CSCI 652/490 Algorithmic Bioinformatics I Term Project Option B Fall 2018

Impact of parameters on pairwise alignment quality

In the second course project, you have studied impact of gap open penalty and alignment score threshold on the alignment accuracy. That serves as an initial experiment to study impact of parameters on alignment quality. In this term project, you will study more parameters including seed pattern on the alignment quality given more simulated sequences based on more complicated evolutionary models.

You won't notice much difference of the evolutionary models used in simulation for this project comparing to project 2, except inversions. Because of the existence of inversions, now true alignments are also recorded in maf-format alignment files. That requires revision to your alignment comparison program that you created in project 2.

Datasets are located at:

/home/turing/mhou/csci652fall2018/data/TermProjectB

In each dataset, there are 9 sequence files of 9 species and a maf-format multi-alignment file that records true alignment: simali.output.proj.maf.

The maf-format is listed here again:

```
a score=????
s genome.chr start-position sequence-size +/- chromosome-size alignment-sequence
s genome.chr start-position sequence-size +/- chromosome-size alignment-sequence
```

The multi-alignment block may contain two or more alignment sequences in each block. The first alignment sequence always has '+' strand. The other alignment sequences may have '-' strand that indicates inversions. When the alignment sequence is on '-' strand, its start-position is relevant to the other end of the sequence. There are two approaches that you can consider to evaluate alignment with inversions.

> Approach 1: You can separate the cases of non-inversion and inversion and add the counts at the end to compute overall sensitivity and specificity for comparing each pair of alignment files.
>
> Approach 2: You can transform the position information to be relevant to '+' strand. But you need to be very careful in such operation. With this approach, both start-position and sequence-size matter in the project.

You can use the same pipeline of alignment tools in project 2 to produce alignment results in this project. You can also use the same measurement of sensitivity and specificity as you used in project 2. The goal here is to find best parameters for each of the pairwise alignments (using blastz) between human and other species. You need to try at least different gap open penalties (O), alignment score thresholds (L) and seed patterns (T and W).