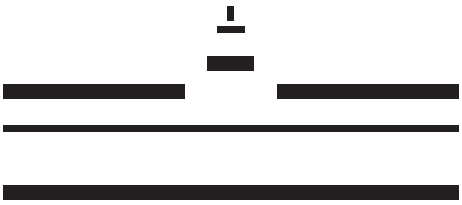

Westfälische
Wilhelms-Universität
Münster

An efficient error correction algorithm using FM-index

Seminararbeit

vorgelegt von:
Eike Gebauer

Matrikelnummer: 408126

Studiengang: MSc. Informatik

Zusammenfassung

An dieser Stelle soll eine Zusammenfassung der wissenschaftlichen Arbeit stehen. Idealerweise ist die Zusammenfassung genau eine Seite lang, was in etwa 350 - 450 Wörtern entspricht. Sie bildet zusammen mit dem Fazit einen Rahmen der gesamten wissenschaftlichen Arbeit und sollte auch für Außenstehende einfach zu verstehen und ansprechend formuliert werden.

Inhaltsverzeichnis

1 Einleitung

DNA Sequenzierung ist aus der biologischen und medizinischen Forschung nicht mehr wegzudenken. Die Entwicklung immer günstigerer und leistungsstärkerer Sequenzierungsverfahren hat in den letzten Jahrzenten für einen rapiden Anstieg der zu verarbeitenden Datenmengen gesorgt. Hinzu kommt, dass die verwendeten Verfahren nicht fehlerfrei sind, was die Verarbeitung der Daten weiter erschwert.

In dieser Arbeit wird ein Paper behandelt, in dem der Fehlerkorrekturalgorithmus FMOE vorgestellt wird. Zu erst wird aber der Hintergrund in den Bereichen Biologie und Bioinformatik erläutert.

1.1 DNA-Sequenzierung

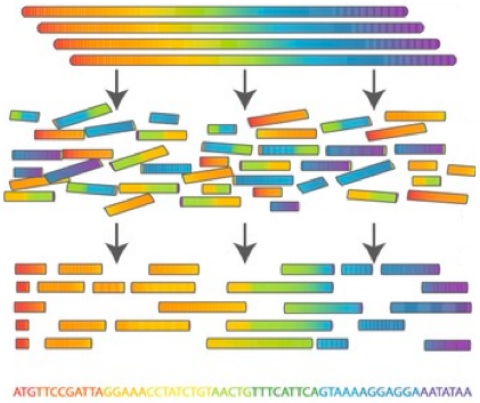

Abbildung 1.1

Das Ziel der de-novo DNA-Sequenzierung ist die Bestimmung der Nukleotidfolge in DNA ohne ein bereits bekanntes Referenzgenom. Wo bei alten Verfahren noch alle Basen nacheinander bestimmt wurden, werden seit dem sogenannten Shotgun Sequencing viele Teilstücke der DNA parallel analysiert. Das macht es möglich, lange DNA Stränge mit Verfahren zu sequenzieren, die nur kurze Fragmente verarbeiten können. Gleichzeitig ermöglicht es die parallele Sequenzierung von mehreren Fragmenten, was den Vorgang extrem beschleunigt.

Die Reihenfolge der DNA-Fragmente aufrecht zu erhalten, ist möglich, aber relativ langsam und aufwendig. Die sogenannten next-gen Sequenzierungsver-

fahren, welche den aktuell höchsten Durchsatz erzielen, verzichten daher darauf. Um trotzdem auf die ursprüngliche Reihenfolge und damit auf die gesamte DNA-Sequenz zurückschließen zu können, wird das Verfahren für mehrere Kopien desselben DNA-Strangs durchgeführt.

Man trennt also mehrere Kopien des Strangs an zufälligen Positionen in ungefähr gleich lange Fragmente. Alle Fragmente können unabhängig von einander - und damit parallel - sequenziert werden. Das Ergebnis der Sequenzierung eines Fragments wird Read genannt. Aufgrund der zufälligen Trennungspositionen überlappen sich die Reads häufig. Vorausgesetzt die Abdeckung reicht aus, kann man mit den Überlappungen die Reads in die ursprüngliche Reihenfolge bringen und so die gesuchte DNA-Sequenz bestimmen. Dieser Prozess heißt Assembly. Wie bereits erwähnt, sind die aktuellen Verfahren zur Bestimmung der Reads nicht fehlerfrei. Es kann zum Überspringen (Deletion) oder Vertauschen (Substitution) richtiger Basen oder zur Einfügung (Insertion) falscher Basen kommen Kollektiv werden diese Fehler Indels genannt Indels können für unterschiedlich große Probleme während der Assembly sorgen Fehlerhafte Reads können aufgrund mangelnder Überlappung nicht korrekt eingeordnet werden, dann spricht man von fragmentierter Assembly Größere Probleme entstehen aber, wenn Reads trotz Fehlern zu einer validen Assembly führen Bei dieser sog. Misassembly erhält man zum Schluss unter Umständen eine valide, aber falsche Sequenz Diese Komplikationen zu verhindern, ist Aufgabe der Fehlerkorrektur

Man sollte direkt jedes Kapitel, Unterkapitel und jede Formel mit einem Label versehen `\label{}` um eine konsistente Referenzierung im gesamten Dokument zu ermöglichen. Eine Referenzierung im Fließtext lässt sich mittels `\ref{}` umsetzen.

1.2 Zeilenumbrüche und Absätze

Zeilenumbrüche sollten im Quelltext immer mittels einer Leerzeile umgesetzt werden, um eine automatische Texteinrückung in der darauffolgenden Zeile zu ermöglichen. Dies macht das Lesen der Arbeit einfacher. Auf die Verwendung des LaTeX-Kommandos `\\` sollte verzichtet werden.

1.3 Einfügen von Grafiken

Grafiken sollten mittels der `\figure`-Umgebung eingebettet werden. Um die Druckqualität und Wiederverwendbarkeit der Grafiken für Vorträge, Poster, etc. zu erhöhen sind Vektorgrafiken (z.B. .eps oder .pdf) zu bevorzugen. Zur Erzeugung und Konvertierung von Vektorgrafiken ist die OpenSource Software *Inkscape* zu empfehlen.

Um mehrere Bilder horizontal anzuordnen, sollte die `\subfigure`-Umgebung

(a) Das Logo (b) Noch einmal das Logo

Abbildung 1.2: Das Logo der WWU

verwendet werden. Diese Bilder können dann mit `\subref` oder `\ref` referenziert werden und erscheinen im Text als (a) oder 1.2(a).

1.4 Wissenschaftliches Zitieren

Für das Referenzieren von wissenschaftlicher Literatur wie Fachbüchern, Konferenzpapern und Veröffentlichungen in Wissenschaftsmagazinen gibt es unterschiedliche Konventionen. Je nach Fachrichtung weichen Layout und Zitationsstil sehr stark voneinander ab [1]. Wir empfehlen aus Gründen der Einheitlichkeit die Verwendung der *Bibtex*-Umgebung. Die zitierte Literatur kann ausgelagert in einer Datei (z.B.: Quellen.bib) gepflegt werden und mittels des `\cite`-Kommandos referenziert werden.

1.5 Zusammenfassung

Zum Ende eines längeren Kapitels bietet es sich häufig an eine Zusammenfassung der wichtigsten Punkte zu liefern. Dies erleichtert das Lesen und den Übergang zum nächsten Kapitel.

2 Methodik

Im Methodik-Kapitel werden die mathematischen Ausführungen des Verfahrens bzw. des Algorithmus vorgestellt, jedoch technische Details zur Umsetzung und Implementierung (falls nötig) auf ein darauffolgendes Kapitel verlagert.

2.1 Mathematische Notation

Mathematische Formeln können mittels der `\begin{align}...\end{align}` Umgebung gesetzt werden:

$$f(n) = \begin{cases} n/2, & \text{wenn } n \text{ gerade,} \\ 3n+1, & \text{wenn } n \text{ ungerade.} \end{cases} \tag{2.1}$$

$$g(n) = \frac{n}{2} \tag{2.2}$$

2.2 Algorithmus

Eigene Algorithmen beschreibt man am Besten mit Hilfe von Pseudo-Code und dem Paket `algorithm`.

Algorithm 1 Algorithmus

Require: Argument $n \in \mathbb{N}$
  $a = 0$
  for $i = 0, \ldots, n$ **do**
    $a = a + 1$
  end for
  return a

3 Implementierung

In diesem Kapitel werden Technische Details, wie zum Beispiel die Verwendete Laufzeitumgebung, Laufzeitanalyse oder wichtige Implementierungsdetails behandelt. Dabei sollte beachtet werden inwiefern diese Angaben für die Arbeit von Bedeutung sind. Die Namen einzelner Funktionsaufrufe oder Klassendiagramme sind zum Beispiel im Allgemeinen für den Leser uninteressant.

3.1 Quelltext-Abschnitte

Eigene Auszüge aus dem Quelltext bindet man am Einfachsten mit dem Befehl `\lstinputlisting{}` ein. Das Ergebnis ist in 3.1 zu sehen. Optionen für die Sprache, Tab-Breite etc. von `\lstinputlisting` können auch am Anfang des Dokuments mit `\lstset{}` gesetzt werden. Das Paket erlaubt mit `firstline=...` und `lastline=...` auch die Einbindung von einzelnen Zeilen einer Datei.

Quelltext 3.1: Codebeispiel

```
int myTest(n) {
  a = 0;
  for (int i = 0; i <= n; i++)
      a = a + 1;
  return a;
}
```

4 Ergebnisse

Dieses Kapitel sollte die Ergebnisse beinhalten, die mit den Methoden aus Kapitel 2 erstellt wurden.

Tabelle 4.1: Beispieltabelle

Spalte1	Spalte2	Spalte3
1	2	3

5 Diskussion

In diesem Kapitel werden die zuvor vorgestellten Ergebnisse der Arbeit diskutiert. Häufig wird es mit dem Fazit zusammengelegt.

6 Fazit

Dieses Kapitel bildet die abschließende Zusammenfassung der Arbeit. Dazu können die folgende Punkte behandelt werden:

- Reflexion: wurden die Ziele der Arbeit erreicht?
- mögliche Erweiterungen und Verbesserungen („future work“)

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[1] Jele, Harald: *Wissenschaftliches Arbeiten: Zitieren.* R. Oldenbourg Verlag, 2010

Eidesstattliche Erklärung

Hiermit versichere ich, dass die vorliegende Arbeit über *„Titel“* selbstständig verfasst worden ist, dass keine anderen Quellen und Hilfsmittel als die angegebenen benutzt worden sind und dass die Stellen der Arbeit, die anderen Werken – auch elektronischen Medien – dem Wortlaut oder Sinn nach entnommen wurden, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht worden sind.

Vorname Nachname, Münster, 13. Januar 2019

Ich erkläre mich mit einem Abgleich der Arbeit mit anderen Texten zwecks Auffindung von Übereinstimmungen sowie mit einer zu diesem Zweck vorzunehmenden Speicherung der Arbeit in eine Datenbank einverstanden.

Vorname Nachname, Münster, 13. Januar 2019