An overview of search and match algorithms complexity and performance

sarah Mohsen fahmy / DR:sara el-metwally /genomics

ABSTRACT

DNA extracted from cells has huge amount of information about us. It distinguishes creatures and make them different. So, to do sequence matching that helps us to understand evaluation and genetic relationships, we have a lot of algorithms each of them has its performance and complexity.

INTRODUCTION

When we discover pattern in a specific string ,we have two main approaches of matching:

a. Exact matching:

1. For instance: Smith-waterman(SW),Needleman wunsch (NW), Boyer moore horspool(BMH).
2. Dynamic programming: Knuth morris pratt(KMP).

b. Approximate matching (fuzzy string searching):for instance like Rabin karp and Brute force.

We will analyze in this paper matching in protein, DNA and RNA.

RELATED WORK

Pattern P can be matched in String T by adding four empty spaces before the pattern and two after.

```
String   T:  A C C T C G A G T
                     | | |
Pattern P:   _ _ _ _ C G A _ _
```

In (Yeh and Cheng 2008), They used Levenshtein distance applied to images and videos to determine feature vectors. For instance:

```
Input A: □□ △ □ △
Input B: □□ △ □
```

To find maximum matches we remove last triangle in input A.

In (Amir et al. 2004),They proposed new string distance like Levenshtein distance implemented with Message Passing Interface (MPI).

In (Knuth et al. 1977), This traditional algorithm is now known as KMP string matching algorithm which used for pattern matching in strings.

In (Hussain et al. 2013), named Bidirectional Exact Pattern Matching (BDEPM) which uses pointers in string matching.

```
String   T:  A C C T C G A G T
                    ↑| | |↑
Pattern P:   _ _ _ _ C G A _ _
```

In (Alsmadi and Nuser 2012), They evaluated two algorithms for DNA string comparison. The Longest Common Substring (LCS) algorithm, and Longest Common Sub-Sequence (LCSS) algorithms. In the following example, the highlighted letters, CTCT, in the sequences is LCSS of the specified sequences.

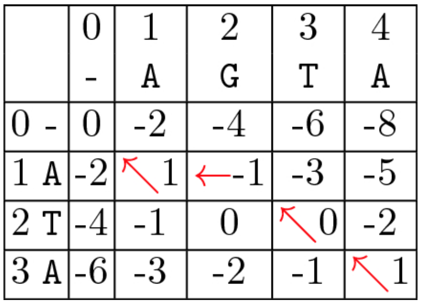
```
String   T:  A C G T C G A G T
               |   | |       |
Pattern P:   _ C _ T C _ _ _ T
```

Different types of string matching algorithms are explored in (Singla and Garg 2012), concluding that for string matching, Boyer Moore algorithm is the best. In Aho-Corasick performs better than the CommentZ-Walter algorithm.

Methodology

Needleman Wunsch Algorithm

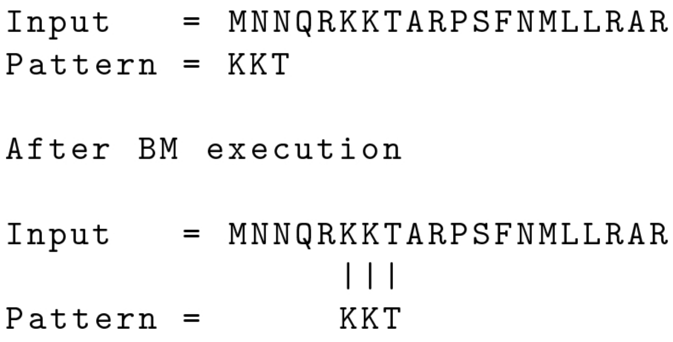
The Needleman-Wunsch (NW) algorithm (Needleman and Wunsch 1970) is a Dynamic Programming (DP) algorithm that solves the problem of sequence alignment. for finding the optimal alignment between two sequences, the maximum score of this function is needed to compute and the alignment that yields it.



	0	1	2	3	4
	-	A	G	T	A
0 -	0	-2	-4	-6	-8
1 A	-2	↖1	←-1	-3	-5
2 T	-4	-1	0	↖0	-2
3 A	-6	-3	-2	-1	↖1

Boyer Moore Algorithm

Boyer Moore algorithm (BM) for string search and match is a standard benchmark algorithm, considered one of the most efficient when the alphabet comprises a small size of characters, used on standard editors to perform string search. So this algorithm is often used in bioinformatics for disease detection.



```
Input   = MNNQRKKTARPSFNMLLRAR
Pattern = KKT

After BM execution

Input   = MNNQRKKTARPSFNMLLRAR
               |||
Pattern =      KKT
```

Results

In this survey are analyzed different string matching algorithms in the context of biological sequence, DNA and Proteins.

Boyer More is faster when using large sequences, avoiding many comparisons. Boyer More in best case scenario complexity is sub-linear. As future work, it is proposed a parallel algorithm for fuzzy string matching, using artificial intelligence neural networks for better performance and accuracy.