# Use of Soft Topographic Maps for Clustering Bacteria on the basis of their 16S rRNA gene sequence

ID - 170

Urso Alfonso[1], Fiannaca Antonino[2], Gaglio Salvatore[2,3], Giammanco Giovanni M.[4], La Rosa Massimo[2], Rizzo Riccardo[1]

[1]Consiglio Nazionale delle Ricerche - Istituto di CAlcolo e Reti ad alte prestazioni (ICAR/CNR) v Palermo, Italy
[2]Consiglio Nazionale delle Ricerche - Istituto di CAlcolo e Reti ad alte prestazioni (ICAR/CNR) - Palermo, Italy
[3]Dipartimento di Ingegneria Informatica - University of Palermo, Italy
[4]Dipartimento di Igiene e Microbiologia - University of Palermo, Italy

**Motivation**

Microbial identification is crucial for the study of infectious diseases. The classical method to attribute a specific name to a bacterial isolate to be identified relays on the comparison of morphologic and phenotypic characters to those described for type or typical strains. In the last years a new standard for identifying bacteria using genotypic information began to be developed. In this new approach phylogenetic relationships of bacteria could be determined by comparing a stable part of the genetic code, and the part of the DNA most commonly used for taxonomic purposes for bacteria is the 16S rRNA housekeeping gene. The goal of this work is to show that clustering of bacteria can be obtained using genotypic features and a topographic representation of the clusters allows to better understand the relationships among them.

**Methods**

In order to show the effectiveness of the proposed method, the bacteria belonging to Phylum BXII, Proteobacteria; Class III, Gammaproteobacteria, according to the current taxonomy of Bergeys Manual of Systematic Bacteriology, were considered for the analysis. This class includes 14 orders, each of them containing one or more family further subdivided in genera. A total of 147 16S rRNA gene sequences belonging to type strains representative of every single genus within the class was downloaded from the GenBank database. In order to cluster and visualize the bacteria dataset we used an algorithm that can obtain a topographic mapping of a set of objects, starting from proximity data (i.e. mutual distance or dissimilarity measures among the objects). This algorithm builds the map using a set of units (neurons) organized in a rectangular lattice that defines their neighbourhood relationships. This algorithm was developed as an extension of the Self Organizing Map (SOM), a widely used neural network for visualization purposes. The bacteria in the dataset are labelled according to their order and they are characterized by mutual dissimilarities. The dissimilarity matrix is the input for the topographic mapping algorithm, based on a cost function minimized using the deterministic annealing technique.

**Results**

We carried out experiments with maps of different dimensions, in particular 8x8, 9x9, 10x10 neurons, trained with the dissimilarity matrix described above, and we noticed that clustering obtained in 10x10 maps is the most accurate. The results are shown in figure: most of the bacteria are clustered according to their order in the present taxonomy. However, some interesting situations are conserved in all the experiments regardless to the map dimensions. Bacteria belonging to orders 5, 6, 9 and 10 are split into two areas of adjacent clusters indicating that they should be divided into different orders. One or two atypical strains, standing alone with respect to the rest of their respective order, were found within orders 1, 3, 5, 13 and 14. These bacteria could deserve a new order attribution or otherwise the sequence was wrongly registered in the GenBank. Bacteria belonging to order 7 and to one of the two clusters of order 5 are very close to each other and they may form a unique order. In general, the proposed method was able to highlight situations in which bacteria could form new orders or they might be incorrectly registered in the GenBank. The topographic map we obtained showed a clustering that generally reflects the present bacterial taxonomy, but highlighted some peculiar cases where this innovative instrument could provide a tool to both update the current taxonomy based on genotypic features and correct the GenBank sequence submission system.

**Image:** http://www.pa.icar.cnr.it/urso/figura_ridotta.jpg

**Email:** urso@pa.icar.cnr.it