Bachelor project: Synopsis
DNA/RNA-sequence clustering algorithm

Anders Kiel Hovgaard Nikolaj Dybdahl Rathcke

February 23, 2015

1 Problem statement

Can we implement an algorithm that can cluster DNA/RNA-sequencing data, of sizes at least half a million sequences of 50-500 characters each, and compete with the performance of `UCLUST`? If not, how close can we get?

1.1 Restrictions on the scope of the project

The algorithm to be implemented might either be one that we design from scratch or a modified version of an existing algorithm. Similarly for the implementation, we might reuse parts of existing open implementation to some extent.

Clustering of protein sequences will not be considered in this project, i.e. only RNA and DNA sequences will be considered.

2 Motivation

The main motivation for this project comes from a need for efficient tools for clustering of sequence data, and related techniques, in the microbiology department at the University of Copenhagen. In particular, the idea for this project comes from a collaboration between the supervisor of this project, Rasmus Fonseca, and Martin Asser Hansen, who is a bioinformatician in the Molecular Microbial Ecology Group[1].

The problem consists of clustering huge amounts of DNA/RNA-sequences (up to 500 million strings of 50-500 characters each), based on a given similarity threshold, so that any one sequence only appears in exactly one cluster.

There are not many available algorithms and tools for efficient clustering of sequencing data. The one tool `UCLUST`[2] that does the job is closed-source and considered too expensive.

[1] `http://www2.bio.ku.dk/microbiology/`

[2] `http://drive5.com/usearch/`

The goal of this project is to research the possibilities of creating an open tool that can match the performance of the proprietary version of `UCLUST`.

3 Tasks and time planning

- Survey existing literature, algorithms and implementations of sequence clustering. (∼1-2 weeks)
- Choose distance metric and parameters, design algorithm for calculating distance. (∼1 week)
- Design initial version of clustering algorithm. (∼1-2 weeks)
- Implement distance metric. (∼1-2 weeks)
- Implement clustering algorithm. (∼3-4 weeks)
- Analyze and profile the implementation, investigate possibilities for optimization and perform these on the implementation. (∼3-4 weeks)
- Perform ongoing testing after implementation of initial algorithm and in-between optimizations and modifications. (∼3-4 weeks)
- Analyze the time and space complexity of the algorithm. (∼1-2 weeks)

Initial algorithm design

- Product: *An initial design of an algorithm that performs clustering on DNA/RNA sequences. Choice of distance metric and parameters.*
- Dependencies: *None.*
- Workload: *10 workdays*

Implementation of distance metric

- Product: *An implementation of the chosen distance metric in* `C/C++` *or similar.*
- Dependencies: *Distance metric from "Initial algorithm design". Ongoing testing and optimization.*
- Workload: *10 workdays*

Algorithm implementation

- Product: *An implementation of the algorithm in `C/C++` or similar.*

- Dependencies: *Clustering algorithm from ”Initial algorithm design” and the product from the task ”Implementation of distance metric”. Ongoing testing and optimization*

- Workload: *20 workdays*

Optimization

- Product: *A modified algorithm with reduced running time.*

- Dependencies: *Ongoing process with the tasks ”Implementation of distance metric”, ”Algorithm implementation” and ”Testing”.*

- Workload: *20 workdays*

Testing

- Product: *Ongoing feedback on the algorithm and optimization.*

- Resources: *Real life data and a powerful computer with similar specifications to what would be used.*

- Dependencies: *Ongoing process with the tasks ”Implementation of distance metric”, ”Algorithm implementation” and ”Optimization”*

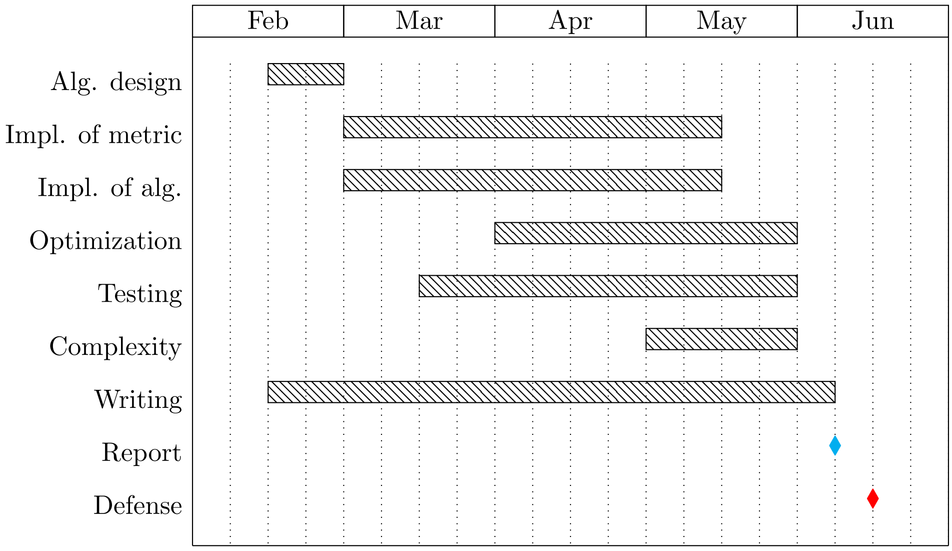
- Workload: *10 workdays*

Time complexity analysis

- Product: *An analysis of the algorithm and possibly major optimizations.*

- Dependencies: *Algorithm and major optimizations.*

- Workload: *5 workdays*


Feb
Mar
Apr
May
Jun
Alg. design
Impl. of metric
Impl. of alg.
Optimization
Testing
Complexity
Writing
Report
Defense