Bachelor Thesis

Synopsis
Optimized pattern matching in genomic data

Martin Westh Petersen - mqt967
Kasper Myrtue - vkl275

23. Februar 2015

1 Problem definition

Is it possible create a program with the same functionality as scan_for_matches, but with an average increase in performance of at least 50% for patterns that consists of more than 1 pattern unit?

2 Limitations

3 Motive

Pattern matching functionality for strings in genomic data is very useful, but requires a good performance due to huge amounts of data. Scan_for_matches serves this purpose, but has never been formally analyzed with respect to complexity and performance and the code is poorly documented and hard to read. A significant increase in performance is also possible (as the author noted himself in the README file in the source package). These optimizations are the foundation for this project.