華中科技大學

生物信息学上机实验

院　　系　生命科学与技术学院

专业班级　登峰 1901 班

姓　　名　张皓鸿

学　　号　U201912537

2021 年 5 月 3 日

目　　录

1　基因组分析

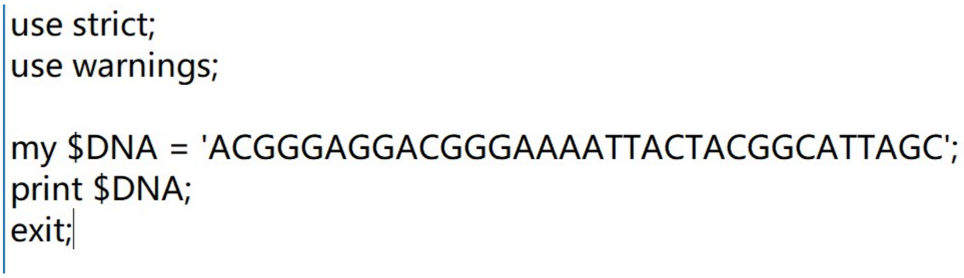
1.1　总结β属冠状病毒和SARS-CoV-2（2019-nCoV）的主要特点

1、都是RNA病毒

2、冠状病毒的s蛋白分为两个功能单元S1和S2。S1通过与宿主受体结合促进病毒感染。它包括两个结构域，n端结构域和c端RBD结构域，它们通过s蛋白与ACE2受体结合而感染人类。

1.2　编写并运行example4-1.pl

编写的example4-1.pl如下

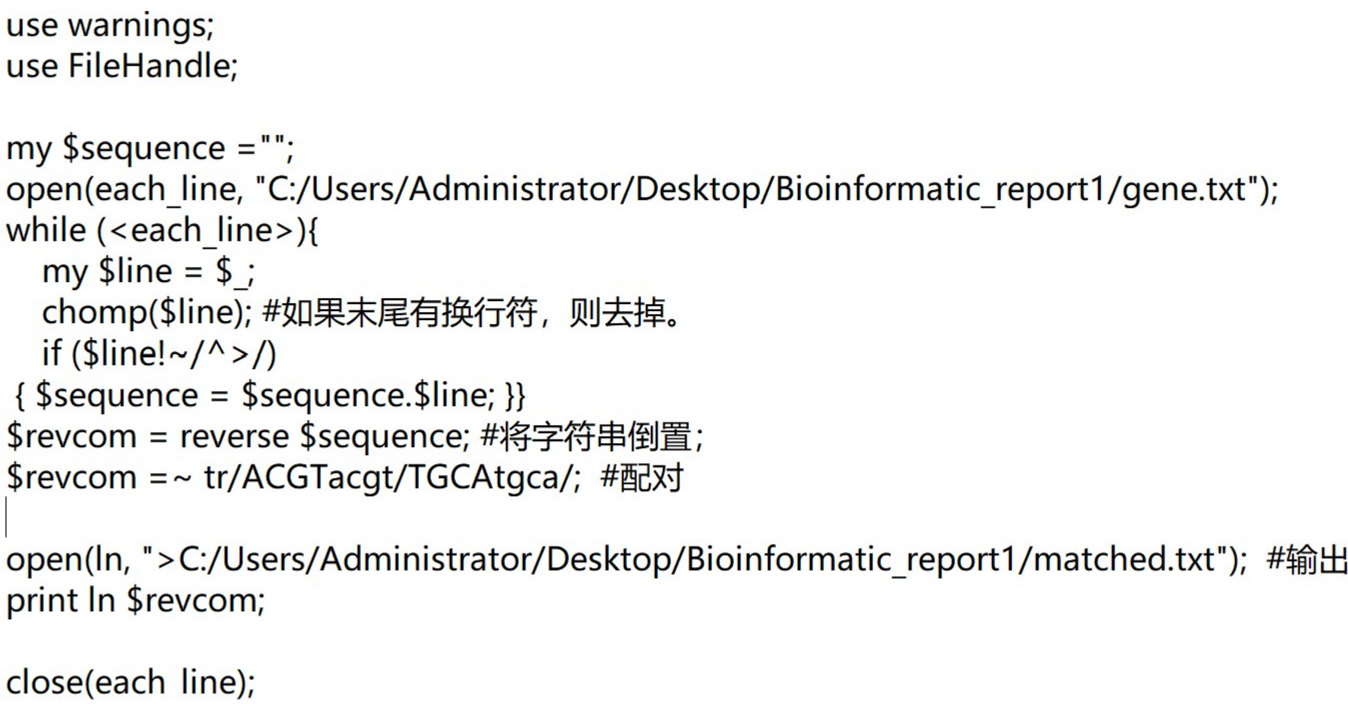


```
use strict;
use warnings;

my $DNA = 'ACGGGAGGACGGGAAAATTACTACGGCATTAGC';
print $DNA;
exit;
```

图1-1　example4-1.pl

1.3　SARS-CoV-2的基因组序列

得到新冠病毒基因的fasta文件后，通过如下perl程序得到其互补序列。

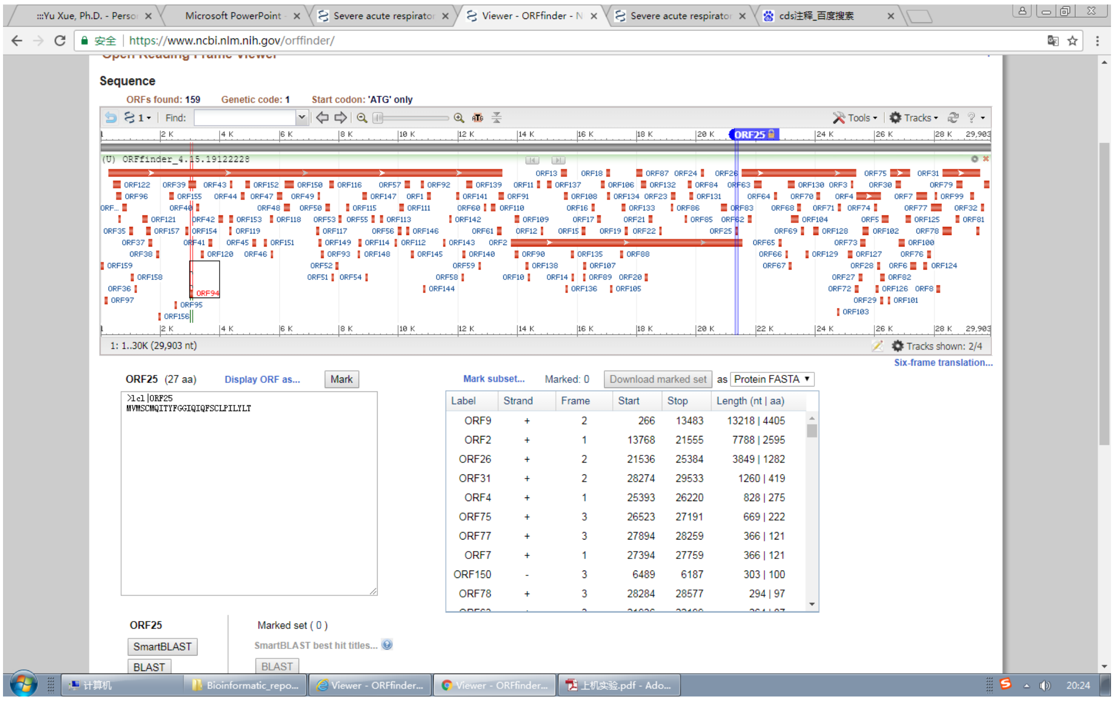


```
use warnings;
use FileHandle;

my $sequence = "";
open(each_line, "C:/Users/Administrator/Desktop/Bioinformatic_report1/gene.txt");
while (<each_line>){
    my $line = $_;
    chomp($line); #如果末尾有换行符，则去掉。
    if ($line!~/^>/)
 { $sequence = $sequence.$line; }}
$revcom = reverse $sequence; #将字符串倒置；
$revcom =~ tr/ACGTacgt/TGCAtgca/;  #配对

open(In, ">C:/Users/Administrator/Desktop/Bioinformatic_report1/matched.txt");  #输出
print In $revcom;

close(each_line);
```

图 1-2　gene_match.pl

1.4　SARS-CoV-2 潜在编码序列的预测

预测开放阅读框如下



https://www.ncbi.nlm.nih.gov/orffinder/

Sequence

ORFs found: 159　Genetic code: 1　Start codon: 'ATG' only

(U) ORFfinder_4.15.19122228

1: 1..30K (29,903 nt)

ORF25 (27 aa)

```
>lcl|ORF25
MVMSCMQITYFGGIQIQFSCLPILYLT
```

Mark subset...　Marked: 0　Download marked set　as Protein FASTA

Label	Strand	Frame	Start	Stop	Length (nt | aa)
ORF9	+	2	266	13483	13218 | 4405
ORF2	+	1	13768	21555	7788 | 2595
ORF26	+	2	21536	25384	3849 | 1282
ORF31	+	2	28274	29533	1260 | 419
ORF4	+	1	25393	26220	828 | 275
ORF75	+	3	26523	27191	669 | 222
ORF77	+	3	27894	28259	366 | 121
ORF7	+	1	27394	27759	366 | 121
ORF150	-	3	6489	6187	303 | 100
ORF78	+	3	28284	28577	294 | 97

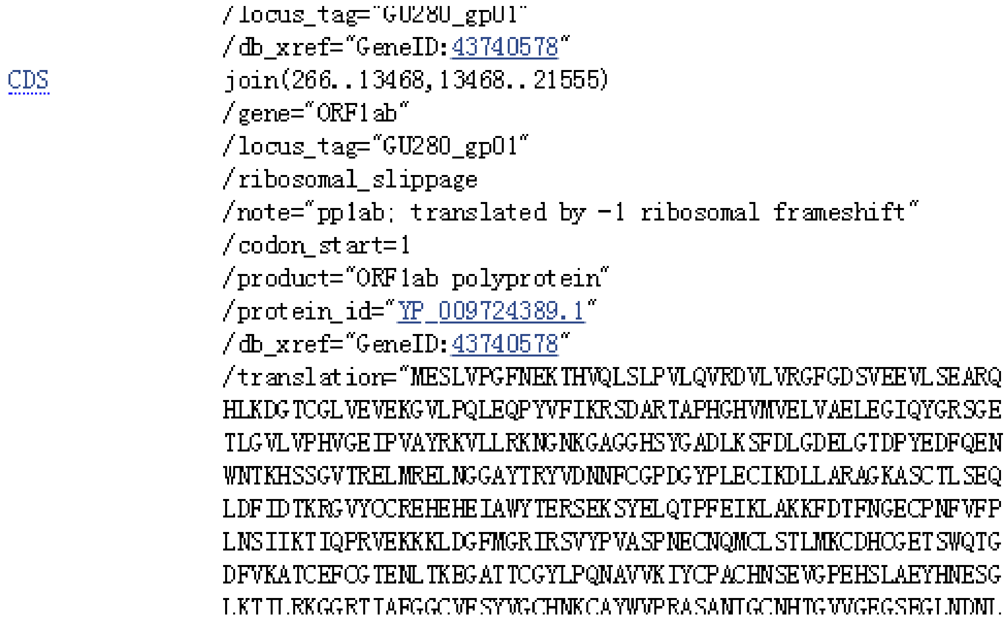
ORF25

Marked set (0)

SmartBLAST best hit titles...

图 1-3　ORF_prediction.PNG

CDS 注释如下



```
                     /locus_tag="GU280_gp01"
                     /db_xref="GeneID:43740578"
CDS                  join(266..13468,13468..21555)
                     /gene="ORF1ab"
                     /locus_tag="GU280_gp01"
                     /ribosomal_slippage
                     /note="pp1ab; translated by -1 ribosomal frameshift"
                     /codon_start=1
                     /product="ORF1ab polyprotein"
                     /protein_id="YP_009724389.1"
                     /db_xref="GeneID:43740578"
                     /translation="MESLVPGFNEKTHVQLSLPVLQVRDVLVRGFGDSVEEVLSEARQ
                     HLKDGTCGLVEVEKGVLPQLEQPYVFIKRSDARTAPHGHVMVELVAELEGIQYGRSGE
                     TLGVLVPHVGEIPVAYRKVLLRKNGNKGAGGHSYGADLKSFDLGDELGTDPYEDFQEN
                     WNTKHSSGVTRELMRELNGGAYTRYVDNNFCGPDGYPLECIKDLLARAGKASCTLSEQ
                     LDFIDTKRGVYCCREHEHEIAWYTERSEKSYELQTPFEIKLAKKFDTFNGECPNFVFP
                     LNSIIKTIQPRVEKKKLDGFMGRIRSVYPVASPNECNQMCLSTLMKCDHCGETSWQTG
                     DFVKATCEFCGTENLTKEGATTCGYLPQNAVVKIYCPACHNSEVGPEHSLAEYHNESG
                     LKTILRKGGRTIAFGGCVFSYVGCHNKCAYWVPRASANIGCNHTGVVGEGSEGLNDNL
```

图 1-4 CDS_annotation.PNG

通过比较预测的 ORF 与 CDS 注释可以发现，预测的 ORF 几乎占据了整段基因，但实际 CDS 仅为其中的一部分，说明基因并不是整段表达的。

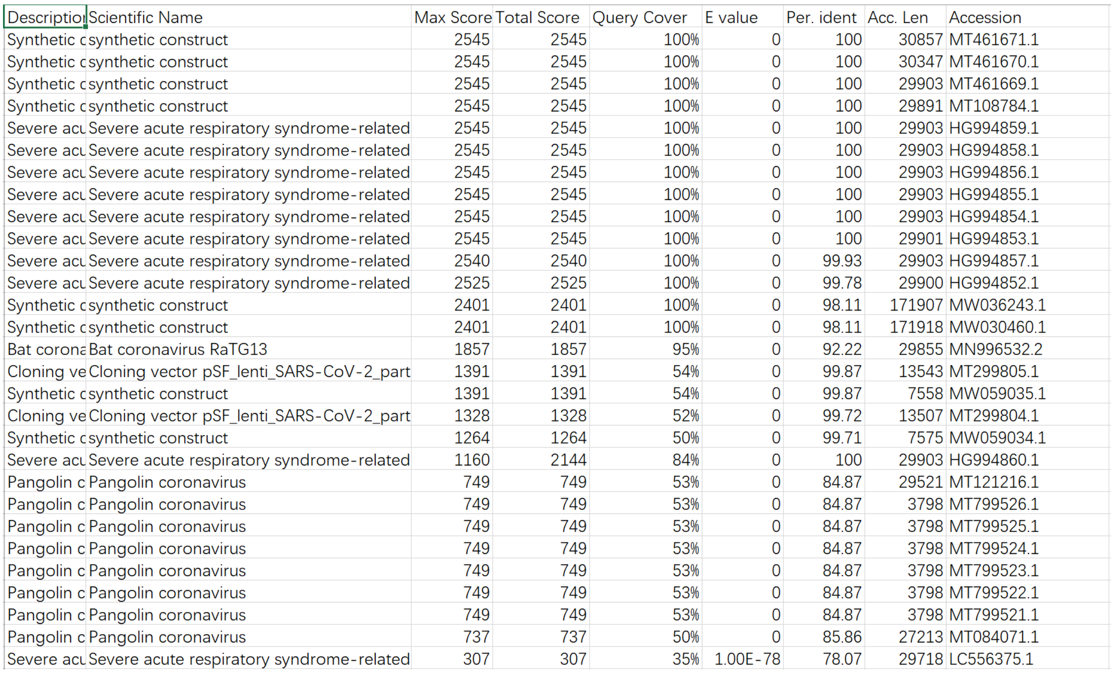
1.5 发现与 SARS-CoV-2 同源的冠状病毒

MT461669.1	MT108784.1	HG994854.1	HG994852.1
HG994857.1	HG994855.1	MT461671.1	MT461670.1
MN996532.2	HG994858.1	HG994859.1	HG994856.1
HG994853.1	MT121216.1	MW703458.1	MT040335.1
MT040333.1	MT072864.1	MT040334.1	MT040336.1

表 1-1 20 个 SARS-CoV-2 的同源冠状病毒的序列号

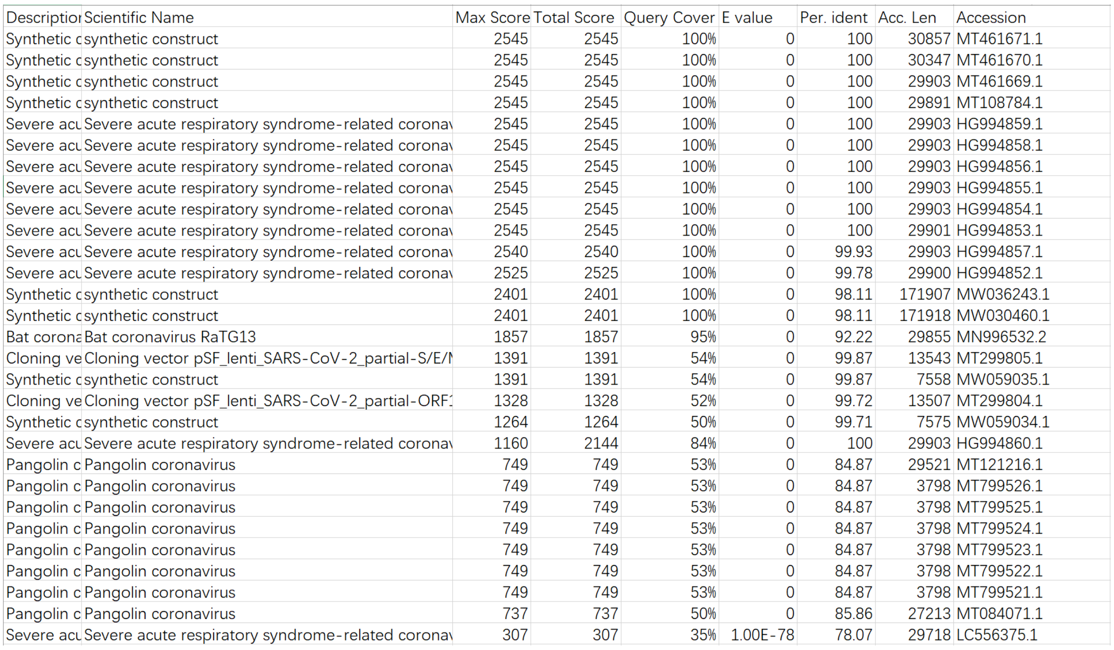
1.6 插入片段分析

使用 megablast, dimegablast, blastn 搜索结果如下



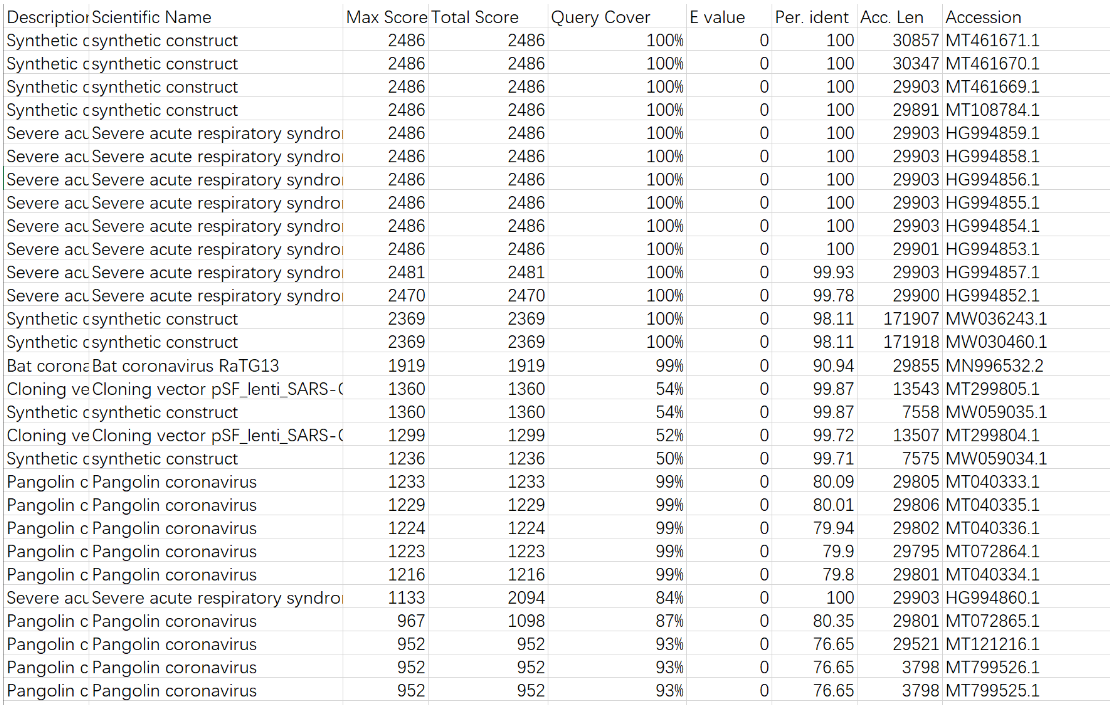
Descriptio	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. ident	Acc. Len	Accession
Synthetic c	synthetic construct	2545	2545	100%	0	100	30857	MT461671.1
Synthetic c	synthetic construct	2545	2545	100%	0	100	30347	MT461670.1
Synthetic c	synthetic construct	2545	2545	100%	0	100	29903	MT461669.1
Synthetic c	synthetic construct	2545	2545	100%	0	100	29891	MT108784.1
Severe acu	Severe acute respiratory syndrome-related	2545	2545	100%	0	100	29903	HG994859.1
Severe acu	Severe acute respiratory syndrome-related	2545	2545	100%	0	100	29903	HG994858.1
Severe acu	Severe acute respiratory syndrome-related	2545	2545	100%	0	100	29903	HG994856.1
Severe acu	Severe acute respiratory syndrome-related	2545	2545	100%	0	100	29903	HG994855.1
Severe acu	Severe acute respiratory syndrome-related	2545	2545	100%	0	100	29903	HG994854.1
Severe acu	Severe acute respiratory syndrome-related	2545	2545	100%	0	100	29901	HG994853.1
Severe acu	Severe acute respiratory syndrome-related	2540	2540	100%	0	99.93	29903	HG994857.1
Severe acu	Severe acute respiratory syndrome-related	2525	2525	100%	0	99.78	29900	HG994852.1
Synthetic c	synthetic construct	2401	2401	100%	0	98.11	171907	MW036243.1
Synthetic c	synthetic construct	2401	2401	100%	0	98.11	171918	MW030460.1
Bat corona	Bat coronavirus RaTG13	1857	1857	95%	0	92.22	29855	MN996532.2
Cloning ve	Cloning vector pSF_lenti_SARS-CoV-2_part	1391	1391	54%	0	99.87	13543	MT299805.1
Synthetic c	synthetic construct	1391	1391	54%	0	99.87	7558	MW059035.1
Cloning ve	Cloning vector pSF_lenti_SARS-CoV-2_part	1328	1328	52%	0	99.72	13507	MT299804.1
Synthetic c	synthetic construct	1264	1264	50%	0	99.71	7575	MW059034.1
Severe acu	Severe acute respiratory syndrome-related	1160	2144	84%	0	100	29903	HG994860.1
Pangolin c	Pangolin coronavirus	749	749	53%	0	84.87	29521	MT121216.1
Pangolin c	Pangolin coronavirus	749	749	53%	0	84.87	3798	MT799526.1
Pangolin c	Pangolin coronavirus	749	749	53%	0	84.87	3798	MT799525.1
Pangolin c	Pangolin coronavirus	749	749	53%	0	84.87	3798	MT799524.1
Pangolin c	Pangolin coronavirus	749	749	53%	0	84.87	3798	MT799523.1
Pangolin c	Pangolin coronavirus	749	749	53%	0	84.87	3798	MT799522.1
Pangolin c	Pangolin coronavirus	749	749	53%	0	84.87	3798	MT799521.1
Pangolin c	Pangolin coronavirus	737	737	50%	0	85.86	27213	MT084071.1
Severe acu	Severe acute respiratory syndrome-related	307	307	35%	1.00E-78	78.07	29718	LC556375.1

图 1-5　megablast_result.PNG



Descriptio	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. ident	Acc. Len	Accession
Synthetic c	synthetic construct	2545	2545	100%	0	100	30857	MT461671.1
Synthetic c	synthetic construct	2545	2545	100%	0	100	30347	MT461670.1
Synthetic c	synthetic construct	2545	2545	100%	0	100	29903	MT461669.1
Synthetic c	synthetic construct	2545	2545	100%	0	100	29891	MT108784.1
Severe acu	Severe acute respiratory syndrome-related corona	2545	2545	100%	0	100	29903	HG994859.1
Severe acu	Severe acute respiratory syndrome-related corona	2545	2545	100%	0	100	29903	HG994858.1
Severe acu	Severe acute respiratory syndrome-related corona	2545	2545	100%	0	100	29903	HG994856.1
Severe acu	Severe acute respiratory syndrome-related corona	2545	2545	100%	0	100	29903	HG994855.1
Severe acu	Severe acute respiratory syndrome-related corona	2545	2545	100%	0	100	29903	HG994854.1
Severe acu	Severe acute respiratory syndrome-related corona	2545	2545	100%	0	100	29901	HG994853.1
Severe acu	Severe acute respiratory syndrome-related corona	2540	2540	100%	0	99.93	29903	HG994857.1
Severe acu	Severe acute respiratory syndrome-related corona	2525	2525	100%	0	99.78	29900	HG994852.1
Synthetic c	synthetic construct	2401	2401	100%	0	98.11	171907	MW036243.1
Synthetic c	synthetic construct	2401	2401	100%	0	98.11	171918	MW030460.1
Bat corona	Bat coronavirus RaTG13	1857	1857	95%	0	92.22	29855	MN996532.2
Cloning ve	Cloning vector pSF_lenti_SARS-CoV-2_partial-S/E/	1391	1391	54%	0	99.87	13543	MT299805.1
Synthetic c	synthetic construct	1391	1391	54%	0	99.87	7558	MW059035.1
Cloning ve	Cloning vector pSF_lenti_SARS-CoV-2_partial-ORF	1328	1328	52%	0	99.72	13507	MT299804.1
Synthetic c	synthetic construct	1264	1264	50%	0	99.71	7575	MW059034.1
Severe acu	Severe acute respiratory syndrome-related corona	1160	2144	84%	0	100	29903	HG994860.1
Pangolin c	Pangolin coronavirus	749	749	53%	0	84.87	29521	MT121216.1
Pangolin c	Pangolin coronavirus	749	749	53%	0	84.87	3798	MT799526.1
Pangolin c	Pangolin coronavirus	749	749	53%	0	84.87	3798	MT799525.1
Pangolin c	Pangolin coronavirus	749	749	53%	0	84.87	3798	MT799524.1
Pangolin c	Pangolin coronavirus	749	749	53%	0	84.87	3798	MT799523.1
Pangolin c	Pangolin coronavirus	749	749	53%	0	84.87	3798	MT799522.1
Pangolin c	Pangolin coronavirus	749	749	53%	0	84.87	3798	MT799521.1
Pangolin c	Pangolin coronavirus	737	737	50%	0	85.86	27213	MT084071.1
Severe acu	Severe acute respiratory syndrome-related corona	307	307	35%	1.00E-78	78.07	29718	LC556375.1

图 1-6　dimegablast_result.PNG



Descriptior	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. ident	Acc. Len	Accession
Synthetic c	synthetic construct	2486	2486	100%	0	100	30857	MT461671.1
Synthetic c	synthetic construct	2486	2486	100%	0	100	30347	MT461670.1
Synthetic c	synthetic construct	2486	2486	100%	0	100	29903	MT461669.1
Synthetic c	synthetic construct	2486	2486	100%	0	100	29891	MT108784.1
Severe acu	Severe acute respiratory syndro	2486	2486	100%	0	100	29903	HG994859.1
Severe acu	Severe acute respiratory syndro	2486	2486	100%	0	100	29903	HG994858.1
Severe acu	Severe acute respiratory syndro	2486	2486	100%	0	100	29903	HG994856.1
Severe acu	Severe acute respiratory syndro	2486	2486	100%	0	100	29903	HG994855.1
Severe acu	Severe acute respiratory syndro	2486	2486	100%	0	100	29903	HG994854.1
Severe acu	Severe acute respiratory syndro	2486	2486	100%	0	100	29901	HG994853.1
Severe acu	Severe acute respiratory syndro	2481	2481	100%	0	99.93	29903	HG994857.1
Severe acu	Severe acute respiratory syndro	2470	2470	100%	0	99.78	29900	HG994852.1
Synthetic c	synthetic construct	2369	2369	100%	0	98.11	171907	MW036243.1
Synthetic c	synthetic construct	2369	2369	100%	0	98.11	171918	MW030460.1
Bat corona	Bat coronavirus RaTG13	1919	1919	99%	0	90.94	29855	MN996532.2
Cloning ve	Cloning vector pSF_lenti_SARS-(	1360	1360	54%	0	99.87	13543	MT299805.1
Synthetic c	synthetic construct	1360	1360	54%	0	99.87	7558	MW059035.1
Cloning ve	Cloning vector pSF_lenti_SARS-(	1299	1299	52%	0	99.72	13507	MT299804.1
Synthetic c	synthetic construct	1236	1236	50%	0	99.71	7575	MW059034.1
Pangolin c	Pangolin coronavirus	1233	1233	99%	0	80.09	29805	MT040333.1
Pangolin c	Pangolin coronavirus	1229	1229	99%	0	80.01	29806	MT040335.1
Pangolin c	Pangolin coronavirus	1224	1224	99%	0	79.94	29802	MT040336.1
Pangolin c	Pangolin coronavirus	1223	1223	99%	0	79.9	29795	MT072864.1
Pangolin c	Pangolin coronavirus	1216	1216	99%	0	79.8	29801	MT040334.1
Severe acu	Severe acute respiratory syndro	1133	2094	84%	0	100	29903	HG994860.1
Pangolin c	Pangolin coronavirus	967	1098	87%	0	80.35	29801	MT072865.1
Pangolin c	Pangolin coronavirus	952	952	93%	0	76.65	29521	MT121216.1
Pangolin c	Pangolin coronavirus	952	952	93%	0	76.65	3798	MT799526.1
Pangolin c	Pangolin coronavirus	952	952	93%	0	76.65	3798	MT799525.1

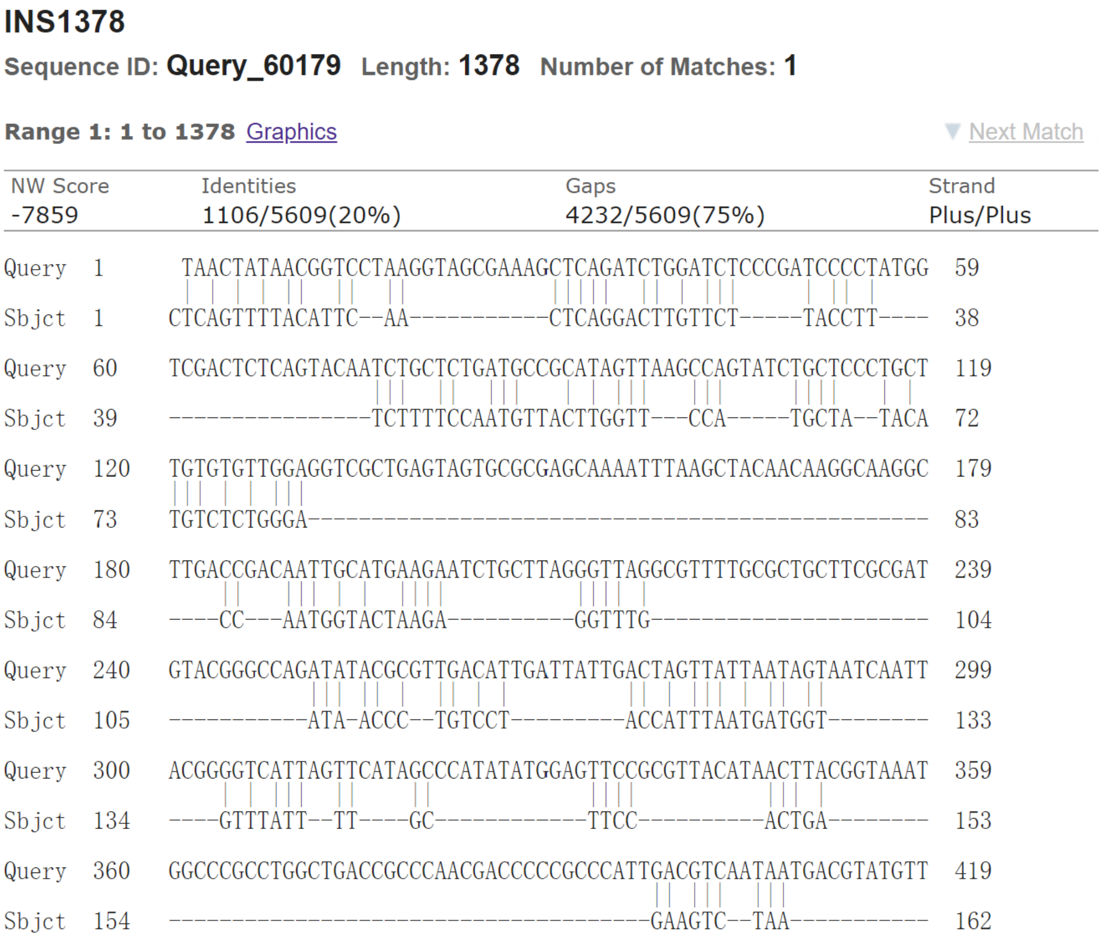
图 1-7　blastn_result.PNG

可以看出 megablast 的结果较少，但使用 dimegablast 与 blastn 允许错配后，可用结果变多，故该基因在冠状病毒中保守，不为人工插入序列。

2　序列分析

2.1　INS1378 与 pShuttle-SN 载体的相似性

Global Align 与 EMBOSS Water 比对结果如下



INS1378

Sequence ID: **Query_60179** Length: **1378** Number of Matches: **1**

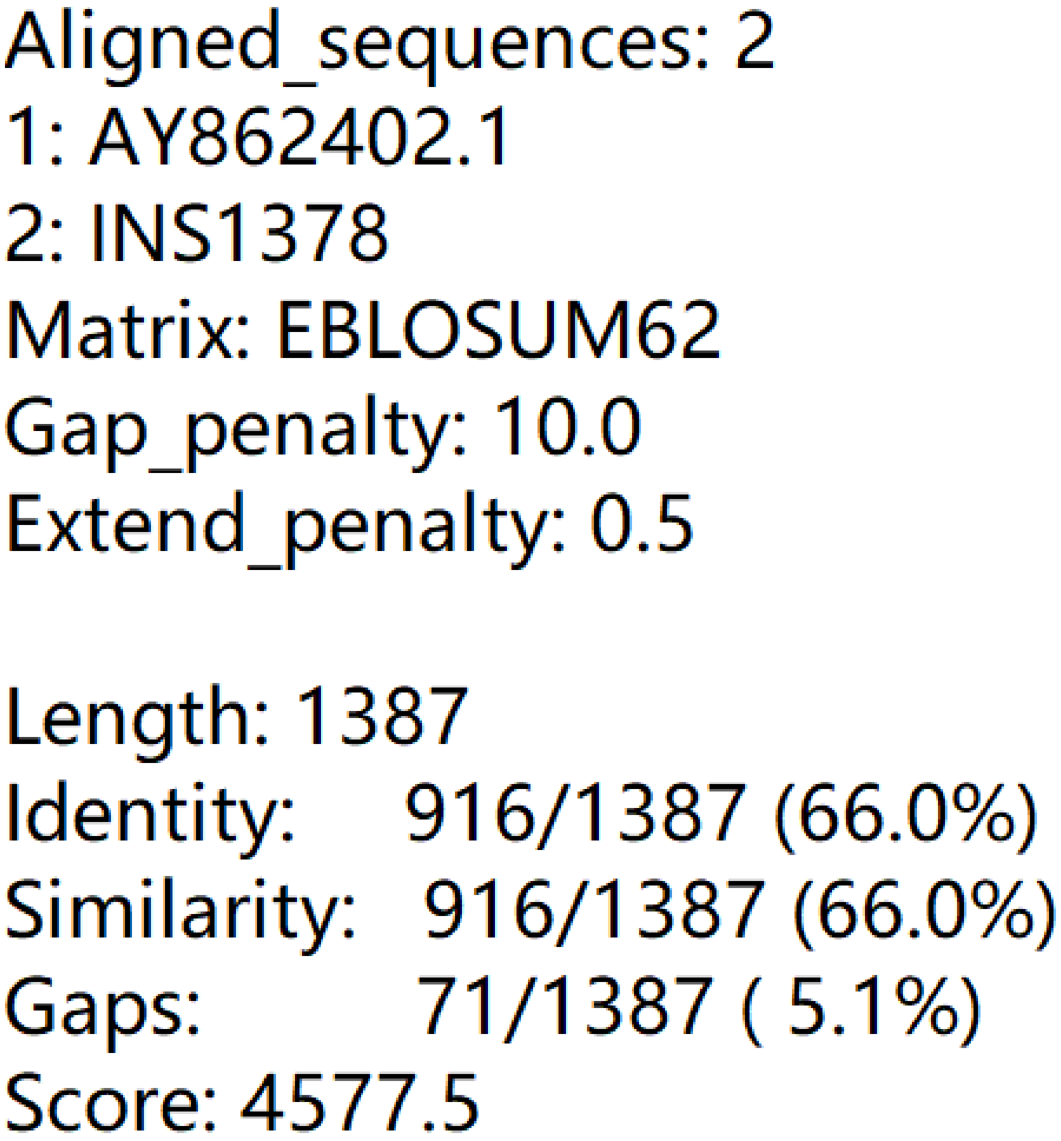
Range 1: 1 to 1378 Graphics

▼ Next Match

NW Score	Identities	Gaps	Strand
-7859	1106/5609(20%)	4232/5609(75%)	Plus/Plus

```
Query  1     TAACTATAACGGTCCTAAGGTAGCGAAAGCTCAGATCTGGATCTCCCGATCCCCTATGG  59
             | | | | ||  ||  ||           |||||  || | |||     | || |
Sbjct  1    CTCAGTTTTACATTC--AA-----------CTCAGGACTTGTTCT-----TACCTT----  38

Query  60   TCGACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATCTGCTCCCTGCT  119
                            |||  ||  |||   | | |||   |||     ||||   | |
Sbjct  39   ----------------TCTTTTCCAATGTTACTTGGTT---CCA-----TGCTA--TACA  72

Query  120  TGTGTGTTGGAGGTCGCTGAGTAGTGCGCGAGCAAAATTTAAGCTACAACAAGGCAAGGC  179
            ||| | | |||
Sbjct  73   TGTCTCTGGGA-------------------------------------------------  83

Query  180  TTGACCGACAATTGCATGAAGAATCTGCTTAGGGTTAGGCGTTTTGCGCTGCTTCGCGAT  239
                ||   ||| | |  ||||          |||| |
Sbjct  84   ----CC---AATGGTACTAAGA----------GGTTTG----------------------  104

Query  240  GTACGGGCCAGATATACGCGTTGACATTGATTATTGACTAGTTATTAATAGTAATCAATT  299
                       ||| || |  || | |         || | ||| | || ||
Sbjct  105  -----------ATA-ACCC--TGTCCT---------ACCATTTAATGATGGT--------  133

Query  300  ACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACATAACTTACGGTAAAT  359
                | | |||  ||    ||            ||||          ||| |
Sbjct  134  ----GTTTATT--TT----GC------------TTCC----------ACTGA--------  153

Query  360  GGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGACGTATGTT  419
                                                  || |||  |||
Sbjct  154  --------------------------------------GAAGTC--TAA-----------  162
```

图 2-1　global_align.PNG



```
Aligned_sequences: 2
1: AY862402.1
2: INS1378
Matrix: EBLOSUM62
Gap_penalty: 10.0
Extend_penalty: 0.5

Length: 1387
Identity:     916/1387 (66.0%)
Similarity:   916/1387 (66.0%)
Gaps:          71/1387 ( 5.1%)
Score: 4577.5
```

图 2-2　EMBOSS_Water.PNG

Global Align 使用的是 Needleman Wunsh 算法，出现负分，且结果序列 gap 较多；EMBOSS Water 使用 Smith Waterman 算法，出现负分则记为零分，结果序列 gap 较少。

2.2 SARS-CoV-2 的蛋白质序列

2.3 等电点与分子量分析

序号	等电点 **pI**	分子量 **Mw**
CDS_1	6.32	794057.79
CDS_2	6.04	489988.91
CDS_3	6.24	141178.47
CDS_4	5.55	31122.94
CDS_5	8.57	8365.04
CDS_6	9.51	25146.62
CDS_7	4.60	7272.54
CDS_8	8.23	13744.17
CDS_9	4.17	5180.27
CDS_10	5.42	13831.01
CDS_11	10.07	45625.70
CDS_12	7.93	4449.23

表 2-1　编码蛋白等电点与分子量

2.4 功能结构域分析

2.5 细胞亚定位分析

3 进化分析

4 Spike 蛋白分析