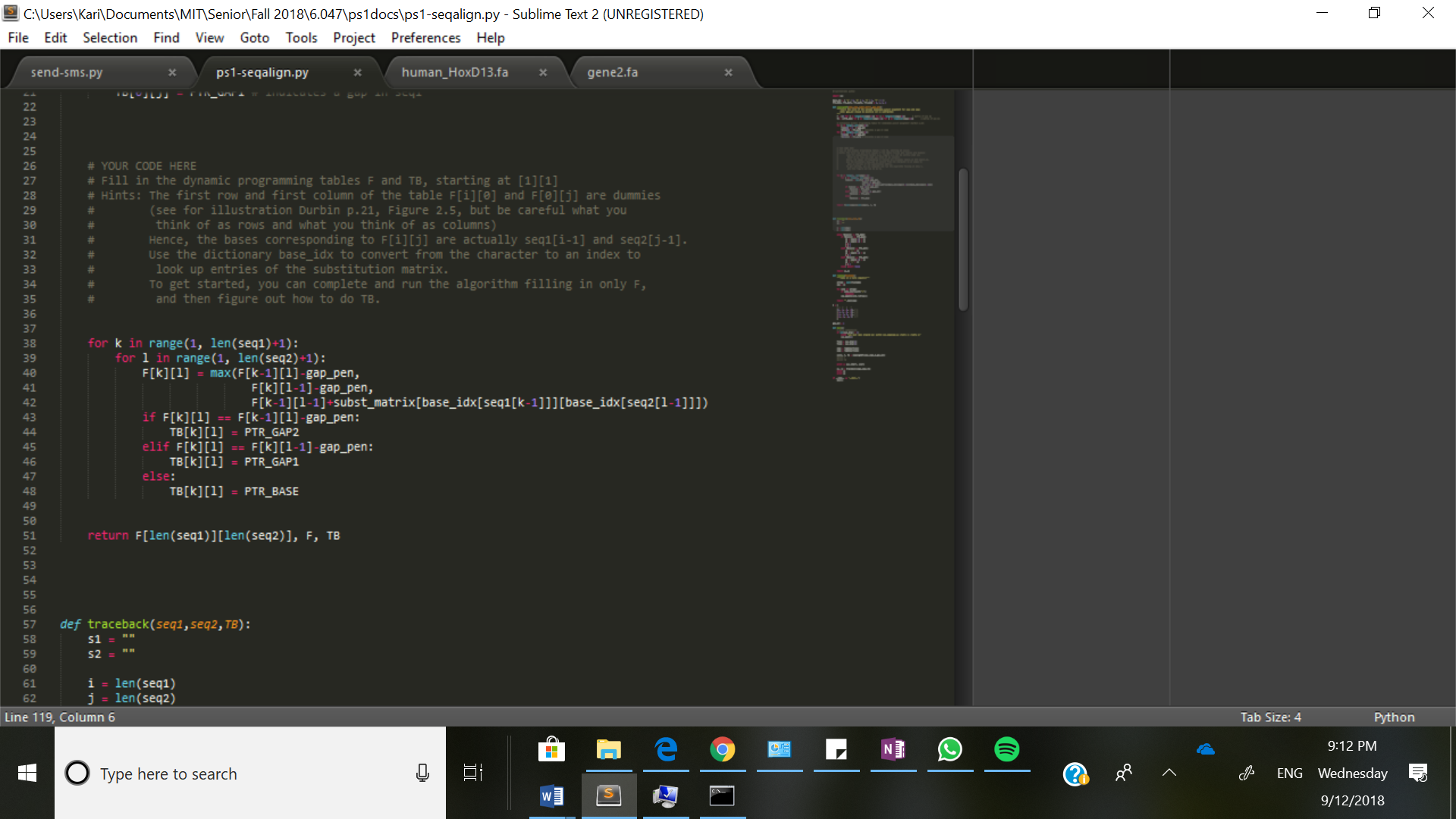
Kari Stromhaug

Collaborators: Claire Goul, Ronit Langer

Problem Set 1

Problem 1:

Part a:

1. 
2. CTAAGTACT

C-AT-TA--

1. 2971

Part b:

* Went through and made everything positive, set the diagonals to 0.
* Calculate the min instead of the max
* Set the initialization to positive as well
* Added the gap penalty instead of subtracting it
* Subtracted the base\_idx value instead of adding
* Alignments were the same so that was good

Part c:

* I get a score of 197

Part d:

* Using HoxD, I get a score of 258
* Human\_hoxD and Human hoxA: 1145
* Mouse hoxD and Mouse hoxA: 1095
* Averaging the two above I get 1120
* Averaging the human-mouse scores I get 227.5
* So we use a ratio: 1120/x = 227.5/70,000,000 🡪 x = 345,000,000 years
* Assumptions: we had the same rate of change over all of those years

Problem 2:

Part a:

* There are 62829 hits and 24.70197% hit the diagonal
* The off-diagonal hits are structured as horizontal and vertical lines, and probably come from repetitive sequences in the genome. There are certain sequences that have multiple matches in the other animal.
* **What types of genomic elements could cause such a pattern?**
* They could be transposable elements, regulatory motifs or UTR’s. They could also possibly be repetitive satellite DNA regions.
* **Why are matches that are close to the diagonal more likely than off-diagonal matches to represent “correct”, or orthologous alignments?** Matches close to the diagonal are roughly in the same location with respect to the rest of the sequence in both animals. It is more likely, and makes sense if both organisms had a common ancestor.

Part b:

1. Changed kmerlen to 100. There are only a few hits, all of which lie along the diagonal. There are 1198 hits, and 100% hits on the diagonal.
2. Changed kmerlen to 60

Changed key assignment to key = seq1[i:i+kmerlen:2] to skip every other nucleotide.

The graph is more sparse than the original. There are still some hits that are off the diagonal, but not nearly as many.

23933 hits, 38.74149% on the diagonal

1. Changed kmerlen to 90, and 2 to 3 in the key-assignment line

The graph is much more sparse outside the diagonal. There are only 5 off-diagonal hits.

8887 hits, 93.85619% hit the diagonal

1. Changed kmerlen to 120, and 3 to 4 in key-assignment line

The graph is less sparse now. There are more off-diagonal hits, and they seem to be along 4 or 5 lines. So it looks like 5 sequences in seq1 matched multiple locations in seq2.

6044 hits, 82.13104% hit the diagonal

1. **How to implement checking for at most two mismatches in each contiguous sequence of 6**

You would have to roll each 100mer through the sequence to align to the reference.

At each 100mer, start rolling through and check the alignment of the 6 nucleotides. If there are three mismatches, roll the 100mer to start right after the first mismatch. Keep checking as such.

Part c: **Although parts a, b.ii, b.iii, and b.iv require the same number of matching bases (30 = 60/2 = 90/3 = 120/4), one of them is more specific to the diagonal. Explain why this might be so.**

* 90/3 is likely better than 60/2 and 30 because it checks over a longer sequence.
* 90/3 is much more specific to the diagonal. I would guess that as you look at every interval of 4 and higher, there is more room for random coincidental matches, as opposed to matches that are specific to just one area
* 90/3 therefore strikes a good balance between checking over a larger area but not being too far spaced.

Part d: **Explain the trade-off you see between number of hits near the diagonal (sensitivity) and the percentage of hits near the diagonal (specificity). How is the trade-off affected by the hashing parameters?**

* As specificity decreases,
* As sensitivity increases, specificity decreases. So there will be fewer hits, and a higher percentage will be on the diagonal.
* As the space between checks increases, the (sensitivity) number of hits on the diagonal increases. Overall, the specificity decreases as the sensitivity increases.

Part e:

* The inversion is at location 800000

Question 3:

Part a:

1. This is a geometric sum, which comes out to: 1/(1-akk)
2. It’s a geometric distribution:

Part b:

* The high-GC expected duration is 100, and the low-GC expected duration is 1.01
* Can confirm that I got 83.31% accuracy

Part c:

* Mystery1:
  + 71.96% accuracy
  + highGC: 234 --- expected 100
  + lowGC: 220 --- expected 101
  + In both, the majority of the lengths lie between 40 and 150. However, the Viterbi algorithm predicted some tat were much larger, extending all the way up to 1000
* Mystery2:
  + 68.80% accuracy
  + highGC: 214 --- expected 100
  + lowGC: 212 --- expected 99
  + In both, the majority lie between 50 and 120, but the Viterbi alg has peaks at around 350-300, and determined sequences up to 1100 long
* Mystery3:
  + 67.72% accuracy
  + highGC: 221 --- expected 100
  + lowGC: 207 --- expected 100
  + for the authoritative plot, all lengths are between 92 and 98, which is where the largest peaks in the Viterbi plot are as well, but again it comes with a few lengths above 1000.

Part d:

* As long as everything is correctly annotated and we have enough data then retraining will produce a better markov model. This is because we would generate accurate transition and emission probabilities.
* If we don’t have enough samples to train on, the model could be very overfit. We would have to make sure we have a large, variable set of data.

Part e:

* A problem with our Viterbi model is that gene component lengths do not follow a geometric distribution.
* They use a semi-markov architecture that takes into account the typical number of exons per gene
* They maintain a reading frame
* The complete number of codons is generated from the appropriate length distribution. There are three different kinds of exons, and each has a slightly different length distribution. So they use a different length dist depending on the type of sequence.
* We should use a non-geometric distribution, one that we find more suitable for the actual composition of the genome. We could use a fixed-length distribution or a non-parametric distribution, which uses histograms.