Genes clustering on large, mixed microarray data sets

Tulipano A (1,2), Marangi C (4), Angelini L (3), Pellicoro M (3), Donvito G (2), Maggi G (2), Gisel A (1)

(1) CNR, Istituto Tecnologie Biomediche Sezione di Bari, via Amendola 122/D, 70126 Bari (Italy)
(2) INFN Sezione di Bari, via Amendola 173, 70126 Bari (Italy)
(3) Dipartimento Interateneo di Fisica, Università di Bari, via Amendola 173, 70126 Bari (Italy)
(4) CNR, Istituto per le Applicazioni del Calcolo Sezione di Bari, via Amendola 122/D, 70126 Bari (Italy)

Motivation

Every single microarray data set is an image of the transcriptional level of ten thousands genes during a specific biological experiment at a specific time. An increase in the number of data sets within the biological experiment multiplies the number of images and forms an overall picture of the processes monitored during the biological experiment. Every picture contains information of some general processes found also in other biological experiments and some information on some processes which are specific for that biological experiment. Public databases such as GEO (http://ncbi.nlm.nih.gov/geo) and ArrayExpress (http://www.ebi.ac.uk/arrayexpress) contain a large amount of such biological experiments including several data sets. Those large numbers of data sets can be used to monitor and differentiate general processes and specific processes within such biological experiments.

Methods

As a first test we were downloading from GEO a data collection derived from the Affimetrix microarray design 'Human Genome U133 Array Set HG-U133A' all normalized by MAS 5.0. The total number of data sets we included in our analysis is 587 covering more than 20 biological experiments. In order to have a comparable set of data, we scaled each data set point by means of a global normalization, doing a logarithmic transformation on it and setting the median of the values distribution of each microarray experiment to zero. For the global clustering analysis we have chosen a clustering algorithm based on the cooperative behaviour of an inhomogeneous lattice of coupled chaotic maps, the Chaotic Map Clustering (CMC, Angelini et al. 2000). A chaotic map is assigned to each data point and the strength of the coupling between pairs of maps is a decreasing function of their distance. The mutual information between pairs of maps, in the stationary regime, is then used as the similarity index for clustering the data set. For our analysis we set the parameter for the cluster resolution high enough to observe clusters as stable as possible. Running CMC under such stringent conditions on the whole set of 587 microarray experiments generated few clusters containing no more than 40 genes per cluster.

Results

From the analysis of the members of each cluster by the Gene Ontology (http://www.geneontology.org/) it is clear that those clusters contain genes representing biological processes very general, such as metabolism of different compounds, different transport processes, RNA processing, but also clearly different among each other. Limiting now the analysis to a subset of biologically similar experiments, we find the same clusters as in the global analysis in addition to some new and therefore specific clusters for that subset of microarray experiments. In this way, by choosing different subsets of microarray experiments, we are able to assign to each subset specific biological processes and find new annotations for genes little annotated within those clusters. Such an analysis takes advantage of the large information within those high throughput data publicly available to improve the knowledge of every single gene represented in those data set.

Contact email: angelica.tulipano@ba.infn.it

References

- Angelini L., De Carlo F., Marangi C., Pellicoro M. and Stramaglia S., 2000 Clustering data by inhomogeneous chaotic map lattices. Phys. Rev. Letters 85(3); 554-557.