# Synopsis for Bachelorproject

## Regular Expression Matching In Genomic Data

Rasmus Haarslev - nkh877
Troels Thomsen - qvw203

23. Februar 2015

Department of Computer Science
University of Copenhagen

# 1 Problem definition

We wish to determine the possibility of converting sequence analysis patterns used for scan-for-matches[2], into regular expressions and test their efficiency against the KMC[1] engine.

Specifically we wish to solve the following problems:

- Is it possible to programatically convert patterns used by the scan-for-matches program into regular expressions for the KMC engine? If not all patterns used by scan-for-matches then which ones?
- Is it possible to achieve speeds matching or exceeding scan-for-matches with the generated regular expressions and the KMC engine?
- Are there features missing from the KMC engine (such as backtracking), which if they were present would yield better performance in the case of these specific patterns?

## 1.1 Limits

- We will not attempt to modify the KMC engine in any regard.

# 2 Motivation

Institute for Bioinformatics have a allocated, and are still allocating, a lot of DNA sequencing data. Currently they have around a total of 2 petabytes of data. These sequences of DNA contain a lot of information, but searching through the data currently uses the scan-for-matches program, which while performing very well, is not very user friendly and have some unfortunate limitations when running many consecutive scans, since it performs I/O operations for every run.

Recently Fritz' group have developed a regular expression engine called KMC, which so far have performed five times better than current industry standard engines. Since scan-for-matches outperforms NR-grep[1], we are hoping that optimized regular expressions running on KMC will be able to outperform NR-grep and subsequently scan-for-matches.

[1] Kleene Meets Church

# References

[1] Gonzalo Navarro. Nr-grep: A fast and flexible pattern matching tool. http://www.dcc.uchile.cl/~gnavarro/ps/spe01.pdf. Visited 18th February 2015.

[2] The SEED Team. Scan for matches. http://blog.theseed.org/servers/2010/07/scan-for-matches.html. Visited 18th February 2015.