FCA AS A MEANS FOR CONSOLIDATION OF CLUSTERING RESULTS DERIVED FROM MULTIPLE EXPERIMENTS

VESELKA BOEVA

Computer Systems and Technologies Department
Technical University of Sofia-branch Plovdiv
Tsanko Dyustabanov 25, 4000 Plovdiv, Bulgaria
vboeva@tu-plovdiv.bg

ANNA HRISTOSKOVA

Department of Information Technology
Ghent University - IBBT
Gaston Crommenlaan 8 (201), 9050 Ghent, Belgium
anna.hristoskova@intec.UGent.be

ELENA TSIPORKOVA

Sirris, ICT and Software Engineering group
The Collective Center for the Belgian technological industry
Brussels, Belgium
elena.tsiporkova@sirris.be

ELENA KOSTADINOVA

Computer Systems and Technologies Department
Technical University of Sofia-branch Plovdiv
Tsanko Dyustabanov 25, 4000 Plovdiv, Bulgaria
elli@tu-plovdiv.bg

Gene clustering is one of the most important top-down microarray analysis techniques when it comes to extracting meaningful information from gene expression profiles. Clustering algorithms are used to divide genes into groups according to the degree of their expression similarity. Such a grouping indicates that the respective genes are correlated and/or co-regulated, and subsequently indicates that the genes could possibly share a common biological role.

Presently, with the increasing number and complexity of the available gene expression data sets the combination of data from multiple microarray studies addressing a similar biological question is gaining high importance. The integration and evaluation of multiple datasets yields more reliable and robust results since they are based on a larger number of samples and the effects of the individual study-specific biases are diminished. One useful way for integration analysis of the data from different experiments is to aggregate their clustering results into a consensus clustering which both

emphasizes the common organization in all the datasets and at the same time reveals the significant differences among them.

In this work, we examine and demonstrate the potential of Formal Concept Analysis (FCA) for consolidation and analysis of clustering results derived separately from a set of microarray experiments studying the same biological phenomenon. We consider two approaches to consensus clustering of gene expression data across multiple experiments. The first algorithm consists of two distinctive steps: 1) a preliminary selected clustering algorithm (*e.g. k*-means) is initially applied to each experiment separately, which produces a list of different clustering solutions, one per experiment; 2) these clustering solutions are further transformed into a single clustering result by employing FCA, which allows to analyze and extract valuable insights from the data. In the second algorithm, the available microarray experiments are initially divided into groups of related datasets with respect to a predefined criterion (*e.g.* experimental settings). The rationale behind this is that if experiments are closely related to one another, then these experiments may produce more accurate and robust clustering solution. Subsequently, the Particle Swarm Optimization (PSO)-based clustering algorithm is applied to each group of experiments separately. The result is a list of different clustering solutions, one for each group. These clustering solutions are pooled together and again analyzed by employing FCA. Notice that FCA produces a concept lattice where each concept represents a subset of genes that belong to a number of clusters. In both algorithms the generated concepts compose the final clustering partition.

The foregoing clustering approaches have been demonstrated to have certain advantages with respect to the traditional consensus clustering techniques, namely both methods: 1) use all data by allowing potentially each experiment (or group of related experiments) to have a different set of genes, *i.e.* the total set of studied genes is not restricted to those contained into all datasets; 2) are better tuned to each experimental condition by identifying the initial number of clusters for each experiment (or group of related experiments) separately depending on the number, composition and quality of the gene profiles; 3) avoid the problem with ties (*i.e.* a case when a gene is randomly assigned to a cluster because it belongs to more than one cluster) by employing FCA in order to analyze together all the partitioning results and find the final clustering solution representative of the whole experimental compendium.