Approximate Regular Expressions: A Comparison of Exact and Approximate Matching Algorithms

Umme Salma Gadriwala, Tasnim Noshin, Rumsha Siddiqui
{gadriwau, noshint, siddiqur}@mcmaster.ca
April 2019

MOTIVATION, PROBLEM AND SOLUTION

MOTIVATION

In bioinformatics, DNA sequences are often represented by regular expressions to capture different variations of the same structure. Efficient approximate string matching would allow us to capture more sequences and optimize time and cost of resources.

PROBLEM

Is it viable to replace an implementation of exact regular expression matching with one of approximate matching for added functionality?

SOLUTION

A comparison of the running times of exact matching using Thompson's NFA to the Myers and Miller's approximate matching construction.

Performance tests compare the running times for both algorithm using sequences and regular expressions of various length.

EXACT MATCHING: THOMPSON'S CONSTRUCTION

INPUT: Regular expression, r over Σ; and string, s

OUTPUT: True, if s satisfies r

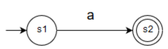
METHOD:

- Construct a Thompson's NFA by recursively applying ε, symbol, union, concatenation and Kleene closure rules over r.
- Traverse the NFA for s.
- Return *true* if traversal ends at a terminating state.

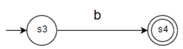
COMPLEXITY: NFA construction: $O(|r|)$ steps, $O(|r|)$ memory;
String traversal: $O(|r|\cdot|s|)$ steps

Thompson's NFA for r = (a|b)a*

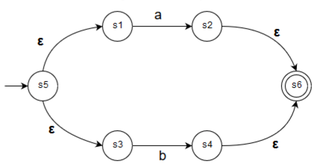
Step 1: r = a



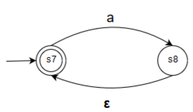
Step 2: r = b



Step 3: r = a | b



Step 4: r = a*



Step 5: r = (a | b) a*

APPROXIMATE MATCHING: MYER'S & MILLER'S CONSTRUCTION

INPUT: Regular expression, r over Σ; string, s; and error value, k

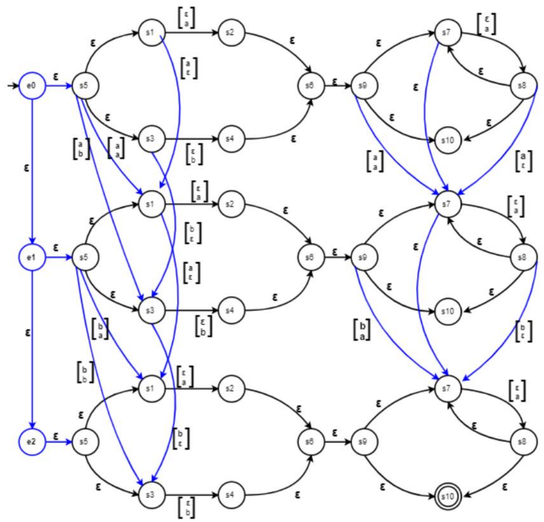
OUTPUT: True, if s satisfies r with at most k errors

METHOD:

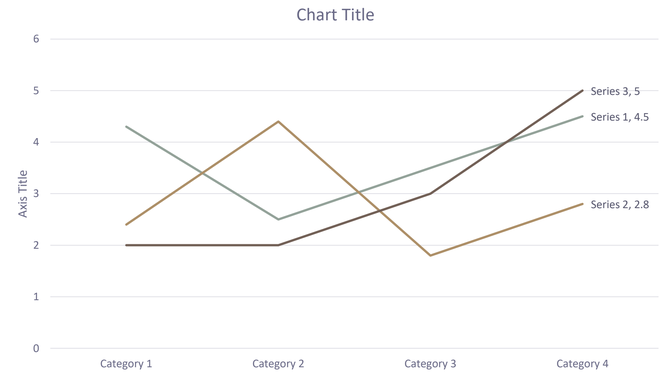
- Construct a Myer's and Miller's NFA by combining $|s|+1$ instances of Thompson's NFA construction of r by adding: deletion, insertion, and substitution edges based on s.
- Traverse the NFA for s, tallying each error transition.
- Return *true* if $k \leq$ counter.

COMPLEXITY: NFA construction: $O(|r|+|s|)$ steps,
$O(|r|\cdot|s|)$ memory;
String traversal: $O(|r|\cdot|s|)$ steps

Myer's & Miller's NFA for r = (a|b)a*, s = "ab"



PERFORMANCE TESTS



- Testing environment details

CONCLUSIONS

- Conclusion 1
- Conclusion 2
- Conclusion 3