Midterm exam. Genome Informatics, Spring 2021. Closed book, calculators but no cell phones allowed. Answers may include e^2, $\sqrt{2}$, etc. but simplify when possible.

Problem 1.

Consider the following sequences:

1. `MTSP*`
2. `HVYHAADYDH*`
3. `aggacgtatcgacg`
4. `augacuucucccauu`
5. `augacuucucccc`
6. `augatuututttt`
7. `tactgaagagggtaa`

DNA sequence ____**7**____ is transcribed to produce mRNA sequence ____**4**____, which is then translated to amino acid sequence ____**1**____.

(Fill in each blank with numbers 1-7 from the table above.)

Problem 2.

A. hydrophobicity
B. agoraphobia
C. hydrogen bonding
D. bail bonds
E. James Bond

____**C**____ is very important in understanding the structure of both nucleic acids (DNA,RNA) and proteins, while ____**A**____ is especially relevant to understanding protein structure.

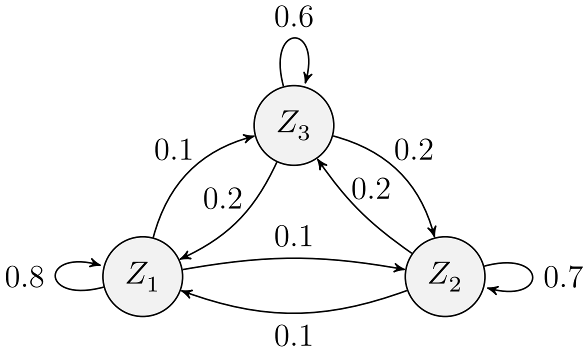
Problem 3.

A. hypergeometric cube
B. epigenetic state
C. epicurean doctrine
D. genome sequence
E. Fibonacci sequence

During cell division ____**B**____ is usually approximately copied to the daughter cells, along with ____**D**____ which is copied with high fidelity (i.e. very few errors). Abnormalities in ____**D**____ are typically responsible for inherited disease, while abnormalities in both ____**D**____ and ____**B**____ play a major role in cancer.

(You may use the same choice multiple times.)

Problem 4.



	Z_1	Z_2	Z_3
H	0.8	0.2	0.5
T	0.2	0.8	0.5

Emission probabilities of {H,T} for the 3 states.

Suppose the model *always* starts in Z_1, i.e. $P[S_1 = Z_1] = 1$, and output sequence $X =$ HHHTTT.

Define: $\alpha_{ij} \overset{\text{def}}{=} P[X_{1..i}, S_i = Z_j \mid \lambda]$ $\beta_{ij} \overset{\text{def}}{=} P[X_{i+1..n} | S_i = Z_j, \lambda]$
with λ meaning the HMM model and its parameter values described above.

Task: Fill in the values missing from this table.

i	1	2	3	4	5	6
X	H	H	H	T	T	T
α_{i1}	0.8000	0.5120	0.3354	0.0555	0.0109	0.0030
α_{i2}	0.0000	0.0160	0.0141	0.0410	0.0322	0.0214
α_{i3}	0.0000	0.0400	0.0392	0.0299	0.0159	0.0085
β_{i1}	0.0412	0.0572	0.0686	0.1258	0.2900	1.0000
β_{i2}	0.0198	0.0595	0.2741	0.4366	0.6800	1.0000
β_{i3}	0.0321	0.0687	0.1560	0.2704	0.5000	1.0000

Question. Given $X =$ HHHTTT, what is the probability that the model emitted the 4th character 'T', from state Z_2?

Solution: The missing values in the table can be computed using the recurrences for α_{ij} and β_{ij} given at top.

$$P[S_4 = 2 | x_1 x_2 x_3 x_4 x_5 x_6] = \frac{\alpha_{42}\beta_{42}}{P[x_1 x_2 x_3 x_4 x_5 x_6]} = \frac{\alpha_{42}\beta_{42}}{\alpha_{61} + \alpha_{62} + \alpha_{63}} \approx \frac{(0.041)(0.4366)}{0.003 + 0.0214 + 0.0085} \approx 0.5441$$

Continued on other side...

i	1	2	3	4	5	6
X	H	H	H	T	T	T
Z_1	0.8000	0.5120	0.3277	0.0524	0.0084	0.0013
Z_2	0.0000	0.0160	0.0102	0.0262	0.0147	0.0082
Z_3	0.0000	0.0400	0.0256	0.0164	0.0049	0.0015

Question. The values in the table above are probabilities computed using the input X and the parameter values for the HMM shown on the other side of this page.
Guess what probability the numbers in the table represent and give the general mathematical formula for that quantatity. e.g. your answer should look something like:

$$M[i,j] \quad = \quad \text{something} \;\; \text{P[some formula or statement including } i \text{ and } j)]$$

Solution: The probabilities are those computed for so called Viterbi decoding.

$$M[i,j] = \max_{S_1 \ldots S_{i-1}} \mathrm{P}[X_1 \ldots X_i, S_1 \ldots S_{i-1}, S_i = j]$$

Contrast this to α_{ij} which can be defined as:

$$\alpha_{ij} = \sum_{S_1 \ldots S_{i-1}} \mathrm{P}[X_1 \ldots X_i, S_1 \ldots S_{i-1}, S_i = j]$$

Problem 5.
Consider aligning 2 sequences $x = x_1...x_n$ and $y = y_1...y_m$. Let $s(x_i, y_j)$ denote the score of aligning character x_i to character y_j. Let d denote the gap opening cost, and $e < d$ the gap extension cost, so that a gap of length l costs $d + (l-1)e$.
You may assume that for any character pair (a,b): $s(a,b) > -2e$

Task: Describe dynamic programming to align $x = x_1...x_n$ and $y = y_1...y_m$ under two variations regarding what kind of alignments are desired.
In each case you should use one or more matrix, and precisely describe
0: The matrix(s) (i.e. dynamic programming table) to use and their size
1: The recursion formula(s)
2. Initialization (i.e. base case of recursion, or boundary condition)
3. How to compute the alignment score

Case 1. Global alignment of x and y

For example the alignment:
```
aaaab
a---b
```
would have a score of: $s(\texttt{a}, \texttt{a}) + s(b, \texttt{b}) - d - 2e$

Solution: The BSA textbook gives a solution with separate insertion matrices I_x, I_y (page 29) and then later the one below with a single insert matrix (page 31). Either is acceptable, but here I adopt the simpler one.
Matrices: Matrices M and I, both of size $n + 1 \times m + 1$
Recursion:

$$M(i,j) = \max \begin{cases} M(i-1, j-1) + s(x_i, y_j) \\ I(i-1, j-1) + s(x_i, y_j) \end{cases}$$

$$I(i,j) = \max \begin{cases} M(i, j-1) - d \\ M(i-1, j) - d \\ I(i, j-1) - e \\ I(j-1, i) - e \end{cases}$$

Initialization:
$\forall i \in \{1, ..., n\}$ $M(i, 0) \longleftarrow -\infty$, $I(i, 0) \longleftarrow -d - (i-1)e$
$\forall j \in \{1, ..., m\}$ $M(0, j) \longleftarrow -\infty$, $I(0, j) \longleftarrow -d - (j-1)e$
$M(0,0) \longleftarrow 0$, $I(0,0) \longleftarrow -\infty$
Alignment score: $\max \{M(n,m),\ I(n,m)\}$.
Note: the best alignment may end in an indel.

Continued on other side...

Case 2. Alignment of part of x to all of y.

In other words, the best possible global alignment score of y to a substring of x.

For example:

```
--babb--ab-
aaaabbbaaba
```

would have a score of: $2s(\mathtt{a},\mathtt{a}) + s(\mathtt{a}, b) + 3s(\mathtt{b},\mathtt{b}) - d - e$

Solution: Matrices: Matrices M and I, both of size $n+1 \times m+1$
Recursion:

$$M(i,j) = \max \begin{cases} M(i-1,j-1) + s(x_i, y_j) \\ I(i-1,j-1) + s(x_i, y_j) \end{cases}$$

$$I(i,j) = \max \begin{cases} M(i,j-1) - d \\ M(i-1,j) - d \\ I(i,j-1) - e \\ I(j-1,i) - e \end{cases}$$

Initialization:
$\forall i \in \{1, ..., n\}\ \ M(i,0) \longleftarrow 0,\ \ I(i,0) \longleftarrow -\infty$
$\forall j \in \{1, ..., m\}\ M(0,j) \longleftarrow -\infty,\ \ I(0,j)\ \longleftarrow -d-(j-1)e$
$M(0,0) \longleftarrow 0,\ \ I(0,0) \longleftarrow -\infty$
Alignment score: $\max_{0<i\leqq n} \{M(i,m), I(i,m)\}$.
Note: $I(i,0) \longleftarrow -\infty$ dissallows using the first column of I, since a penalty free initial gap in x is allowed by $M(i,0) \longleftarrow 0$.