

第 4 次作业

崔晏菲 自硕 21 2021210976

1. 使用动态规划对 AGWGHEE 和 AWHEA 进行匹配

使用 BLOSUM 矩阵作为打分矩阵。Gap Score 得分为-8。

a) Globule Alignment:

	-	A	W	H	E	A
-	0	-8	-16	-24	-32	-40
A	-8	4	-4	-12	-20	-28
G	-16	-4	2	-6	-14	-20
W	-24	-12	7	0	-8	-16
G	-32	-20	-1	5	-2	-8
H	-40	-28	-9	7	5	-3
E	-48	-36	-17	-1	12	4
E	-56	-44	-25	-9	4	11

匹配结果为 AGWGHEE 和 A-W-HEA，得分为 11。

b) Local Alignment:

	-	A	W	H	E	A
-	0	0	0	0	0	0
A	0	4	0	0	0	4
G	0	0	2	0	0	0
W	0	0	11	3	0	0
G	0	0	3	9	1	0
H	0	0	0	11	9	1
E	0	0	0	3	16	8
E	0	0	0	0	8	15

匹配结果为 WGHE 和 W-HE，得分为 15。

2. 使用 NCBI BLAST WEB 进行序列匹配

在 NCBI BLAST 网站上传 sequence1.fa 并选择相关参数进行比对，比对结果如下：

BLAST® » blastn suite » results for RID-R9EMHEN2013

Home Recent Results Saved Strategies Help

< Edit Search Save Search Search Summary How to read this report? BLAST Help Videos Back to Traditional Results Page

Your search is limited to records that include: Viruses (taxid:10239)

Job Title homework_sequence

RID R9EMHEN2013 Search expires on 11-17 17:25 pm Download All

Program BLASTN Citation

Database refseq_representative_genomes (2 databases) See details

Query ID lclQuery_20541

Description homework_sequence

Molecule type dna

Query Length 29751

Other reports Distance tree of results MSA viewer

Filter Results

Organism only top 20 will appear exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity E value Query Coverage

to to to

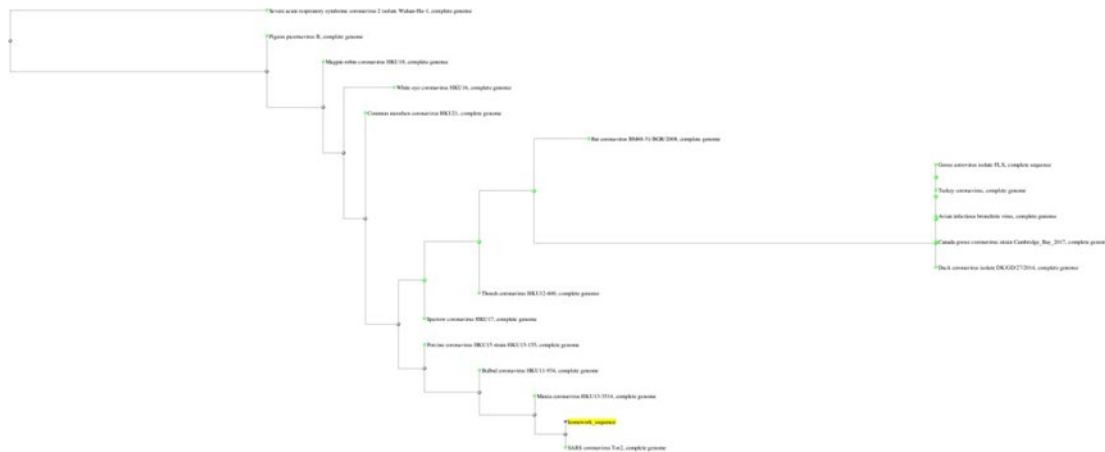
Filter Reset

Descriptions Graphic Summary Alignments Taxonomy

Reports Lineage Organism Taxonomy

23 sequences selected

Organism	Blast Name	Score	Number of Hits	Description
Pseudoviridae	viruses		23	
• Pseudoviridae	viruses		22	
• • Orthocoronavirinae	viruses		21	
• • • Betacoronavirus	viruses		6	
• • • • Sarbecovirus	viruses		3	
• • • • • Severe acute respiratory syndrome-related coronavirus	viruses		2	
• • • • • • SARS coronavirus Tor2	viruses	54940	1	SARS coronavirus Tor2 hits
• • • • • • Severe acute respiratory syndrome coronavirus 2	viruses	15175	1	Severe acute respiratory syndrome coronavirus 2
• • • • • • • Bat coronavirus BM48-31/BGR/2008	viruses	15389	1	Bat coronavirus BM48-31/BGR/2008
• • • • • • • Bat Ho-betacoronavirus/Zhejiang/2013	viruses	1507	1	Bat Ho-betacoronavirus/Zhejiang/2013
• • • • • • • Pterodroma bat coronavirus HKU5	viruses	398	1	Pterodroma bat coronavirus HKU5
• • • • • • • Rabbit coronavirus HKU14	viruses	263	1	Rabbit coronavirus HKU14 hits



a) Which species does the sequence belong to?

根据比对结果，比对分数最高的是 SARS coronavirus Tor2，因此该序列属于 SARS coronavirus Tor2。

b) List at least 3 kinds of species whose genome is similar with this sequence.

比对结果的距离树如上图所示，根据距离树，和该物种序列最接近的三个物种为：Munia coronavirus HKU13-3514, Bulbul coronavirus HKU11-934 和 Porcine coronavirus HKU15 strain HKU15-155。

c) How many protein-coding genes are there in this genome?

SARS coronavirus Tor2 的 Sequence ID 为 NC_004718.3，对应的网址为：
https://www.ncbi.nlm.nih.gov/nuccore/NC_004718.3/。

根据网页上的信息可知该序列共有 12 个蛋白编码基因，分别为：

ORF1ab(265..21485), S(21492..25259), ORF3a (25268..26092), ORF3b (25689..26153), E(26117..26347), M(26398..27063), ORF6 (27074..27265), ORF7a (27273..27641), ORF7b (27638..27772), ORF8a (27779..27898), ORF8b (27864..28118), N(28120..29388)。

d) What is the percentage of the non-coding sequence in this genome?

根据上一小问的基因进行计算编码区域的长度：

$$\begin{aligned}
 & (21485 - 265 + 1) + (25259 - 21492 + 1) + (26092 - 25268 + 1) \\
 & + (26347 - 25689 + 1) + (27063 - 26398 + 1) \\
 & + (27265 - 27074 + 1) + (27772 - 27273 + 1) \\
 & + (28118 - 27779 + 1) + (29388 - 28120 + 1) \\
 & = 29440bp
 \end{aligned}$$

非编码序列的比例为

$$\frac{29751 - 29440}{29751} = 1.045\%$$

3. 使用 blast 和 bowtie 进行序列匹配

a) 使用 Blast 进行序列匹配

命令行如下：

```
D:\硕士课程\计算分子生物学引论\作业\第4次作业_崔晏菲_2021210976\ncbi-blast-2.13.0+-x64-win64\ncbi-blast-2.13.0+\bin>blastn -query ../../sequence2.fa -db ../../ref/blast/hg19 -out ../../seq2fa_blast.txt

D:\硕士课程\计算分子生物学引论\作业\第4次作业_崔晏菲_2021210976\ncbi-blast-2.13.0+-x64-win64\ncbi-blast-2.13.0+\bin>blastn -query ../../seq2fastq.fa -db ../../ref/blast/hg19 -out ../../seq2fastq_blast.txt
```

其中 sequence2.fastq 在比对之前，先使用了 fastq2fa.py 脚本进行转换，转换为了 .fa 类型的文件。

在命令行参数中，-query 后是待比对序列的文件位置，-db 后是 genome reference 的文件位置，-out 后是输出文件的位置。

匹配结果分别见 seq2fa_blast.txt 和 seq2fastq_blast.txt 文件。

Blast 对比结果如下图所示，该序列和 chrM 以及 chr5 最相似。

BLASTN 2.13.0+

Reference: Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb Miller (2000), "A greedy algorithm for aligning DNA sequences", J Comput Biol 2000; 7(1-2):203-14.

Database: /data/database/UCSC/hg19/genome/hg19.fa
25 sequences; 3,095,693,983 total letters

Query= SRR891268.1 HWI-ST281:266:C1LTTACXX:1:1101:1238:1978 length=50

Length=50

	Score	E	(Bits)	Value
Sequences producing significant alignments:				
chrM	91.6	2e-17		
chr5	91.6	2e-17		

b) 使用 Bowtie 进行序列匹配

命令行如下：

```
D:\硕士课程\计算分子生物学引论\作业\第4次作业_崔晏菲_2021210976\bowtie2-2.1.0-mingw-win64\bowtie2-2.1.0>bowtie2-align.exe --local -x ../../ref/bowtie2/hg19 -f ../../sequence2.fa -S ../../seq2fa_bowtie.txt
300 reads; of these:
  300 (100.00%) were unpaired; of these:
    3 (1.00%) aligned 0 times
   108 (36.00%) aligned exactly 1 time
   189 (63.00%) aligned >1 times
99.00% overall alignment rate

D:\硕士课程\计算分子生物学引论\作业\第4次作业_崔晏菲_2021210976\bowtie2-2.1.0-mingw-win64\bowtie2-2.1.0>bowtie2-align.exe --local -x ../../ref/bowtie2/hg19 -fq ../../sequence2.fastq -S ../../seq2fastq_bowtie.txt
3000 reads; of these:
  3000 (100.00%) were unpaired; of these:
    40 (1.33%) aligned 0 times
   973 (32.43%) aligned exactly 1 time
  1987 (66.23%) aligned >1 times
98.67% overall alignment rate
```

--local 表示在本地运行；-x 后为 genome reference 的文件位置；-f 表示待比对序列文件格式为(multi-)FASTA 文件，文件后缀为 .fa/.mfa，-q 表示待比对序列为 FASTQ 格式文件，文件后缀为 .fq/.fastq，-f/-q 后为待比对序

列的文件位置。-S 后为比对结果的存储位置。

对于 `sequence2.fa` 序列，300 个 reads 中有 189 个完成超过一次匹配，有 108 个完成 1 次匹配，有 3 个没有匹配，overall alignment rate 为 99%。

对于 `sequence2.fastq` 序列，3000 个 reads 中有 1987 个完成超过一次匹配，有 973 个完成 1 次匹配，有 40 个没有匹配，overall alignment rate 为 98.67%。

匹配结果分别见 `sequence2fa.txt` 和 `sequence2fastq.txt` 文件。