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