Computer Science C122 / 222 - Algorithms in Computational Genomics

Project 1a, 1b - Read Mapping

In this project, the goal is to identify the variants in a donor genome using reads from that genome and a reference genome sequence.

Genomic Simulator:

The reference sequence is a randomly generated sequence. A donor genome sequence is simulated by applying a mutation process on the reference sequence. In the mutation process, each position has a chance of a substitution, insertion or deletion. The mutations are logged as they are being generated resulting in a donor genome "true mutation" list relative to the reference genome. Reads are then simulated from the donor genome by sampling uniformly from locations in the genome. Read errors are inserted into the reads with each position in the read having a chance of substitution (read error), insertion or deletion. Reads are generated in pairs with the insert size between the pairs uniformly generated between a minimum and maximum insert size.

Project Goal:

The goal of the project is to predict the donor mutations list from the reads and reference genome.

Data formats:

The reference genome and reads format are in "fasta" format. Each sequence in a fasta file contains an identifier line starting with a ">" and then the sequence on following lines.

An example of the reference sequence format is:

```
>genome_1000
AAGTGGGTCTCGGCGGAACTGGCTACGAGAATATGCAGTTGGCAATGGTACCACTTTTGTAAGTACATAGTTCATGAGTC
CGTTTTGACGTGGTGGCCATCTTTGTCACACCTCGATCCACGCCCTATAATACTTAGTTAACGCCTTCTATGTCGTGTAA
TCCACAAATTAATTCGAGAACATCCTGCCCGTAGGTTTCAGATGGATTCATAGTGCCCCATTTGGTGACGAGCGCTTGAG
GCAACTATTTAGCTTTGCGGCGTGACCCGCACTACCGTATCCGTTGGGCCTGTTTTAAGGAAAAATATAGGGCAAGACTG
ACTTGGCCCAGTGCAATCGCGCACCCCGCCTCGCAGCAGGCCTCTAGAAGCAGCAGGTCTCGTGTTAAGGTTGTACTAGG
TTCGGTTAGTACACTTTGACTACACCATCAGTAACTATTAAGATCAGATTCGTTCGTGTTAGTAGGATCCATGGATTCCG
ACATCGGCCCGAAGCCCCCTGGGTCACAATGAGGCAGGCGATCGGAGCGACATACGACCCCACTCCACATTAATGCGATG
```

Reads which are part of a pair are denoted with the same identifier ending with a "/1" or "/2". An example of a reads format is:

```
>read_0/1
TGACTACACCATCAGTAACATTAAGATCAGATTCGTTCGTGTTACTAGG
>read_0/2
CCACTCCACATTCATGCGATGGATATGATCCCACGGCAAGTCGCCTTTGA
>read_1/1
ACTACGTATCCGTTGGACCTGTTTTAAGGAAAAATATAGGGCAAGACTGAC
>read_1/2
AGGAAAAATATAGGGCAAGACTGATTTGGCCCAGTGCAATCGCGCACCCC
```

```
>read_2/1
AGCGCGCACCCCGCCTCTCAGCAGGCCTCTAGAAGCAGCAGGTCTCGTGT
>read_2/2
ATTCGTTCGTGTCACTAGGATCCTGGATTACCGACATCGGCCCTATGCCCC
>read_3/1
CAGGCGAACGGAGCGACATACGACCCCACTCCACATTAATGCGATGGATA
>read_3/2
CAGTAATTAGATGGGATAATTTCGTTCGGGGTCCAACCACCTATAGGTAG
>read_4/1
GTTGCATGTCCAAGTAGAGAAGAGCCAGTCCCCCGGACACGCTCCAAAACG
>read_4/2
GAGATACATCTCGAGGATGGGCCTGCGCGTCAGCTAATACATTAAATTCA
```

The mutation list format has a line for each mutation between the reference and the donor genome. The format starts with a ">" symbol followed by a "S", "I", or "D" for a substitution, insertion and deletion followed by the position in the reference followed by a space. For substitutions, the format continues with the reference and donor bases are provided. For insertions, the new sequence that is added is included. For deletions, the sequence that is removed is included. An example of the format is:

```
>S732 G A
>S851 A G
>I260 T
>I274 T
>D96 C
>D471 T
```

Evaluation:

The predicted mutation list is evaluated by comparing entries between the predicted mutation list and the true mutation list. An entry is considered correct only if it matches exactly between the two lists. We evaluate the predictions by computing the F-score for overall mutations, substitutions only, insertions only and deletions only.

Sample Genome:

A sample genome of length 1000 is provided with the true mutation list. A set of reads with errors and without errors is provided. This genome can be used to develop and evaluate a solution to the project before applying it to the larger problems.

Project 1a:

A reference genome of 10,000 is provided with paired reads containing errors is provided.

Project 1b:

There are two versions of this project which differ by genome sizes. The two genome sizes are 1,000,000, and 100,000,000.