УДК 530.17, 519.76, 519.876.5, 578.8
PACS 89.75.-k, 87.14.gn, 87.19.xd

<https://doi.org/10.30970/vph.58.2021.72>

Параметризація рангово-частотних розподілів послідовностей нуклеотидів у вірусних РНК

М. Гусєв, А. Ровенчак

*Фізичний факультет,
Львівський національний університет імені Івана Франка
вул. Кирила і Мефодія, 8, 79005 Львів, Україна
e-mail: {mhusev, andrij.rovenchak}@gmail.com*

Проаналізовано параметризацію рангово-частотних розподілів та частотних спектрів для нуклеотидних послідовностей у вірусних РНК, які отримано шляхом заміни найчастотнішого нуклеотиду в геномі на розділювач («пробіл»). На підставі рангово-частотних розподілів розраховано такі параметри: ентропія, середня довжина послідовності, дисперсія довжини послідовності, коефіцієнт дисперсії. Частотні спектри апроксимовано модифікованим розподілом Бозе, в якому звичайну експоненту замінено на каппа-експоненту Каніадакіса, що дозволяє точніше відтворювати спостережувані «хвости» розподілів. Ми очікуємо, що запропоновані параметри можуть бути корисними для майбутніх досліджень, наприклад, через встановлення кореляції їхніх значень з іншими характеристиками вірусів.

Ключові слова: РНК-вірус, нуклеотидна послідовність, рангово-частотний розподіл, частотний спектр, каппа-експонента Каніадакіса, модифікований розподіл Бозе.

1. Вступ

Застосування міждисциплінарних підходів є притаманною рисою сучасних досліджень у різних галузях науки. У цій праці ми зосередимося на фізичних моделях для опису біологічних об'єктів. Таке моделювання може стосуватися, наприклад, реальних процесів у клітинах чи взаємодії макромолекул [1–4] або ж і ґрунтуватися на абстрактніших методах теорії складних мереж [5–7].

Вивчення нуклеотидних послідовностей часто відбувається на пограниччі фізики, лінгвістики та біології, що пов'язано з певними аналогіями між азотистими основами, кодонами, генами тощо та літерами, словами, реченнями та іншими текстовими одиницями [10–14]. З іншого боку, саме методи статистичної фізики використовують для пояснення закономірностей у галузі квантитативної лінгвістики [15–18] (див. також <https://www.maths.usyd.edu.au/u/ega/physicist-language/>). У цій

праці ми продовжуємо дослідження, розпочаті в [8, 9]. Означивши спосіб розбиття ланцюжка РНК на послідовності нуклеотидів, ми пропонуємо набір параметрів, який можна застосувати для допоміжної класифікації РНК-вірусів.

Стаття має таку структуру. У другому розділі описано процедуру розбиття макромолекули РНК на послідовності нуклеотидів та наведено результати деяких розрахунків, що ґрунтуються на рангово-частотних розподілах цих послідовностей. Розділ 3 присвячено апроксимації частотних спектрів, які отримують із рангово-частотних розподілів, за допомогою модифікованого розподілу Бозе. У розділі 4 проаналізовано залежність одного з параметрів від кількості послідовностей у геномі. Коротке обговорення одержаних результатів наведено в розділі 5.

2. Параметризація рангово-частотних розподілів

Між ДНК і РНК є три основні відмінності: ДНК містить цукор дезоксирибозу, РНК — рибозу, у якій є додаткова, порівняно з дезоксирибозою, гідроксильна група. Ця група збільшує ймовірність гідролізу молекули, тобто зменшує стабільність молекули РНК. Молекули РНК, у середньому, набагато коротші й переважно одноланцюжкові [22, 23].

Для цього дослідження було взято геноми РНК-вірусів із Національного Центру Біотехнології (NCBI, <https://www.ncbi.nlm.nih.gov>), для обчислення ми використали 52 вірусні РНК, де більшість вірусів мають пандемійний характер [9].

Щоби проаналізувати РНК запланованими в цій праці підходами, треба певним способом поділити її на частини. Азотисті основи, що формують нуклеотиди у складі ДНК та РНК, позначають за першими літерами їхніх назв: **a** — аденін, **c** — цитозин, **g** — гуанін, **t** — тимін. Тут варто зазначити, що насправді в РНК замість тиміну входить урацил, однак з міркувань уніфікованого способу подання інформації в базах NCBI цієї різниці не відображено. Для поділу на частини ми використовуємо заміну аденіну (**a**) на «пробіл», а порожню послідовність, що знаходиться між двома роздільниками, позначимо як **x**, тобто послідовність нуклеотидів

$$\text{aacctaccactcaccctagcattacttatatgatatgtctccatacccattacaatc} \quad (1)$$

перейде в

$$\text{x cct cc ctc ccct gc tt ctt t tg t tgtctcc t ccc tt c x tc.} \quad (2)$$

Отримані внаслідок такої процедури послідовності нуклеотидів і є одиницями для аналізу. Їхні рангово-частотні залежності отримуємо стандартним способом: найчастотніша має ранг 1, наступна за частотою ранг 2 тощо. Послідовностям з однаковим частотами присвоюємо послідовні ранги [8].

Особливістю цього розподілу є те, що йому притаманна обернено пропорційна залежність між рангом та частотою, і це зображено на рис. 1 для розподілу кількох вірусних РНК, що були у таблиці 1. Виходячи із цього, будемо аналізувати ці розподіли як оберненопропорційні [19].

Цей розподіл можемо охарактеризувати ентропією:

$$S = - \sum_r^{r_{\max}} p_r \ln p_r, \quad (3)$$

Рис. 1: Рангово-частотні залежності для кількох досліджуваних вірусів
 Fig. 1: Range-frequency dependences for several studied viruses

де $p_r = \frac{f_r}{N}$, $N = \sum_r f_r$, f_r — кількість послідовностей для r -го рангу.

Статистична ентропія є одним із параметрів для класифікації розподілу нуклеотидних послідовностей у вірусних РНК [9, 19]. Разом з нею серед параметрів можна виділити середнє значення довжини послідовності (перший центральний момент):

$$m_1 = \frac{1}{N} \sum_i L_i, \quad (4)$$

де N — кількість послідовностей нуклеотидів у геномі, L_i — кількість нуклеотидів для i -ї послідовності. Зрозуміло, що x має довжину $L = 0$. Наступним параметром для цього розподілу буде дисперсія статистичного розподілу, що є мірою розсіяння статистичної величини відносно середнього значення розподілу, її також називають другим центральним моментом [20, 21]:

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

Коефіцієнт дисперсії означимо так:

$$d = \frac{m_2}{m_1 - 1}. \quad (6)$$

Після проведення обчислень отримано статистичні параметри, що наведені у таблиці 2.

Табл. 1: Рангово-частотні списки для деяких вірусів. Тут r — ранг, f_r — частота
Table 1: Rank–frequency lists for some viruses. Here, r is rank and f_r stands for frequency

r	<i>Ebola</i>		<i>H1N1</i>		<i>hepatitis A</i>		<i>HIV-1</i>		<i>SARS-CoV</i>	
	посл.	f_r	посл.	f_r	посл.	f_r	посл.	f_r	посл.	f_r
1	x	1978	x	1511	x	692	x	1087	x	2515
2	c	506	g	437	g	212	g	360	c	780
3	g	421	c	340	c	130	c	264	g	538
4	t	352	t	222	t	109	t	221	t	397
5	tc	147	tg	128	tg	69	gg	100	tt	191
6	tt	137	gg	111	tt	57	gc	99	tg	191
7	tg	115	tc	91	gg	46	gt	93	ct	167
8	cc	110	gc	86	tc	43	tt	63	cc	162
9	gg	109	cc	62	ttg	31	cc	60	tc	140
10	gc	92	tgg	52	gc	28	tc	50	gg	139
11	ct	70	ct	51	tgg	22	tg	49	gt	116
12	gt	67	gt	51	cc	22	ct	32	gc	106
13	cg	59	tt	42	ttt	21	tgg	26	ctc	52
14	ttg	55	ttg	40	ttc	19	ggg	25	tgc	52
15	ttt	40	tgc	33	gt	16	gcc	22	tgg	51
16	ttc	37	ggc	28	ctc	14	gct	14	ctg	48
17	ctc	34	ttc	27	tcc	14	ccc	13	ttc	46
18	tcc	32	ggg	26	ct	14	tcc	13	ctt	46
19	gtg	31	cg	23	gct	11	tgt	13	ttg	45
20	ctg	29	ctg	22	tgc	10	ggt	12	ttt	44
...	...		...		...		...		...	

3. Частотні спектри

Означимо n_j як кількість послідовностей, що мають частоту, рівну j . Тоді, побудувавши залежність n_j , апроксимуємо її за модифікованим Бозе-розподілом:

$$n_j = \frac{1}{z^{-1}X\left(\frac{(j-1)^\gamma}{T}\right) - 1}, \quad (7)$$

де $X(t) = e^t$,

$$z = \frac{n_1}{n_1 + 1}, \quad (8)$$

тут n_1 — найбільша кількість послідовностей для певної частоти, а в нашому випадку — кількість послідовностей, що трапляються у розподілі лише один раз [8].

Хоч і розподіл (7) підходить для апроксимації, але все ж таки залишаються «хвости», які не накладаються на розподіл, що пов'язано зі швидким спаданнями експоненти на великих значеннях аргумента. Тому ліпшим варіантом для апроксимації буде, наприклад, розподіл такого вигляду:

$$n_j = \frac{1}{z^{-1} \exp_{\varkappa} \left(\frac{(j-1)^\gamma}{T} \right) - 1}, \quad (9)$$

де $\exp_{\varkappa}(x)$ — каппа-експонента Каніадакіса [24, 25]

$$\exp_{\varkappa}(x) = \left(\sqrt{1 + \varkappa^2 x^2} + \varkappa x \right)^{\frac{1}{\varkappa}}, \quad (10)$$

що застосовується для канонічного розподілу Гіббса [26], і добре підходить для апроксимації частотних спектрів [8].

Як можна побачити з рис. 2, апроксимація за розподілом (9) справді більше підходить, ніж за (7). Під час розрахунків для простоти вважаємо, що $\gamma = 1.5$ і обчислюємо параметри $\varkappa$ і T . Технічно це реалізовано за допомогою процедури `fit` у програмі `GNUPlot`.

Рис. 2: Апроксимація частотного спектра для вірусу свинячого грипу (*H1N1*) за розподілами (7) та (9). Треба зазначити такий момент: геноми вірусів грипу складаються з 7–8 сегментів [27]. Під час розрахунків «пробіли» додано на межах сегментів

Fig. 2: Approximation of the frequency spectrum for swine influenza virus (*H1N1*) by distributions (7) and (9). It should be noted that the genomes of influenza viruses consist of 7–8 segments [27]. In the calculations, “spaces” are added at the segment boundaries

Результати розрахунків параметрів $\varkappa$ і T для досліджуваних у цій роботі вірусів продемонстровано на рис. 3 і в табл. 2.

Табл. 2: Статистичні параметри для усіх досліджуваних вірусів.
 Table 2: Statistical parameters for all the studied viruses

Назва	основи	N	S	m_1	m_2	T	κ
Australian bat lyssavirus	11822	3433	3.952	2.720	7.231	81.2	4.20
Ball python nidovirus	33452	11118	3.691	2.338	4.574	167.5	4.03
Coxsackievirus A	7460	2205	3.857	2.657	6.972	45.5	7.57
Coxsackievirus B	7402	2119	3.913	2.776	8.237	43.9	6.83
Dengue	10723	3554	3.520	2.347	5.183	60.5	5.12
Ebola	18962	6056	3.770	2.455	5.598	99.2	4.44
Enterovirus71	7405	2009	4.139	2.948	8.021	54.1	4.86
Feline a-CoV	29355	8572	4.038	2.734	7.524	170.8	2.87
H1N1	13371	4456	3.545	2.340	5.292	66.7	5.50
H3N2	13332	4456	3.518	2.344	5.538	67.6	6.11
H5N1	13559	4438	3.614	2.394	5.333	73.5	7.21
Hantaan	11845	3835	3.662	2.395	5.480	71.2	7.00
HCoV-229E	27317	7421	4.141	2.997	10.324	146.8	3.84
HCoV-HKU1	29926	8332	4.065	2.893	8.350	163.5	3.76
HCoV-NL63	27553	7254	4.208	3.105	10.306	169.6	2.05
HCoV-OC43	30741	8503	4.187	2.907	8.352	183.2	2.84
Hendra henipavirus	18234	6015	3.649	2.337	5.156	88.9	5.42
Hepatitis A	7478	2189	3.713	2.730	8.322	48.7	5.49
Hepatitis C	9646	1890	4.742	4.302	31.255	47.7	4.99
Hepatitis D	1682	340	3.760	4.195	28.428	9.0	5.34
Hepatitis E	7176	1231	4.957	4.984	28.894	32.2	4.66
HIV-1	9181	3273	3.302	2.133	4.347	56.1	4.29
HIV-2	10359	3507	3.412	2.272	5.356	70.9	4.41
HRV-A1	7137	2389	3.461	2.313	5.295	39.8	8.60
HRV-B3	7208	2339	3.582	2.378	5.286	26.1	9.82
HRV-C	6944	2177	3.636	2.488	6.345	43.0	5.99
Japanese encephalitis	10965	3013	4.023	2.930	9.580	53.9	6.93
Kunjin	10795	2955	4.093	2.930	8.698	73.9	4.20
Lassa mammarenavirus	10659	3038	3.900	2.816	8.628	58.7	5.48
Marburgvirus	19114	6256	3.662	2.396	5.418	115.5	3.34
Measles	15894	4639	4.068	2.697	6.702	85.5	5.62
MERS-CoV	30119	7901	4.394	3.089	9.693	171.8	3.08
Mumps	15384	4729	3.867	2.551	5.987	77.2	4.82
Murray Valley encephalitis	11014	3182	3.928	2.761	7.924	70.7	4.41
Nipah	18246	6176	3.580	2.277	4.543	80.6	7.32
Norovirus	7600	2094	3.931	2.919	9.136	42.4	4.85
Phage-MS2	3569	836	4.138	3.528	13.413	22.4	7.59
Planidovirus	41178	16361	3.036	1.885	2.756	143.9	4.80
Poliovirus	7440	2207	3.824	2.654	7.163	29.4	12.88
Puumala	12062	3940	3.608	2.382	5.217	61.2	7.48
Rabies	11928	3419	3.976	2.772	7.784	86.5	3.40
Rift Valley	11979	3343	4.043	2.839	8.221	64.0	7.49
Rubella	9762	1459	5.308	5.831	36.851	49.9	3.44
Saint Louis encephalitis	10939	3194	3.897	2.733	6.917	71.9	3.83
SARS-CoV	29751	8482	4.111	2.804	8.201	165.3	3.00
SARS-CoV-2	29903	8955	3.937	2.661	7.367	153.7	3.42
SFTS	11490	3076	4.135	2.988	8.975	66.1	7.40
Tobacco mosaic	6395	1862	3.806	2.734	7.759	41.2	5.84
West Nile	10962	2997	4.038	2.936	9.576	63.2	6.16
Western equine encephalomyelitis	11484	3251	4.087	2.796	7.635	65.0	7.12
Yellow fever	10862	2964	3.985	2.955	9.866	61.7	4.99
Zika	10794	2992	4.011	2.873	8.163	75.2	3.83

Рис. 3: Значення κ і T для досліджуваних вірусівFig. 3: The values of κ and T for the studied viruses

4. Масштабування параметра T

Аналізуючи значення параметра T , можна зауважити, що вони більші для довших геномів. Це спостереження підтверджується зокрема значенням коефіцієнта кореляції Пірсона, розрахованим для пари змінних T – N :

$$r(T, N) = \frac{\sum_{i=0}^n (T_i - \langle T \rangle)(N_i - \langle N \rangle)}{\sqrt{\sum_{i=0}^n (T_i - \langle T \rangle)^2 \sum_{i=0}^n (N_i - \langle N \rangle)^2}}, \quad (11)$$

де $\langle \dots \rangle$ позначає середнє значення величини. Обчислення дає $r(T, N) \simeq 0.87$, тобто вказує на високу кореляцію. Цікаво, що параметр κ , на відміну від T , доволі слабо корелює з кількістю послідовностей N у геномі, до того ж кореляція негативна: $r(\kappa, N) \simeq -0.42$.

Така кореляція між T і N дає підстави говорити про певне масштабування (або «скейлінг»), порівн. [28],

$$T = aN^b. \quad (12)$$

Значення показника b виявляється близьким до одиниці, $b = 0.92 \pm 0.05$. Результати розрахунків проілюстровано на рис. 4. Урахування такого масштабування дає змогу запропонувати на додаток до T інший параметр, $t = TN^{-b}$, який не буде суттєво корелювати з довжиною генома, як показано на рис. 5.

Рис. 4: Залежність значення параметра T від кількості послідовностей N у геномах. Суцільна лінія відповідає наближенню (12) із $b = 0.921$

Dependence of the value of the parameter T on the number of sequences N in the genomes. The solid line corresponds to approximation (12) with $b = 0.921$

5. Обговорення

Ми представили кілька підходів до простої параметризації РНК вірусів на основі аналізу послідовностей нуклеотидів у вірусних РНК. Додатково до аналізованого раніше набору параметрів, що включає ентропії та середні довжини послідовностей [9], у цій праці застосовано апроксимацію так званих частотних спектрів за допомогою модифікованого розподілу Бозе [8]. Ми також розширили початкову базу даних із 30 до 52 вірусів.

Наведені результати вказують на те, що схожі віруси знаходяться ближче один до одного. Зокрема, з рис. 3 можна помітити, що віруси з родини коронавірусів знаходяться разом, окреміше від інших вірусів; для них притаманні великі значення T та малі значення λ , також їх можна класифікувати за параметрами, що наведені в табл. 2, оскільки для них притаманна невелика розбіжність за цими параметрами. Подібне спостерігаємо також для представників родини філовірусів — *Ebola-* та *Marburgvirus*.

З іншого боку, потрібно мати на увазі, що класифікація таких складних об'єктів за двома параметрами може бути лише допоміжним інструментом для визначення

Рис. 5: Значення $\varkappa$ та $t = TN^{-b}$ для досліджуваних вірусів. Позначення такі самі, як на рис. 3. Вірус поліомієліту перебуває поза межами вертикального діапазону

Fig. 5: The values of $\varkappa$ and $t = TN^{-b}$ for the studied viruses. The notations are the same as in Fig. 3. The polio virus is outside the vertical range

схожості видів, тобто з близькості значень параметрів безпосередньо не впливає подібність організмів. Справді, таке зображення є лише проєкцією на двовимірну площину точок певного багатовимірного простору, де координатами виступають значення різноманітних параметрів [9].

Накопичення даних про параметри розподілу нуклеотидних послідовностей в РНК різних вірусів сприятиме розширенню бази порівняльних досліджень, і в може стати допоміжним матеріалом для подальшого вивчення цих вірусів та їх властивостей [9, 19].

-
1. Sung W. *Statistical Physics for Biological Matter*. Springer, 2018. doi:10.1007/978-94-024-1584-1.
 2. Kuzmak A., Carmali Sh., von Lieres E., Russell A. J., Kondrat S. Can enzyme proximity accelerate cascade reactions? *Sci. Rep.* 2019. Vol. 9, Article 455. Doi:10.1038/s41598-018-37034-3.
 3. Abetz V., Kremer K., Müller M., Reiter G. Functional macromolecular systems: Kinetic pathways to obtain tailored structures. *Macromol. Chem. Phys.* 2019. Vol. 220, No. 2. Article 1800334. Doi:10.1002/macp.201800334.
 4. Honchar Yu., von Ferber C., Holovatch Yu. Variety of scaling laws for DNA thermal denaturation. *Physica A.* 2021. Vol. 573. Article 125917. Doi:10.1016/j.physa.2021.125917

5. *Analysis of Complex Networks: From Biology to Linguistics* / edited by M. Dehmer, F. Emmert-Streib. Weinheim: Wiley, 2009.
6. Qian H. Stochastic physics, complex systems and biology. *Quant. Biol.* 2013. Vol. 1, No. 1. P. 50–53. Doi:10.1007/s40484-013-0002-6.
7. Babič M., Mihelič J., Calì M. Complex network characterization using graph theory and fractal geometry: The case study of lung cancer DNA sequences. *Appl. Sci.* 2020. Vol. 10, No. 9. Article 3037. Doi:10.3390/app10093037.
8. Rovenchak A. Telling apart *Felidae* and *Ursidae* from the distribution of nucleotides in mitochondrial DNA. *Mod. Phys. Lett. B.* 2018. Vol. 32, No. 5. Article 1850057. Doi:10.1142/S0217984918500574
9. Husev M., Rovenchak A. On the verge of life: Distribution of nucleotide sequences in viral RNAs. *Biosemitotics.* 2021. Doi:10.1007/s12304-021-09403-5.
10. Brendel V., Beckmann J. S., Trifonov E. N. Linguistics of nucleotide sequences: Morphology and comparison of vocabularies. *J. Biomol. Struct. Dyn.* 1986. Vol. 4, No. 1. P. 11–21. Doi:10.1080/07391102.1986.10507643.
11. Botstein D., Cherry J. M. Molecular linguistics: Extracting information from gene and protein sequences. *Proc. Natl Acad. Sci.* 1997. Vol. 94, No. 11. P. 5506–5507. Doi:10.1073/pnas.94.11.5506.
12. Ferrer-i-Cancho R., Hernández-Fernández A., Baixeries J., Dębowski L., Mačutek J. When is Menzerath–Altmann law mathematically trivial? A new approach. *Complexity.* 2014. Vol. 13, No. 6. P. 633–644. Doi:10.1002/cplx.21429.
13. Faltýnek D., Matlach V., Lacková L. Bases are not letters: On the analogy between the genetic code and natural language by sequence analysis. *Biosemitotics.* 2019. Vol. 12, No. 2. P. 289–304. Doi:10.1007/s12304-019-09353-z.
14. Ji S. The molecular linguistics of DNA: Letters, words, sentences, texts, and their meanings. *Theoretical Information Studies: Information in the World* / edited by M. Burgin, G. Dodig-Crnkovic. Singapore: World Scientific, 2020. P. 187–231. Doi:10.1142/9789813277496_0010.
15. Corral Á., Font-Clos F. Dependence of exponents on text length versus finite-size scaling for word-frequency distributions. *Phys. Rev. E.* 2017. Vol. 96, No. 2. Article 022318. Doi:10.1103/PhysRevE.96.022318.
16. Yan X., Yang S.-G., Kim B. J., Minnhagen P. Benford's law and first letter of words. *Physica A.* 2018. Vol. 512. P. 305–315. Doi:10.1016/j.physa.2018.08.133.
17. Mazzolini A., Grilli J., De Lazzari E., Osella M., Cosentino Lagomarsino M., Gherardi M. Zipf and Heaps laws from dependency structures in component systems *Phys. Rev. E.* 2018. Vol. 98, No. 1. Article 012315. Doi:10.1103/PhysRevE.98.012315.
18. DeGiuli E. Random language model *Phys. Rev. Lett.* 2019. Vol. 122, No. 12. Article 128301. Doi:10.1103/PhysRevLett.122.128301.
19. Гусев М., Ровенчак А. Розподіл нуклеотидних послідовностей у вірусних РНК. *Міжнародна конференція студентів і молодих науковців з теоретичної та експериментальної фізики “Еврика-2021”*: тези доповідей, Львів, 18–19 травня 2021 р. С. Н2.
20. Kelih E., Antić G., Grzybek P., Stadlober E. Classification of author and/or genre? The impact of word length. *Classification – The Ubiquitous Challenge* / edited by C. Weihs, W. Gaul. Heidelberg: Springer, 2005. P. 498–505. Doi:10.1007/3-540-28084-7_58.

21. Buk S., Humenych O., Mal'tseva L., Rovenchak A. Word-length-related parameters of text genres in the Ukrainian language. A pilot study. *Text and Language: Structures - Functions - Interrelations. Quantitative perspectives* / edited by P. Grzybek, E. Kelih, J. Mačutek. Wien: Praesens, 2010. P. 13–19.
22. Alberts B., Johnson A., Lewis J., Raff M., Roberts K., Walter P. *Molecular Biology of the Cell*, 6th edn. New York: Garland Science, 2015.
23. Patton J. T. *Segmented Double-stranded RNA Viruses: Structure and Molecular Biology*. Caister Academic Press, 2008.
24. Kaniadakis G. Non-linear kinetics underlying generalized statistics. *Physica A*. 2001. Vol. 296. P. 405–425. Doi:10.1016/S0378-4371(01)00184-4.
25. Kaniadakis G., Theoretical foundations and mathematical formalism of the power-law tailed statistical distributions., *Entropy*. 2013. Vol. 15, No. 10. P. 3983–4010. Doi:10.3390/e15103983.
26. Колесниченко А. В. К построению статистической термодинамики неэкстенсивных систем на основе каппа-энтропии Каниадакиса. *Math. Montisnigri*. 2020. Vol. 48. P. 118–144. Doi:10.20948/mathmontis-2020-48-10.
27. Bouvier N. M., Palese P. The biology of influenza viruses. *Vaccine*. 2008. Vol. 26, Suppl. 4. P. D49–D53. Doi:10.1016/j.vaccine.2008.07.039.
28. Rovenchak A., Buk S. Application of a quantum ensemble model to linguistic analysis. *Physica A*. 2011. Vol. 390, No. 7. P. 1326–1331. Doi:10.1016/j.physa.2010.12.009.

References

1. W. Sung, *Statistical Physics for Biological Matter* (Springer, 2018). Doi:10.1007/978-94-024-1584-1.
2. A. Kuzmak, Sh. Carmali, E. von Lieres, A. J. Russell, S. Kondrat, *Sci. Rep.* **9**, 455 (2019). Doi:10.1038/s41598-018-37034-3.
3. V. Abetz, K. Kremer, M. Müller, G. Reiter, *Macromol. Chem. Phys.* **220**, 1800334 (2019). Doi:10.1002/macp.201800334.
4. Yu. Honchar, C. von Ferber, Yu. Holovatch, *Physica A* **573**, 125917 (2021). Doi:10.1016/j.physa.2021.125917.
5. *Analysis of Complex Networks: From Biology to Linguistics*, edited by M. Dehmer, F. Emmert-Streib (Weinheim: Wiley, 2009).
6. H. Qian, *Quant. Biol.* **1**, 50–53 (2013). Doi:10.1007/s40484-013-0002-6.
7. M. Babič, J. Mihelič, M. Calì, *Appl. Sci.* **10**, 3037 (2020). Doi:10.3390/app10093037.
8. A. Rovenchak, *Mod. Phys. Lett. B* **32**, 1850057 (2018). Doi:10.1142/S0217984918500574.
9. M. Husev, A. Rovenchak, On the verge of life: Distribution of nucleotide sequences in viral RNAs. *Biosemitotics*, in press (2021). Doi:10.1007/s12304-021-09403-5.
10. V. Brendel, J. S. Beckmann, E. N. Trifonov, *J. Biomol. Struct. Dyn.* **4**, 11–21 (1986). Doi:10.1080/07391102.1986.10507643.
11. D. Botstein, J. M. Cherry, *Proc. Natl Acad. Sci.* **94**, 5506–5507 (1997). Doi:10.1073/pnas.94.11.5506.

12. R. Ferrer-i-Cancho, A. Hernández-Fernández, J. Baixeries, L. Dębowski, J. Mačutek, Complexity **13**, 633–644 (2014). Doi:10.1002/cplx.21429.
13. D. Faltýnek D., V. Matlach, L. Lacková, Biosemiotics **12**, 289–304 (2019). Doi:10.1007/s12304-019-09353-z.
14. S. Ji, in *Theoretical Information Studies: Information in the World*, edited by M. Burgin, G. Dodig-Crnkovic (Singapore: World Scientific, 2020), p. 187–231. Doi:10.1142/9789813277496_0010.
15. Á. Corral, F. Font-Clos, Phys. Rev. E **96**, 022318 (2017). Doi:10.1103/PhysRevE.96.022318.
16. X. Yan, S.-G. Yang, B. J. Kim, P. Minnhagen, Physica A **512**, 305–315 (2018). Doi:10.1016/j.physa.2018.08.133.
17. A. Mazzolini, J. Grilli, E. De Lazzari, M. Osella, M. Cosentino Lagomarsino, M. Gherardi, Phys. Rev. E **98**, 012315 (2018). Doi:10.1103/PhysRevE.98.012315.
18. E. DeGiuli, Phys. Rev. Lett. **122**, 128301 (2019). Doi:10.1103/PhysRevLett.122.128301.
19. M. Husev, A. Rovenchak, in *International Conference of Students and Young Researchers in Theoretical and Experimental Physics “HEUREKA-2021”*: abstracts (Lviv, 18–19 May 2021), p. H2.
20. E. Kelih, G. Antić, P. Grzybek, E. Stadlober, in *Classification – The Ubiquitous Challenge*, edited by C. Weihs, W. Gaul (Heidelberg: Springer, 2005), p. 498–505. Doi:10.1007/3-540-28084-7_58.
21. S. Buk, O. Humenych, L. Mal'tseva, A. Rovenchak, in *Text and Language: Structures - Functions - Interrelations. Quantitative perspectives*, edited by P. Grzybek, E. Kelih, J. Mačutek (Wien: Praesens, 2010), p. 13–19.
22. B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, *Molecular Biology of the Cell*, 6th edn. (New York: Garland Science, 2015).
23. J. T. Patton, *Segmented Double-stranded RNA Viruses: Structure and Molecular Biology* (Caister Academic Press, 2008).
24. G. Kaniadakis, Physica A **296**, 405–425 (2001). Doi:10.1016/S0378-4371(01)00184-4.
25. G. Kaniadakis, Entropy **15**, 3983 (2013). Doi:10.3390/e15103983.
26. A. V. Kolesnichenko, Math. Montisnigri **48**, 118–144 (2020). Doi:10.20948/mathmontis-2020-48-10.
27. N. M. Bouvier, P. Palese, Vaccine **26**[S4], D49–D53 (2008). Doi:10.1016/j.vaccine.2008.07.039.
28. A. Rovenchak, S. Buk, Physica A **390**, 1326–1331 (2011). Doi:10.1016/j.physa.2010.12.009.

Статтю отримано: 08.12.2021
Прийнято до друку: 20.01.2022

Parametrization of rank-frequency distributions of nucleotide sequences in virus RNAs

M. Husev, A. Rovenchak

*Faculty of Physics, Ivan Franko National University of Lviv
8, Kyrylo and Mefodiy St., 79005 Lviv, Ukraine
e-mail: {mhusev, andrij.rovenchak}@gmail.com*

The paper analyzes the parameterization of rank-frequency distributions and frequency spectra of nucleotide sequences in viral RNA defined in a special way. To achieve homogeneity of data, we analyze genomes of 52 single-stranded RNA viruses only. Such sequences are obtained by replacing the most frequent nucleotide in the genome with a separator (“space”). This can be illustrated as follows. The sequence **aacctaccactcaccctagcattacttatatgatatgtctccataccattacaatc**, with adenine (a) being the most frequent nucleotide, becomes **x cct cc ctc ccct gc tt ctt t tg t tgtctcc t ccc tt c x tc**. Note that a zero-length sequence named **x** is inserted between two consecutive adenines. Based on the rank-frequency distributions, the following parameters are calculated: entropy, mean sequence length, sequence length variance, variance coefficient.

The frequency spectra N_j are defined as the number of different sequences within a particular genome having absolute frequency exactly equal to j . They are fitted by a modified Bose distribution, in which the ordinary exponential is replaced with the Kaniadakis kappa-exponential $\exp_{\kappa}(x) = (\sqrt{1 + \kappa^2 x^2} + \kappa x)^{\frac{1}{\kappa}}$. Such an approach allows to more accurately reproduce the observed “tails” of the distributions. The calculated values of the fitting parameters, κ and T , demonstrate moderate grouping of related species. We also observed that there exists a high positive correlation between T and the genome length N (measured as the number of defined nucleotide sequences). The fitting with the function $T = aN^b$ over all the 52 genomes yielded $b = 0.92 \pm 0.05$, so another classification parameter, $t = TN^{-b}$ was suggested.

We expect that the proposed parameters may be useful for future studies, for example, by correlating their values with other characteristics of viruses.

Key words: RNA virus, nucleotide sequence, rank-frequency distribution, frequency spectrum, Kaniadakis kappa-exponential, modified Bose-distribution.