

# 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[1], into regular expressions and test their efficiency against the KMC<sup>1</sup> 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.

---

<sup>1</sup>Kleene Meets Church

## References

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