JBAAR

Journal of Bioscience and Applied Research
https://jbaar.journals.ekb.eg
SPBH-IC: Biomedical pISSN: 2636-4077, eISSN: 2636-4085

SPBH

Genomic analysis of SARS-COV2 using Biopython and Simplot comparison.

Mohammed Khodja[1]*, Azzedine Melouki[2], Benazi Nabil[3], Dehimat Abdelouahab[4]

[1] Microbiology and Biochemistry Department, University of M'sila, Sciences Faculty, Algeria
E-mail: mohamedabdellah.khodja@univ-msila.dz

[2] Chemistry department, University of M'sila, Sciences Faculty, Algeria
E-mail: azzedine.melouki@univ-msila.dz

[3]Institut PASTEUR Algérie, Annexe M'sila 28000, Alegria
E-mail: benmsila@hotmail.fr

[4]Natural and Life Sciences Department, University of M'sila, Sciences Faculty, Algeria
E-mail: a-ouahab.dehimat@univ-msila.dz

***Corresponding author:** Mohammed Khodja
E-mail: mohamedabdellah.khodja@univ-msila.dz

DOI:10.21608/jbaar.2024.332656.1098

Abstract:

The presented work case interests pairwise sequence alignment algorithms for two analysis cases. The first analysis is SARS-COV2 Omicron in Algerian compared to the neighbor countries using Global and local alignment algorithms, and the second compared coronaviruses from animal sources like Pangolin, Bat, Civet, and Camal, to predict its origin. Of course, with real applications and samples from GISAID, NCBI, and GenBank, and deal with them using the Biopython command line in the Colab platform. The presented work confirms that the local alignment algorithm for the spike region provides better family classification and grouping by matching score sorting. The identity percentage of SARS-COV-2 compared to animal coronaviruses published previously using Simplot software from other authors have wrong values and reviewed in the presented work with the right values using the Biopython library. A high matching score found in bats and pangolins confirms that the origin of SARS-COV-2 is from animal wildlife.

Keywords: Pairwise sequence alignment, Needleman-Wunsch algorithm, Smith-Waterman's algorithm, Biopython, Simplot software, coronaviruses.

Introduction

This work will deal with genomic software and biomolecular databases to verify the concept of DNA sequencing alignment theory and algorithm, for this work is divided into two parts. The first part checks the effectiveness of pairwise sequence alignment in our case with SARS-COV-2 Omicron of Algerian and its matching with another version of neighboring countries, of course with real samples published in famous international databases specialized for collecting genomes of SARS-COV2. The Biopython software and 21 SARS-COV2 samples from different countries are used in our case, three of them collected by Dr. Benazi as he is a member of Pasteur Laboratory of Msila and another three Algerian samples shared by GISAID website(1-5), the left 15 also obtained from the same website from different countries like England, Italy, Pakistan, France …etc. the second part will review what published previously about similarity percentage of SARS-COV2 compared to animal coronaviruses like Pangolin, Bat, Civet and Camal. The global and local alignments were performed in this study to check their effectiveness

Received: September 30, 2024. Accepted: December 1, 2024. Published: December 15, 2024

related to the structural and evolutionary relationship between the samples used in this study. Even though the development of calculators and genomic software is currently attracting an excellent level, the processing of genomic sequences remains long and tiring in addition to the calculation processing the genomic sequence is not deprived of achieving optimal results. Several algorithms tend to minimize the time of calculations and improve storage capacity management as in (1,6,7), where the challenge of tools and genomic data analyses improves, suggesting a fast method of sequencing data query by command line. Faced with these problems propose ideas towards sequence treatment with a minimum of time and costs with accurate results, to obtain a genomic classification of living organisms in particular the viruses our case study concentrates sure the classification of SARSCOV2 coronavirus.

The Biopython Global and Local Pairwise alignment algorithm was used in the presented study to confirm their concept with real example applications. Pairwise Alignment is a method for aligning two sequences that aims to find the most optimum pairwise alignments by searching for the highest similarity score. The applied methodology has been extensively utilized in the examination of sequences to investigate their functional, evolutionary, and structural characteristics. When two sequences exhibit a significant level of similarity, they might be classified as belonging to the same family(1). The first method used is Global alignments based on the Needleman-Wunsch algorithm was introduced in 1970 (2), It attempts to align entire sequences, which are typically of equal length. The second method called Local alignment is alignments describing the most similar region within the sequences to be aligned, short sequence versus longer one, the alignment is based on Smith-Waterman's algorithm which first published in 1980(3).

Researchers' interest in sequencing alignment and phylogenetics trees build, it's clear from many published papers, like in (6) develop a web application called NEXTRAIN Grasping the distribution and pathogen evolution with interactive data visualizations of SARSCOV2. Also in (7) Examine Pakistani SARS-CoV-2 genetic variations using GISAID samples and some NEXTSTAIN analysis. Another research (8,9) depict a new strain of SARS-CoV-2 that has different spike mutations and the importance of this region for virus classifications. In the presented case interest in validating the sequence alignment method by using real samples published in NEXTRAIN and GISAID of Algerian Omicron SARS-COV2. The present work answers the next question: does the Global and Local alignment give information about the structural and evolutionary relationship? The present work it's the first step for future work related to searching minimum coding regions by using an Artificial intelligence algorithm helping in SARS-COV2 classification that positively on computing ability, especially for phylogenetic tree building purposes.

The identity of animal coronaviruses compared to Human SARS-COV2 was performed previously in many literature and reviewed in this work. In (10) mention comparison percentage of the whole genome of SARS-COV2 compared to animal coronaviruses: Human SARS-COV has 79.6% similarity to SARS-COV2, and 96.2% of a bat coronaviruses 'RaTG13' to SARS-COV2, and 88% of bat coronavirus 'ZC45 and ZXC21'. In (11) mention comparison percentage of whole genome of SARS-COV2 compared to animal coronaviruses: Human SARS-COV has 82% of similarity to SARS-COV2, and 96.2% of a bat coronaviruses 'RaTG13' to SARS-COV2, and 88% of bat coronavirus 'ZC45 and ZXC21', 90% coronavirus of Pangolin, and confirm the possibility of a zoonotic origin of SARS-COV2. In (12) mentions the comparison percentage of the whole genome of SARS-COV2 compared to animal coronaviruses: Human SARS-COV has 79% similarity to SARS-COV2 and 50% of Middle East respiratory syndrome coronavirus (MERS-CoV) to SARS-COV2. In (13) mentions comparison percentage of the whole genome of SARS-COV2 compared to animal coronaviruses: Human SARS-COV has 79% of similarity to SARS-COV2, and 50% of Middle East respiratory syndrome coronavirus (MERS-CoV) to SARS-COV2, and 88% of a bat coronaviruses 'RaTG13' to SARS-COV2, and 88% of bat coronavirus 'ZC45 and ZXC21', the hidden virus reservoir in wild animals and their potential to occasionally spill over into human populations. In (14) suggest the origin of SARS-COV-2 is Pangolin and whole genome similarity to SARS-COV-2 are: Pangolin coronavirus 91%, and 90% bat RaTG13 identical to SARS-COV2, and 85% of bat coronavirus 'ZC45 and ZXC21'. In (15) asked the scientific community to review all that was published about sequences of coronavirus RaTG13 and RmYN02(16), and critic what was published in (10) and other research papers, said the samples used to support the natural origin of SARS-COV-2 were sequenced during pandemic and did not previously. It should be confirmed that the identity percentage in these published papers (10–14,16) were computed using Simplot V3.5.1 2003 (17) which is technically too old software and may not be too accurate, which differs from than presented work using Biopython considered more accurate than Simplot.

The presented work has four sections: the introduction and the literature review. The second section presents the algorithm of Global and Local Alignment used in this study. The third section contains the results and discussion. Lastly, the conclusion of the presented work and analysis.

Software tools and methods

In this case, the study uses the Biopython command line to compute the score match of 21 samples obtained from the GISAID website(18,19)] and 19 samples from GenBank and NCBI websites, all used samples in this work in "Table 1". Pairwise sequence alignment needs two sequences to do a comparison, and the resulting will be the matching score, the bigger implies the best match and that indicates there is a structural, functional and evolutionary relationship between the two sequences. Two analyses performed the first pairwise sequence alignment just between different country SARS-COV-2 comparison, and the second analysis pairwise sequence alignment of SARS-COV-2 with animal coronavirus (bat, pangolin, civet, MERS). It should be confirmed that the identity percentage in these published papers (10–14,16) were computed using Simplot V3.5.1 2003 (17) which is technically too old software, and may not be too accurate, which differs from than presented work Biopython is considered more accurate than Simplot.

The first analysis

For Global sequence alignment "GA" compares between whole genome reference of SARS-COV-2, and the query sequence. Which is one sequence from 21 samples, and record the corresponding score in "Table 2", and repeat that for Local sequence alignment "LA" for each sample the obtained results are shown in "Table 3", after filling the tables will perform analysis on the obtained results, and check if there is the relationship between the obtained results, and what presented about sequence alignment in (2,3,20). The reference genome sequence used is Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome NCBI Reference Sequence: NC_045512.2 (21). The Global alignment performs matching from end-to-end sequences its algorithm in "Fig 1", But Local alignment tries to find the highest score matching of the subregion of the query sequence compared to the whole sequence reference, its algorithm shown in "Fig 2" In the Local alignment of the query sequence, Spike region chosen to use with alignment because Spike region length around 3800 nucleotides between the next positions 21,563-25,384 equals 3,822 bp. For that 5000-nucleotide length of Motif was chosen around the Spike region.

Table 1. Genomic sequences of coronaviruses from different hosts

Coronaviruses	The host	GenBank, NCBI, GISAID, Accession number	Length Pb
AY390556_GZ02	Human SARS-COV 2003	AY390556.1	29760
SARSCOV_BJ01	Human SARS-COV 2003	AY278488.2	29725
SARScovGD01	Human SARS-COV 2003	AY278489.2	29757
sarscov_Tor2	Human SARS-COV 2003	NC_004718.3	29751
Bat SARS-cov WIV1	Bat Rhinolophus sinicus 2013	KF367457.1	30309
Bat RaTG13	Bat Rhinolophus affinis 2020	MN996532.2	29855
bat-SL-CoVZC45	Bat Rhinolophus pusillus 2018	MG772933.1	29802
bat-SL-CoVZXC21	Bat Rhinolophus pusillus 2018	MG772934.1	29732
BtRs-BetaCoVGX2013	Bat Rhinolophus sinicus 2014	KJ473815.1	29161
BatSARSlikecov Rs4231	Bat Rhinolophus sinicus 2016	KY417146.1	29782
SARS_CoV_A022	Palm Civet 2004	AY686863.1	29499
SARS_CoV_SZ3	Civet_2003	AY304486.1	29741
SARScov_HCGZ3203	Himalayan palm civets (Paguma larvata) 2004	AY545918.1	29737
PangolinGD	Manis javanica 2020	MT121216.1	29521
PangolinGX	Pangolin 2020	MT072864.1	29795
Merscov	Middle East respiratory syndrome-related coronavirus (MERS-CoV) 2012	NC_019843.3	30119
Bat CpYunnan2011	Bat Chaerephon plicata 2012	JX993988.1	29452
Bat Hp-betacoronavirus	Bat Hipposideros pratti 2013	NC_025217.1	31491
bat RmYN02	Bat Rhinolophus malayanus 2019	GISAID:412977	29671

Global alignments algorithm	***Local alignments algorithm***
Ref=call the reference sequence	***Ref=call the reference sequence***
Queryseq=Call 21 samples	***Queryseq=Call 21 samples***
Set configuration of alignments	***Set configuration of alignments***
For each sample do	***For each sample do***
Score= call global align (ref, queryseq)	***Score= call global align (ref, queryseq)***
Record the score in table	***Record the score in table***
end	***end***

Fig. 1 Global and local sequences alignment algorithms

Received: September 30, 2024. Accepted: December 1, 2024. Published: December 15, 2024

Main Python code part

```
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)
q1=q[21000:26000]
Localscore = pairwise2.align.localms(ref1, q1,match=1, mismatch= -1,
open=-1, extend=-0.5,score_only=True)
```

Fig. 2 Python program code for global and local sequences alignment

The second analysis

In this section performs pairwise sequence alignment to get a matching score between SARS-COV-2 and coronavirus from animal sources. The obtained matching score was converted to a percentage to compare it with what was published previously in (10–14,16) are summarized in “Table 4” and the recalculation performed in the presented work are shown in “Table 5”. Different coronavirus hosts used in this study are presented in the next “Table 1”. All genomic sequences were collected from NCBI, GenBank, or Gisaid. The exact Python program to perform Global and Local alignment and to compute matching scores is shown in “Fig. 2”.

Results and discussion

The first analysis

The next three samples (EPI_ISL_15946159, EPI_ISL_16242292, EPI_ISL_16242296) were chosen specifically, because one member of this study participated in the collection of samples and three other Algerian samples to compare them with Omicron of other countries like England, Italy, France. “Table 2” represents the obtained results using Global Pairwise Sequences Alignment which performs an end-to-end alignment of the query sequence with the reference sequence Wuhan-Hu-1, After calculating the matching score of all samples compared with Wuhan-Hu-1 reference genome assembly, filled the results in the “Table 2” and sorted from lowest to highest matching score GA% (Global alignment matching score in percentage). In a general analysis, the matching score is very high, and that confirms the principle of Global alignment which mentions in many literatures, that the technique is most suitable for closely related sequences of similar lengths, it is very clear in “Table 2” such all calculated score is around 98% of the match, but can notice Omicron SARS-COV2 BA.5.2 distributed by four groups mentioned by Blod font and line, between them different Omicron version that may be indicating the Global aligning give just general idea about sequences how related, with less classification between sequences.

Table 2. Global Sequence Alignment sorted from lowest to highest matching score GA_%

Countries	**GSAID Accession ID**	**Clad**	**Length**	**Global Score vs Wuhan-Hu-1**	**GA_%**	**LA_ Score**	**LA_%**
Pakistan	15111973	**BA.5.2 22B**	28727	28009	97.50%	3564	93.27%
Algeria	16454585	**BA.5.2**	28804	28115	97.61%	3739.5	97.87%
Algeria	17182683	22E BQ.1.1.59	29086	28516.5	98.04%	3724.5	97.47%
Algeria	16242296	**BA.5.2**	29099	28555.5	98.13%	3739.5	97.87%
England	13841928	**BA.5.2 22B**	29229	28743.5	98.34%	3739.5	97.87%
Algeria	16242292	**BA.5.1.30**	29235	28759.5	98.37%	3737.5	97.81%
Ukraine	16637095	**BA.5.1.3 22B**	29271	28804.5	98.41%	3640	95.26%
Switzerland	18828202	23I (BA.2.86)	29518	29089	98.55%	3666.5	95.96%
Algeria	18830343	XBB.1.5.63 23A	29418	28995.5	98.56%	3720	97.36%
Algeria	15946159	**BA.5.2.27 22B**	29399	28998	98.64%	3739.5	97.87%
France	18913704	23I (BA.2.86)	29685	29309	98.73%	3662	95.84%
England	18969306	23I (BA.2.86)	29737	29408	98.89%	3664.5	95.90%
England	18910386	23I (BA.2.86)	29738	29412	98.90%	3666.5	95.96%
England	15192184	23I (BA.2.86)	29740	29414	98.90%	3666.5	95.96%
Germany	18829988	23I (BA.2.86)	29766	29458.5	98.97%	3644.5	95.38%
England	16440149	**23C (CH.1.1)**	29724	29444.5	99.06%	3725	97.49%
Canada	15978247	BA.5.1.30 22B	29646	29370.5	99.07%	3739.5	97.87%
PuertoRico	15229630	BA.5.1.30 22B	29652	29385.5	99.10%	3711	97.12%
England	15192184	BA.5.2	29681	29416.5	99.11%	3739.5	97.87%
SouthAfrica	13830427	BE.1 22B (BA.5)	29715	29476.5	99.20%	3733.5	97.71%
Italy	14971363	BA.5.2 22B	29768	29543	99.24%	3737.5	97.81%

"Table 3" represents the obtained results using Local Pairwise Sequence 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 filling the results in the "Table 3" and sorted from lowest to highest matching score LA_% (Local alignment matching score in percentage). In general analysis of the matching score is very high, and that confirms the principle of Local sequence alignment which mentions 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 3 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 sequence alignment using for Spike region gives better grouping of SARS-COV2 family, that may infer to the importance of Spike region.

The second analysis

As suggest in (15), should the researcher community review all that was published about SARS-COV-2 origin which is the main goal of the present work, next table 4 collects what published in (10) (11) (12) (13) (14,16) about the similarity percentage of SARS-COV-2 compared with animal coronaviruses.

Table 3. Local Sequence Alignment sorted from lowest to highest matching score LA_%

Countries	GSAID Accession ID	Clad	Length	Global Score vs Wuhan-Hu-1	GA_%	LA_ Score	LA_%
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%

Received: September 30, 2024. Accepted: December 1, 2024. Published: December 15, 2024

Table 4. Identity of coronaviruses percentage from different sources compared to SARSCOV2

	Identity type	**From** (8)	**From** (9)	**From** (10)	**From** (11)	**From** (12)	**From** (14)
Human SARS-COV 2003	Whole genome	79.60%	82.00%	79.00%	79.00%	73.00%	79.00%
	Spike identity	73.40%	-	73.00%	72.00%		72.30%
Bat COV 2013-2016	Whole genome	79.60%	-	-	-	73.00%	-
	Spike identity	73.00%	-	75.00%	-	-	-
bat-SL-CoVZC45 2018	Whole genome	88.10%	88.00%	-	88.00%	85.00%	87.00%
	Spike identity	77.80%	-	-	75.00%	-	75.00%
bat-SL-CoVZXC21 2018	Whole genome	88.00%	89.00%	-	88.00%	85.00%	87.00%
	Spike identity	77.10%	-	-	75.00%	-	75.00%
Civet COV 2004	Whole genome	-	-	-	-	-	-
	Spike identity	-	-	76.00%	-	-	-
Pangolin COV 2020	Whole genome	-	91.00%	85.50%	-	93.00%	85.20%
	Spike identity	-	-	91.00%	-	-	83.20%
MERS-COV 2012	Whole genome	-	-	50.00%	50.00%	-	-
	Spike identity	-	-	-	-	-	-
Bat RaTG13 2020	Whole genome	96.20%	96.20%	-	-	96.00%	96.10%
	Spike identity	93.00%	-	96.20%	-	-	93.00%
bat RmYN02 2019	Whole genome	-	-	93.30%	-	-	93.30%
	Spike identity	-	-	-	-	-	72.00%

After applying Algorithms in “Fig. 1” and “Fig. 2” the obtained results in the next “Table 5”:
In this work, Biopython is used for matching score calculation. Which is considered more accurate than Simplot used in (10) (11) (12) (13) (14,16). Needleman and Waterman Algorithms were well implemented successfully in the Biopython library, which helped to obtain the results in “Table 5”.

By doing a comparison between “Table 4” and “Table 5”, big differences were noticed between what was published, and what was recalculated in this study.

In (10) (11) (12) (13) (14,16) as an average state, the whole genome comparison gives 80% SARS-COV similar to SARS-COV2, but after reviewing that with a deep check of pairwise sequence alignment in the Biopyhon library, well implementation of the matching score algorithm gives the percentage just 63%. For Spike region in (10) (11) (12) (13) (14,16) state that 72% of the Spike region of SARS-COV is similar to the Spike region of SARS-COV-2 after recalculation using Biopython gives just 53%.

In in (10) (11) (12) (13) (14,16) state the identity percentage of the whole genome and Spike of Bat RaTG13 has (96%, 93%), and Bat RmYN02 has (93%, 72%), that means RaTG13 has a bigger identity percentage in both cases, wherein the presented work the matching score for whole genome and Spike of RaTG13 has (92%, 86%) and RmYN02 has (88%,88%), that means inverse results of what publish about them previously.

Table 5. Matching score different coronavirus sources compared to SARS-COV-2 obtained by Biopython library.

Coronaviruses	Global Matching score GA%		Local Matching score		
			Spike length	Score	LA%
Merscov	8984.5	30.05%	4061	1013	26.51%
Bat Hpcoronavirus	10029.5	33.54%	3953	1125	29.44%
BtCoVGX2013	17992	60.17%	3728	2003.5	52.43%
Bat CpYunnan2011	18421.5	61.60%	3725	1998	52.29%
SARS_CoV_A022	18511	61.90%	3767	2046	53.55%
Bat SARS-cov WIV1	18720.5	62.60%	3770	2064.5	54.03%
SARScov_HCGZ3203	18864	63.08%	3767	2053	53.73%
SARScovGD01	18904.5	63.22%	3767	2061.5	53.95%
SARSCOV_BJ01	18905	63.22%	3767	2060.5	53.93%
SARS_CoV_SZ3	18909.5	63.24%	3767	2063	53.99%
AY390556_GZ02	18934	63.32%	3767	2063.5	54.00%
sarscov_Tor2	18939.5	63.34%	3767	2053.5	53.74%
BatSARScov Rs4231	18947.5	63.36%	3767	2056.5	53.82%
PangolinGX	21656.5	72.42%	3809	2623	68.65%
bat-SL-CoVZXC21	22983.5	76.86%	3737	2196	57.47%
bat-SL-CoVZC45	23111	77.29%	3740	2251.5	58.92%
PangolinGD	23867	79.81%	3797	2728.5	71.41%
bat RmYN02	26442	88.43%	3953	3385.5	88.60%
Bat RaTG13	27591.5	92.27%	3809	3298	86.31%

MERS-COV is stated to have a 50% identity to SARS-COV-2, but after recalculation, the obtained matching score percentage is just 30%.

It is very clear that there is something wrong with what was published in (10) (11) (12) (13) (14,16), may there be a misunderstanding on how to use Simplot, or the software itself has the wrong identity algorithm implementation.

Conclusion

The presented work is related to the Global and Local Pairwise sequences alignment algorithm "Needleman and Waterman Algorithms". In case Local sequence alignment has better family SARS-COV2 grouping and classification, that may infer the importance of the Spike region. This case study also confirms the challenge of computing ability where even the free Colab platform meets difficulties just to computer Pairwise sequence alignment of SARS-COV2 with 30K nucleotide base long, should the researcher confirm the good computer performance to complete all required sequence genome analysis. The presented work also confirms that there are wrong Identity percentages published previously using Simplot software, some of them stated that the whole genome SARS-COV 2003 has 80% identity with SARS-COV2 2019, which confirmed the wrong value, after recalculation using Biopython, it has just 63%, and repeated that check for spike region were stated in previous published work that has 73% to SARS-COV2, but the right value is just 54%. The cause of the wrong published identity percentage may be for misunderstanding of Simplot software, or the software itself has wrong identity algorithm implementation. the effectiveness of the Biopython pairwise sequence alignment algorithm is confirmed in the presented work, and it is better than Simplot software. Hight matching scores found in bats and pangolins that confirm the origin of SARS-COV-2 is from animal wildlife, specifically from bats.

Acknowledgments

All figures were reprinted/adapted under CC BY license and are property of respective authors. This research did not receive any kind of funding.

Conflict of Interest: NIL

Funding: NIL

References

1. Hasija Y. All About Bioinformatics: From Beginner to Expert. London: Academic Press, an imprint of Elsevier; 2023.

2. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology. 1970 Mar;48(3):443–53.

3. Smith TF, Waterman MS. Identification of common molecular subsequences. Journal of Molecular Biology. 1981 Mar;147(1):195–7.

4. Yameny, A. Characterization of SARS-CoV-2 Omicron XBB.1.5 sub-lineage: A review. *Journal of Medical and Life Science*, 2023; 5(2): 96-101. doi: 10.21608/jmals.2023.305080

5. Yameny, A. COVID-19 Laboratory diagnosis methods. *Journal of Bioscience and Applied Research*, 2023; 9(2): 94-101. doi: 10.21608/jbaar.2023.311827

6. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Kelso J, editor. Bioinformatics. 2018 Dec 1;34(23):4121–3.

7. Basheer A, Zahoor I. Genomic epidemiology of SARS-CoV-2 divulge B.1, B.1.36, and B.1.1.7 as the most dominant lineages in first, second, and third wave of SARS-CoV-2 infections in Pakistan [Internet]. 2021 [cited 2024 Mar 29]. Available from: http://medrxiv.org/lookup/doi/10.1101/2021.07.28.21261233

8. Gladkikh A, Dolgova A, Dedkov V, Sbarzaglia V, Kanaeva O, Popova A, et al. Characterization of a Novel SARS-CoV-2 Genetic Variant with Distinct Spike Protein Mutations. Viruses. 2021 May 29;13(6):1029.

9. Nabil B, Sabrina B, Abdelhakim B. Transmission route and introduction of pandemic SARS-CoV-2 between China, Italy, and Spain. Journal of Medical Virology. 2021 Jan;93(1):564–8.

10. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020 Mar 12;579(7798):270–3.

11. Wang S, Trilling M, Sutter K, Dittmer U, Lu M, Zheng X, et al. A Crowned Killer's Résumé: Genome, Structure, Receptors, and Origin of SARS-CoV-2. Virol Sin. 2020 Dec;35(6):673–84.

12. Hu B, Guo H, Zhou P, Shi ZL. Characteristics of SARS-CoV-2 and COVID-19. Nat Rev Microbiol. 2021 Mar;19(3):141–54.

13. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet. 2020 Feb;395(10224):565–74.

14. Zhang T, Wu Q, Zhang Z. Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. Current Biology. 2020 Apr;30(7):1346-1351.e2.

15. Deigin Y, Segreto R. SARS-CoV-2′s claimed natural origin is undermined by issues with genome sequences of its relative strains: Coronavirus sequences RaTG13, MP789 and RmYN02 raise multiple questions to be critically addressed by the scientific community. BioEssays. 2021 Jul;43(7):2100015.

16. Zhou H, Chen X, Hu T, Li J, Song H, Liu Y, et al. A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. Current Biology. 2020 Jun;30(11):2196-2203.e3.

17. Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, et al. Full-Length Human Immunodeficiency Virus Type 1 Genomes from Subtype C-Infected Seroconverters in India, with Evidence of Intersubtype Recombination. J Virol. 1999 Jan;73(1):152–60.

18. Khare S, Gurry C, Freitas L, B Schultz M, Bach G, Diallo A, Akite N, Ho J, Tc Lee R, Yeo W, Core Curation Team G, Maurer-Stroh S, GISAID Global Data Science Initiative (GISAID), Munich, Germany, (2021) GISAID's Role in Pandemic Response. China CDC Weekly 3:1049–1051

19. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance [Internet]. 2017 Mar 30 [cited 2024 Mar 29];22(13). Available from: https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2017.22.13.30494

20. Hasija Y. Algorithms in computational biology. In: All About Bioinformatics [Internet]. Elsevier; 2023 [cited 2024 Mar 29]. p. 77–104. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780443152504000046

21. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020 Mar;579(7798):265–9.

Graphical abstract

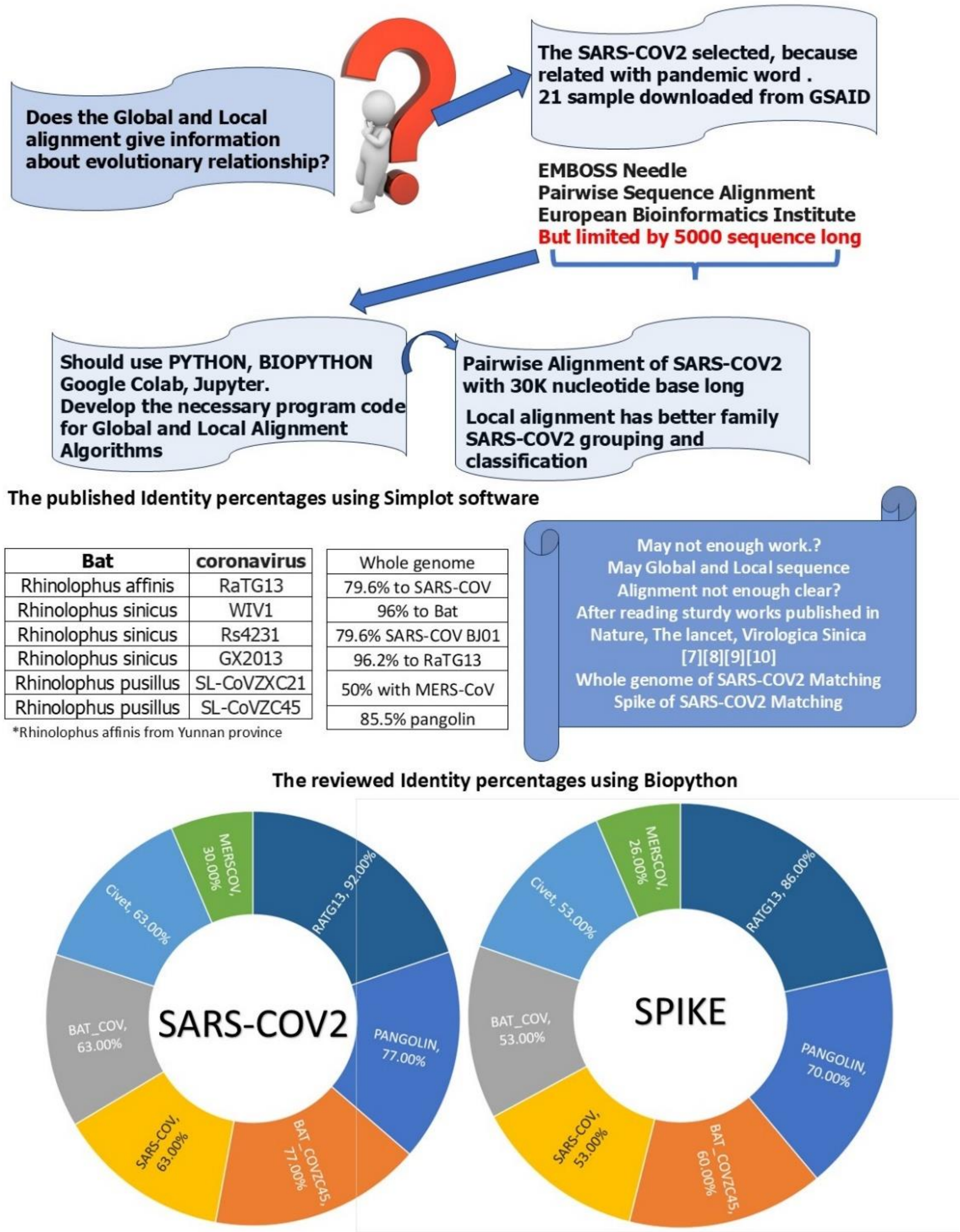

Figure 1. Global Alignment "Needleman Algorithm" **Figure 2.** Local Alignment "Waterman Algorithm".

Highlights

- Matching score of Spike region give better classification of SARS-COV-2.
- Identity percentage of coronaviruses from different sources published previously reviewed.
- The obtained Identity percentage differ than other published papers previously.
- Biopython library more accurate than **Simplot software** for identity calculation and comparison.
- 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.

Simplot used by the most interest publication related with SARS-COV2 origin