Chaotic map clustering on large mixed microarray data sets: modularity and coupled two-ways

ID - 206

Tulipano Angelica[1,2], Marangi Carmela[3], Angelini Leonardo[4], De Sario Giulia[1], Maggi Giorgio[2], Gisel Andreas[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]CNR, Istituto per le Applicazioni del Calcolo Sezione di Bari, via Amendola 122/D, 70126 Bari (Italy)

[4]Dipartimento Interateneo di Fisica, Università di Bari, via Amendola 173, 70126 Bari (Italy)

Motivation

Microarray data are a rich source of information because they contain the expression values of thousands of genes and in addition, especially in public repositories, hundreds of experiments with the same array design are available. Comparing expression levels over a wide range of experiments can reveal new and valuable information about behaviours of genes. Furthermore, because of the vast amount of experiments available, technical errors can be filtered out. Clustering is a good and challenging analysis method for data sets of such size and complexity.

Methods

A collection of Affimetrix microarray, Human Genome U133 Array Set HG-U133A, were selected and 587 data sets covering more than 20 biological experiments were included. Comparing data sets of different experiments requires an adjustment of the data. Each data set point was scaled by means of a global normalization, doing a logarithmic transformation on it and setting the median of each microarray experiment to zero. We have chosen a hierarchical clustering algorithm based on the cooperative behaviour of an inhomogeneous lattice of coupled chaotic maps, the Chaotic Map Clustering [1]. 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. To verify the optimal level of the partition we use the modularity [2] as a measure of the quality of the division in clusters of a hierarchical level. This quantity measures the number of connections between elements of the same cluster minus the expected value of the same quantity in a set of data points with the same cluster divisions but random connections between the elements, weighting each connection with the mutual information. We calculated the modularity for each level and looked for its peak.

The clustering analysis of the expression matrix D = ng x ns, where ng is the number of genes and ns is the number of samples (experiments), was performed in two ways.

The first way considers the samples as the objects to be clustered, with the ng levels of expression playing the role of the features, representing each sample as a point in a ng-dimensional space, grouping the samples with similar expression profiles. The other way considers the genes as the objects to be clustered, as measured over all of the samples, as a point in a ns-dimensional space, grouping genes acting correlatively on different samples. The analysis was drastically improved with a two-way coupled clustering approach [3]. Using the groups of genes and samples obtained with the two-way clustering, this method identifies submatrices of the total expression matrix, whose further clustering analysis reveals partitions of samples (and genes) into biologically relevant classes. By focusing on small subsets we lowered the noise induced by the other samples and genes and we were able to discover partitions and correlations that were masked and hidden when the full dataset was used in the analysis.

Results

As efficiency test we applied our method to a set of well classified microarray expression data [4]. Varying the resolution of the partition we obtained an efficiency range of 80%-87% and considering all the hierarchical levels of clusters we always observed the correspondence maximum value of efficiency-maximum value of modularity. Analyzing our 587 data sets performing the two-ways approach we were able to retrieve stable groups of genes. By means of statistical analysis of the expression values of the stable clusters we clearly identified groups of genes over-expressed in small groups of samples. Using the biological knowledge of the gene ontology annotating the involved genes, we could show, applying a Fisher exact test, that each of the clusters have a set of over-represented functionalities and in most of the cases also clearly different functionalities from cluster to cluster. With this work we demonstrated that a coupled two-way approach using the CMC as clustering algorithm classifies data sets from different experiments in groups of specific biological functionality. This results can be used to learn more about the involved gene products analyzing all the applied experiments but also to improve the annotation quality of each gene product within a cluster.

Email: angelica.tulipano@ba.infn.it