

## CS196-1 HW1: Solutions

### Problem 2

We gave a specific database to align the protein sequences against, so these matches should be identical. However, we were not specific with what exact database to align the DNA sequences, so many answers are acceptable. The relative order of the species should be the same and the results should be similar. The following are the results obtained by the preferred method (aligning the human BRCA1 gene against the Nucleotide collection database).

#### BLASTN

<i>Organism</i>	<i>Max Score</i>
Bos taurus	5040
Rattus norvegicus	747
Drosophila melanogaster	None or 41
Mus musculus	1471
Xenopus laevis	None or 167

#### BLASTP

<i>Organism</i>	<i>Score</i>
Bos taurus	2587
Rattus norvegicus	1854
Drosophila melanogaster	52.4
Mus musculus	1847
Xenopus laevis	293

#### Conclusion

Bos taurus, Rattus norvegicus, and Mus musculus clearly have genes with high similarity to the BRCA1 gene. Xenopus laevis, while its scores are much lower, is still very similar—at least at the protein level. A glance at the database entry of the match (AAL13037) leaves no doubt: its definition is “breast and ovarian cancer susceptibility protein”.

Drosophila melanogaster has low scores by both protein and nucleotide matching, and a look at the database entry for the match confirms that it is not related to BRCA1.

Therefore the organisms with high similarity are Bos taurus, Rattus norvegicus, Mus musculus, and Xenopus laevis.

### Problem 3

$\delta_U$	-	A	C	G	T
-	<i>nil</i>	0	0	0	0
A	0	1	$-\infty$	$-\infty$	$-\infty$
C	0	$-\infty$	1	$-\infty$	$-\infty$
G	0	$-\infty$	$-\infty$	1	$-\infty$
T	0	$-\infty$	$-\infty$	$-\infty$	1

The above similarity matrix is preferred. Generally, the matrix must have a superior score given to matches, an intermediate score given to gaps, and a very low score for mismatches (such that they should never be able to occur since we are seeking subsequences). We accept any correct matrix given proper justification.

### Problem 4

There is one set of letters which denote single amino acids (one letter for each of the 20 amino acids) which exclude B, J, O, U, X, and Z. There is another set which includes B, Z, J, and X (each stands for multiple amino acids). 'String Length' can be defined in many different ways, all of which should yield the same results but sometimes do not (therefore, this part was not graded):

<i>String</i>	<i>Amino Acids</i>	<i>Highest Score</i>
computerscience	cmpterscience	27.8
biology	ilgy	16.8
	bilgy	19.3

*cmpterscience* has the greater string match.

BLASTing *ilgy* returns only the message "no significant similarity found". Why is that? The sequence is so short that it returns too many results to be useful! Try searching for the sequence in a single genome instead of the whole database, and you will see matches.

<i>String</i>	<i>Amino Acids</i>	<i>Highest Score</i>
protein	prtein	23.5
aminoacid	aminacid	25.7

*aminacid* has the greater string match.

<i>String</i>	<i>Amino Acids</i>	<i>Highest Score</i>
dynamicprogramming	dynamicprgramming	30.3
divideandconquer	divideandcnqer	29.1

*divideandcnqer* has the greater string match.

## Problem 5

For this problem, we accepted any answers that were well justified. Here are some possible solutions:

1. Yes. Align phrases typically used by Shakespeare in his known works to the questionable document.
2. No. Alignment does not give you the frequency of phrases.
3. Yes. The slang words are similar to standard English words, so aligning a dictionary to a slang word may find its standard spelling.
4. No. Deciphering the code is akin to choosing a similarity matrix, not to alignment.
5. Yes. Plagiarized papers would have large regions which align.

## Problem 6

There are many possible answers for this problem. Anything fulfilling the requirements with a good explanation was accepted.

Both of these matrices disallow gaps by making the cost of a gap be negative infinity, ensuring that any alignment without gaps will have a higher score. They prioritize matches by crediting a match with a positive amount and giving mismatches a negative value.

	-	A	T	C	G
-	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$
A	$-\infty$	1.0	-1.0	-1.0	-1.0
T	$-\infty$	-1.0	1.0	-1.0	-1.0
C	$-\infty$	-1.0	-1.0	1.0	-1.0
G	$-\infty$	-1.0	-1.0	-1.0	1.0

	-	A	T	C	G
-	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$
A	$-\infty$	1.0	-0.5	-0.5	-0.5
T	$-\infty$	-0.5	1.0	-0.5	-0.5
C	$-\infty$	-0.5	-0.5	1.0	-1.0
G	$-\infty$	-0.5	-0.5	-1.0	1.0

The difference between these two matrices lies in the penalty for mismatches between A or T and any other base. Since there are fewer hydrogen bonds between A and T versus C and G, it could make sense to penalize mutations of As or Ts less.