Grade: 100

Algorithms in computational Biology - HW #3

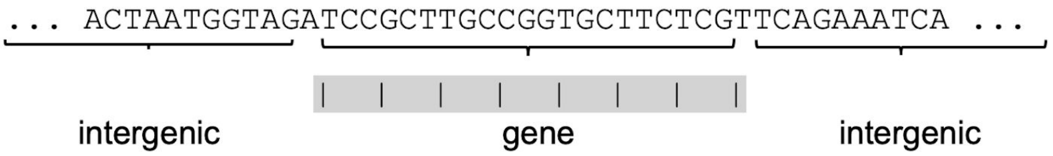
Submit:

- Ilay Anais,
- Hadar Pur

Researchers discovered a new virus (*oy vei*) whose DNA has the following peculiar features:

- Protein coding genes have a higher composition of C’s and G’s (40% each) than T’s (20%), and no A’s.
- DNA sequences outside of protein coding genes (termed intergenic) have the opposite bias in composition: 20% C’s and G’s and 30% A’s and T’s.
- Protein coding genes are always flanked by an A (before the gene) and T (after the gene). The flanking bases are not part of the protein coding sequence.
- Similar to other organisms, protein coding genes in this virus consist of a series of codons of length 3. The length of a protein coding gene is geometrically distributed with an average length of 5 codons (see note below). The length of an inter-genic segment (between terminating T and next starting A and) is also geometrically distributed with an average length of 20 bases. Note that a gene is never empty, but two genes may be separated by a terminating T followed by a starting A.

The following is a typical sequence in the virus’ DNA with its gene annotation given below (including the boundaries of the seven codons in the gene):



Note: a random variable *X* is said to be geometrically distributed *X*~*Geom*(*p*) if: P(*X*=*x*) = (1-*p*)*x*-1*p*. The mean of such a variable is E[*X*]=1/*p*. Notice that the length of a consecutive sequence of annotation of the same state in an HMM is geometrically distributed with parameter *p* determined by the probability of transition out of that state.

Section A:

The nine possible annotations of the following fragment of viral DNA sequence:

1. CCATCGCACTCCGATGTGGCCGGTGCTCACGTTGCCT
2. CCATCGCACTCCGATGTGGCCGGTGCTCACGTTGCCT
3. CCATCGCACTCCGATGTGGCCGGTGCTCACGTTGCCT
4. CCATCGCACTCCGATGTGGCCGGTGCTCACGTTGCCT
5. CCATCGCACTCCGATGTGGCCGGTGCTCACGTTGCCT
6. CCATCGCACTCCGATGTGGCCGGTGCTCACGTTGCCT
7. CCATCGCACTCCGATGTGGCCGGTGCTCACGTTGCCT
8. CCATCGCACTCCGATGTGGCCGGTGCTCACGTTGCCT
9. CCATCGCACTCCGATGTGGCCGGTGCTCACGTTGCCT

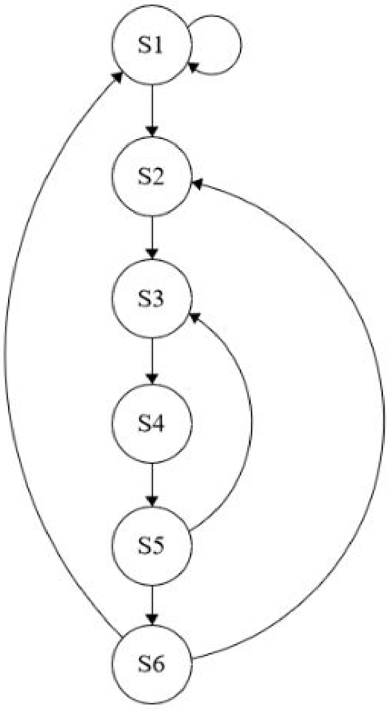
We marked the potential genes of the virus in blue, and the rest of the sequence are the inter-genics.

- 4,5,7,8: We assume that there are inter-genic segments between the genes.
- 6,7,8: We can assume that it will continue further to the right in the full sequence, and that we will have a valid continuation after this.
- 9: We assume that there are simply no genes of the virus in this sequence, and that the entire segment is an inter-genic segment.

Section B:

Hidden Markov model (HMM) that models the DNA sequence of the virus together with the appropriate annotations:

State machine:



S1: allow inter-genic
S2: allow flank to start with an A
S3: allow codon letter 1
S4: allow codon letter 2
S5: allow codon letter 3
S6: allow flank to end with a T

Emission probability matrix:

From\To	A	C	G	T
allow inter-genic	0.3	0.2	0.2	0.3
allow flank to start with an A	1	0	0	0
allow codon letter 1	0	0.4	0.4	0.2
allow codon letter 2	0	0.4	0.4	0.2
allow codon letter 3	0	0.4	0.4	0.2
allow flank to end with a T	0	0	0	1

Transition probability matrix:

Since the length of inter-genics' average is 20 bases:

- p = 1/20 = 0.05 (to move to another state that there is not inter-genic)
- 1-p = 19/20 = 0.95 (to move to state that is inter-genic or stay in this state)

Since the length of the protein gene is average of 5 codons, each of which is 3 letters long:

- p = 1/5 = 0.2 (to move to the end of the gene)
- 1-p = 4/5 = 0.8 (to stay in the gene)

From\To	S1	S2	S3	S4	S5	S6
S1	0.95	0.05	0	0	0	0
S2	0	0	1	0	0	0
S3	0	0	0	1	0	0
S4	0	0	0	0	1	0
S5	0	0	0.8	0	0	0.2
S6	0.95	0.05	0	0	0	0

Section C:

1. CCATCGCACTCCGATGTGGCCGGTGCTCACGTTGCCT

C	C	A	T	C	G	C	A	C	T	C	C	G	A	T	G	T	G	G	C	C	G	G	T	G	C	T	C	A	C	G	T	T	G	C	C	T
1	1	1	1	1	1	1	1	1	1	1	1	1	2	3	4	5	3	4	5	3	4	5	6	1	1	1	1	1	1	1	1	1	1	1	1	1

2. CCATCGCACTCCGATGTGGCCGGTGCTCACGTTGCCT

C	C	A	T	C	G	C	A	C	T	C	C	G	A	T	G	T	G	G	C	C	G	G	T	G	C	T	C	A	C	G	T	T	G	C	C	T
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	3	4	5	6	1	1	1	1

3. CCATCGCACTCCGATGTGGCCGGTGCTCACGTTGCCT

C	C	A	T	C	G	C	A	C	T	C	C	G	A	T	G	T	G	G	C	C	G	G	T	G	C	T	C	A	C	G	T	T	G	C	C	T
1	1	1	1	1	1	1	1	1	1	1	1	1	2	3	4	5	3	4	5	3	4	5	3	4	5	6	1	1	1	1	1	1	1	1	1	1

4. CCATCGCACTCCGATGTGGCCGGTGCTCACGTTGCCT

C	C	A	T	C	G	C	A	C	T	C	C	G	A	T	G	T	G	G	C	C	G	G	T	G	C	T	C	A	C	G	T	T	G	C	C	T
1	1	1	1	1	1	1	1	1	1	1	1	1	2	3	4	5	3	4	5	3	4	5	6	1	1	1	1	2	3	4	5	6	1	1	1	1

5. CCATCGCACTCCGATGTGGCCGGTGCTCACGTTGCCT

C	C	A	T	C	G	C	A	C	T	C	C	G	A	T	G	T	G	G	C	C	G	G	T	G	C	T	C	A	C	G	T	T	G	C	C	T
1	1	1	1	1	1	1	1	1	1	1	1	1	2	3	4	5	3	4	5	3	4	5	3	4	5	6	1	2	3	4	5	6	1	1	1	1

6. CCATCGCACTCCGATGTGGCCGGTGCTCACGTTGCCT

C	C	A	T	C	G	C	A	C	T	C	C	G	A	T	G	T	G	G	C	C	G	G	T	G	C	T	C	A	C	G	T	T	G	C	C	T
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	3	4	5	3	4	5	3	4

7. CCATCGCACTCCGATGTGGCCGGTGCTCACGTTGCCT

C	C	A	T	C	G	C	A	C	T	C	C	G	A	T	G	T	G	G	C	C	G	G	T	G	C	T	C	A	C	G	T	T	G	C	C	T
1	1	1	1	1	1	1	1	1	1	1	1	1	2	3	4	5	3	4	5	3	4	5	6	1	1	1	1	2	3	4	5	3	4	5	3	4

8. CCATCGCACTCCGATGTGGCCGGTGCTCACGTTGCCT

C	C	A	T	C	G	C	A	C	T	C	C	G	A	T	G	T	G	G	C	C	G	G	T	G	C	T	C	A	C	G	T	T	G	C	C	T
1	1	1	1	1	1	1	1	1	1	1	1	1	2	3	4	5	3	4	5	3	4	5	3	4	5	6	1	2	3	4	5	3	4	5	3	4

9. CCATCGCACTCCGATGTGGCCGGTGCTCACGTTGCCT

C	C	A	T	C	G	C	A	C	T	C	C	G	A	T	G	T	G	G	C	C	G	G	T	G	C	T	C	A	C	G	T	T	G	C	C	T
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Section D:

Link for google colab:
https://colab.research.google.com/drive/1tqQDUl3zu9RNlpboWf4yvD6wyaJ3aWIF?usp=sharing
We added the .py file just in case you can’t open the Google collab link.

Section E:

Viterbi's max-prob annotation:

C	C	A	T	C	G	C	A	C	T	C	C	G	A	T	G	T	G	G	C	C	G	G	T	G	C	T	C	A	C	G	T	T	G	C	C	T
1	1	1	1	1	1	1	1	1	1	1	1	1	2	3	4	5	3	4	5	3	4	5	3	4	5	6	1	2	3	4	5	3	4	5	3	4

max log(P(S|X,HMM)) = -52.77840110563882
max P(S|X,HMM) = 1.1984823322410419e-23

Section F:

Link for google colab:
https://colab.research.google.com/drive/1tqQDUl3zu9RNlpboWf4yvD6wyaJ3aWIF?usp=sharing
We added the .py file just in case you can’t open the Google collab link.

Section G:

Maximum A-Posteriori Probability:

P(S_1 = S1|X,HMM) = 1.0
P(S_2 = S1|X,HMM) = 1.0
P(S_3 = S1|X,HMM) = 0.9999999999999982
P(S_4 = S1|X,HMM) = 0.9999999999999974
P(S_5 = S1|X,HMM) = 0.9999999999999983
P(S_6 = S1|X,HMM) = 0.9999999999999986
P(S_7 = S1|X,HMM) = 0.9999999999999986
P(S_8 = S1|X,HMM) = 0.999999999999983
P(S_9 = S1|X,HMM) = 0.9999999999999827
P(S_10 = S1|X,HMM) = 0.9999999999999819
P(S_11 = S1|X,HMM) = 0.9999999999999847
P(S_12 = S1|X,HMM) = 0.9999999999999847
P(S_13 = S1|X,HMM) = 0.999999999999985
P(S_14 = S2|X,HMM) = 0.9612537613310652
P(S_15 = S3|X,HMM) = 0.9612537613310734
P(S_16 = S4|X,HMM) = 0.9612537613310747
P(S_17 = S5|X,HMM) = 0.9612537613310755
P(S_18 = S3|X,HMM) = 0.9612537613310772
P(S_19 = S4|X,HMM) = 0.9612537613310779
P(S_20 = S5|X,HMM) = 0.9612537613310783
P(S_21 = S3|X,HMM) = 0.9612537613310784
P(S_22 = S4|X,HMM) = 0.9612537613310788
P(S_23 = S5|X,HMM) = 0.9612537613310788
P(S_24 = S3|X,HMM) = 0.6856819396206495
P(S_25 = S4|X,HMM) = 0.6856819396206496
P(S_26 = S5|X,HMM) = 0.6856819396206495
P(S_27 = S6|X,HMM) = 0.6856819396206422
P(S_28 = S1|X,HMM) = 0.9999999999999937
P(S_29 = S2|X,HMM) = 0.6652848561799922
P(S_30 = S3|X,HMM) = 0.6652848561799958
P(S_31 = S4|X,HMM) = 0.6652848561799961
P(S_32 = S5|X,HMM) = 0.6652848561799946
P(S_33 = S3|X,HMM) = 0.5371156901773128
P(S_34 = S4|X,HMM) = 0.5371156901773116
P(S_35 = S5|X,HMM) = 0.5371156901773114
P(S_36 = S3|X,HMM) = 0.5371156901773111
P(S_37 = S4|X,HMM) = 0.537115690177309

C	C	A	T	C	G	C	A	C	T	C	C	G	A	T	G	T	G	G	C	C	G	G	T	G	C	T	C	A	C	G	T	T	G	C	C	T
1	1	1	1	1	1	1	1	1	1	1	1	1	2	3	4	5	3	4	5	3	4	5	3	4	5	6	1	2	3	4	5	3	4	5	3	4

likelihood log(P(X|HMM)) = -51.77951793226804
likelihood P(X|HMM) = 3.2541763644144575e-23

There was no change in the results from E for the most likely path.

Section H:

New Emission probability matrix:

From\To	A	C	G	T
allow inter-genic	0.3	0.2	0.2	0.3
allow flank to start with an A	1	0	0	0
allow codon letter 1	0.05	0.4	0.4	0.15
allow codon letter 2	0.05	0.4	0.4	0.15
allow codon letter 3	0.05	0.4	0.4	0.15
allow flank to end with a T	0	0	0	1

New Transition probability matrix stays the same as before:

From\To	S1	S2	S3	S4	S5	S6
S1	0.95	0.05	0	0	0	0
S2	0	0	1	0	0	0
S3	0	0	0	1	0	0
S4	0	0	0	0	1	0
S5	0	0	0.8	0	0	0.2
S6	0.95	0.05	0	0	0	0

Yes, there was a change:

Before:

Viterbi's max-prob annotation:

C	C	A	T	C	G	C	A	C	T	C	C	G	A	T	G	T	G	G	C	C	G	G	T	G	C	T	C	A	C	G	T	T	G	C	C	T
1	1	1	1	1	1	1	1	1	1	1	1	1	2	3	4	5	3	4	5	3	4	5	3	4	5	6	1	2	3	4	5	3	4	5	3	4

max log(P(S|X,HMM)) = -52.77840110563882

max P(S|X,HMM) = 1.1984823322410419e-23

```
Maximum A-Posteriori Probability:
P(S_1 = S1|X,HMM) = 1.0
P(S_2 = S1|X,HMM) = 1.0
P(S_3 = S1|X,HMM) = 0.9999999999999982
P(S_4 = S1|X,HMM) = 0.9999999999999974
P(S_5 = S1|X,HMM) = 0.9999999999999983
P(S_6 = S1|X,HMM) = 0.9999999999999986
P(S_7 = S1|X,HMM) = 0.9999999999999986
P(S_8 = S1|X,HMM) = 0.999999999999983
P(S_9 = S1|X,HMM) = 0.9999999999999827
P(S_10 = S1|X,HMM) = 0.9999999999999819
P(S_11 = S1|X,HMM) = 0.9999999999999847
P(S_12 = S1|X,HMM) = 0.9999999999999847
P(S_13 = S1|X,HMM) = 0.999999999999985
P(S_14 = S2|X,HMM) = 0.9612537613310652
P(S_15 = S3|X,HMM) = 0.9612537613310734
P(S_16 = S4|X,HMM) = 0.9612537613310747
P(S_17 = S5|X,HMM) = 0.9612537613310755
P(S_18 = S3|X,HMM) = 0.9612537613310772
P(S_19 = S4|X,HMM) = 0.9612537613310779
P(S_20 = S5|X,HMM) = 0.9612537613310783
P(S_21 = S3|X,HMM) = 0.9612537613310784
P(S_22 = S4|X,HMM) = 0.9612537613310788
P(S_23 = S5|X,HMM) = 0.9612537613310788
P(S_24 = S3|X,HMM) = 0.6856819396206495
P(S_25 = S4|X,HMM) = 0.6856819396206496
P(S_26 = S5|X,HMM) = 0.6856819396206495
P(S_27 = S6|X,HMM) = 0.6856819396206422
P(S_28 = S1|X,HMM) = 0.9999999999999937
P(S_29 = S2|X,HMM) = 0.6652848561799922
P(S_30 = S3|X,HMM) = 0.6652848561799958
P(S_31 = S4|X,HMM) = 0.6652848561799961
P(S_32 = S5|X,HMM) = 0.6652848561799946
P(S_33 = S3|X,HMM) = 0.5371156901773128
P(S_34 = S4|X,HMM) = 0.5371156901773116
P(S_35 = S5|X,HMM) = 0.5371156901773114
P(S_36 = S3|X,HMM) = 0.5371156901773111
P(S_37 = S4|X,HMM) = 0.537115690177309
```

C	C	A	T	C	G	C	A	C	T	C	C	G	A	T	G	T	G	G	C	C	G	G	T	G	C	T	C	A	C	G	T	T	G	C	C	T
1	1	1	1	1	1	1	1	1	1	1	1	1	2	3	4	5	3	4	5	3	4	5	3	4	5	6	1	2	3	4	5	3	4	5	3	4

```
likelihood log(P(X|HMM)) =  -51.77951793226804
likelihood P(X|HMM) =  3.2541763644144575e-23
```

After:

Viterbi's max-prob annotation:

C	C	A	T	C	G	C	A	C	T	C	C	G	A	T	G	T	G	G	C	C	G	G	T	G	C	T	C	A	C	G	T	T	G	C	C	T
1	1	1	1	1	1	1	2	3	4	5	3	4	5	3	4	5	3	4	5	3	4	5	3	4	5	6	1	1	1	1	1	1	1	1	1	1

max log(P(S|X,HMM)) = -53.965226242952205
max P(S|X,HMM) = 3.657632166123875e-24

Maximum A-Posteriori Probability:
P(S_1 = S1|X,HMM) = 1.0
P(S_2 = S1|X,HMM) = 0.9999999999999997
P(S_3 = S1|X,HMM) = 0.8225236665011396
P(S_4 = S1|X,HMM) = 0.8225236665011378
P(S_5 = S1|X,HMM) = 0.8225236665011372
P(S_6 = S1|X,HMM) = 0.8225236665011366
P(S_7 = S1|X,HMM) = 0.8225236665011363
P(S_8 = S2|X,HMM) = 0.43086615563856207
P(S_9 = S3|X,HMM) = 0.4308661556385641
P(S_10 = S4|X,HMM) = 0.43086615563856395
P(S_11 = S1|X,HMM) = 0.46162621511831947
P(S_12 = S1|X,HMM) = 0.4616262151183194
P(S_13 = S1|X,HMM) = 0.4616262151183193
P(S_14 = S2|X,HMM) = 0.43307759078914154
P(S_15 = S3|X,HMM) = 0.8596166302735987
P(S_16 = S4|X,HMM) = 0.8596166302735995
P(S_17 = S5|X,HMM) = 0.8596166302735989
P(S_18 = S3|X,HMM) = 0.8596166302736004
P(S_19 = S4|X,HMM) = 0.8596166302736008
P(S_20 = S5|X,HMM) = 0.8596166302736008
P(S_21 = S3|X,HMM) = 0.8596166302736009
P(S_22 = S4|X,HMM) = 0.8596166302736009
P(S_23 = S5|X,HMM) = 0.8596166302736014
P(S_24 = S3|X,HMM) = 0.6329163447316773
P(S_25 = S4|X,HMM) = 0.6329163447316781
P(S_26 = S5|X,HMM) = 0.6329163447316783
P(S_27 = S6|X,HMM) = 0.42305928778781265
P(S_28 = S1|X,HMM) = 0.6826353138130123
P(S_29 = S1|X,HMM) = 0.34754377061031105
P(S_30 = S3|X,HMM) = 0.5449486001465671
P(S_31 = S4|X,HMM) = 0.544948600146567
P(S_32 = S5|X,HMM) = 0.5449486001465654
P(S_33 = S3|X,HMM) = 0.3826289708723117
P(S_34 = S1|X,HMM) = 0.5098633998845622
P(S_35 = S1|X,HMM) = 0.5098633998845612
P(S_36 = S1|X,HMM) = 0.5098633998845605
P(S_37 = S1|X,HMM) = 0.5098633998845611

C	C	A	T	C	G	C	A	C	T	C	C	G	A	T	G	T	G	G	C	C	G	G	T	G	C	T	C	A	C	G	T	T	G	C	C	T
1	1	1	1	1	1	1	2	3	4	1	1	1	2	3	4	5	3	4	5	3	4	5	3	4	5	6	1	1	3	4	5	3	1	1	1	1

likelihood log(P(X|HMM)) = -51.72913052589137
likelihood P(X|HMM) = 3.422347136736786e-23

Section I:

The original DNA sequence we got in section A written below is actually a sequence that works with these criteria:

- Protein coding genes have a higher composition of C's and G's (40% each) than T's (15%), and A's (5%).
- DNA sequences outside of protein coding genes (termed intergenic) have the opposite bias in composition: 20% C's and G's and 30% A's and T's.
- Protein coding genes are always flanked by an A (before the gene) and T (after the gene). The flanking bases are not part of the protein coding sequence.
- Similar to other organisms, protein coding genes in this virus consist of a series of codons of length 3. The length of a protein coding gene is geometrically distributed with an average length of 5 codons (see note below). The length of an inter-genic segment (between terminating T and next starting A and) is also geometrically distributed with an average length of 20 bases. Note that a gene is never empty, but two genes may be separated by a terminating T followed by a starting A

DNA sequence:

'CCATCGCACTCCGATGTGGCCGGTGCTCACGTTGCCT'

Viterbi's max-prob annotation:

C	C	A	T	C	G	C	A	C	T	C	C	G	A	T	G	T	G	G	C	C	G	G	T	G	C	T	C	A	C	G	T	T	G	C	C	T
1	1	1	1	1	1	1	2	3	4	5	3	4	5	3	4	5	3	4	5	3	4	5	3	4	5	6	1	1	1	1	1	1	1	1	1	1

Maximum A-Posteriori Probability:

C	C	A	T	C	G	C	A	C	T	C	C	G	A	T	G	T	G	G	C	C	G	G	T	G	C	T	C	A	C	G	T	T	G	C	C	T
1	1	1	1	1	1	1	2	3	4	1	1	1	2	3	4	5	3	4	5	3	4	5	3	4	5	6	1	1	3	4	5	3	1	1	1	1

Where the highlighted red cells show an impossible (0 probability) transition.