Soft topographic map for clustering and classification of bacteria

In the Proc. of the 7th Int.l Symposium on Intelligent Data Analysis (IDA)
Ljubljana, 6-8 Sept. 2007

Soft Topographic Map for Clustering and Classification of Bacteria

Massimo La Rosa[1], Giuseppe Di Fatta[2], Salvatore Gaglio[13],
Giovanni M. Giammanco[4] , Riccardo Rizzo[1] and Alfonso M. Urso[1]

[1] ICAR-CNR, Consiglio Nazionale delle Ricerche, Palermo, Italy.
[2] School of Systems Engineering, University of Reading, UK
[3] Dipartimento di Ingegneria Informatica, Universitá di Palermo, Italy
[4] Dipartimento di Igiene e Microbiologia, Universitá di Palermo, Italy

Abstract. In this work a new method for clustering and building a topographic representation of a bacteria taxonomy is presented. The method is based on the analysis of stable parts of the genome, the so-called "housekeeping genes". The proposed method generates topographic maps of the bacteria taxonomy, where relations among different type strains can be visually inspected and verified. Two well known DNA alignement algorithms are applied to the genomic sequences. Topographic maps are optimized to represent the similarity among the sequences according to their evolutionary distances. The experimental analysis is carried out on 147 type strains of the Gammaprotebacteria class by means of the 16S rRNA housekeeping gene. Complete sequences of the gene have been retrieved from the NCBI public database. In the experimental tests the maps show clusters of homologous type strains and present some singular cases potentially due to incorrect classification or erroneous annotations in the database.

1 Introduction

Microbial identification is crucial for the study of infectious diseases. 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. Recently a new naming approach based on bacteria genotype has been proposed and is currently under development. In this new approach phylogenetic relationships of bacteria could be determined by comparing a stable part of the genetic code. The part of the genetic code commonly used for taxonomic purposes for bacteria is the 16S rRNA "housekeeping" gene. The 16S rRNA gene sequence analysis can be used to obtain a classification for rare or poorly described bacteria, to classify organisms with an unusual phenotype in a well defined taxon, to find misclassification that can lead to the discovery and description of new pathogens.

The aim of this work is to obtain a topographic representation of bacteria clusters to visualize the relations among them. Moreover, we intend to achieve this objective by using directly the genotype information, without building a

feature space. Many clustering approaches are based on a feature space where objects are represented. Biological datasets usually contain large objects (long nucleotides sequences or images); a vector space representation of such objects can be difficult and typically results in a high dimensional space where the euclidean distance is a low contrast metric. The definition of a vector space also requires the choice of a set of meaningful axes that represent some measurable qualities of the objects. In DNA sequences this approach is not straightforward and may be hindered by an arbitrary choice of features. According to these considerations we do not adopt a vector space representation, but a matrix of pairwise distances obtained directly from the genetic sequences. Such a matrix can be computed in terms of string distances by means of well understood and theoretically sound techniques commonly used in genomics.

The paper is organized as follows: in section 2 we refer to works that focus on similar classification problems of biological species; in sections 3 and 4 we describe the algorithms we have adopted for the similarity measure and the generation of topographic maps; in section 5 we present an experimental analysis of the proposed method and provide an interpretation of the results.

2 Related Work

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 prokaryothic phyla and adopted the Principal Component Analysis method [2] on matrices of evolutionary distances. Clarridge [3], Drancourt et al. [4, 5] 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 [7] and for protein sequences [9] adopted Median Som, an extension of the Self-Organizing Map (SOM) to non-vectorial data. Chen et al. [11] proposed a protein sequence clustering method based on the Optic algorithm [12]. Butte and Kohane [8] described 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. INPARANOID [13] is another related approach that performs a clustering based on BLAST [14] scores to find orthologs and in-paralogs in two species.

Among other algorithms for the clustering of pairwise proximity data, it is worth to mention an approach to segment textured images [29]. Dubnov et al. [16] proposed a nonparametric pairwise clustering algorithm that iteratively extracts the two most prominent clusters in the dataset, thus generating a hierarchical clustering structure. A hierarchical approach was also followed in [17, 18]. Other works, e.g. [19, 20], adopted Multidimensional Scaling [22] to embed dissimilarity data in a Euclidean space.

3 Genetic Sequence Similarity

3.1 Sequence Alignment

Sequence alignment allows to compare homologous sites of the same gene between two different species. For this purpose, we used two of the most popular alignment algorithms: ClustalW [23] for multiple-alignment; and Needleman-Wunsch [24] for pairwise alignment. The ClustalW algorithm aims to produce the best alignment configuration considering all the sequences at the same time, whereas Needleman-Wunsch algorithm provides a global optimum alignment between two sequences even of different length. Sequence alignment algorithms usually insert gaps in the input sequences in order to stretch them and to find the best matching configuration: gaps represent nucleotide insertions or deletions and are very important in terms of molecular evolution. An example of pairwise alignment is shown in Figure 1.

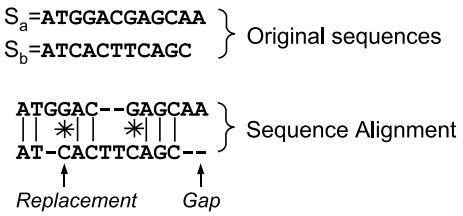

Fig. 1. Pairwise alignment between two gene sequences

3.2 Evolutionary Distance

The evolutionary distance is a distance measure between two homologous sequences, previously aligned. There are several kinds of evolutionary distances: the simplest one is the number of nucleotide substitutions per site. The number of substitutions observed between sequences 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 retro-mutations. As a consequence, it is important to use stochastic methods in order to obtain an exact estimate of evolutionary distances. Many stochastic models exist that differ from each other on the basis of their a priori assumptions.
The most common a priori assumptions are:

- all sites evolve in an independent manner;
- all sites can change with the same probability;
- all kinds of substitution are equally probable;

– substitution speed is constant over time.

In our study, we used the method proposed by Jukes and Cantor [25], where all the assumptions above are valid. According to [25], the evolutionary distance d between two nucleotide sequences is equal to:

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

where p is the number of substitutions per site, defined as:

$$p = \frac{\text{number of different nucleotides}}{\text{total number of compared nucleotides}}\ . \tag{2}$$

It is important to note that sites containing gaps or undefined bases are not considered in the computation of distances.
Evolutionary distances computed with (1) constitute the elements of a dissimilarity matrix that represents the input for the algorithm described in the next section.

4 Soft Topographic Map Algorithm

A widely used algorithm for topographic maps is the Kohonen's Self Organizing Map (SOM) algorithm [31], but it does not operate with dissimilarity data.

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 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, Burger and Obermayer [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 neighbourhood relationships.

The cost function for soft topographic mapping of proximity data (in our case a dissimilarity matrix) can be formulated as follows:

$$E(\{c_{t\mathbf{r}}\}) = \frac{1}{2}\sum_{t,t'}\sum_{\mathbf{r},\mathbf{s},\mathbf{u}} \frac{c_{t\mathbf{r}}h_{\mathbf{rs}}c_{t'\mathbf{u}}h_{\mathbf{us}}}{\sum_{t''}\sum_{\mathbf{v}} c_{t''\mathbf{v}}h_{\mathbf{vs}}} d_{tt'}, \tag{3}$$

where $d_{tt'}$ is the generic element of the dissimilarity matrix, namely the pairwise distance among nucleotide sequences of bacteria t and t'. Two constraints hold in (3): $\sum_{\mathbf{r}} c_{t\mathbf{r}} = 1, \forall t$, i.e. each data vector can belong only to one neuron $\mathbf{r}$, and $\sum_{\mathbf{s}} h_{\mathbf{rs}} = 1, \forall \mathbf{r}$. The function $h_{\mathbf{rs}}$ is equivalent to the neighborhood function of classic SOM algorithm and represents the coupling between neurons $\mathbf{r}$ and $\mathbf{s}$ in the map grid. $h_{\mathbf{rs}}$ is usually chosen as a normalized Gaussian function such as:

$$h_{\mathbf{rs}} \propto \exp\left(-\frac{|\mathbf{r}-\mathbf{s}|^2}{2\sigma^2}\right), \forall \mathbf{r}, \mathbf{s}. \tag{4}$$

Table 1. Soft Topographic Map algorithm.

1. Initialization Step:
   (a) $e_{t\mathbf{r}} \leftarrow n_{t\mathbf{r}}, \forall t, \mathbf{r}, n_{t\mathbf{r}} \in [0,1]$
   (b) compute lookup table for $h_{\mathbf{rs}}$ as in Eq. (4)
   (c) compute dissimilarity matrix from input data as in Eq. (1)
   (d) put $\beta \cong \beta^*$
   (e) choose β_{final}, increasing temperature factor η, convergence threshold ϵ
2. Training Step:
   (a) while $\beta < \beta_{final}$ (Annealing cycle)
      i. repeat (EM cycle)
         A. E step: compute $P(\mathbf{x}_t \in \mathbf{C_r})\forall t, \mathbf{r}$ as in Eq. (5)
         B. M step: compute $a_{t\mathbf{r}}^{new}, \forall t, \mathbf{r}$ as in Eq. (7)
         C. M step: compute $e_{t\mathbf{r}}^{new}, \forall t, \mathbf{r}$ as in Eq. (6)
      ii. until $\left\| e_{t\mathbf{r}}^{new} - e_{t\mathbf{r}}^{old} \right\| < \epsilon$
      iii. put $\beta \leftarrow \eta\beta$
   (b) end while

In order to optimize the cost function the deterministic annealing [29, 30] technique has been used. This technique is based on the optimization of a family of cost functions, representing free energy, that depend on the parameter β, the so called inverse temperature. This parameter represents the amount of smoothing that is done to the original cost function.

The minimization of this function leads to the probability of the assignment of the data vector t to the node $\mathbf{r}$ (i.e. to its cluster $\mathbf{C_r}$):

$$P(\mathbf{x}_t \in \mathbf{C_r}) = \frac{\exp(-\beta e_{t\mathbf{r}})}{\sum_{\mathbf{u}} \exp(-\beta e_{t\mathbf{u}})}, \forall t, \mathbf{r}. \tag{5}$$

In Equation (5), $e_{t\mathbf{r}}$ is the partial assignment cost of data vector $\mathbf{x}_t$ to be assigned to cluster $\mathbf{C_r}$, and it is defined as:

$$e_{t\mathbf{r}} = \sum_{\mathbf{s}} h_{\mathbf{rs}} \sum_{t'} a_{t'\mathbf{s}} \left(d_{tt'} - \frac{1}{2} \sum_{t''} a_{t''\mathbf{s}} d_{t't''} \right), \forall t, \mathbf{r}. \tag{6}$$

Equation (6) is obtained considering that diagonal elements of the dissimilarity matrix are equal to zero and that the dissimilarity matrix is symmetric. The weighting factors $a_{t\mathbf{r}}$ are given by:

$$a_{t\mathbf{r}} = \frac{\sum_{\mathbf{s}} h_{\mathbf{rs}} P(\mathbf{x}_t \in \mathbf{C_s})}{\sum_{t'} \sum_{\mathbf{s}} h_{\mathbf{rs}} P(\mathbf{x}_{t'} \in \mathbf{C_s})}, \forall t, \mathbf{r} \tag{7}$$

and can be seen as weighted averages over data vectors.

The Soft Topographic Map algorithm for proximity data described above can be summarized in the pseudo code of Table 1. Minimization procedure can be done in two steps, formed by two nested loops. The inner loop 2(a)i constitutes an expectation-maximization (EM) algorithm: starting from a random initialization of partial costs, equations (5), (7), (6) are computed in sequence for a fixed value of β until the difference between current partial costs and previous partial costs is lower than a certain threshold. Then, in the outer loop 2a, in order to find the global minimum of the cost function, β is gradually increased and the inner loop repeated. β is increased according to the annealing scheme $\beta \leftarrow \eta\beta$, with $\eta = 1.1 \ldots 2.0$, up to a previously chosen β_{final}.
As seen in [27], the initial value of β should be just above a certain value β^* calculated as:

$$\beta^* = \frac{1}{\lambda^{\mathbf{C}}_{max}\lambda^{\mathbf{G}}_{max}}, \tag{8}$$

where $\lambda^{\mathbf{C}}_{max}$ is the largest eigenvalue of the covariance matrix $\mathbf{C}$ of the data and $\lambda^{\mathbf{G}}_{max}$ is the largest eigenvalue of a matrix $\mathbf{G}$, whose elements are equal to:

$$g_{\mathbf{rt}} = \sum_{\mathbf{s}} h_{\mathbf{rs}} \left(h_{\mathbf{st}} - \frac{1}{M} \right). \tag{9}$$

5 Experimental Analysis

5.1 Bacteria dataset

In order to test our approach, we have built a database of 16S rRNA bacteria gene sequences. The choice of the bacteria set has been done according to the current taxonomy [1]. We focused on the bacteria belonging to Phylum BXII, Proteobacteria; Class III, Gammaproteobacteria: this class includes some of the most common and dangerous bacteria related to human pathologies. In the Gammaproteobacteria class there are 14 orders, each of them containing one or more family. Each family is divided in genera; for each genus we selected the type strains, as shown in Figure 2.

For each type strain we selected the 16S rRNA gene sequence, which contains approximately 1400 nucleotides. The resulting 147 sequences were retrieved from GenBank [33] in FASTA format [15].

Each gene sequence is labelled according to its order in the actual taxonomy.

5.2 Experimental results

We carried out a set of experimental tests using the algorithm described in the section 4 with the bacteria dataset of section 5.1. We used both the dissimilarity matrices obtained from multiple alignment of sequences and pairwise alignment of sequences in order to compare the results. More specifically, we used two well known bioinformatic tools: Mega software [34], that implements ClustalW algorithm, and Emboss tools [35] for Needleman-Wunsch algorithm. In both situations, we used default options.

		Order Name	Number of Families	Number of Type Strains
Gammaproteobacteria	□	Chromatiales	3 Families	25 Type Strains
	△	Xanthomonodales	1 Family	11 Type Strains
	●	Thiotrichales	3 Families	11 Type Strains
	▣	Methylococcales	1 Families	7 Type Strains
	◎	Pseudomonadales	2 Families	7 Type Strains
	★	Vibrionales	1 Family	3 Type Strains
	◉	Enterobacteriales	1 Family	39 Type Strains
	○	Acidithobacillales	2 Families	2 Type Strains
	■	Cardiobacteriales	1 Family	3 Type Strains
	▲	Legionellales	2 Families	3 Type Strains
	◎	Oceanospirillales	4 Families	11 Type Strains
	⊠	Alteromonadales	1 Family	13 Type Strains
	☆	Aeromonadales	2 Families	7 Type Strains
	◭	Pasteurellales	1 Families	6 Type Strains
				147 Type Strains

Fig. 2. Actual taxonomy of the bacteria dataset

We applied a slightly tuned version of Soft Topographic Map algorithm: in order to speed up processing time, neighbourhood 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 side of maps. 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 25 learning epochs; finally we put the width of neighbourhood functions σ to 0.5.

We generated several maps of different dimensions, from 8×8 up to 20×20 neurons. The dimensions of the maps were set by considering the number of input patterns (147 gene sequences) and the number of expected clusters (14 orders in the taxonomy). We compared each pair of maps of the same dimension obtained from multiple alignment and pairwise alignment. The results were quite similar and we can state that the alignment technique does not affect final results.

In Figures 3, 4, 5, we show the results provided by 12×12, 16×16, 20×20 maps, trained with the dissimilarity matrix using the pairwise alignment. In the maps, bright areas denote proximity and dark zones represent distance, according to the U-Matrix style [32].

It should be noticed that in larger maps the units tend to classify homogeneous patterns better. Namely, comparing the maps we can observe that the number of bacteria belonging to mixed clusters, i.e units containing bacteria of

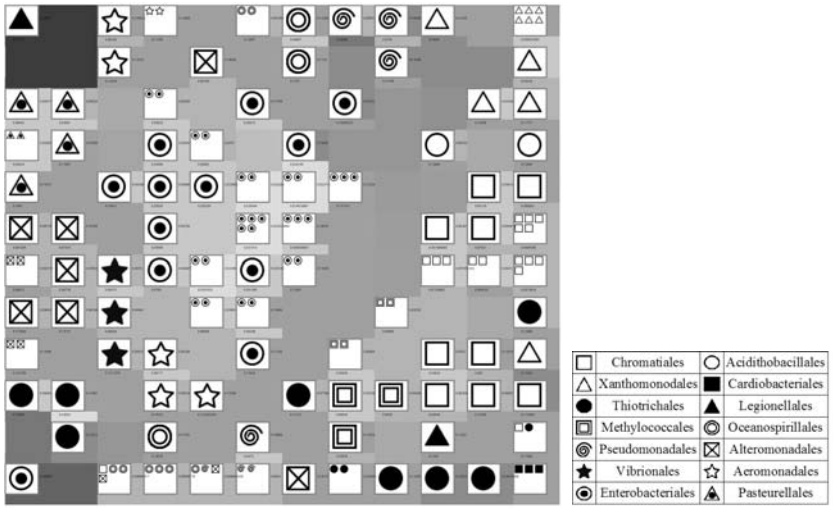

Fig. 3. 12×12 topographic map of bacteria dataset

different orders, decreases as the number of neurons increases (Figure 6). Therefore, the 20×20 map is the most accurate. In all the maps most of the bacteria are classified according to their order in the actual taxonomy. We can also see that bacteria belonging to the order "Enterobacteriales" are split into a series of adjacent clusters in the central part of the map. This could mean that the order "Enterobacteriales" could be subdivided into distinct families rather than the single one of the actual taxonomy (see Figure 2).

Finally, an interesting result is that there are some anomalies that are constant for all the tests regardless of the chosen map dimension and alignment algorithm. For example, in small maps (not shown here) the "Alterococcus agarolyticus" 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 (e.g. at the lower left corner of Figure 3 and at upper left conrner of Figures 4 and 5). Another interesting example is given by "Legionella pneumophila" bacterium of "Legionellales" order: that in all maps is located in a corner of the grid and surrounded by a dark grey area. This would suggest that it can be considered to have an order of its own. In general, we noticed that in the transition from smaller maps to larger ones there is always a set of bacteria that show the following anomalies:

- bacteria belonging to mixed clusters and far from their homologous bacteria,
- isolated bacteria in a single cluster far from their homologous bacteria.

In the former case, it is possible that those bacteria were either incorrectly classified or incorrectly registered into GenBank. In the latter, it is very likely

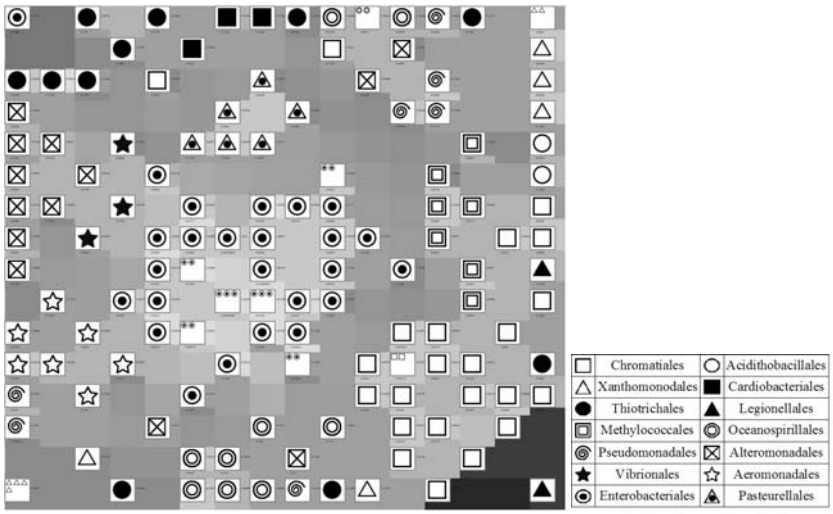

Fig. 4. 16 × 16 topographic map of bacteria dataset

that those bacteria could form new orders or families that have not been discovered by analyzing only phenotypic features.

In conclusion, although the topographic maps have shown a clustering that generally reflects the current taxonomy, some singular cases have been detected. The proposed approach is a first attempt to provide an innovative tool to support the correction of genetic sequence submission systems (e.g. GenBank) and to build a genotypic features based taxonomy.

6 Conclusions

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 clustering and classification 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 directly from the genetic data, without 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. The map allows an easy identification of cases that could represent incorrect classification or incorrect registration in the database. 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.

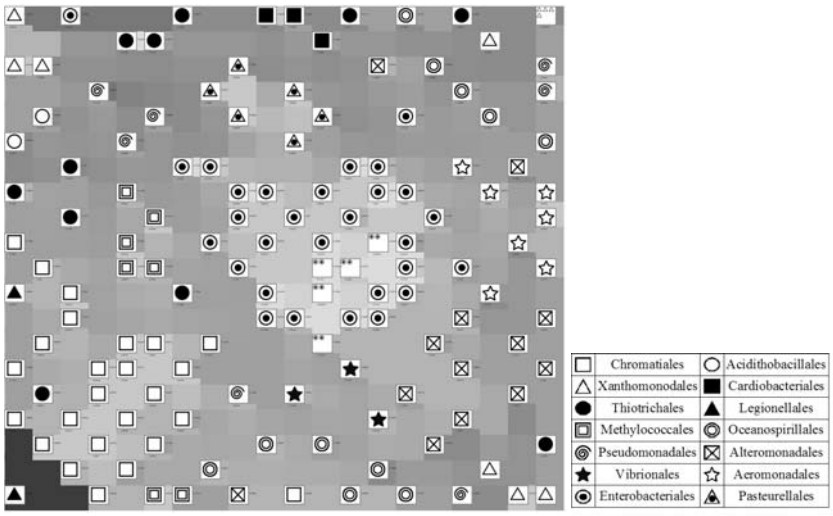

Fig. 5. 20 × 20 topographic map of bacteria dataset

References

1. Garrity, G. M., Julia B. A. and Lilburn T. 2004. The revised road map to the manual, p. 159-187. *In* G. M. Garrity (ed), Bergey's manual of systematic bacteriology. Springer-Verlag, New York, N.Y.
2. I. T. Joliffe, *Principal Component Analysis,* Springer-Verlag, New York, 1986.
3. Clarridge, Jill E., III. Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases Clin. Microbiol. Rev. 2004 17: 840-862
4. Drancourt, Michel, Bollet, Claude, Carlioz, Antoine, Martelin, Rolland, Gayral, Jean-Pierre, Raoult, Didier. 16S Ribosomal DNA Sequence Analysis of a Large Collection of Environmental and Clinical Unidentifiable Bacterial Isolates J. Clin. Microbiol. 2000 38: 3623-3630
5. Drancourt, M., Berger, P., Raoult, D. Systematic 16S rRNA Gene Sequencing of Atypical Clinical Isolates Identified 27 New Bacterial Species Associated with Humans J. Clin. Microbiol. 2004 42: 2197-2202
6. Drancourt, M., Raoult, D. Sequence-Based Identification of New Bacteria: a Proposition for Creation of an Orphan Bacterium Repository J. Clin. Microbiol. 2005 43: 4311-4315
7. M. Oja, P. Somervuo, S. Kaski, and T. Kohonen, *"Clustering of human endogenous retrovirus sequences with median self-organizing map"*, in WSOM'03 Workshop on Self-Organizing Maps, 9-14 Sep 2003.
8. Butte, A.J., and Kohane, I.S. (2000) *Mutual information relevance networks: functional genomics clustering using pairwise entropy measurements.* Proc. Pacific Symposium on Biocomputing, 5, 415-426.
9. P. Somervuo and T. Kohonen, *Clustering and visualization of large protein sequence databases by means of an extension of the self-organizing map,* in Discovery Science. Proceedings of the Third International Conference (2000), pp. 76–85.

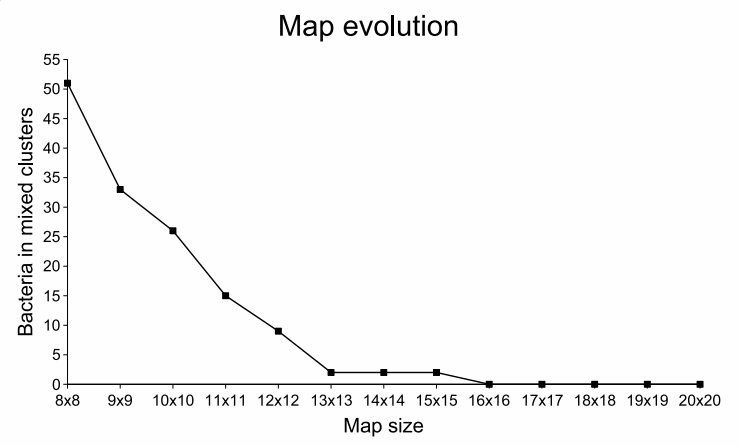

Fig. 6. Bacteria in mixed clusters w.r.t. map size

10. Teuvo Kohonen and Panu Somervuo, How to make large self-organizing maps for nonvectorial data, Neural Networks, Volume 15, Issues 8-9, October-November 2002, Pages 945-952.
11. Yonghui Chen; Reilly, K.D.; Sprague, A.P.; Zhijie Guan. SEQOPTICS: A Protein Sequence Clustering Method. Computer and Computational Sciences, 2006. IMSCCS '06. First International Multi-Symposiums on, Vol.1, Iss., 20-24 June 2006 Pages: 69- 75
12. Ankerst M, Breunig MM, Kriegel HP, Sander J: OPTICS: Ordering Points To Identify the Clustering Structure. In *SIGMOD 1999, Proceedings ACM SIGMOD International Conference on Management of Data, June 13, 1999, Philadelphia, Pennsylvania, USA*, 1999:49-60.
13. Maido Remm, Christian E. V. Storm and Erik L. L. Sonnhammer, Automatic clustering of orthologs and in-paralogs from pairwise species comparisons, Journal of Molecular Biology, Volume 314, Issue 5, 14 December 2001, Pages 1041-1052.
14. Altschul S, Gish W, Miller W, Myers E, Lipman D: Basic local alignment search tool. *J Mol Biol* 1993, **232**:584-99.
15. http://www.ncbi.nlm.nih.gov/blast/fasta.shtml
16. S. Dubnov, R. El-Yaniv, Y. Gdalyahu, E. Schneidman, N. Tishby, and G. Yona. *A new nonparametric pairwise clustering algorithm based on iterative estimation of distance profiles* Machine Learning, 47, 3561, 2002
17. J. Buhmann, T. Zoller, "Active Learning for Hierarchical Pairwise Data Clustering," *icpr* p. 2186, 2000.
18. Thomas Hofmann and Joachim M. Buhmann. *Hierarchical pairwise data clustering by mean–field annealing.* In Proceedings of ICANN'95, NEURON IMES'95, volume II, pages 197–202. EC2 & Cie, 1995.
19. Graepel, T., Herbrich, R., Bollmann-Sdorra, P., and Obermayer, K., "*Classification on Pairwise Proximity Data,* " NIPS.
20. T. Hofmann and J. Buhmann. *Multidimensional scaling and data clustering.* In G. Tesauro, D. S. Touretzky, and T. K. Leen, editors, Advances in Neural Information Processing Systems 7, pages 459–466. Cambridge, Mass: MIT Press, 1995.

21. Hansjrg Klock and Joachim M. Buhmann. *Multidimensional scaling by deterministic annealing* In Springer Lecture Notes in Computer Science Venice, editor, Proceedings of the International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition, EMMCVPR'97, volume 1223, pages 246–260, May 1997.
22. W. S. Torgerson, "*Multidimensional scaling: I. Theory and method,*" Psychometrika, vol. 17, pp. 401–419, 1952.
23. 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, 22:4673–4680, 1994.
24. Needleman, S. B. and Wunsch, C. D. (1970) J. Mol. Biol. 48, 443-453.
25. T. H. Jukes and C. R. Cantor, *Mammalian Protein Metabolism,* H. N. Munro, editors, Academic Press, New York, 1969, ch. Evolution of Protein Molecules, pp. 21– 132.
26. S. P. Luttrell, "*A Bayesian analysis of self-organizing maps,*" Neural Comput., vol. 6, pp. 767–794, 1994.
27. T. Graepel, M. Burger, and K. Obermayer. *Self-organizing maps: generalizations and new optimization techniques.* Neurocomputing, 21:173–190, 1998.
28. Graepel, T. and Obermayer, K. (1999). *A stochastic self organizing map for proximity data.* Neural Computation, 11:139–155.
29. T. Hofmann and J. M. Buhmann, "*Pairwise data clustering by deterministic annealing,*" IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, pp. 1–14, 1997. 154
30. Rose, K., "*Deterministic Annealing for Clustering, Compression, Classification, Regression, and Related Optimization Problems,* " Proc. of the IEEE, Vol. 86:11, pp.2210-2239, 1998.
31. Teuvo Kohonen. *Self-organizing maps.* Springer, Berlin; Heidelberg; New-York, 1995.
32. Ultsch, A. U*-Matrix: a Tool to visualize Clusters in high dimensional Data, Technical Report No. 36, Dept. of Mathematics and Computer Science, University of Marburg, Germany, (2003)
33. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide
34. S Kumar, K Tamura, and M Nei (2004) "MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment" *Briefings in Bioinformatics* 5:150-163.
35. Rice,P. Longden,I. and Bleasby,A. *EMBOSS: The European Molecular Biology Open Software Suite* (2000) Trends in Genetics **16**, (6) *pp276–277*