Michelle Meng

BENG 182

Prof. Bafna

9 May 2020

Assignment 3

1a. E(l-mer matches) = (m-l+1)(n-l+1)($\frac{1}{4^l}$)

1b. The expected number of l-mer matches between the query sequence and the database will decrease as m goes from a large number to l. For example, let's say n = 100 and l = 3. If we have a large m = 20, our expected number of l-mer matches is 27.56. Let's lower m to m=10. Our new expected number of l-mer matches is now 12.25. Now let's say that m=3, then we can expect only 1.53 l-mers to match.

1c. E(l-mer matches) = $\frac{m}{r}$ ((r-l+1)(n-l+1)($\frac{1}{4^l}$)) if we are only interested in matches occurring between a query string and the database. For our example above, the change would not be significant as query string lengths would be rather short, however if m is large (and we have many queries), the number of expected matches would decrease.

1d. E(5-mer matches for query string "AATAAGCCGC") =

(0.1)^5+((0.1)^4)(0.4))+((0.1)^3)(0.4)^2)+((0.1)^2)(0.4)^3)+((0.1)(0.4)^4) +(0.4)^5

= 0.0137

2. The method follows the following steps:

- Build a Aho-Corasick trie using queries
- Search through database
    - For matches, run global alignment

The time complexities are as follow:

- O(n + m) = build Aho-Corasick trie
- O(E) = expected number of matches
- O(r^2) = global alignment

Our runtime is the time it takes to build to Aho-Corasick trie and the time it takes the globally align each match (($\frac{1}{4^l}$ mn) expected matches that each take r^2 to run.

Speed-up = $\frac{mn}{T(l,m,n) + r^2(\frac{1}{4})^l mn}$ = $\frac{mn}{(m+n) + r^2(\frac{1}{4^l})mn}$

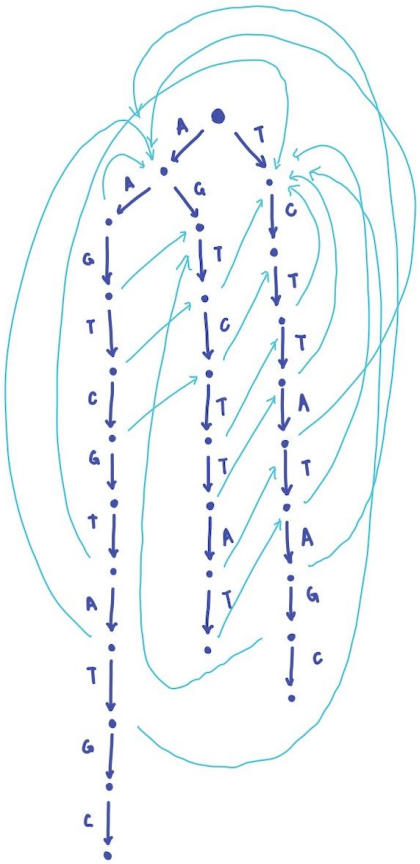
Sensitivity is the Pr(at least one l-mer match|alignment with edit 1- Ɛ)

Sensitivity <= $(1 - (1 - \text{Ɛ})^l)^{r-1+1}$

= 1 - $(1 - (1 - 0.15)^l)^{r-l+1}$

L	Speed-up	Sensitivity
5	0.102	1
11	419.395	0.99
15	105116.815	0.99
20	4782516.337	0.959
25	4999777.965	0.732
30	4999999.783	0.419
35	4999999.999	0.2
40	5000000	0.087

3.



4. E(hits) = $n(\frac{1}{4^{11}}) + n(\frac{1}{4^{8}}) + n(\frac{1}{4^{9}})$

P-val(finding one or more hits to any keyword) = 1 - P-val(no hits) = 1 - $\frac{e^{-E}E^{0}}{0!} = 1 - e^{-E}$

where E is the E(hits) above.

P-val(finding one of more hits to any keyword) = 1 - $e^{-(n(\frac{1}{4^{11}})+n(\frac{1}{4^{8}})+n(\frac{1}{4^{9}}))}$

To find at what size of a database matches are no longer significant, we want to find what value of n gives us a E-value >= 1. Let's deduce this below:

N	E-value
10,000	0.193
20,000	0.386
30,000	0.589
40,000	0.772
50,000	0.966
51,000	0.985
52,000	1.004
53,000	1.024
55,000	1.062
60,000	1.159

n = 52,000

5. The Aho-Corasick script and trie file are included in the code section, as well as the result files (results_query1.txt and results_query2).

Number of matches found in queries.txt: 22 (4 strings)

Number of matches found in queries2.txt: 452 (379 strings)

Queries.txt expected vs. observed (queries with no matches are not listed)

n = 13385190

Expected = $(n-l+1)(\frac{1}{4^l})$

l = length of match

	Expected	Observed
AATAGCTAACA	3.191	12
GCCTGGGTGACAGAGTGAGACCCTGTCTC	4.64 x 10^-11	8
ACCAAACTATAGAAT	0.012	1
CTCTTAATATTTATGAAGAAGAACATGGT	4.64 x 10^-11	1

For all of the queries whose matches were found had more matches than expected. This might be a result of Alu sequences. Alu sequences are known as selfish sequences who are short stretches of DNA that are the most transposable elements in the genome. This means that they reproduce and insert themselves throughout the genome. The human genome contains over one million copies of these pesky sequences. It is possible that the section of chr 12 that we examined on the reference genome hg-19 contains these Alu sequences. This is plausible because we expect to see 0 matches to

“GCCTGGGTGACAGAGTGAGACCCTGTCTC” and “CTCTTAATATTTATGAAGAAGAACATGGT”, however we observe 8 matches to the first and 1 to the second. “GCCTGGGTGACAGAGTGAGACCCTGTCTC” and “AATAGCTAACA” seem to show drastically more matches observed than expected, making it plausible that these are Alu sequences that reproduced into the section of DNA we observed.

*** Collaborated with Sabeel Mansuri