Synopsis

Mehdi Asser Husum Nadif

February 22, 2015

DIKU

Title

Effort to make a quick and precise clustering algorithm for 16S rRNA sequences.

Problem statement

I will attempt to create a clustering algorithm, that can compete with, or even surpass, uClust in terms of speed, without losing precision, when working with a dataset of 500.000 RNA/DNA sequences of length 500-1500 characters.

Limitations

1. I will only attempt to make a clustering algorithm that works for RNA/DNA sequences.
2. I will assume that all biological data is correct, and will not account for any discrepancies in the given data.
3. The efficiency of my solution will be measured against the 64 bit version of uClust; Newer, more efficient algorithms will not be taken into account.

Reason

The amount of biological sequence data is increasing at a higher speed than the growth of computer efficiency, as predicted by Moore's Law. Algorithms that run quicker are therefore indispensable. For instance, in the Microbiology Department of the University of Copenhagen, more than 500 million unprocessed RNA and DNA sequence strings have amassed. They need to be clustered. The most efficient algorithm for clustering huge amounts of sequences is uClust [15], but it is unfortunately quite costly. There is therefore a need for an open source clustering algorithm that is as efficient speed- and precision wise as uClust.

Work Plan

I have set the following assignments for myself:

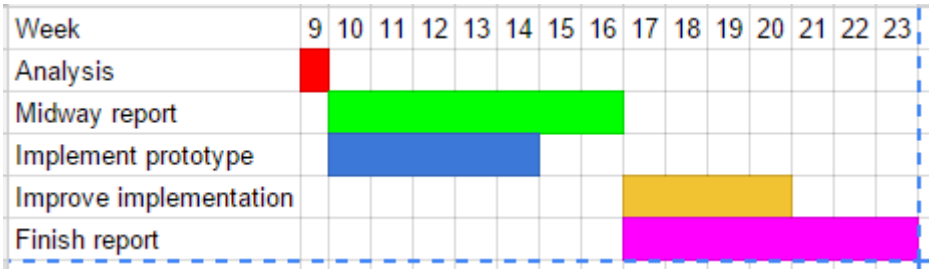
1. **Implement Prototype**
   Product: A working clustering algorithm using the method i hope to finish with. It does not have to be quick or precise, nor parallelized, but it should be working.
   Resources: C (programming language), Notepad++ as editor.
   Dependencies: Analysis of clustering algorithms.
   Workload: 5-7 days.
2. **Improve implementation**
   Product: Improving upon the working algorithm by testing, debugging and parallellization. Might also need to improve small parts of the algorithm, such as database search and the distance metric.
   Resources: C, Notepad++, Articles on optimization of database search

etc.
Dependencies: A finished prototype.
Workload: 8-10 days.

3. **Write midway report**
   Product: Having written the midway report. It should include an introduction to the subjects i've used for the prototype, such as statistics, clustering algorithms, and whatever else might be used.
   Resources: Articles about the subjects that i use, Latex.
   Dependencies: Analysis of clustering algorithms.
   Workload: 10-15 days.

4. **Analysis of clustering algorithms.**
   Product: An analysis of a diverse number of clustering algorithms to find out which method to use for the prototype.
   Resources: Many articles about clustering.
   Workload: 2 days

5. **Write rest of report**
   Product: Any additional background material will be added, otherwise the analysis of the algorithms speed comparatively with other similar algorithms will be added, and the conclusion. Goes along with the improvement of the algorithm.
   Dependencies: Midway Report, Prototype.
   Resources: Latex, Results from tests.
   Workload: 8-10 days.



References

[1] Michail Kazimianec, Arturas Mazeika, *Clustering of Short Strings in Large Databases*, 2009

[2] Trevor Hastie, Robert Tibshirani, Jerome Friedman, *The Elements of Statistical Learning: Datamining, Inference, and Prediction*, 2. ed., Springer Series in Statistics, 2008

[3] Anil K. Jain, Richard C. Dubes, *Algorithms for Clustering Data*, Michigan State University, 1988.

[4] Yaqing Si, Peng Liu, *et al.*, *Model-based clustering for RNA-seq data.*, November 4. 2013.

[5] David J Russell, Samuel F Way, Andrew K Benson, Khalid Sayood, *A grammar-based distance metric enables fast and accurate clustering of large sets of 16S sequences*, BMC 2010.

[6] Robert C. Edgar, *Search and clustering order of magnitude faster than BLAST*, 2010.

[7] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, Weizhong Li, *CD-HIT: accelerated for clustering the next-generation sequencing data*, 2012

[8] Weizhong Li *et al.*, *Ultrafast clustering algorithm for metagenomic sequence analysis*, 2012

[9] Xiao Yang, J. Zola, *et al.*, *Large Scale Metagenomic Sequence Clustering via Sketching and Maximal Quasi-clique Enumeration on Map-Reduce Clusters*, 2012

[10] Zeehasham Rasheed, Huzefa Rangwala, *A Map-Reduce Framework for Clustering Metagenomes*, 2013

[11] Michael Farrar, *Striped Smith-Waterman speeds database searches six times over other SIMD implementations*, 2006

[12] Andrei Z. Broder, *et al.*, *Min-Wise Independent Permutations*, 1998

[13] Mohammadreza Ghodsi, Bo Liu, Mihai Pop, *DNACLUST: accurate and efficient clustering of phylogenetic marker genes*, 2011

[14] Chen W, Zhang CK, Cheng Y, Zhang S., Zhao H, *A Comparison of Methods for Clustering 16S rRNA Sequences into OTUs*, 2013

[15] Leonard M. Adleman, *Computing with DNA*, Scientific America, 1998.

[16] Jianqiu Ji and Jianmin Li and Shuicheng Yan and Qi Tian and Bo Zhang, *Min-Max Hash for Jaccard Similarity*, 2013