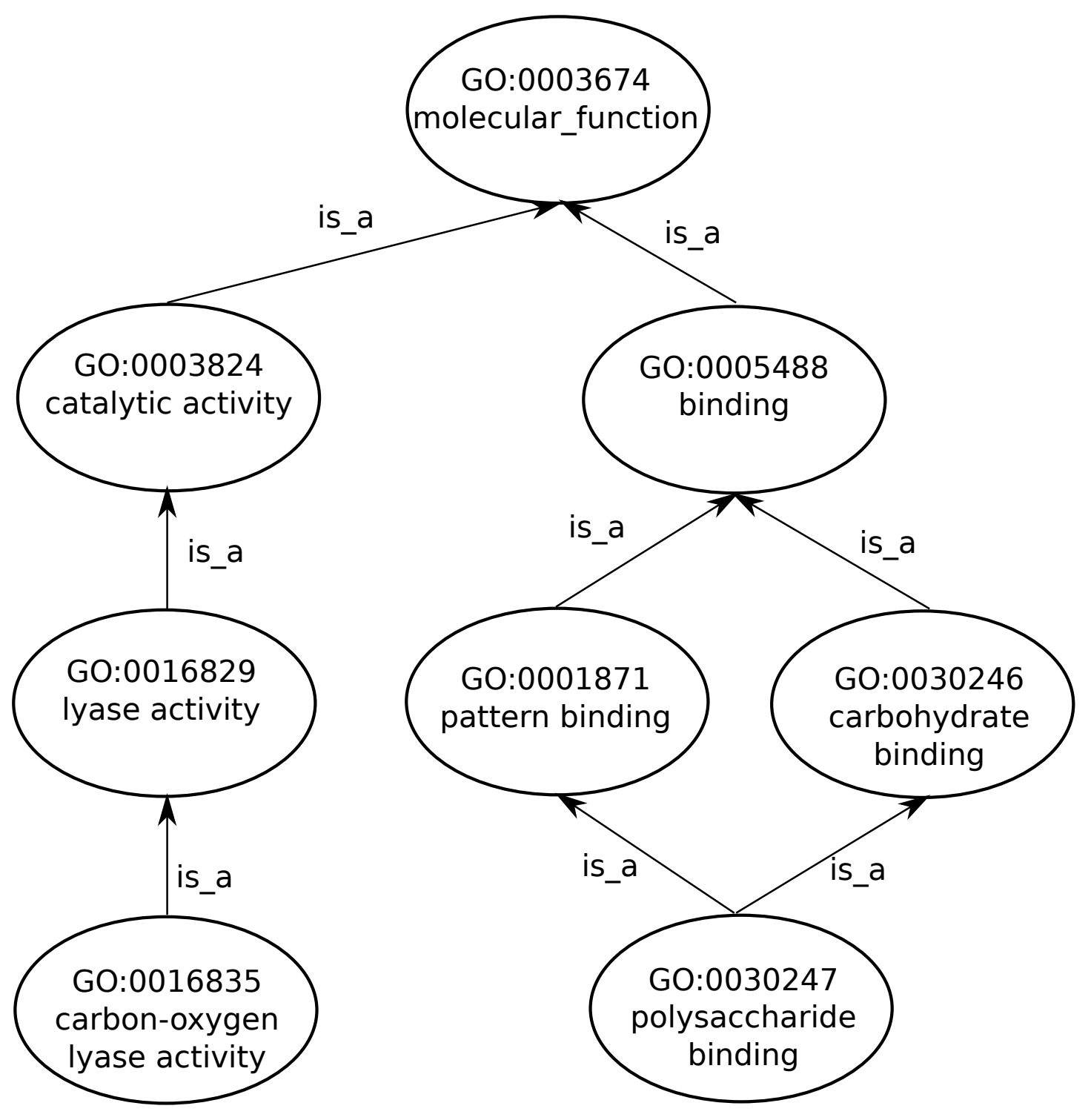
Contact: Andrés Rabinovich <arabinovich@leloir.org.ar>

Gene ontology

Provides a controlled vocabulary to describe gene atributes[4].
Genes are mapped to the most specific GO category.
There are different ways of defining distances between genes g_1 and g_2 on GO space. In particular we use[5]:

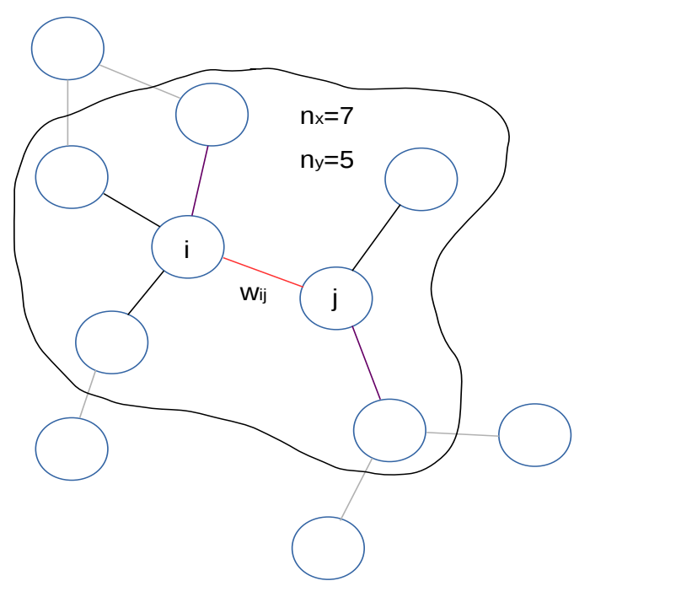
$$Sim_{rcmax}(GO(g_1), GO(g_2)) = \max\left(\frac{1}{N}\sum_i \max_{1\le j\le M} S_{ij}, \frac{1}{M}\sum_j \max_{1\le i\le N} S_{ij}\right)$$

With S_{ij} the common shared information of terms i and j, deffined as the information content of the most informative common ancestor .



Local Neighborhoods

We built a 30 mutual nearest neighbor network using gene expression similarity. For every edge we can define a local neighborhood and quantify the degree of agreement between the two spaces with a local version of KTA



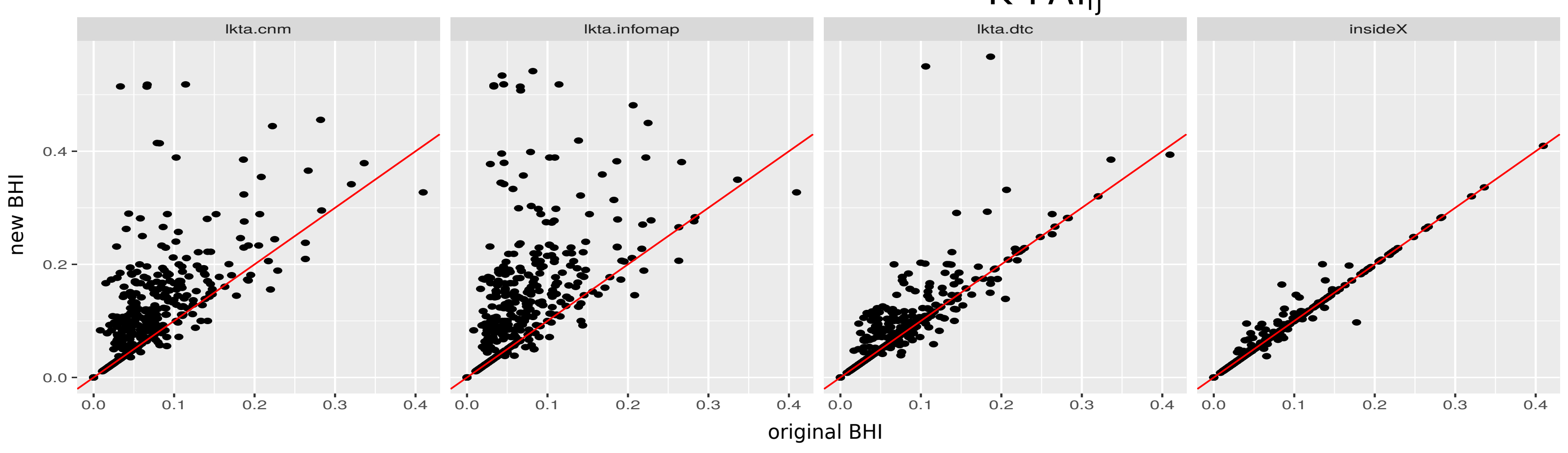
Kernel Target Alignment

KTA quantifies the degree of agreement between two spaces. We expect expression space and GO space to be different but not orthogonal. It is defined for kernels K1 and K2 as $KTA(C,k_1,k_2) = \frac{\langle K_1,K_2\rangle_F}{\sqrt{\langle K_1,K_1\rangle_F \langle K_2,K_2\rangle_F}}$

With $\langle K_1,K_2\rangle_F = \sum_{i,j=1}^{m} K_1(x_i,x_j)K_2(x_i,x_j)$ the Frobenius inner product [6].

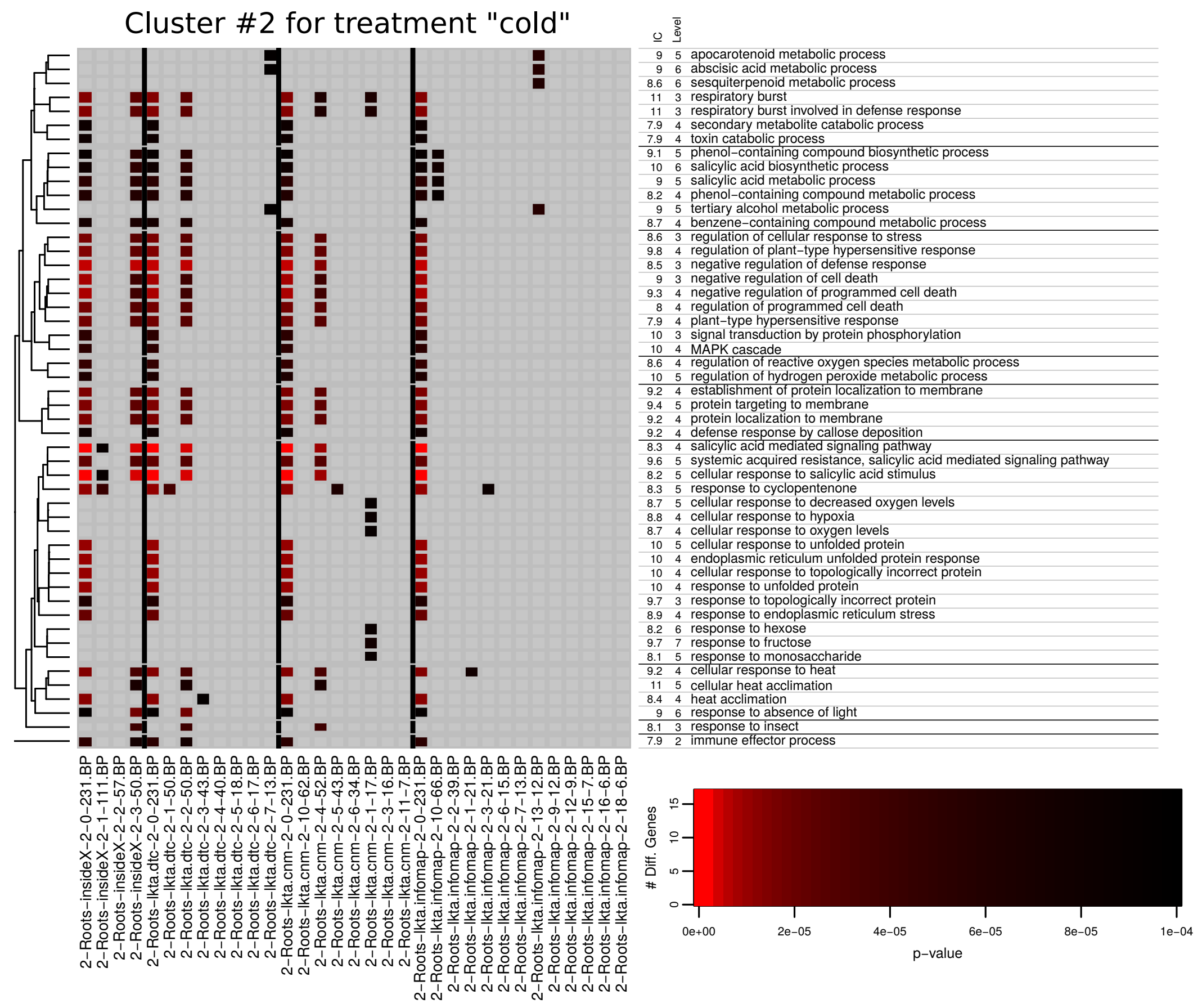
Mixed metrics

Given an edge and the average weights of the GO neighborhood, we can quantify the biological coherence of that neighborhood. In order to find transcriptional clusters with increased biological coherence, we can modify the original local transcriptional weight Wij between edges i and j with this information: $w_{ij} = simcor_{ij}^{\beta * stress_{ij}}$ With $stress_{ij} = \frac{KTA_{background}}{KTAl_{ij}}$



Statistical over-representation test

Is a usual way of detecting enrichment of gene annotations in clusters[7].



In most cases, the heuristics we developed using mixed metrics were able to detect substructure inside the clusters. For example, in cluster #2 for treatment "cold", a medium size subcluster explains most of the enrichment of the original cluster.
This also shows that it is hard to assign biological meaning to small clusters, even when they are biologically homogeneus.

Conclusions

A mixed metric induced by local KTA in network neighborhoods in both expression space and GO space was presented, along with three heuristics.
This methods allow for the simultaneous mining of meaningful structure in both expression and functional spaces, finding highly correlated clusters with increased biological homogeneity. A statistical over-representation test was used in order to detect enrichment of gene annotations in clusters and subclusters, with medium sized subclusters explaining most of the enrichment of the original clusters.