Comparison of Genomic Sequences Clustering Using Normalized Compression Distance and Evolutionary Distance

Massimo La Rosa[1,2], Riccardo Rizzo[2], Alfonso Urso[2],
and Salvatore Gaglio[1,2]

[1] Dipartimento di Ingegneria Informatica, Universitá di Palermo, Italy
[2] ICAR-CNR, Consiglio Nazionale delle Ricerche, Palermo, Italy

Abstract. Genomic sequences are usually compared using evolutionary distance, a procedure that implies the alignment of the sequences. Alignment of long sequences is a long procedure and the obtained dissimilarity results is not a metric. Recently the normalized compression distance was introduced as a method to calculate the distance between two generic digital objects, and it seems a suitable way to compare genomic strings. In this paper the clustering and the mapping, obtained using a SOM, with the traditional evolutionary distance and the compression distance are compared in order to understand if the two distances sets are similar. The first results indicate that the two distances catch different aspects of the genomic sequences and further investigations are needed to obtain a definitive result.

1 Introduction

In recent years, the growing availability of biological data has driven computer scientists to develop algorithms and methodologies able to support the analysis of this kind of information. Typical biological data are, for instance, genetic and protein sequences, molecular structures, chemical compounds, gene expressions data. Given this large amount of raw data, one of the most used approach to extract useful and functional information is to perform data mining techniques, such as unsupervised or supervised clustering.

Our work focuses, in particular, on biological data made up of bacteria DNA sequences, that can be easily found in very large databases like GenBank [1] or EMBL [2], and aims at finding a simple and reliable clustering tool that can help biologists in their studies concerning species identification and phylogenesis.

One of the major issue is related to the information content of nucleotide sequences: they are not directly representable using a feature space, so the only information we have is in terms of pairwise similarities/dissimilarities.

There exists several types of distances for comparing two protein or nucleotide sequences, the so called "evolutionary distances" [3], and all of these methods are based on the concept of sequence alignment [4,5]. This methodology has, however, at least two major drawbacks: first of all, the algorithms that perform

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part III, LNAI 5179, pp. 740–746, 2008.

the alignments are often computationally expensive, especially when aligning very long sequences, and depend on several parameters; secondly, evolutionary distances are not metrics, that is triangular inequality does not hold.

For all of these reasons, it is more advantageous the use of a distance metric, alignment-free, that overcomes the limitations remarked above. Recently, Li et al. [6] have developed a metric, based on Kolmogorov complexity [7]. This similarity metric, called "Universal Similarity Metric" (USM), can be generally applied to compare any two objects that can be represented as a binary string; applying USM to nucleotide sequences, it is possible to obtain a parameter-free similarity measure, having the properties of a metric, without any *apriori* alignment.

The aim of this work is to use the USM in order to compute a dissimilarity matrix of a bacteria dataset, made up of a single "housekeeping gene"; from this matrix we shall perform an unsupervised clustering, using an extension of the classical Self-Organizing Map algorithm [8], tuned up to cope with input data expressed only in terms of their pairwise dissimilarities. The choice of a dataset containing only one "housekeeping" gene, more specifically the 16S rRNA gene, is justified by recent trends in bacterial genomics that have shown how this gene is particularly suitable for clustering and classification purposes [9,10]. The obtained clustering results will be compared with the ones coming from a dissimilarity matrix computed with evolutionary distance on the same dataset. With this approach, we want to test the effectiveness of USM applied to short DNA sequences, about 1400 base pairs, rather than whole genomes, as a valid input data in order to perform a meaningful unsupervised clustering of bacteria.

2 Related Work

The use of USM for DNA sequences comparison has been done in [6] and [11]. In those papers, a dissimilarity matrix of complete mammalian mithocondrial genomes has been computed in order to build a phylogenetic tree or to visualize the data through the Multidimensional Scaling technique [20].

A deeper study of USM and its applications to biological datasets is presented in [23]. Authors of this paper tested three approximations of USM by using 25 different compressors (see next section) on six biological datasets. Then, from the dissimilarity matrices obtained, they built several phylogenetic trees and compared them with the corresponding trees created using the classical evolutionary distance. From all of these experimental results, they further validated the use of USM for biological sequences.

A clustering approach, using Median Som [18], based on protein sequences and evolutionary distances was carried out in [12]. Median Som was also adopted by [13] for clustering of human endogenous retrovirus sequences, given a distance matrix based on the FASTA similarity scores [14].

The generation of a topographic map was at the basis of the work of [15], in which the algorithm proposed by [16] was used to obtain a visualization of a well defined bacteria dataset, containing only the DNA sequences of *type strains*, that is sample species, of the Gammaproteobacteria class.

3 Universal Similarity Metric

Li et al. introduced the concept of Universal Similarity Metric in [6], using the theory of Kolmogorov complexity [7]. The USM represents a class of distance measures obtained as an approximation of Kolmogorov complexity, since this one is not computable. Among this class of measures, we will consider the one called "Normalized Compression Distance" (NCD), in which the Kolmogorov complexity of an object x is approximated with the size of its compressed version.

So, given a real-word reference compressor C, NCD is defined as:

$$\mathrm{NCD}(x,y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}} \tag{1}$$

where $C(xy)$ is the compressed size of the concatenation of x and y, and $C(x)$ and $C(y)$ are the compressed sizes respectively of x and y. Roughly speaking, NCD is a number between 0 and 1: the smaller the number, the more similar are the objects. In our system, the generic objects x and y are the nucleotide sequences, that can be seen as strings composed of the four letters of DNA alphabet: $\{a, c, g, t\}$.

Using a normal compressor C, according to the definitions in [11], NCD has the properties of a similarity metric. A large set of real-world compressors are normal compressors: we have decided to use *GenCompress* [17], a normal compression algorithm designed to work with DNA sequences and able to give the best compression ratios, providing this way the best approximation of Kolmogorov complexity.

4 Clustering Algorithms

One of the most used algorithms for data clustering is Kohonen's Self-Organizing Map (SOM) [8]. SOM algorithm has also the great advantage to provide a direct visualizations of clusters, because it defines a set of neurons, also called models, that are arranged in a rectangular lattice, creating a topographic map. During learning phase, each input pattern is mapped to its closest neuron, known as best matching unit (*bmu*): when the learning is complete, the coordinates of each pattern into the two-dimensional lattice are the same of its own *bmu*.

Unfortunately, SOM does not work with dataset expressed only in terms of pairwise dissimilarities, as in our situation. An adaptation of SOM to non-vectorial data is the so called Median SOM [18]. Median SOM, however, has a strong limitation: the neurons are selected among the set of patterns, providing this way a severe restriction especially for small datasets. For this reason, we used a novel technique, called Relational Topographic Map [19], that is able to work with nonvectorial data and, at the same time, overcomes the limitation introduced by Median SOM.

Relational Topographic Map [19], differently from Median SOM, represents the neurons as linear combinations of input patterns. If we indicate with $\boldsymbol{x}^j$ the generic pattern for which the distance $\left\|\boldsymbol{x}^i - \boldsymbol{x}^j\right\|^2$ exists, and and if we consider the generic neuron $\boldsymbol{w}^i$ as a linear combination of input patterns, that is $\boldsymbol{w}^i = \sum_j \alpha_{ij}\boldsymbol{x}^j$, with $\sum_j \alpha_{ij} = 1$, then $\left\|\boldsymbol{w}^i - \boldsymbol{x}^j\right\|^2$ is equal to:

$$\left\|\boldsymbol{w}^i - \boldsymbol{x}^j\right\|^2 = (D \cdot \alpha_i)_j - 1/2 \cdot \alpha_i^t \cdot D \cdot \alpha_i, \tag{2}$$

where D is the pairwise distance matrix among patterns (more details in [19]). Moreover, the distance between two neurons on the map can be computed as follows:

$$\left\|\boldsymbol{w}^i - \boldsymbol{w}^j\right\|^2 = \alpha_j^t \cdot D \cdot \alpha_i - 1/2 \cdot \alpha_j^t \cdot D \cdot \alpha_j - 1/2 \cdot \alpha_i^t \cdot D \cdot \alpha_i \tag{3}$$

Using this transformation, the distance between neurons and patterns will depend only on the pairwise dissimilarities among input patterns. During learning phase, the weights of these linear combinations are updated and, as usual, input patterns are mapped to their *bmu*. As demonstrated by its authors, Relational Topographic Map gives better clustering results than Median SOM because the neurons are not constrained to overlap to the patterns.

5 Experimental Results

5.1 Dataset Description

All of our experiments were conducted on a test dataset derived from the one used by Garrity and Lilburn in [24]. From the original dataset, consisting of 1436 small subunit ribosomal RNA sequences, we selected only the 16S rRNA sequences, whose length is about 1400 bp, obtaining a total of 1285 gene sequences. The sequences were labeled with their own order, using the same classification as in [24]. There is a total of 16 orders, and 18 bacteria sequences are unclassified. From this bacteria sequences dataset, two dissimilarity matrices were computed: the first matrix was computed using the evolutionary distance with the Jukes and Cantor method [21]; the second dissimilarity matrix was obtained using NCD, according to (1), and GenCompress compressor.

5.2 Algorithm Setup

We trained, with the Relational Topographic Map algorithm, and with both dissimilarity matrices, 40 square maps of 10×10 dimension, and for each configuration we obtained 40 maps. We chose to consider 10×10 maps, since having 16 preclassified orders, with 100 neurons there are about six neurons per order. Each pair of corresponding maps, trained with the two input matrices, had the same initialization and a learning process of 40 epochs.

5.3 Map Analysis

A first comparison between maps was done as follows: each cell in a map was given a unique label according to the Order of the most of the elements belonging

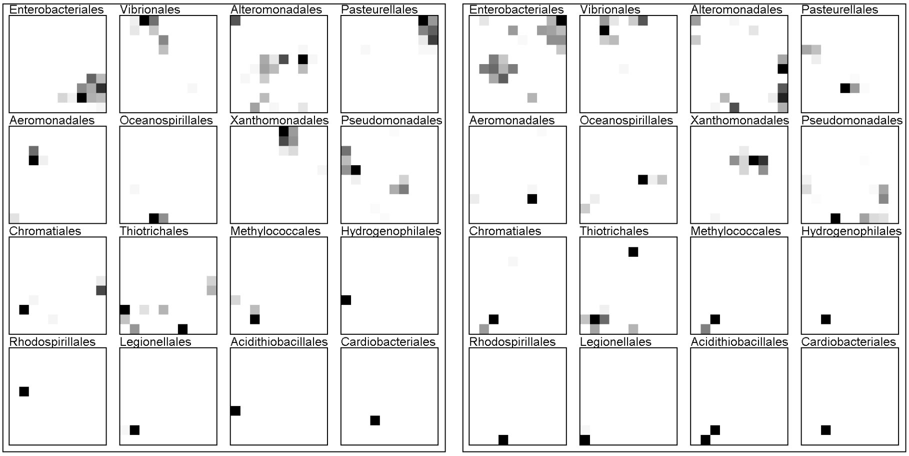

Fig. 1. Map trained with evolutionary distance, on the left, and with NCD, on the right. Both maps contains the lowest number of misplaced patterns. Considering the two dissimilarity matrices, we obtained a quite similar clustering, but different topological placement.

to that cell; then, for each cell, we counted up the number of misplaced patterns, according to the label previously assigned. We found that the maps trained with evolutionary distances and NCD have, on average, respectively 29.75 and 37.25 misplaced patterns, and considering the total number of 1285 elements of our dataset, that means we obtain an error rate, respectively, of 2.3% and 2.8%. This means the two distances alow to obtain a quite similar clustering results.

Furthermore, we applied the measure of dissimilarity defined in [22], which allows to compare two Self-Organizing Maps.

This measure has been adapted to work with Relational Topographic Map, using (2) to compute the distance between each input pattern and its own *bmu*, and (3) to compute the distance between neurons on the map; then it was computed for every pair of maps trained with the two input matrices: the result we obtained is that there is a mean similarity of 21.61% among the two sets of maps referred to the two dissimilarity matrices. This low value of similarity can be interpreted considering that pairwise matrix computed through NCD represents a very strong perturbation with respect to the pairwise matrix computed through evolutionary distance; so, even if the two clusterings are quite similar, according to the error rate decribed above, the two maps, from a topographic point of view, are very different.

In Fig. (1), we show the two maps, on the left the one trained with evolutionary distance and on the right the other one trained with NCD, with the lowest number of misplaced elements. The visualization is organized in submaps: each submap represents the distribution of one order, written at the top of the submap itself. The gray level of each cell is proportional to the number of sequences belonging to that node: in each submap, the maximum number of elements in a node is given the darkest shade of gray. It is possible to see that, for example, a large group like “Enterobacteriales”, in the map obtained with NCD is split into two main clusters, whereas in the map obtained with evolutionary distance it preserves an evident compactness. Another difference is between “Thiotricales” orders: in the upper map it is spread through the map, while in the lower map it forms two clusters: one at the bottom left corner and another one at the top of the map. As for the other orders, in both maps they exhibit well defined clusters, although their positions upon maps do not coincide.

This situation reflects the fact that, considering the two dissimilarity matrices, we obtained a quite similar clustering, but different topological placement.

6 Conclusions

In previous studies, Normalized Compression Distance has been used to create phylogenetic trees of complex species based on whole genomes. In this work, we applied NCD and classical evolutionary distance in order to obtain two dissimilarity matrices of a dataset composed of the 16S rRNA gene sequences of bacteria. Our main goal was to use these two dissimilarity matrices as input data for an unsupervised clustering process, performed with an extension of Self-Organizing Map, and to compare the clusterings computed this way. We found that the two matrices provide a quite similar clustering result. On the other hand, using a distance measure to compare the maps, we found that, although with the same initialization, the maps, in terms of topological positioning, are different. From these two contrasting results we can state that further investigations are needed in order to understand the bind between Normalized Compression Distance and evolutionary distance and, subsequently, if NCD can substitute evolutionary distance in genome comparisons.

References

1. National Center for Biotechnology Information, Entrez Nucleotide query, `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide`
2. European Molecular Biology Laboratory, `http://www.ebi.ac.uk/embl/`
3. Nei, M., Kumar, S.: Molecular Evolution and Phylogenetics. Oxford University Press, New York (2000)
4. Needleman, S.B., Wunsch, C.D.: J. Mol. Biol. 48, 443–453 (1970)
5. Thompson, J.D., Higgins, D.G., Gibson, T.J.: 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)
6. Li, M., Chen, X., Li, X., Ma, B., Vityi, P.M.B.: The similarity metric. IEEE Trans. Inf. Theory 50(12), 3250–3264 (2004)
7. Li, M., Vitanyi, P.M.B.: An Introduction to Kolmogorov Complexity and its Applications, 2nd edn. Springer, New York (1997)
8. Kohonen, T.: Self-organizing maps. Springer, Heidelberg (1995)

9. Drancourt, M., Bollet, C., Carlioz, A., Martelin, R., Gayral, J., Raoult, D.: 16S Ribosomal DNA Sequence Analysis of a Large Collection of Environmental and Clinical Unidentifiable Bacterial Isolates. J. Clin. Microbiol. 38, 3623–3630 (2000)
10. Drancourt, M., Berger, P., Raoult, D.: Systematic 16S RNA Gene Sequencing of Atypical Clinical Isolates Identified 27 New Bacterial Species Associated with Humans. J. Clin. Microbiol. 42, 2197–2202 (2004)
11. Cilibrasi, R., Vitanyi, P.M.B.: Clustering by Compression. IEEE Trans. Inf. Theory 51(4), 1523–1545 (2005)
12. Somervuo, P., Kohonen, T.: Clustering and visualization of large protein sequence databases by means of an extension of the self-organizing map. In: Proceedings of the Third International Conference on Discovery Science, pp. 76–85 (2000)
13. Oja, M., Somervuo, P., Kaski, S., Kohonen, T.: Clustering of human endogenous retrovirus sequences with median self-organizing map. In: WSOM 2003 Workshop on Self-Organizing Maps, September 9-14, 2003 (2003)
14. Pearson, W., Lipman, D.: Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85, 2444–2448 (1988)
15. La Rosa, M., Di Fatta, G., Gaglio, S., Giammanco, G.M., Rizzo, R., Urso, A.: Soft Topographic Map for Clustering and Classification of Bacteria. In: R. Berthold, M., Shawe-Taylor, J., Lavrač, N. (eds.) IDA 2007. LNCS, vol. 4723, pp. 332–343. Springer, Heidelberg (2007)
16. Graepel, T., Burger, M., Obermayer, K.: Self-organizing maps: generalizations and new optimization techniques. Neurocomputing 21, 173–190 (1998)
17. Chen, X., Kwong, S., Li, M.: A compression algorithm for DNA sequences. Engineering in Medicine and Biology Magazine 20(4), 61–66 (2001)
18. Kohonen, T., Somervuo, P.: How to make large self-organizing maps for nonvectorial data. Neural Networks 15(8-9), 945–952 (2002)
19. Hasenfuss, A., Hammer, B.: Relational Topographic Maps. In: R. Berthold, M., Shawe-Taylor, J., Lavrač, N. (eds.) IDA 2007. LNCS, vol. 4723, pp. 93–105. Springer, Heidelberg (2007)
20. Torgerson, W.S.: Multidimensional scaling: I. Theory and method. Psychometrika 17, 401–419 (1952)
21. Jukes, T.H., Cantor, R.R.: Evolution of protein molecules. In: Munro, H.N. (ed.) Mammalian Protein Metabolism, pp. 21–132. Academic Press, New York (1969)
22. Kaski, S., Lagus, K.: Comparing Self-Organizing Maps. In: Proceedings of the 1996 International Conference on Artificial Neural Networks (1996)
23. Ferragina, P., Giancarlo, R., Greco, V., Manzini, G., Valiente, G.: Compression-based classification of biological sequences and structures via the Universal Similarity Metric: experimental assessment. BMC Bioinformatics 8, 252 (2007)
24. Garrity, G.M., Lilburn, T.G.: Self-organizing and self-correcting classifications of biological data. Bioinformatics 21(10), 2309–2314 (2005)