Genome Informatics 2020 Midterm exam.

Problem 1.
Assume the DNA sequence below may be transcribed from either strand. Find any open reading frames (ORF：開放閱讀框) longer than 10. If you find any, write their amino acid sequences using the one letter code from the table at bottom.

`atttaattattcaatcgatgcagaatcagcagtcatcgacacgactactaaactgactagcatcaacgac`

	u	c	a	g
u.u	F	S	Y	C
c	F	|	Y	C
a	L	|	終	終
g	|	S	終	W
c.u	|	P	H	R
c	|	|	H	|
a	|	|	Q	|
g	L	P	Q	R
a.u	I	T	N	S
c	|	|	N	S
a	I	|	K	R
g	M	T	K	R
g.u	V	A	D	G
c	|	|	D	|
a	|	|	E	|
g	V	A	E	G

Genetic table in abbreviated form.
The three stop codons are `uaa`, `uag`, and `uga`; these also code for methionine.

Problem 2.
This question asked about a semi-global alignment, under affine gap scores.
Let g_o and g_e represent gap opening and gap extension scores. so, for example, that a gap of length 3 has a score of $g_o + 2g_e$.
Let x and y denote the sequences to align. The alignment requirement is to use all of y, but the start and end of x can be freely skipped. Let $S(a, b)$ represent the score of aligning character a with b in the same column (including the case when a and b match each other).

General Question
Let $x[i]$ denote the $i-1$th character of string x (so the first character is $x[0]$, like in computer programming). For general sequences x and y describe the following:

- The size and structure of the dynamic programming table D.
- The recursive relation used to update D.
- How D should be initialized.
- From what cell should trace-back start and when should trackback terminate.

Question given sequences x and y

Given the scoring system: Match +8, Mismatch -9; $g_o = -10$, $g_e = -1$. And the sequences:

$x =$ `ctctcaagttata`
$y =$ `tcgtt`

To compute the best semi-global alignment of these sequences.

- Fill in the DP table on the next page.
- Show which values are involved in the traceback
- Show the alignment

, , , , , , , , , , , ,
, , , , , , , , , , , ,
, , , , , , , , , , , ,
, , , , , , , , , , , ,
, , , , , , , , , , , ,
, , , , , , , , , , , ,
, , , , , , , , , , , ,
, , , , , , , , , , , ,
, , , , , , , , , , , ,
, , , , , , , , , , , ,
, , , , , , , , , , , ,

Problem 3.
Membrane spanning proteins are proteins in which part of the protein goes through a cellular membrane. The problem models that with a 2-state HMM. The two states are:

`Aqua`	Represents parts of the protein in aqueous solution (water)
`Memb`	Represents parts of the protein inside cellular membranes

Assume the HMM always starts in state `Aqua` and emits the initial M (Methionine). Then the model repeat cycles of (transition?, emit).
The following page shows computation for the single path with the maximum probability of generating the 150 amino acid sequence of the protein Glycophorin. The numbers are all lg ($\log_2$) probabilities.
From the numbers on the following page infer the emission probabilities of each amino acid in each state, and the transition probabilities between states. To make the problem easier I provide a table below to fill in.

	Lg Emission Probability			
State	-3	-4	-5	-6
`Aqua`	__	__ __ __ __ __ __ __ __ __ __	__ __ __ __ __ __ __	__ __
`Memb`	__ __ __ __	__ __ __ __	__ __ __ __	__ __ __ __ __ __ __ __

Lg Transition Probability

	`Aqua`	`Memb`
`Aqua`	____	____
`Memb`	____	____

In the top table giving lg emission probabilities each __ represents an amino acid. So you can tell immediately that exactly one of the amino acids in the `Aqua` state has a probability of $\frac{1}{8}$ (lg = -3). In the bottom table representing transition probabilities, each ___ represents a lg probability.

Maximum Probability Path lg Probabilities:

		Aqua	Memb			Aqua	Memb			Aqua	Memb
M	1	0	-∞	D	51	-202.0	-207.8	A	101	-416.0	-412.8
Y	2	-5.1	-8.9	T	52	-206.1	-209.9	G	102	-420.1	-416.9
G	3	-9.2	-13.0	Y	53	-211.2	-215.0	V	103	-424.2	-420.0
K	4	-13.3	-19.1	A	54	-215.3	-218.1	I	104	-427.9	-423.1
I	5	-17.4	-20.2	A	55	-219.4	-221.2	G	105	-431.0	-427.2
I	6	-21.5	-23.3	T	56	-223.5	-225.3	T	106	-435.1	-431.3
F	7	-26.6	-27.4	P	57	-227.6	-231.4	I	107	-439.2	-434.4
V	8	-30.7	-30.5	R	58	-232.7	-237.5	L	108	-442.3	-437.5
L	9	-34.8	-33.6	A	59	-236.8	-239.6	L	109	-445.4	-440.6
L	10	-38.9	-36.7	H	60	-241.9	-245.7	I	110	-448.5	-443.7
L	11	-43.0	-39.8	E	61	-246.0	-251.8	S	111	-450.6	-447.8
S	12	-46.1	-43.9	V	62	-250.1	-252.9	Y	112	-455.7	-452.9
E	13	-50.2	-50.0	S	63	-253.2	-257.0	G	113	-459.8	-457.0
I	14	-54.3	-53.1	E	64	-257.3	-263.1	I	114	-463.9	-460.1
V	15	-58.4	-56.2	I	65	-261.4	-264.2	R	115	-469.0	-466.2
S	16	-61.5	-60.3	S	66	-264.5	-268.3	R	116	-474.1	-472.3
I	17	-65.6	-63.4	V	67	-268.6	-271.4	L	117	-478.2	-475.4
S	18	-68.7	-67.5	R	68	-273.7	-277.5	I	118	-482.3	-478.5
A	19	-72.8	-70.6	T	69	-277.8	-281.6	K	119	-486.4	-484.6
S	20	-75.9	-74.7	V	70	-281.9	-284.7	K	120	-490.5	-490.7
S	21	-79.0	-78.8	Y	71	-287.0	-289.8	S	121	-493.6	-494.8
T	22	-83.1	-82.9	P	72	-291.1	-295.9	P	122	-497.7	-500.9
T	23	-87.2	-87.0	P	73	-295.2	-301.0	S	123	-500.8	-505.0
G	24	-91.3	-91.1	E	74	-299.3	-305.1	D	124	-504.9	-510.7
V	25	-95.4	-94.2	E	75	-303.4	-309.2	V	125	-509.0	-511.8
A	26	-99.5	-97.3	E	76	-307.5	-313.3	K	126	-513.1	-517.9
M	27	-104.6	-102.4	T	77	-311.6	-315.4	P	127	-517.2	-523.0
H	28	-109.7	-108.5	G	78	-315.7	-319.5	L	128	-521.3	-524.1
T	29	-113.8	-112.6	E	79	-319.8	-325.6	P	129	-525.4	-530.2
S	30	-116.9	-116.7	R	80	-324.9	-329.7	S	130	-528.5	-533.3
T	31	-121.0	-120.8	V	81	-329.0	-331.8	P	131	-532.6	-538.4
S	32	-124.1	-124.9	Q	82	-334.1	-337.9	D	132	-536.7	-542.5
S	33	-127.2	-129.0	L	83	-338.2	-341.0	T	133	-540.8	-544.6
S	34	-130.3	-133.1	A	84	-342.3	-344.1	D	134	-544.9	-550.7
V	35	-134.4	-136.2	H	85	-347.4	-350.2	V	135	-549.0	-551.8
T	36	-138.5	-140.3	H	86	-352.5	-356.3	P	136	-553.1	-557.9
K	37	-142.6	-146.4	F	87	-357.6	-360.4	L	137	-557.2	-560.0
S	38	-145.7	-150.5	S	88	-360.7	-364.5	S	138	-560.3	-564.1
Y	39	-150.8	-154.6	E	89	-364.8	-370.6	S	139	-563.4	-568.2
I	40	-154.9	-157.7	P	90	-368.9	-374.7	V	140	-567.5	-570.3
S	41	-158.0	-161.8	E	91	-373.0	-378.8	E	141	-571.6	-576.4
S	42	-161.1	-165.9	I	92	-377.1	-379.9	I	142	-575.7	-578.5
Q	43	-166.2	-171.0	T	93	-381.2	-384.0	E	143	-579.8	-584.6
T	44	-170.3	-174.1	L	94	-385.3	-387.1	N	144	-584.9	-589.7
N	45	-175.4	-180.2	I	95	-389.4	-390.2	P	145	-589.0	-594.8
D	46	-179.5	-185.3	I	96	-393.5	-393.3	E	146	-593.1	-598.9
T	47	-183.6	-187.4	F	97	-398.6	-397.4	T	147	-597.2	-601.0
H	48	-188.7	-193.5	G	98	-402.7	-401.5	S	148	-600.3	-605.1
K	49	-192.8	-198.6	V	99	-406.8	-404.6	D	149	-604.4	-610.2
R	50	-197.9	-202.7	M	100	-411.9	-409.7	Q	150	-609.5	-614.3