Hindawi Publishing Corporation
Advances in Artificial Neural Systems
Volume 2011, Article ID 617427, 8 pages
doi:10.1155/2011/617427

Research Article

Soft Topographic Maps for Clustering and Classifying Bacteria Using Housekeeping Genes

Massimo La Rosa, Riccardo Rizzo, and Alfonso Urso

ICAR-CNR, Consiglio Nazionale delle Ricerche, Viale delle Scienze, Ed.11, 90128 Palermo, Italy

Correspondence should be addressed to Riccardo Rizzo, ricrizzo@pa.icar.cnr.it

Received 11 May 2011; Revised 13 July 2011; Accepted 26 July 2011

Academic Editor: Tomasz G. Smolinski

The Self-Organizing Map (SOM) algorithm is widely used for building topographic maps of data represented in a vectorial space, but it does not operate with dissimilarity data. Soft Topographic Map (STM) algorithm is an extension of SOM to arbitrary distance measures, and it creates a map using a set of units, organized in a rectangular lattice, defining data neighbourhood relationships. 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 bacteria genetic code, the so-called "housekeeping genes." The goal of this work is to build a topographic representation of bacteria clusters, by means of self-organizing maps, starting from genotypic features regarding housekeeping genes.

1. Introduction

Microbial identification is a fundamental topic for the study of infectious diseases, and new approaches in the analysis of bacterial isolates, for identification purposes, are currently under development. The classical method to identify bacterial isolates is based on the comparison of morphologic and phenotypic characteristics to those described as type or typical strains. On the other hand, recent trends focus on the analysis of bacteria genotype, taking into account the "housekeeping" genes, representing a very stable part of DNA. One of the most used genes, that in many studies has proven to be especially suitable for taxonomic and identification goals (see Section 2), is the 16S rRNA gene. Employing genotypic features allows to obtain a classification for rare or poorly described bacteria, to classify organisms with an unusual phenotype in a well-defined taxon, and to find misclassification that can lead to the discovery and description of new pathogens.

In this work, we present a method to make a topographic representation of bacteria clusters and to visualize the relations among them. This topographic map is obtained considering a single gene of bacteria genome, the 16S rRNA gene. Since the definition of a vector space to represent nucleotide sequences is not reliable and well structured, the information provided by the gene sequences is in the form of a pairwise dissimilarity matrix. We computed such a matrix in terms of string distances by means of well understood and theoretically sound techniques commonly used in genomics, and, in order to produce the topographic representation, we adopted a modified version of Self-Organizing Map that is able to work with input dataset expressed in terms of dissimilarity distances.

2. Background

The job of putting scientific names to microbial isolates, namely the bacteria identification, is within the practice of clinical microbiology. The aim is to give insight into the etiological agent causing an infectious disease, in order to find possible effective antimicrobial therapy. The traditional method for performing this task is dependent on the comparison of an accurate morphologic and phenotypic description of type strains or typical strains with the accurate morphologic and phenotypic description of the isolate to be identified. Microbiologists used standard references such as Bergey's Manual of Systematic Bacteriology [1]. In the 1980s, a new standard for identifying bacteria began to be

developed by Woese et al. [2]. It was shown that phylogenetic relationships of bacteria could be determined considering genotypic methods by comparing a stable part of the genetic code. The identification of bacteria based on genotypic methods is generally more accurate than the traditional identification on the basis of phenotypic characteristics. The preferred genetic technique that has emerged is based on the comparison of the bacterial 16S rRNA gene sequence, and, in recent years, several attempts to reorganize actual bacteria taxonomy have been carried out by adopting 16S rRNA gene sequences.

Authors in [1] focused on the study of bacteria belonging to the prokaryotic phyla and adopted the Principal Component Analysis method [3] on matrices of evolutionary distances. Authors in [4–6] carried out an analysis of 16S rRNA gene sequences to classify bacteria with atypical phenotype: they proposed that two bacterial isolates would belong to different species if the dissimilarity in the 16S rRNA gene sequences between them was more than 1% and less than 3%. Clustering approaches for DNA sequences were carried out by [7, 8]: the authors considered human endogenous retrovirus sequences and a distance matrix based on the FASTA similarity scores [9]; then they adopted Median SOM [10], an extension of the Self-Organizing Map (SOM) [11], to nonvectorial data.

As the authors said, the Median SOM has a better convergence if the patterns are roughly ordered. This is not an issue for the Soft Topographic Map and the Deterministic Annealing approach. The Median SOM was also used to cluster protein sequences from SWISS-PROT database in [12]. Authors in [13] proposed a protein sequence clustering method based on the Optic algorithm [14]. In [15], a technique to find functional genomic clusters in RNA expression data by computing the entropy of gene expression patterns and the mutual information between RNA expression patterns for each pair of genes is described. INPARANOID [16] is another related approach that performs a clustering based on BLAST [17] scores to find orthologs and inparalogs in two species. The use of maps for organization of biological data was also used in [18], where a map of gammaproteobacteria is reported; the obtained map is based on a reorganization of the dissimilarity matrix, and some of the results can be obtained with the approach proposed in this work.

Topological representations are not restricted, however, only to biological data, but they can be adopted, for instance, with video and audio data [19], as well.

The proposed work represents an extended version of our preliminary results presented in [20].

3. Methods

3.1. Sequence Alignment and Evolutionary Distance. Sequence alignment is a well-known bioinformatics technique useful to compare genomic sequences, even of different length, between two different species. In our system, we used two of the most popular alignment algorithms: ClustalW [21], implementing a multiple alignment among all sequences at the same time; Needleman and Wunsch [22], that provide a pairwise alignment, that is the best alignment configuration between two sequences.

Once aligned, it is possible to compute a distance between two homologous sequences. In bioinformatics domain, there are many types of distances, usually called "evolutionary distance"; these distances differ from each other on the basis of their a priori assumptions.

The simplest kind of distance is the number of substitutions per site, defined as

$$p = \frac{\text{number of different nucleotides}}{\text{total number of compared nucleotides}}. \quad (1)$$

The number of substitutions observed is often smaller than the number of substitutions that have actually taken place. This is due to many genetic phenomena such as multiple substitutions on the same site (*multiple hits*), convergent substitutions or retromutations. For these reasons, a series of stochastic methods has been introduced in order to obtain a better estimate of evolutionary distances. In our study, we considered the method proposed by [23], whose a priori assumptions are

(1) all sites evolve in an independent manner;

(2) all sites can change with the same probability;

(3) all kinds of substitution are equally probable;

(4) substitution speed is constant over time.

According to [23], the evolutionary distance d between two nucleotide sequences is equal to

$$d = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right), \quad (2)$$

where p is the number of substitutions per site (1).

3.2. Soft Topographic Map Algorithm. A widely used algorithm for topographic maps is the Kohonen's Self-Organizing Map (SOM) algorithm [11], but it does not operate with dissimilarity data. The SOM network builds a projection from an input space to a lattice (usually 2D) of neurons, visualized as a 2D map. Each neuron is a pointer to a position in the input space and is a tile on the map. The input patterns are distributed on the map because they are associated to the nearest neuron in the input space: that neuron is usually referred to as the best-matching unit (bmu). SOM networks are trained using the unsupervised learning paradigm: the label of input patterns, if present, will not be considered during training phase.

The SOM is widely used to project input data into a low-dimensional space [24, 25].

Many studies on the SOM algorithm have been carried after the original paper: according to Luttrell's work [26], the generation of topographic maps can be interpreted as an optimization problem based on the minimization of a cost function. This cost function represents an energy function, and it takes its minimum when each data point is mapped to the best matching neuron, thus providing the optimal set of parameters for the map.

An algorithm based on this formulation of the problem was developed by Graepel et al. [27, 28] and provides an extension of SOM to arbitrary distance measures. This algorithm is called Soft Topographic Map (STM) and creates a map using a set of units (neurons or models) organized in a rectangular lattice that defines their neighborhood relationships. STM is able to work with data whose features are expressed in terms of dissimilarity measures among each other. Algorithm full description, along with theoretical and practical details, can be read in [27].

4. Implementation

The Soft Topographic Map algorithm described in this paper needs some tuning; in this section, we give all the necessary information for a fruitful use of the algorithm.

4.1. Dataset. The main purpose of our work is to demonstrate that STM algorithm can be applied to a biological dataset in order to obtain a topographic map useful to visualize clusters of bacteria belonging to the same order, according to actual taxonomy. Biological dataset is composed of 16S rRNA gene sequences. Each sequence is a text string containing only four types of characters: "A," "C," "G," "T" corresponding to the four DNA nucleotides. Information content of the dataset is expressed in terms of a dissimilarity measure, computed according to (2). According to the actual taxonomy [1], we focused our attention on a class containing some of the most common and dangerous bacteria related to human pathologies: Gammaproteobacteria, belonging to the Proteobacteria phylum. In Table 1, a brief description of the experimental dataset is shown: the dataset is composed of 147 type strains, and the resulting 16S gene sequences were downloaded from NCBI public nucleotide database, GenBank [29], in FASTA format [30].

4.2. Parameters Setup. A Soft Topographic Map is an array of many neural units where patterns to classify are associated to these units at the end of the training phase. In order to speed up processing time, we applied a slightly tuned version of the Soft Topographic Map algorithm: neighborhood functions associated to each neuron have been set to zero if they referred to neurons outside a previously chosen radius in the grid. The radius has been put to 1/3 of the map dimensions. As for the other parameters of the algorithm, we put the annealing increasing factor $\eta = 1.1$ and threshold convergence $\epsilon = 10^{-5}$, as suggested by [27]. After several tests we chose, as a good compromise between processing time and clustering quality, the final value of inverse temperature equal to 10 times the initial value, leading as a consequence to 25 learning epochs; finally, we put the width of neighborhood functions σ to 0.5.

The maps have been drawn using a gray-level scale to represent the distances between the units: the color between two near occupied cells, both horizontally and vertically, is proportional to the average distance of the patterns being in those neurons. To be more precise, empty cells are filled with a gray level proportional to the mean distance among the four closest occupied neurons, along the vertical and horizontal axes referred to that empty cell. Gray scale is calibrated so that bright values denote proximity and dark values represent distance. Two sample maps and the distance scale are shown in Figure 1.

4.3. Map Evaluation Criteria. In order to select the map dimension, it is useful to evaluate the evolution of the clustering process with regard to map size. To this end, it is possible to compare the number of neural units and the number of patterns defining the following ratio:

$$K = \frac{\text{number of pattern to classify}}{\text{number of neural units}}. \tag{3}$$

If $K \geq 1$, then each neural unit can have many input patterns so that each neural unit can be considered as a cluster. In this case, the focus is on the use of all neural units, and the ones that are not used are often referred to as "dead units".

If $K < 1$, then the single neural unit cannot be a cluster center and the cluster is constituted by many neural units separated by a set of dead units. The maps with $K \geq 1$ are sometimes called KNN-SOM, while the ones with $K < 1$ are visualized with a technique called U-Matrix [31].

In our implementation, we started with a ratio $K \approx 2$ (using a square 8×8 STM map) in order to understand if it was possible to identify the neural units as clusters. It was difficult to find this correspondence due to the high number of units that were associated with sequences of different order. This is highlighted in the center diagram in Figure 2, that shows the number of mixed clusters (i.e., units that have associated sequences of different orders).

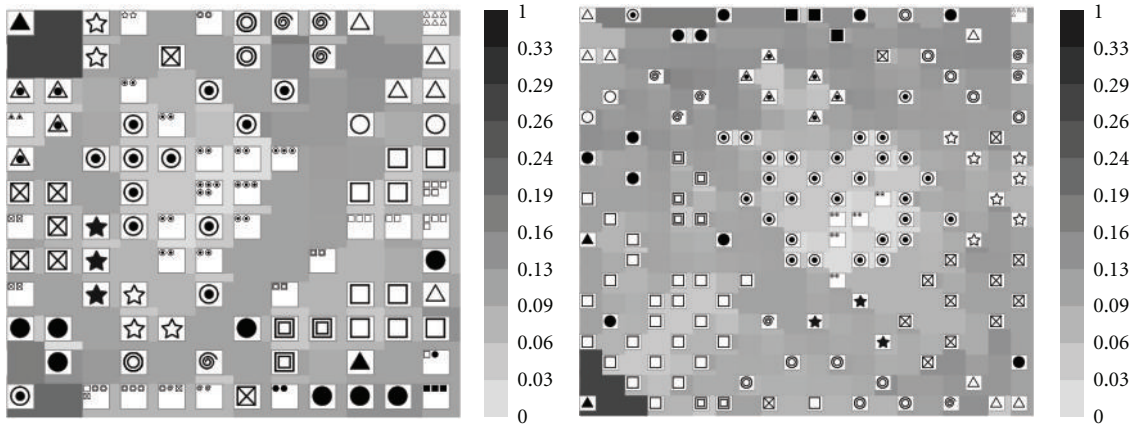
Usually neural network results are determined by initial weight values. A common procedure to filter this noise is to train many networks with different initial set up. In our experiments, we used 20 different network initializations for each experiment.

For the evaluation of the quality of the mapping, several methods are reported in literature, but these methods need a metric space (a vector space) where the patterns and the units of the map are represented as vectors. For a short review on topology preservation, see [32]. In our problem, we have not a feature space where the patterns are placed, and we have only a dissimilarity matrix that reports the pattern organization.

In order to establish an evaluation criteria for the obtained maps, we noticed that the rows and the columns of the map represent a linear ordering of the patterns, order that should be present also in the dissimilarity matrix. For example, selecting a row on the map, we have a set of ordered patterns; the same patterns are used to select the corresponding subset of rows and columns of the dissimilarity matrix. These dissimilarity values can be considered as distance values and allow to order the patterns in a linear fashion. A pattern sequence can be easily obtained using the Sammon mapping technique on a linear space starting from the data of dissimilarity matrix. This sequence should be identical to the one obtained from the map; the

Table 1: Actual taxonomy of the bacteria dataset. We focused on Gammaproteobacteria class, which is divided into 14 orders. Each order has one or more families. Inside each family, we considered only the type strains, that is, sample species.

Gammaproteobacteria	Order name	Number of families	Number of type strains	Code numbers
□	Chromatiales	3 families	25 type strains	1–25
○	Acidithiobacillales	2 families	2 type strains	26,27
△	Xanthomonodales	1 family	11 type strains	28–38
■	Cardiobacteriales	1 family	3 type strains	39, 40, 41
●	Thiotrichales	3 families	11 type strains	42–52
▲	Legionellales	2 families	2 type strains	53, 54
▣	Methylococcales	1 families	7 type strains	55–61
◎	Oceanospirillales	4 families	11 type strains	62–72
֎	Pseudomonadales	2 families	7 type strains	73–79
⊠	Alteromonadales	1 family	13 type strains	80–92
★	Vibrionales	1 family	3 type strains	93, 94, 95
☆	Aeromonadales	2 families	7 type strains	96–102
⊙	Enterobacteriales	1 family	39 type strains	103–141
◬	Pasteurellales	1 families	6 type strains	142–147
	14 orders	25 families	147 type strains	



(a) (b)

Figure 1: 12 × 12 (left) and 20 × 20 (right) topographic maps of bacteria dataset. In the legend under the figures, the dissimilarity values corresponding to each gray level are shown. 0.33 is the max distance in our dataset, so this darkest gray level available.

two sequences can be compared using the Spearman's rank correlation coefficient [33] defined as

$$\rho = 1 - \left[\frac{6 \sum d_i^2}{n(n^2 - 1)} \right], \tag{4}$$

where d_i is the difference between each rank of corresponding values of the compared variables x and y; n is the number of pairs of values. In the above equation, we consider only the term in the square brackets because we discard the possible inversion between the pattern sequence of the map and the one of the dissimilarity matrix. Averaging all the Spearman coefficients for each column and each row, we obtain a score for a given map. All these scores, calculated for each map geometry and for initialization, are reported in the upper diagram of Figure 2 as a box plot.

Evaluating this coefficient, we can decide which geometry can be used. Maps with few neural units are discarded, because there are units with many patterns that create ties in the ordering; in fact, patterns associated with the same unit do not have any order, while very large maps present a naturally decreasing value due to the fact that the patterns are very sparse. This effect can be seen in Figure 2 on the right of the thin vertical line.

5. Results

In this Section, we present the results we obtained applying the techniques described in Section 3 to the bacteria dataset described in Section 4.

Given the dataset described above, we carried out several experiments. We obtained several maps of different

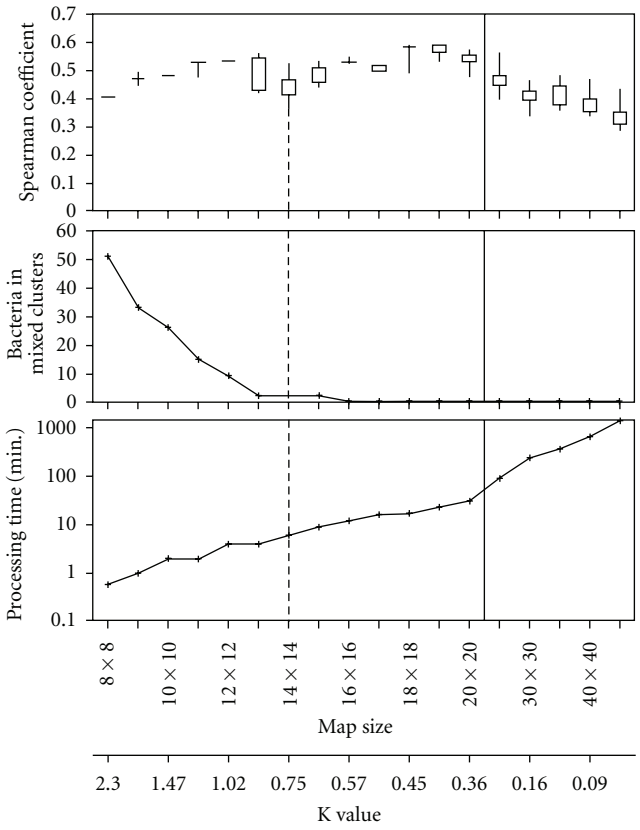


Figure 2: In the upper graph is the box plot graph of the Spearman coefficient. In the chart on the center, we can observe that the number of bacteria belonging to mixed clusters, that is, cells in the map labeled with bacteria of different orders, aims at decreasing as the size of maps increases. In the lower graph is the processing time in minutes (logarithmic scale).

dimensions, from 8×8 up to 45×45 neurons, and for every configuration, we trained 20 maps in order to avoid the dependence from the initial conditions.

Comparing the results provided from pairwise and multiple alignment, we saw that there are not meaningful differences in the corresponding maps, so we focused only on the evolutionary distances computed from pairwise alignment.

In Figure 1, we can see the evolution of clustering process with regard to map size: first of all, we can notice how most of the bacteria are classified according to their order in the actual taxonomy; then, we can observe that the number of bacteria belonging to mixed clusters, that is, cells in the map labeled with bacteria of different orders, aims at decreasing as the size of maps increases. We can state that small maps, according to the chart until about 10×10, do not provide useful results because there are too few available neurons and consequently the maps are not able to correctly discriminate among different patterns. If we look, in fact, at the charts of Figure 2, there are too many mixed clusters and high values of the Spearman coefficient. On the other side, we noticed that in very large maps (not shown in this paper), from 25×25 and so on, the topographic maps "lose" their clustering properties because input patterns aim at spreading all over the grid, filling all the available space. Considering the definition of parameter K given in (3), maps with $K < 1$ and $K \gg 1$ are meaningless.

The map size and the optimal K parameter value are also a function of the method used to produce the dissimilarity matrix. For example, using Normalized Compression Distance (NCD) [34], the optimum size of the map can be different, as stated in one of our previous work about this topic [35].

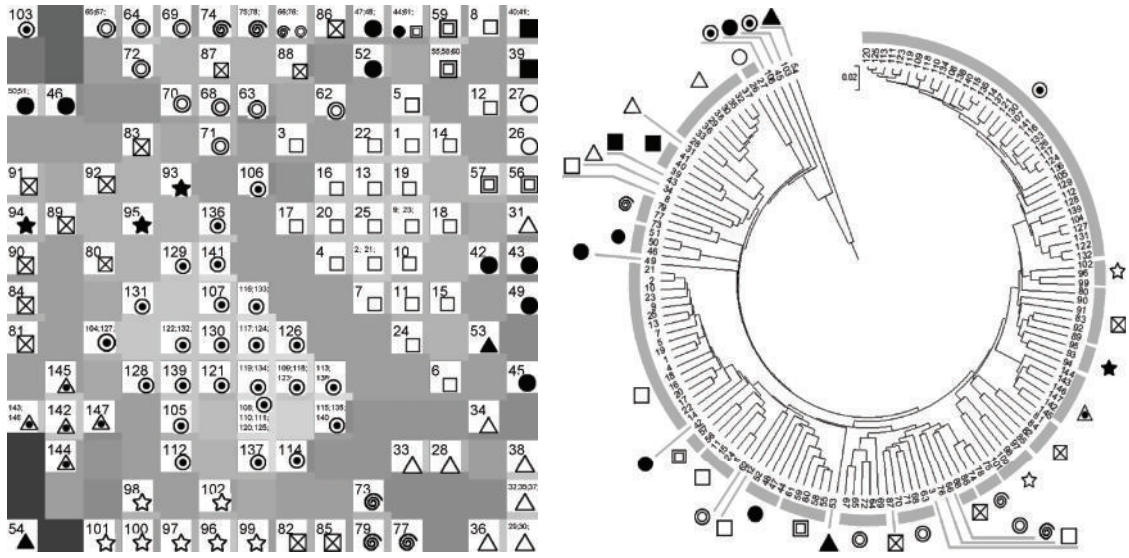
One of the most interesting result is that there are some anomalies that are constant for all the tests regardless the dimension of the maps. For example, in small maps (not shown here), the "*Alterococcus agarolyticus*" (number 103 in Figure 3) bacterium of the "Enterobacteriales" order is incorrectly clustered together with bacteria of other orders, whereas, in larger maps, it is isolated in an individual cluster, usually at the border of the map and far from its homologous strains (see Figure 3). Another interesting example is given by "*Legionella pneumophila*" (number 54 in Figure 3) bacterium of "Legionellales" order: in all maps, it is located in a corner of the grid and surrounded by a dark gray area. This would suggest that these two bacteria could form new orders, not present in actual taxonomy, or at least new families. The same anomalies are confirmed by the Multidimensional Scaling and the evolutionary tree.

Since the maps provide a visualization of bacteria datasets, if there are some "anomalies", they are clearly highlighted as isolated elements standing at the border or in the corners of the map. These anomalies can suggest biologists to do further experimental trials in order to determine if, eventually, there are some misclassifications in the taxonomy. That does not mean the proposed method should mainly be used in order to perform identification or annotation of unknown bacterial species, but that the visualization is also able to detect anomalies and if there are unknown elements, to project them in the map because of unsupervised learning feature of STM algorithm (see Section 3).

The bacteria organization in the map finds some other confirmation in [18], for example, the neighborhood of Xanthomonas (33 in Figure 3), Pseudomonas (73), and Enterobacteriales (114); notice also that the position of Buchnera (106) is not in the same compact group of the other Enterobacteriales in the map center, although not so distant as depicted in [18].

Considering the evaluation of the map, reported in Figure 2, we choose in the set of the 14×14 maps the one that presents the absolute minimum of the Spearman coefficient before of its natural decreasing on the right side of the thin vertical line. This choice also minimizes the number of mixed clusters, as can be seen in the center diagram of Figure 2.

5.1. Comparison with Phylogenetic Tree. We compared the chosen 14×14 map with the phylogenetic tree referred to our dataset. In Figure 3, it is possible to notice that there are four outliers bacteria: "*Francisella tularensis*" (45), "*Legionella pneumophila*" (54), "*Alterococcus agarolyticus*" (103), and "*Buchnera aphidicola*" (106). The first three



(a) (b)

Figure 3: Comparison between the phylogenetic tree and the selected 14×14 map. It is possible to notice that there are four outliers bacteria: "*Francisella tularensis,*" "*Alterococcus agarolyticus,*" "*Legionella pneumophila,*" "*Buchnera aphidicola.*" The first three bacteria are clustered in the border of the map and far from their homologous strains; the remaining one lies in a single cell surrounded by a dark gray area that indicates its actual distance from its neighbors is bigger than the one shown in the map. There are other bacteria far from their homologous strains: "*Schineria larvae,*" "*Arhodomonas aquaeolei,*" "*Halothiobacillus neapolitanus,*" "*Nitrosococcus nitrosus.*" "Enterobacteriales" and "Pasteurellales" form compact group in both representations.

bacteria are positioned on the border of the map and far from their homologous strains; the remaining one lies in a single cell surrounded by a dark gray area that indicates its actual distance from its neighbors is bigger than it appears. Apart from these four elements, in the phylogenetic tree, we found other bacteria far from their order, for instance, "*Schineria larvae*" (34), "*Halothiobacillus neapolitanus*" (8), "*Nitrosococcus nitrosus*" (12), and "*Arhodomonas aquaeolei*" (3): once again these elements are at the border or in a zone on the map surrounded by a dark gray level. Although "*Schineria larvae*" (34) and "*Halothiobacillus neapolitanus*" (8) are coupled in the dendrogram, we can see in the map how they are actually far away: that happens because some pairings in the tree are forced, and, in this case, do not give useful information. "*Schineria larvae*" (34) and "*Francisella tularensis*" (45), whose actual distance is 0.1339, are close in the map, but their surrounding gray level explains their real distance, as we can also see in the phylogenetic tree.

If we consider entire orders, for example, "Enterobacteriales" and "Pasteurellales", they form compact groups both in the tree and in the map. Moreover, the "Methylococcales" order that in actual taxonomy has one family, in the map, is divided in two clusters (56, 57 and 55, 58, 59, 60, 61) as reported in the phylogenetic tree and in [18].

Our visualization method allows, then, not only to detect some singular situations, but also to understand their relative positions with regards to all the patterns in the dataset. At a first look to the phylogenetic tree, in fact, it should be possible to wrongly realize that the four outliers described above are far from all the other bacteria, but near each other. Using the map, instead, we can see how the four outliers are completely isolated. At the same time our method provides a very simple system to immediately visualize compact orders and/or families, as previously explained.

This is clear looking at Figure 3 where the map and the tree contain the same objects but the map is far more readable.

5.2. Comparison with Multidimensional Scaling. Multidimensional Scaling (MDS) is a widely used technique for embedding a dataset, defined only in terms of pairwise distances, in an euclidean space and plotting it in a 2D (or 3D) plane [36]. For this reason, we compared our two-dimensional topographic representation with a 2D plot, obtained through MDS, of our bacteria dataset, presented in Figure 4.

First of all, we can notice that the four outliers bacteria, "*Francisella tularensis,*" "*Buchnera aphidicola,*" "*Alterococcus agarolyticus,*" "*Legionella pneumophila,*" are separated from all the other elements. Apart from this evident result, there are not many other similarities with our map nor with the phylogenetic tree. Bacteria belonging to "Pasteurellales" order, for example, forming in the previous visualizations a well-defined group, in MDS plot, stand in very distant zones without any observable relationship. There are some dislocated elements even inside "Enterobacteriales," though most of them still form a compact group in the center part of the diagram. Moreover, it is difficult to give a clue on the distance among the patterns.

In conclusion, the use of MDS plotting gives less information with respect to the ones obtained by means of topographic map and phylogenetic tree. Because of the

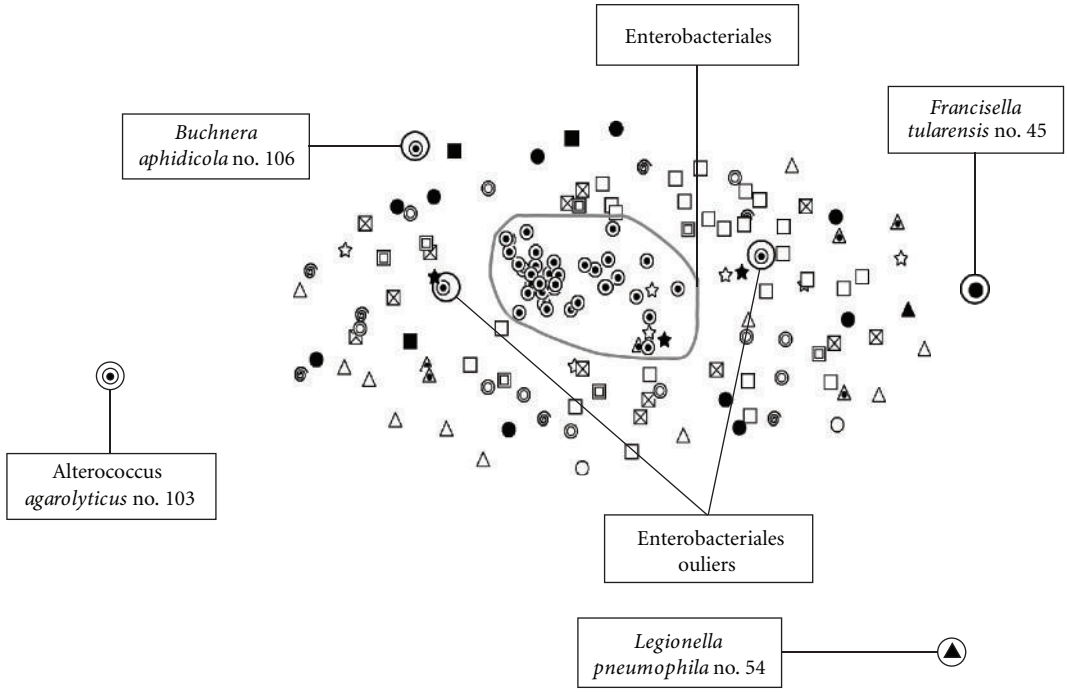


Figure 4: 2D representation of bacteria dataset obtained through Multidimensional Scaling. Apart from the four outliers, "*Legionella pneumophila,*" "*Alterococcus agarolyticus,*" "*Francisella tularensis,*" "*Buchnera aphidicola,*" that are separated from all the other elements, the remaining bacteria do not show meaningful similarities with the visualizations obtained through topographic map and phylogenetic tree.

distortion introduced by MDS, in fact, most of the patterns, with the exception of the four outliers, have lost their distinctive properties already discussed in the previous paragraph.

6. Conclusion

In recent trends for the definition of bacteria taxonomy, genotypical characteristics are considered very important and type strains are compared on the basis of the stable part of the genetic code. In this paper, the Soft Topographic Map algorithm has been applied to the visualization and clustering of bacteria according to their genotypic similarity. In the similarity measure, we have adopted the 16S rRNA gene sequence, as commonly used for taxonomic purposes. A characteristic of the proposed approach is that the topographic map is built from the genetic data, using the Soft Topographic Map algorithm working on proximity data, rather than using a vector space representation. The generated maps show that the proposed approach provides a clustering that generally reflects the current taxonomy with some singular cases. Moreover, the results depend on the size of the maps, since small and large maps, with regards to the number of input patterns, do not give meaningful information. The size of the maps should be chosen so that the ratio between input elements and neurons is $K \approx 1$, with a corresponding value of Spearman coefficient representing a local minimum.

The visualization of bacteria dataset through the map also allows an easy identification of cases representing some "anomalies" in input dataset. These anomalies should be further investigated because they, eventually, could represent incorrect classification or incorrect registration in the database. It also provides a compact representation, in one image, useful to visualize bacteria clusters and their mutual separation, although the evaluation of distance between clusters is still inaccurate. Furthermore our system has proved to be a valid alternative to the traditional visualizing tool used in bioinformatics, like phylogenetic trees and 2D plot obtained through MDS.

In future research activities, we intend to extend the analysis to other "housekeeping" genes and to combine different genotypical characteristics in order to obtain finer clustering and classification. We would like also to use other distance measures, eventually alignment-free, and different clustering algorithm in order to improve execution time and the quality of clustering.

References

[1] G. M. Garrity, B. A. Julia, and T. Lilburn, "The revised road map to the manual," in *Bergey's Manual of Systematic Bacteriology*, G. M. Garrity, Ed., pp. 159–187, Springer, New York, NY, USA, 204.

[2] C. R. Woese, E. Stackebrandt, T. J. Macke, and G. E. Fox, "A phylogenetic definition of the major eubacterial taxa," *Systematic and Applied Microbiology*, vol. 6, no. 2, pp. 143–151, 1985.

[3] I. T. Joliffe, *Principal Component Analysis*, Springer, New York, NY, USA, 1986.

[4] J. E. Clarridge, "Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases," *Clinical Microbiology Reviews*, vol. 17, no. 4, pp. 840–862, 2004.

[5] M. Drancourt, C. Bollet, A. Carlioz, R. Martelin, J.-P. Gayral, and D. Raoult, "16S ribosomal DNA sequence analysis of a large collection of environmental and clinical unidentifiable bacterial isolates," *Journal of Clinical Microbiology*, vol. 38, pp. 3623–3630, 2000.

[6] M. Drancourt, P. Berger, and D. Raoult, "Systematic 16S rRNA gene sequencing of atypical clinical isolates identified 27 new bacterial species associated with humans," *Journal of Clinical Microbiology*, vol. 42, no. 5, pp. 2197–2202, 2004.

[7] M. Oja, P. Somervuo, S. Kaski, and T. Kohonen, "Clustering of human endogenous retrovirus sequences with median self-organizing map," in *Proceedings of the Workshop on Self-Organizing Maps (WSOM '03)*, 2003.

[8] M. Oja, G. O. Sperber, J. Blomberg, and S. Kaski, "Self-organizing map-based discovery and visualization of human endogenous retroviral sequence groups," *International Journal of Neural Systems*, vol. 15, no. 3, Article ID 163179, 2005.

[9] W. R. Pearson and D. J. Lipman, "Improved tools for biological sequence comparison," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 85, no. 8, pp. 2444–2448, 1988.

[10] T. Kohonen and P. Somervuo, "How to make large self-organizing maps for nonvectorial data," *Neural Networks*, vol. 15, no. 8-9, pp. 945–952, 2002.

[11] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin, Germany, 1995.

[12] P. Somervuo and T. Kohonen, "Clustering and visualization of large protein sequence databases by means of an extension of the self-organizing map," in *Proceedings of the 3rd International Conference on Discovery Science*, pp. 76–85, 2000.

[13] Y. Chen, K. D. Reilly, A. P. Sprague, and Z. Guan, "Seqoptics: a protein sequence clustering method," in *Proceedings of the 1st International Multi- Symposiums on Computer and Computational Sciences (IMSCCS'06)*, vol. 1, pp. 69–75, June 2006.

[14] M. Ankerst, M. M. Breunig, H. P. Kriegel, and J. Sander, "Optics: ordering points to identify the clustering structure," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 49–60, Philadelphia, Pa, USA, June 1999.

[15] A. J. Butte and I. S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," in *Proceedings of the Pacific Symposium on Biocomputing*, vol. 5, pp. 415–426, 2000.

[16] M. Remm, C. E. V. Storm, and E. L. L. Sonnhammer, "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons," *Journal of Molecular Biology*, vol. 314, no. 5, pp. 1041–1052, 2001.

[17] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 232, pp. 584–599, 1993.

[18] G. M. Garrity and T. G. Lilburn, "Self-organizing and self-correcting classifications of biological data," *Bioinformatics*, vol. 21, no. 10, pp. 2309–2314, 2005.

[19] C. Fyfe, W. Barbakh, W. C. Ooi, and H. Ko, "Topological mappings of video and audio data," *International Journal of Neural Systems*, vol. 18, no. 6, pp. 481–489, 2008.

[20] M. La Rosa, G. Di Fatta, S. Gaglio, G. M. Giammanco, R. Rizzo, and A. M. Urso, "Soft topographic map for clustering and classification of bacteria," in *Advances in Intelligent Data Analysis VII*, vol. 4723 of *Lecture Notes in Computer Science*, pp. 332–343, 2007.

[21] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[22] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[23] T. H. Jukes and C. R. Cantor, "Evolution of protein molecules," in *Mammalian Protein Metabolism*, H. N. Munro, Ed., pp. 21–132, Academic Press, New York, NY, USA, 1969.

[24] A. N. Gorban and A. Zinovyev, "Principal manifolds and graphs in practice: from molecular biology to dynamical systems," *International Journal of Neural Systems*, vol. 20, no. 3, pp. 219–232, 2010.

[25] W. Barbakh and C. Fyfe, "Online clustering algorithms," *International Journal of Neural Systems*, vol. 18, no. 3, pp. 185–194, 2008.

[26] S. P. Luttrell, "A Bayesian analysis of self-organizing maps," *Neural Computation*, vol. 6, pp. 767–794, 1994.

[27] T. Graepel, M. Burger, and K. Obermayer, "Self-organizing maps: generalizations and new optimization techniques," *Neurocomputing*, vol. 21, no. 1–3, pp. 173–190, 1998.

[28] T. Graepel and K. Obermayer, "A stochastic self-organizing map for proximity data," *Neural Computation*, vol. 11, no. 1, pp. 139–155, 1999.

[29] GenBank, 2007, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide.

[30] Fasta, 2007, http://www.ncbi.nlm.nih.gov/blast/fasta.shtml.

[31] A. Ultsch, "Maps for the visualization of high dimensional data spaces," in *Proceedings of the Workshop on Self-Organizing Maps (WSOM '03)*, vol. 3, pp. 225–230, 2003.

[32] D. Vidaurre and J. Muruzábal, "A quick assessment of topology preservation for SOM structures," *IEEE Transactions on Neural Networks*, vol. 18, no. 5, pp. 1524–1528, 2007.

[33] E. W. Weisstein, *The CRC Concise Encyclopedia of Mathematics*, CRC Press, New York, NY, USA, 1999.

[34] M. Li, X. Chen, X. Li, B. Ma, and P. M. B. Vitanyi, "The similarity metric," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3250–3264, 2004.

[35] M. La Rosa, S. Gaglio, R. Rizzo, and A. Urso, "Normalised compression distance and evolutionary distance of genomic sequences: comparison of clustering results," *International Journal of Knowledge Engineering and Soft Data Paradigms*, vol. 1, no. 4, pp. 345–362, 2009.

[36] W. S. Torgerson, "Multidimensional scaling: I. Theory and method," *Psychometrika*, vol. 17, pp. 401–419, 1952.