



Unsupervised Classification of Some Bacteria with 16S RNA Genes

Agnia Teterleva¹, Vladislav Abramov⁴, Andrey Morgun³,
Irina Larionova³, and Michael Sadovsky^{2,3,4}

¹ School of Fundamental Biology and Biotechnology of Siberian Federal University,
Krasnoyarsk, Russia

² Institute of Computational Modelling of the Siberian Branch of the Russian
Academy of Sciences, Krasnoyarsk, Russia
msad@icm.krasn.ru

³ Krasnoyarsk State Medical University named after Prof. V.F. Voino-Yasenetsky,
Krasnoyarsk, Russia

⁴ Federal Siberian Research and Clinical Center of FMBA of Russia,
Krasnoyarsk, Russia

Abstract. We used unsupervised nonlinear clustering to reveal the interplay between structure of nucleotide sequences and the taxonomy of their bearers. Triplet frequency composition is referred to a structure, and taxonomy is determined through standard morphology and physiology of bacteria. Soft 16×16 elastic map has been used for clustering. Some preliminary results are presented here approving the high efficiency of such approach to phylogeny analysis. Further applications to medicine are discussed.

Keywords: Triplet frequency · Clustering · Elastic map · Structure

1 Introduction

An interplay between structure and function of genetic entities, and the taxonomy of their bearers still challenges researchers. A lot has been done here (see e.g., [4–6, 14, 20] and much more others). Obviously, the answer depends on the genetic matter taken into consideration: some entities show the strong prevalence of the taxonomy over function [19], while another matter shows the prevalence of the function over taxonomy [3]. This paper aims to further the studies of the interplay mentioned above.

Here we study the relation of the structure encoded in 16 SRNA (that is a triplet frequency dictionary) to taxonomy of the bearers of those molecules. In general, there are three entities: structure of a genetic sequence, the function encoded in it, and taxonomy of the bearer of that former. Ultimately, we aim to study an interplay between all of them. To do it, one must define rigorously what is a structure. Hereafter we shall refer a structure as a triplet frequency dictionary $W^{(j)}$ of the j^{th} 16 SRNA sequence.

The index j enlists the genetic entities to be considered, with respect to their taxonomy. Frequency dictionary (also known as k -mer ensemble) is well-known

object in the studies of nucleotide sequences; it opposes to the widely spread sequences comparison methods based on alignment. The main disadvantage of alignment consists in the necessity to set up some informally determined parameters. On the contrary, k -mer based methods are free from that disadvantage thus providing a researcher with more reliable and formally defined results. The k -mer approach to sequence comparison has a long story and still goes through a progress. We use here classical approach based on the count of the triplet frequency determined over a sequence; however, some expansions of the method could be found in [13], see also papers [1, 2, 21].

One may not expect a dependence on a function of the anticipated patterns in interplay; here the function is the same. Indeed, we study the ribosomal RNA (16S RNA, specifically) genes, hence all of them encode the same function. Thus, we aim to reveal the dependence between triplet composition of the genes and the taxonomy of their bearers.

To reveal the interplay between structure and taxonomy, we do the following steps:

- choose the genetic entities with clearly determined and controlled function;
- convert them into a triplet frequency dictionary $W^{(j)}$ each;
- use up-to-date and powerful methods to cluster the points (frequency dictionaries) in the relevant metric space and identify the clusters;
- check whether a taxonomy of DNA donating organisms or determine the composition of the clusters (if any).

Suppose, the clusters are observed (otherwise no interplay takes place). Here three possible outputs may be:

1. the clusters are apparent, and each cluster comprises the sequences belonging to a specific taxon (or taxa);
2. the clusters are apparent, and each cluster comprises the sequences belonging to organisms of various taxa (maybe, rather distant);
3. an hierarchy in the clusters composition takes place: e.g. there are super-clusters gathering the higher taxa with fine pattern of each super-cluster determined by lower taxonomy position of the organisms.

Here we present some preliminary results on the study of the relation between triplet composition of 16S RNA genes and taxonomy of some bacteria. Ultimately, this work aims to reveal the medically sounding effects in such pattern appearance.

The medical value of a tool to retrieve knowledge from 16S pyrosequencing and the determination of patterns characterizing healthy people vs. patients with various neurological diseases or their predisposition is very high. The reliable changes in qualitative and quantitative diversity of the microbiota for inflammatory bowel diseases (Crohn's disease and ulcerative colitis, Parkinson's disease, Alzheimer's disease, multiple sclerosis, and other neurodegenerative and neuroinflammatory diseases) are reported.

However, the lack of correct and convenient interpretation follows in a severe expansion of time spent on analysis; one must rigorously follow the same protocol that is not always possible elsewhere. However, a diagnosis of a number of

gastroenterological, neurological, and possibly other diseases may be improved. In the future, it will significantly contribute the personalized medical care based on microbiota records. The most ambitious goal here is to create a preventive strategy to correct the human microbiota due to targeted drugs prescription: either eliminating harmful microflora or activating the necessary one. It is necessary to assess the adequacy of the correction being carried out during this treatment procedure.

2 Materials and Methods

2.1 Genetic Material

To reveal the interplay between structure and taxonomy over a set of 16S RNA bacterial genes we use SILVA database¹. It is freely accessible database gathering SRNAs of a great variety of organisms, including bacteria. For the purposes of our study we downloaded 52474 sequences of large subunits of bacterial 16S RNA. The distribution of the genes over taxons is extremely inhomogeneous: some of higher taxa comprise a few species (or strains), while others comprise hundreds or more. Such bias results in a “signal loss”: numerous entries representing higher taxa with few species fail to produce a signal, but make a noise just deteriorating a cluster pattern. To avoid this effect, we hashed the database: we eliminated both over-represented and under-represented taxa. Finally, we tried to balance the representativeness of various taxa in the dataset, so that the entries representing various lower taxa ranged in number from a hundred to tens. The final size of the database was 2143 entries. Taxonomic composition of the database is shown in Table 1. Of course, the composition of the dataset is far from an ideally balanced; however, it represents to some extent the natural distribution of taxa. it should be borne in mind that any database is filled not according to nature, but following the preferences in the choice of species to be sequenced.

2.2 Triplet Frequency Dictionary

Triplet frequency dictionary $W^{(j)}$ is the list of all 64 triplets ω_k , $k = AAA, \dots, TTT$ accompanied with their frequency f_{ω_k} ; index j here enlists the sequences in the dataset. To make a dictionary, place the reading frame of the length 3 at the very beginning of a sequence and count all the triplets identified by the frame as it moves along a sequence from left to right (for determinacy), with the given step t . Within this paper, $t = 1$. Obvious constraint

$$\sum_{k=AAA}^{TTT} f_{\omega} = 1 \quad (1)$$

holds true.

¹ <https://www.arb-silva.de/>.

Table 1. Abundances of various taxa and genetic entries in the dataset; N stands for the number of genetic entities in the family.

Subdomain	Class*	Order	Family	N
<i>Acidobacteriota</i>	<i>Acidobacteriae</i>	<i>Acidobacteriales</i>	<i>Acidobacteriaceae</i>	31
<i>Acidobacteriota</i>	<i>Acidobacteriae</i>	<i>Acidobacteriales</i>	<i>Koribacteraceae</i>	1
<i>Acidobacteriota</i>	<i>Acidobacteriae</i>	<i>Solibacterales</i>	<i>Solibacteraceae</i>	2
<i>Actinobacteriota</i>	<i>Acidimicrobiia</i>	<i>Acidimicrobiales</i>	<i>Acidimicrobiaceae</i>	24
<i>Verrucomicrobiota</i>	<i>Chlamydiae</i>	<i>Chlamydiales</i>	<i>Chlamydiaceae</i>	49
<i>Verrucomicrobiota</i>	<i>Chlamydiae</i>	<i>Chlamydiales</i>	<i>Parachlamydiaceae</i>	39
<i>Firmicutes</i>	<i>Bacilli</i>	<i>Mycoplasmatales</i>	<i>Mycoplasmataceae</i>	13
<i>Bacteroidota</i>	<i>Bacteroidia</i>	<i>Bacteroidales</i>	<i>Bacteroidaceae</i>	163
<i>Verrucomicrobiota</i>	<i>Chlamydiae</i>	<i>Chlamydiales</i>	<i>Chlamydiaceae</i>	101
<i>Bacteroidota</i>	<i>Bacteroidia</i>	<i>Bacteroidales</i>	<i>Prevotellaceae</i>	147
<i>Bacteroidota</i>	<i>Bacteroidia</i>	<i>Bacteroidales</i>	<i>Tannerellaceae</i>	94
<i>Bacteroidota</i>	<i>Bacteroidia</i>	<i>Bacteroidales</i>	<i>Porphyromonadaceae</i>	106
<i>Bacteroidota</i>	<i>Bacteroidia</i>	<i>Bacteroidales</i>	<i>Rikenellaceae</i>	69
<i>Bacteroidota</i>	<i>Bacteroidia</i>	<i>Bacteroidales</i>	<i>Dysgonomonadaceae</i>	58
<i>Bacteroidota</i>	<i>Bacteroidia</i>	<i>Bacteroidales</i>	<i>Marinifilaceae</i>	58
<i>Firmicutes</i>	<i>Bacilli</i>	<i>Staphylococcales</i>	<i>Staphylococcaceae</i>	150
<i>Firmicutes</i>	<i>Bacilli</i>	<i>Bacillales</i>	<i>Bacillaceae</i>	151
<i>Firmicutes</i>	<i>Bacilli</i>	<i>Lactobacillales</i>	<i>Listeriaceae</i>	148
<i>Firmicutes</i>	<i>Bacilli</i>	<i>Paenibacillales</i>	<i>Paenibacillaceae</i>	124
<i>Firmicutes</i>	<i>Bacilli</i>	<i>Bacillales</i>	<i>Planococcaceae</i>	101
<i>Firmicutes</i>	<i>Bacilli</i>	<i>Brevibacillales</i>	<i>Brevibacillaceae</i>	147
<i>Firmicutes</i>	<i>Bacilli</i>	<i>Exiguobacteriales</i>	<i>Exiguobacteraceae</i>	191
<i>Firmicutes</i>	<i>Bacilli</i>	<i>Alicyclobacillales</i>	<i>Alicyclobacillaceae</i>	176

*We use the term *class* to denote any subdomain higher than *order*

The transformation of a sequence into the triplet frequency dictionary converts that latter into a point in 63-dimensional metric space; the constraint (1) allows to eliminate a triplet, since there are 63 ones linearly independent only. In theory, any triplet might be excluded from the analysis; practically, we have excluded the triplet CAC, since it has the least standard deviation figure determined over the dataset. An idea standing behind such choice is that the triplet with the minimal standard deviation contributes less of all into the distinguishability of the genetic entities.

The transformation maps symbol sequence into more convenient mathematical object that is the points in metric space, thus allowing to implement the effective methods of analysis. To do it, one must introduce a metrics; further, we use Euclidean metrics

$$\rho(W_j, W_l) = \sqrt{\sum_{k=AAA}^{TTT} (f_k^{(j)} - f_k^{(l)})^2}. \quad (2)$$

Thus, we investigate the distribution of the points corresponding to genetic sequences in this metric space revealing patterns and clusters, if any.

2.3 Clustering and Visualization

A variety of methods to cluster the multidimensional data is huge. We used k -means and elastic map technique to cluster the data. k -means is well known linear classification method [7, 9], so let's focus on elastic map technique. It is the non-linear statistics method based on the approximation of the multidimensional data by a manifold of the lower dimension; further we shall use two-dimensional manifolds [8].

The idea of this method consists in jamming the originally plain manifold (a square in our case) in the manner to minimise the total deformation energy of the elastic manifold, and mathematical springs connected to the manifold in the projection points. It is highly powerful and efficient method to cluster multidimensional data and visualise them.

Non-linear clustering of genes was provided by local density technique. In simple words, local density is a specific number of point in a small site on map. To calculate the local density, we supply each point on the map with bell-shaped function

$$h(r, r^{(j)}) = \exp \left\{ -\frac{(r - r^{(j)})^2}{\sigma^2} \right\}, \quad (3)$$

where r is a point position on the map, $r^{(j)}$ is the coordinate of a gene converted into a point through triplet frequency transformation, and σ is the contrast parameter. The function (3) looks like a normal distribution function, however it is not.

As soon as all the points on the map are supplied with the function (3), one should calculate the sum

$$H(r) = \sum_{j \in \Omega} h(r, r^{(j)}). \quad (4)$$

Here Ω is the set of all the points from the dataset. One should plot the function (4) over the map to see the density of the points distribution (see Fig. 1(b)); Ω is the set all the points representing the considered genes.

3 Results and Discussion

16SRNAs are typically used in the studies of the relations (phylogeny as well) of bacteria [10, 18, 22]. Usually, the comparison of the sequences is provided by alignment; here we present some preliminary results of the structure identification provided through the implementation of alignment-free approach, namely the unsupervised clustering based on elastic map technique.

Speaking in advance, we tried the unsupervised clustering to reveal a pattern in taxa distribution of some bacteria; we aimed mainly to prove that such pattern

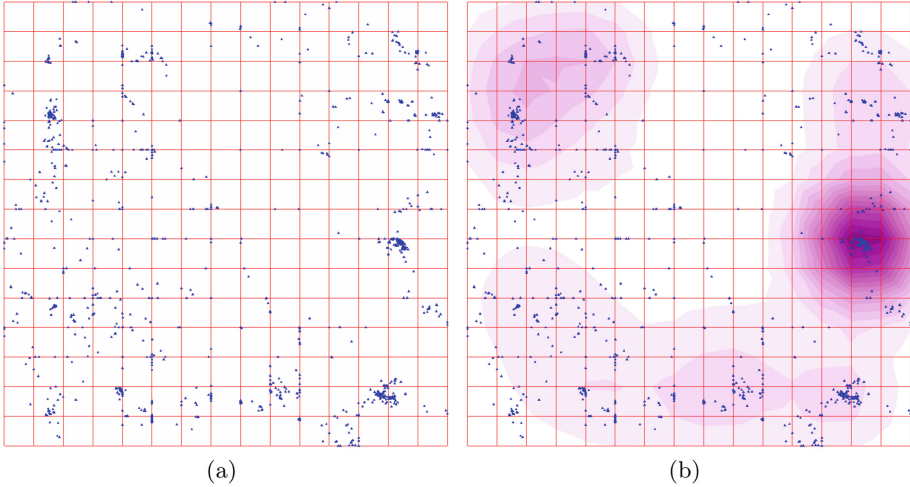


Fig. 1. A distribution of 2143 points over the elastic map with no local density indication (Fig. 1(a)) and with that former (Fig. 1(b)).

exists. Figure 1 shows the raw distribution of 2143 genes of 16S RNA over the soft elastic map. Figure 1(a) shows the distribution itself, and Fig. 1(b) shows the same distribution over the local density mapped at the same map.

Figure 1 shows the overall distribution of the genes over the soft elastic map 16×16 (Fig. 1(a)). To compare with, Fig. 1(b) shows this distribution together with the local density. Of course, the cluster pattern depends on the contrast radius σ from (3); the choice of that latter is quite informal. We used by default figure of 0.25 for this parameter. Doubtlessly, there is one highly dense cluster located at the right of the map. There are three to four clusters more, as well.

Figure 2 shows the individual distributions of specific orders over the elastic map. To do it, we made all markers of genes except those belonging to a specific order invisible; however, the elastic map as well as the local density chart is developed for the entire set of genes (these are 2143 entries). For technical reasons, we had to merge two orders (these are *Mycoplasmatales* and *Solibacterales*) into a single map (see Fig. 2(i)).

We explored the distinguishability of rather high taxa through the clustering of 16S RNA bacterial genes converted into triplet frequency dictionaries. Thus, a question arises what happens with lower taxa? In other words, if one implements the same procedure for the set of genes belonging, say to the same order, then what kind of clustering could be observed? Again, here two options may take place: the former is that lower taxa yield the distinct clustering (regardless of the peculiarities of the clusters composition, at the first step), and the latter is a decomposition of a cluster pattern resulting in more or less uniform distribution of the genes over the elastic map.

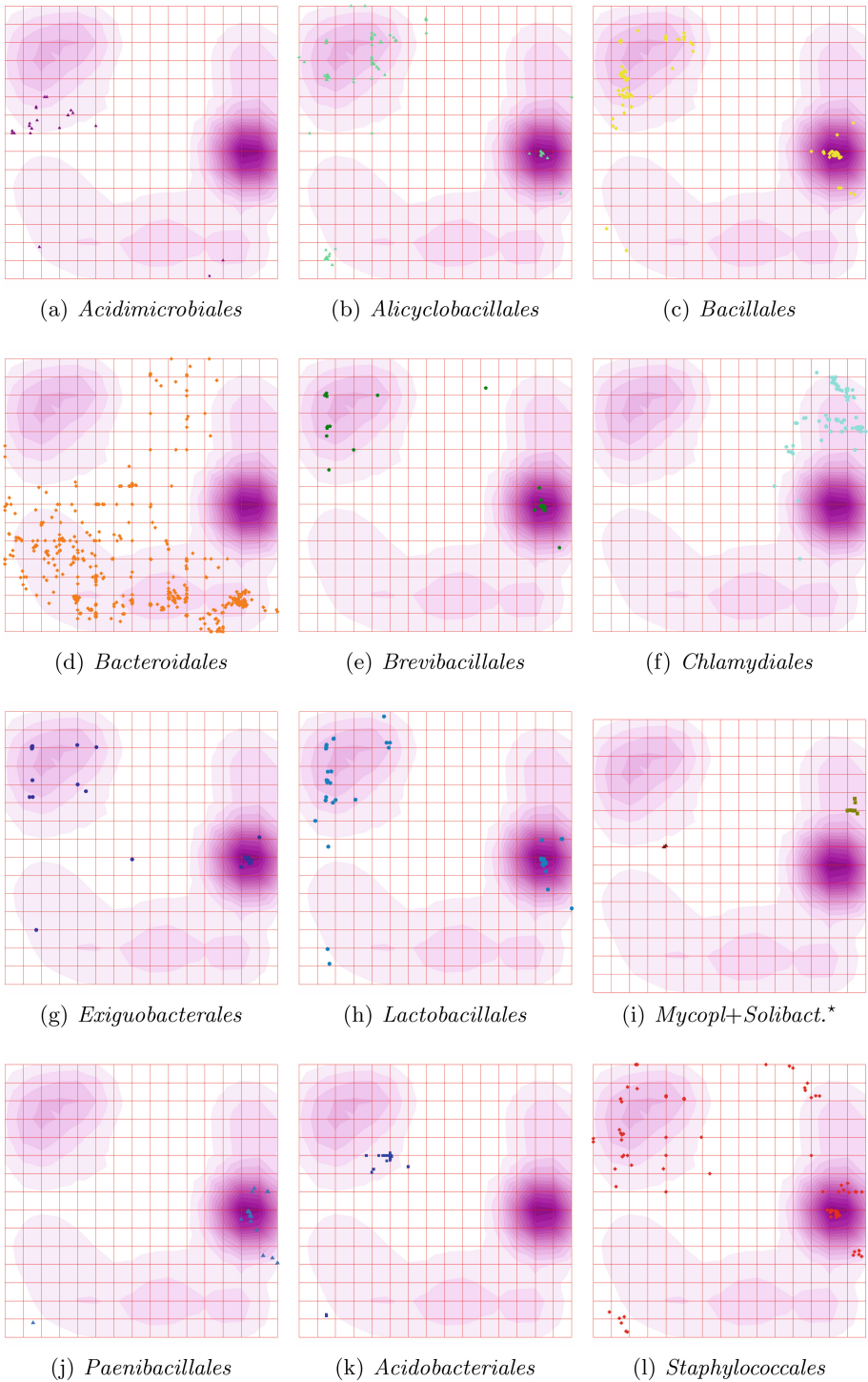


Fig. 2. Individual distributions of various orders.

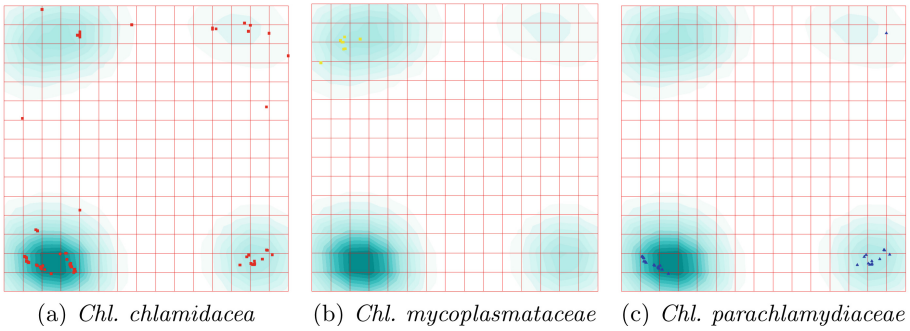


Fig. 3. Family distributions of *Chlamydiales* order.

The first option means a scalability of the cluster pattern observed through the triplet composition analysis of the genes; the second one means the absence of fine structure in the lower taxa distributions developed due to triplet composition approach. Figure 3 illustrates the lower level distinguishability of the genes, for *chlamydiales* order. There are three families comprised into the dataset, for this order. Obviously, the distribution of the families is highly specific and the species show significant speciality in the mutual location over the elastic map. The orders *Acidobacteriales* and *Acidomicrobiales* comprise a single lower suborder with 34 and 24 entries each, respectively, so we just omitted these orders from consideration here.

Space limitation makes it impossible to show the in-order distribution for all five orders shown above; however, two other orders (these are *Bacillales*, 1118 entries and *Bacteroidia*, 695 entries) have eight and seven families each, so we studied the distribution of the families for them. Surprisingly, these two orders show opposite patterns in the behaviour. *Bacillales* order shows three apparent clusters: the first one is the most dense, and two others are less dense. So, the distribution of the families over the clusters is pretty close to a uniform one: the genes belonging to various families are distributed quite homogeneously over these three clusters. It means that no dependence between lower taxonomy and triplet composition of the genes for this order is observed.

The cluster comprising *Chlamydiales* order makes a clear and apparent group located separately from other considered bacteria orders, in elastic map (see Fig. 3(f) and 3). Such isolation of pathogenic bacteria makes a promising result concerning the reliable diagnostics, in future. On the contrary, the order *Bacteroidia* exhibits very good and clear speciality in the cluster composition. It comprises seven families and they are distributed over the elastic map separately. The genes of this order yield four clusters; however, the genes are separated, for each family.

The ultimate goal is to identify and verify the early predictors of some neurological diseases, in particular the multiple sclerosis through the analysis of microbiota [11, 12, 15–17]. This ambitious goal requires an implementation of

the tool for fast and efficient analysis of some genetic markers of the microbiota, and 16S RNA seems to be the best one here. A diagnostics of the mentioned diseases requires a study of a normal pattern of the gut microbiota occurrence; hence, we prove an efficiency in the microbial population investigation and a reference value, for further medically sounding studies. The method of clustering and/or classification could complement the currently practising techniques of for diagnostics and curation strategy implementation.

Here we present some preliminary results aimed to demonstrate the feasibility and efficiency of the diagnostics based on 16S RNA analysis of the microbiota of healthy and sick people. To implement such diagnostic tool, one should make sure that a genetic marker used to distinguish sick people from healthy ones really supports this distinguishability. The results provided here unambiguously prove the efficiency of such approach, in principle. Doubtlessly, our current results do not comprise a diagnostic tool; they just approve the feasibility of the tool if it is implemented.

4 Conclusion

Here we explored the interplay between triplet composition of 16S RNA bacterial genes of five orders and taxonomy of those bacteria. Some preliminary results are present aimed to approve the feasibility of triplet composition based clustering of 16S RNA bacterial genes to identify the distinguishability of various taxa in the 63-dimensional Euclidean space of triplets frequency. The results unambiguously show that various taxa differ in terms of the triplets frequency so that more detailed and exhaustive investigation of the interplay for sure makes sense and may bring a lot. Moreover, the interplay is *scalable*: a transition from higher taxon to lower ones reveals the new and more fine structuredness in the clustering.

The study of interplay between taxonomy and k -tuple composition of genes is of great interest and value itself. However, these studies may contribute a lot in various applied areas including e.g. medicine.

Thus, a design and implementation of a tool for early diagnostics such hard to detect diseases based on comparative analysis of formally identified structures in bacterial 16S RNA is feasible.

References

1. Bonnici, V., Cracco, A., Franco, G.: A k -mer based sequence similarity for pangenomic analyses. In: International Conference on Machine Learning, Optimization, and Data Science, pp. 31–44. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-95470-3_3
2. Das, S., Das, A., Bhattacharya, D., Tibarewala, D.: A new graph-theoretic approach to determine the similarity of genome sequences based on nucleotide triplets. *Genomics* **112**(6), 4701–4714 (2020)

3. Fedotovskaya, V., Sadovsky, M., Kolesnikova, A., Shpagina, T., Putintseva, Y.: Function vs. taxonomy: further reading from fungal mitochondrial ATP synthases. In: IWBBIO, pp. 438–444 (2020)
4. Frappat, L., Sciarrino, A.: Conspiracy in bacterial genomes. *Physica A Stat. Mech. Appl.* **369**(2), 699–713 (2006)
5. Gorban, A.N., Zinovyev, A.Y.: The mystery of two straight lines in bacterial genome statistics. *Bull. Math. Biol.* **69**(7), 2429–2442 (2007)
6. Gorban, A., Popova, T., Zinovyev, A.: Codon usage trajectories and 7-cluster structure of 143 complete bacterial genomic sequences. *Physica A Stat. Mech. Appl.* **353**, 365–387 (2005)
7. Gorban, A.N., Zinovyev, A.: Principal manifolds and graphs in practice: from molecular biology to dynamical systems. *Int. J. Neural Syst.* **20**(03), 219–232 (2010). <https://doi.org/10.1142/S0129065710002383>. PMID: 20556849
8. Gorban, A.N., Zinovyev, A.Y.: Fast and user-friendly non-linear principal manifold learning by method of elastic maps. In: 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015, Campus des Cordeliers, Paris, France, 19–21 October 2015, pp. 1–9 (2015). <https://doi.org/10.1109/DSAA.2015.7344818>
9. Gorban, A., Sumner, N., Zinovyev, A.: Topological grammars for data approximation. *Appl. Math. Lett.* **20**(4), 382–386 (2007)
10. Janda, J.M., Abbott, S.L.: 16S RRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J. Clin. Microbiol.* **45**(9), 2761–2764 (2007)
11. Kozhieva, M., Naumova, N., Alikina, T., Boyko, A., Vlassov, V., Kabilov, M.R.: Primary progressive multiple sclerosis in a Russian cohort: relationship with gut bacterial diversity. *BMC Microbiol.* **19**(1), 1–9 (2019)
12. Li, B., et al.: Mild cognitive impairment has similar alterations as Alzheimer’s disease in gut microbiota. *Alzheimer’s Dement.* **15**(10), 1357–1366 (2019)
13. Luczak, B.B., James, B.T., Girgis, H.Z.: A survey and evaluations of histogram-based statistics in alignment-free sequence comparison. *Brief. Bioinform.* **20**(4), 1222–1237 (2019)
14. Mascher, M., Schubert, I., Scholz, U., Friedel, S.: Patterns of nucleotide asymmetries in plant and animal genomes. *Biosystems* **111**(3), 181–189 (2013)
15. Mirza, A., et al.: The multiple sclerosis gut microbiota: a systematic review. *Multiple Sclerosis Related Disorders* **37**, 101427 (2020)
16. Mulak, A., Bonaz, B.: Brain-gut-microbiota axis in Parkinson’s disease. *World J. Gastroenterol. WJG* **21**(37), 10609 (2015)
17. Nair, A.T., Ramachandran, V., Joghee, N.M., Antony, S., Ramalingam, G.: Gut microbiota dysfunction as reliable non-invasive early diagnostic biomarkers in the pathophysiology of Parkinson’s disease: a critical review. *J. Neurogastroenterol. Motil.* **24**(1), 30 (2018)
18. Purkhold, U., Pommerening-Röser, A., Juretschko, S., Schmid, M.C., Koops, H.P., Wagner, M.: Phylogeny of all recognized species of ammonia oxidizers based on comparative 16S RRNA and AMOA sequence analysis: implications for molecular diversity surveys. *Appl. Environ. Microbiol.* **66**(12), 5368–5382 (2000)
19. Sadovsky, M., Putintseva, Y., Chernyshova, A., Fedotova, V.: Genome structure of organelles strongly relates to taxonomy of bearers. In: Ortuño, F., Rojas, I. (eds.) *Bioinform. Biomed. Eng.*, pp. 481–490. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16483-0_47
20. Sadovsky, M.G., Putintseva, J.A., Shchepanovsky, A.S.: Genes, information and sense: complexity and knowledge retrieval. *Theory Biosci.* **127**(2), 69–78 (2008)

21. Sievers, A., Wenz, F., Hausmann, M., Hildenbrand, G.: Conservation of k -mer composition and correlation contribution between introns and intergenic regions of animalia genomes. *Genes* **9**(10), 482 (2018)
22. Yang, B., Wang, Y., Qian, P.Y.: Sensitivity and correlation of hypervariable regions in 16S RRNA genes in phylogenetic analysis. *BMC Bioinform.* **17**(1), 1–8 (2016)