GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

ISSN 1612-6793

Bachelor's Thesis

submitted in partial fulfillment of the
requirements for the course "Applied Computer Science"

My Title

Robin William Hundt

Institute of Computer Science

Bachelor's and Master's Theses
of the Center for Computational Sciences
at the Georg-August-Universität Göttingen

09. May 2020

Georg-August-Universität Göttingen
Institute of Computer Science

Goldschmidtstraße 7
37077 Göttingen
Germany

☎ +49 (551) 39-172000
FAX +49 (551) 39-14403
✉ office@informatik.uni-goettingen.de
www.informatik.uni-goettingen.de

First Supervisor: Prof. Dr. Burkhard Morgenstern
Second Supervisor: Dr. Peter Meinicke

I hereby declare that I have written this thesis independently without any help from others and without the use of documents or aids other than those stated. I have mentioned all used sources and cited them correctly according to established academic citation rules.

Göttingen, 09. May 2020

Abstract

Here comes the abstract...

Contents

Chapter 1

Introduction

Chapter 2

Basics

2.1 Multiple sequence alignment

Chapter 3

Prior Work

Chapter 4

Algorithm

In this chapter, the analysis of ...

Chapter 5

Implementation

In this chapter, the implementation of ...

Chapter 6

Evaluation

6.1 BAliBASE 3

The third version of the BAliBASE benchmark protein alignment database has been released in 2005 and is widely employed for the comparison of multiple alignment programs [1, 2]. It is constructed in a semi automatic process as shown in fig. 6.1 and suitable to evaluate global and local alignment programs. The database is split into 5 reference sets with different characteristics representing distinctive multiple alignment problems.

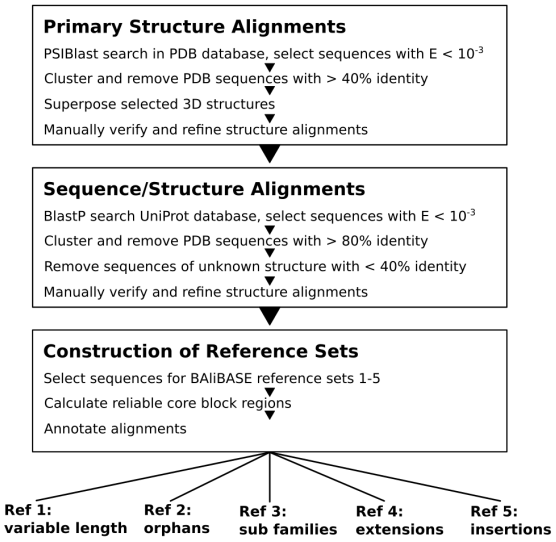

Figure 6.1: Flow chart showing the semi automatic process used to establish the reference sets TODO cite self

- reference set 1 subset V1, for which any two sequences share <20% identity and no internal insertions over 35 residues long
- reference set 1 subset V2, consisting of families with at least four equidistant sequences for which any two sequences share 20-40% identity and no large insertions
- reference set 2, for which all sequences share >40% identity and at least one 3D structure is known. Additionally an "Orphan" sequence with <20% identity is chosen per family
- for reference set 3, all sequences in the same subfamily have >40% identity, whereas sequences from different subfamilies share <20% identity
- for reference sets 4 and 5, every sequence shares

at least 20% with one other sequence, including sequences with large N/C-terminal extensions (ref 4) or internal insertions (ref 5)

6.1.1 Core blocks

6.2 Sum-of-pairs and column score

6.3 MAFFT

6.4 Results

Chapter 7

Conclusion

7.0.1 Further work

Bibliography

[1] J. D. Thompson, P. Koehl, R. Ripp, and O. Poch, "Balibase 3.0: latest developments of the multiple sequence alignment benchmark," *Proteins: Structure, Function, and Bioinformatics*, vol. 61, no. 1, pp. 127–136, 2005.

[2] D. J. Russell, *Multiple Sequence Alignment Methods* -, softcover reprint of the original 1st ed. 2014 ed. unbekannt: Humana Press, 2016.