Bertrand Chazot

UPM-64
Linear-Time String Matching Algorithms
The Z Algorithm

The goal of the Z Algorithm is to find in a text T, the positions -if any- of the pattern P.

The Z algorithm is based on the computation, for each character 'k' of the text, of zk, the size of the biggest prefix of P.$.T starting at 'k'. A little more in computed (l and r) in order to get a linear algorithm. r corresponds to the right end of the current prefix that ends the more at the right. l is the left end of this same prefix.

After that, it only remains to check if it exists zk such as zk=|P|. If so, P has been found in T at the position k.

When there is wildcard, this has to be done for each subpattern. The result consists of the positions of the subpatterns that occur in the right order without overlapping.

I implemented this algorithm in c++. The class Zalgo knows T, a method is called to find the occurrences of P -with or without wildcard- and return either a vector of positions of each occurrence of P for the no-wildcard case, either trees of positions, each path from the root to a leaf contains the position of each subpatern in T.

I tested this implementation on 5 tests : 2 different alphabets (binary and DNA), with or without wildcard.

Here are the results :

TEST 1 Binary, no wildcard
There is 2 occurrences of P in T
Percentage of case 1 : 0.300781
Percentage of case 2a : 0.638281
Percentage of case 2b1 : 0.0546875
Percentage of case 2b2 : 0.00625

TEST 3 Binary, no wildcard
P is not in T
Percentage of case 1 : 0.559507
Percentage of case 2a : 0.322027
Percentage of case 2b1 : 0.052503
Percentage of case 2b2 : 0.0659628

TEST 4 DNA, no wildcard
There is 5 occurrences of P in T
Percentage of case 1 : 0.917795
Percentage of case 2a : 0.0318405
Percentage of case 2b1 : 0.00289836
Percentage of case 2b2 : 0.0474665

TEST 2 Binary, wildcard
There is 1 occurrences of P in T
Percentage of case 1 : 0.564565
Percentage of case 2a : 0.346774
Percentage of case 2b1 : 0.0507553
Percentage of case 2b2 : 0.0379057

TEST 5 DNA, wildcard
P is not in T
Percentage of case 1 : 0.916677
Percentage of case 2a : 0.0618126
Percentage of case 2b1 : 1.9988e-06
Percentage of case 2b2 : 0.0215088

Bertrand Chazot

This algorithm is very efficient, except in tests 4 and 5 where the percentage of case 1 (the most time-consuming) is very high.