https://doi.org/10.58245/ipsi.tir.2502.10

Surface Glycoprotein Classification Using Repeat Sequences

Alshafah, Samira A.; Beljanski, Miloš V.; and Mitić, Nenad S.

Abstract: Correct recognition of sequence characteristics is one of the most important steps in bioinformatics. The paper presents a method for the characterization of surface glycoprotein in different viruses. The method uses amino-acid and nucleotide repeat sequences to form a virus/protein profile that is used to characterize and predict the type of protein. Based on a set of characteristic repetitive sequences, a classification model for predicting the type of virus was constructed.

Index Terms: classification, coronavirus, filoviridae, repeat sequences, surface glycoprotein

1. *Introduction*

The great SARS-CoV-2 coronavirus pandemic of 2020 gave a new impulse to reveal in more detail the mechanisms that viruses use to attack their hosts. Of special interest is the SARS-CoV-2 spike glycoprotein (spike protein, S-protein), the protein responsible for attachment to and fusion with the specific host cell. S-protein is a trimmer belonging to Type I transmembrane proteins containing a single transmembrane domain, N-terminally oriented on the external side of the virion. SARS-CoV-2 S-protein shares many of the features of other viral fusion proteins [1], such as Ebola virus (EBOV), influenza hemagglutinin and HIV envelope glycoprotein. EBOV is a member of the Filoviridae family, a genome consisting of negative-sense single-stranded RNA (-ssRNA), encoding even proteins that include GP (glycoprotein - spike), as a type I transmembrane fusogenic protein [2].

Protein recognition based on the input sequence, or its part, as well as the determination of its characteristics, can be of great importance in pharmacology, and in the production of new drugs and vaccines [3]. In response to COVID-19, several vaccines as well as Monoclonal Antibodies have been developed using as a target various domains of the S-protein [4].

Manuscript received December 4, 2024.
Alshafah S. A. is with Computer Department, Faculty of Education, University of Zawia, Libya (e-mail: Samira_7_8@yahoo.com).
Beljanski M.V. is with Institute of General and Physical Chemistry, Belgrade, Serbia (e-mail: mbel@matf.bg.ac.rs).
Mitic N.S is with Faculty of Mathematics, University of Belgrade, Serbia (e-mail: nenad.mitic@matf.bg.ac.rs)

2. *Problem Definition and Existing Solutions*

The problem of determining the characteristics of the S-protein consists of recognizing for a given input sequence the type of protein and virus to which that protein belongs, as well as finding its key characteristics that can be used in further processing. The problem is increased in the situation when the input sequences are incomplete or consist of parts, as well as due to the existence of a large number of mutations of S-protein sequences in both the nucleotide and amino acid (AA) sequences.

The existing solutions applied by biologists are either experimental (and therefore quite expensive and time-consuming) or are based on a comparison (e.g. using BLAST) of the input sequence with known sequences from the database with subsequent determination of specific characteristics. Processing in this way requires a large amount of computing resources. Therefore, it would be very important to determine some specific sequences (motifs) that are not found in other viruses and/or proteins, which would significantly shorten the processing time and increase the reliability of the later results.

3. *Proposed Solution*

This work aims to further characterize S-protein according to the presence of various types of nucleotide and amino-acid repeat sequences [5, 6]. Characterization will be done by determination [7] of the "repetitive" signature of surface glycoprotein protein. The assumption is that the S-protein of each virus has its unique profile determined by a set of repetitive sequences. A set of various viruses (human Coronaviruses: SARS CoV 1, SARS CoV 2, MERS, OC43 and NL63, and as outgroup Filoviruses Ebola and Marburg) is selected for verification of this assumption. For each element of the set, a repeat signature is obtained and analyzed. These Coronaviruses and Filoviruses were chosen because they all are pathogenic. Spike glycoprotein exists with the same function in all four previously mentioned viruses. As additional verification that repetitive sequences can be used for the characterization of S-proteins

for individual virus types, a data mining classification model [8] has been constructed.

3.1 Basic Idea

The basis of the proposed method is the use of repeating sequences (repeats) [5, 9, 10]. Nucleotide and amino acid sequences often consist of repeat sequences of various lengths, types and content of constituents. There are 4 types of repeats in nucleotide sequences: direct non-complementary (DN), direct complementary (DC), indirect non-complementary (IN), indirect complementary IC, and two protein types of repeats (DN and IN). Additionally, repeats can be classified using different criteria in at least two categories: low complexity, consisting of adjacent tandem repeats, and more complex interspersed repeats. RNA and protein repeats may be involved in transposition, in formation of secondary structures, recognition by RNA-binding proteins, etc. [11].

Several sources contain complete data sources for viruses, especially for SARS-COV-2 (for example see [12]). We downloaded viral spike glycoprotein sequences from the NCBI database [13] in nucleotide and amino acid form because surface glycoprotein sequences are available in the same format not only for SARS-COV-2 but also for other viruses used in the research. For each of the sequences, direct and palindromic repeating sequences in the amino acid record of the protein were determined using the StatRepeats program [10], as well as all 4 types of nucleotide repeating sequences: DN, DC, IN, IC. Two sets of repeating sequences were used in the research: statistically significant repeats and all repeats regardless of the statistical significance of their occurrence in the observed sequences. The resulting set of repetitive sequences was then analyzed to determine the characteristic repetitive sequences for a particular virus type and used as input to a classification algorithm for virus type prediction. Although the set of observed repeats was different for both sets of sequences very similar results were obtained. In the following, only the results for the set of all repeating sequences will be presented.

3.2 Material Used

Considering the very large number of sequenced isolates of both SARS CoV 2 ([12], [13]) and other coronaviruses as well as Ebola/Marburg ([14]), only reference isolates of each analyzed virus were taken for analysis. The list of retrieved isolates is shown in Table 1. For SARS CoV 2, in addition to the reference isolate (Wuhan), isolates representing WHO strains (ALFA, BETA, GAMA, DELTA, EPSILON, LAMBDA, MU and OMIKRON) were taken to include the widest possible set of sequences of this virus. For bioinformatics analysis, we used a minimum length of 6 for nucleotide sequences and 2 for amino acid sequences as a parameter of the StatRepeats program.

Table 1. Virus types, their NCBI identifications and isolates used in research. SARS CoV 1, MERS, Hcov OC43, Hcov NL63, Ebola and Marburg isolates are reference isolates for the corresponding virus type; For SARS CoV 2 reference isolate is Wuhan while other isolates are selected based on corresponding WHO variances. S-proteins that correspond to those isolates are used as input sequences in StatRepeats program [10].

Virus Type	NCBI identification	Isolate
SARS CoV 1	NC_004718.3	SARS coronavirus Tor2
SARS CoV 1	JX163923.1	SARS coronavirus Tor2
MERS	NC_038294.1	Betacoronavirus England 1
MERS	NC_019843.3	Middle East respiratory syndrome-related coronavirus
SARS CoV 2	NC_045512.2	Wuhan
SARS CoV 2	MZ047082	ALPHA
SARS CoV 2	OM095211	BETA
SARS CoV 2	MZ611957	GAMMA
SARS CoV 2	OQ718959	DELTA
SARS CoV 2	OP143426	EPSILON
SARS CoV 2	MW713651	LAMBDA
SARS CoV 2	MZ467310	MU
SARS CoV 2	OM080680	OMIKRON
HCOV OC43	NC_006213.1	Human coronavirus OC43
HCOV NL63	NC_005831.2	Human coronavirus NL63

Virus Type	NCBI identification	Isolate
Ebola	NC_039345.1	Bombali ebolavirus
Ebola	NC_014373.1	Bundibugyo ebolavirus
Ebola	NC_014372.1	Tai Forest ebolavirus
Ebola	NC_006432.1	Sudan ebolavirus
Ebola	NC_004161.1	Reston ebolavirus
Ebola	NC_002549.1	Zaire ebolavirus
Marburg	NC_024781.1	Orthomarburgvirus – RAVN
Marburg	NC_001608.3	Orthomarburgvirus - Mt. Elgon-Musoke

4. RESULTS

Normalized numbers of repeats in Spike proteins by virus and type of repeat are shown in Table 2 (per 100 nucleotides) and Table 3 (per 100 amino acids). The total number of repetitive sequences of all types found in the material is 146,764 nucleotide ones (of which 12,414 repeats are different) and 99,389 AA repeats (of which 2,043 repeats are different). As expected, the number of AA repeats is significantly lower than the number of nucleotide repeats of the same length. It is also observed that Ebola and Marburg have a smaller number of shorter repeats in the S protein compared to coronaviruses.

Additional analysis was performed by examining whether there are repeat sequences that occur in all observed viruses, whether there are sequences that are characteristic only for certain viruses, as well as whether there are pairs of nucleotide and amino acid repeats that start at the same position in the virus (i.e. nucleotide repeats represent amino acid codons in the AA repeat).

Table 2A. Normalized number of repeats per 100NA by repeat length, repeat type (DC, DN,) and virus type

REPEATS/100NA	DC								DN									
	6	7	8	9	10	11	12	13	6	7	8	9	10	11	12	13	14	15
Ebola	12.959	3.318	0.813	0.197	0.016			0.008	19.184	5.289	1.371	0.452	0.123	0.025		0.016		
Marburg	14.809	3.641	0.831	0.318	0.049		0.024		21.359	6.158	1.784	0.489	0.147	0.049	0.024			
Mers	28.804	8.9	1.883	0.419	0.148	0.049			40.066	10.844	2.942	0.652	0.197	0.098	0.025			
SARS CoV 1	32.59	8.891	1.99	0.557	0.106				42.105	11.372	3.477	0.902	0.212	0.08	0.027		0.027	0.027
SARS CoV 2	38.428	11.397	3.285	1.038	0.189	0.032	0.026	0.023	43.089	13.46	3.638	1.055	0.324	0.102	0.003	0.017		
Hcov OC43	39.488	10.537	2.782	0.739	0.246	0.049			47.981	14.993	4.727	1.231	0.394	0.074	0.049	0.025		
Hcov NL63	37.485	9.408	2.481	0.590	0.147	0.025	0.025		56.448	18.767	5.944	1.916	0.590	0.123				

Table 2B. Normalized number of repeats per 100NA by repeat length, repeat type (IC, IN,) and virus type

REPEATS/100NA	IC										IN									
	6	7	8	9	10	11	12	14	16	18	6	7	8	9	10	11	12	13	14	15
Ebola	16.671	4.114	1.24	0.328	0.123	0.025	0.025			0.008	18.773	5.65	1.807	0.649	0.246	0.131	0.033	0.041	0.016	0.016
Marburg	17.937	4.497	1.564	0.318	0.171		0.024				19.257	6.867	1.686	0.538	0.049	0.122	0.024	0.073	0.024	0.049
Mers	34.589	8.198	2.24	0.431	0.271	0.049	0.012	0.025			35.34	9.281	2.622	0.997	0.332			0.025	0.025	
SARS CoV 1	35.828	9.873	2.68	0.425	0.239	0.053	0.186				35.191	10.005	2.866	1.008	0.345	0.08	0.08	0.053	0.053	0.053
SARS CoV 2	41.005	11.945	3.614	1.009	0.361	0.07	0.157		0.023	0.003	39.967	12.671	2.961	1.055	0.405	0.076	0.079	0.026		0.076
Hcov OC43	44.141	12.358	3.274	0.739	0.320	0.049	0.074				46.356	14.599	3.520	1.108	0.468	0.123			0.025	0.025
Hcov NL63	41.489	10.636	3.807	0.933	0.246	0.098	0.049		0.025		53.918	16.728	4.348	1.621	0.590	0.221	0.049	0.049	0.049	0.025

Table 3. Normalized number of repeats per 100AA by repeat length, repeat type (DN, IN) and virus type

REPEATS/100AA	DN				IN					
	2	3	4	5	2	3	4	5	6	7
Ebola	113.2	7.131	0.543	0.025	115.544	13.965	0.839	0.666		0.049
Marburg	129.736	10.059	0.954	0.147	131.278	15.565	1.468	0.734	0.147	
Mers	213.636	11.789	0.665		220.51	17.664	1.256	0.407		0.074
SARS CoV 1	189.801	12.351	0.717	0.08	192.749	17.689	1.514	0.478		
SARS CoV 2	195.476	11.744	1.024	0.061	196.316	17.782	1.356	0.42	0.079	
Hcov OC43	212.712	13.308	0.813	0.074	221.064	18.256	1.035	0.591		0.074
Hcov NL63	250.959	16.445	1.549		251.475	21.681	1.254	0.664	0.074	0.074

The analysis showed that 1095 (different) AA repeats of both types (489 DN, 606 IN) and 6663 (different) nucleotide repeats (1453 DC, 1775 DN, 1694 IC, 1741 IN) occur in at least 2 different types of viruses. In addition, the number of different repeats occurring in all 7 virus types is 203 (AA) and 22 (nucleotide). Such repeats are short repeats of length 2 for AA repeats and length 6 for nucleotide repeats. The result of this analysis shows that longer repeating sequences usually occur in only one type of virus (and they do not occur only once, i.e. by chance), which supports the set hypothesis about the characterization of the S-protein of each virus, and the existence of a classifying based on repetitive sequences.

There are a total of 4646 paired nucleotide and amino acid repeats located at the same positions in the protein. Of these, only one repeat (in Hcov OC43, amino-acid VBSG or nucleotide GTAGGTAGTGGT) has a length of 12 NA (4 AA), 139 repeats are 3(AA) or 9(NA) long, while the rest are 2(AA) or 6(NA) long. This set consists of 277 IN repeats, while the rest of them are DN-type. Although these repeats are unique to each type of virus, research has shown that for a more accurate classification, it is necessary to include unique repeats of longer length (that do not have their AA or NA pair). Nevertheless, it is important to discover all such repeats because they are of interest to biologists involved in the analysis of protein and RNA secondary structures and interactions.

The number of repeating sequences that occur in only one type of virus is shown in Table 4. The obtained results show that there are a significant number of repeats of all types that characterize individual types of viruses, except possibly SARS-CoV-1. Ebola and Marburg, as well as Hcov OC43 and Hcov NL63 have a significantly higher number of repeats that characterize them, which is quite striking. Examples of some of the longest repeats (left component) are:

- NUC repeats: CGACAAACTAGTTTGTCG, TGGTAATTATAATTACCA, GTGTAGTTAACTACAC, ATTATAATTATAAAT, ..., AACGAAACGACAC, AATCATTTATCGG, ...
- AA repeats: DANINAD, TTPAPTT, VLLSLLV, LSTPS, TFYAYFT, VSVPVSV, IRSSRI, ...

At the same time, in most inverse non-complementary repeats, the left and right components match, that is, the repeats are real palindromes.

Table 4. Number of amino acids and nucleotide repeat sequences in S-proteins by length that occur only in one type of virus. These repeats are the core of successful classification.

Repeat number		AA						NUC										
		2	3	4	5	6	7	6	7	8	9	10	11	12	13	14	15	18
Sars_CoV_2	DN		11	2				11	25	14	9	4	2	1				
	IN		15	3		1		17	29	18	9	5		1				
	DC							9	33	21	8	1						
	IC							7	27	13	2	2						1
Ebola	DC							240	275	94	23	2			1			
	DN	9	96	16	1			269	327	136	53	15	3		2			
	IC							274	316	126	40	13	3	3				1
	IN	2	118	20	19		2	271	307	154	60	30	16	4	5	2	2	
Marburg	DC							70	88	34	13	2		1				
	DN		32	9	2			56	109	63	18	6	2	1				
	IC							69	85	55	13	7		1				
	IN	2	25	11	6	2		76	121	51	20	2	5	1	3	1	2	
Mers	DC							9	13	5								
	DN		2					10	19	11	3							
	IC							5	9	6	3	2		1				
	IN		4		1			9	16	11	1	1						
Sars_CoV_1	DN								1									
	IN								2									
Hcov OC43	DN	5	105	11	1			82	174	140	50	16	3	2	1			
	DC							95	189	92	28	10	2					
	IN	9	89	13	6		1	92	184	106	43	18	5		1		1	
	IC							81	204	100	30	13	2	3				
Hcov NL63	DN	4	82	21				70	172	143	70	20	5					
	DC							74	152	84	20	6	1	1				
	IN	3	93	17	8	1	1	69	177	113	63	21	9	2	2	2	1	
	IC							69	160	112	38	10	4	2			1	

The S-protein profile that characterizes each of the observed viruses is obtained by extracting unique repetitive sequences found only in that type of virus. Protein profiles can serve as a basis for virus type determination using machine learning techniques. For this purpose, a classification model was created to predict the type of virus. The input material is divided into two parts, stratified by virus type and repeat length. The ratio of material distribution is 60:40 (training/test).

After preprocessing, the data for each repeat occurring in the material includes five attributes: the type of repeat (DN, DC, IN, IN for nucleotide, or DN, IN for amino acid repeats), its sequence (left and right components), and its position within the protein (first, second, or third part). All attributes in the input material are categorical. Six algorithms were used to create the classification model: C5.0, CART, CHAID, Neural network and Tree-AS [15] from the SPSS Modeler package [16] and SPRINT from the Intelligent Miner package [17]. All algorithms were applied with default parameters only. The obtained classification accuracy is shown in Figure 1. The numerical values of predictions are shown in Table 5. Using only repeats longer than 10 (NA) or 4(AA) don't produce results because of a small number of repeats that serve as input records (see Table 4).

As expected, different algorithms gave different prediction quality. It varies from stable models (C5.0, CART and some predictions from CHAID and Neural Network) to possible overfitting (CHAID and Neural network for AA repeats length 4, SPRINT and Tree-AS). Exploring the cause of the possible overfitting we have found that the reason lies in the way the material was divided into training and test. The division was done automatically (stratified to virus type and repeat length) using a programming tool that did not take into account which repeat is unique to a specific virus. The result was that for some viruses most of the unique repeats used for virus characterization fall in the training part, compared to a very small percentage in the test part. This is especially true for Hcov OC43 and Hcov NL63 viruses. We made some experiments in which we (a) manually removed and (b) redistributed NL63 virus repeats which resulted in increasing precision by about 7-9% and decreasing the gap between training and test precision. For some further research which will include more different viruses additional attributes related to unique/non-unique repeats should be added and included in the material division process. Adding new sequences can, in general, lead both to the increase of the quality of the results as well as to its decrease (and, of course, the quality of the results can remain at the same level). If the sequence to be added is a mixture of two existing sequences (e.g. an artificial protein sequence made as a mixture of proteins from two SARS viruses) the quality may decrease. Otherwise (e.g. if surface glycoprotein of non-human coronavirus is added) the quality should not decrease and may increase.

According to the presented results, more accurate results are obtained using nucleotide repeats than amino acid ones. Although algorithms were applied with default parameters and without any optimization, existing stable predictions with an accuracy of over 50% confirm that repeats can be used for constructing virus profiles. Research related to improving accuracy can move in three directions: (1) expanding the number of viruses/number of virus isolates and the number of different protein types that serve as a generator of the set of unique repeats, (2) combining amino acid and nucleotide unique repeats as the basis for a classification model, and (3) adding attribute related to unique/non-unique virus repeat and include it in the material division process.

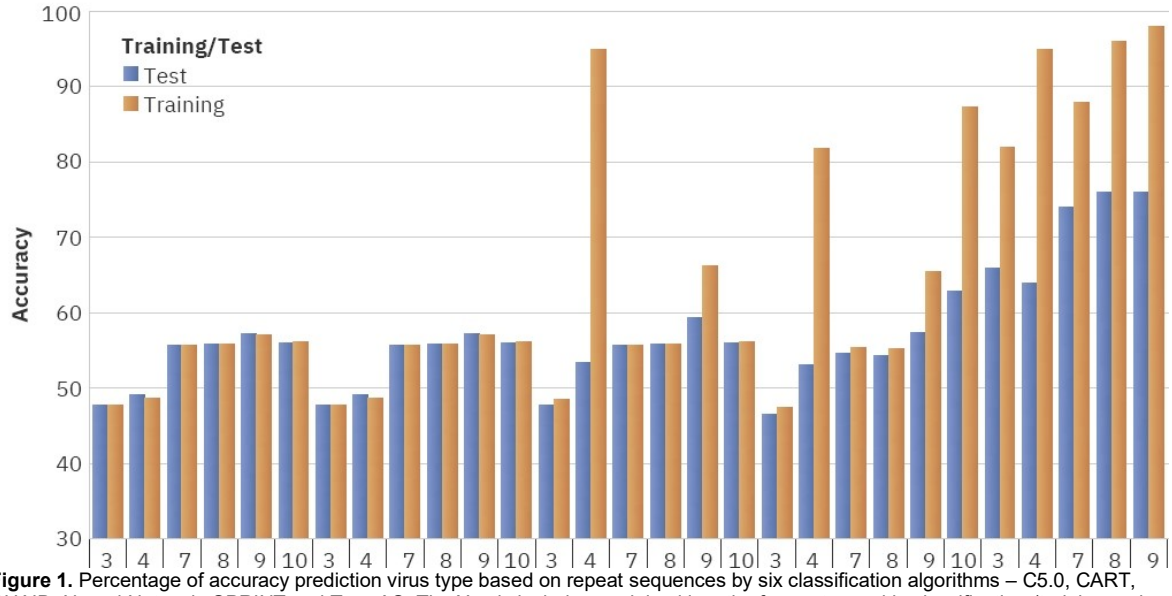

Figure 1. Percentage of accuracy prediction virus type based on repeat sequences by six classification algorithms – C5.0, CART, CHAID, Neural Network, SPRINT and Tree-AS. The X-axis includes a minimal length of repeats used in classification (training and test) and the type of repeats (amino acids or nucleotides). Tree-AS did not successfully complete the work for nucleotide repeats with minimal length 7.

Table 5. Percentage of accuracy prediction virus type based on repeat sequences by six classification algorithms – numerical values. Tree-AS did not successfully complete the work for nucleotide repeats with minimal length 7.

Algorithm		Amino acid		Nucleotide			
		Minimal repeat length					
		3	4	7	8	9	10
C5.0	Training	47.8	48.67	55.68	55.84	57.14	56.21
	Test	47.81	49.09	55.68	55.83	57.16	56.08
CART	Training	47.8	48.67	55.68	55.84	57.14	56.21
	Test	47.81	49.09	55.68	55.83	57.16	56.08
CHAID	Training	48.51	94.92	55.68	55.84	66.25	56.21
	Test	47.75	53.45	55.68	55.83	59.33	56.08
Neural_network	Training	47.39	81.84	55.37	55.24	65.52	87.31
	Test	46.62	53.09	54.67	54.32	57.37	62.89
SPRINT	Training	82	95	88	96	98	98
	Test	66	64	74	76	76	69
Tree-AS	Training	82.43	77		96.23	98.29	88.95
	Test	49.29	49.09		71.34	70.6	65.36

5. CONCLUSION

Repeat sequences may be used as a tool to characterize the surface glycoprotein for individual viruses. It turns out that many repeats uniquely characterize certain viruses. This is very important, especially in cases where the input material to be recognized is unknown, incomplete, or fragmentary. Characteristic repeat sequences can serve as a basis for the formation of efficient classification models for virus-type prediction. Finding characteristic sequences for certain proteins (in this case S-protein) is important in the production of new drugs and vaccines. The described method is the first step in repeat analysis of other proteins of pathogenic viruses. Further analysis will include, based on results obtained, more isolates and also a comparison to more outgroups (both DNA and RNA viruses).

REFERENCES

[1] Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, Graham BS, McLellan JS. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science. 2020 Mar 13;367(6483):1260-1263. doi: 10.1126/science.abb2507. Epub 2020 Feb 19. PMID: 32075877; PMCID: PMC7164637

[2] Jain S, Martynova E, Rizvanov A, Khaiboullina S, Baranwal M.: Structural and Functional Aspects of Ebola Virus Proteins. Pathogens. 2021 Oct

15;10(10):1330. doi: 10.3390/pathogens10101330. PMID: 34684279; PMCID: PMC8538763.

[3] Huang, Y., Yang, C., Xu, Xf. *et al.* Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. Acta Pharmacol Sin 41, 1141–1149 (2020). https://doi.org/10.1038/s41401-020-0485-4

[4] Kyriakidis NC, López-Cortés A, González EV, Grimaldos AB, Prado EO. SARS-CoV-2 vaccines strategies: a comprehensive review of phase 3 candidates. NPJ Vaccines. 2021 Feb 22;6(1):28. doi: 10.1038/s41541-021-00292-w. PMID: 33619260; PMCID: PMC7900244

[5] Holmudden M, Gustafsson J, Bertrand YJK, Schliep A, Norberg P. Evolution shapes and conserves genomic signatures in viruses. Commun Biol. 2024 Oct 30;7(1):1412. doi: 10.1038/s42003-024-07098-1. PMID: 39478059; PMCID: PMC11526014.

[6] Luo H, Nijveen H. Understanding and identifying amino acid repeats. Brief Bioinform. 2014 Jul;15(4):582-91. doi: 10.1093/bib/bbt003. PMID: 23418055; PMCID: PMC4103538.

[7] Jelovic A, Mitic N, Eshafah S, Beljanski M. Finding Statistically Significant Repeats in Nucleic Acids and Proteins. J Comput Biol. 2018 Apr;25(4):375-387. doi: 10.1089/cmb.2017.0046.

[8] Tan, P-N, Steinbach, M., Karpatne, A., Kumar, V. Introduction to Data Mining (2nd Edition), ISBN 0133128903, Pearson, 2018

[9] Silva JM, Pratas D, Caetano T, Matos S. The complexity landscape of viral genomes. Gigascience. 2022 Aug 11;11:giac079. doi: 10.1093/gigascience/giac079. PMID: 35950839; PMCID: PMC9366995.

[10] Jelović A.: RepeatsPlus - program for finding motifs and repeats in data sequences, J. Bioinf. Comput. Biol., 19 (3), p. 2150010, doi: 10.1142/S0219720021500104, 2021

[11] Gebauer F, Schwarzl T, Valcárcel J, Hentze MW. RNA-binding proteins in human genetic disease. Nat Rev Genet. 2021 Mar;22(3):185-198. doi: 10.1038/s41576-020-00302-y. Epub 2020 Nov 24. PMID: 33235359

[12] Coronavirus 19 database https://ngdc.cncb.ac.cn/ncov/

[13] NCBI Virus: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/find-data/virus, accessed 30.10.2024.

[14] Hume AJ, Mühlberger E. Distinct Genome Replication and Transcription Strategies within the Growing Filovirus Family. J Mol Biol. 2019 Oct 4;431(21):4290-4320. doi: 10.1016/j.jmb.2019.06.029. Epub 2019 Jun 29. PMID: 31260690; PMCID: PMC6879820.

[15] IBM SPSS Modeler 18.3 Algorithms Guide, https://ibm.com/docs/en/SS3RA7_18.3.0/pdf/AlgorithmsGuide.pdf, IBM, 2021

[16] SPSS Modeler v18.3 https://www.ibm.com/products/spss-modeler

[17] IBM Intelligent Miner v10.5 http://www- 01.ibm.com/software/data/infosphere/warehouse/mining.html

Samira Al-Shafah is an assistant professor in the Department of Computer Science at the Faculty of Education, Az-Zawiya University. She obtained her Master's degree in 2007 in Computer Science from the Academy of Graduate Studies and her PhD in Computer Science, specializing in Information Technology, from the Faculty of Computer Science and Mathematics, University of Belgrade, Serbia. Her research interests focus on bioinformatics and data mining.

Miloš Beljanski is a Molecular Biologist and Principal research fellow, retired from the Institute of General and Physical Chemistry/Biolab, Belgrade. He received his BSc, MSc and DrSci from the University of Belgrade, Faculty of Sciences, Dept. of Biology. His research interests cover areas of Biophysical Chemistry and Bioinformatics.

Nenad Mitić is a full Professor at the Department of Computer Science, Faculty of Mathematics, University of Belgrade. He received his BSc, MSc and PhD from the University of Belgrade, Faculty of Mathematics. From 1983 to 1991 he was a system analyst on an IBM mainframe in the Statistical Office of the Republic of Serbia, Belgrade. From 1991 till now he has been with the Faculty of Mathematics, Belgrade. His research interests cover areas of bioinformatics, data mining, big data and functional programming. Currently, he is the president of the Serbian Society for Bioinformatics and Computational Biology.