Lecture Notes in Computer Science 4580

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

Editorial Board

David Hutchison
Lancaster University, UK

Takeo Kanade
Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler
University of Surrey, Guildford, UK

Jon M. Kleinberg
Cornell University, Ithaca, NY, USA

Friedemann Mattern
ETH Zurich, Switzerland

John C. Mitchell
Stanford University, CA, USA

Moni Naor
Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz
University of Bern, Switzerland

C. Pandu Rangan
Indian Institute of Technology, Madras, India

Bernhard Steffen
University of Dortmund, Germany

Madhu Sudan
Massachusetts Institute of Technology, MA, USA

Demetri Terzopoulos
University of California, Los Angeles, CA, USA

Doug Tygar
University of California, Berkeley, CA, USA

Moshe Y. Vardi
Rice University, Houston, TX, USA

Gerhard Weikum
Max-Planck Institute of Computer Science, Saarbruecken, Germany

Bin Ma Kaizhong Zhang (Eds.)

Combinatorial Pattern Matching

18th Annual Symposium, CPM 2007
London, Canada, July 9-11, 2007
Proceedings

Springer

Volume Editors

Bin Ma
Kaizhong Zhang
University of Western Ontario
Department of Computer Science
London, Ontario, N6A 5B7, Canada
E-mail: {bma; kzhang}@csd.uwo.ca

Library of Congress Control Number: 2007929746

CR Subject Classification (1998):
F.2.2, I.5.4, I.5.0, H.3.3, J.3, E.4, G.2.1, E.1

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

ISSN 0302-9743
ISBN-10 3-540-73436-8 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-73436-9 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springer.com

© Springer-Verlag Berlin Heidelberg 2007
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12086467 06/3180 5 4 3 2 1 0

Preface

The papers contained in this volume were presented at the 18th Annual Symposium on Combinatorial Pattern Matching (CPM 2007) held at the University of Western Ontario, in London, Ontario, Canada from July 9 to 11, 2007.

All the papers presented at the conference are original research contributions on computational pattern matching and analysis, data compression and compressed text processing, suffix arrays and trees, and computational biology. They were selected from 64 submissions. Each submission was reviewed by at least three reviewers. The committee decided to accept 32 papers. The programme also included three invited talks by Tao Jiang from the University of California, Riverside, USA, S. Muthukrishnan from Rutgers University, USA, and Frances Yao from City University of Hong Kong, Hong Kong.

Combinatorial Pattern Matching 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 non-trivial combinatorial properties of such structures and to exploit these properties in order to either achieve superior performance for the corresponding computational problems or pinpoint conditions under which searches cannot be performed efficiently.

The Annual Symposium on Combinatorial Pattern Matching started in 1990, and has since taken place every year. The objective of the annual CPM meetings is to provide an international forum for research in combinatorial pattern matching and related applications. Previous CPM meetings were held in Paris, London, Tucson, Padova, Asilomar, Helsinki, Laguna Beach, Aarhus, Piscataway, Warwick, Montreal, Jerusalem, Fukuoka, Morelia, Istanbul, Jeju Island, and Barcelona. Selected papers from the first meeting appeared in volume 92 of *Theoretical Computer Science*, from the 11th meeting in volume 2 of the *Journal of Discrete Algorithms*, from the 12th meeting in volume 146 of *Discrete Applied Mathematics*, and from the 14th meeting in volume 3 of the *Journal of Discrete Algorithms*. Starting from the 3rd meeting, the proceedings of all meetings have been published in the LNCS series, volumes 644, 684, 807, 937, 1075, 1264, 1448, 1645, 1848, 2089, 2373, 2676, 3109, 3537, 4009, and 4580.

The whole submission and review process, as well as the production of this volume, was carried out with the help of the EasyChair system. The conference was sponsored by the University of Western Ontario and by the Fields Institute.

April 2007

Bin Ma
Kaizhong Zhang

Conference Organization

Programme Chairs

Bin Ma
Kaizhong Zhang

Programme Committee

Tatsuya Akutsu
Amihood Amir
Raffaele Giancarlo
Gregory Kucherov
Ming Li
Guohui Lin
Heikki Mannila
Gonzalo Navarro
Ron Pinter
Mathieu Raffinot
Cenk Sahinalp
David Sankoff
Steve Skiena
James Storer
Masayuki Takeda
Gabriel Valiente
Martin Vingron
Lusheng Wang

Local Organization

Meg Borthwick
Jingping Liu
Cheryl McGrath
Roberto Solis-Oba (Co-chair)
Kaizhong Zhang (Co-chair)

External Reviewers

Cagri Aksay
Hideo Bannai
Petra Berenbrink
Guillaume Blin
Marie-Pierre Béal
Cedric Chauve

Shihyen Chen
Phuong Dao
Guillaume Fertin
Tom Friedetzky
Liliana Félix
Aris Gionis
Stefan Haas
Niina Haiminen
Iman Hajiresouliha
Tzvika Hartman
Rodrigo Hausen
Hannes Heikinheimo
Yasuto Higa
Fereydoun Hormozdiari
Lucian Ilie
Shunsuke Inenaga
Hossein Jowhari
Oren Kapah
Emre Karakoc
Orgad Keller
Takuya Kida
Roman Kolpakov
Tsvi Kopelowitz
Dennis Kostka
Juha Kärkkäinen
Gad Landau
Michael Lappe
Thierry Lecroq
Avivit Levy
Weiming Li
Yury Lifshits
Jingping Liu
Mercè Llabrés
Antoni Lozano
Giovanni Manzini
Conrado Martínez
Igor Nor
Pasi Rastas
Tobias Rausch
David Reese
Hugues Richard
Jairo Rocha
Oleg Rokhlenko
Francesc Rosselló
Dominique Rossin
Wojciech Rytter
Kunihiko Sadakane
Hiroshi Sakamoto
Rahaleh Salari
Marcel Schulz
Jouni Seppanen
Dina Sokol
Jens Stoye
Takeyuki Tamura
Eric Tannier
Helene Touzet
Antti Ukkonen
Tomas Vinar
Robert Warren
Lei Xin

Table of Contents

Invited Talks (Abstracts)

A Combinatorial Approach to Genome-Wide Ortholog Assignment: Beyond Sequence Similarity Search 1
Tao Jiang

Stringology: Some Classic and Some Modern Problems 2
S. Muthukrishnan

Algorithmic Problems in Scheduling Jobs on Variable-Speed Processors 3
Frances F. Yao

Session 1: Alogirthmic Techniques I

Speeding Up HMM Decoding and Training by Exploiting Sequence Repetitions 4
Shay Mozes, Oren Weimann, and Michal Ziv-Ukelson

On Demand String Sorting over Unbounded Alphabets 16
Carmel Kent, Moshe Lewenstein, and Dafna Sheinwald

Session 2: Approximate Pattern Matching

Finding Witnesses by Peeling 28
Yonatan Aumann, Moshe Lewenstein, Noa Lewenstein, and Dekel Tsur

Cache-Oblivious Index for Approximate String Matching 40
Wing-Kai Hon, Tak-Wah Lam, Rahul Shah, Siu-Lung Tam, and Jeffrey Scott Vitter

Improved Approximate String Matching and Regular Expression Matching on Ziv-Lempel Compressed Texts 52
Philip Bille, Rolf Fagerberg, and Inge Li Gørtz

Self-normalised Distance with Don't Cares 63
Peter Clifford and Raphaël Clifford

Session 3: Data Compression I

Move-to-Front, Distance Coding, and Inversion Frequencies Revisited . . . 71
Travis Gagie and Giovanni Manzini

A Lempel-Ziv Text Index on Secondary Storage 83
Diego Arroyuelo and Gonzalo Navarro

Dynamic Rank-Select Structures with Applications to Run-Length Encoded Texts 95
Sunho Lee and Kunsoo Park

Most Burrows-Wheeler Based Compressors Are Not Optimal 107
Haim Kaplan and Elad Verbin

Session 4: Computational Biology I

Non-breaking Similarity of Genomes with Gene Repetitions 119
Zhixiang Chen, Bin Fu, Jinhui Xu, Boting Yang, Zhiyu Zhao, and Binhai Zhu

A New and Faster Method of Sorting by Transpositions 131
Maxime Benoît-Gagné and Sylvie Hamel

Finding Compact Structural Motifs 142
Jianbo Qian, Shuai Cheng Li, Dongbo Bu, Ming Li, and Jinbo Xu

Session 5: Computational Biology II

Improved Algorithms for Inferring the Minimum Mosaic of a Set of Recombinants 150
Yufeng Wu and Dan Gusfield

Computing Exact p-Value for Structured Motif 162
Jing Zhang, Xi Chen, and Ming Li

Session 6: Algorithmic Techniques II

Improved Sketching of Hamming Distance with Error Correcting 173
Ely Porat and Ohad Lipsky

Deterministic Length Reduction: Fast Convolution in Sparse Data and Applications 183
Amihood Amir, Oren Kapah, and Ely Porat

Guided Forest Edit Distance: Better Structure Comparisons by Using Domain-knowledge 195
Zeshan Peng and Hing-fung Ting

Space-Efficient Algorithms for Document Retrieval 205
Niko Välimäki and Veli Mäkinen

Session 7: Data Compression II

Compressed Text Indexes with Fast Locate 216
Rodrigo González and Gonzalo Navarro

Processing Compressed Texts: A Tractability Border 228
Yury Lifshits

Session 8: Computational Biology III

Common Structured Patterns in Linear Graphs: Approximation and Combinatorics 241
Guillaume Fertin, Danny Hermelin, Romeo Rizzi, and Stéphane Vialette

Identification of Distinguishing Motifs 253
WangSen Feng, Zhanyong Wang, and Lusheng Wang

Algorithms for Computing the Longest Parameterized Common Subsequence 265
Costas S. Iliopoulos, Marcin Kubica, M. Sohel Rahman, and Tomasz Waleń

Fixed-Parameter Tractability of the Maximum Agreement Supertree Problem 274
Sylvain Guillemot and Vincent Berry

Session 9: Pattern Analysis

Two-Dimensional Range Minimum Queries 286
Amihood Amir, Johannes Fischer, and Moshe Lewenstein

Tiling Periodicity 295
Juhani Karhumäki, Yury Lifshits, and Wojciech Rytter

Fast and Practical Algorithms for Computing All the Runs in a String . . 307
Gang Chen, Simon J. Puglisi, and W.F. Smyth

Longest Common Separable Pattern Among Permutations 316
Mathilde Bouvel, Dominique Rossin, and Stéphane Vialette

Session 10: Suffix Arrays and Trees

Suffix Arrays on Words 328
Paolo Ferragina and Johannes Fischer

Efficient Computation of Substring Equivalence Classes with Suffix Arrays 340
Kazuyuki Narisawa, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda

A Simple Construction of Two-Dimensional Suffix Trees in Linear Time 352
Dong Kyue Kim, Joong Chae Na, Jeong Seop Sim, and Kunsoo Park

Author Index 365