Clustering of bacteria by the triplet composition of 5S RNA

Iuliia Ovchinnikova[1*], *Michael* Sadovsky[2], and *Vladislav* Sokolov[1]

[1]Siberian Federal University, 79, Svobodny av., Krasnoyarsk, 660041, Russian Federation
[2]Institute for Computational Modeling SB RAS, 50/44, street Akademgorodok, Krasnoyarsk, 660036, Russian Federation

Abstract. The relationship between the structure of nucleotide sequences, the function they encode, and the taxonomy of the host is an important subject of study for molecular biologists and bioinformaticians, as well as specialists in processing big data arrays. There are various approaches in this direction; one of them is connected with the study of the relationship be-tween the structure of biological macromolecules (specifically, bacterial ribosomal RNA molecules) and the taxonomy of their carriers. For these organisms, 16S RNA sequences are a classic subject of study. However, of no less interest is the study of the relationship be-tween structure and taxonomy for other sequences. In the framework of this work, we studied the relationship between structure and taxonomy using bacterial 5S RNA sequences as an example.

1 Introduction

The structure of biological macromolecules can be determined in many different ways; in the framework of this work, we will understand the structure as a frequency dictionary of triplets (sometimes also called a frequency profile); here the index lists the sequences in question. Under the frequency dictionary of triplets, we mean a list of all triplets occurring in the studied sequence, indicating the frequencies of these triplets. Building a frequency dictionary converts a nucleotide (symbolic) sequence into points in a 63-dimensional metric space, making them mathematical objects that allow you to use the entire arsenal of appropriate tools for your study.

This work has two goals:

1) to study the peculiarities of clustering of frequency dictionaries of bacterial 5S RNA gene triplets in terms of the relationship between such clustering and the taxonomic composition of clusters (if any are found). Note that bacterial 5S RNA is not the most widely used object for this kind of research, and comparison of the results with others, for example, those based on clustering based on bacterial 16S RNA genes, determines the relevance of this work;
2) to study the stability of observed clustering (if there is any) to random selection of excluded overrepresented taxa.

* Corresponding author: july.l4o6@mail.ru

2 Materials and methods

Previously [1 – 3], it was shown that the triplet composition of various genetic systems correlates very well with the taxonomic position of the carriers of the corresponding genes. However, the species composition of genetic databases (in our case, *SILVA* databases) is often very displaced: some low-level taxa are represented by a disproportionately large number of records (for example, at the level of strains, etc.), which leads to a distortion of the distribution pattern of frequency dictionaries in the metric space.

The standard way out of this situation is to index the database: deleting some of the records that are in overrepresented groups. In this case, a natural question arises: what is the influence of the choice of deleted objects on clustering? The answer to the question is not known in advance, so one of the objectives of this work is to check how great such an influence is.

Let's proceed to the description of the work itself and the results. The *SILVA* gene database contains a total of 182697 bacterial 5S RNA gene entries; this number of genes was contained in the database prior to indexing. The number of records used in the work is presented Table 1. This table presents the values of the number of genes in the original database before and after indexing. For one of the implementations of the indexed database, the number of entries in it was 49535 genes.

Table 1. The composition of the 5S RNA gene base of bacteria *SILVA* at the phylum level. N is the number of genes in the original database, M is the number of genes after indexing.

Phylum	***N***	***M***	**Phylum**	***N***	***M***
Actinobacteriota	562	291	Fusobacteriota	694	694
Bacteroidota	855	491	Gemmatimonadota	110	55
Campylobacterota	7744	2589	Myxococcota	732	555
Chloroflexi	3180	2429	Nitrospirota	131	33
Cyanobacteria	3090	2316	Patescibacteria	776	360
Deferribacterota	118	59	Planktomycetota	432	296
Deferrisomatota	101	44	Proteobacteria	158610	27612
Deinococcota	432	212	Spirochaetota	1567	1433
Desulfobacterota	708	451	Synergistota	78	78
Elusimicrobiota	126	68	Thermotogota	179	179
Fibrobacterota	74	65	Verrucomicrobiota	1318	1180
Firmicutes	12748	8336			

The stability of clustering with respect to random selection of elements for indexing the bacterial 5S RNA gene base was studied using the elastic map method — see [4–6] for more details. This is a non-linear method that allows you to reduce the data dimension and identify clusters. It should also be emphasized that for the initial data, only 63 triplets are linearly independent - this is due to the fact that the sum of the frequencies of all triplets determined for each gene is equal to one. Thus, one triplet should be excluded from the analysis; formally, any triplet can be excluded, but in practice, one should be excluded for which the standard deviation, determined for the entire database of the studied genes, is minimal. This choice is due to the fact that this triplet makes the least contribution to the distinguishability of genes.

3 Results and discussion

Clustering of bacterial 5S RNA genes according to their frequency dictionaries of triplets was carried out using the freely distributed *VidaExpert* software [5 – 7]. Figure 1 shows the distribution of all twenty-three divisions of the studied bacteria on an elastic map; a soft (16 × 16) map was chosen to study the distribution; all map options were selected by default. The most important conclusion from the results presented in Figure 1 is that bacteria's 5S RNAs are just as taxonomically sensitive as their 16S RNAs; the issue of relative sensitivity and accuracy requires further research.

Figure 1 shows the elastic map clustering results for six different versions of the bacterial 5S RNA gene base. Recall that a comparative study of these clusters obtained on different databases is the main result of our work. In this figure, the points corresponding to each of the phylum listed in Table 1 are shown in different colors (Table 2). You might get the impression that in this figure the number of colors used to highlight the markers is less than 23. However, this is not the case. The apparent decrease in the number of colors is explained by the fact that some of the points are projected onto each other and on elastic maps (Figure 1), presented in internal coordinates, are superimposed on one another and become invisible.

Table 2. Deciphering the color coding of the phylum of bacteria considered in the work.

Phylum	R	G	B	color
Actinobacteriota	255	0	0	
Bacteroidota	0	255	0	
Campylobacterota	128	0	0	
Chloroflexi	255	255	0	
Cyanobacteria	0	255	255	
Deferribacterota	128	128	128	
Deferrisomatota	128	0	128	
Deinococcota	0	0	128	
Desulfobacterota	255	0	128	
Elusimicrobiota	128	128	255	
Fibrobacterota	128	255	0	
Firmicutes	255	0	255	
Fusobacteriota	255	128	0	
Gemmatimonadota	0	128	128	
Myxococcota	128	0	255	
Nitrospirota	255	128	128	
Patescibacteria	0	128	0	
Planctomycetota	0	64	0	
Proteobacteria	0	128	255	
Spirochaetota	128	128	0	
Synergistota	128	64	64	
Thermotogota	64	0	128	
Verrucomicrobiota	255	128	255	

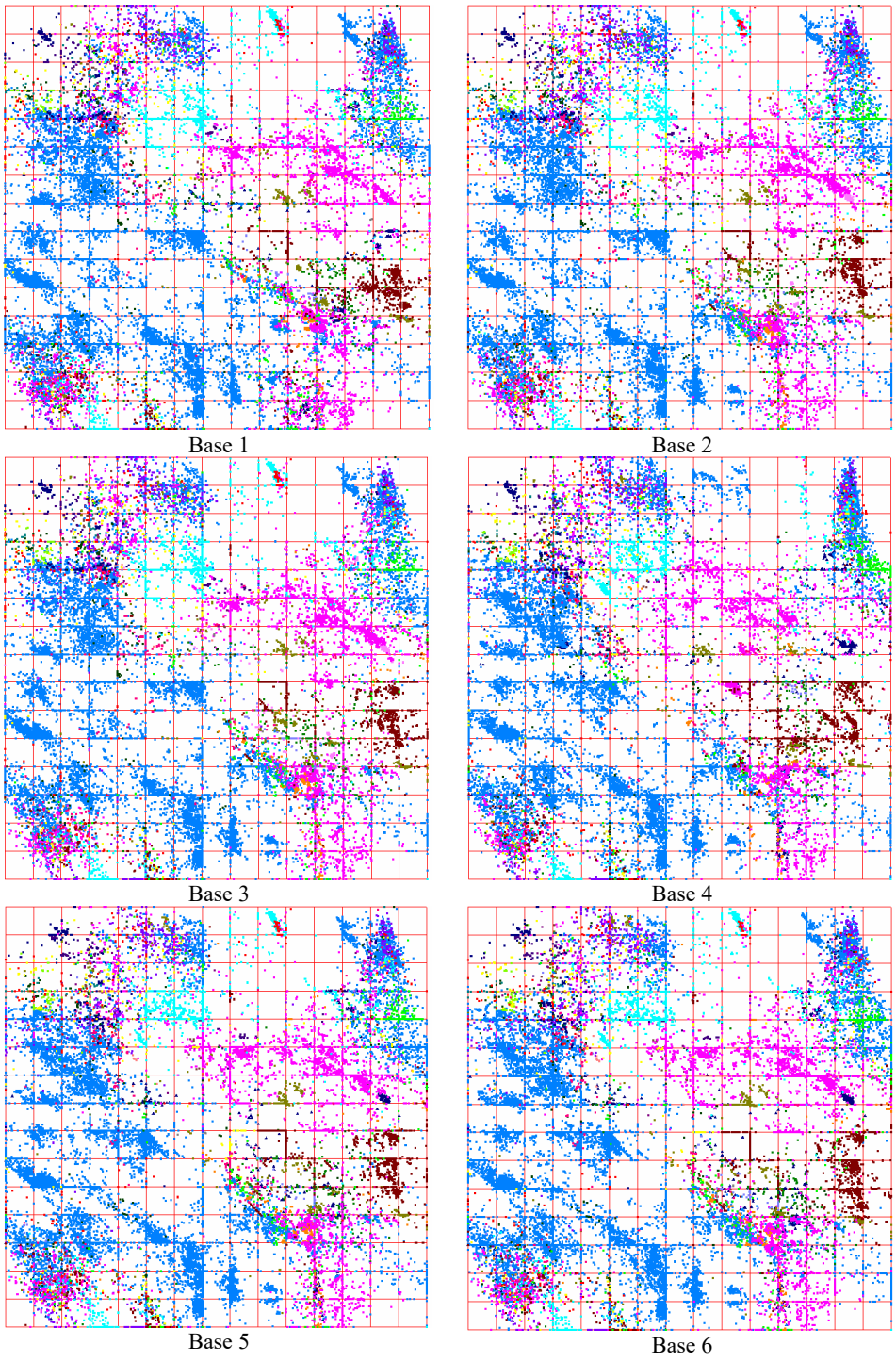

Fig. 1. Distribution of all twenty-three phylums on an elastic map in six different versions of the bacterial 5S RNA gene base.

Before describing the results of studying the stability of clustering to the composition of the gene base, we point out one important and nontrivial result. Previously [2, 3], it was

shown that clustering by the method of elastic maps and classification by the method of dynamic nuclei of bacterial 16S RNA genes reveals a very strong relationship between classes or clusters and the taxonomy of the carriers of these genes. It is natural to expect this kind of behavior from other types of genes; The results presented in this work unambiguously prove that a completely similar pattern of clustering is observed for the bacterial 5S RNA genes. Obviously, the clusters identified by the frequencies of 5S RNA gene triplets in bacteria are just as sensitive to the taxonomic position of the carriers of these genes. In other words, just as for bacterial 16S RNA genes, clusters isolated on the basis of the frequencies of bacterial 5S RNA gene triplets are also formed with a clear taxonomic preference. In particular, it follows from this coincidence that a simultaneous comparative study of the results of classification (clustering) of the same organisms—bacteria in our case—obtained for different genes of their S RNA deserves special attention. However, a detailed discussion of these studies is beyond the scope of this work.

Let us now turn to the description of the main result of this work — the study of the stability of clustering results to accidental deletion of records in those taxa that are overrepresented in the original database. As mentioned above, six databases were created for this purpose, in which gene records in taxa with low representation (i.e., with a small number of species in them, up to 30 records) were excluded from the analysis, since such underrepresented data are not able to generate signal of the proper strength, however, they form noise, which significantly distorts the clustering pattern.

Gene records in overrepresented taxa were partially deleted, and the choice of deleted genes was carried out randomly (indexing). Indexing of overrepresented taxa is necessary in order to suppress too strong a signal from such a group of organisms, which can also significantly distort the results of clustering.

Each of the six databases (Figure 1) were treated with an elastic map with the same parameter values. A comparative analysis of the distributions shown in Figure 1 allows us to state that there is an influence of the composition of the base on which clustering is built; however, this effect is not large. Indeed, if we take any pair of cards presented in Figure 1 and compare them with each other, we can see that there is no literal, pixel-by-pixel correspondence. In other words, if two such cards are placed one on top of the other, they will not match.

On the other hand, the visible differences between these cards are very small. Such discrepancies in the observed distributions can be strictly described and formalized, but we will not deal with this in this paper. We confine ourselves to the fact that all these maps are obviously visually very close to each other. Actually, this observation provides an answer to the main question of this work: is it true that different bases obtained randomly (according to the procedure described above) give a close or even identical result.

This is fully confirmed by Figure 2, which shows maps showing the same department (Campylobacterota) for different database implementations. This figure shows the distribution of bacterial 5S RNA genes, which is exactly the same as that shown in Figure 1. However, in this figure, markers of all departments, except for Campylobacterota, have zero size. In other words, the pattern of distribution density and the distribution itself remain the same, but only the genes of the specified division are visible. Representatives of this department are numerous enough to see characteristic changes in the form of clusters. It is clearly seen that, in general, all six maps are very close to each other: the cluster shown on them has an almost identical shape. Moreover, it can be seen that the pairs of bases (2 – 3) and (5 – 6) give virtually identical patterns of the distribution of this department.

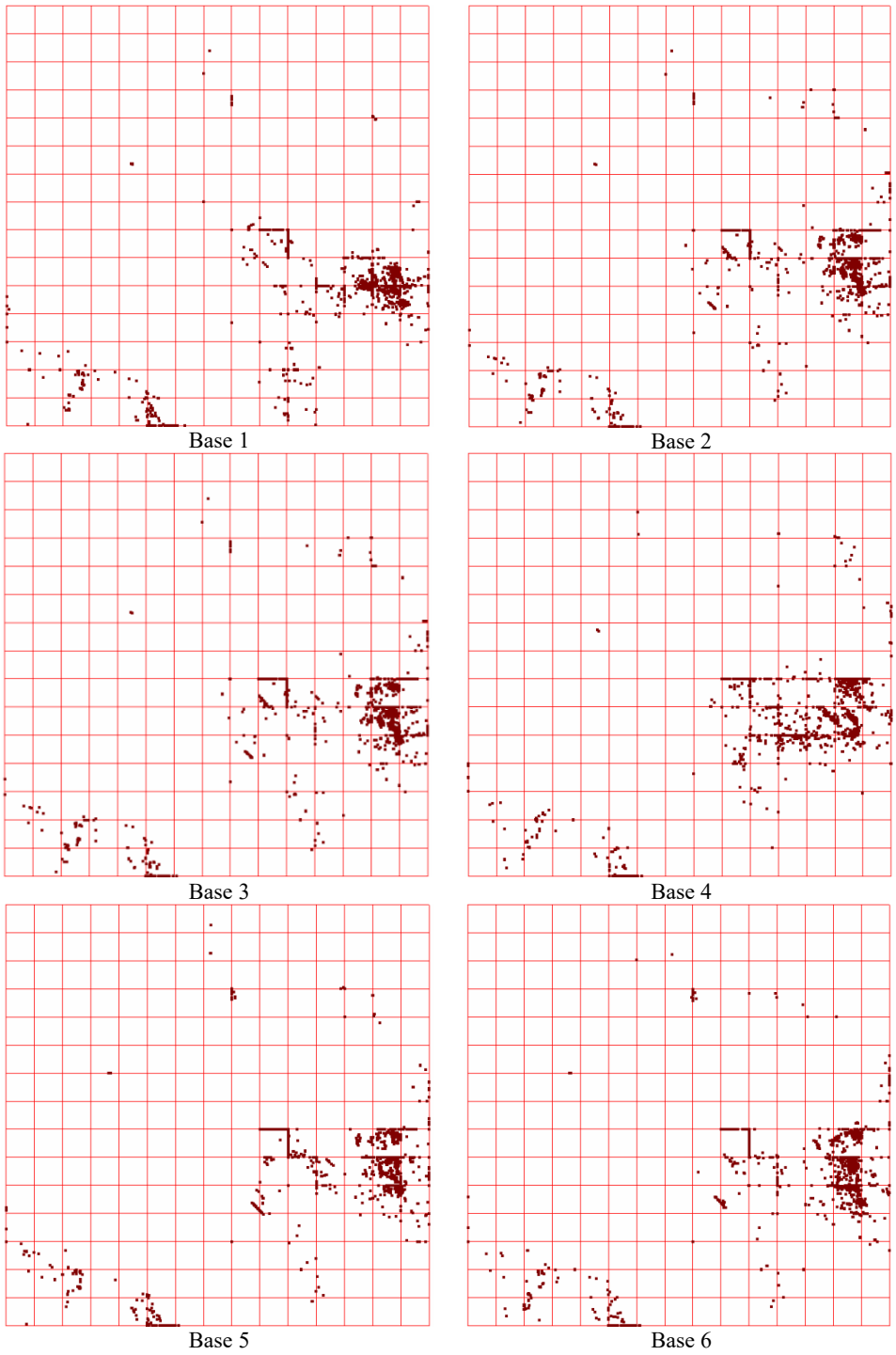

Fig. 2. Individual distribution of 5S RNA genes of Campylobacterota bacteria on a 16 × 16 soft elastic map, for six different database implementations.

4 Conclusion

As can be seen from the presented results, the answer to the question about the absence of a strong influence of random indexing on the clustering pattern from the point of view of the taxonomic composition of the identified clusters is positive: such an influence should be considered very weak and insignificant for further research. The results obtained in this work prove that the composition of the database, determined by the random exclusion of overrepresented records, does not have any significant effect on the results of clustering. This observation allows henceforth to use any of the indexed bases obtained by the random exclusion of overrepresented taxa to investigate the relationship of structure, function, and taxonomy.

This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Grant No.075-15-2022-1121).

References

1. M. Sadovsky, Yu. Putintseva., A. Chernyshova., V. Fedotova, *Genome structure of organelles strongly related to taxonomy of bearers*, International Conference on Bioinformatics and Biomedical Engineering, Springer, pp. 481-490 (2015)
2. A. Gorban, T. Popova, M. Sadovsky, Open Systems & Information Dynamics **7(1)**, 1-17 (2000)
3. A. Teterleva, V. Abramov, A. Morgun, I. Larionova, M. Sadovsky, LNBI **13346**, **Part I,** 205-215 (2022)
4. A. Gorban, A. Zinovyev, *Principal Manifolds for Data Visualization and Dimension Reduction*. Lecture Notes in Computational Science and Engineering (Berlin - Heidelberg - New York, Springer, 2007), **58**, pp. 153-176
5. A. Gorban, A. Zinovyev, *Fast and user-friendly non-linear principal manifold learning by method of elastic maps*. IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015, Campus des Cordeliers, Paris, France, 1-9 (2015)
6. A. Gorban, A. Zinovyev, Computing **75(4)**, 359-379 (2005)