

## Problem 1A

Since there are a lot of results for the total 15 readsets, I just paste some of them here:

```

the score for # 6read is :97
*****results for # 7 read:*****
GGACACTTCGCATGGTGGACAGCCTTTGTTACTAATGTGAATGCGTCATC
|||||
GGACACTTCGCATGGTGGACAGCCTTTGTTACTAATGTGAATGCGTCATC

the score for # 7read is :100
*****results for # 8 read:*****
GATGTAATTATCTTGGCAAACACGCGAACAAATAGATGTTATGTCATG
|||||||x|||||||x|||||||x|||
GATGTAATTATCTTGGCTAACACGCGAACAAATTGATGTTATGTTATG

the score for # 8read is :91
*****results for # 9 read:*****
GAGGAATACAAATCCAATTCAGTTGTCTTCCTATTCTTTATTTGACATGA
|||||||x|||||||x|||||x|||||||x|||||||
GAGGAATACCAATCCAATTCAGTTGTCTCCCTATTCTTTCTTTGACATGA

the score for # 9read is :88
*****results for # 10 read:*****
AGGGGTACTGCTGTTATGTCTTTAAAAGAAGGTCAAATCAATGATATGAT
||||||| | |||||
AGGGGTACTGCTGTTATGTCTTTA__GAAGGTCAAATCAATGATATGAT

AGGGGTACTGCTGTTATGTCTTTAAAAGAAGGTCAAATCAATGATATGAT
||||||| | |||||
AGGGGTACTGCTGTTATGTCTTT_A__GAAGGTCAAATCAATGATATGAT

AGGGGTACTGCTGTTATGTCTTTAAAAGAAGGTCAAATCAATGATATGAT
||||||| | |||||
AGGGGTACTGCTGTTATGTCTTT__A__GAAGGTCAAATCAATGATATGAT

AGGGGTACTGCTGTTATGTCTTTAAAAGAAGGTCAAATCAATGATATGAT
||||||| | |||||
AGGGGTACTGCTGTTATGTCTTT__AGAAGGTCAAATCAATGATATGAT

the score for # 10read is :85
*****results for # 11 read:*****
GTAGACTTATAATTAGAGAAAACAACAGAGTTGTTATTTCTAGTGATGTT
|||||||
GTAGACTTATAATTAGAGAAAACAACAGAGTTGTTATTTCTAGTGATGTT

the score for # 11read is :100
*****results for # 12 read:*****
CAATGTTTGTCTTCTTG__TTTTATTGCCACTAGTCTCTAGTCAGTGTG
||||||| |||||
CAATGTTTGTCTTCTTGCCCTTTTATTGCCACTAGTCTCTAGTCAGTGTG

the score for # 12read is :90
*****results for # 13 read:*****
ATTACCCCTGCATACACTA__ATTCTTTCACACGTGGTGTGTTTATTACC
||||||| |||||
ATTACCCCTGCATACACTAGGGATTCTTTCACACGTGGTGTGTTTATTACC

the score for # 13read is :85
*****results for # 14 read:*****
GTTTTACATTCAACTCAGGACTTGTTCTTACCTTTCTTTTCCAATGTTAC
|||||||
GTTTTACATTCAACTCAGGACTTGTTCTTACCTTTCTTTTCCAATGTTAC

the score for # 14read is :100

```

All the results can be found in the txt file "output\_1A.txt".

## Problem 1B

```
[yg336@wind ~/inf503/Homework4]$ tail output_1B_1000.txt
TAGGA_GAGTGTGGACACTTATG_AATGTCT_TGACACTCGTTTATAAAAGT
|x||| |xxx||xx|x||| ||| |xx||x| |x|||||||x|xx|x|||
TCGGACGCACGTCAAGACT_ATGTAGGGTATGTAACACTCGCTCGTCAAGT

TAGGA_GAGTGTGGACACTTATG_AATGTCT_TGACACTCGTTTATAAAAGT
|x||| |xxx||xx|x||| ||| |xx||x| |x|||||||x|xx|x|||
TCGGACGCACGTCAAGAC_TATGTAGGGTATGTAACACTCGCTCGTCAAGT

the score for # 999read is :37
total time for 1000random sequences is 55s
[yg336@wind ~/inf503/Homework4]$ tail output_1B_10000.txt
TGTTCCTTGCTCG_CAAACATACAACGTGTTGTAGCTTGTACACCCGTTTCTAT
||| x||| ||| |x||| x||||x| ||x|x|xx|||| x| |x||||x|||
TGT_ATTG_TCGCCGAAC_CACAAAG_GTCGGACTTTGT_GC_CGGTTTGTAT

TGTTCCTTGCTCG_CAAACATACAACGTGTTGTAGCTTGTACACCCGTTTCTAT
|| |x||| ||| |x||| x||||x| ||x|x|xx|||| x| |x||||x|||
TG_TATTG_TCGCCGAAC_CACAAAG_GTCGGACTTTGT_GC_CGGTTTGTAT

the score for # 9999read is :38
total time for 10000random sequences is 547s
[yg336@wind ~/inf503/Homework4]$ tail output_1B_100000.txt
CTCAGGTTTTGCTGCAT_ACAGTCGC_TACAGGATTGGCAACTATAAATTAA
||||x |x||||x|||| ||x||x|| x|x||||x|x|xxx||| |||||||
CTCAT_TCTTGCGGCATGACCGTTGCGGAAAGGTTGGGTCCCTA_AAATTAA

CTCAGGTTTTGCTGCAT_ACAGTCGC_TACAGGATTGGCAACTATAAATTAA
|||| x|x||||x|||| ||x||x|| x|x||||x|x|xxx||| |||||||
CTCA_TTCTTGCGGCATGACCGTTGCGGAAAGGTTGGGTCCCTA_AAATTAA

the score for # 99999read is :48
total time for 100000random sequences is 5543s
[yg336@wind ~/inf503/Homework4]$ tail output_1B_1000000.txt
GCTACTGTAGTAATTGGAACAAGCAAATTCTATG_GTGGTTGGCACAA
|x|x||x||x||||x||| xx||x||||x|xx|| x|x||x||||x|x|
GTTGCTCTATTAATCGGA_GGAGAAAATCCGGTGCCTCGTCGGCGCTA

GCTACTGTAGTAATTGGAACAAGCAAATTCTATG_GTGGTTGGCACAA
|x|x||x||x||||x|| |xx||x||||x|xx|| x|x||x||||x|x|
GTTGCTCTATTAATCGG_AGGAGAAAATCCGGTGCCTCGTCGGCGCTA

the score for # 999999read is :38
total time for 1000000random sequences is 53983s
```

## Problem 2A

Since there are a lot of results for the total 15 readsets, I just paste some of them here:

```

*****results for the # 2 read:*****

*****results for the # 3 read:*****

ATGCGTTAGCTTACTACACACAAACAAAGGGAGGTAGGT
|x|| |x|||x|| | |||||xx|x|xx|x|x|
AGGC_TAAGCTAAC_CAACACACACAGAATCTCGCT

CTAACTTTAGAGTC_CAACCAACAGAATCT
|||xx|x|xxx| ||| |||||
CTAAGCTAACCAACACACAC_AACAGAATCT

CTAACTTTAG_AGTCCAACCAACAGAATCT
|||xx|x|x|xxx||| |||||
CTAAGCTAACCAACACACAC_AACAGAATCT

CTAACTTTA_GAGTCCAACCAACAGAATCT
|||xx|x|x|xxx||| |||||
CTAAGCTAACCAACACACAC_AACAGAATCT

CTAA_CTTTAGAGTCCAACCAACAGAATCT
||| ||xxx|xxx||| |||||
CTAAGCTAACCAACACACAC_AACAGAATCT

CTAACTTTAGAGTC_CAACCAACAGAATCT
|||xx|x|xxx| ||| |||||
CTAAGCTAACCAACACAA_CAACAGAATCT

CTAACTTTAG_AGTCCAACCAACAGAATCT
|||xx|x|x|xxx||| |||||
CTAAGCTAACCAACACAA_CAACAGAATCT

CTAACTTTA_GAGTCCAACCAACAGAATCT
|||xx|x|x|xxx||| |||||
CTAAGCTAACCAACACAA_CAACAGAATCT

CTAA_CTTTAGAGTCCAACCAACAGAATCT
||| ||xxx|xxx||| |||||
CTAAGCTAACCAACACAA_CAACAGAATCT

*****results for the # 4 read:*****

AACATGGCAAGGAAGACCTTAAAT
||||||| |x|x| |||
AACATGGCAAGGAA_ATCA_AAAT

AACATGGCAAGGAAGACCTTAAAT
||||||| |x| x|||
AACATGGCAAGGAA_ATC_AAAAT

```

All the results can be found in the txt file "output\_2A.txt".

## Problem 2B

The BLAST method is faster than Smith-Waterman algorithm.



```
[yg336@wind ~/inf503/Homework4 ]$ tail output_2B_1000.txt

TTTCTTGGCACTGATAACACTCGCTACT
|x|x|x|xxx| | | | | | | | | |xx|
TATGTAGGGTATG_TAACACTCGCTCGT

TTTCTTGGCACTGATAACACTCGCTACTT
|x|x|x|xxx| | | | | | | | | |x|
TATGTAGGGTATG_TAACACTCGCT_CGT

total time for 1000random sequences is 0s
[yg336@wind ~/inf503/Homework4 ]$ tail output_2B_10000.txt
|x| |x|x|x|x x| | | | | | | | | | |x| | | |x|x|x|x|
CGAG_TGCACCAAGTGACCAAGCAAGAAAGAGACAATTTCTAGAGGGT

CCAGATCCATCA_AAACCAAGCAAG__AG_GTC_ATTTATTGAAGAT
|x| |x|x|x|xx| | | | | | | | | | |x| | | |x|x|x|x|
CGAG_TGCACCAAGTGACCAAGCAAGAAAGAGACAATTTCTAGAGGGT

*****results for the # 9999 read:*****

total time for 10000random sequences is 0s
[yg336@wind ~/inf503/Homework4 ]$ tail output_2B_100000.txt

CTACAGTTTCTGTTTC_TTCACCTGATG_CTGTTAC_AGCGTATAAT_GG
| | | | |xxx|xxx| | | | | | | | | | |x| |x| x|x| | |
CTA_AGT_CAAGACCCTTTCACCTGATGCCTG_GACGAAC_CACAATCGG

CTACAGTTTCTGTTTC_TTCACCTGATG_CTGTTAC_AGCGTATAAT_GG
| | | | |xxx|xxx| | | | | | | | | | |x| |x| x|x| | |
CTA_AG_TCAAGACCCTTTCACCTGATGCCTG_GACGAAC_CACAATCGG

total time for 100000random sequences is 8s
[yg336@wind ~/inf503/Homework4 ]$ tail output_2B_1000000.txt

*****results for the # 999998 read:*****
*****results for the # 999999 read:*****

CAAAGAATACTGT_TAAGAGTGTG
| | |x|x| | | | | | | | | | |
CAAGGGAT_C_GTATAAGAGTGTG

total time for 1000000random sequences is 79s
```

## Problem 2C

```
yg336@wind:~/inf503/Homework4
The number of arguments passed: 5
The first argument is: main
The second argument is: problem2C
The third argument is: /common/contrib/classroom/inf503/SARS_COV2.txt
The fourth argument is: /home/yg336/inf503/Homework4/readsets.txt
The fifth argument is: 1000
random select 100,000 fragments from SARS-COV2 genome:
fragment found in SARS-COV2: 99998
random select 100,000 fragments from SARS-COV2 genome with 5% mutation:
fragment found in SARS-COV2: 8907
```

**Note:** All the fragments should be found in the SARS-COV2 genome if they are randomly selected with no mutation. The mismatch here is because of the boundary condition.