Proseminar:
Burrows Wheeler Transformation, Suffix Arrays und ihre Anwendungen in der Bioinformatik

Suffix Arrays

Nong, Ge, Sen Zhang, and Wai Hong Chan. "**Two efficient algorithms for linear time suffix array construction.**" IEEE transactions on computers 60.10 (2010): 1471-1484.

Nong, Ge; Zhang, Sen; Chan, Daricks Wai Hong: **Linear Suffix Array Construction by Almost Pure Induced-Sorting**.” Proceedings of the Data Compression Conference, 2009.

Zitiert durch: 180/187

Kurzbeschreibung:
SA-IS:

- Splitten des Strings in die “LMS-Strings“
- Rekursive Sortierung mit Radix-Sort

Kärkkäinen, Juha, Peter Sanders, and Stefan Burkhardt. "**Linear work suffix array construction**." Journal of the ACM (JACM) 53.6 (2006): 918-936.

Zitiert durch: 461

Kurzbeschreibung:
Skew-Algorithm:

- Suffix Array der Positionen MOD 3==0
- Suffix Array der übrigen Positionen
- Mergen der Suffix Arrays

Ko, Pang, and Srinivas Aluru. "**Space efficient linear time construction of suffix arrays**." Annual Symposium on Combinatorial Pattern Matching. Springer, Berlin, Heidelberg, 2003.

Zitiert durch: 396

Kurzbeschreibung:

- Grundidee ähnlich zu SA-IS, aber zeitlich früher
- Teilt den String auch in L- und S-Type Suffixes auf

- Rekursive Sortierung nach dem Aufteilen, dann aber anders als bei SA-IS (über die so-genannten S-Distances)

Schürmann, Klaus-Bernd, and Jens Stoye. "**An incomplex algorithm for fast suffix array construction.**" Software: Practice and Experience 37.3 (2007): 309-329.

Zitiert durch: 91

Kurzbeschreibung:

- Suffixe werden anhand ihrer Präfixe mit fixer Länge d in Buckets eingeteilt; Buckets bekommen je nach repräsentiertem String eine ID, die die lexikographische Ordnung wiedergibt (Phase 1)
- In Phase 2 werden Buckets der Größe >1 rekursiv sortiert. Hierbei wird auf die Sortierung der Buckets untereinander zurückgegriffen. Dadurch entstehen immer kleinere Buckets und immer längere sortierte Präfixe

---

Kärkkäinen, Juha, and Dominik Kempa. "**Engineering a lightweight external memory suffix array construction algorithm**." Mathematics in Computer Science 11.2 (2017): 137-149.

Kärkkäinen, Juha, Dominik Kempa, and Simon J. Puglisi. "**Parallel external memory suffix sorting**." Annual Symposium on Combinatorial Pattern Matching. Springer, Cham, 2015.

Zitiert durch: 34/42

Kurzbeschreibung:

- Suffix Arrays werden blockweise berechnet und dann gemischt
- Das zweite Paper erweitert diesen Ansatz, indem die SAs der Blöcke parallel und mit Hilfe von externem Speicher berechnet werden, um so einen sehr schnelles Verfahren für sehr große Daten zu erhalten.

BWT

Kärkkäinen, Juha. "**Fast BWT in small space by blockwise suffix sorting**." Theoretical Computer Science 387.3 (2007): 249-257.

Zitiert durch: 109

Kurzbeschreibung:
BWT-Erzeugung durch SAs

- Um das Verfahren effizient zu halten, werden die SAs nicht für die komplette Eingabe berechnet, sondern die Eingabe wird in Blöcke geteilt, dafür jeweils die SAs berechnet und daraus dann die BWT des gesamtem Strings abgeleitet
- Verfahren ist zwar langsamer als andere Verfahren, aber benötigt dafür deutlich weniger Speicher

---

Sirén, Jouni. "**Burrows-Wheeler transform for terabases.**" 2016 Data Compression Conference (DCC). IEEE, 2016.

Zitiert durch: 23

Kurzbeschreibung:

- BWT für Textfragmente berechnen und dann zu einer “Gesamt-BWT” mischen

---

Kufleitner, Manfred. "**On bijective variants of the Burrows-Wheeler transform.**" arXiv preprint arXiv:0908.0239 (2009).

Zitiert durch: 59

Kurzbeschreibung:
Variante der BWT:

- Es wird nicht nach komplettem Suffix rotiert und sortiert, sondern es wird eine Dekomposition in „Lyndon Words“ (=Substrings mit aufsteigender Buchstabenfolge) durchgeführt, und für jedes Lyndon Word die BWT einzeln konstruiert.

DeBruijn Graphs

Baier, Uwe, Timo Beller, and Enno Ohlebusch. "**Graphical pan-genome analysis with compressed suffix trees and the Burrows–Wheeler transform**." Bioinformatics 32.4 (2016): 497-504.

Zitiert durch: 58

Kurzbeschreibung:

- Erzeugung des „Compressed deBruijn Graph“ (Einmalige Pfade aus Knoten werden zu einem Knoten zusammengefasst) aus „Pangenomen“ (Vielzahl ähnlicher Genome bzw. Genom+Variationen)
- 2 Varianten
  - Suffix Tree
  - BWT
- Online verfügbar unter: https://academic.oup.com/bioinformatics/article/32/4/497/1743785

Simpson, Jared T., and Richard Durbin. "**Efficient construction of an assembly string graph using the FM-index**." Bioinformatics 26.12 (2010): i367-i373.

Zitiert durch: 322

Kurzbeschreibung:

- Erzeugung des „Overlap Graphs“ (Knoten für den überlappenden Teil zweier Reads, Kanten mit Label für den Präfix/das Suffix vor/nach dem gemeinsamen Teil)
- Wird mit Hilfe des FM-Indexes berechnet.
- Online verfügbar unter: https://academic.oup.com/bioinformatics/article/26/12/i367/286190

Bowe, Alexander, et al. "**Succinct de Bruijn graphs**." International workshop on algorithms in bioinformatics. Springer, Berlin, Heidelberg, 2012.

Zitiert durch: 194

Kurzbeschreibung:

- Komprimierte (Succinct) Darstellung des DeBruijn Graphen
- 5 Operationen werden unterstützt:
  - Outdegree(v)
  - Outgoing(c,v)
  - Indegree(v)
  - Incoming(c,v)
  - Index(s)

Genome applications

Li, Heng, and Richard Durbin. "**Fast and accurate short read alignment with Burrows–Wheeler transform.**" bioinformatics25.14 (2009): 1754-1760.

Zitiert durch: 35254

Kurzbeschreibung:

- Benutzt BWT (und verdeutlicht die Ideen an einer Art Präfix Tree)
- Alignment von Short Reads auf großem Referenzgenom, unter Berücksichtigung von kleinen Abweichungen

---

Sirén, Jouni, Niko Välimäki, and Veli Mäkinen. "**Indexing graphs for path queries with applications in genome research**." IEEE/ACM Transactions on Computational Biology and Bioinformatics 11.2 (2014): 375-388.

Zitiert durch: 214

Kurzbeschreibung:

- Alignment von Reads auf Pangenom-Repräsentation (Pangenom = Referenzgenom + bekannte Variationen)
- Generalized Compressed Suffix Array (GCSA) und eine BWT-basierte Variante für den Pangenom-Graphen um so einen Index zu erhalten, auf dem effizient nach möglichen Alignment-Positionen der Reads gesucht werden kann.

Maciuca, Sorina, et al. "**A natural encoding of genetic variation in a burrows-wheeler transform to enable mapping and genome inference**." International Workshop on Algorithms in Bioinformatics. Springer, Cham, 2016.

Zitiert durch: 62

Kurzbeschreibung:

- BWT-basierte Darstellung eines Referenzgenoms + Variationen
- Kodierung mittels FM-Index (BWT + WT)
- Erlaubt eine „Variation-aware“ Backward-Suche auf der BWT