T.C.

MARMARA UNIVERSITY

FACULTY of ENGINEERING

COMPUTER ENGINEERING DEPARTMENT

CSE4065 – Introduction to Computational Genomics

Assignment I Report

Prepared by

150114032 – Şükrü Gümüştaş

Given by

Betül Demiröz Boz, PhD

Assistant Professor

1. Project Structure

Python programming language is used to perform all operations in the project. Python’s Regular Expression library “re” and CSV library “csv” are imported. Code sections are explained below:

```
def get_and_validate_input(filename):
    file = open("./res/" + filename, "r")
    lines = file.read()
    file.close()
    lines = lines.upper()
    lines = re.sub(r'[\s]', "", lines)
    if not re.match(r'[AGCT]{5,}', lines):
        raise ValueError("[" + filename + "] Input is invalid! [Must only contain A, G, C,
                         T and be longer than 5!]")
    return lines
```

This function aims to get input from a file. It takes the filename as an argument. If the file doesn’t exist under the “res” folder it gives a standard error. There is not an extra control for that. If the file exists, it should contain all the sequence from a sample. Since it is stated that at least 5-mers should be investigated in the project description, if the file content is empty or has the length less than 5, the program raises an error and terminates. If the file contains other characters than nucleobases it also raises error and terminates. Otherwise the content of the file is returned as String.

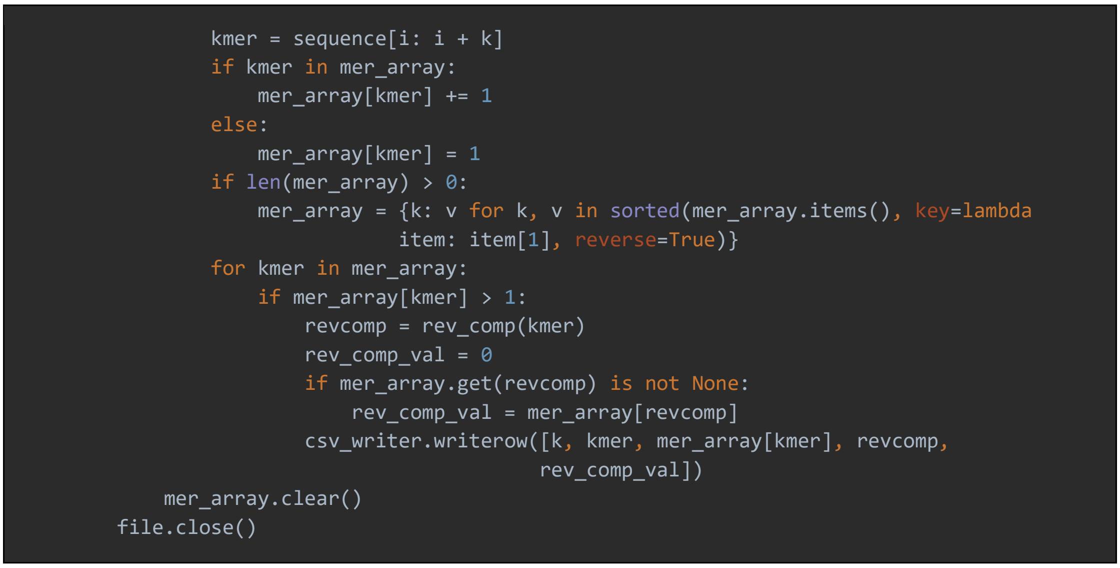
```
complements = str.maketrans("ACTG", "TGAC")
```

This variable is declared globally to get the complements of each nucleobase by using the following function:

```
def rev_comp(sub_sequence):
    return sub_sequence.translate(complements)[::-1]
```

The operation in this function is a standard Python operation to get a sequence’s reverse complement.

```
def main():
    sequences = {"NEPAL": get_and_validate_input("nepal_COVID19.txt"),
                "WUHAN": get_and_validate_input("wuhan_COVID19.txt")}
    mer_array = {}
    for origin in sequences:
        sequence = sequences[origin]
        len_seq = len(sequence)
        print('Origin [%s] -> Output file: %s_output.csv' % (origin, origin))
        file = open('%s_output.csv' % origin, 'w')
        csv_writer = csv.writer(file)
        csv_writer.writerow(['Length', 'K-Mer', 'Count', 'Revcomp', 'Count'])
        for k in range(5, len_seq):
            for i in range(len_seq - k + 1):
```



```
                kmer = sequence[i: i + k]
                if kmer in mer_array:
                    mer_array[kmer] += 1
                else:
                    mer_array[kmer] = 1
                if len(mer_array) > 0:
                    mer_array = {k: v for k, v in sorted(mer_array.items(), key=lambda
                                item: item[1], reverse=True)}
                for kmer in mer_array:
                    if mer_array[kmer] > 1:
                        revcomp = rev_comp(kmer)
                        rev_comp_val = 0
                        if mer_array.get(revcomp) is not None:
                            rev_comp_val = mer_array[revcomp]
                        csv_writer.writerow([k, kmer, mer_array[kmer], revcomp,
                                            rev_comp_val])
            mer_array.clear()
        file.close()
```

In the main function, a dictionary is declared with origins of samples as key and file contents as values.

Another dictionary "mer_array" is declared to hold the k-mers and their frequency.

For each file content, starting from 5-mers, each k-mers frequency is added to mer_array. Then this mer_array is sorted according to their values in descending order. Finally all resulst for each file stored in csv files.

For some memory issues, each k-mer is written to file with its reverse complement in its section and mer_array is cleared for the next iteration.
Since results are too long, they are stored in csv files and will not be shown in this document.

2. Comparison of 2 Samples

There are distinct k-mers which appear in sequence from the patient form Nepal. These are:

K-mer	Frequency	RevComp	Frequency
TAAAGTGA	2	TCACTTTA	0
TCAATAAA	2	TCACTTTA	3
TTCAATAA	2	TTATTGAA	3
TTCAATAAA	2	TTTATTGAA	2
TTTCAATAA	2	TTATTGAAA	2
TTTCAATAAA	2	TTTATTGAAA	1

There are distinct k-mers which appear in sequence from the patient form Wuhan. These are:

K-mer	Frequency	RevComp	Frequency
AAAAAAA	6	TTTTTTT	2
AAAAAAAA	5	TTTTTTTT	1
AAAAAAAAA	4	TTTTTTTTT	0
AAAAAAAAAA	3	TTTTTTTTTT	0
AAAAAAAAAAA	2	TTTTTTTTTTT	0
AAAGGTTTA	2	TAAACCTTT	1
AACAAAGTG	2	CACTTTGTT	1
AAGGTTTAT	2	ATAAACCTT	0
AATAGCT	2	AGCTATT	3
AATGACAAA	2	TTTGTCATT	0
ACAAAGTG	2	CACTTTGT	2
ACCTTCC	2	GGAAGGT	1
AGCTTCT	2	AGAAGCT	11
AGGTTTAT	2	ATAAACCT	2
AGTGCTATC	2	GATAGCACT	1
ATACCTTC	2	GAAGGTAT	0
ATAGCTT	2	AAGCTAT	4
ATTTTAATA	2	TATTAAAAT	2
CAAAGTGAC	2	GTCACTTTG	0
CATGTGAT	2	ATCACATG	0
CCATGTGA	2	TCACATGG	0
CCCATG	2	CATGGG	3
CCTTCCC	2	GGGAAGG	0
CTTAGGAG	2	CTCCTAAG	0
GCTTCTT	2	AAGAAGC	6
GGTTTATAC	2	GTATAAACC	1
GTGCTATC	2	GATAGCAC	1
GTTTATAC	2	GTATAAAC	3
TAAAGGTTT	2	AAACCTTTA	0
TAAAGGTTTA	2	TAAACCTTTA	0
TAATAGC	2	GCTATTA	5
TACCTTCC	2	GGAAGGTA	0
TAGTAGT	2	ACTACTA	7
TCAACAAA	2	TTTGTTGA	4
TCCCC	2	GGGGA	4
TCTTAGGA	2	TCCTAAGA	1
TTAAAGGT	2	ACCTTTAA	1
TTAAAGGTT	2	AACCTTTAA	0
TTAAAGGTTT	2	AAACCTTTAA	0
TTAAAGGTTTA	2	TAAACCTTTAA	0
TTAATAGC	2	GCTATTAA	1
TTAGGAG	2	CTCCTAA	0
TTAGTAG	2	CTACTAA	10
TTCAACAA	2	TTGTTGAA	3
TTCTTAGGA	2	TCCTAAGAA	0
TTTCAACA	2	TGTTGAAA	1
TTTTAATA	2	TATTAAAA	2

There are common values with both different k-mer frequencies and different reverse complement frequencies in each file. These values are:

K-mer	Freq. In Nepal	Freq. In Wuhan	Revcomp	Freq. in Nepal	Freq. in Wuhan
AAGGT	45	46	ACCTT	52	53
ACCTT	52	53	AAGGT	45	46
ATAAA	58	57	TTTAT	74	75
ATAGC	20	21	GCTAT	44	45
GCTAT	44	45	ATAGC	20	21
TTAAA	94	95	TTTAA	90	91
TTTAA	90	91	TTAAA	94	95
TTTAT	74	75	ATAAA	58	57

There are common values with only different k-mer frequencies. These values are:

K-mer	Freq. in Nepal	Freq. in Wuhan	Revcomp	Freq. in Nepal	Freq. in Wuhan
AAAAA	56	64	TTTTT	61	61
AAAAAA	2	9	TTTTTT	6	6
AAAGG	44	45	CCTTT	44	44
AAAGGT	17	18	ACCTTT	22	22
AAAGGTT	7	8	AACCTTT	2	2
AAAGGTTT	2	3	AAACCTTT	2	2
AACAA	98	99	TTGTT	102	102
AACAAA	29	30	TTTGTT	28	28
AACAAAG	10	11	CTTTGTT	6	6
AACAAAGT	3	4	ACTTTGTT	1	1
AAGGTT	18	19	AACCTT	11	11
AAGGTTT	5	6	AAACCTT	5	5
AAGGTTTA	3	4	TAAACCTT	2	2
AATAA	43	42	TTATT	72	72
AATAAA	16	15	TTTATT	27	27
AATAAAG	4	3	CTTTATT	8	8
AATAG	32	33	CTATT	66	66
AATAGC	6	7	GCTATT	15	15
AATGA	49	50	TCATT	37	37
AATGAC	9	10	GTCATT	8	8
AATGACA	4	5	TGTCATT	3	3
AATGACAA	2	3	TTGTCATT	1	1
ACAAA	87	89	TTTGT	88	88
ACAAAA	19	20	TTTTGT	32	32
ACAAAAA	4	5	TTTTTGT	9	9
ACAAAG	23	24	CTTTGT	15	15
ACAAAGT	6	7	ACTTTGT	7	7
ACCTTC	11	12	GAAGGT	16	16
AGAAT	48	49	ATTCT	58	58
AGAATG	12	13	CATTCT	15	15
AGAATGA	2	3	TCATTCT	2	2
AGCTT	44	45	AAGCT	51	51
AGCTTC	4	5	GAAGCT	14	14
AGGAG	24	25	CTCCT	8	8
AGGAGA	6	7	TCTCCT	3	3
AGGTT	50	51	AACCT	39	39
AGGTTT	19	20	AAACCT	12	12
AGGTTTA	5	6	TAAACCT	5	5
AGTAG	23	24	CTACT	59	59
AGTAGT	5	6	ACTACT	16	16
AGTAGTG	2	3	CACTACT	4	4
AGTGC	39	40	GCACT	36	36
AGTGCT	16	17	AGCACT	6	6
AGTGCTA	4	5	TAGCACT	4	4
AGTGCTAT	3	4	ATAGCACT	2	2
ATAAAG	15	14	CTTTAT	19	19
ATAAAGT	4	3	ACTTTAT	7	7
ATACC	24	25	GGTAT	24	24
ATACCT	11	12	AGGTAT	7	7
ATACCTT	4	5	AAGGTAT	1	1
ATAGCT	6	7	AGCTAT	11	11
ATCCC	2	3	GGGAT	7	7
ATGAC	38	39	GTCAT	28	28
ATGACA	15	16	TGTCAT	14	14
ATGACAA	7	8	TTGTCAT	4	4
ATGACAAA	3	4	TTTGTCAT	1	1

ATGTG	48	49	CACAT	36	36
ATGTGA	13	14	TCACAT	12	12
ATGTGAT	4	5	ATCACAT	4	4
ATTTT	76	77	AAAAT	70	70
ATTTTA	25	26	TAAAAT	22	22
ATTTTAA	8	9	TTAAAAT	10	10
ATTTTAAT	2	3	ATTAAAAT	3	3
CAAAA	60	61	TTTTG	97	97
CAAAAA	14	15	TTTTTG	22	22
CAAAG	45	46	CTTTG	60	60
CAAAGT	12	13	ACTTTG	18	18
CAAAGTG	5	6	CACTTTG	4	4
CAAAGTGA	2	3	TCACTTTG	1	1
CAACA	74	75	TGTTG	97	97
CAACAA	28	29	TTGTTG	22	22
CAACAAA	7	8	TTTGTTG	6	6
CAACAAAG	3	4	CTTTGTTG	0	0
CAATA	29	28	TATTG	60	60
CAATAA	7	6	TTATTG	17	17
CAATAAA	4	3	TTTATTG	7	7
CATGT	40	41	ACATG	50	50
CATGTG	11	12	CACATG	10	10
CATGTGA	4	5	TCACATG	2	2
CCATG	20	21	CATGG	27	27
CCATGT	7	8	ACATGG	10	10
CCATGTG	3	4	CACATGG	4	4
CCCAT	7	8	ATGGG	19	19
CCCCA	5	6	TGGGG	7	7
CCTTC	21	22	GAAGG	25	25
CCTTCC	3	4	GGAAGG	2	2
CTATC	18	19	GATAG	13	13
CTATCC	2	3	GGATAG	0	0
CTTAG	28	29	CTAAG	21	21
CTTAGG	6	7	CCTAAG	6	6
CTTAGGA	2	3	TCCTAAG	1	1
CTTCT	56	57	AGAAG	63	63
CTTCTT	25	26	AAGAAG	21	21
CTTCTTA	5	6	TAAGAAG	1	1
CTTCTTAG	2	3	CTAAGAAG	1	1
GAATG	29	30	CATTC	33	33
GAATGA	3	4	TCATTC	8	8
GACAA	52	53	TTGTC	47	47
GACAAA	15	16	TTTGTC	12	12
GACAAAA	2	3	TTTTGTC	2	2
GACAAAAA	2	3	TTTTTGTC	0	0
GAGAA	31	32	TTCTC	21	21
GAGAAT	6	7	ATTCTC	6	6
GATTT	48	49	AAATC	37	37
GATTTT	17	18	AAAATC	12	12
GATTTTA	5	6	TAAAATC	2	2
GCTATC	5	6	GATAGC	1	1
GCTTC	26	27	GAAGC	24	24
GCTTCT	7	8	AGAAGC	14	14
GGAGA	18	19	TCTCC	9	9
GGAGAA	5	6	TTCTCC	4	4
GGTTT	54	55	AAACC	42	42
GGTTTA	16	17	TAAACC	15	15
GGTTTAT	4	5	ATAAACC	4	4
GGTTTATA	2	3	TATAAACC	1	1
GTAGT	44	45	ACTAC	38	38
GTAGTG	16	17	CACTAC	9	9
GTAGTGC	5	6	GCACTAC	3	3
GTAGTGCT	2	3	AGCACTAC	0	0

GTGAT	52	53	ATCAC	27	27
GTGATT	10	11	AATCAC	8	8
GTGATTT	2	3	AAATCAC	4	4
GTGCT	53	54	AGCAC	22	22
GTGCTA	15	16	TAGCAC	8	8
GTGCTAT	5	6	ATAGCAC	3	3
GTTTA	68	69	TAAAC	56	56
GTTTAT	23	24	ATAAAC	13	13
GTTTATA	5	6	TATAAAC	6	6
TAAAGG	12	13	CCTTTA	7	7
TAAAGGT	4	5	ACCTTTA	4	4
TAAAGGTT	2	3	AACCTTTA	0	0
TAAAGT	17	16	ACTTTA	25	25
TAAAGTG	4	3	CACTTTA	0	0
TAATA	44	45	TATTA	68	68
TAATAG	11	12	CTATTA	22	22
TACCT	35	36	AGGTA	28	28
TACCTT	15	16	AAGGTA	11	11
TACCTTC	4	5	GAAGGTA	4	4
TAGCT	39	40	AGCTA	40	40
TAGCTT	8	9	AAGCTA	12	12
TAGGA	16	17	TCCTA	28	28
TAGGAG	5	6	CTCCTA	2	2
TAGGAGA	3	4	TCTCCTA	1	1
TAGTA	23	24	TACTA	66	66
TAGTAG	4	5	CTACTA	25	25
TAGTG	51	52	CACTA	45	45
TAGTGC	16	17	GCACTA	9	9
TAGTGCT	5	6	AGCACTA	2	2
TAGTGCTA	3	4	TAGCACTA	0	0
TAGTGCTAT	2	3	ATAGCACTA	0	0
TATAC	28	29	GTATA	25	25
TATACC	5	6	GGTATA	5	5
TATACCT	4	5	AGGTATA	2	2
TATACCTT	2	3	AAGGTATA	0	0
TATCC	8	9	GGATA	8	8
TCAAC	59	60	GTTGA	62	62
TCAACA	20	21	TGTTGA	29	29
TCAACAA	9	10	TTGTTGA	10	10
TCAAT	43	42	ATTGA	36	36
TCAATA	11	10	TATTGA	15	15
TCAATAA	3	2	TTATTGA	5	5
TCTTA	57	58	TAAGA	34	34
TCTTAG	10	11	CTAAGA	7	7
TCTTAGG	3	4	CCTAAGA	3	3
TGACA	64	65	TGTCA	48	48
TGACAA	16	17	TTGTCA	17	17
TGACAAA	7	8	TTTGTCA	4	4
TGATT	56	57	AATCA	43	43
TGATTT	17	18	AAATCA	20	20
TGATTTT	6	7	AAAATCA	5	5
TGATTTTA	2	3	TAAAATCA	0	0
TGCTA	65	66	TAGCA	24	24
TGCTAT	20	21	ATAGCA	10	10
TGCTATC	3	4	GATAGCA	1	1
TGTGA	45	46	TCACA	39	39
TGTGAT	18	19	ATCACA	8	8
TTAAAG	20	21	CTTTAA	25	25
TTAAAGG	2	3	CCTTTAA	2	2
TTAAT	73	74	ATTAA	59	59
TTAATA	15	16	TATTAA	17	17
TTAATAG	5	6	CTATTAA	9	9
TTAGG	16	17	CCTAA	37	37

TTAGGA	3	4	TCCTAA	9	9
TTAGTA	8	9	TACTAA	26	26
TTATA	48	49	TATAA	59	59
TTATAC	10	11	GTATAA	10	10
TTATACC	3	4	GGTATAA	1	1
TTATACCT	2	3	AGGTATAA	0	0
TTCAAC	25	26	GTTGAA	23	23
TTCAACA	4	5	TGTTGAA	9	9
TTCAAT	13	12	ATTGAA	10	10
TTCAATA	3	2	TATTGAA	5	5
TTCTT	96	97	AAGAA	82	82
TTCTTA	27	28	TAAGAA	14	14
TTCTTAG	6	7	CTAAGAA	2	2
TTCTTAGG	2	3	CCTAAGAA	0	0
TTTAAT	27	28	ATTAAA	25	25
TTTAATA	6	7	TATTAAA	5	5
TTTAATAG	2	3	CTATTAAA	2	2
TTTAGTA	3	4	TACTAAA	7	7
TTTATA	13	14	TATAAA	20	20
TTTATAC	2	3	GTATAAA	5	5
TTTCAAC	7	8	GTTGAAA	6	6
TTTCAAT	6	5	ATTGAAA	2	2
TTTCAATA	3	2	TATTGAAA	2	2
TTTTA	89	90	TAAAA	79	79
TTTTAA	34	35	TTAAAA	33	33
TTTTAAT	8	9	ATTAAAA	8	8

There are common values with only different reverse complement frequencies. These values are:

K-mer	Freq. in Nepal	Freq. in Wuhan	Revcomp	Freq. in Nepal	Freq. in Wuhan
AAAAT	70	70	ATTTT	76	77
AAAATC	12	12	GATTTT	17	18
AAAATCA	5	5	TGATTTT	6	7
AAACC	42	42	GGTTT	54	55
AAACCT	12	12	AGGTTT	19	20
AAACCTT	5	5	AAGGTTT	5	6
AAACCTTT	2	2	AAAGGTTT	2	3
AAATC	37	37	GATTT	48	49
AAATCA	20	20	TGATTT	17	18
AAATCAC	4	4	GTGATTT	2	3
AAATCACA	2	2	TGTGATTT	0	1
AACCT	39	39	AGGTT	50	51
AACCTT	11	11	AAGGTT	18	19
AACCTTT	2	2	AAAGGTT	7	8
AAGAA	82	82	TTCTT	96	97
AAGAAG	21	21	CTTCTT	25	26
AAGAAGC	6	6	GCTTCTT	1	2
AAGAAGCT	5	5	AGCTTCTT	0	1
AAGAAGCTA	2	2	TAGCTTCTT	0	1
AAGCT	51	51	AGCTT	44	45
AAGCTA	12	12	TAGCTT	8	9
AAGCTAT	4	4	ATAGCTT	1	2
AAGGTA	11	11	TACCTT	15	16
AATCA	43	43	TGATT	56	57
AATCAC	8	8	GTGATT	10	11
AATCACA	2	2	TGTGATT	0	1
ACATG	50	50	CATGT	40	41
ACATGG	10	10	CCATGT	7	8
ACCTTT	22	22	AAAGGT	17	18

ACCTTTA	4	4	TAAAGGT	4	5
ACTAC	38	38	GTAGT	44	45
ACTACT	16	16	AGTAGT	5	6
ACTACTA	7	7	TAGTAGT	1	2
ACTTTA	25	25	TAAAGT	17	16
ACTTTAT	7	7	ATAAAGT	4	3
ACTTTATT	4	4	AATAAAGT	1	0
ACTTTATTG	2	2	CAATAAAGT	1	0
ACTTTG	18	18	CAAAGT	12	13
ACTTTGT	7	7	ACAAAGT	6	7
AGAAG	63	63	CTTCT	56	57
AGAAGC	14	14	GCTTCT	7	8
AGAAGCT	11	11	AGCTTCT	1	2
AGAAGCTA	2	2	TAGCTTCT	0	1
AGCAC	22	22	GTGCT	53	54
AGCACT	6	6	AGTGCT	16	17
AGCACTA	2	2	TAGTGCT	5	6
AGCTA	40	40	TAGCT	39	40
AGCTAT	11	11	ATAGCT	6	7
AGCTATT	3	3	AATAGCT	1	2
AGCTATTA	2	2	TAATAGCT	0	1
AGGTA	28	28	TACCT	35	36
AGGTAT	7	7	ATACCT	11	12
AGGTATA	2	2	TATACCT	4	5
ATAAAC	13	13	GTTTAT	23	24
ATAAACC	4	4	GGTTTAT	4	5
ATAAACCT	2	2	AGGTTTAT	1	2
ATAGCA	10	10	TGCTAT	20	21
ATAGCAC	3	3	GTGCTAT	5	6
ATAGCACT	2	2	AGTGCTAT	3	4
ATCAC	27	27	GTGAT	52	53
ATCACA	8	8	TGTGAT	18	19
ATCACAT	4	4	ATGTGAT	4	5
ATGGG	19	19	CCCAT	7	8
ATGGGG	3	3	CCCCAT	0	1
ATTAA	59	59	TTAAT	73	74
ATTAAA	25	25	TTTAAT	27	28
ATTAAAA	8	8	TTTTAAT	8	9
ATTAAAAT	3	3	ATTTTAAT	2	3
ATTCT	58	58	AGAAT	48	49
ATTCTC	6	6	GAGAAT	6	7
ATTGA	36	36	TCAAT	43	42
ATTGAA	10	10	TTCAAT	13	12
ATTGAAA	2	2	TTTCAAT	6	5
CACAT	36	36	ATGTG	48	49
CACATG	10	10	CATGTG	11	12
CACATGG	4	4	CCATGTG	3	4
CACTA	45	45	TAGTG	51	52
CACTAC	9	9	GTAGTG	16	17
CACTACT	4	4	AGTAGTG	2	3
CACTTTG	4	4	CAAAGTG	5	6
CACTTTGT	2	2	ACAAAGTG	1	2
CATGG	27	27	CCATG	20	21
CATGGG	3	3	CCCATG	1	2
CATTC	33	33	GAATG	29	30
CATTCT	15	15	AGAATG	12	13
CATTCTC	2	2	GAGAATG	0	1
CCTAA	37	37	TTAGG	16	17
CCTAAG	6	6	CTTAGG	6	7
CCTAAGA	3	3	TCTTAGG	3	4
CCTTT	44	44	AAAGG	44	45
CCTTTA	7	7	TAAAGG	12	13
CCTTTAA	2	2	TTAAAGG	2	3

CTAAG	21	21	CTTAG	28	29
CTAAGA	7	7	TCTTAG	10	11
CTAAGAA	2	2	TTCTTAG	6	7
CTACT	59	59	AGTAG	23	24
CTACTA	25	25	TAGTAG	4	5
CTACTAA	10	10	TTAGTAG	1	2
CTACTAAA	3	3	TTTAGTAG	0	1
CTACTAAAA	2	2	TTTTAGTAG	0	1
CTATT	66	66	AATAG	32	33
CTATTA	22	22	TAATAG	11	12
CTATTAA	9	9	TTAATAG	5	6
CTATTAAA	2	2	TTTAATAG	2	3
CTCCT	8	8	AGGAG	24	25
CTCCTA	2	2	TAGGAG	5	6
CTTTAA	25	25	TTAAAG	20	21
CTTTAT	19	19	ATAAAG	15	14
CTTTATT	8	8	AATAAAG	4	3
CTTTATTG	2	2	CAATAAAG	1	0
CTTTG	60	60	CAAAG	45	46
CTTTGT	15	15	ACAAAG	23	24
CTTTGTT	6	6	AACAAAG	10	11
GAAGC	24	24	GCTTC	26	27
GAAGCT	14	14	AGCTTC	4	5
GAAGCTA	2	2	TAGCTTC	0	1
GAAGG	25	25	CCTTC	21	22
GAAGGT	16	16	ACCTTC	11	12
GAAGGTA	4	4	TACCTTC	4	5
GATAG	13	13	CTATC	18	19
GCACT	36	36	AGTGC	39	40
GCACTA	9	9	TAGTGC	16	17
GCACTAC	3	3	GTAGTGC	5	6
GCTATT	15	15	AATAGC	6	7
GCTATTA	5	5	TAATAGC	1	2
GGAAGG	2	2	CCTTCC	3	4
GGATA	8	8	TATCC	8	9
GGGAT	7	7	ATCCC	2	3
GGGGA	4	4	TCCCC	1	2
GGTAT	24	24	ATACC	24	25
GGTATA	5	5	TATACC	5	6
GTATA	25	25	TATAC	28	29
GTATAA	10	10	TTATAC	10	11
GTATAAA	5	5	TTTATAC	2	3
GTATAAAC	3	3	GTTTATAC	1	2
GTCAT	28	28	ATGAC	38	39
GTCATT	8	8	AATGAC	9	10
GTCATTC	2	2	GAATGAC	0	1
GTTGA	62	62	TCAAC	59	60
GTTGAA	23	23	TTCAAC	25	26
GTTGAAA	6	6	TTTCAAC	7	8
GTTGAAAA	3	3	TTTTCAAC	0	1
TAAAA	79	79	TTTTA	89	90
TAAAAT	22	22	ATTTTA	25	26
TAAAATC	2	2	GATTTTA	5	6
TAAAC	56	56	GTTTA	68	69
TAAACC	15	15	GGTTTA	16	17
TAAACCT	5	5	AGGTTTA	5	6
TAAACCTT	2	2	AAGGTTTA	3	4
TAAGA	34	34	TCTTA	57	58
TAAGAA	14	14	TTCTTA	27	28
TACTA	66	66	TAGTA	23	24
TACTAA	26	26	TTAGTA	8	9
TACTAAA	7	7	TTTAGTA	3	4
TACTAAAA	3	3	TTTTAGTA	0	1

TACTAAAAT	2	2	ATTTTAGTA	0	1
TAGCA	24	24	TGCTA	65	66
TAGCAC	8	8	GTGCTA	15	16
TAGCACT	4	4	AGTGCTA	4	5
TATAA	59	59	TTATA	48	49
TATAAA	20	20	TTTATA	13	14
TATAAAC	6	6	GTTTATA	5	6
TATTA	68	68	TAATA	44	45
TATTAA	17	17	TTAATA	15	16
TATTAAA	5	5	TTTAATA	6	7
TATTAAAA	2	2	TTTTAATA	1	2
TATTAAAAT	2	2	ATTTTAATA	1	2
TATTG	60	60	CAATA	29	28
TATTGA	15	15	TCAATA	11	10
TATTGAA	5	5	TTCAATA	3	2
TATTGAAA	2	2	TTTCAATA	3	2
TCACA	39	39	TGTGA	45	46
TCACAT	12	12	ATGTGA	13	14
TCACATG	2	2	CATGTGA	4	5
TCATT	37	37	AATGA	49	50
TCATTC	8	8	GAATGA	3	4
TCATTCT	2	2	AGAATGA	2	3
TCCTA	28	28	TAGGA	16	17
TCCTAA	9	9	TTAGGA	3	4
TCTCC	9	9	GGAGA	18	19
TCTCCT	3	3	AGGAGA	6	7
TGGGG	7	7	CCCCA	5	6
TGGGGA	2	2	TCCCCA	0	1
TGTCA	48	48	TGACA	64	65
TGTCAT	14	14	ATGACA	15	16
TGTCATT	3	3	AATGACA	4	5
TGTTG	97	97	CAACA	74	75
TGTTGA	29	29	TCAACA	20	21
TGTTGAA	9	9	TTCAACA	4	5
TTAAAA	33	33	TTTTAA	34	35
TTAAAAT	10	10	ATTTTAA	8	9
TTATT	72	72	AATAA	43	42
TTATTG	17	17	CAATAA	7	6
TTATTGA	5	5	TCAATAA	3	2
TTATTGAA	3	3	TTCAATAA	2	1
TTATTGAAA	2	2	TTTCAATAA	2	1
TTCTC	21	21	GAGAA	31	32
TTCTCC	4	4	GGAGAA	5	6
TTCTCCT	2	2	AGGAGAA	0	1
TTGTC	47	47	GACAA	52	53
TTGTCA	17	17	TGACAA	16	17
TTGTCAT	4	4	ATGACAA	7	8
TTGTT	102	102	AACAA	98	99
TTGTTG	22	22	CAACAA	28	29
TTGTTGA	10	10	TCAACAA	9	10
TTGTTGAA	3	3	TTCAACAA	1	2
TTTATT	27	27	AATAAA	16	15
TTTATTG	7	7	CAATAAA	4	3
TTTATTGA	3	3	TCAATAAA	2	1
TTTATTGAA	2	2	TTCAATAAA	2	1
TTTGT	88	88	ACAAA	87	89
TTTGTC	12	12	GACAAA	15	16
TTTGTCA	4	4	TGACAAA	7	8
TTTGTT	28	28	AACAAA	29	30
TTTGTTG	6	6	CAACAAA	7	8
TTTGTTGA	4	4	TCAACAAA	1	2
TTTTG	97	97	CAAAA	60	61
TTTTGT	32	32	ACAAAA	19	20

TTTTGTC	2	2	GACAAAA	2	3
TTTTT	61	61	AAAAA	56	64
TTTTTG	22	22	CAAAAA	14	15
TTTTTGT	9	9	ACAAAAA	4	5
TTTTTT	6	6	AAAAAA	2	9
TTTTTTG	2	2	CAAAAAA	0	1
TTTTTTT	2	2	AAAAAAA	0	6

According to these results, there are some distinct values in each sample, and there are some common values with different frequencies.