Alexander S. Kulikov
Sergei O. Kuznetsov
Pavel Pevzner (Eds.)

LNCS 8486

Combinatorial Pattern Matching

25th Annual Symposium, CPM 2014
Moscow, Russia, June 16–18, 2014
Proceedings

Springer

Lecture Notes in Computer Science 8486

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Alexander S. Kulikov Sergei O. Kuznetsov
Pavel Pevzner (Eds.)

Combinatorial Pattern Matching

25th Annual Symposium, CPM 2014
Moscow, Russia, June 16-18, 2014
Proceedings

Springer

Volume Editors

Alexander S. Kulikov
St. Petersburg Department of Steklov Institute of Mathematics
27 Fontanka
St. Petersburg 191023, Russia
E-mail: kulikov@logic.pdmi.ras.ru

Sergei O. Kuznetsov
National Research University Higher School of Economics
3 Bolshoy Trekhsvyatitelskiy pereulok
Moscow 109028, Russia
E-mail: skuznetsov@hse.ru

Pavel Pevzner
University of California at San Diego
9500 Gilman Drive, EBU3b 4236
La Jolla, CA 92093-0404, USA
E-mail: ppevzner@cs.ucsd.edu

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-319-07565-5 e-ISBN 978-3-319-07566-2
DOI 10.1007/978-3-319-07566-2
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014939431

LNCS Sublibrary: SL 1 – Theoretical Computer Science and General Issues

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This volume contains the papers presented at the 25th Annual Symposium on Combinatorial Pattern Matching held during June 15–17, 2014 in Moscow, Russia. The hosting university was National Research University Higher School of Economics.

There were 54 submissions from 25 countries. Each submission was reviewed by at least three Program Committee members. The committee decided to accept 28 papers, corresponding to an acceptance rate of 54%. We thank the members of the Program Committee and all additional external reviewers for their hard work that resulted in this excellent program. Their names are listed on the following pages. The whole submission and review process was carried out with the invaluable help of the EasyChair conference system.

The year 2014 marks the quarter-of-a-century milestone for the CPM symposium series. CPM 2014 celebrated this event with invited talks by the co-founders of the conference series: Alberto Apostolico (Georgia Institute of Technology, USA, and IASI-CNR, Italy), Maxime Crochemore (King's College London, UK, and Université Paris-Est, France), Zvi Galil (Georgia Institute of Technology, USA), and Udi Manber (Google, USA). In addition, CPM 2014 featured a special keynote lecture by Gene Myers (Max Planck Institute, Germany) that highlighted the contribution of the CPM community to bioinformatics.

The objective of the annual CPM meetings is to provide an international forum for research in combinatorial pattern matching and related applications. It addresses issues of searching and matching strings and more complicated patterns such as trees, regular expressions, graphs, point sets, and arrays. The goal is to derive combinatorial properties of such structures and to exploit these properties in order to achieve superior performance for the corresponding computational problems. The meeting also deals with problems in computational biology, data compression and data mining, coding, information retrieval, natural language processing, and pattern recognition.

The Annual Symposium on Combinatorial Pattern Matching started in 1990, and has since taken place every year. Previous CPM meetings were held in Paris, London (UK), Tucson, Padova, Asilomar, Helsinki, Laguna Beach, Aarhus, Piscataway, Warwick, Montreal, Jerusalem, Fukuoka, Morelia, Istanbul, Jeju Island, Barcelona, London (Ontario, Canada), Pisa, Lille, New York, Palermo, Helsinki, Bad Herrenalb. This years' meeting was the first in Russia. Starting from the third meeting, proceedings of all meetings have been published in the LNCS series, as volumes 644, 684, 807, 937, 1075, 1264, 1448, 1645, 1848, 2089, 2373, 2676, 3109, 3537, 4009, 4580, 5029, 5577, 6129, 6661, 7354, and 7922.

We thank Russian Foundation for Basic Research, National Research University Higher School of Economics, and Yandex for their financial support.

March 2014

Alexander S. Kulikov
Sergei Kuznetsov
Pavel Pevzner

Organization

Program Committee

Max Alekseyev	George Washington University, USA
Amihood Amir	Bar-Ilan University, Israel
Maxim Babenko	Moscow State University, Russia
Martin Farach-Colton	Rutgers University, USA
Paolo Ferragina	University of Pisa, Italy
Johannes Fischer	Technical University of Dortmund, Germany
Dan Gusfield	University of California at Davis, USA
Roman Kolpakov	Moscow State University, Russia
Gregory Kucherov	Université Paris-Est Marne-la-Vallée, France
Alexander S. Kulikov	St. Petersburg Department of Steklov Institute of Mathematics, Russia (Co-chair)
Juha Kärkkäinen	University of Helsinki, Finland
Gad M. Landau	University of Haifa, Israel
Stefano Lonardi	University of California at Riverside, USA
Ian Munro	University of Waterloo, Canada
S. Muthukrishnan	Rutgers University, USA
Gonzalo Navarro	University of Chile, Chile
Kunsoo Park	Seoul National University, South Korea
Pavel Pevzner	University of California San Diego, USA (Co-chair)
Nadia Pisanti	University of Pisa, Italy
Mikhail Roytberg	Higher School of Economics, Russia
Tatiana Starikovskaya	Higher School of Economics, Russia
Jim Storer	Brandeis University, USA
Jens Stoye	University of Bielefeld, Germany
Esko Ukkonen	University of Helsinki, Finland

Steering Committee

Alberto Apostolico	Georgia Institute of Technology, USA, and IASI-CNR, Italy
Maxime Crochemore	Université Paris-Est, France, and Kings College London, UK
Zvi Galil	Georgia Institute of Technology, USA

Organizing Committee

Stepan Artamonov	Moscow State University, Russia
Maxim Babenko	Moscow State University, Russia
Dmitry Ignatov	Higher School of Economics, Russia
Dmitry Ilvovsky	Higher School of Economics, Russia
Alexander S. Kulikov	St. Petersburg Department of Steklov Institute of Mathematics, Russia
Sergei Kuznetsov	Higher School of Economics, Russia, (Chair)
Dmitry Morozov	Higher School of Economics, Russia
Kamil Salihov	Moscow State University, Russia
Ruslan Savchenko	Moscow State University, Russia
Tatiana Starikovskaya	Higher School of Economics, Russia

Additional Reviewers

Aganezov, Sergey
Amit, Mika
Antipov, Dmitry
Artamonov, Stepan
Bankevich, Anton
Belazzougui, Djamal
Bowe, Alexander
Braga, Marilia
Butman, Ayelet
Chikhi, Rayan
Cicalese, Ferdinando
Cording, Patrick Hagge
Cunial, Fabio
Fedorov, Sergey
Feijao, Pedro
Fertin, Guillaume
Frid, Yelena
Gagie, Travis
Gawrychowski, Pawel
Giaquinta, Emanuele
Gysel, Rob
He, Meng
Hu, Fei
I, Tomohiro
Inenaga, Shunsuke
Jahn, Katharina
Jiang, Shuai
Kempa, Dominik
Kolesnichenko, Ignat
Kopczynski, Dominik
Labarre, Anthony
Levy, Avivit
Matsieva, Julia
Melsted, Pall
Mirebrahim, Seyed Hamid
Moret, Bernard
Nekrich, Yakov
Nielsen, Jesper Sindahl
Ottaviano, Giuseppe
Paparo, Omer
Pizzi, Cinzia
Polishko, Anton
Prencipe, Giuseppe
Puglisi, Simon
Pérez-Lantero, Pablo
Radoszewski, Jakub
Rogulenko, Sergey
Russo, Luis M. S.
Salikhov, Kamil
Savchenko, Ruslan
Sirotkin, Alexander
Sirén, Jouni
St. John, Katherine
Stevens, Kristian
Tarasov, Pavel
Thankachan, Sharma

Tomescu, Alexandru I.
Vasilevskaya, Maria
Venturini, Rossano
Vildhøj, Hjalte Wedel
Vind, Søren
Välimäki, Niko
Willing, Eyla
Wittler, Roland

Abstracts of Invited Talks

Sequence Comparison in the Time of the Deluge

Alberto Apostolico

Georgia Institute of Technology, USA and IASI-CNR, Italy

It is almost fifty years since the appearance of the famous paper entitled "Binary codes capable of correcting deletions, insertions, and reversals" (1966 English translation, Soviet Physics Doklady) by which Vladimir Levenshtein introduced his eponymous measure of string distance. The notion was since to be re-discovered, variously dithered and more or less efficiently computed in distant domains of application. In particular, it provided the platform for half a century of analysis, alignment, search, taxonomy and phylogeny of molecular sequences. As the size and multiplicity of the sequences produced expand to an unprecedented scale, many elegant techniques of the past no longer work. In molecular taxonomy and phylogeny, for instance, the alignment of whole genomes proves both computationally unbearable and hardly significant. In metagenomics, the elucidation of microbiome compositions under conditions of noisy and largely unidentified reference sequences faces steep barriers during assembly and assignment. In recent studies, classical notions are increasingly being complemented or even supplanted by global similarity measures that refer, implicitly or explicitly, to the composition of sequences in terms of their constituent patterns. Such measures hinge more or less uniformly on an underlying notion of relative compressibility, whereby two sequences are similar if either one can be described using mostly pieces from the other. They can free from chores of alignment and assembly. Their computation poses interesting and variously affordable algorithmic problems. This talk will review some such measures, their computation, and their applications.

Repeats in Strings

Maxime Crochemore

King's College London, UK, and Université Paris-Est, France

Large amounts of text are generated every day in the cyberspace via Web sites, emails, social networks, and other communication networks. These text streams need to be analysed to detect critical events or the monitor business for example. An important characteristics to take into account in this setting is the existence of repetitions in texts. Their study constitutes a fundamental area of combinatorics on words due to major applications to string algorithms, data compression, music analysis, and biological sequences analysis, etc. The talk surveys algorithmic methods used to locate repetitive segments in strings. It discusses the notion of runs that encompasses various types of periodicities considered by different authors, as well as the notion of maximal-exponent factors that captures the most significant repeats occurring in a string. The design and analysis of repeat finders rely on combinatorial properties of words and raise a series of open problems in combinatorics on words.

"Stringology" is 30 Years Old

Zvi Galil

Georgia Institute of Technology, USA

This year we have two anniversaries: This is the 25th CPM. CPM1, the first one, was held in Paris in 1990. But in June 1984 there was a sort of a predecessor of CPM1 which can be considered as CPM0. It was a Nato Workshop on combinatorial algorithms on words in Maratea, Italy. The first paper was my paper: "Open Problems in Stringology". Thus, the term Stringology is exactly 30 years old. The field of stringology is much older. Even though the word stringology cannot be found in any respectable dictionary, Google search yields 23,800 hits; in addition 1640 articles use it, 123 in their title.

In celebration of this 30th anniversary, we will review the status of the 13 open problems of the paper and will introduce some new ones.

How to Think Big

Udi Manber

Google Inc.

"If you have to ask, you'll never know"
Louis Armstrong (answering a question
about the meaning of "swing")

In business, making more impact used to mean making more money. But that changed. In 2010 Twitter's revenues were 40x less than cat litter, but its impact was enormous. Same with Amazon, Google, and Facebook at their beginnings. The key to all of them was that they were doing completely new things. No one thought there was a need to share 140 characters before Twitter, but it turned out to be extremely useful. Very few people thought at the time that shopping on the web, highly relevant search, or communication between friends were important business needs.

When the telephone was introduced to telegraph companies they discounted it. They had a good reason. More than a century later AT&T discounted voice over IP for similar reasons. Yet the 100+ years old AT&T was sold for less than the 5 years old WhatsApp. Things change rather quickly nowadays.

There are some common insights from these stories that can be applied to academic research and I will try to highlight them in this talk.

What's Behind Blast

Gene Myers

Max Planck Institute for Molecular Cell Biology and Genetics

While Blast is one of the most widely used search engines for molecular biology, and while there is a general understanding of how it works, few know the story of how it came about and the theoretical algorithmic result from which it was derived. I will tell the story and explain the theoretical algorithm — an $O(DN^{\mathrm{pow}(D/P)} \log N)$ expected-time algorithm for finding all matches to a query of length P with not more than D differences in a database of length $N \gg P$. Surprisingly, this result, published in early 1994, has to my knowledge not been improved upon to this day.

Table of Contents