String Overlap Finder

Introduction

This is an implementation for a solution for all-pairs suffix prefix using compact prefix tree. We called our solution String overlap finder (SOF). This solution uses OpenMp to support multithreading.

To compile : make

Running the program

```
Apsp filename
```

The program has one parameter and five optional parameters:

filename: is the name of the file. Here is an example for the contents of the file:

```
ACCCCAT
TTTCCAGG
TTTGGCCAAA
```

where '\n' (new line) is the separator between input strings. The separator can be changed.

Optional parameters:

-p   the number of threads which are used. (The maximum is the default)

-d   distribution method (The default is 1).

-o   Output. The default is 0 (no output)

       1 : results are shown in two dimensional array (k2)

       2 : outputting all suffix prefix matches (not only the maximum).

-m   Minimal match length. (The default is 1).

-s   Sorting (The default is 0 (no sorting)).

Examples

This command will find overlaps using 4 threads. The results will be put in two dimensional array. Minimal length is 10.

```
Apsp  test.txt -p 4  -m 10 -o 1
```

Run the code Sequentially

to run the code sequentially:

```
Apsp test.txt -p 1
```

Important

- you can generate random cases to test the code. The program 'gen' will generate a random string. The user specifies 3 parameters:

1- K (number of strings)
2- N (total size of all strings)
3- if the generated strings have equal sizes.

The resulted file, test.txt, includes a string with the right format.

- if you have a fasta file, please use the program 'converter' to convert a fasta file to a file with the right format. To run:

```
converter t1.fasta t1.txt
```

- you may supply your own file. An example:

```
AACCCCAAAA
CCCGGTTTAAAAAA
AAGTCCCC
```

- In Apsp.cpp, there is a constant MAX_K which determines the maximum number of strings which the program can accept. Please feel free to increase it and run make again. You will notice a waste of memory if you use big MAX_K value for small samples (N is small). Please note that MAX_N determines the maximum length of N that SOF can handle. So if you test with large data sets, make sure that you increase this value.

- Make sure that you DON'T run the program with output=1 when you k >10000 since a two dimensional array is required to store results (k2).

- if you have any problem, please contact us:

Maan Haj Rachid
Qatar University
mh1108047@qu.edu.qa