Approximation Algorithms for Multiple Sequence Alignment based on *l*-stars

By: Yidan Chen

Supervisor: Christian Nørgaard Storm Pedersen

June 2022

Abstract

Multiple sequence alignment is one of the most challenging problems in bioinformatics and is proven to be NP-hard. Heuristic approaches are often used to find approximation solutions with bounded error in a feasible time.

This thesis discusses the l-star structures, their application in multiple sequence alignment problem, and three related approximation algorithms mentioned in *Approximation algorithms for multiple sequence alignment* (Bafna et al., 1994). Those three algorithms, the optimized algorithm, the (2l-1)-stars algorithm and the randomized algorithm are proved to have an approximation ratio of $2 - l/k$ for certain configurations.

Experiments were carried out to examine performance of the approximation algorithms in practice. All of them worked as expected and for fixed l, the running time of the optimized algorithm is exponential while that of the (2l-1)-stars algorithm and the randomized algorithm is polynomial. Due to fact that the calculation of an exact alignment for each clique in an l-star is very time consuming, those algorithms are not considered to be practical when l is large.

Table of Contents

Abstract ..... **iii**

1 Introduction **1**
1.1 Sequence Alignment 1
1.2 Exact Algorithms 4
1.3 Approximation Algorithms 6
1.4 Thesis Outline 7

2 The *l*-stars Method **9**
2.1 *l*-stars 9
2.2 Compactible Alignments 11
2.3 Weighted Edges 12
2.4 Optimal Weighted Alignments 14
2.5 Approximation Ratio 15
2.6 A Naïve Approach 16

3 Optimized *l*-stars Algorithm **19**
3.1 All Possible *l*-stars 19
3.2 Optimization 21
3.3 Summary and Analysis 23

4 (2*l*-1)-stars Algorithm **26**
4.1 A Smaller Balanced Set 26
4.2 Matching Problem 28
4.3 Summary and Analysis 30

5 Randomized *l*-stars Algorithm **32**
5.1 Probability of Bad Performance 32
5.2 A Randomized Algorithm 33
5.3 Summary and Analysis 34

6 Experiments **36**
6.1 Approximation Ratios 36
6.2 Running Time and SP Score 37
6.3 Summary 49

7 Conclusion **51**

References **53**