应用 Clustal Omega 分析 SARS CoV-2 不同谱系的进化和变异

中山大学生命科学学院 董承志 19331027

1 工具测试

在进行自选的序列比对前，使用 Clustal Omega 提供的 sample sequence 进行测试，测试中输入的数据为 DNA 序列，该组 DNA 序列如下图中所示。

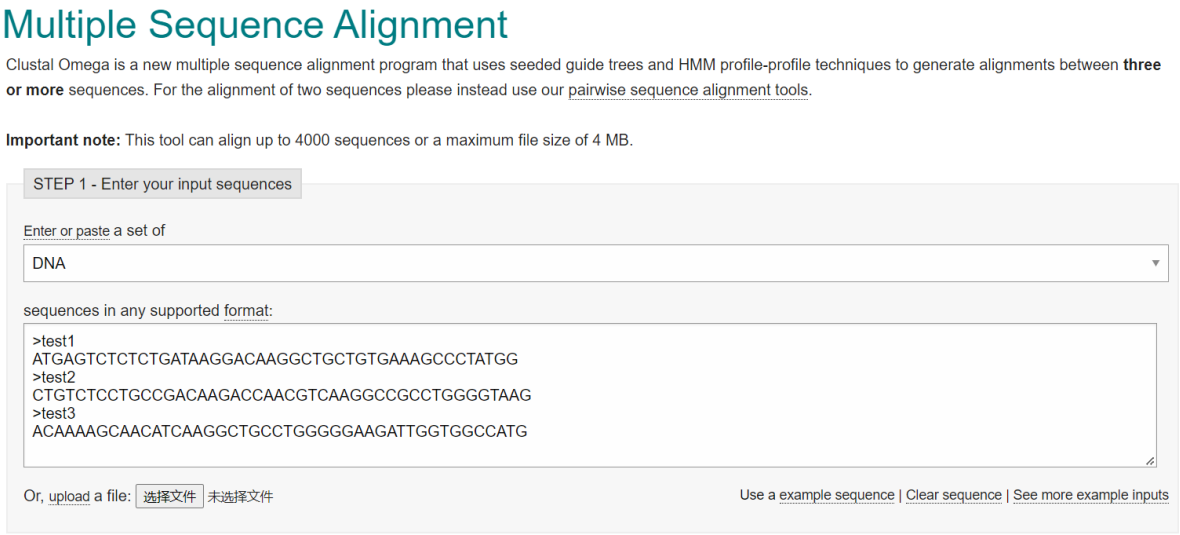

Figure 1: Clustal Omega 提供的 DNA 序列

在进行进一步的参数设置时，使用 Clustal Omega 提供的默认参数（Figure 2），并点击提交。

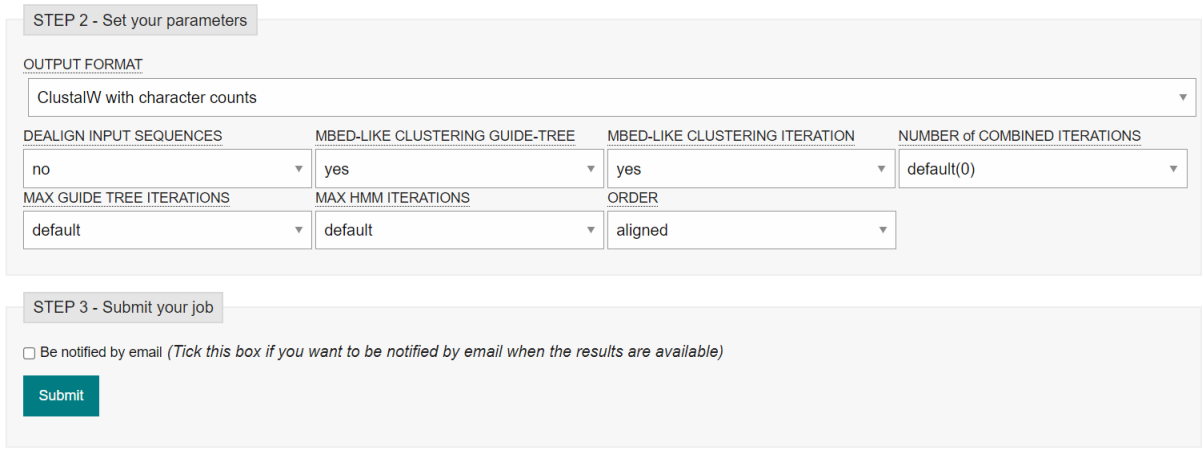

Figure 2: 参数设置

示例序列的比对结果如下图，Clustal Omega 使用了符号“*”来表示输入序列全部匹配的位置，以符号“-”表示插入空位。

```
CLUSTAL O(1.2.4) multiple sequence alignment


test1      ATGAGTCTCTCTGATAAGGACAAGGCTGCTGTGAAAGCCCTATGG---------------- 45
test2      ------CTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAG---------- 45
test3      --------------------ACAAAAGCAACATCAAGGCTGCCTGGGGGAAGATTGGTGGC 41
                              ****        * ** **    ***

test1      ---- 45
test2      ---- 45
test3      CATG 45
```

Figure 3: 示例结果

2 数据搜集

选择冠状病毒作为本次研究对象，探究新型冠状病毒不同谱系的刺突蛋白的进化关系，从而初步了解 SARS CoV-2 的进化和变异情况。首先，通过“VIRALZONE”网站获得 20 个冠状病毒参考毒株基因组中编码刺突蛋白的核苷酸序列，数据搜集结果如下表 1 所示。

表 1 20 种冠状病毒参考毒株基因组中编码刺突蛋白的核苷酸序列

谱系	命名	参考毒株位点
B.1.1.7 + Q.*	Alpha	MW633953
B.1.351 + B.1.351.*	Beta	MW598413
P.1 + P.1.*	Gamma	MW642250
BA.1	Omicron	OL672836
BA.2	Omicron	OM371884
B.1.621 + B.1.621.*	Mu	OK005482
C.37 + C.37.1	Lambda	MW850639
B.1.427/B.1.429	Epsilon	MW643426
B.1.525	Eta	MW560924
B.1.526	Iota	MW643362
B.1.617.1	Kappa	MW966601
P.2	Zeta	MW523796
B.1.1.318	-	MW809039
B.1.1.519	-	MW644499
B.1.466.2	-	MZ006524
R.1	-	MW598432
HCoV_229E	-	NC_002645
HCoV_OC43	-	NC_006213
SARS-CoV	-	NC_004718
Wuhan-Hu-1	-	NC_045512

3 序列比对

本次研究以武汉原始 SARS CoV-2 毒株 Wuhan-Hu-1 作为 SARS CoV-2 的原始对照，HCoV_229E、HCoV_OC43、SARS-CoV 毒株作为外部对照进行序列比对。Clustal Omega 使用过程中选择文件上传序列，保持默认参数。

Figure 4: 上传冠状病毒序列文件

4 比对结果

20 种冠状病毒毒株的编码刺突蛋白的核苷酸序列经过 Clustal Omega 运算后得到如下汇总结果文件：

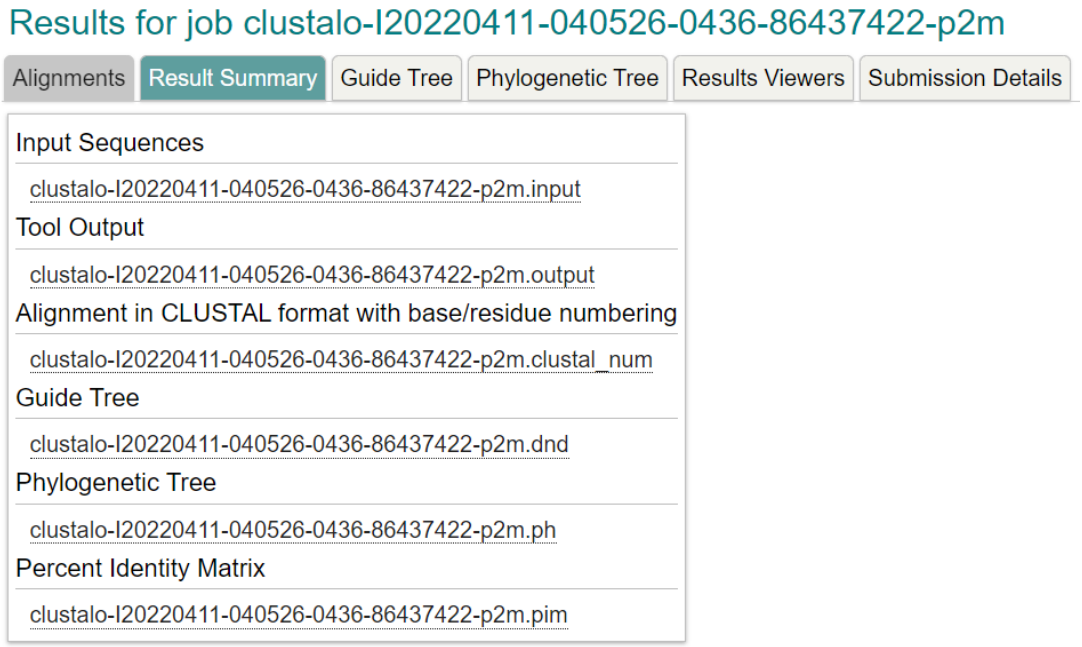

Figure 5: 冠状病毒序列分析汇总结果

4.1 Tool Output

“Tool Output”文件显示共用时 9.41s 完成本次 20 条序列比对。

```
Using 8 threads
Read 20 sequences (type: DNA) from clustalo-I20220411-040526-0436-86437422-p2m.upfile
not more sequences (20) than cluster-size (100), turn off mBed
Calculating pairwise ktuple-distances...
Ktuple-distance calculation progress: 0 % (0 out of 210)
Ktuple-distance calculation progress: 1 % (4 out of 210)
Ktuple-distance calculation progress: 2 % (5 out of 210)
Ktuple-distance calculation progress: 3 % (7 out of 210)
Ktuple-distance calculation progress: 4 % (9 out of 210)
Ktuple-distance calculation progress: 29 % (61 out of 210)
Ktuple-distance calculation progress: 33 % (71 out of 210)
Ktuple-distance calculation progress: 52 % (111 out of 210)
Ktuple-distance calculation progress: 55 % (117 out of 210)
Ktuple-distance calculation progress: 59 % (124 out of 210)
Ktuple-distance calculation progress: 70 % (149 out of 210)
Ktuple-distance calculation progress: 73 % (155 out of 210)
Ktuple-distance calculation progress: 76 % (161 out of 210)
Ktuple-distance calculation progress: 77 % (163 out of 210)
Ktuple-distance calculation progress: 80 % (170 out of 210)
Ktuple-distance calculation progress: 83 % (176 out of 210)
Ktuple-distance calculation progress: 85 % (180 out of 210)
Ktuple-distance calculation progress done. CPU time: 2.03u 0.00s 00:00:02.02 Elapsed: 00:00:01
Guide tree written to clustalo-I20220411-040526-0436-86437422-p2m.dnd
Guide-tree computation done.
Progressive alignment progress: 5 % (1 out of 19)
Progressive alignment progress: 10 % (2 out of 19)
Progressive alignment progress: 15 % (3 out of 19)
Progressive alignment progress: 21 % (4 out of 19)
Progressive alignment progress: 26 % (5 out of 19)
Progressive alignment progress: 31 % (6 out of 19)
Progressive alignment progress: 36 % (7 out of 19)
Progressive alignment progress: 42 % (8 out of 19)
Progressive alignment progress: 47 % (9 out of 19)
Progressive alignment progress: 52 % (10 out of 19)
Progressive alignment progress: 57 % (11 out of 19)
Progressive alignment progress: 63 % (12 out of 19)
Progressive alignment progress: 68 % (13 out of 19)
Progressive alignment progress: 73 % (14 out of 19)
Progressive alignment progress: 78 % (15 out of 19)
Progressive alignment progress: 84 % (16 out of 19)
Progressive alignment progress: 89 % (17 out of 19)
Progressive alignment progress: 94 % (18 out of 19)
Progressive alignment progress: 100 % (19 out of 19)
Progressive alignment progress done. CPU time: 46.39u 9.41s 00:00:55.80 Elapsed: 00:00:53
Alignment written to clustalo-I20220411-040526-0436-86437422-p2m.clustal_num
```

Figure 5: 序列分析所用时长

4.2 Alignment in CLUSTAL format with base/residue numbering

比对结果显示除 P.2 谱系外，其它 SARS CoV-2 谱系均具有较高的相似度。

```
CLUSTAL O(1.2.4) multiple sequence alignment

HCoV_229E           ATGTTTGTTTTGCTTGTTGC---ATATGCCTTGTT------GCATATTGCTGGTTGTCAA      51
HCoV_OC43           ATGTTTTTGATACTTTTAATTTCCTTACCAACGGCTTTTGCTGTTATAGGAGATTTAAAG      60
SARS-CoV            ATGTTTATTTTCT---TATT---ATTTCTTACTCTCACTAGTGGTAGTG---ACCTTGAC      51
P.2                 ------------------------------------------GTGTTAATNNNNNA      14
BA.1                ATGTTTGTTTTTC---TTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATCTTACA      57
BA.2                ATGTTTGTTTTTC---TTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATCTTATA      57
C.37                ATGTTTGTTTTTC---TTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATCTTACA      57
B.1.351             ATGTTTGTTTTTC---TTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATTTTACA      57
P.1                 ATGTTTGTTTTTC---TTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATTTTACA      57
B.1.617.1           ATGTTTGTTTTTC---TTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATCTTACA      57
B.1.526             ATGTTTGTTTTTTT---TTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATCTTACA      57
B.1.427/B.1.429     ATGTTTGTTTTTC---TTGTTTTATTGCCACTAGTCTCTATTCAGTGTGTTAATCTTACA      57
B.1.466.2           ATGTTTGTTTTTC---TTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATCTTACA      57
R.1                 ATGTTTGTTTTTC---TTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATCTTACA      57
B.1.1.519           ATGTTTGTTTTTC---TTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATCTTACA      57
Wuhan-Hu-1          ATGTTTGTTTTTC---TTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATCTTACA      57
B.1.1.7             ATGTTTGTTTTTC---TTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATCTTACA      57
B.1.525             ATGTTTGTTTTTC---TTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATCTTACA      57
B.1.621             ATGTTTGTTTTTC---TTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATCTTACA      57
B.1.1.318           ATGTTTGTTTTTC---TTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATCTTACA      57
                                                                    *

HCoV_229E           ACTACAAATGGGCTGAACAC---TAGTTACTCTGTTTGCAACGGCTGTGTTGGTTATTCA     108
HCoV_OC43           TGTACTTCAGATAATATTAATGATAAAGACACCGGTCCTCCTCCTATAAGTACTGATACT     120
SARS-CoV            CGGTGCACCACTTTTGATGATGTTCAAGCTCCTAATTACACTCAACATACTTCATCTATG     111
P.2                 ACCAGAACTCAATTACNN------------------NNNNNNNNNNNNNNNNNNNNACA      56
BA.1                ACCAGAACTCAATTACCC------------------CCTGCATACACTAATTCTTTCACA      99
BA.2                ACCAGAACTCAAT--------------------------CATACACTAATTCTTTCACA      90
C.37                ACCAGAACTCAATTACCC------------------CCTGCATACACTAATTCTTTCACA      99
B.1.351             ACCAGAACTCAATTACCC------------------CCTGCATACACTAATTCTTTCACA      99
P.1                 AACAGAACTCAATTACCC------------------TCTGCATACACTAATTCTTTCACA      99
B.1.617.1           ACCAGAACTCAATTACCC------------------CCTGCATACACTAATTCTTTCACA      99
B.1.526             ACCAGAACTCAATTACCC------------------CCTGCATACACTAATTCTTTCACA      99
B.1.427/B.1.429     ACCAGAACTCAATTACCC------------------CCTGCATACACTAATTCTTTCACA      99
B.1.466.2           ACCAGAACTCAATTACCC------------------CCTGCATACACTAATTCTTTCACA      99
R.1                 ACCAGAACTCAATTACCC------------------CCTGCATACACTAATTCTTTCACA      99
B.1.1.519           ACCAGAACTCAATTACCC------------------CCTGCATACACTAATTCTTTCACA      99
Wuhan-Hu-1          ACCAGAACTCAATTACCC------------------CCTGCATACACTAATTCTTTCACA      99
B.1.1.7             ACCAGAACTCAATTACCC------------------CCTGCATACACTAATTCTTTCACA      99
B.1.525             ACCAGAACTCAATTACCC------------------CCTGCATACACTAATTCTTTCACA      99
B.1.621             ACCAGAACTCAATTACCC------------------CCTGCATACACTAATTCTTTCACA      99
B.1.1.318           ACCAGAACTCAATTACCC------------------CCTGCATACACTAATTCTTTCACA      99
```

Figure 6: 序列比对结果

4.3 Guide Tree

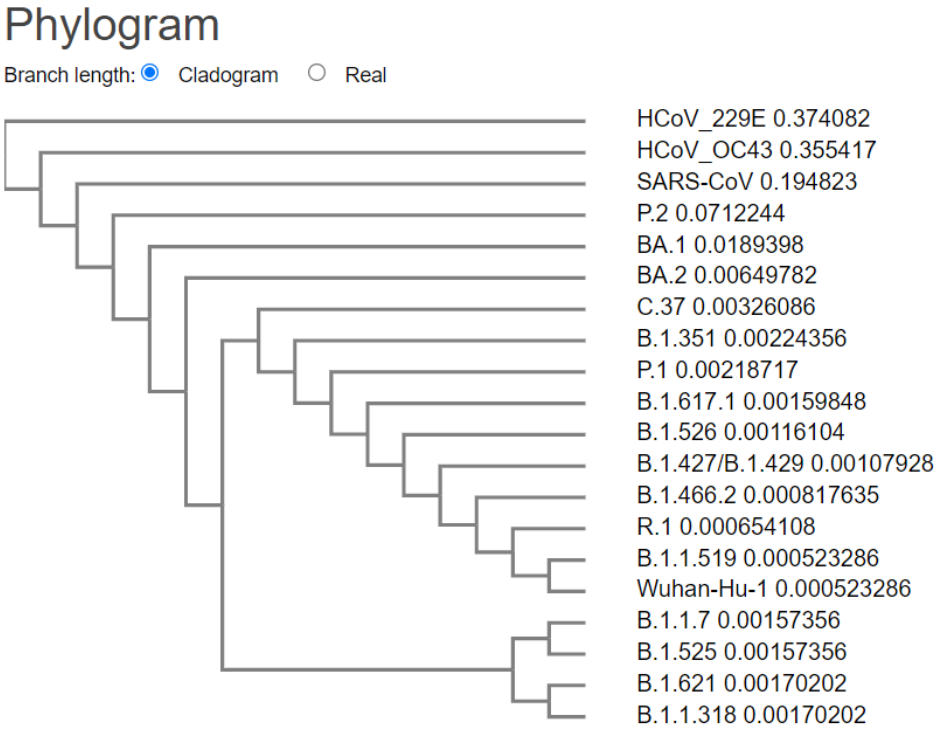

Figure 7: Guide Tree

4.4 Phylogenetic Tree

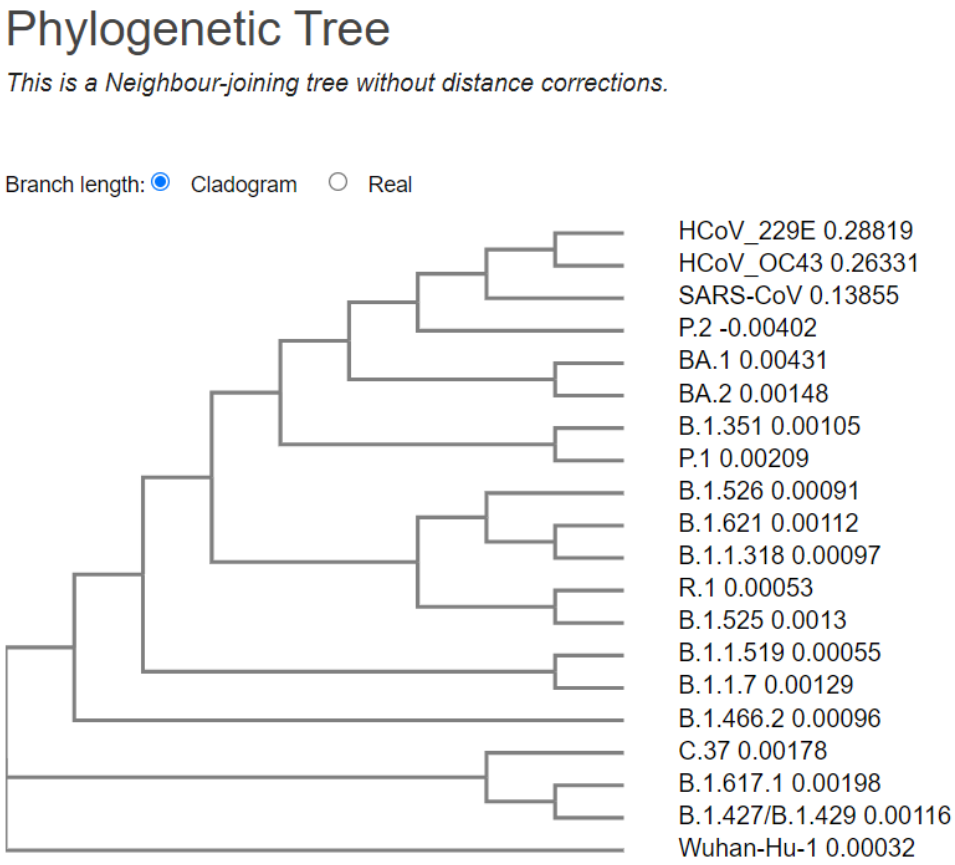

Figure 7: Phylogenetic Tree

4.5 Percent Identity Matrix

Percent Identity Matrix - created by Clustal2.1

1: HCoV_229E	100.00	44.85	44.19	41.28	45.84	45.90	45.96	45.87	45.87	45.83	45.77	45.83	45.77	45.87	45.80	45.87	45.91	45.88	45.83	45.77
2: HCoV_OC43	44.85	100.00	47.14	42.53	48.27	48.33	48.27	48.22	48.37	48.21	48.24	48.26	48.26	48.26	48.32	48.32	48.25	48.27	48.33	48.25
3: SARS-CoV	44.19	47.14	100.00	66.36	73.06	73.45	73.77	73.58	73.51	73.70	73.59	73.59	73.56	73.59	73.67	73.64	73.52	73.57	73.50	73.65
4: P.2	41.28	42.53	66.36	100.00	86.44	86.87	87.70	87.29	87.24	87.16	87.32	87.29	87.32	87.32	87.32	87.37	87.21	87.21	87.12	87.20
5: BA.1	45.84	48.27	73.06	86.44	100.00	99.42	98.78	98.89	98.79	98.77	98.90	98.84	98.90	98.90	99.00	98.95	98.92	98.87	98.98	98.92
6: BA.2	45.90	48.33	73.45	86.87	99.42	100.00	99.10	99.21	99.13	99.08	99.16	99.16	99.21	99.21	99.32	99.27	99.24	99.13	99.19	99.19
7: C.37	45.96	48.27	73.77	87.70	98.78	99.10	100.00	99.63	99.53	99.61	99.68	99.71	99.71	99.74	99.74	99.79	99.66	99.66	99.61	99.63
8: B.1.351	45.87	48.22	73.58	87.29	98.89	99.21	99.63	100.00	99.69	99.61	99.79	99.69	99.71	99.79	99.74	99.79	99.71	99.71	99.71	99.69
9: P.1	45.87	48.37	73.51	87.24	98.79	99.13	99.53	99.69	100.00	99.50	99.63	99.58	99.61	99.69	99.63	99.69	99.61	99.61	99.61	99.58
10: B.1.617.1	45.83	48.21	73.70	87.16	98.77	99.08	99.61	99.61	99.50	100.00	99.66	99.69	99.71	99.71	99.71	99.74	99.63	99.63	99.61	99.66
11: B.1.526	45.77	48.24	73.59	87.32	98.90	99.16	99.68	99.79	99.63	99.66	100.00	99.74	99.76	99.84	99.79	99.84	99.71	99.76	99.76	99.79
12: B.1.427/B.1.429	45.83	48.26	73.59	87.29	98.84	99.16	99.71	99.69	99.58	99.69	99.74	100.00	99.76	99.79	99.79	99.84	99.71	99.71	99.66	99.69
13: B.1.466.2	45.77	48.26	73.56	87.32	98.90	99.21	99.71	99.71	99.61	99.71	99.76	99.76	100.00	99.82	99.84	99.87	99.76	99.74	99.71	99.74
14: R.1	45.87	48.26	73.59	87.32	98.90	99.21	99.74	99.79	99.69	99.71	99.84	99.79	99.82	100.00	99.84	99.90	99.76	99.82	99.76	99.79
15: B.1.1.519	45.80	48.32	73.67	87.32	99.00	99.32	99.74	99.74	99.63	99.71	99.79	99.79	99.84	99.84	100.00	99.90	99.82	99.76	99.76	99.79
16: Wuhan-Hu-1	45.87	48.32	73.64	87.37	98.95	99.27	99.79	99.79	99.69	99.74	99.84	99.84	99.87	99.90	99.90	100.00	99.82	99.82	99.76	99.79
17: B.1.1.7	45.91	48.25	73.52	87.21	98.92	99.24	99.66	99.71	99.61	99.63	99.71	99.71	99.76	99.76	99.82	99.82	100.00	99.69	99.76	99.71
18: B.1.525	45.88	48.27	73.57	87.21	98.87	99.13	99.66	99.71	99.61	99.63	99.76	99.71	99.74	99.82	99.76	99.82	99.69	100.00	99.71	99.71
19: B.1.621	45.83	48.33	73.50	87.12	98.98	99.19	99.61	99.71	99.61	99.61	99.76	99.66	99.71	99.76	99.76	99.76	99.76	99.71	100.00	99.79
20: B.1.1.318	45.77	48.25	73.65	87.20	98.92	99.19	99.63	99.69	99.58	99.66	99.79	99.69	99.74	99.79	99.79	99.79	99.71	99.71	99.79	100.00

5 结果分析

综上多个结果，可得出以下结论：

1. 除 P.2 谱系外，其它 SARS CoV-2 谱系编码刺突蛋白的核苷酸序列均具有较高的相似度；
2. Omicron 的 BA.1 和 BA.2 谱系毒株序列与武汉原始 SARS CoV-2 毒株 Wuhan-Hu-1 相似度最高，与先前的研究结果基本一致，排除了从其它已知谱系进一步进化而来的可能。但是 BA.1 和 BA.2 谱系毒株序列与武汉

原始 SARS CoV-2 毒株 Wuhan-Hu-1 相似度并不是很高，故其有可能是从其它未知的谱系进化而来。

3. 武汉原始 SARS CoV-2 毒株 Wuhan-Hu-1 与其他 SARS CoV-2 谱系均表现出最高的相似度，故目前已知的谱系基本上由最早发现的 SARS CoV-2 毒株进化而来。

6 数据

本文中所有数据均可从 https://github.com/Dongchengzhi/SYSU-Biotechnology/tree/main/Bioinformatics 获取。