CSCI 652/490 Algorithmic Bioinformatics I Project 1 100 pts

Substitution is a type of genomic mutation where one base is replaced by another type (e.g., A→ G, A→ T, C→ T, etc.) There are two categories of substitutions: transition and transversion. Transition refers to substitutions between A and G, and C and T. Transversion refers to all other types of substitutions. The ratio between transition (t_i) and transversion (t_v) is a characteristics between a pair of species. Indel is another type of genomic mutation, including insertion and deletion. An indel is represented as contiguous '-' characters in an alignment. Below is an example of short alignment between two species.

```
Species1   TTCTGATGACTAACTGGACTGA
Species2   TTCCG-----TAGCTGGACAGA
```

In above example, there are three substitutions: two transitions and one transversion. There is an indel of five bases. When we analyze an alignment, substitutions are often referred to as "*mismatches*". Positions in the alignment with identical bases from each species are called "*matches*". Similarly, an indel is often referred to as a "*gap*". A sequence of contiguous '-' characters is counted as one gap. The gap length refers to the number of '-' characters in the gap.

For a given pairwise alignment between two species, we have following definitions:

Substitution rate = number-of- mismatches / (number-of-mismatches + number-of-matches)

t_i/t_v = number of transitions / number of transversions

gap rate = number of gaps / (number of matches + number of mismatches + number of gaps)

In this project you will study the difference between human and other two reference genomes based on pairwise alignment result. The alignment files human.chr22.chimp.maf and human.chr22.mouse.maf are located at the department hopper server: /home/turing/mhou/csci652fall2018/data/ human.chr22.chimp.maf

In the alignment files, some bases are in lower case. They shall be treated the same as upper case letters.

1) (40 points) For each above file, compute substitution rate and t_i/t_v ratio. Please also output original counts that can be filled in the following table. Each cell records the number of positions from the alignment file.

	Species 2				
Species 1		A	C	G	T
	A				
	C				
	G				
	T				

2) (30 points) For each above file, compute overall gap rate. Compute the gap frequency of each gap length. Fill your data in the following table. "…" in the table indicates you can expand the number of rows in the table as necessary.

Gap length (bases)	Gap count	Gap frequency
1		
2		
3		
…		
Total		

3) (30 points) Write a report regarding your project. Include your last name, first name, zid, assignment number on the top of your report. Specify the programming language you use and the approximate amount of time you spend on programming (including debugging, testing, running program, etc.). Summarize your result and give brief insight about what you learn from your result.

4) Submit your program source file to Blackboard before due time. Submit your printed written report before class on the due day.