Theory and Application of Bio-Informatics, Spring 2017

Assignment 2: BLAST

In this assignment, you will implement a Heuristic Local Sequence Alignment Search algorithm, using filtration concepts as in BLAST. You will compare your local alignment algorithm from Assignment 1 to the one you'll implement here.
Please read the entire Assignment before starting your implementation.

Input files:

text.fasta – containing text sequences t1 to t10
queries.fasta – containing query sequences q1 to q10
score.matrix – a scoring matrix with the same format as in Assignment 1

BLAST overview:

From Wiki:

"In bioinformatics, **BLAST** for **B**asic **L**ocal **A**lignment **S**earch **T**ool is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences. A BLAST search enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold."

For this assignment, BLAST will be defined according to this variant:

1. Text pre-process – mapping each w-mer (a substrings of length w) from the text to its index in the text string. Mapping can be done in various ways – you can implement it using a data structure of your choice. This is done once for the text sequence, and used for multiple queries.
2. Query pre-process – construct a data structure, which holds for each w-mer from the query a list of neighbour words. Reminder: A neighbour word s' of a w-mer s is a word of length $|s| = w$, with $score(s, s') \geq T$ (using a scoring matrix). T is a chosen threshold.
3. Finding hits – for each word in the query ("sliding window"), find the indexes of its (and its neighbour words) appearances in the text, using the data structures constructed in stages 1 and 2. A hit is called a HSP (High-Scoring Segment Pair).
4. Extension of HSPs – extend the hits found in stage 3 to obtain MSP's (Maximal Scoring Pairs). Extension stops when the extended HSP's score drops X below the maximal score obtained so far.

A point to consider: How will you evaluate the results? How will you decide if the query has a significant match in the text?

PART 1 - Local alignment:

In this part, locally align each query sequence (q1 to q10) to all text sequences (t1 to t10), using your algorithm from Assignment 1. For each query sequence, save 2 text sequences that were best aligned. You will use this information in later parts to test your algorithm's results.

PART 2 – Your BLAST algorithm:

Implement your BLAST algorithm according to the stages mentioned in the BLAST overview. Feel free to vary from specific guidelines, but make sure you can rationalize your changes. Find a way to score your MSPs, so you can compare scores between different pairs.

In the output file include all query sequences against all text sequences.

As in part 1, for each sequence in the query file, keep track of the 2 best scoring text sequences.

Change your parameters (w, T and X) to achieve:

1. The highest accuracy possible compared to part 1.
2. 20% (roughly) less accurate, but a faster running time.

In a PDF file, compare the running times of the 3 algorithms:

- Local alignment (Part 1).
- Blast implementation with maximum accuracy (compared to part 1).
- Blast implementation with lower accuracy, but higher running time.

*The meaning of 100% accuracy is that for each query sequence, the 2 best scoring text sequences are the same as in part 1.

Also discuss in the PDF:

1. The changes you've made in the parameters, and how did they change the speed and accuracy of the algorithm.
2. What is the bottleneck of your algorithm, running time wise?
3. What is the major memory requirement of your algorithm? Can it be improved?
4. Can closely related sequences be accurately found, using an algorithm which disregards indels (insertions/deletions), as you just did?

PART 3 - Some Biology:

Using NCBI BLAST, find out which sequences belong to which species, and what is the gene's name. Include a PDF file which states for each query sequence, which are the 2 best matching text sequences you've found (in Part 1). Do not write the whole sequence, but rather the label you've found in NCBI.

For example:

q2: House cat - alpha kinase:

- t4: Yellow eyed Siamese cat – alpha kinase.
- t6: Tiger – protease p1.

Submission information:

The submission can be alone or in pairs but not triplets.
You must implement the assignment using python 2.7

Submit zip file named BLAST.zip with the files:

- **part2.pdf**
- **part3.pdf**
- Source files in folder named ***src*** and a python file named ***blast.py***
- Output files inside a folder named ***output***

Submit your assignment using the submission system.