

Denmark
Ministry of Higher Education and Science

Aarhus University
Faculty of Natural and Technical Sciences



Bioinformatics Research Center

Master's Thesis

Master in Bioinformatics

MASTER'S THESIS

Submitted to

Faculty of Natural and Technical Sciences
Bioinformatics Research Center

In partial fulfillment of the requirements for the degree of
MASTER IN BIOINFORMATICS

By

Pietro Mariani

ALGORITHMS FOR MULTIPLE SEQUENCE ALIGNMENT

Christian Storm Pedersen

Supervisor

June 2022

Author's Declaration of Originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Pietro Mariani

15-06-2022

Abstract

The **Multiple Sequence Alignment** problem is a fundamental problem for both Biological and Computer Sciences. Gusfield's approximation algorithm is a progressive alignment method that utilizes dynamic programming to compute a MSA. Gusfield's algorithm uses a star-like guide tree to conduct the construction process of the alignment but this guide tree can lose significant information from the sequences. Optimizing and improving the solution to this problem is the aim of many researches. In this project I will explore two new approaches to the problem. Instead of using a star tree I will implement Prim and Kruskal's algorithms to compute a **Minimum Spanning Tree** from a complete graph with nodes representing the sequences and edges representing its pairwise alignment scores. This new guide trees will then be used to direct the merging of new sequences into the MSA. The sequences will be simulated according to three different evolutionary trees with different mutation rates. These two new approaches to the MSA problem show good promise in capturing signals from the underlying structure of the sequences and creating more optimal alignments.

List of Abbreviations and Terms

MSA	Multiple Sequence Alignment
MST	Minimum Spanning Tree
mu	mutation rate

Table of Contents

List of Figures	6
General Introduction	7
1 Theory	8
1.1 Multiple Sequence Alignment	8
1.1.1 Gusfield's algorithm	8
1.2 Minimum Spanning Tree	10
1.2.1 Kruskal	11
1.2.2 Prim	12
2 Methods	13
2.1 General implementation	13
2.2 Data Simulation - <i>Pyvolve</i>	14
2.3 Implementing MST construction algorithms	15
2.3.1 Prim	15
2.3.2 Kruskal	16
2.4 Implementing MSA based on MST	17
2.4.1 MSA merger	18
2.5 MSA score & MSA checker	21
2.5.1 MSA score	21
2.5.2 MSA checker	21
3 Experiments	22
3.1 MST construction time	22
3.1.1 Time experiments	22
3.1.2 Time results	23
3.2 Score	24
3.2.1 Score experiments	24
3.2.2 Score results	25
4 Discussion	33
4.1 MST construction time	33
4.2 MSA score	34
4.2.1 Star tree	34
4.2.2 Ladder tree	35

4.2.3 Symmetrical tree	36
Genreal Conclusion	38
4.3 Conclusion	38
4.4 Future Work	39
References	40

List of Figures

1	Kruskal's first steps, [2]	11
2	Prim's first steps, [2]	12
3	Star tree	15
4	Ladder tree	15
5	Symmetrical tree	15
6	Prim implementation	16
7	Kruskal implementation	17
8	msa_merger() cases	18
9	K_merge() cases example	20
10	Prim's mst construction time based on MSA size	23
11	Kruskal's mst construction time based on MSA size	23
12	Star tree scores with $\mu=0.1$	25
13	Star tree scores with $\mu=0.5$	25
14	Star tree scores with $\mu=1$	25
15	Star tree scores from Prim and Gusfield, $\mu=0.1$	26
16	Ladder tree scores with $\mu=0.1$	27
17	Ladder tree scores with $\mu=0.5$	27
18	Ladder tree scores with $\mu=1$	27
19	Kruskal and Gusfield scores from ladder tree with $\mu=0.1$	28
20	Symmetrical tree scores with $\mu=0.1$	29
21	Symmetrical tree scores with $\mu=0.5$	29
22	Symmetrical tree scores with $\mu=1$	29
23	Star tree with different mutation rates, Prim and Gusfield scores. (a) $\mu=0.1$ (b) $\mu=0.5$ (c) $\mu=1$	30
24	Average scores between trees with $\mu=0.1$	31
25	Average scores between trees with $\mu=0.5$	31
26	Average scores between trees with $\mu=1$	31
27	Star tree	34
28	MSA scores based on mutation rates (star tree)	35
29	Ladder tree	35
30	Symmetrical tree	36
31	MSA scores based on mutation rates (symmetrical tree)	37