

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. augacuucucucc
6. augatuututttt
7. tactgaagagggtaa

DNA sequence \_\_\_\_\_ is transcribed to produce mRNA sequence \_\_\_\_\_, which is then translated to amino acid sequence \_\_\_\_\_.

(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

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

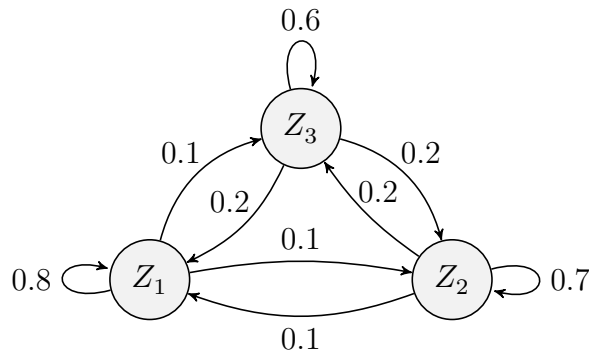
**Problem 3.**

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

During cell division \_\_\_\_\_ is usually approximately copied to the daughter cells, along with \_\_\_\_\_ which is copied with high fidelity (i.e. very few errors). Abnormalities in \_\_\_\_\_ are typically responsible for inherited disease, while abnormalities in both \_\_\_\_\_ and \_\_\_\_\_ 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  $X = \text{HHHTTT}$ .

Define:  $\alpha_{ij} \stackrel{\text{def}}{=} P[X_{1..i}, S_i = Z_j | \lambda]$   $\beta_{ij} \stackrel{\text{def}}{=} P[X_{i+1..n} | S_i = Z_j, \lambda]$   
 with  $\lambda$  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
$\alpha_{i1}$		0.5120	0.3354		0.0109	0.0030
$\alpha_{i2}$		0.0160	0.0141		0.0322	0.0214
$\alpha_{i3}$		0.0400	0.0392		0.0159	0.0085
$\beta_{i1}$	0.0412	0.0572	0.0686			1.0000
$\beta_{i2}$	0.0198	0.0595	0.2741			1.0000
$\beta_{i3}$	0.0321	0.0687	0.1560			1.0000

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

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 quantity. e.g. your answer should look something like:

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

**Problem 5.**

Consider aligning 2 sequences  $x = x_1 \dots x_n$  and  $y = y_1 \dots y_m$ . 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  $x = x_1 \dots x_n$  and  $y = y_1 \dots y_m$  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:  $s(\mathbf{a}, \mathbf{a}) + s(\mathbf{b}, \mathbf{b}) - d - 2e$

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:  $\begin{array}{c} \text{--babb--ab-} \\ \text{aaaabbaaba} \end{array}$  would have a score of:  $2s(\mathbf{a}, \mathbf{a}) + s(\mathbf{a}, \mathbf{b}) + 3s(\mathbf{b}, \mathbf{b}) - d - e$