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

- show table of MSA with gaps inserted, maybe picture from balibase?

Chapter 3

Prior Work

3.1 Gabios-Lib

3.2 Dialign

3.3 Spaced Word Matches

3.3.1 Multi dimension matches

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 alignment benchmark dataset

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. It is divided into:

- 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

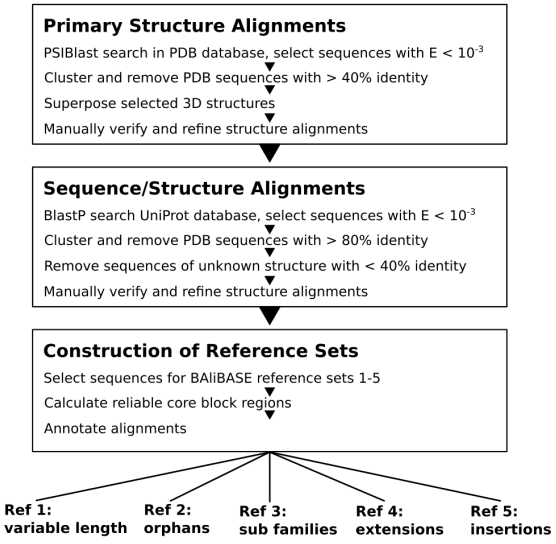

Figure 6.1: Flow chart showing the semi automatic process used to establish the reference sets TODO cite self

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

Evaluating and comparing alignment programs is a difficult problem due to the uncertainty of supposedly "real" alignments of actual sequences. The BAliBASE database marks alignment columns which can be reliably aligned as so called "core blocks". These core blocks are calculated and manually verified, making up 19% of the full length sequences which are used in the evaluation of *spam-align* [1].

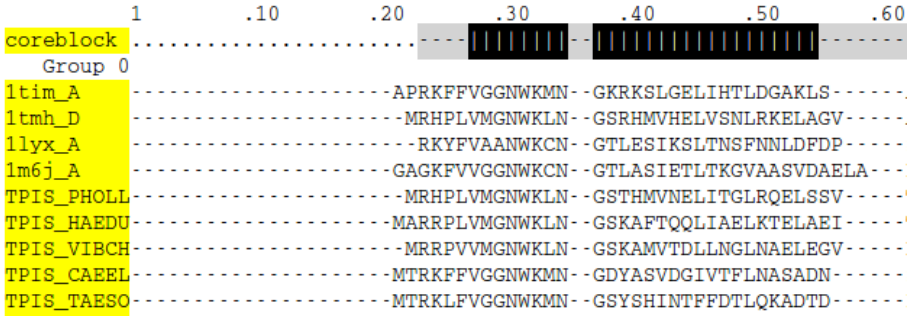

Figure 6.2: BAliBASE web interface. Black columns indicate core blocks. lbgi.fr

6.2 Sum-of-pairs and column score

Comparing the alignment output of different methods can be done by computing the sum-of-pairs and column scores.
Given a test alignment A_t and a reference alignment A_r with M sequences and N_t, N_r columns respectively, the sum-of-pairs and column score is defined according to Thompson et al. [3].

Definition 6.2.1 (Sum-of-pairs score)
The sum of pairs score is the ratio of correctly aligned individual residues. Formally it is defined as:

$$p_{ijk} = \begin{cases} 1 \textit{ if residues } A_{t_ij} \textit{ and } A_{t_ik} \textbf{\textit{ are}} \textit{ aligned in } A_r \\ 0 \textit{ otherwise} \end{cases}$$

$$S_i = \sum_{j=1}^{M} \sum_{k=i+1}^{M} p_{ijk}$$

$$SPS = \frac{\sum_{i=1}^{N_t} S_i}{\sum_{i=1}^{N_r} S_{r_i}}$$

with S_{r_i} being the number of correctly aligned residues in the reference.

Definition 6.2.2 (Column score)
The column score is the ratio of correctly aligned columns.

$$C_i = \begin{cases} 1 \textit{ if all the residues in the i-th column are aligned correctly} \\ 0 \textit{ otherwise} \end{cases}$$

$$CS = \frac{\sum_{i=1}^{N_t} C_i}{N_r}$$

Note that $C_i = 1$ only if all the residues in the i-th column are aligned correctly and no residue belonging to this column is part of another one. For this reason, the numerator is smaller or equal to the denominator.
The definition of the column score is slightly different than that provided by the authors of BAliBASE [3] but resembles the actual implementation in the included BaliScore tool and its reimplementation provided with this thesis.
These scores are only calculated for the core blocks of the BAliBASE alignments, meaning that for the following evaluation A_t is an alignment over the full sequences, while A_r contains only the aligned residues inside the core blocks.

6.3 Evaluated programs

6.3.1 Mafft

Additionally to *Dialign2.2* and *spam-align* the widely used multiple alignment program *MAFFT* (version 7) is evaluated. It employs a progressive alignment strategy

- progressive alignment
- guide tree from all pairwise alignments

6.3.2 Dialign

6.3.3 Spam

6.4 Results

Chapter 7

Conclusion

7.1 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.

[3] J. D. Thompson, F. Plewniak, and O. Poch, "A comprehensive comparison of multiple sequence alignment programs," *Nucleic acids research*, vol. 27, no. 13, pp. 2682–2690, 1999.