Karlsruhe Institute of Technology

Text Indexing

Lecture 00: Course Overview

Florian Kurpicz

The slides are licensed under a Creative Commons Attribution-ShareAlike 4.0 International License: www.creativecommons.org/licenses/by-sa/4.0 | commit 05dc783 compiled at 2023-10-23-09:05

Organizational Matters

Lectures

- Monday 14:00–15:30 (50.34, 236)
- lecture only

Organizational Matters

Lectures

- Monday 14:00–15:30 (50.34, 236)
- lecture only

Project (mandatory)

- topics will be handed out 06.11.2021
- coding project and small presentation
- 20 % of the final grade

Organizational Matters

Lectures

- Monday 14:00–15:30 (50.34, 236)
- lecture only

Project (mandatory)

- topics will be handed out 06.11.2021
- coding project and small presentation
- 20 % of the final grade

Oral Exam

- 20 minutes
- 80 % of the final grade
- pizza marks content not relevant for exam

Organizational Matters

Lectures

- Monday 14:00–15:30 (50.34, 236)
- lecture only

Project (mandatory)

- topics will be handed out 06.11.2021
- coding project and small presentation
- 20 % of the final grade

Oral Exam

- 20 minutes
- 80 % of the final grade
- pizza marks content not relevant for exam 

Office Hours (Room 210)

- Monday 15:30–16:00 (lecture period)
- by appointment (otherwise)

Materials

Slides

- published shortly before the lecture (`https://algo2.iti.kit.edu/4612.php`)

Videos

- online for old lectures, new topics will be recorded

Materials

Slides

- published shortly before the lecture (`https://algo2.iti.kit.edu/4612.php`)

Videos

- online for old lectures, new topics will be recorded

Additional Material

- references to literature included
- books
  - Gonzalo Navarro. *Compact Data Structures - A Practical Approach*. Cambridge University Press, 2016. ISBN: 978-1-10-715238-0
  - Enno Ohlebusch. *Bioinformatics Algorithms: Sequence Analysis, Genome Rearrangements, and Phylogenetic Reconstruction*. Oldenbusch Verlag, 2013. ISBN: 978-3000413162
- most likely no script

Content

Fundamentals

- tries
- suffix tree
- suffix array
- longest common prefix array
- Burrows-Wheeler transform (BWT)
- wavelet tree (+ bit vector rank/select)
- FM-index

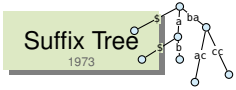
Compressed Indices

- compressing the BWT and wavelet trees
- Lempel-Ziv 77/78 compression
- LZ compression vs. BWT compression
- compressed suffix trees and suffix arrays
- r-index

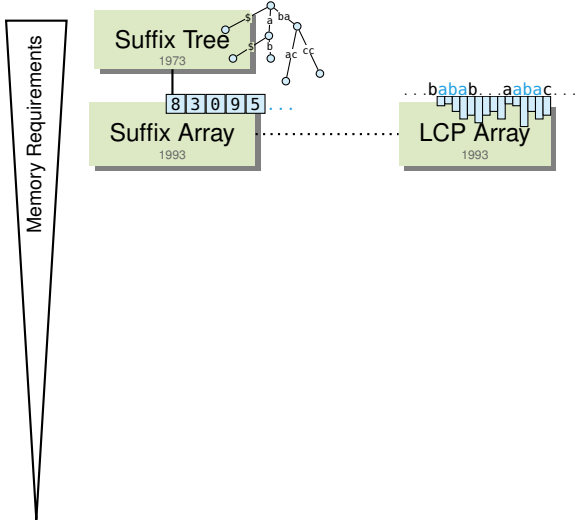
Additional Topics

- parallel construction
- different query types

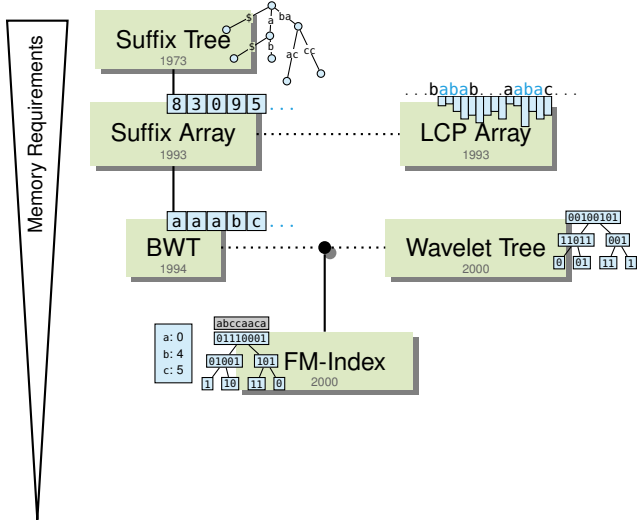
From the Suffix Tree to the *r*-Index



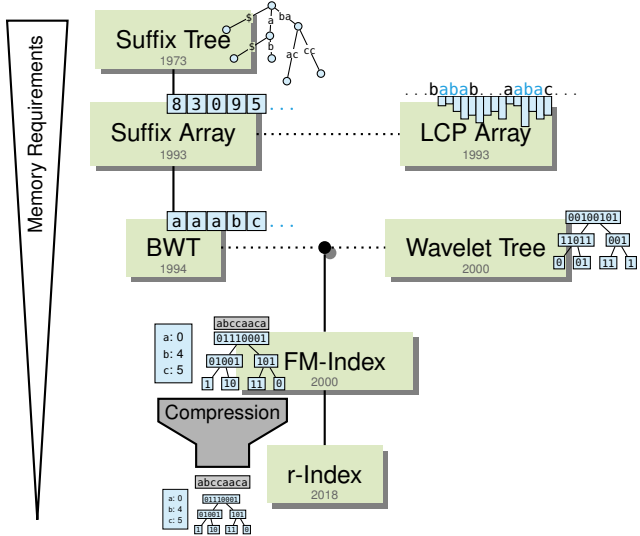
From the Suffix Tree to the *r*-Index



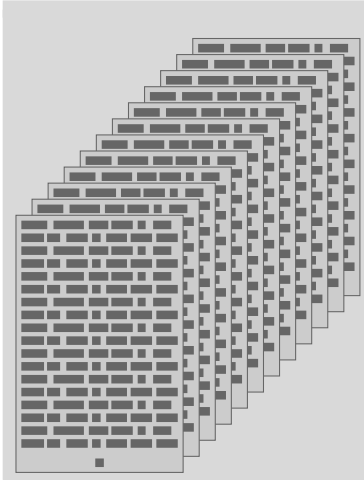
From the Suffix Tree to the *r*-Index



From the Suffix Tree to the *r*-Index

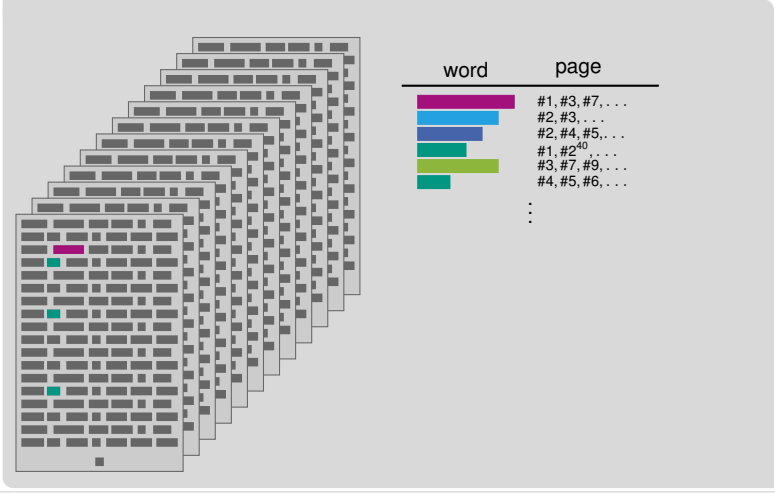


Motivation for Text Indices



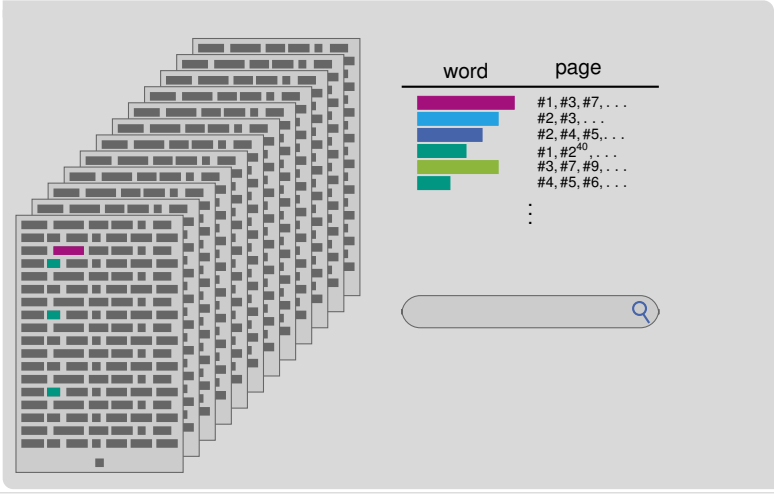
- collection of text
- scanning not feasible

Motivation for Text Indices



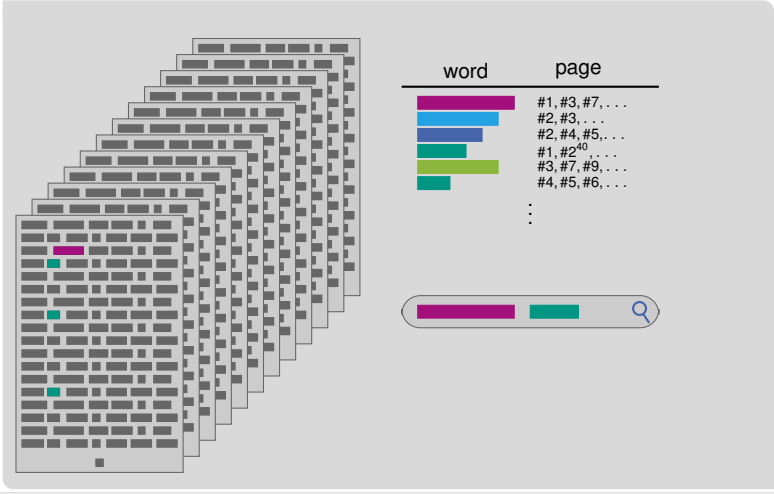
- collection of text
- scanning not feasible
- inverted index (word based)

Motivation for Text Indices



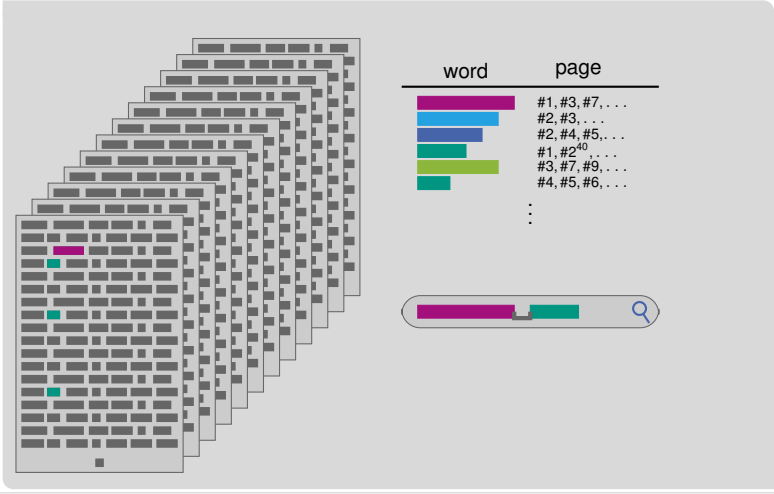
- collection of text
- scanning not feasible
- inverted index (word based)

Motivation for Text Indices



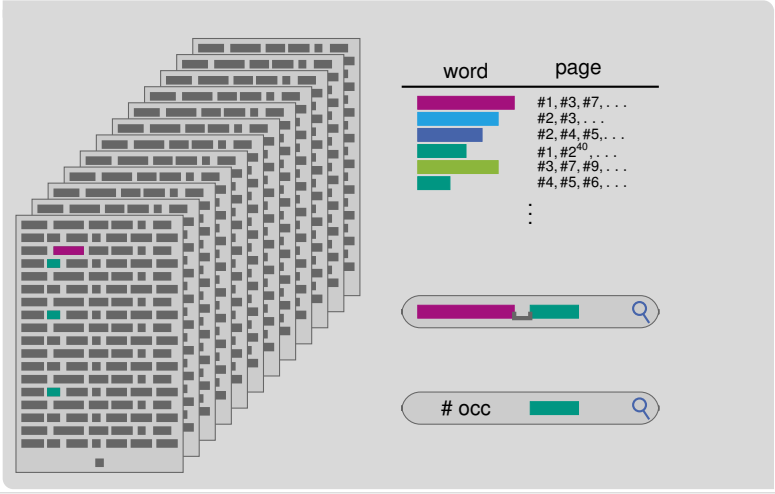
- collection of text
- scanning not feasible
- inverted index (word based)

Motivation for Text Indices



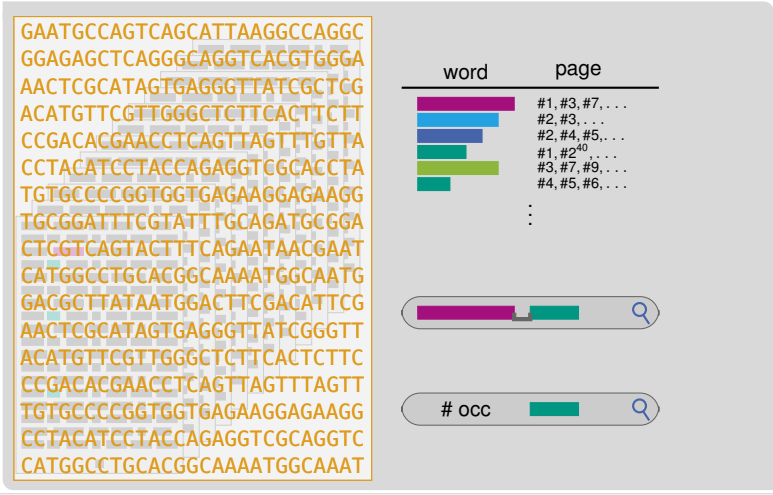
- collection of text
- scanning not feasible
- inverted index (word based)
- phrase search

Motivation for Text Indices



- collection of text
- scanning not feasible
- inverted index (word based)
- phrase search
- counting queries

Motivation for Text Indices

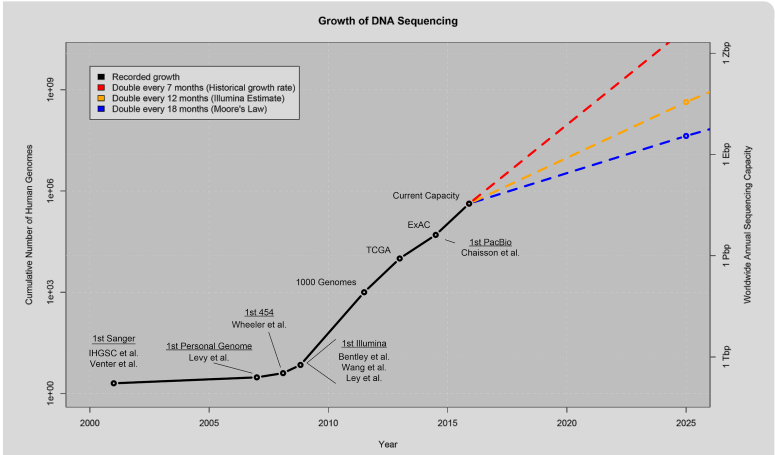


- collection of text
- scanning not feasible
- inverted index (word based)
- phrase search
- counting queries
- what if there are no “words”

Why Texts?

Text is Everywhere

- Text-based Information
  - Wikipedia $\mathbb{W}$
  - dblp
  - books
  - news articles
  - code
- Very Important in Bioinformatics
  - DNA
  - proteins



[Ste+15]

Bibliography

[Nav16] Gonzalo Navarro. *Compact Data Structures - A Practical Approach*. Cambridge University Press, 2016. ISBN: 978-1-10-715238-0.

[Ohl13] Enno Ohlebusch. *Bioinformatics Algorithms: Sequence Analysis, Genome Rearrangements, and Phylogenetic Reconstruction*. Oldenbusch Verlag, 2013. ISBN: 978-3000413162.

[Ste+15] Zachary D Stephens., Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, and Gene E. Robinson. “Big Data: Astronomical or Genomical?” In: *PLOS Biology* 13.7 (July 2015), pages 1–11. DOI: 10.1371/journal.pbio.1002195.