Homework1

Ruilong Chen

September 28, 2024

1 Question1

In 1990, Michael Crichton published the book "Jurassic Park" about the resurrection of dinosaurs using the blood from the stomachs of insects which had been encased in tree sap, later turned into the mineral, amber.

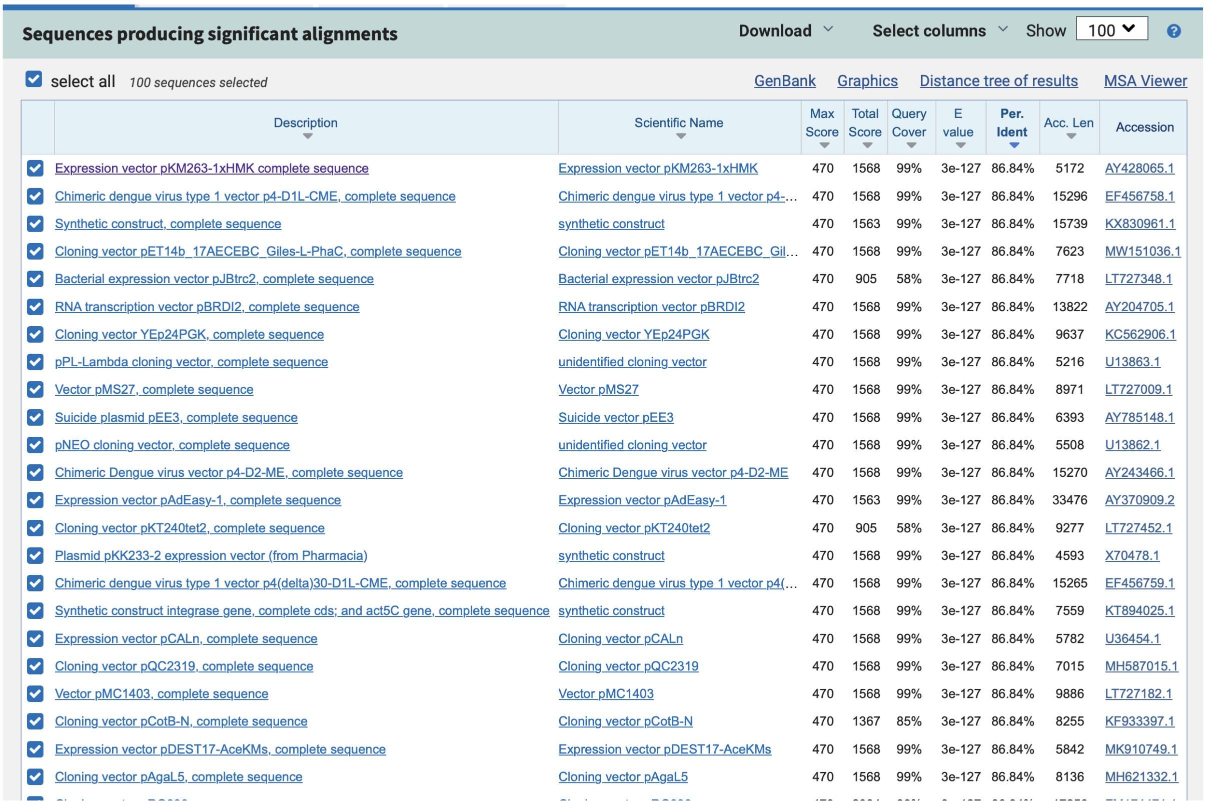
At one point in the book, Dr. Henry Wu is asked to explain some of DNA techniques used in reconstructing the extinct dinosaur genomes. Dr. Wu describes the use of restriction enzymes and how the fragmented pieces of dino DNA can be spliced together with these enzymes. He also alludes to the fact that they don't have the entire genome but that they "fill in the gap" with modern day frog DNA.

At one point during his discussion he points to a computer screen and remarks "Here you see the actual structure of a small fragment of dinosaur DNA."

>JurassicPark DinoDNA from the book Jurassic Park

```
gcgttgctgg cgtttttcca taggctccgc ccccctgacg agcatcacaa aaatcgacgc
ggtggcgaaa cccgacagga ctataaagat accaggcgtt tccccctgga agctccctcg
tgttccgacc ctgccgctta ccggatacct gtccgccttt ctcccttcgg gaagcgtggc
tgctcacgct gtaggtatct cagttcggtg taggtcgttc gctccaagct gggctgtgtg
ccgttcagcc cgaccgctgc gccttatccg gtaactatcg tcttgagtcc aacccggtaa
agtaggacag gtgccggcag cgctctgggt cattttcggc gaggaccgct ttcgctggag
atcggcctgt cgcttgcggt attcggaatc ttgcacgccc tcgctcaagc cttcgtcact
ccaaacgttt cggcgagaag caggccatta tcgccggcat ggcggccgac gcgctgggct
ggcgttcgcg acgcgaggct ggatggcctt ccccattatg attcttctcg cttccggcgg
cccgcgttgc aggccatgct gtccaggcag gtagatgacg accatcaggg acagcttcaa
cggctcttac cagcctaact tcgatcactg gaccgctgat cgtcacggcg atttatgccg
caagtcagag gtggcgaaac ccgacaagga ctataaagat accaggcgtt tcccctggaa
gcgctctcct gttccgaccc tgccgcttac cggatacctg tccgcctttc tcccttcggg
ctttctcatt gctcacgctg taggtatctc agttcggtgt aggtcgttcg ctccaagctg
acgaaccccc cgttcagccc gaccgctgcg ccttatccgg taactatcgt cttgagtcca
acacgactta acgggttggc atggattgta ggcgccgccc tataccttgt ctgcctcccc
gcggtgcatg gagccgggcc acctcgacct gaatggaagc cggcggcacc tcgctaacgg
ccaagaattg gagccaatca attcttgcgg agaactgtga atgcgcaaac caacccttgg
ccatcgcgtc cgccatctcc agcagccgca cgcggcgcat ctcgggcagc gttgggtcct
gcgcatgatc gtgctagcct gtcgttgagg acccggctag gctggcgggg ttgccttact
atgaatcacc gatacgcgag cgaacgtgaa gcgactgctg ctgcaaaacg tctgcgacct
atgaatggtc ttcggtttcc gtgtttcgta aagtctggaa acgcggaagt cagcgccctg
```

Select, copy, and paste the sequence shown above into NCBI BLAST web portal to run BLAST. Comment on the results. (3pts)



Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Expression vector pKM263-1xHMK complete sequence	Expression vector pKM263-1xHMK	470	1568	99%	3e-127	86.84%	5172	AY428065.1
☑	Chimeric dengue virus type 1 vector p4-D1L-CME, complete sequence	Chimeric dengue virus type 1 vector p4-...	470	1568	99%	3e-127	86.84%	15296	EF456758.1
☑	Synthetic construct, complete sequence	synthetic construct	470	1563	99%	3e-127	86.84%	15739	KX830961.1
☑	Cloning vector pET14b_17AECEBC_Giles-L-PhaC, complete sequence	Cloning vector pET14b_17AECEBC_Gil...	470	1568	99%	3e-127	86.84%	7623	MW151036.1
☑	Bacterial expression vector pJBtrc2, complete sequence	Bacterial expression vector pJBtrc2	470	905	58%	3e-127	86.84%	7718	LT727348.1
☑	RNA transcription vector pBRDI2, complete sequence	RNA transcription vector pBRDI2	470	1568	99%	3e-127	86.84%	13822	AY204705.1
☑	Cloning vector YEp24PGK, complete sequence	Cloning vector YEp24PGK	470	1568	99%	3e-127	86.84%	9637	KC562906.1
☑	pPL-Lambda cloning vector, complete sequence	unidentified cloning vector	470	1568	99%	3e-127	86.84%	5216	U13863.1
☑	Vector pMS27, complete sequence	Vector pMS27	470	1568	99%	3e-127	86.84%	8971	LT727009.1
☑	Suicide plasmid pEE3, complete sequence	Suicide vector pEE3	470	1568	99%	3e-127	86.84%	6393	AY785148.1
☑	pNEO cloning vector, complete sequence	unidentified cloning vector	470	1568	99%	3e-127	86.84%	5508	U13862.1
☑	Chimeric Dengue virus vector p4-D2-ME, complete sequence	Chimeric Dengue virus vector p4-D2-ME	470	1568	99%	3e-127	86.84%	15270	AY243466.1
☑	Expression vector pAdEasy-1, complete sequence	Expression vector pAdEasy-1	470	1563	99%	3e-127	86.84%	33476	AY370909.2
☑	Cloning vector pKT240tet2, complete sequence	Cloning vector pKT240tet2	470	905	58%	3e-127	86.84%	9277	LT727452.1
☑	Plasmid pKK233-2 expression vector (from Pharmacia)	synthetic construct	470	1568	99%	3e-127	86.84%	4593	X70478.1
☑	Chimeric dengue virus type 1 vector p4(delta)30-D1L-CME, complete sequence	Chimeric dengue virus type 1 vector p4(...	470	1568	99%	3e-127	86.84%	15265	EF456759.1
☑	Synthetic construct integrase gene, complete cds; and act5C gene, complete sequence	synthetic construct	470	1568	99%	3e-127	86.84%	7559	KT894025.1
☑	Expression vector pCALn, complete sequence	Cloning vector pCALn	470	1568	99%	3e-127	86.84%	5782	U36454.1
☑	Cloning vector pQC2319, complete sequence	Cloning vector pQC2319	470	1568	99%	3e-127	86.84%	7015	MH587015.1
☑	Vector pMC1403, complete sequence	Vector pMC1403	470	1568	99%	3e-127	86.84%	9886	LT727182.1
☑	Cloning vector pCotB-N, complete sequence	Cloning vector pCotB-N	470	1367	85%	3e-127	86.84%	8255	KF933397.1
☑	Expression vector pDEST17-AceKMs, complete sequence	Expression vector pDEST17-AceKMs	470	1568	99%	3e-127	86.84%	5842	MK910749.1
☑	Cloning vector pAgaL5, complete sequence	Cloning vector pAgaL5	470	1568	99%	3e-127	86.84%	8136	MH621332.1

There are several matched gene sequence have the same best results. The 'Total Score' is 1568; the 'Query Cover' is 99%; The 'E Value' is 3e-127; The 'Per. Ident' is 86.84%. After checking the source of these matched gene sequences, we are able to find that most of the results are plasmid sequences. It's likely that this gene has been inserted into plasmid vectors for research purposes, such as cloning, expression, or gene transfer experiments.

2 Question2

Mark's published article was brought to Michael Crichton's attention. In his second book, "The Lost World", Mr. Crichton used Mark as a consultant.

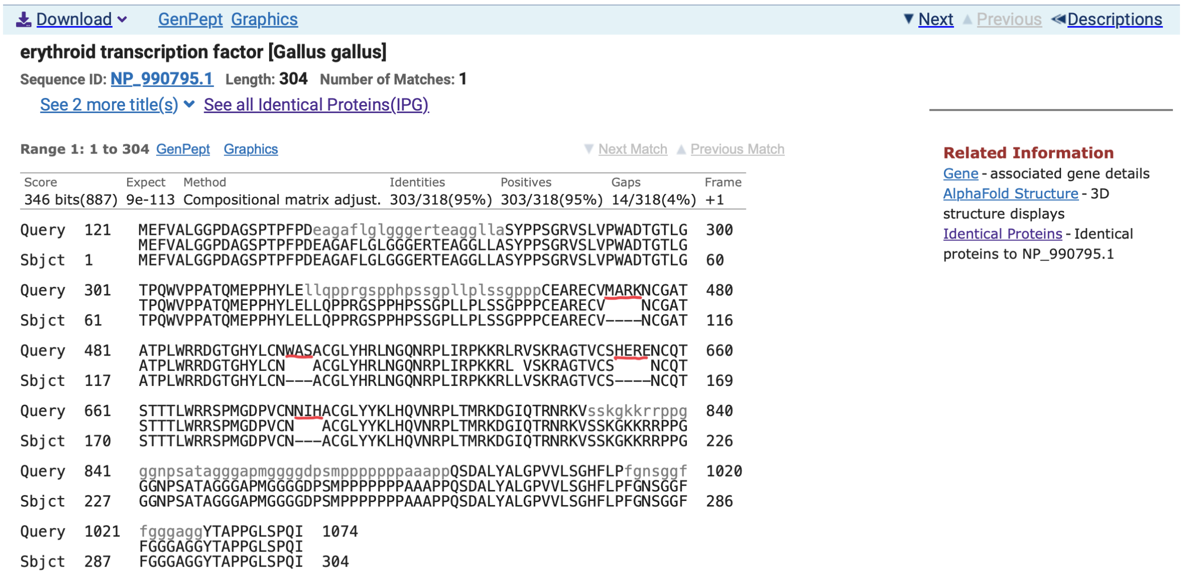
Here is the sequence Mark gave Michael Crichton for the book "The Lost World":

>LostWorld DinoDNA from the book The Lost World

gaattccgga agcgagcaag agataagtcc tggcatcaga tacagttgga gataaggacg
gacgtgtggc agctcccgca gaggattcac tggaagtgca ttacctatcc catgggagcc
atggagttcg tggcgctggg ggggccggat gcgggctccc ccactccgtt ccctgatgaa
gccggagcct tcctggggct gggggggggc gagaggacgg aggcgggggg gctgctggcc
tcctaccccc cctcaggccg cgtgtccctg gtgccgtggg cagacacggg tactttgggg
accccccagt gggtgccgcc cgccacccaa atggagcccc cccactacct ggagctgctg
caaccccccc ggggcagccc cccccatccc tcctccgggc ccctactgcc actcagcagc
gggcccccac cctgcgaggc ccgtgagtgc gtcatggcca ggaagaactg cggagcgacg
gcaacgccgc tgtggcgccg ggacggcacc gggcattacc tgtgcaactg ggcctcagcc
tgcgggctct accaccgcct caacggccag aaccgcccgc tcatccgccc caaaaagcgc
ctgcgggtga gtaagcgcgc aggcacagtg tgcagccacg agcgtgaaaa ctgccagaca
tccaccacca ctctgtggcg tcgcagcccc atgggggacc ccgtctgcaa caacattcac
gcctgcggcc tctactacaa actgcaccaa gtgaaccgcc ccctcacgat gcgcaaagac
ggaatccaaa cccgaaaccg caaagtttcc tccaagggta aaaagcggcg cccccccgggg
gggggaaacc cctccgccac cgcgggaggg ggcgctccta tggggggagg gggggacccc

tctatgcccc ccccgccgcc ccccccggcc gccgcccccc ctcaaagcga cgctctgtac
gctctcggcc ccgtggtcct ttcgggccat tttctgccct ttggaaactc cggagggttt
tttggggggg gggcgggggg ttacacggcc cccccggggc tgagcccgca gatttaaata
ataactctga cgtgggcaag tgggccttgc tgagaagaca gtgtaacata ataatttgca
cctcggcaat tgcagagggt cgatctccac tttggacaca acagggctac tcggtaggac
cagataagca ctttgctccc tggactgaaa aagaaaggat ttatctgttt gcttcttgct
gacaaatccc tgtgaaaggt aaaagtcgga cacagcaatc gattatttct cgcctgtgtg
aaattactgt gaatattgta aatatatata tatatatata tatatctgta tagaacagcc
tcggaggcgg catggaccca gcgtagatca tgctggattt gtactgccgg aattc

Select, copy, and paste the “Lost World” sequence into NCBI BLAST web portal to run blastx. Can you find Mark’s hidden message? Hine: look at the best pairwise alignment. (3 pts)



erythroid transcription factor [Gallus gallus]

Sequence ID: NP_990795.1 Length: 304 Number of Matches: 1

See 2 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 304 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps	Frame
346 bits(887)	9e-113	Compositional matrix adjust.	303/318(95%)	303/318(95%)	14/318(4%)	+1

```
Query  121   MEFVALGGPDAGSPTPFPDeagaflglgggerteagglla SYPPSGRVSLVPWADTGTLG  300
             MEFVALGGPDAGSPTPFPDEAGAFLGLGGGERTEAGGLLASYPPSGRVSLVPWADTGTLG
Sbjct  1     MEFVALGGPDAGSPTPFPDEAGAFLGLGGGERTEAGGLLASYPPSGRVSLVPWADTGTLG  60

Query  301   TPQWVPPATQMEPPHYLEllqpprgspphpssgpllplssgpppCEARECVMARKNCGAT  480
             TPQWVPPATQMEPPHYLELLQPPRGSPPHPSSGPLLPLSSGPPPCEARECV    NCGAT
Sbjct  61    TPQWVPPATQMEPPHYLELLQPPRGSPPHPSSGPLLPLSSGPPPCEARECV----NCGAT  116

Query  481   ATPLWRRDGTGHYLCNWASACGLYHRLNGQNRPLIRPKKRLRVSKRAGTVCSHERENCQT  660
             ATPLWRRDGTGHYLCN   ACGLYHRLNGQNRPLIRPKKRL VSKRAGTVCS    NCQT
Sbjct  117   ATPLWRRDGTGHYLCN---ACGLYHRLNGQNRPLIRPKKRLLVSKRAGTVCS----NCQT  169

Query  661   STTTLWRRSPMGDPVCNNIHACGLYYKLHQVNRPLTMRKDGIQTRNRKVsskgkkrrppg  840
             STTTLWRRSPMGDPVCN   ACGLYYKLHQVNRPLTMRKDGIQTRNRKVSSKGKKRRPPG
Sbjct  170   STTTLWRRSPMGDPVCN---ACGLYYKLHQVNRPLTMRKDGIQTRNRKVSSKGKKRRPPG  226

Query  841   ggnpsatagggapmggggdpsmppppppppaaappQSDALYALGPVVLSGHFLPfgnsggf  1020
             GGNPSATAGGGAPMGGGGDPSMPPPPPPPPAAAPPQSDALYALGPVVLSGHFLPFGNSGGF
Sbjct  227   GGNPSATAGGGAPMGGGGDPSMPPPPPPPPAAAPPQSDALYALGPVVLSGHFLPFGNSGGF  286

Query  1021  fgggaggYTAPPGLSPQI  1074
             FGGGAGGYTAPPGLSPQI
Sbjct  287   FGGGAGGYTAPPGLSPQI  304
```

Related Information

Gene - associated gene details
AlphaFold Structure - 3D structure displays
Identical Proteins - Identical proteins to NP_990795.1

Looking at the best pairwise alignment, we can find that Mark’s hidden massage is “MARK WAS HERE NIH”

Also, the result shows that the best matched sequence refers to a protein called GATA-1(erythroid transcription factor) from Gallup gallus (chicken). GATA-1 is a critical transcription factor involved in the regulation of genes necessary for erythropoiesis (the production of red blood cells).

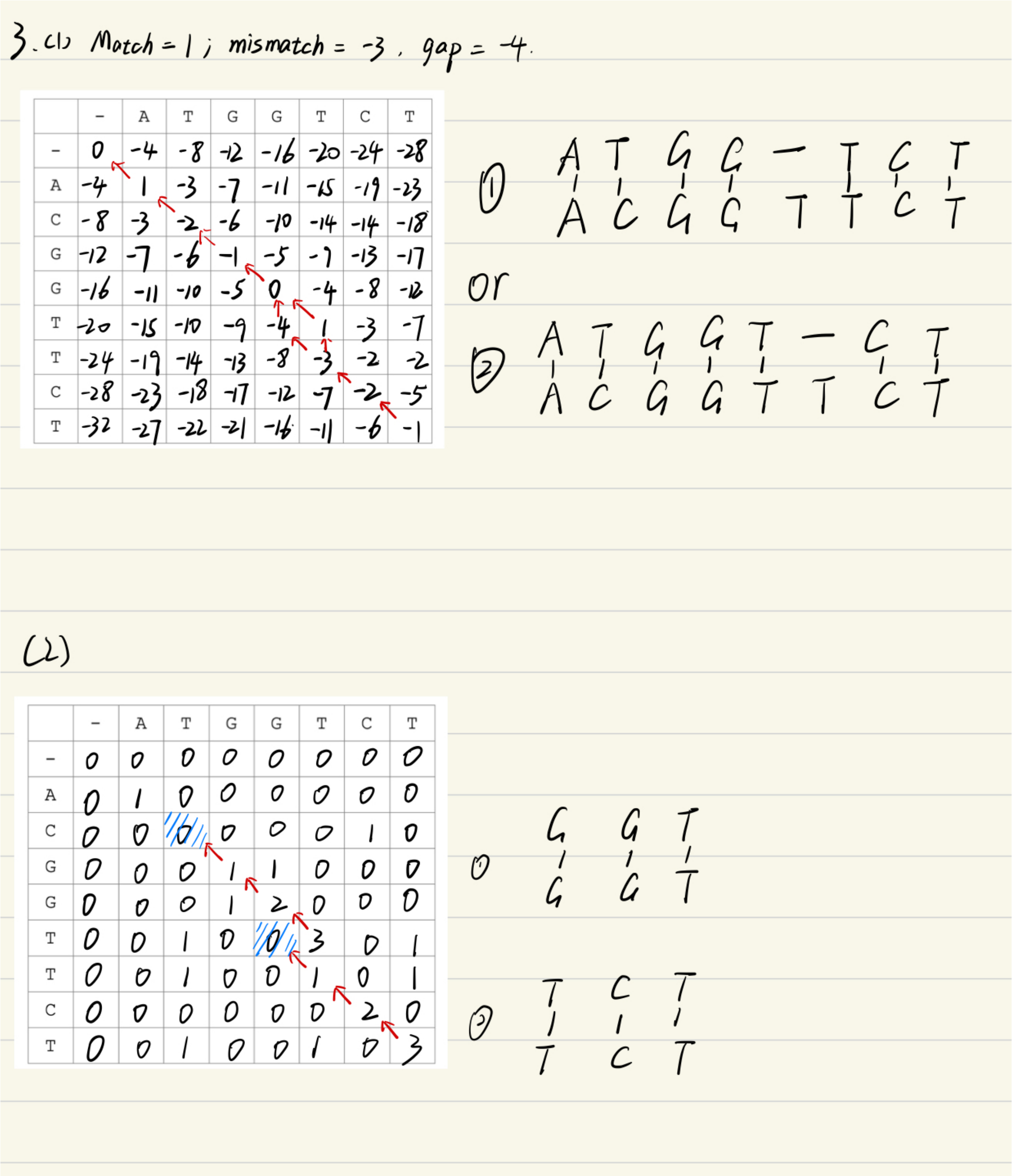
3 Question3

Pairwise global alignment. Suppose the alignment scoring function is the following: Match = 1; mismatch = -3, gap = -4.

Suppose the two DNA sequences are

ATGGTCT ACGGTTCT

- Align these two sequences using the Needleman-Wunsch algorithm manually by filling in the table below and show the optimal path. (3pts)
- Use Smith-Waterman algorithm to perform a local alignment and show the path. (3 pts)



3. (1) Match = 1 ; mismatch = -3 , gap = -4.

	-	A	T	G	G	T	C	T
-	0	-4	-8	-12	-16	-20	-24	-28
A	-4	1	-3	-7	-11	-15	-19	-23
C	-8	-3	-2	-6	-10	-14	-14	-18
G	-12	-7	-6	-1	-5	-9	-13	-17
G	-16	-11	-10	-5	0	-4	-8	-12
T	-20	-15	-10	-9	-4	1	-3	-7
T	-24	-19	-14	-13	-8	-3	-2	-2
C	-28	-23	-18	-17	-12	-7	-2	-5
T	-32	-27	-22	-21	-16	-11	-6	-1

```
①  A T G G - T C T
    | | | |   | | |
    A C G G T T C T
```

or

```
②  A T G G T - C T
    | | | | |   | |
    A C G G T T C T
```

(2)

	-	A	T	G	G	T	C	T
-	0	0	0	0	0	0	0	0
A	0	1	0	0	0	0	0	0
C	0	0	▨0	0	0	0	1	0
G	0	0	0	1	1	0	0	0
G	0	0	0	1	2	0	0	0
T	0	0	1	0	▨0	3	0	1
T	0	0	1	0	0	1	0	1
C	0	0	0	0	0	0	2	0
T	0	0	1	0	0	1	0	3

```
①  G G T
    | | |
    G G T
```

```
②  T C T
    | | |
    T C T
```

4 Question4

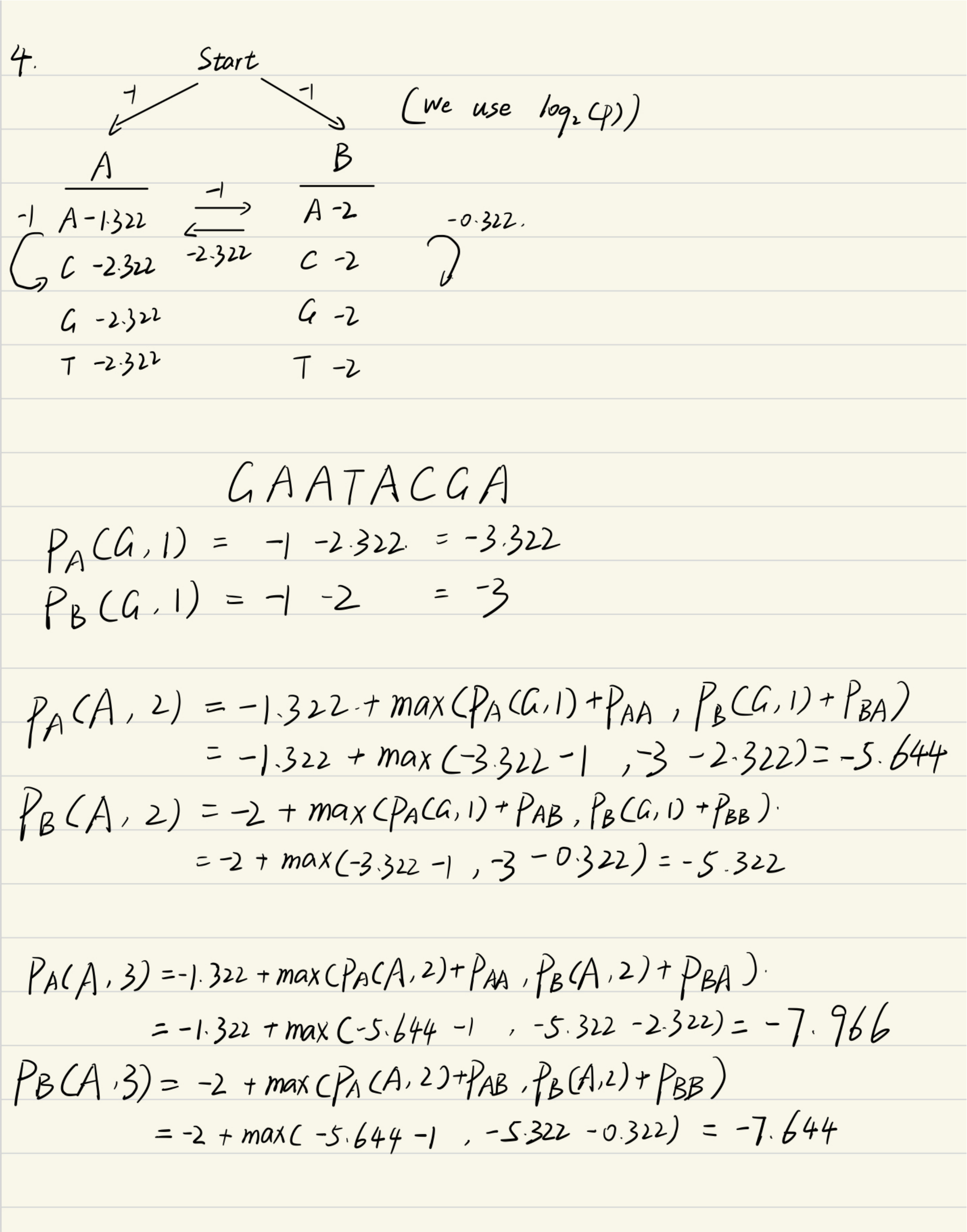
Hidden Markov model. Consider a two-state HMM: A (enrichment of A nucleotide) and B (background). The emission and transition probabilities are

A	B
A 0.4	A 0.25
C 0.2	C 0.25
G 0.2	G 0.25
T 0.2	T 0.25

	A	B
A	0.5	0.5
b	0.2	0.8

And the start probabilities for both states are 0.5.

Please infer the hidden states of sequence “GAATACGA” using the Viterbi algorithm. Please show your steps. (3 pts)



4.

Start

-1 -1

(we use $\log_2(p)$)

A

-1 A -1.322

C -2.322

G -2.322

T -2.322

-1

-2.322

B

A -2

C -2

G -2

T -2

-0.322.

GAATACGA

$P_A(G,1) = -1 -2.322 = -3.322$

$P_B(G,1) = -1 -2 = -3$

$P_A(A,2) = -1.322 + \max(P_A(G,1)+P_{AA}, P_B(G,1)+P_{BA})$

$= -1.322 + \max(-3.322 -1, -3 -2.322) = -5.644$

$P_B(A,2) = -2 + \max(P_A(G,1)+P_{AB}, P_B(G,1)+P_{BB})$

$= -2 + \max(-3.322 -1, -3 -0.322) = -5.322$

$P_A(A,3) = -1.322 + \max(P_A(A,2)+P_{AA}, P_B(A,2)+P_{BA})$

$= -1.322 + \max(-5.644 -1, -5.322 -2.322) = -7.966$

$P_B(A,3) = -2 + \max(P_A(A,2)+P_{AB}, P_B(A,2)+P_{BB})$

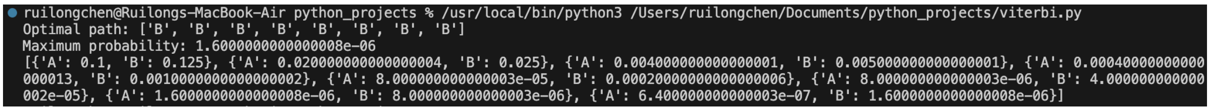
$= -2 + \max(-5.644 -1, -5.322 -0.322) = -7.644$

Sometimes the sequence is long. Calculating by hand is inconvenient. We can use python to solve this. Here is the code:

```
import numpy as np
```

```python

def viterbi(obs, states, start_p, trans_p, emit_p):
    # obs: observation sequence
    # states: hidden states
    # start_p: initial probabilities
    # trans_p: state transition probabilities
    # emit_p: emission probabilities
    # Initialize variables
    V = [{}]  # Viterbi table, V[t][s] represents the maximum probability of being in state s at time t
    path = {}  # To save the optimal path

    # Initialize the Viterbi table for t=0
    for s in states:
        V[0][s] = start_p[s] * emit_p[s][obs[0]]
        path[s] = [s]

    # Update the Viterbi table for each time step t
    for t in range(1, len(obs)):
        V.append({})
        new_path = {}

        for s in states:
            # Select the optimal previous state, meaning which previous state is most likely to result in the current observation and state
            (prob, state) = max((V[t-1][prev_state] * trans_p[prev_state][s] * emit_p[s][obs[t]], prev_state) for prev_state in states)
            V[t][s] = prob
            new_path[s] = path[state] + [s]

        path = new_path

    # Termination: Select the optimal path for the last time step
    (prob, state) = max((V[len(obs) - 1][s], s) for s in states)
    return prob, path[state], V

# Example data
states = ('A', 'B')
observations = ('G', 'A', 'A', 'T', 'A', 'C', 'G', 'A')
start_probability = {'A': 0.5, 'B': 0.5}
transition_probability = {
   'A': {'A': 0.5, 'B': 0.5},
   'B': {'A': 0.2, 'B': 0.8},
}
emission_probability = {
   'A': {'A': 0.4, 'C': 0.2, 'G': 0.2, 'T': 0.2},
   'B': {'A': 0.25, 'C': 0.25, 'G': 0.25, 'T': 0.25},
}

# Run the Viterbi algorithm
prob, optimal_path, V = viterbi(observations, states, start_probability, transition_probability, emission_probability)
print(f"Optimal path: {optimal_path}")
print(f"Maximum probability: {prob}")
print(V)
```

We get:



```
● ruilongchen@Ruilongs-MacBook-Air python_projects % /usr/local/bin/python3 /Users/ruilongchen/Documents/python_projects/viterbi.py
  Optimal path: ['B', 'B', 'B', 'B', 'B', 'B', 'B', 'B']
  Maximum probability: 1.6000000000000008e-06
  [{'A': 0.1, 'B': 0.125}, {'A': 0.020000000000000004, 'B': 0.025}, {'A': 0.004000000000000001, 'B': 0.005000000000000001}, {'A': 0.00040000000000000
  000013, 'B': 0.0010000000000000002}, {'A': 8.000000000000003e-05, 'B': 0.00020000000000000006}, {'A': 8.000000000000003e-06, 'B': 4.000000000000
  002e-05}, {'A': 1.6000000000000008e-06, 'B': 8.000000000000003e-06}, {'A': 6.400000000000003e-07, 'B': 1.6000000000000008e-06}]
```

	G	A	A	T	A	C	G	A
A	0.1	0.02	0.004	0.0004	8×10^{-5}	8×10^{-6}	1.6×10^{-6}	6.4×10^{-7}
B	0.125	0.025	0.005	0.001	0.0002	4×10^{-5}	8×10^{-6}	1.6×10^{-6}

The most probable path is: BBBBBBBB. The probability is: 1.6×10^{-6}**.**

5 Question5

CTCF (CCCTC-binding factor) is a zinc-finger protein that functions as a transcription factor. It also has insulator activity and is important for the 3D structure of chromatin, through formation of chromatin loops. Using the frequency matrix found at JASPAR using ID MA0139.1

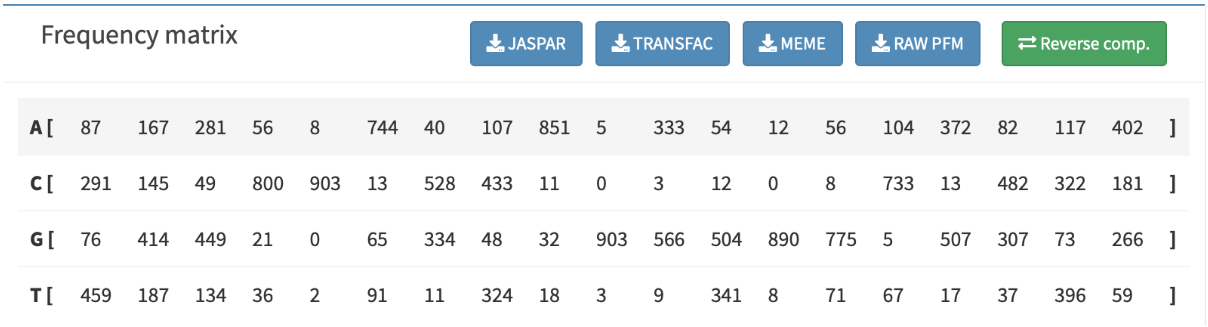
- Represent the CTCF motif using IUPAC code (assume a nucleotide is absent from a position if its proportion is less than 10%. (1 pt)

Read Stormo and Hartzell PNAS 1989 paper (can be found in the references folder under Files of the course Canvas website. Generate Fig 1 B, C and D using the CTCF frequency matrix from the above.

- Fig 1B. Derive the position-specific weight matrix (PSWM). (2 pts)
- Fig 1C. Derive the specific matrix. Note that the number "23" used in the calculation 0.5/23 when fb = 0 need to be modified for CTCF. You need to figure out what number should be used. Also use pb = 0.25 for all b. (2 pts)
- Fig 1D. Draw Iseq for CTCF (hand draw is fine). (2 pts)
- Fig 1D. What is the sum of all positions for CTCF (in bits)? (1 pt)

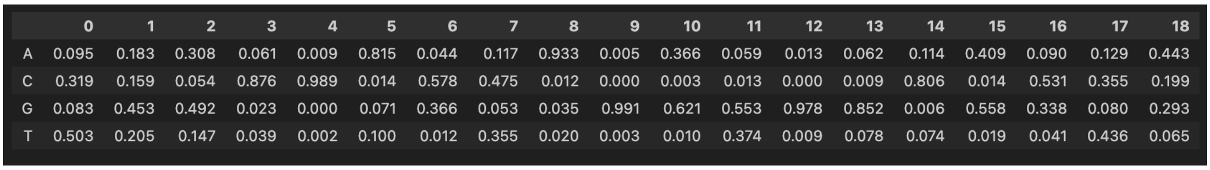
Simulate 10 CTCF motifs using the PSWM, i.e., generate 10 CTCF motif sequences. Put them in FASTA format, use weblogo web server http://weblogo.threeplusone.com/ to generate a logo plot. (1 pt)

We can find the frequency matrix at JASPAR:



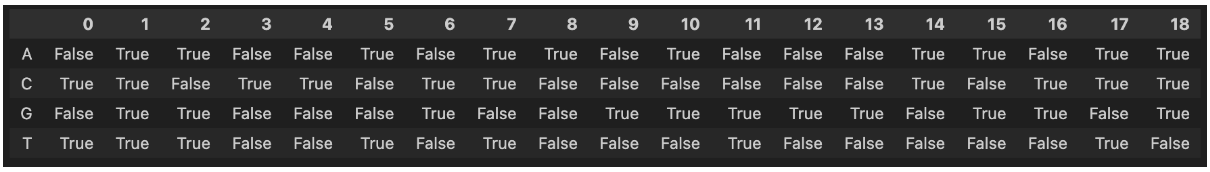
```
A [  87   167  281  56   8    744  40   107  851  5    333  54   12   56   104  372  82   117  402  ]
C [  291  145  49   800  903  13   528  433  11   0    3    12   0    8    733  13   482  322  181  ]
G [  76   414  449  21   0    65   334  48   32   903  566  504  890  775  5    507  307  73   266  ]
T [  459  187  134  36   2    91   11   324  18   3    9    341  8    71   67   17   37   396  59   ]
```

Then we can calculate the proportion, and assume a nucleotide is absent if its proportion is less than 10%



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
A	0.095	0.183	0.308	0.061	0.009	0.815	0.044	0.117	0.933	0.005	0.366	0.059	0.013	0.062	0.114	0.409	0.090	0.129	0.443
C	0.319	0.159	0.054	0.876	0.989	0.014	0.578	0.475	0.012	0.000	0.003	0.013	0.000	0.009	0.806	0.014	0.531	0.355	0.199
G	0.083	0.453	0.492	0.023	0.000	0.071	0.366	0.053	0.035	0.991	0.621	0.553	0.978	0.852	0.006	0.558	0.338	0.080	0.293
T	0.503	0.205	0.147	0.039	0.002	0.100	0.012	0.355	0.020	0.003	0.010	0.374	0.009	0.078	0.074	0.019	0.041	0.436	0.065

Figure 1: PSWM

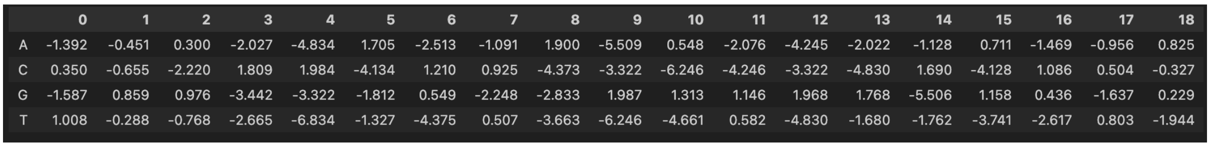


	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
A	False	True	True	False	False	True	False	True	True	False	True	False	False	False	True	True	False	True	True
C	True	True	False	True	True	False	True	True	False	False	False	False	False	False	True	False	True	True	True
G	False	True	True	False	False	False	True	False	False	True	True	True	True	True	False	True	True	False	True
T	True	True	True	False	False	True	False	True	False	False	False	True	False	False	False	False	False	True	False

Finally, we can use IUPAC code to represent the motif:

YNDCCWSHAGRKGGMRSHV

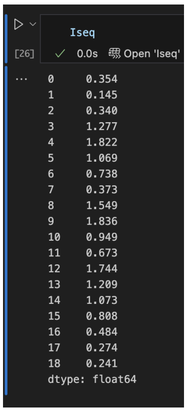
We have gained the PSWM, we can then calculate the specific matrix. At positions for which $f_b = 0$, we use 0.5/20 to estimate the frequency. The specific matrix is calculated as $log_2(f_b/p_b)$. We can get:



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
A	-1.392	-0.451	0.300	-2.027	-4.834	1.705	-2.513	-1.091	1.900	-5.509	0.548	-2.076	-4.245	-2.022	-1.128	0.711	-1.469	-0.956	0.825
C	0.350	-0.655	-2.220	1.809	1.984	-4.134	1.210	0.925	-4.373	-3.322	-6.246	-4.246	-3.322	-4.830	1.690	-4.128	1.086	0.504	-0.327
G	-1.587	0.859	0.976	-3.442	-3.322	-1.812	0.549	-2.248	-2.833	1.987	1.313	1.146	1.968	1.768	-5.506	1.158	0.436	-1.637	0.229
T	1.008	-0.288	-0.768	-2.665	-6.834	-1.327	-4.375	0.507	-3.663	-6.246	-4.661	0.582	-4.830	-1.680	-1.762	-3.741	-2.617	0.803	-1.944

Figure 2: specific matrix

Then, use $I_{seq} = \sum_{b=A}^{T} f_b log_2(\frac{f_b}{p_b})$ to calculate the I_{seq} and draw it:



```
Iseq
```

```
0     0.354
1     0.145
2     0.340
3     1.277
4     1.822
5     1.069
6     0.738
7     0.373
8     1.549
9     1.836
10    0.949
11    0.673
12    1.744
13    1.209
14    1.073
15    0.808
16    0.484
17    0.274
18    0.241
dtype: float64
```

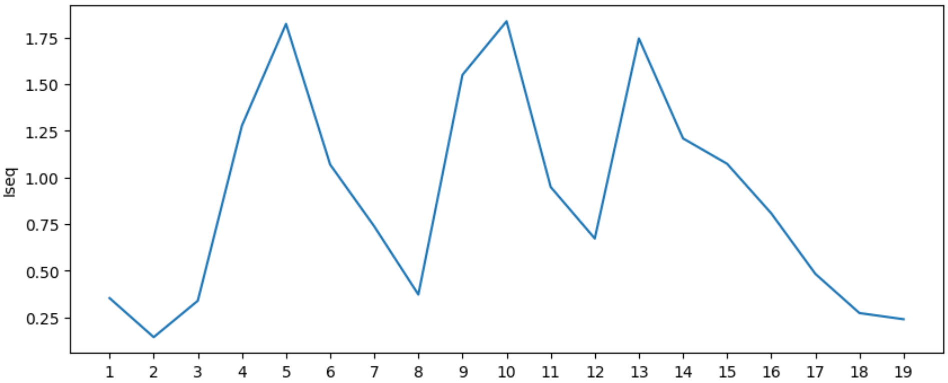


Figure 3: Iseq

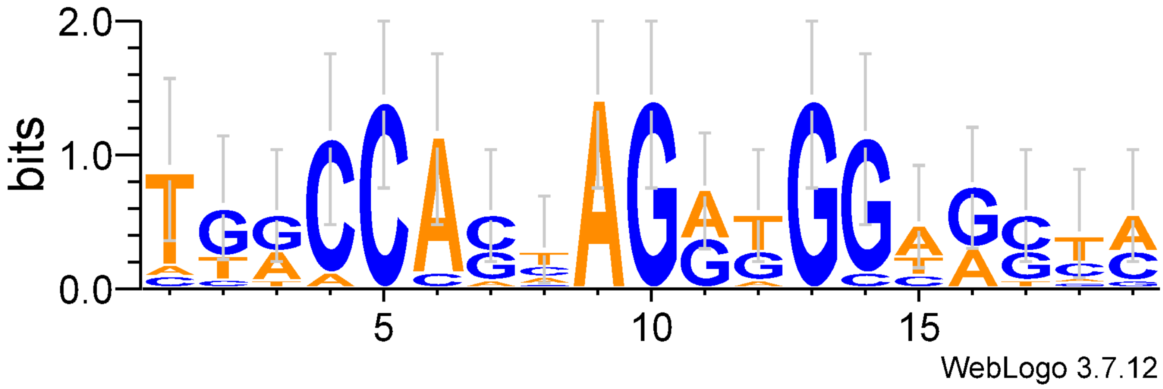
So, the sum of all positions is 16.96 bits.

We use python to do the simulation:

```python
import pandas as pd
import numpy as np
A = [87,167,281,56,8,744,40,107,851,5,333,54,12,56,104,372,82,117,402]
C = [291, 145, 49, 800, 903, 13, 528, 433, 11, 0 , 3, 12, 0 ,8, 733, 13, 482, 322,181]
G = [76, 414, 449 , 21, 0 ,65, 334, 48, 32, 903, 566, 504, 890, 775, 5, 507, 307, 73, 266]
T = [459, 187, 134, 36 , 2, 91,11,324,18,3,9,341,8,71,67,17,37,396,59]

df=pd.DataFrame({'A':A, 'C':C,'G':G,'T':T})
ctcf_pswm=df.divide(df.sum(axis=1),axis=0)

# Define the nucleotides and create a function to generate a sequence based on the PSWM
```

```python
nucleotides = ['A', 'C', 'G', 'T']

def generate_sequence(pswm, length):
    sequence = []
    for i in range(length):
        # Choose a nucleotide based on the PSWM probabilities for the current position
        nucleotide = np.random.choice(nucleotides, p=pswm.iloc[i])
        sequence.append(nucleotide)
    return ''.join(sequence)

# Simulate 10 motifs
motifs = [generate_sequence(ctcf_pswm, ctcf_pswm.shape[0]) for _ in range(10)]

# Write motifs to FASTA format
with open('ctcf_motifs.fasta', 'w') as fasta_file:
    for i, motif in enumerate(motifs):
        fasta_file.write(f">motif_{i+1}\n{motif}\n")
```

Put file into the website, we get:

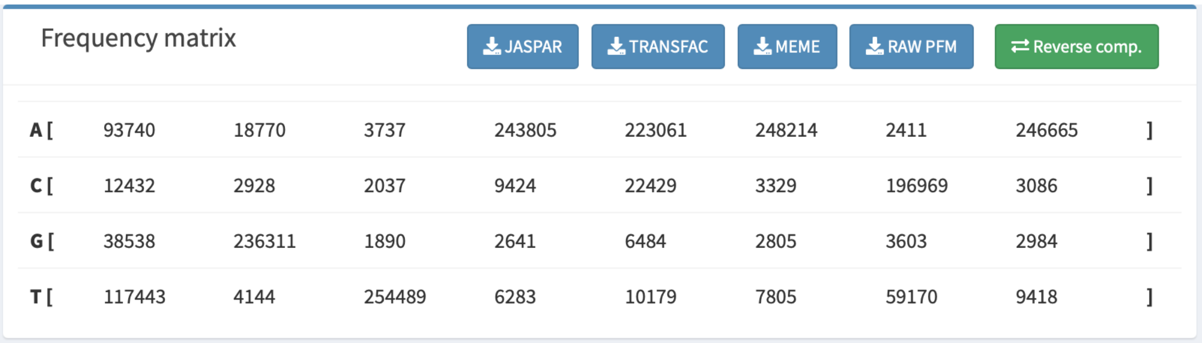


6 Question6

Locate the FOXA2 motif using JASPAR ID MA0047.4. Using the motif's PSWM to can the motif against the following sequence:

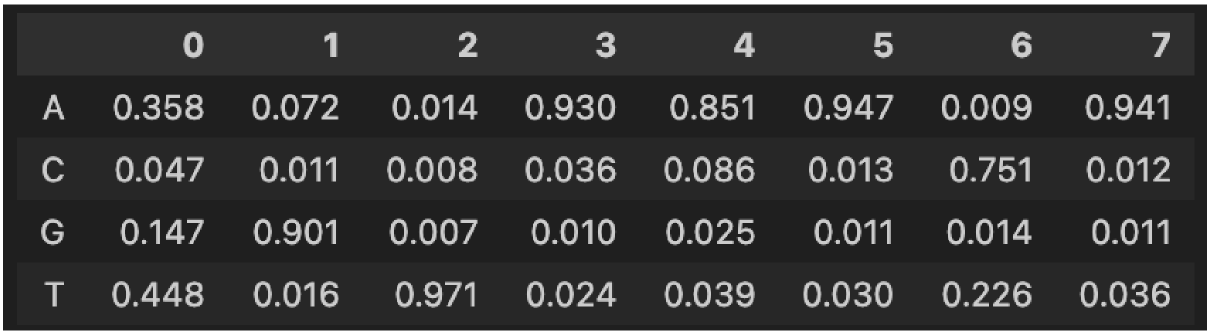
ACGTGCTAAG

Write down the matching probability for all possible motif start positions. Show your work. (3 pts).



Frequency matrix

A [	93740	18770	3737	243805	223061	248214	2411	246665	]
C [	12432	2928	2037	9424	22429	3329	196969	3086	]
G [	38538	236311	1890	2641	6484	2805	3603	2984	]
T [	117443	4144	254489	6283	10179	7805	59170	9418	]



	0	1	2	3	4	5	6	7
A	0.358	0.072	0.014	0.930	0.851	0.947	0.009	0.941
C	0.047	0.011	0.008	0.036	0.086	0.013	0.751	0.012
G	0.147	0.901	0.007	0.010	0.025	0.011	0.014	0.011
T	0.448	0.016	0.971	0.024	0.039	0.030	0.226	0.036

There are 3 start positions, denote as S_1,S_2,S_3.

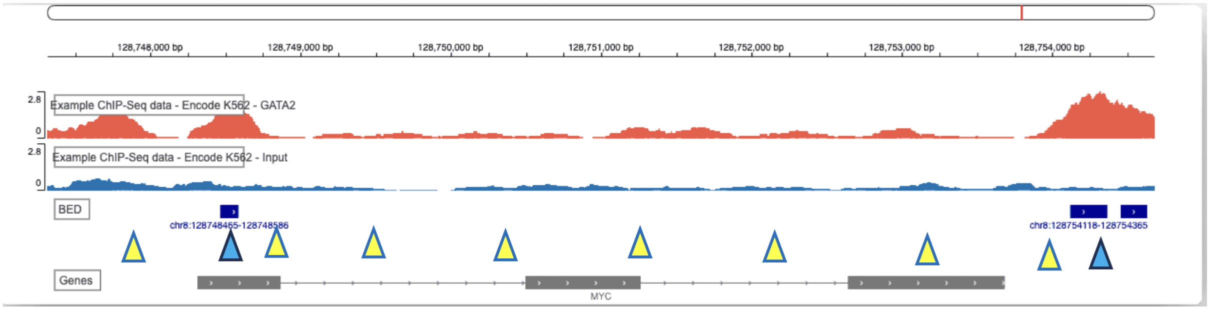
$P(S_1) = 0.358 * 0.011 * 0.007 * 0.024 * 0.025 * 0.013 * 0.226 * 0.941 \approx 4.57 * 10^{-11}$

$P(S_2) = 0.047 * 0.901 * 0.971 * 0.01 * 0.086 * 0.030 * 0.009 * 0.941 \approx 8.98 * 10^{-9}$

$P(S_3) = 0.147 * 0.016 * 0.007 * 0.036 * 0.039 * 0.947 * 0.009 * 0.011 \approx 2.17 * 10^{-12}$

7 Quesion7

Given the snapshot of called peaks from a TF ChIP-seq experiment in a part of the genome below. Suppose a colored triangle indicates a motif site. Please indicate which motif (color) is likely to be the binding site of the TF and explain why? (3 pts)



The blue triangle likely represents the binding site of the transcription factor because it is positioned directly under the highest ChIP-seq peak, which is indicative of strong TF-DNA binding activity.