600.639 Computational Genomics
Final Project

Ashleigh Thomas and Jed Estep

Abstract

Easily searchable representations of genome strings are useful for many kinds of analysis, but in practice their usability is often limited on commodity hardware due to their high memory requirements. Suffix arrays are one of the least memory-intensive commonly used representations, but its space requirements may still be prohibitive in the case of indexing numerous genomes. In this paper we investigate the suffix array compression scheme described in [1]. We attempt to apply the compressed suffix arrays as a searchable database of multiple genomes with use in the context of metagenomics. The design of our database is similar to that of QUASAR [2].

1 Introduction

We arrived at this design while investigating multiple topics. From one end, we were interested in pursuing the applicability of compressed data structures to representing genomes. Many implementations of useful index structures like suffix trees are extremely memory intensive, so decreasing the size of their representation is paramount if they are to be used on commodity hardware. Literature on the topic of succinct data structures often neglects to discuss practical versions of their structures, and as such we explore how well the methods of [1] work in a real program.

From an alternative angle, we noted that most approaches to metagenomics rely on probabilistic methods, such as [3], and less attention is given to index search methods that are commonly used for read alignment.

2 Prior Work

Burkhardt [2] points out that, in the operation of QUASAR, numerous special methods are necessary to accommodate suffix arrays which are too large to fit in main memory. As such, we attempted to apply the compression methods of Grossi and Vitter [1] to a search index similar to QUASAR.

References

[1] Roberto Grossi and Jeffrey Vitter, *Compressed Suffix Arrays and Suffix Trees with Applications to Text Indexing and String Matching.* Society for Industrial and Applied Mathematics Journal of Computing, Vol. 35, No. 2, pp. 378-407, 2005.

[2] Stefan Burkhardt, et al., *q-gram Based Database Searching Using a Suffix Array (QUASAR).* Proceedings of the third annual international conference on Computational molecular biology, pp. 77-83, 1999.

[3] Arthur Brady and Steven Salzberg, *Phymm and PhymmBL: Metagenomic Phylogenetic Classification with Interpolated Markov Models.* Nature Methods, 2009.