Applied
Microbiology
International

LAM ECS Research Symposium 2024
University of the West of England

Pairwise sequence alignment analysis of Algerian SARS-COV2 Omicron

Mohammed Khodja[1], Benazi Nabil[2] , Azzedine Melouki[1]
[1]University of Msila, [2]Pasteur Institute of Algeria

Does the Global and Local alignment give information about evolutionary relationship?

The SARS-COV2 selected, because related with pandemic word .
21 sample downloaded from GSAID

EMBOSS Needle
Pairwise Sequence Alignment
European Bioinformatics Institute
But limited by 5000 sequence long

Should use PYTHON, BIOPYTHON Google Colab, Jupyter.
Develop the necessary program code for Global and Local Alignment Algorithms

Pairwise Alignment of SARS-COV2 with 30K nucleotide base long

Local alignment has better family SARS-COV2 grouping and classification

Bat	coronavirus
Rhinolophus affinis	RaTG13
Rhinolophus sinicus	WIV1
Rhinolophus sinicus	Rs4231
Rhinolophus sinicus	GX2013
Rhinolophus pusillus	SL-CoVZXC21
Rhinolophus pusillus	SL-CoVZC45

*Rhinolophus affinis from Yunnan province

Whole genome
79.6% to SARS-COV
96% to Bat
79.6% SARS-COV BJ01
96.2% to RaTG13
50% with MERS-CoV
85.5% pangolin

May not enough work.?
May Global and Local sequence Alignment not enough clear?
After reading sturdy works published in Nature, The lancet, Virologica Sinica [7][8][9][10]
Whole genome of SARS-COV2 Matching
Spike of SARS-COV2 Matching

Host of Coronavirus: Bat

Host of Coronavirus: Pangolin

Host of Coronavirus: Civet

Highlights

- Matching score of Spike region give better classification of SARS-COV-2.
- Rhinolophus affinis-COV "d RaTG13" most closest to Human SARS-COV2.
- Human-COV, Civet and some BAT-COV have same matching score 63% to Human SARS-COV2.
- Matching score in both cases Global and Local approximately stay same that refer to the importance of Spike region.
- All type of coronavirus found in Bat, that can considered the main host of coronavirus

Figure 2. Global Alignment "Needleman Algorithm"

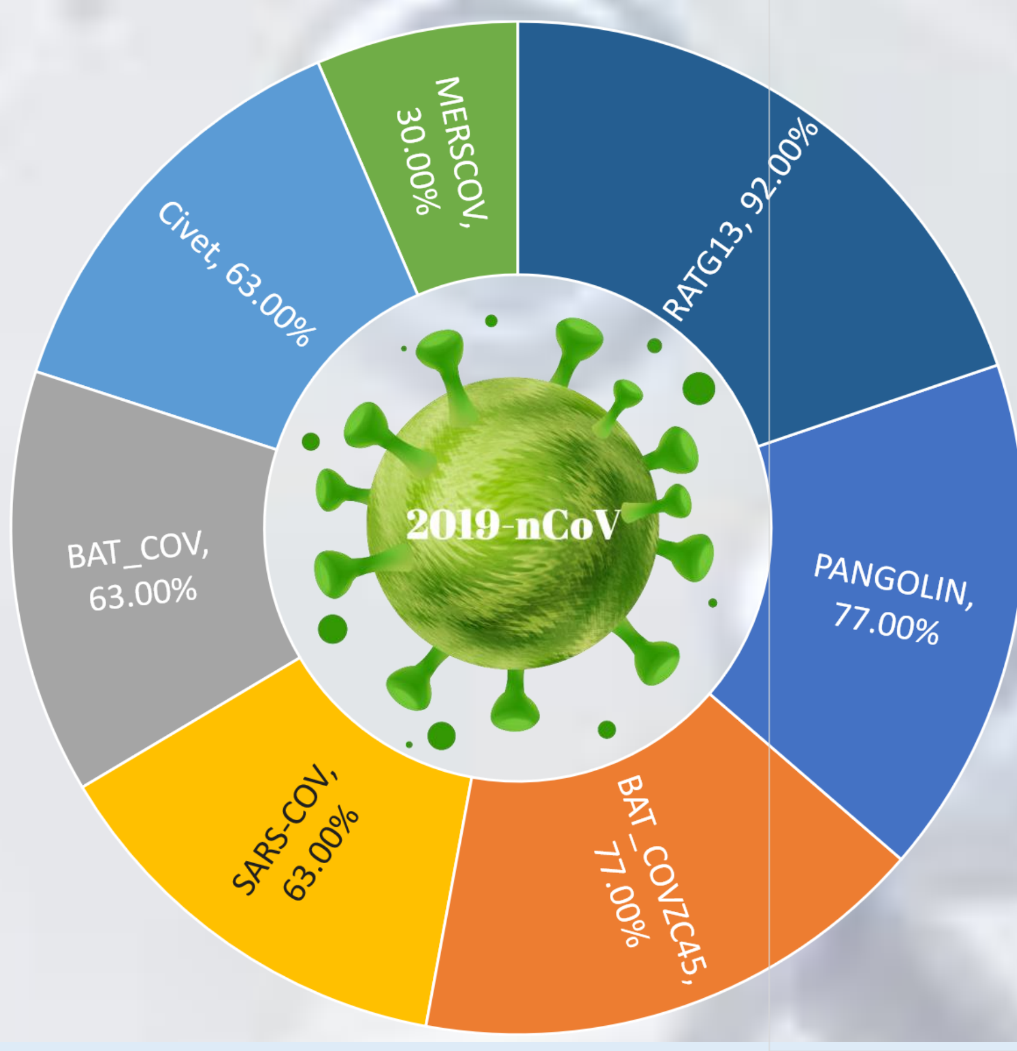
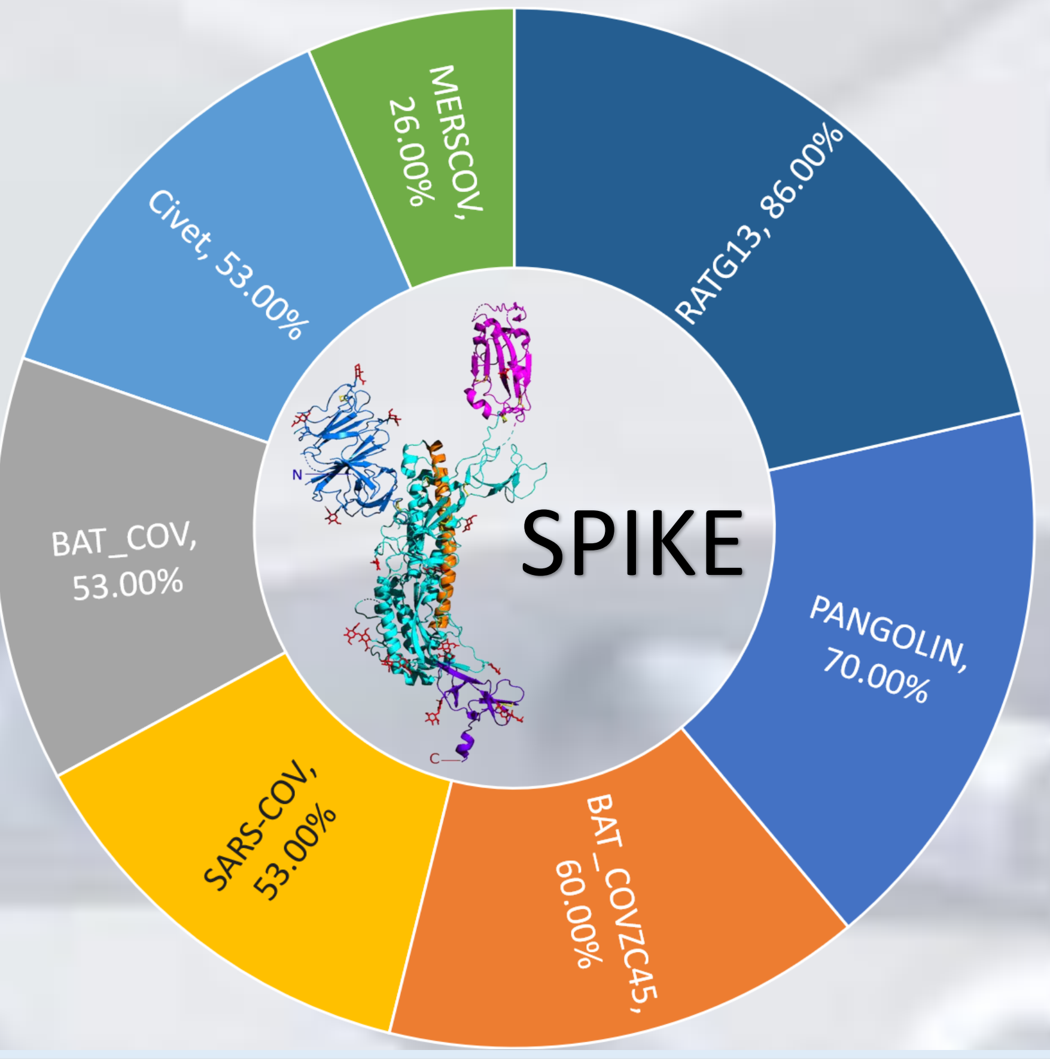


Figure 3. Local Alignment "Waterman Algorithm".



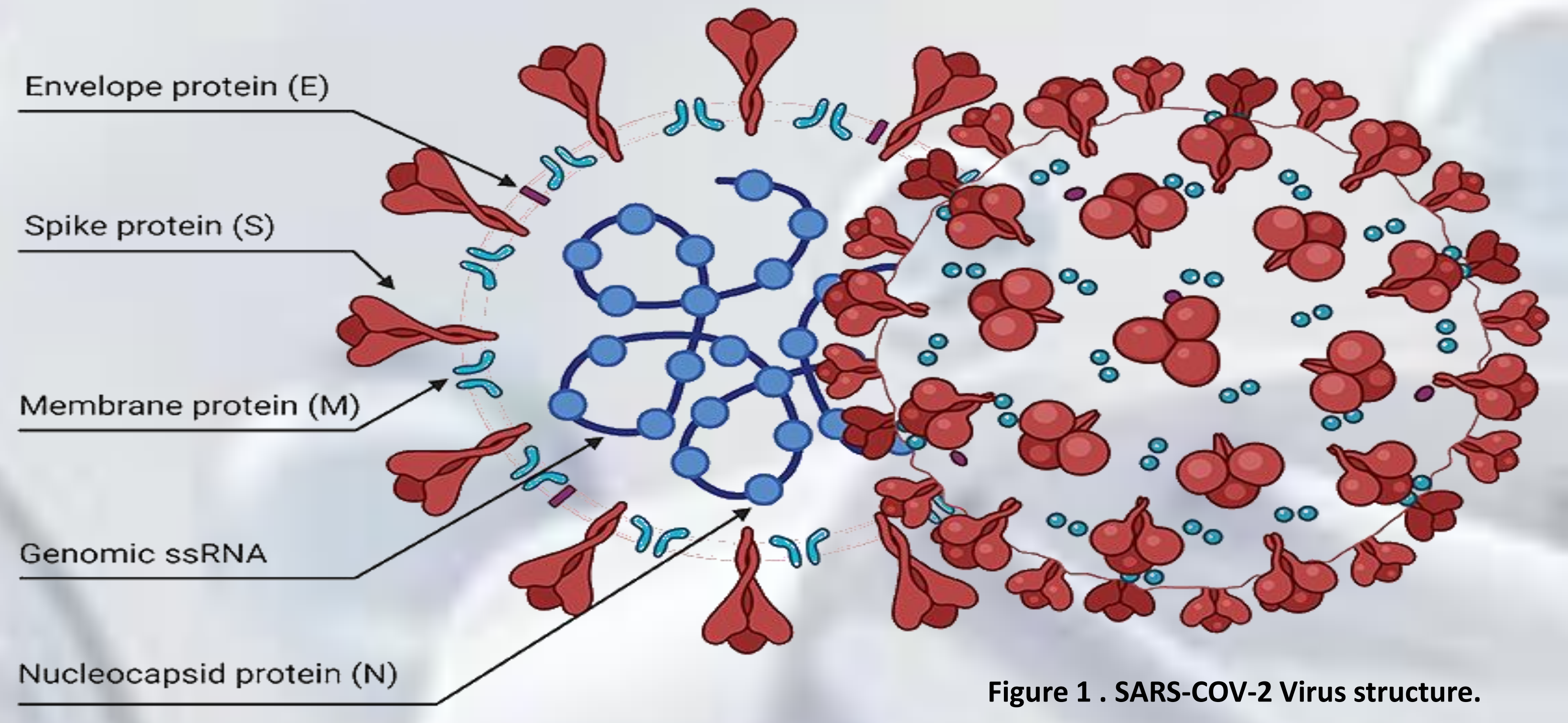


Figure 1 . SARS-COV-2 Virus structure.

Results and discussion

The Table 1 represent the obtained results using Local Pairwise Alignment which performs a subregion of the query sequence with the whole reference sequence Wuhan-Hu-1 After calculating the scoring match of all samples compared with Wuhan-Hu-1 reference genome assembly and filled the results in the table 1 and sorted from lowest to highest matching score LS_% (Local alignment matching score in percentage). As general analysis the matching score very high and that confirm the principle of Local alignment which mention in many literature that this technique is most for closely related subsequences of different sequence lengths and determining the position of Motif regarding the whole reference sequence, that it is very clear in the table 2 such all calculated score are around 98% of match, but Can notice Omicron SARS-COV2 BA.5.2 distributed by two groups by Blod font and line, between them different Omicron version that may be indicating the Local aligning using for Spike region gives better grouping of SARS-COV2 family which is less then Global alignment case that may related with special features of Spike region that related with mane mutations effect directly on its classifications or which SARS-COV2 family belong.

Table 1. Local Alignment sorted from lowest to highest matching score LA_% .

Countries	GSAID_Accession_ID	Clad	Length	Global Score vs Wuhan-Hu-1	GS_%	LA_Score	LS_%
France	18913704	23I (BA.2.86)	29685	29309	98.73%	4849	96.98%
Germany	18829988	23I (BA.2.86)	29766	29458.5	98.97%	4860.5	97.21%
England	18969306	23I (BA.2.86)	29737	29408	98.89%	4862	97.24%
England	18910386	23I (BA.2.86)	29738	29412	98.90%	4864.5	97.29%
England	15192184	23I (BA.2.86)	29740	29414	98.90%	4864.5	97.29%
Switzerland	18828202	23I (BA.2.86)	29518	29089	98.55%	4865.5	97.31%
Pakistan	15111973	**BA.5.2 22B**	28727	28009	97.50%	4878.5	97.57%
Ukraine	16637095	**BA.5.1.3 22B**	29271	28804.5	98.41%	4897	97.94%
Algeria	18830343	XBB.1.5.63 23A	29418	28995.5	98.56%	4901	98.02%
England	16440149	23C (CH.1.1)	29724	29444.5	99.06%	4905	98.10%
Algeria	17182683	22E BQ.1.1.59	29086	28516.5	98.04%	4912.5	98.25%
PuertoRico	15229630	**BA.5.1.30 22B**	29652	29385.5	99.10%	4918	98.36%
Algeria	16242296	**BA.5.2**	29099	28555.5	98.13%	4923.5	98.47%
England	13841928	**BA.5.2 22B**	29229	28743.5	98.34%	4924.5	98.49%
SouthAfrica	13830427	**BE.1 22B (BA.5)**	29715	29476.5	99.20%	4925.5	98.51%
Algeria	16242292	**BA.5.1.30**	29235	28759.5	98.37%	4925.5	98.51%
Algeria	15946159	**BA.5.2.27 22B**	29399	28998	98.64%	4927.5	98.55%
England	15192184	**BA.5.2**	29681	29416.5	99.11%	4929.5	98.59%
Canada	15978247	**BA.5.1.30 22B**	29646	29370.5	99.07%	4929.5	98.59%
Italy	14971363	**BA.5.2 22B**	29768	29543	99.24%	4929.5	98.59%
Algeria	16454585	**BA.5.2**	28804	28115	97.61%	4930	98.60%

Abstract

The presented study case interest by pairwise sequence alignment of SARS-COV2 Omicron in Algerian compared to the neighbor's countries using Global and local alignment algorithm, to check what introduced in different literature about scoring matching of alignment between two sequences of course with real application and samples from GISAID with Biopython command line in Colab platform. The study case confirms with real application that Global alignment technique is most suitable for closely related sequences of similar lengths, and Local alignment method gives better family classification and grouping of scoring match of alignment that may be inferred to the importance of the Spike region chosen for Local alignment for virus family classification.

Methods and Materials

Needs **Biopython** command line to compute score match. Needs **21 SARSCOV2** samples obtained from **GISAID** website[1][2]. Pairwise alignment needs two sequences to do comparison and the resulting will be the matching score, the bigger implies the best match and that indicate there is structural, functional and evolutionary relationship between the two sequences. Verify if there is relationship between the obtained results and what presented about sequence alignment in [3],[4][5]. **The reference** genome sequence used SARS-COV2 isolate **Wuhan-Hu-1**, complete genome NCBI Reference Sequence [6]. **The Global alignment performs matching from end to end sequences, But Local alignment try find highest score matching of subregion of query sequence compared to whole sequence reference.** In **the Local alignment** of the query sequence, **Spike region see Figure 1** chosen to use with alignment, because Spike region length around 3800 nucleotide between the next **positions 21,563-25,384** equals **3,822 bp**. For that 5000 nucleotide length of Motif chosen to be 5000 inside it Spike region.

```
ref=SeqIO.parse("Wuhan-Hu.fasta", "fasta")
q=SeqIO.parse("querysequence_file.fasta", "fasta")
q=q.replace('N',")
Globalscore = pairwise2.align.globalms(ref1, q1,match=1, mismatch= -1, open=-1, extend=-0.5,score_only=True)
Localscore = pairwise2.align.localms(ref1, q1,match=1, mismatch= -1, open=-1, extend=-0.5,score_only=True)
```

Conclusions

The presented study related with **Global and Local Pairwise Alignment**, to check the effectiveness what defined about their **Algorithms Needlman and Waterman for matching score calculation**, and does it has **relationship** with **sequence structure, function and evolutionary**. relationship proved by real DNA sequences of SARS-COV2 genome sequence from different countries shared by **GISAID** website. In case **Local** alignment has better family **SARS-COV2 grouping and classification**, that may infer **to the importance of Spike region** for SARS-COV2 classification. This case study also confirms the challenging of computing ability where even free Colab platform meets difficulties just to computer **Pairwise Alignment of SARS-COV2 with 30K nucleotide base long**. Similar to the presented can considered as first step into searching low and optimal sequence regions using Artificial Intelligence, helpful for microorganism and viruses classification and phylogenetic analysis.

Contact

M. Khodja. Phd in intelligent control engineering.
Instructor of Bioinformatics in faculty of science.
Msila University in Algeria
Email: mohamedabdellah.khodja@univ-msila.dz
Phone: 00213549885011

References

1. S. Khare et al., "GISAID's Role in Pandemic Response," China CDC Weekly, vol. 3, no. 49, pp. 1049–1051, 2021, doi: 10.46234/ccdcw2021.255.
2. Y. Shu and J. McCauley, "GISAID: Global initiative on sharing all influenza data – from vision to reality," Eurosurveillance, vol. 22, no. 13, Mar. 2017.
3. T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," Journal of Molecular Biology, vol. 147, no. 1, pp. 195–197, Mar. 1981.
4. Y. Hasija, "Algorithms in computational biology," in All About Bioinformatics, Elsevier, 2023, pp. 77–104.
5. S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," Journal of Molecular Biology, vol. 48, no. 3, pp. 443–453, Mar. 1970.
6. F. Wu et al., "A new coronavirus associated with human respiratory disease in China," Nature, vol. 579, no. 7798, pp. 265–269, Mar. 2020.
7. Hu, Ben, et al. "Characteristics of SARS-CoV-2 and COVID-19." Nature Reviews Microbiology 19.3 (2021): 141-154.
8. Lu, Roujian, et al. "Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding." The lancet 395.10224 (2020): 565-574.
9. Wang, Shichuan, et al. "A crowned Killer's resume: genome, structure, receptors, and origin of SARS-CoV-2." Virologica Sinica 35 (2020): 673-684.
10. Zhou, Peng, et al. "A pneumonia outbreak associated with a new coronavirus of probable bat origin." nature 579.7798 (2020): 270-273.