

CS195L HW1: Solutions

Problem 2

We gave a specific database to align the protein sequences against, so these matches should be identical. 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). Note that these are

BLASTN

<i>Organism</i>	<i>Max Score</i>
Bos taurus	5750 and 2675
Rattus norvegicus	3386 and 1086
Drosophila melanogaster	41.0 and 41
Mus musculus	3283 and 3279
Xenopus laevis	310 and 310

BLASTP

<i>Organism</i>	<i>Score</i>
Bos taurus	2592 and 258
Rattus norvegicus	1854 and 1849
Drosophila melanogaster	52.4 and 52.0
Mus musculus	1847 and 1846
Xenopus laevis	293 and 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, still has relevant similarity to the others. 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

δ_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). Best scores are highlighted.

<i>String</i>	<i>Amino Acids</i>	<i>Highest Score</i>	<i>Length</i>
computerscience	cmpterscience	28.2	12
biology	ilgy	16.3	4

cmpterscience has the greater string match.

<i>String</i>	<i>Amino Acids</i>	<i>Highest Score</i>	<i>Length</i>
protein	prtein	23.1	6
aminoacid	aminacid	26.5	8

aminacid has the greater string match.

<i>String</i>	<i>Amino Acids</i>	<i>Highest Score</i>	<i>Length</i>
dynamicprogramming	dynamicprgramming	31.2	14
divideandconquer	divideandcnqer	32.0	14

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. *2008 answer:* Yes. Align phrases typically used by Shakespeare in his known works to the questionable document.
2010: Many students rightfully disagree that alignments are not proof; they may spuriously just confirm that it is a 16th century book. The phrases that are signatures of Shakespeare are probably too unique to detect by alignment.
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



Figure 1:

Problem Extra-credit

There are a number of ways to reach the optimal alignment for these polypeptide sequences. One acceptable answer is to choose - by heuristic - a good ordering of pairwise alignments to perform such that you get an optimal answer. Any reasonable technique using the NCBI tools is accepted. The simplest way to get the answer is to use NCBI's COBALT, a constraint-based multiple protein alignment tool.

```

1  TPNVSVVDLTVRLGKG----- 16
1  -----LEKPAKYDDIK-- 11
1  -----LDDDVTESDVNAA 13

```

1	-----KGASYEDVCAA	11
1	--DVSVDLTV-----	9
1	-----LTCRLEKPAKY-----	11
1	-----NKETTYDEIKKV	12