

# Identification and visualisation of drug mode of action from molecular signatures using an unsupervised classification approach

Francesco Iorio, Diego Di Bernardo

Systems Biology Group,  
Telethon Institute of Genetics and Medicine - Naples, Italy  
{iorio,dibernardo}@tigem.it

Identifying pathways mediating a drug mode of action is a key challenge in biomedicine. We demonstrated that using gene expression profiles in yeast, it is possible to detect the mode of action of a drug candidate[4]. Recently Lamb et al[1] developed a large public database of expression signatures of drugs and genes, called Connectivity Map (CMap). This is a collection of gene-expression profiles from cultured human cells, treated with 164 compounds at different concentration levels for a total of about 500 profiles.

In this work we focus on a classification-based approach that, starting from the expression profiles of compounds in the CMap, identifies those compounds whose mode-of-action (MOA) is the most similar to that of a drug candidate of interest with unknown MOA.

The first step is to represent each gene expression profile in the CMap as a point in a metric space, making use of appropriate distance metrics (see below). In this space the distance between points is proportional to the similarity between the compounds the points correspond to.

In the second step, a *clusterisation* of the dataset is performed in the considered metric space. We assigned to each point, i.e. each gene expression profile, a set of “membership scores” quantifying how much a profile “belongs” to each of the clusters. The membership scores are computed by means of random projection techniques [3].

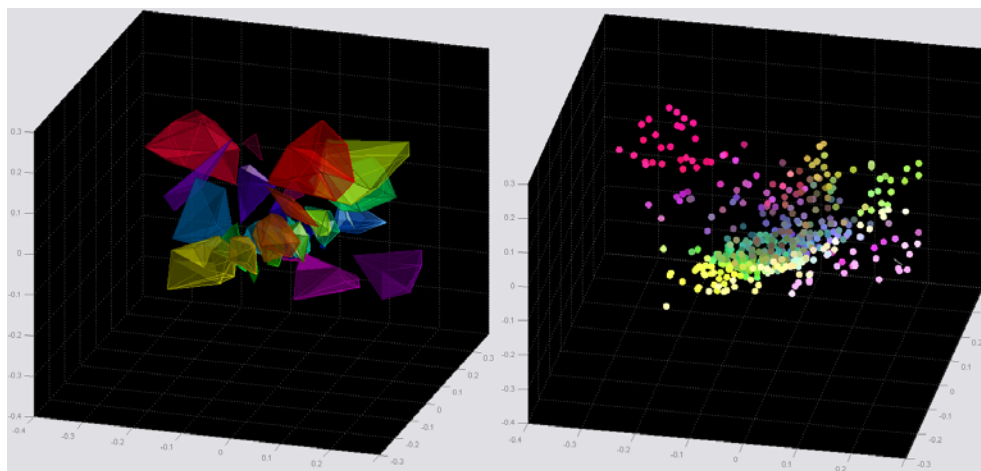


Figure 1: Gene expression profiles clustered using euclidean distance in one of the tested metric spaces (a) and painted according their membership function values (b).

Combining clustering with a visualization tool we developed, it is possible to explore the clustering space and the similarities between different drugs. In the left side of Figure 1 an example of visualisation via cluster analysis is shown. In the right side of the same figure the gene expression profiles are coloured according their membership scores.

In the third step, to identify similarity between a drug candidate and the compounds contained in the reference dataset, we embedded the gene expression profile obtained using the drug candidate in the metric space. Its neighbouring points can be observed and, on the basis of their composition, the new drug can be classified, and a set of membership scores can be assigned to it.

To implement this classification approach, we tested a variety of distance metrics based on gene rankings[2] in order to find those that best captured similarities due to the drug effect and discarded similarities due to other sources. In order to test step 3, we chose a subset of the CMap dataset as a test set, and considered, for each expression profile in the test set, the closest N neighbours as positive predictions.

On the basis of the labels of the closest neighbours, the predictions have been assessed and performances have been compared. We considered the positive predicted points True Positives (TP) if their labels have the same label of the test profile (i.e. profiles of the same drug but applied to a different cell line, or at a different concentration or timing).

Results of this comparison are shown in the figure 2.

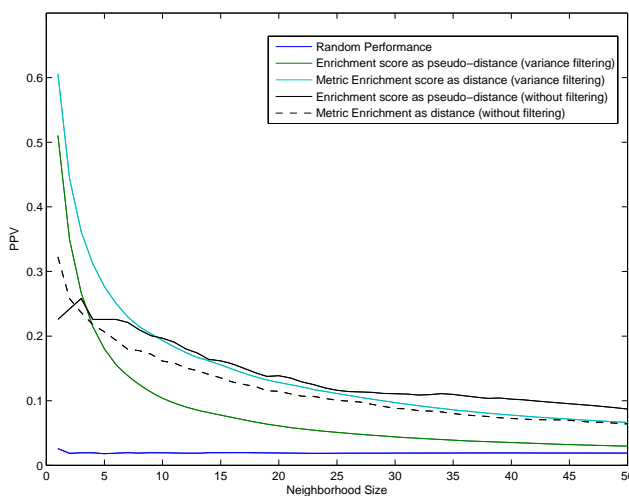


Figure 2: Classification results. Test performed on a subset of the cMap dataset considering neighbours as positive predictions and the profiles sharing the same drug labels as true positive. We used the enrichment score (ES), proposed by Lamb et al[1], as a pseudo-distance (solid black line), or a novel *metric version* of the ES as distance we developed (dashed black line), with and without a variance filtering in order to select only those genes that varied the most across the different treatments. The best performance are provided by the variance-filtered metric ES.

We introduced a classification based approach allowing investigations on the mode of action of new drugs. Our analysis approach overcomes the problems of the experimental conditions influence and allows the use of microarray data without preliminary normalization procedures.

## References

- [1] J. Lamb et Al. *The Connectivity Map dataset: Using gene-expression signatures to connect small molecules, genes and diseases*. Science - September 29, 2006, Vol. 313, 1929-1935.
- [2] A. Subramaninan et Al. *Gene set enrichment analysis: A knowledgebased approach for interpreting genome-wide expression profiles*. PNAS - October 25, 2005, Vol. 102 n. 43, 15545-15550.
- [3] A. Bertoni et Al. *Random projections for assessing gene expression cluster stability*. Proceedings of IJCNN 2005, The IEEE-INNS International Joint Conference on Neural Networks, Montreal, 2005.
- [4] di Bernardo et Al. *Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks*. Nature Biotechnology - 2005, 23(3):377-83.