**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.

***Problem 3***

(a) *What is the expected (mean) state duration of state k as a function of the transition probability akk ?*

Let N be the number of consecutive states, j and k be the two states, and axy be the transition probabilities between states *x* and *y*, where *x,y* = j,k

Mean state duration = [akk\*P(pii=k) + ajk\*P(pii = j)] \* N

*What is the distribution of state durations P(Dk = d)?*

Given M possible states, the probability distribution of state *d* is

P(Dk = d) = add\*P(Dk = d) + sum{i = all states except d}[ aid\*P(Dk = i) ]

(b) The expected state durations for high and low GC regions are 99 and 101 base pairs respectively (according to the authoritative statistics). According to the Viterbi analysis of hmmgen, the expected state durations for high and low GC regions are 338 and 365 base pairs respectively.

(c) Statistics from mystery files

*mystery1*

Authoritative annotation statistics

-----------------------------------

High-GC mean region length: 100

High-GC base composition: A=19.94% G=29.87% C=30.20% T=20.00%

Low-GC mean region length: 101

Low-GC base composition: A=29.87% G=20.27% C=19.73% T=30.13%

Viterbi annotation statistics

-----------------------------

High-GC mean region length: 221

High-GC base composition: A=20.21% G=29.45% C=29.60% T=20.74%

Low-GC mean region length: 232

Low-GC base composition: A=29.43% G=20.85% C=20.48% T=29.24%

Accuracy: 71.89%

*mystery2*

Authoritative annotation statistics

-----------------------------------

High-GC mean region length: 100

High-GC base composition: A=19.85% G=29.78% C=30.07% T=20.30%

Low-GC mean region length: 99

Low-GC base composition: A=29.84% G=19.86% C=19.99% T=30.31%

Viterbi annotation statistics

-----------------------------

High-GC mean region length: 204

High-GC base composition: A=20.33% G=29.35% C=29.71% T=20.61%

Low-GC mean region length: 223

Low-GC base composition: A=28.97% G=20.68% C=20.76% T=29.59%

Accuracy: 68.66%

*mystery3*

Authoritative annotation statistics

-----------------------------------

High-GC mean region length: 100

High-GC base composition: A=19.81% G=29.71% C=30.56% T=19.91%

Low-GC mean region length: 100

Low-GC base composition: A=29.56% G=20.09% C=20.11% T=30.24%

Viterbi annotation statistics

-----------------------------

High-GC mean region length: 207

High-GC base composition: A=20.24% G=29.33% C=30.17% T=20.26%

Low-GC mean region length: 221

Low-GC base composition: A=28.84% G=20.75% C=20.81% T=29.59%

Accuracy: 67.85%

The compositions and mean lengths are relatively similar between all three mystery sequences, with differences being very minute. Mystery3 has equal mean lengths for high and low GC-rich areas, while mystery 1 is slightly higher in low GC regions and mystery2 is slightly higher in high GC rich regions.

The accuracy levels of Viterbi was lower on these samples compared to hmmgen, with the highest being mystery1 at 72%, mystery2 at 69% and mystery3 at 68%.

The Viterbi state distributions differed only slightly from the predicted distributions. In all cases, the GC levels were slightly lower in high GC regions than expected, and slightly higher in low GC regions than expected.

(d) Retraining the HMM parameters would likely not improve the accuracy of this algorithm. While retraining guarantees that we will converge on a local maximum, it does not guarantee that we will maximize P(x|pi), and performance is worse than Baum-Welch. Why?