

Diego Arroyuelo  
Barbara Poblete (Eds.)

LNCS 13617

# String Processing and Information Retrieval

29th International Symposium, SPIRE 2022  
Concepción, Chile, November 8–10, 2022  
Proceedings



Springer

## Founding Editors

Gerhard Goos

*Karlsruhe Institute of Technology, Karlsruhe, Germany*

Juris Hartmanis

*Cornell University, Ithaca, NY, USA*

## Editorial Board Members

Elisa Bertino

*Purdue University, West Lafayette, IN, USA*

Wen Gao

*Peking University, Beijing, China*

Bernhard Steffen 

*TU Dortmund University, Dortmund, Germany*

Moti Yung 

*Columbia University, New York, NY, USA*

More information about this series at <https://link.springer.com/bookseries/558>

Diego Arroyuelo · Barbara Poblete (Eds.)

# String Processing and Information Retrieval

29th International Symposium, SPIRE 2022  
Concepción, Chile, November 8–10, 2022  
Proceedings

*Editors*

Diego Arroyuelo   
Universidad Técnica Federico Santa María  
Valparaíso, Chile

Barbara Poblete  
Universidad de Chile  
Santiago, Chile

Millennium Institute for Foundational  
Research on Data  
Santiago, Chile

ISSN 0302-9743                      ISSN 1611-3349 (electronic)  
Lecture Notes in Computer Science  
ISBN 978-3-031-20642-9              ISBN 978-3-031-20643-6 (eBook)  
<https://doi.org/10.1007/978-3-031-20643-6>

© The Editor(s) (if applicable) and The Author(s), under exclusive license  
to Springer Nature Switzerland AG 2022

Chapter “Engineering Compact Data Structures for Rank and Select Queries on Bit Vectors” is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>). For further details see license information in the chapter.

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

The 29th International Symposium on String Processing and Information Retrieval, SPIRE 2022, was held during November 8–10, 2022, in Concepción, Chile. SPIRE started in 1993 as the South American Workshop on String Processing, and therefore it was held in Latin America until 2000. Then, SPIRE moved to Europe, and from then on it has been held in Australia, Japan, the UK, Spain, Italy, Finland, Portugal, Israel, Brazil, Chile, Colombia, Mexico, Argentina, Bolivia, Peru, the USA, and France. In this edition, SPIRE was back in Chile, continuing the long and well-established tradition of encouraging high-quality research at the broad nexus of string processing, information retrieval, and computational biology. After two years running online (because of the COVID-19 pandemic), this year SPIRE returned to onsite mode (allowing also online attendants).

This volume contains the accepted papers presented in SPIRE 2022. There was a total of 43 submissions. We thank all authors who submitted their work for consideration to SPIRE 2022. Each submission received at least three single blind reviews and, after intensive discussion, the Program Committee decided to accept 23 papers. These were classified into seven tracks: string algorithms, string data structures, string compression, information retrieval, computational biology, space-efficient data structures, and pattern matching. Authors of accepted papers come from 14 countries across four continents (Asia, Europe, North America, and South America). We thank the authors for their valuable contributions and presentations at the conference. We also want to especially thank the Program Committee members and the external reviewers for their valuable work during the review and discussion phases. The SPIRE 2022 program also included two invited talks:

- “De Bruijn Graphs: Solving Biological Problems in Small Space”, by Leena Salmela, and
- “LZ-End Parsing: Upper Bounds”, by Dominik Kempa,

and the tutorial “Graph Databases” by Aidan Hogan and Domagoj Vrgoč. We thank them for accepting our invitation and for their enlightening presentations.

We are also grateful to the organizing committee, chaired by José Fuentes and Cecilia Hernández (Universidad de Concepción), whose excellent work allowed SPIRE 2022 to become a reality. Also, we want to thank the financial support of the Institute for Foundational Research on Data (IMFD), the Centre for Biotechnology and Bioengineering (CeBiB), the Vicerrectoría and the Facultad de Ingeniería of Universidad de Concepción, and R9 Ingeniería, which was crucial to fund the invited speakers, tutorial, streaming service, free student registration (to encourage onsite student participation, after two years of online activities), and the auditorium for the conference.

To complete the event, SPIRE 2022 had a Best Paper Award sponsored by Springer, which was announced at the conference.

November 2022

Diego Arroyuelo  
Barbara Poblete

# Organization

## Program Committee Chairs

Diego Arroyuelo	Universidad Técnica Federico Santa María and Millennium Institute for Foundational Research on Data, Chile
Barbara Poblete	University of Chile, Chile, and Amazon, USA

## Program Committee

Amihood Amir	Bar-Ilan University, Israel
Ricardo Baeza-Yates	Northeastern University, USA, Pompeu Fabra University, Spain, and University of Chile, Chile
Hideo Bannai	Tokyo Medical and Dental University, Japan
Altigran da Silva	Universidade Federal do Amazonas, Brazil
Antonio Fariña	University of A Coruña, Spain
Gabriele Fici	University of Palermo, Italy
Travis Gagie	Dalhousie University, Canada
Pawel Gawrychowski	University of Wrocław, Poland
Marcos Goncalves	Federal University of Minas Gerais, Brazil
Inge Li Gørtz	Technical University of Denmark, Denmark
Meng He	Dalhousie University, Canada
Wing-Kai Hon	National Tsing Hua University, Taiwan
Shunsuke Inenaga	Kyushu University, Japan
Dominik Köppl	Tokyo Medical and Dental University, Japan
Thierry Lecroq	University of Rouen Normandy, France
Zsuzsanna Lipták	University of Verona, Italy
Felipe A. Louza	Universidade Federal de Uberlândia, Brazil
Giovanni Manzini	University of Pisa, Italy
Joao Meidanis	University of Campinas and Scylla Bioinformatics, Brazil
Alistair Moffat	University of Melbourne, Australia
Viviane P. Moreira	Universidade Federal do Rio Grande do Sul, Brazil
Gonzalo Navarro	University of Chile, Chile
Nadia Pisanti	University of Pisa, Italy
Solon Pissis	Centrum Wiskunde & Informatica, The Netherlands
Nicola Prezza	Ca' Foscari University of Venice, Italy
Simon Puglisi	University of Helsinki, Finland
Rajeev Raman	University of Leicester, UK
Kunihiko Sadakane	The University of Tokyo, Japan



Srinivasa Rao Satti	Norwegian University of Science and Technology, Norway
Marinella Sciortino	University of Palermo, Italy
Diego Seco	University of A Coruña, Spain
Sharma V. Thankachan	University of Central Florida, USA
Rossano Venturini	University of Pisa, Italy
Nivio Ziviani	Federal University of Minas Gerais, Brazil

## Steering Committee

Ricardo Baeza-Yates	Northeastern University, USA, Pompeu Fabra University, Spain, and University of Chile, Chile
Christina Boucher	University of Florida, USA
Nieves R. Brisaboa	University of A Coruña, Spain
Thierry Lecroq	University of Rouen Normandy, France
Simon Puglisi	University of Helsinki, Finland
Berthier Ribeiro-Neto	Google Inc. and Federal University of Minas Gerais, Brazil
Sharma Thankachan	University of Central Florida, USA
Hélène Touzet	CNRS, France
Nivio Ziviani	Federal University of Minas Gerais, Brazil

## Organizing Committee

José Fuentes-Sepúlveda	Universidad de Concepción and Millennium Institute for Foundational Research on Data, Chile
Cecilia Hernández	Universidad de Concepción, Chile

## Additional Reviewers

Paniz Abedin	Daniel Gibney
Fabiano Belém	Sara Giuliani
Luciana Bencke	Adrián Gómez-Brandón
Giulia Bernardini	Keisuke Goto
Itai Boneh	Veronica Guerrini
Davide Cenzato	Tomohiro I
Dustin Cobas	Michael Itzhaki
Guillermo De Bernardo	Varunkumar Jayapaul
Daniel Xavier De Sousa	Seungbum Jo
Jonas Ellert	Serikzhan Kazi
Massimo Equi	Eitan Kondratovsky
Celso França	William Kuszmaul
José Fuentes-Sepúlveda	Francesco Masillo
Younan Gao	Takuya Mieno
Samah Ghazawi	Yuto Nakashima

Takaaki Nishimoto  
Francisco Olivares  
Lucas Oliveira  
Kunsoo Park  
Pierre Peterlongo  
Giulio Ermanno Pibiri  
Jakub Radoszewski  
Giuseppe Romana  
Yoshifumi Sakai

Ayumi Shinohara  
Tatiana Starikovskaya  
Guilherme Telles  
Cristian Urbina  
Adriano Veloso  
Felipe Viegas  
Kaiyu Wu  
Wiktor Zuba

## **Abstracts of Invited Talks**

# De Bruijn Graphs: Solving Biological Problems in Small Space

Leena Salmela 

Department of Computer Science, University of Helsinki, Helsinki, Finland  
leena.salmela@helsinki.fi

**Abstract.** De Bruijn graphs have become a standard data structure in analysing sequencing data due to its ability to represent the information in a sequencing read set in small space. They represent the sequencing reads by the  $k$ -mers, i.e., substrings of length  $k$  occurring in the reads. Classically, the edges of a de Bruijn graph are defined to be the  $k$ -mers and the nodes are the  $k - 1$ -length prefixes and suffixes of the  $k$ -mers. The construction of a de Bruijn graph starts by counting the  $k$ -mers occurring in the reads. Many good methods exist for extracting exact  $k$ -mers from read data and counting the number of their occurrences. However, sequencing read sets can contain a significant number of sequencing errors, which limits the usefulness of counting exact  $k$ -mers to short  $k$ -mers. Recently, we have developed methods for extracting longer  $k$ -mers from noisy data by using spaced seeds and strobemers.

De Bruijn graphs were originally introduced for solving the genome assembly problem, where the goal is to reconstruct the genome based on sequencing reads. In practice, genome assembly is solved with de Bruijn graphs by reporting unitigs, which are non-branching paths in the de Bruijn graphs. The choice of  $k$  is a crucial matter in de-Bruijn-graph-based genome assembly. A too small  $k$  will make the graph tangled, resulting in short unitigs, while a too large  $k$  will fragment the graph, again resulting in short unitigs. A variable-order de Bruijn graph, which represents de Bruijn graphs of all orders  $k$  in a single data structure, has been presented as a solution to the choice of  $k$ . However, it is not clear how the definition of unitigs can be extended to variable-order de Bruijn graphs.

In this talk, we present a robust definition of assembled sequences in variable-order de Bruijn graphs and an algorithm for enumerating them. Apart from genome assembly, de Bruijn graphs are used in many other problems such as sequencing error correction, reference free variant calling, indexing read sets, and so on. At the end of this talk, we will review some of these applications and their de-Bruijn-graph-based solutions.

**Keywords:** de Bruijn graph ·  $k$ -mer · Genome assembly

# LZ-End Parsing: Upper Bounds and Algorithmic Techniques

Dominik Kempa

Stony Brook University,  
Stony Brook, New York, USA  
kempa@cs.stonybrook.edu

**Abstract.** Lempel–Ziv (LZ77) compression is the most commonly used lossless compression algorithm. The basic idea is to greedily break the input string into blocks (called “phrases”), every time forming as a phrase the longest prefix of the unprocessed part that has an earlier occurrence. In 2010, Krefl and Navarro introduced a variant of LZ77 called LZ-End, that additionally requires the previous occurrence of each phrase to end at the boundary of an already existing phrase. Due to its excellent practical performance as a compression algorithm and a compressed index, they conjectured that it achieves a compression that can be provably upper-bounded in terms of the LZ77 size. Despite the recent progress in understanding such relation for other compression algorithms (e.g., the run-length encoded Burrows–Wheeler transform), no such result is known for LZ-End. In this talk, we give an overview of the recent progress on the above problem. More precisely, we prove that for any string of length  $n$ , the number  $z_e$  of phrases in the LZ-End parsing satisfies  $z_e = \mathcal{O}(z \log^2 n)$ , where  $z$  is the number of phrases in the LZ77 parsing. This is the first non-trivial upper bound on the size of LZ-End parsing in terms of LZ77, and it puts LZ-End among the strongest dictionary compressors. Using our techniques, we also derive bounds for other variants of LZ-End and with respect to other compression measures. Our second contribution is a data structure that implements random access queries to the text in  $\mathcal{O}(z_e)$  space and  $\mathcal{O}(\text{poly } \log n)$  time. This is the first linear-size structure on LZ-End that efficiently implements such queries. All previous data structures either incur a logarithmic penalty in the space or have slow queries. We also show how to extend these techniques to support longest-common-extension (LCE) queries. This work was carried out in collaboration with Barna Saha and was presented at the 2022 ACM-SIAM Symposium on Discrete Algorithms (SODA 2022).

**Keywords:** LZ-End · LZ77 · Dictionary compression

# Contents

## String Algorithms

Subsequence Covers of Words . . . . .	3
<i>Panagiotis Charalampopoulos, Solon P. Pissis, Jakub Radoszewski, Wojciech Rytter, Tomasz Waleń, and Wiktor Zuba</i>	
Maximal Closed Substrings . . . . .	16
<i>Golnaz Badkobeh, Alessandro De Luca, Gabriele Fici, and Simon J. Puglisi</i>	
Online Algorithms for Finding Distinct Substrings with Length and Multiple Prefix and Suffix Conditions . . . . .	24
<i>Laurentius Leonard, Shunsuke Inenaga, Hideo Bannai, and Takuya Mieno</i>	
The Complexity of the Co-occurrence Problem . . . . .	38
<i>Philip Bille, Inge Li Gørtz, and Tord Stordalen</i>	

## String Data Structures

Reconstructing Parameterized Strings from Parameterized Suffix and LCP Arrays . . . . .	55
<i>Amihood Amir, Concettina Guerra, Eitan Konradovsky, Gad M. Landau, Shoshana Marcus, and Dina Sokol</i>	
Computing the Parameterized Burrows–Wheeler Transform Online . . . . .	70
<i>Daiki Hashimoto, Diptarama Hendrian, Dominik Köppl, Ryo Yoshinaka, and Ayumi Shinohara</i>	
Accessing the Suffix Array via $\phi^{-1}$ -Forest . . . . .	86
<i>Christina Boucher, Dominik Köppl, Herman Perera, and Massimiliano Rossi</i>	
On the Optimisation of the GSACA Suffix Array Construction Algorithm . . .	99
<i>Jannik Olbrich, Enno Ohlebusch, and Thomas Büchler</i>	

## String Compression

Balancing Run-Length Straight-Line Programs . . . . .	117
<i>Gonzalo Navarro, Francisco Olivares, and Cristian Urbina</i>	

Substring Complexities on Run-Length Compressed Strings . . . . . 132  
*Akiyoshi Kawamoto and Tomohiro I*

**Information Retrieval**

How Train–Test Leakage Affects Zero-Shot Retrieval . . . . . 147  
*Maik Fröbe, Christopher Akiki, Martin Potthast, and Matthias Hagen*

**Computational Biology**

Genome Comparison on Succinct Colored de Bruijn Graphs . . . . . 165  
*Lucas P. Ramos, Felipe A. Louza, and Guilherme P. Telles*

Sorting Genomes by Prefix Double-Cut-and-Joins . . . . . 178  
*Guillaume Fertin, Géraldine Jean, and Anthony Labarre*

KATKA: A KRAKEN-Like Tool with *k* Given at Query Time . . . . . 191  
*Travis Gagie, Sana Kashgouli, and Ben Langmead*

Computing All-vs-All MEMs in Run-Length-Encoded Collections of HiFi Reads . . . . . 198  
*Diego Díaz-Domínguez, Simon J. Puglisi, and Leena Salmela*

**Space-Efficient Data Structures**

Internal Masked Prefix Sums and Its Connection to Fully Internal Measurement Queries . . . . . 217  
*Rathish Das, Meng He, Eitan Konratovsky, J. Ian Munro, and Kaiyu Wu*

Compressed String Dictionaries via Data-Aware Subtrie Compaction . . . . . 233  
*Antonio Boffa, Paolo Ferragina, Francesco Tosoni, and Giorgio Vinciguerra*

On Representing the Degree Sequences of Sublogarithmic-Degree Wheeler Graphs . . . . . 250  
*Travis Gagie*

Engineering Compact Data Structures for Rank and Select Queries on Bit Vectors . . . . . 257  
*Florian Kurpicz*

**Pattern Matching on Strings, Graphs, and Trees**

Matching Patterns with Variables Under Edit Distance . . . . . 275  
*Paweł Gawrychowski, Florin Manea, and Stefan Siemer*

**On the Hardness of Computing the Edit Distance of Shallow Trees. . . . .** 290  
*Panagiotis Charalampopoulos, Paweł Gawrychowski, Shay Mozes,  
and Oren Weimann*

**Quantum Time Complexity and Algorithms for Pattern Matching  
on Labeled Graphs . . . . .** 303  
*Parisa Darbari, Daniel Gibney, and Sharma V. Thankachan*

**Pattern Matching Under DTW Distance . . . . .** 315  
*Garance Gourdel, Anne Driemel, Pierre Peterlongo,  
and Tatiana Starikovskaya*

**Author Index . . . . .** 331