MATHEMATICAL MODELING AND COMPUTING, Vol. 11, No. 4, pp. 987–994 (2024)

A statistical approach to coronavirus classification based on nucleotide distributions

Husiev M.[1], Rovenchak A.[1,2]

[1] *Professor Ivan Vakarchuk Department for Theoretical Physics, Ivan Franko National University of Lviv, 12 Drahomanov St., 79005 Lviv, Ukraine*
[2] *SoftServe, Inc., 2d Sadova St., 79021 Lviv, Ukraine*

(Received 24 May 2024; Accepted 18 October 2024)

The objective of this study is to analyze specific genomes, namely the RNA of coronaviruses, based on the parameters obtained from the distributions of nucleotide sequences in their RNA. The viral RNA was subjected to distribution based on nucleotide sequences obtained by changing one nucleotide base (adenine) into a "whitespace", with empty sequences denoted as "x". Statistical spectra were constructed in such cases. They exhibited three distinct peaks that were consistent across the studied species. Parameters based on the rank–frequency distributions of the obtained nucleotide sequences, sequence lengths, and some other statistical parameters were calculated. Based on these parameters, the principal components were built, which were the basis for the grouping of the studied viruses. The most relevant parameters formed the model of a naïve Bayes classifier, which analyzes the probability of the virus belonging to a certain group of viruses in the model.

Keywords: *rank–frequency distribution; parametrization; coronavirus; statistical spectra; principal components; naïve Bayes classifier.*

2010 MSC: 62H30, 92D20, 92-08, 91F20 **DOI:** 10.23939/mmc2024.04.987

1. Introduction

Over the past decades, complex systems have been the focus of attention for scientists and engineers in various fields of science and technology [1]. Such systems, which include biological, economic, and technical components, have a large number of interacting elements that can affect each other. This complicates their analysis, particularly when they are large in size and exhibit low predictability. Various machine-learning methods constitute nowadays one group of approaches to study such systems [2, 3]. On the other hand, the use of physical models or methods grounded in complex network theory becomes increasingly relevant in the classification of complex systems. Such approaches allow for the prediction of a system's behavior by reducing the amount of necessary data for analysis and classification, which is important in the context of the growing volume of data [4, 5].

The analysis of species similarity probabilities and their subsequent classification hold pivotal importance within the biological sciences, necessitating a multidisciplinary approach. By integrating methodologies, researchers can derive more accurate, robust classifications and understand the evolutionary relationships among species [6]. This interdisciplinary strategy enhances the resolution at which species similarities are detected and interpreted, facilitating the identification of cryptic species and refining our understanding of biodiversity [7].

This study was carried out using interdisciplinary approaches to analyze statistical data. The approaches based on linguistics, biology, statistical physics, information theory, economics, and probability theory can be seen there. It should be mentioned that the set of methods employed in this work can be used in the analysis of the distribution of a random variable.

For this study, RNA samples from the most prevalent coronavirus strains were selected to analyze their similarities, which could be attributed to the distribution of nucleotides in their RNA sequences. The data on RNA of the viruses were taken from the National Center for Biotechnology (NCBI,

https://www.ncbi.nlm.nih.gov) [8, 9]. Despite the alleged decline in interest in COVID-19 issues, this subject remains topical due to new strains being constantly discovered [10–12] calling in particular for further development of epidemiological models [13, 14].

The paper is organized as follows: an overview of the data along with detailed method descriptions is available in Section 2 "Methods". Section 3 "Results" contains the presentation of findings, and the paper concludes with a brief discussion provided in Section 4.

2. Methods

To analyze the RNA, we need to divide it into segments. For the segmentation process, we employ the substitution of adenine (**a**) with a "whitespace" symbol, while the empty sequence between two delimiters is denoted as **x** [8, 15]. Therefore, the nucleotide sequence:

$$\textbf{attaaaggtttataccttcccaggtaacaaaccaaccaactttcgatctcttgtagatctg} \tag{1}$$

upon substituting adenine transforms into:

$$\textbf{x tt x x ggttt t ccttccc ggt x c x x cc x cc x ctttcg tctcttgt g tctg}. \tag{2}$$

The first occurrence of **x** is due to the adenine **a** at the beginning of the chain.

Upon doing so, we convert a single RNA sequence into a set of sequences similar to words in texts. This hints, in particular, about possibilities to utilize approaches developed in the domain of quantitative linguistics. Below, we describe the parameters that can be derived based on a rank–frequency distribution of such sequences, as well as those based on their lengths.

To obtain the rank–frequency distribution, items (nucleotide sequences) corresponding to the input RNA are sorted based on their absolute frequency, with the most frequent item given rank 1, the second most frequent given rank 2, and so on.

This distribution can be characterized by Shannon's information entropy [16]:

$$S = -\sum_{r}^{r_{\max}} p_r \ln p_r, \tag{3}$$

where $p_r = \frac{f_r}{N}$, $N = \sum_r f_r$, and f_r is the number of sequences (absolute frequency) for the r-th rank.

Statistical entropy is one of the parameters for classifying the distribution of nucleotide sequences in viral RNA. Along with it, the mean value of the sequence can be highlighted as the first central moment [17]:

$$m_1 = \frac{1}{N}\sum_i L_i, \tag{4}$$

where N is the number of nucleotide sequences, and L_i is the number of nucleotides for the i-th sequence. Clearly, **x**, being an empty sequence, has the length of $L = 0$. The next parameter for this distribution is the variance of the statistical distribution which measures the dispersion of the statistical variable relative to the mean value of the distribution. It can also be referred to as the second central moment:

$$m_2 = \frac{1}{N}\sum_i (L_i - m_1)^2 . \tag{5}$$

The coefficient of variation is defined as:

$$d = \frac{m_2}{m_1 - 1}. \tag{6}$$

The above parameters were used in our previous studies on coronaviruses [8,9]. In the present paper, we extend them with a few more parameters that can be derived from rank–frequency distributions, cf. [18–20]:

— type-token ration TTR $= r_{\max}/N$, where the number of types (different sequences) coincides with the maximum rank $r_{\max}$ while the number of tokens in the total number of sequences N;
— fraction of *hapax legomena* (sequences occurring only once) $p_1 = f_1/N$;

— fraction of *dis legomena* (sequences occurring exactly twice) $p_2 = f_2/N$;
— relation of the numbers of *hapax* and *dis legomena* $f_1/f_2 = p_1/p_2$;
— repeat rate

$$R = \sum_{r=1}^{r_{\max}} p_r^2. \tag{7}$$

3. Results

3.1. Statistical spectra

The statistical spectrum refers to the distribution of a discrete variable that exhibits a spectrum-like pattern. It is considered as such because the variable takes on distinct values or levels forming a discrete set resembling a spectrum. The term "statistical spectrum" is employed to describe the probabilistic distribution of these discrete values within a given system [21].

Let us reflect upon ΔL_i as the deviation from the first central moment for the sequence length L_i described by the equation $\Delta L_i = |L_i - m_1|$. The quantity Q_i will indicate the number of elements with a deviation of ΔL_i. By representing their dependency, one can obtain the statistical spectrum proper.

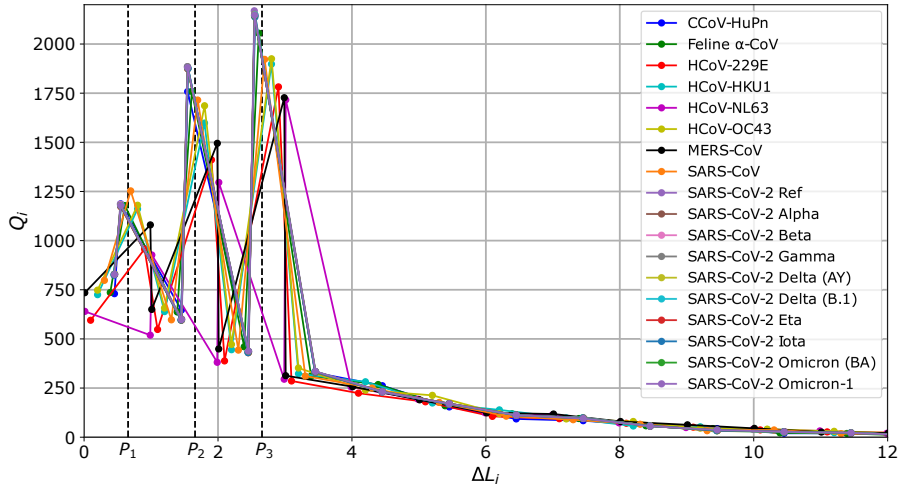

Fig. 1. Statistical spectra representing the relation between deviation from the first central moment and the number of elements having the same deviation length.

By plotting the relationship between Q_i and ΔL_i for different values of L_i, one can construct the statistical spectrum. The spectrum provides a visual representation of the distribution of elements with specific length deviations within the sequence. It illustrates the frequencies or probabilities of occurrence associated with different deviation values. The peaks being typical for every spectrum from the sample can be seen at P_1 ($\Delta L = 0.65694$), P_2 ($\Delta L = 1.65694$) and P_3 ($\Delta L = 2.65694$). The equality of decimal places can be explained by subtracting integer values for sequence lengths from a common value of m_1.

The standard deviation is a statistical measure that quantifies the dispersion or spread of a dataset. It explains how individual data points deviate from the mean or average value. A higher standard deviation indicates greater variability, while a lower standard deviation indicates less variability and a more tightly clustered dataset. For this study, this is one of the distribution parameters [22]. It can be calculated using the equation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i}^{i_{\max}} (Q_i - \langle Q \rangle)^2}, \tag{8}$$

where N is the sample size.

The values of the spectral dispersion σ^2 for these spectra together with other parameters defined above are presented in Table 1.

Table 1. Statistical parameters of the studied genomes.

ID	Name	S	m_1	m_2	d	σ^2	p_1/p_2	NCBI identifier*
1	CCoV-HuPn	4.182	2.546	5.390	3.486	0.388	7.281	MW591993.2
2	Feline a-CoV	4.250	2.609	5.719	3.554	0.487	7.818	315192962
3	HCoV-229E	4.380	2.903	7.176	3.772	0.730	9.594	12175745
4	HCoV-HKU1	4.286	2.798	6.700	3.726	0.360	7.520	85667876
5	HCoV-NL63	4.432	3.012	7.792	3.873	0.475	7.569	49169782
6	HCoV-OC43	4.388	2.799	6.721	3.736	0.369	7.374	1578871709
7	MERS-CoV	4.585	2.990	7.781	3.911	0.487	9.605	667489388
8	SARS-CoV	4.322	2.696	6.199	3.655	0.495	8.345	30271926
	SARS-CoV-2 variants:							
9	Reference	4.171	2.542	5.386	3.493	0.495	9.019	NC_045512
10	Alpha	4.177	2.546	5.413	3.500	0.497	8.691	OL546784.1
11	Beta	4.179	2.548	5.422	3.502	0.496	8.826	MZ314998.1
12	Gamma	4.176	2.545	5.406	3.499	0.496	8.807	2056248244
13	Delta (AY)	4.176	2.547	5.415	3.500	0.497	8.917	OM269121.1
14	Delta (B.1)	4.181	2.554	5.445	3.505	0.593	9.000	OK091006.1
15	Eta	4.175	2.548	5.417	3.500	0.498	8.631	MZ362439.1
16	Iota	4.176	2.552	5.438	3.505	0.659	8.789	MZ702250.1
17	Omicron (BA)	4.177	2.545	5.405	3.500	0.497	8.907	OM283600.1
18	Omicron-1	4.178	2.546	5.409	3.500	0.497	9.114	OM095411.1

*Append this identifier to `https://www.ncbi.nlm.nih.gov/nuccore/` in order to access the data. To directly access the FASTA sequence, use links in the format similar to `https://www.ncbi.nlm.nih.gov/nuccore/NC_045512?report=fasta`.

3.2. Principal component analysis

Previous studies on viral RNAs revealed a set of parameters to distinguish the viruses, including entropy S and the second central moment m_2. Here, we apply a more rigorous approach by performing the Principal Component Analysis (PCA) [23, 24] implemented within the `scikit-learn` library in Python [25, 26].

Initially, a smaller set of variables $\{S, m_1, m_2, d, \sigma^2\}$ was used yielding the following principal components (PC):

$$\mathrm{PC}_1 = 0.128\,S + 0.186\,m_1 + 0.961\,m_2 + 0.160\,d - 0.00345\,\sigma^2, \tag{9a}$$

$$\mathrm{PC}_2 = -0.0451\,S + 0.0241\,m_1 + 0.0141\,m_2 - 0.0554\,d + 0.997\,\sigma^2. \tag{9b}$$

These results suggest that m_2 and σ^2 constitute the best pair of variables describing our data, see Figure 2*a*. Note, however, that two points [they are Delta (B.1) and Iota strains] are not properly grouped with the remaining six SARS-CoV-2 variants.

By using a complemented set of $\{S, m_1, m_2, d, \sigma^2, \mathrm{TTR}, p_1, p_2, p_1/p_2, R\}$ we obtain the following principal components:

$$\begin{aligned}\mathrm{PC}_1 = {} & 0.123\,S + 0.181\,m_1 + 0.934\,m_2 + 0.155\,d - 0.0164\,\sigma^2 \\ & + 0.0145\,\mathrm{TTR} + 0.0120\,p_1 + 0.00187\,p_2 - 0.225\,\frac{p_1}{p_2} + 0.00566\,R,\end{aligned} \tag{10a}$$

$$\begin{aligned}\mathrm{PC}_2 = {} & -0.0413\,S + 0.0420\,m_1 + 0.225\,m_2 - 0.0362\,d + 0.0850\,\sigma^2 \\ & + 0.00806\,\mathrm{TTR} + 0.00830\,p_1 - 0.00082\,p_2 + 0.968\,\frac{p_1}{p_2} - 0.00086\,R,\end{aligned} \tag{10b}$$

suggesting another pair of variables, m_2 and p_1/p_2, see Figure 2*b*. In this case, the cluster of SARS-CoV-2 strains is clearly distinguished.

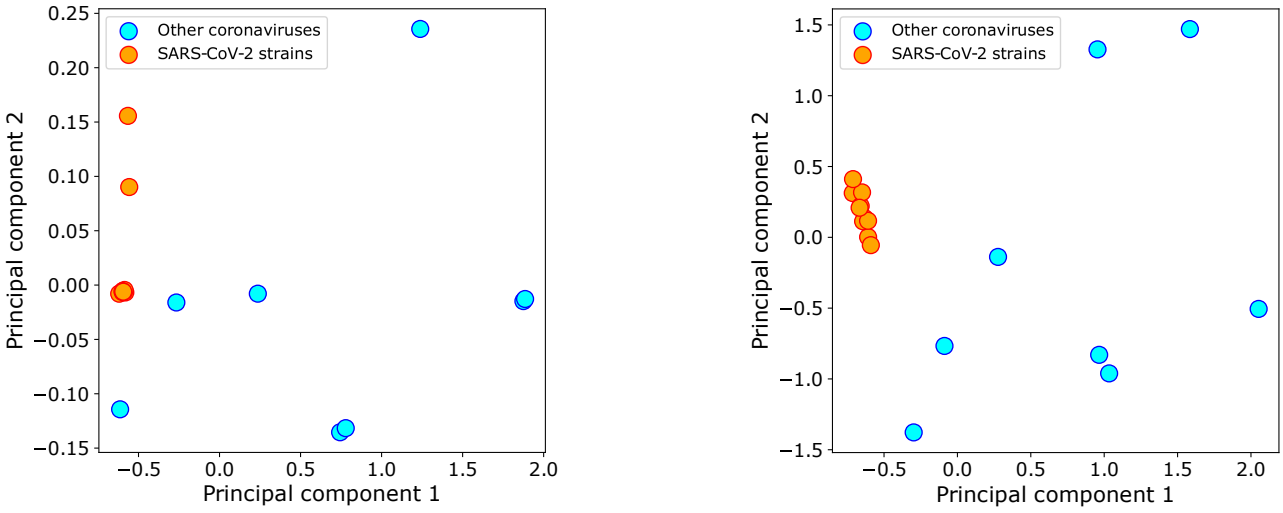

Fig. 2. Grouping of coronaviruses based on PCA using variables $\{S, m_1, m_2, d, \sigma^2\}$ (***a***) and variables $\{S, m_1, m_2, d, \sigma^2, \mathrm{TTR}, p_1, p_2, p_1/p_2, R\}$ (***b***).

3.3. Bayesian analysis

To address the problem of genome classification based on their parameters, the optimal solution is to apply a naïve Bayes classifier, which uses Bayes' theorem with the assumption of independence among variables. This approach is also suitable for classification tasks involving small data samples. The analysis was conducted based on the variables that contribute the most to the principal components, specifically m_2 and p_1/p_2.

For our variables being continuous in nature, the most appropriate choice is the so-called Gaussian Naïve Bayes. In Gaussian Naïve Bayes, the probability density function (PDF) of the Gaussian distribution for a feature x_i given a class y is:

$$P_{\mathrm{GNB}}(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_{yi}^2}} \exp\left(-\frac{(x-\mu_{yi})^2}{2\sigma_{yi}^2}\right), \tag{11}$$

where μ_{yi} is the mean of feature x_i for class y and σ_{yi} stands for the variance of feature x_i for class y.

To perform the analysis of our data, we made use of the `GaussianNB` class from Python's module `sklearn.naive_bayes` [25]. The model was trained on the data from Table 1, with the first 8 genomes belonging to class "0" (not SARS-CoV-2 variants) and the remaining 10 belonging to class "1" (SARS-CoV-2 variants). As features, the two variables with the major contributions to PC_1 and PC_2 given by (10) were used, namely, m_2 and p_1/p_2. The model appeared quite stable yielding correct attributions for the `var_smoothing` parameter [the variance σ_{yi}^2 in Eq. (11)] ranging from the default value of 10^{-9} to 0.3.

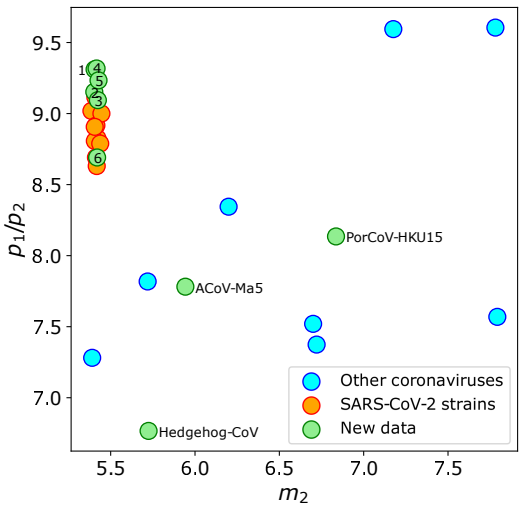

Fig. 3. Grouping of coronaviruses based on variables m_2 and p_1/p_2. The new data in the SARS-CoV-2 domain are labelled by numbers according to Table 2.

The new data include three non-human coronaviruses (avian strain Ma5, porcine HKU15, and hedgehog coronavirus 1) and six strains of SARS-CoV-2 coronaviruses (one of them also identified in domestic cats), see Table 2. As one can see, the model correctly predicts

the class in all new cases. The class probabilities in the Table correspond to $\sigma^2_{yi} = 0.3$. The new data are visualized in Figure 3 together with those used to construct the model.

Table 2. Class probabilities and predicted classes for new data at $\sigma^2_{yi} = 0.3$.

Name	Class "0" probability	Class "1" probability	Predicted class	NCBI identifier*
ACoV-Ma5	0.761	0.239	0	KY626045.1
PorCoV-HKU15	0.986	0.014	0	KJ569769
Hedgehog-CoV	0.999	0.001	0	MK679660.1
SARS-CoV-2 variants:				
1. Omicron JN.1	0.054	0.946	1	PP357646.1
2. Lambda (Felis catus)	0.051	0.949	1	MZ496616.1
3. Kappa	0.052	0.948	1	OM366054.1
4. Omicron XBB.1.16	0.056	0.944	1	OR125680.1
5. Omicron XBB.1.5_nLuc	0.054	0.946	1	OR887438.1
6. Zeta	0.066	0.934	1	OR578389.1

*Append this identifier to `https://www.ncbi.nlm.nih.gov/nuccore/` in order to access the data. To directly access the FASTA sequence, use links in the format similar to `https://www.ncbi.nlm.nih.gov/nuccore/KJ569769?report=fasta`.

4. Discussion

This paper presents an approach to determine the probability of a virus belonging to a group of other viruses based on the parameterization of nucleotide distribution in viral RNA. From the analysis of nucleotide sequence distribution in viral RNAs, ten distribution parameters were obtained. All distribution parameters were orthogonally transformed into principal components to conduct a simpler and more transparent analysis procedure, which forms the basis for virus classification analysis. To calculate probabilities, a naïve Bayes classifier was used, and the analysis was conducted based on two parameters that contribute the most to the principal components.

The results demonstrate both the clustering of similar viruses according to their principal components (see Figure 3) and the calculated probabilities based on the Bayesian classification (see Table 2). Figure 2*b* illustrates that similar viruses are closer to each other, particularly noticeable for SARS-CoV-2 strains. Specifically, among the principal components (9), corresponding to dispersion projections (statistical m_2 and spectral σ^2), a certain correlation can be inferred. Additionally, considering principal components (10), a correlation between statistical dispersion and the relation of the numbers of *hapax* and *dis legomena* can be inferred. Indeed, the Pearson correlation coefficient indicates a moderate correlation $r(m_2, p_1/p_2) \approx 0.7$.

However, it is important to note that the methods presented serve as auxiliary tools for the complex task of species similarity analysis. The authors hope that their methodological developments will be beneficial for broader bioinformatics research endeavors.

Data availability. The complete set of parameters for the analyzed viruses is available at `https://doi.org/10.5281/zenodo.11282386`.

[1] Artime O., De Domenico M. From the origin of life to pandemics: emergent phenomena in complex systems. Philosophical Transactions of the Royal Society A: Mathematical Physical and Engineering Sciences. **380** (2227), 20200410 (2022).

[2] Canfora G., Mercaldo F., Santone A. A novel classification technique based on formal methods. ACM Transactions on Knowledge Discovery from Data. **17** (8), 1–30 (2023).

[3] Raman R., Gupta N., Jeppu Y. Framework for formal verification of machine learning based complex system-of-systems. Insight. **26** (1), 91–102 (2023).

[4] Holovatch Y., Kenna R., Thurner S. Complex systems: physics beyond physics. European Journal of Physics. **38** (2), 023002 (2017).

[5] Newman M. Networks. Oxford University Press; 2nd edition (2018).

[6] Tabish M., Azim S., Hussain M. A., Rehman S. U., Sarwar T., Ishqi H. M. Bioinformatics approaches in studying microbial diversity. In: Malik A., Grohmann E., Alves M. (eds.) Management of Microbial Resources in the Environment, pp. 119–140. Springer, Dordrecht (2013).

[7] Borkin L. J., Litvinchuk S. N., Rosanov Yu. M., Skorinov D. V. On cryptic species (an example of amphibians). Entomological Review. **84** (Suppl 1), S75–S98 (2004).

[8] Husev M., Rovenchak A. On the verge of life: Distribution of nucleotide sequences in viral RNAs. Biosemiotics. **14** (2), 253–269 (2021).

[9] Husev M., Rovenchak A. Parametrization of rank-frequency distributions of nucleotide sequences in virus RNAs. Visnyk Lviv Univ. Ser. Phys. **58**, 72–84 (2021).

[10] Looi M.-K. Covid-19: Scientists sound alarm over new BA.2.86 "Pirola" variant. BMJ. **2023**, p1964 (2023).

[11] Meo S. A., Meo A. S., Klonoff D. C. Omicron new variant BA.2.86 (Pirola): Epidemiological, biological, and clinical characteristics – a global data-based analysis. European Review for Medical and Pharmacological Sciences. **27** (19), 9470–9476 (2023).

[12] Hemo M. K., Islam M. A. JN.1 as a new variant of COVID-19 – editorial. Annals of Medicine & Surgery. **86** (4), 1833–1835 (2024).

[13] Abou-Nouh H., El Khomsi M. Viable control of COVID-19 spread with vaccination. Mathematical Modeling and Computing. **11** (1), 203–210 (2024).

[14] Chen Yuzhou, Gel Y. R., Marathe M. V., Poor H. V. A simplicial epidemic model for COVID-19 spread analysis. Proceedings of the National Academy of Sciences. **121** (1), e2313171120 (2024).

[15] Rovenchak A. Telling apart *Felidae* and *Ursidae* from the distribution of nucleotides in mitochondrial DNA. Modern Physics Letters B. **32** (05), 1850057 (2018).

[16] Shannon C. E. A mathematical theory of communication. The Bell System Technical Journal. **27** (3), 379–423 (1948).

[17] Kelih E., Antić G., Grzybek P., Stadlober E. Classification of author and/or genre? The impact of word length. In: Weihs C., Gaul W. (eds.), Classification – the Ubiquitous Challenge, pp. 498–505. Springer-Verlag, Berlin–Heidelberg (2005).

[18] Zörnig P., Kelih E., Fuks L. Classification of Serbian texts based on lexical characteristics and multivariate statistical analysis. Glottotheory. **7** (1), 41–66 (2016).

[19] Rovenchak A., Rovenchak O. Quantifying comprehensibility of Christmas and Easter addresses from the Ukrainian Greek Catholic Church hierarchs. Glottometrics. **41**, 57–66 (2018).

[20] Rovenchak A. Approaches to the classification of complex systems: Words, texts, and more. In: Holovatch Yu. (ed.), Order, Disorder and Criticality, vol. 7, pp. 209–246. World Scientific (2023).

[21] Chua K. C., Chandran V., Acharya U. R., Lim C. M. Application of higher order statistics/spectra in biomedical signals—A review. Medical Engineering & Physics. **32** (7), 679–689 (2010).

[22] Bland M., Altman D. Statistics notes: Measurement error. BMJ. **312** (7047), 1654 (1996).

[23] Tipping M. E., Bishop C. M. Probabilistic principal component analysis. Journal of the Royal Statistical Society Series B: Statistical Methodology. **61** (3), 611–622 (1999).

[24] Jolliffe I. T., Cadima J. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. **374** (2065), 20150202 (2016).

[25] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. **12**, 2825–2830 (2011).

[26] Principal component analysis (PCA). `https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html`.

Статистичний підхід до класифікації коронавірусів на основі розподілу нуклеотидів

Гусєв М.[1], Ровенчак А.[1,2]

[1] *Кафедра теоретичної фізики імені професора Івана Вакарчука, Львівський національний університет імені Івана Франка, вул. Драгоманова, 12, 79005, Львів, Україна*
[2] *SoftServe, Inc., вул. Садова, 2д, 79021, Львів, Україна*

Метою цього дослідження є аналіз конкретних геномів, а саме РНК коронавірусів, на основі параметрів, отриманих із розподілу нуклеотидних послідовностей у їхніх РНК. Вірусна РНК була розділена на нуклеотидні послідовності, отримані шляхом зміни однієї нуклеотидної основи (аденін) на «пробіл», причому порожні послідовності позначено як «x». Для послідовностей побудовано статистичні спектри. Вони показали три чіткі піки, які були послідовними для досліджуваних видів. Розраховано параметри на основі ранґово-частотного розподілу отриманих нуклеотидних послідовностей, довжини послідовностей та деякі інші статистичні параметри. На підставі цих параметрів було визначено головні компоненти, які лягли в основу групування досліджуваних вірусів. Найбільш релевантні параметри сформували модель наївного класифікатора Баєса, що аналізує ймовірність належності вірусу до певної групи вірусів у моделі.

Ключові слова: *ранґово-частотний розподіл; параметризація; коронавірус; статистичні спектри; головні компоненти; наївний класифікатор Баєса.*