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
```

Solution:

```
                  M  Q  N  Q  Q  S  S  T  R  L  L  N  *
atttaattattcaatcgatgcagaatcagcagtcatcgacacgactactaaactgactagcatcaacgac
taaattaataagttagctacgtcttagtcgtcagtagctgtgctgatgatttgactgatcgtagttgctg
      *  E  I  S  A  S  D  A  T  M  S  V  V  V  L  S  V  L  M
```

	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

Solution:

- D should be a table with cell $D(i, j)$ representing the alignment of the first i characters of x against the first j characters of y; for $i \in [0, l_x]$ and $j \in [0, l_y]$ where l_x, l_y denote the lengths of x and y respectively. Each cell has three parts; D_M for alignments aligning two characters together, and D_X, D_Y for alignments ending in a gap in sequence x or y respectively.

- The general recursion should look like:

$$D_M(i+1,j+1) = S(x[i],y[j]) + max \begin{cases} D_M(i,j) \\ D_X(i,j) \\ D_Y(i,j) \end{cases}$$

$$D_X(i+1,j+1) = max \begin{cases} D_X(i+1,j) + g_e \\ D_M(i+1,j) + g_o \\ D_Y(i+1,j) + g_o \end{cases}$$

$$D_Y(i+1,j+1) = max \begin{cases} D_Y(i,j+1) + g_e \\ D_M(i,j+1) + g_o \\ D_X(i,j+1) + g_o \end{cases}$$

- Set $D_M(0,0) = 0$ to represent an empty alignment. Since the ends of x may be skipped, $D_X(i,0)$ can be initialized to 0 for all i. Opening gaps in y however are to be penalized, so we initialize $D_Y(0,1) = d$ then $D_Y(0,j) = e + D_Y(0,j-1)$ for $j \in [2, l_x]$. To prevent interference, other cells in the first row or column of the table can be set to $-\infty$.

- Since the end of x, but not y can be skipped, traceback should start from the cell in the final column of D with the best score. In other words: $\arg\max_{i,S \in M,X,Y} D_S(i, l_y)$. Traceback should terminate when all of y has been aligned.

Solution:

*	–	t	c	g	t	t
|	---, +0, -9	---,---,-10	---,---,-11	---,---,-12	---,---,-13	---,---,-14
c	+0,---,---	-20, -9,-10	-21, -2,-11	-22,-20,-12	-23,-21,-13	-24,-22,-14
t	+0,---,---	-19, +8,-10	-12,-18, -2	-22,-11, -3	-23, -4, -4	-24, -5, -5
c	+0,---,---	-2, -9,-10	-12,+16,-11	-13,-11, +6	-14,-12, +5	-15,-13, +4
t	+0,---,---	-3, +8,-10	+6,-11, -2	-4, +7, -3	-5,+14, -3	-6,+13, +4
c	+0,---,---	-2, -9,-10	+5,+16,-11	-3, -3, +6	+4, -2, +5	+3, +5, +4
a	+0,---,---	-3, -9,-10	+6,-11,-11	-4, +7, -4	+3, -3, -3	+2, -4, -4
a	+0,---,---	-4, -9,-10	+5,-12,-11	-3, -3, -5	+2, -2, -6	+1, -6, -7
g	+0,---,---	-5, -9,-10	+4,-13,-11	-4,+13, -6	+1,-12, +3	+0, -7, +2
t	+0,---,---	-6, +8,-10	+3,-14, -2	+3, -5, -3	+0,+21, -4	-1,+11,+11
t	+0,---,---	-2, +8,-10	+2, -1, -2	+2, -6, -3	+11,+11, -4	+1,+29, +1
a	+0,---,---	-2, -9,-10	+1, -1,-11	+1, -7, -9	+10, -7, -9	+19, +2, +0
t	+0,---,---	-3, +8,-10	+0,-11, -2	+0, -8, -3	+9, +9, -4	+18,+18, -1
a	+0,---,---	-2, -9,-10	-1, -1,-11	-1, -9,-11	+8, -9,-11	+17, +0, -2

Alignment:
```
tcaagtt
tc--gtt
```

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

State	Lg Emission Probability -3	-4	-5	-6
Aqua	S	AEDGILKPTV	RNQHMFY	CW
Memb	AILV	GFST	CMWY	RNEQDHK

	Lg Transition Probability Aqua	Memb
Aqua	-0.1	-3.9
Memb	-3.9	-0.1

How to Solve

This problem tests basic knowledge of the recursive computation for finding a maximum likelihood path. But is also a sort of puzzle. Some students were able to deduce the answer, probably by heuristically guessing the transition probabilibies based on frequently seen differences in the table given below and then confirming those numbers work.
For completeness, I outline a more methodical solution here. Let $T_{\mathtt{aa}}$, $T_{\mathtt{am}}$, $T_{\mathtt{ma}}$ and $T_{\mathtt{mm}}$ denote the lg transition probabilities, and $E(b)_{\mathtt{a}}$, and $E(b)_m$ denote the emission probabilities of amino acid b in states Aqua and Memb respectively.
From the probabilities of paths leading to Y2 we have:

$$-5.1 = T_{\mathtt{aa}} + E(\mathtt{Y})_{\mathtt{a}}$$
$$-8.9 = T_{\mathtt{am}} + E(\mathtt{Y})_{\mathtt{m}}$$

Fortunately the problem statement tells us all emission probabilitie are $\in \{-3, -4, -5, -6\}$, and we know the transition lg probabilities must be ≤ 0.
So $(T_{\mathtt{aa}}, E(\mathtt{Y})_{\mathtt{a}})$ must be one of: {(-2.1,-3), (-1.1,-4), (-0.1,-5)}
and $(T_{\mathtt{am}}, E(\mathtt{Y})_{\mathtt{m}})$ must be one of: {(-5.9,-3),(-4.9,-4),(-3.9,-5),(-2.9,-6)}.
Recalling that $\lg 0.5 = -1$, only one of $\{T_{\mathtt{aa}}, T_{\mathtt{am}}\}$ can be less than -1, so $T_{\mathtt{aa}} = -0.1$, and $E(\mathtt{Y})_{\mathtt{a}} = -5$

Having deduced $T_{\mathtt{aa}} = -0.1$, Students with a calculator could then solve for $T_{\mathtt{am}}$.

$2^{T_{\mathtt{aa}}} + 2^{T_{\mathtt{am}}} = 1, \;\; 2^{T_{\mathtt{am}}} = 1 - 2^{T_{\mathtt{aa}}}$
$T_{\mathtt{am}} = \lg(1 - 2^{T_{\mathtt{aa}}}) = \lg(1 - 2^{-0.1}) \approx \lg(1 - 0.933) = \lg(0.067) \approx -3.8997 \approx -3.9$

But actually students were not allowed to have calculators, so let's look for another way.

Going to the next amino acid G3 ending in state aqua, we obtain:

$$-9.2 = \max\{-5.1 + T_{\mathtt{aa}} + E(\mathtt{G})_{\mathtt{a}},\; -8.9 + T_{\mathtt{ma}} + E(\mathtt{G})_{\mathtt{m}}\} \tag{1}$$

Since it is given that all emission probabilities are $\leqq -3$, the maximum must come from the first term, i.e.

$$-9.2 = -5.1 + T_{\mathtt{aa}} + E(\mathtt{G})_{\mathtt{a}}, \;\; E(\mathtt{G})_{\mathtt{a}} = -9.2 + 5.1 - T_{\mathtt{aa}} = -9.2 + 5.1 + 0.1 = -4$$

Etc. Moving forward with this is quite tedious, but does give us the idea that the differences in lg probability obtained when emitting an additional amin acid should be the key to solving this somewhat tricky problem. So on the next page let's look at the maximum probabilities lg probabilities again, but listing differences as well.

Maximum Probability Path **negative** lg Probabilities, with differences from previous row.

M001	0		∞		D051	202.0	4.1	207.8	5.1	A101	416.0	4.1	412.8	3.1
Y002	5.1	5.1	8.9	∞	T052	206.1	4.1	209.9	2.1	G102	420.1	4.1	416.9	4.1
G003	9.2	4.1	13.0	4.1	Y053	211.2	5.1	215.0	5.1	V103	424.2	4.1	420.0	3.1
K004	13.3	4.1	19.1	6.1	A054	215.3	4.1	218.1	3.1	I104	427.9	3.7	423.1	3.1
I005	17.4	4.1	20.2	1.1	A 55	219.4	4.1	221.2	3.1	G105	431.0	3.1	427.2	4.1
I006	21.5	4.1	23.3	3.1	T056	223.5	4.1	225.3	4.1	T106	435.1	4.1	431.3	4.1
F007	26.6	5.1	27.4	4.1	P057	227.6	4.1	231.4	6.1	I107	439.2	4.1	434.4	3.1
V008	30.7	4.1	30.5	3.1	R058	232.7	5.1	237.5	6.1	L108	442.3	3.1	437.5	3.1
L009	34.8	4.1	33.6	3.1	A059	236.8	4.1	239.6	2.1	L109	445.4	3.1	440.6	3.1
L010	38.9	4.1	36.7	3.1	H060	241.9	5.1	245.7	6.1	I110	448.5	3.1	443.7	3.1
L011	43.0	4.1	39.8	3.1	E061	246.0	4.1	251.8	6.1	S111	450.6	2.1	447.8	4.1
S012	46.1	3.1	43.9	4.1	V062	250.1	4.1	252.9	1.1	Y112	455.7	5.1	452.9	5.1
E013	50.2	4.1	50.0	6.1	S063	253.2	3.1	257.0	4.1	G113	459.8	4.1	457.0	4.1
I014	54.3	4.1	53.1	3.1	E064	257.3	4.1	263.1	6.1	I114	463.9	4.1	460.1	3.1
V015	58.4	4.1	56.2	3.1	I065	261.4	4.1	264.2	1.1	R115	469.0	5.1	466.2	6.1
S016	61.5	3.1	60.3	4.1	S066	264.5	3.1	268.3	4.1	R116	474.1	5.1	472.3	6.1
I017	65.6	4.1	63.4	3.1	V067	268.6	4.1	271.4	3.1	L117	478.2	4.1	475.4	3.1
S018	68.7	3.1	67.5	4.1	R068	273.7	5.1	277.5	6.1	I118	482.3	4.1	478.5	3.1
A019	72.8	4.1	70.6	3.1	T069	277.8	4.1	281.6	4.1	K119	486.4	4.1	484.6	6.1
S020	75.9	3.1	74.7	4.1	V070	281.9	4.1	284.7	3.1	K120	490.5	4.1	490.7	6.1
S021	79.0	3.1	78.8	4.1	Y071	287.0	5.1	289.8	5.1	S121	493.6	3.1	494.8	4.1
T022	83.1	4.1	82.9	4.1	P072	291.1	4.1	295.9	6.1	P122	497.7	4.1	500.9	6.1
T023	87.2	4.1	87.0	4.1	P073	295.2	4.1	301.0	5.1	S123	500.8	3.1	505.0	4.1
G024	91.3	4.1	91.1	4.1	E074	299.3	4.1	305.1	4.1	D124	504.9	4.1	510.7	5.7
V025	95.4	4.1	94.2	3.1	E075	303.4	4.1	309.2	4.1	V125	509.0	4.1	511.8	1.1
A026	99.5	4.1	97.3	3.1	E076	307.5	4.1	313.3	4.1	K126	513.1	4.1	517.9	6.1
M027	104.6	5.1	102.4	5.1	T077	311.6	4.1	315.4	2.1	P127	517.2	4.1	523.0	5.1
H028	109.7	5.1	108.5	6.1	G078	315.7	4.1	319.5	4.1	L128	521.3	4.1	524.1	1.1
T029	113.8	4.1	112.6	4.1	E079	319.8	4.1	325.6	6.1	P129	525.4	4.1	530.2	6.1
S030	116.9	3.1	116.7	4.1	R080	324.9	5.1	329.7	4.1	S130	528.5	3.1	533.3	3.1
T031	121.0	4.1	120.8	4.1	V081	329.0	4.1	331.8	2.1	P131	532.6	4.1	538.4	5.1
S032	124.1	3.1	124.9	4.1	Q082	334.1	5.1	337.9	6.1	D132	536.7	4.1	542.5	4.1
S033	127.2	3.1	129.0	4.1	L083	338.2	4.1	341.0	3.1	T133	540.8	4.1	544.6	2.1
S034	130.3	3.1	133.1	4.1	A084	342.3	4.1	344.1	3.1	D134	544.9	4.1	550.7	6.1
V035	134.4	4.1	136.2	3.1	H085	347.4	5.1	350.2	6.1	V135	549.0	4.1	551.8	1.1
T036	138.5	4.1	140.3	4.1	H086	352.5	5.1	356.3	6.1	P136	553.1	4.1	557.9	6.1
K037	142.6	4.1	146.4	6.1	F087	357.6	5.1	360.4	4.1	L137	557.2	4.1	560.0	2.1
S038	145.7	3.1	150.5	4.1	S088	360.7	3.1	364.5	4.1	S138	560.3	3.1	564.1	4.1
Y039	150.8	5.1	154.6	4.1	E089	364.8	4.1	370.6	6.1	S139	563.4	3.1	568.2	4.1
I040	154.9	4.1	157.7	3.1	P090	368.9	4.1	374.7	4.1	V140	567.5	4.1	570.3	2.1
S041	158.0	3.1	161.8	4.1	E091	373.0	4.1	378.8	4.1	E141	571.6	4.1	576.4	6.1
S042	161.1	3.1	165.9	4.1	I092	377.1	4.1	379.9	1.1	I142	575.7	4.1	578.5	2.1
Q043	166.2	5.1	171.0	5.1	T093	381.2	4.1	384.0	4.1	E143	579.8	4.1	584.6	6.1
T044	170.3	4.1	174.1	3.1	L094	385.3	4.1	387.1	3.1	N144	584.9	5.1	589.7	5.1
N045	175.4	5.1	180.2	6.1	I095	389.4	4.1	390.2	3.1	P145	589.0	4.1	594.8	5.1
D046	179.5	4.1	185.3	5.1	I096	393.5	4.1	393.3	3.1	E146	593.1	4.1	598.9	4.1
T047	183.6	4.1	187.4	2.1	F 97	398.6	5.1	397.4	4.1	T147	597.2	4.1	601.0	2.1
H048	188.7	5.1	193.5	6.1	G098	402.7	4.1	401.5	4.1	S148	600.3	3.1	605.1	4.1
K049	192.8	4.1	198.6	5.1	V099	406.8	4.1	404.6	3.1	D149	604.4	4.1	610.2	5.1
R050	197.9	5.1	202.7	4.1	M100	411.9	5.1	409.7	5.1	Q150	609.5	5.1	614.3	4.1

The brown entries have a difference of less than 3, therefore we can deduce that they must be the result of a state transition. For example the 1.1 in row I005 implies:

$$\begin{aligned} -20.2 &= -13.3 + T_{\texttt{am}} + E(\texttt{I})_{\texttt{m}} \\ T_{\texttt{am}} &= -20.2 + 13.3 - E(\texttt{I})_{\texttt{m}} \\ &= -6.9 - E(\texttt{I})_{\texttt{m}} \end{aligned}$$

So if $E(\texttt{I})_{\texttt{m}} = -3$, then $T_{\texttt{am}} = -3.9$. To see if $E(\texttt{I})_{\texttt{m}}$ really is -3, we might look for a sequence position with amino acid `I`, in which the previous row has a larger lg probability for `membr` than `aqua`, for example `I`104,`membr` = -423.1 can be traced back to `V`103,`membr` = -420.0, since that is a larger lg probability than `V`103,`aqua` = -424.2. So:

$$\begin{aligned} -423.1 &= -420.0 + T_{\texttt{mm}} + E(\texttt{I})_{\texttt{m}} \\ T_{\texttt{mm}} + E(\texttt{I})_{\texttt{m}} &= -3.1 \end{aligned}$$

Since we know $T_{\texttt{mm}} < 0$ and $E(\texttt{I})_{\texttt{m}} \leqq -3$, we can deduce that indeed $E(\texttt{I})_{\texttt{m}} = -3$ and $T_{\texttt{mm}} = -0.1$; and having confirmed $E(\texttt{I})_{\texttt{m}} = -3$, we know that $T_{\texttt{am}} = -3.9$.
As this point we know all the transition probabilities, so we can pretty easily fill in the lg emission probabilities by looking at differences in lg probability of the maximum likelihood path when adding single residues. The final complication is that this particular protein sequence contains no cysteine `C` or tryptophan `W`; fortunately it turns out that `C` and `W` happen to have the same lg emission probabilities (-6 for the `aqua` state and -5 for the `membr` state), so those can be deduced by elimination after deducing the emission probabilities of the other 18 amino acids.