Midtvejsrapport Bachelorproject

Regular Expression Matching In Genomic Data

Rasmus Haarslev - nkh877
Troels Thomsen - qvw203

Supervisors: Rasmus Fonseca

23. Februar 2015

Department of Computer Science
University of Copenhagen

Contents

1 Problem definition

We wish to determine the possibility of converting sequence analysis patterns used for scan-for-matches[**?**], into regular expressions[**?**] and test their efficiency against the KMC[**?**] 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?
- Can we find weak extensions to regular expressions, which would enable us to support more or all scan-for-matches patterns?

2 Translation of scan-for-matches patterns into regular expressions

2.1 Mismatches, Insertions and Deletions

The subpatterns can have the following form, which allows for mismatches, insertions and deletions in the subpattern.

$$x_1 \dots x_n[m, i, d] \quad \{x_1 \dots x_n \in A \mid m, i, d \in \mathbb{N}_0\} \tag{1}$$

This notation allows for 0 or the given number of mismatches, insertions or deletionsin, or all possible combinations in between. This means we for an example can have m mismatches and i insertions, but not necessarily d deletions.

We can combine the matching of mismatches and deletions to somewhat simplify our expressions. We can express one mismatch and one deletion in the following regular expression.

$$\begin{array}{rcl}
((x_1? \mid [\hat{\ }x_1])\ x_2 \cdots x_n) \mid (x_1?\ [\hat{\ }x_2]\ x_3 \cdots x_n) \mid & \cdots & \mid (x_1? \cdots x_{n-1}\ [\hat{\ }x_n]) \mid \\
([\hat{\ }x_1]\ x_2? \cdots x_n) \mid (x_1\ (x_2? \mid [\hat{\ }x_2])\ x_3 \cdots x_n) \mid & \cdots & \mid (x_1\ x_2? \cdots x_{n-1}\ [\hat{\ }x_n]) \mid \\
& \vdots & \\
([\hat{\ }x_1]\ x_2 \cdots x_n?) \mid (x_1\ [\hat{\ }x_2]\ x_3 \cdots x_n?) \mid & \cdots & \mid (x_1 \cdots x_{n-1}\ (x_n? \mid [\hat{\ }x_n]))
\end{array} \tag{2}$$

This quite large expression quickly grows into an even larger expression, if we have multiple mismatches or deletions. For two mismatches, we have the following. $\hookrightarrow$ denotes the continuation of the current line.

$$\begin{array}{l}
((x_1 \mid [\hat{\ }x_1])\ (((x_2 \mid [\hat{\ }x_2])\ x_3 \cdots x_n) \quad \mid \quad (x_2\ (x_3 \mid [\hat{\ }x_3])\ x_4 \cdots x_n) \\
\hookrightarrow \mid\ \cdots\ \mid (x_2 \cdots x_{n-1}\ (x_n \mid [\hat{\ }x_n])))) \mid \\
(x_1\ (x_2 \mid [\hat{\ }x_2])\ (((x_3 \mid [\hat{\ }x_3])\ x_4 \cdots x_n) \quad \mid \quad (x_3\ (x_4 \mid [\hat{\ }x_4])\ x_5 \cdots x_n) \\
\hookrightarrow \mid\ \cdots\ \mid (x_3 \cdots x_{n-1}\ (x_n \mid [\hat{\ }x_n])))) \mid \\
\qquad\qquad\qquad \vdots \\
(x_1 \cdots x_{n-1}(x_n \mid [\hat{\ }x_n]))
\end{array} \tag{3}$$

In the case of $m, i, d = n$, we have the following case.

$$.?\ (x_1?\ |\ [\hat{}x_1])\ .?\ (x_2?\ |\ [\hat{}x_2]) \cdots .?\ (x_n?\ |\ [\hat{}x_n])\ .? \tag{4}$$

2.2 Backreferencing

One way to simplify our expression, is to use backreferences. With backreferences we can express m mismatches, d deletions and i insertions in the following regular expressions.

2.2.0.1 Mismatches

$$((x_1?|.)\ |\ (x_2?|.)\ |\ \cdots\ |\ (x_n?|.)) \tag{5}$$

2.2.0.2 Deletions

2.2.0.3 Insertions