**Problem Set 1 Written Answers**

***Problem 1***

(d) TODO

***Problem 2***

(a) *How many hits are there and what percentage fall near the diagonal?*

There are 62,829 hits, 24.7% of them are near the diagonal.

*Do you observe any structures in the off-diagonal hits?*

Some hits within the diagonal appear to make “string” or line within this region. However, it does not appear that there is any structure in the non-diagonal regions.

*What types of genomic elements could cause such a pattern?*

Likely regions that would be highly conserved between species, such as promoters, exons, etc.

*Why are matches that are close to the diagonal more likely than off-diagonal matches to represent “correct”, or orthologous, alignments?*

Because they respect the order of genomic elements and sequences, and because they are closely surrounded by other matches. Off diagonal matches tend to be isolated, and do not exhibit any patterns or structures, suggesting they are small random matches.

(b) Firstly, in order to change the k-mer size and number of subsequent base matches, I grouped the seq1 and seq2 hashing and hits code into a function *findHits* which takes as arguments:

*seq1* - string, sequence 1

*seq2* - string, sequence 2

*kmerlen* - int, size of k-mer

*matchspacing*- allowed mismatches between matching bases (ex. 1 indicates exact matching, 2 - every other base must match, 3 - every third base much match, etc.).

For parts i to iv, the only change required is in how the keys are generated. The key for each k-mer is generated only from characters in the sequence following the array indexing [ start : end : increment ]:

*key = seq[ i : i + kmerlen : matchspacing ]*

For part v, a hashing algorithm will not work since you cannot identify where the mismatches would be occurring. This may require another inexact matching type algorithm such as BLAST. (?)

(c) The 90 and 120-mer have equivalent hits on the diagonal. This is because they match longer regions, and allow for smaller mismatches. Between different species, you would expect some regions to be conserved, while some may be different. The 90 and 120-mers allow for small (< 4) mutations, but require conservation of larger areas.

(d) As sensitivity increases, specificity decreases, and vice versa. As the size of k increases, the specificity also increases and the sensitivity decreases because this requires larger areas of conservation. As the size of match spacing increases, the sensitivity increases and the specificity decreases because this requires more exact matching.