A fuzzy cluster analysis model for mining the cMap dataset to investigate common drug modes of action

ID - 124

Iorio Francesco[1], Di Bernardo Diego[1]
[1]Telethon Institute of Genetics and Medicine

Motivation

The Connectivity Map (also know as cMap) is a reference collection of gene-expression profiles from cultured human cells, treated with bioactive small molecules. In [1] the authors shown how to use this resource to find connections among small molecules sharing a mechanism of action and left some open questions. To identify small molecules with similar effects, on the basis of gene-expression profiles, in [1] is shown a nonparamentric, rank-based pattern matching strategy based on Kolmogorov-Smirnov statistic [2] . It is used to produce an Enrichment Score [3], for each profile in the set, that quantify how much each of them is similar to a query signature. A query signature is any list of genes whose expression is correlated with a biological state of interest. Each gene in the query signature carries a sign, indicating whether it is up regulated or down regulated. The profiles are rank-ordered according to their expression values. The query signature is compared to each rank-ordered list to determine whether its up regulated genes tend to appear near the top of the list and its down regulated genes near the bottom (positive connection) or vice versa (negative connection), yielding a connectivity score ranging from +1 to -1. All instances in the database are then ranked according to they connectivity scores; those at the top are most strongly correlated to the query signature, and those at the bottom are most strongly anticorrelated.

Our goal consists in making cluster analysis on profiles contained in the cMap, providing a complete portfolio of small molecules behaviors and discovering how many groups composed by (reasonably) similar elements does the cMap contain. After this first step, the clusters obtained can be used to find connection between an external signature and the profiles clustered by the evaluation of a cluster membership function.

Methods

In agreement with [1], we verified that clustering procedures using classical distance metrics detect dominant clusters related to cell types. In order to discover structures related to the molecule mode of action (MOA), we adopt an approach that, starting by the GSEA [4] definition of enrichment score, will get a novel distance metric that highlights similarities of MOA rather than cell types. We plain to use this metric to check the similarity between all gene profiles pairs in the set building, for each profile, a query signature taking in account only the significant up regulated genes and the significant down regulated ones. The ranking lists obtained for each profile will then be used to build a fuzzy similarity matrices over all them, on which fuzzy clusters can be obtained and membership functions can be evaluated. This will allow MOA identification of small molecules. Visualization techniques using spherical surfaces in several different data-transformation domains will be developed and will make clustering results human-eye understandable.

Results

Email: iorio@tigem.it