

*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*

Alfred Kobsa

*University of California, Irvine, CA, 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*

Gerhard Weikum

*Max-Planck Institute of Computer Science, Saarbruecken, Germany*

Gregory Kucherov Esko Ukkonen (Eds.)

# Combinatorial Pattern Matching

20th Annual Symposium, CPM 2009  
Lille, France, June 22-24, 2009  
Proceedings



Springer

Volume Editors

Gregory Kucherov

Laboratoire d'Informatique Fondamentale de Lille (LIFL)

Bâtiment M3, 59655 Villeneuve d'Ascq CEDEX, France

E-mail: gregory.kucherov@lifl.fr

Esko Ukkonen

University of Helsinki, Department of Computer Science

P.O. Box 68 (Gustaf Hällströmin katu 2b)

00014 University of Helsinki, Finland

E-mail: esko.ukkonen@cs.helsinki.fi

Library of Congress Control Number: 2009930681

CR Subject Classification (1998): F.2, I.5, H.3.3, J.3, I.4.2, E.4, G.2.1, E.1

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

ISSN 0302-9743

ISBN-10 3-642-02440-8 Springer Berlin Heidelberg New York

ISBN-13 978-3-642-02440-5 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.com

© Springer-Verlag Berlin Heidelberg 2009

Printed in Germany

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

Printed on acid-free paper SPIN: 12696680 06/3180 5 4 3 2 1 0

# Preface

It is our great pleasure to introduce the proceedings of the 20th anniversary edition of the Annual Symposium on Combinatorial Pattern Matching (CPM). The meeting was held in Lille, France, hosted by the Laboratoire d'Informatique Fondamentale de Lille (LIFL) affiliated with the Université de Lille 1 and the French Centre National de Recherche Scientifique (CNRS), as well as by INRIA Lille - Nord Europe.

Started in 1990 as a summer school with about 30 invited participants, CPM quickly evolved into a representative annual international conference. Principally motivated by combinatorial algorithms for search problems in strings (texts, sequences), the scope of CPM extended to more complex data structures such as trees, graphs, two-dimensional arrays, or sets of points. Those studies resulted in a rich collection of algorithmic techniques and data structures, making bridges to other parts of the theory of discrete algorithms and algorithm engineering. Today, the area of combinatorial pattern matching is a well-identified active subfield of algorithmic research.

Importantly, this development has been fertilized by a number of major application areas providing direct motivations and fruitful feedback to the CPM problematics. Those applications include data compression, computational biology, Internet search, data mining, information retrieval, coding, natural language processing, pattern recognition, music analysis, and others. On the one hand, all these areas make use of combinatorial pattern matching techniques and, on the other hand, raise new pattern matching problems. For example, the fast progress in computational molecular biology, triggered in the 1990s by the availability of mass genomic data, considerably influenced the combinatorial pattern matching field: as an illustration, about one-third of the papers presented in this volume deal with problems related to bioinformatics applications.

In 2009, the Combinatorial Pattern Matching symposium celebrated its 20th anniversary. 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 (Canada), and Pisa. Starting from the third meeting, proceedings were published in the LNCS series, volumes 644, 684, 807, 937, 1075, 1264, 1448, 1645, 1848, 2089, 2373, 2676, 3109, 3537, 4009, 4580, and 5029. Selected papers from the first meeting in 1990 appeared in volume 92 of *Theoretical Computer Science*, from the 2000 meeting in volume 2 of *Journal of Discrete Algorithms*, from the 2001 meeting in volume 146 of *Discrete Applied Mathematics*, from the 2003 meeting in volume 3 of *Journal of Discrete Algorithms*, from the 2004 meeting in volume 368 of *Theoretical Computer Science* and from the 2005 meeting in volume 5 of *Journal of Discrete Algorithms*. Selected papers from the 2008 edition are to appear in *Theoretical Computer Science*.

To mark the 20th anniversary of CPM, all Program Committee (Co-)Chairs of previous editions of CPM were invited to serve on the 2009 Program Committee. This resulted in a committee of 31 members including many prominent researchers in the area. Moreover, the proceedings open with a special invited contribution entitled “CPM’s 20th Anniversary: A Statistical Retrospective,” tracing the history of the symposium and providing a collection of statistical data and factual information about all the 20 editions of CPM and the presented contributions.

The Program Committee received 63 valid submissions. Each submission was reviewed independently by three committee members, possibly assisted by external reviewers. About 70 external reviewers provided their expertise; they are listed on the pages that follow. The selection process resulted in 27 accepted papers, corresponding to an acceptance rate of about 43%. The Program Committee decided to grant two awards to selected papers: a Best Paper Award and a new Best Student Paper Award. We would like to thank the members of the Program Committee who worked very hard to ensure the timely review of all the submitted manuscripts and participated in the selection process.

The conference program also included three invited talks by Christos Faloutsos (Carnegie Mellon University), Roberto Grossi (University of Pisa), and Ravi Kumar (Yahoo! Research), who graciously accepted the Program Committee’s invitation.

We are indebted to the members of the Steering Committee for their advice and tremendous help in different issues. On behalf of the entire CPM community, we would like to express our gratitude to the institutional sponsors who provided support to CPM 2009. These include the Laboratoire d’Informatique Fondamentale de Lille (UMR CNRS 8022), Université Lille 1, Région Nord-Pas de Calais, GDR Bioinformatique Moléculaire, INRIA Lille - Nord Europe, Yahoo! Research, and the University of Helsinki. The whole submission and review process was carried out with the help of the EasyChair system. Finally, we thank the local organization team headed by Hélène Touzet for carrying out all the laborious work that made the meeting possible.

March 2009

Gregory Kucherov  
Esko Ukkonen

# Organization

## Program Committee

Tatsuya Akutsu	Kyoto University, Japan
Amihood Amir	Bar-Ilan University, Israel and Johns Hopkins University, USA
Alberto Apostolico	Georgia Tech, USA and University of Padova, Italy
Ricardo Baeza-Yates	Yahoo! Research, Barcelona, Spain
Edgar Chávez	Universidad Michoacana, Mexico
Maxime Crochemore	King's College London, UK and Université Paris-Est, France
Martin Farach-Colton	Rutgers University and Tokutek Inc., USA
Paolo Ferragina	University of Pisa, Italy
Zvi Galil	Tel Aviv University, Israel
Raffaele Giancarlo	Università di Palermo, Italy
Dan Gusfield	University of California, Davis, USA
Daniel Hirschberg	University of California, Irvine, USA
Costas Iliopoulos	King's College London, UK
John Kececioglu	University of Arizona, USA
Gregory Kucherov (Co-chair)	CNRS, France
Gad Landau	University of Haifa, Israel
Moshe Lewenstein	Bar-Ilan University, Israel
Stefano Lonardi	University of California, Riverside, USA
Bin Ma	University of Waterloo, Canada
S. Muthukrishnan	Google Inc., New York, USA
Eugene Myers	Howard Hughes Medical Institute, USA
Kunsoo Park	Seoul National University, Korea
Mike Paterson	University of Warwick, UK
Wojciech Rytter	Uniwersytet Warszawski, Poland
S. Cenk Sahinalp	Simon Fraser University, Canada
David Sankoff	University of Ottawa, Canada
Masayuki Takeda	Kyushu University, Japan
Hélène Touzet	Université Lille 1 and INRIA, France
Esko Ukkonen (Co-chair)	University of Helsinki, Finland
Gabriel Valiente	Technical University of Catalonia, Spain
Kaizhong Zhang	University of Western Ontario, Canada

## Steering Committee

Alberto Apostolico	Georgia Tech, USA and University of Padova, Italy
Maxime Crochemore	King's College London, UK and Université Paris-Est, France
Zvi Galil	Tel Aviv University, Israel

## Organizing Committee

Sandrine Catillon	Antoine de Monte
Mathieu Giraud	Laurent Noé
Stéphane Janot	Maude Pupin
Juha Kärkkäinen	Hélène Touzet (Co-chair)
Janne Korhonen	Jean-Stéphane Varré
Gregory Kucherov (Co-chair)	

## External Referees

Can Alkan	Takuya Kida
Pavlos Antoniou	Shmuel Tomi Klein
Hideo Bannai	Pang Ko
Marek Biskup	Mikko Koivisto
Guillaume Blin	Marcin Kubica
Carlos Brizuela	Oded Lachish
Shihyen Chen	Giuseppe Lancia
Hamid Chitsaz	Theodoros Lappas
Manolis Christodoulakis	Thierry Lecroq
David Eppstein	Weiming Li
Antonio Fariña	Hao Lin
Thomas Fernique	Jingping Liu
Fedor Fomin	Xiaowen Liu
Roberto Grossi	Mercè Llabrés
Xi Han	Spiros Michalopoulos
Christophe Hancart	Lukasz Mikulski
Lin He	Igor Nitto
Danny Hermelin	Giulio Pavesi
Farhad Hormozdiari	Marcin Piatkowski
Fereydoun Hormozdiari	Solon Pissis
Lucian Ilie	Wojciech Plandowski
Shunsuke Inenaga	Ely Porat
Jesper Jansson	Jakub Radoszewski
Inuka Jayasekera	M. Sohel Rahman
Hossein Jowhari	Emanuele Raineri
Juha Kärkkäinen	Antonio Restivo

Paolo Ribeca  
Luís M.S. Russo  
Kunihiko Sadakane  
Mert Saglam  
Hiroshi Sakamoto  
Gilles Schaeffer  
Alexander Schönhuth  
Florian Sikora  
Bill Smyth

Wojciech Szpankowski  
Zdenek Tronícek  
Bora Uyar  
Rossano Venturini  
Stéphane Vialette  
Mark Ward  
Oren Weimann  
Michal Ziv-Ukelson

## **Sponsoring Institutions**

Laboratoire d'Informatique Fondamentale de Lille (UMR CNRS 8022)  
Université Lille 1  
Région Nord-Pas de Calais  
GDR Bioinformatique Moléculaire  
INRIA Lille - Nord Europe  
Yahoo! Research  
University of Helsinki



# Table of Contents

CPM's 20th Anniversary: A Statistical Retrospective . . . . .	1
<i>Elena Yavorska Harris, Thierry Lecroq, Gregory Kucherov, and Stefano Lonardi</i>	
Quasi-distinct Parsing and Optimal Compression Methods . . . . .	12
<i>Amihood Amir, Yonatan Aumann, Avivit Levy, and Yuri Roshko</i>	
Generalized Substring Compression . . . . .	26
<i>Orgad Keller, Tsvi Kopelowitz, Shir Landau, and Moshe Lewenstein</i>	
Text Indexing, Suffix Sorting, and Data Compression: Common Problems and Techniques (Invited Talk) . . . . .	39
<i>Roberto Grossi</i>	
Contracted Suffix Trees: A Simple and Dynamic Text Indexing Data Structure . . . . .	41
<i>Andrzej Ehrenfeucht, Ross M. McConnell, and Sung-Whan Woo</i>	
Linear Time Suffix Array Construction Using D-Critical Substrings . . . .	54
<i>Ge Nong, Sen Zhang, and Wai Hong Chan</i>	
On the Value of Multiple Read/Write Streams for Data Compression . . .	68
<i>Travis Gagie</i>	
Reoptimization of the Shortest Common Superstring Problem (Extended Abstract) . . . . .	78
<i>Davide Bilò, Hans-Joachim Böckenhauer, Dennis Komm, Richard Kráľovič, Tobias Mömke, Sebastian Seibert, and Anna Zych</i>	
LCS Approximation via Embedding into Local Non-repetitive Strings . . . . .	92
<i>Gad M. Landau, Avivit Levy, and Ilan Newman</i>	
An Efficient Matching Algorithm for Encoded DNA Sequences and Binary Strings . . . . .	106
<i>Simone Faro and Thierry Lecroq</i>	
Fast Searching in Packed Strings . . . . .	116
<i>Philip Bille</i>	
New Complexity Bounds for Image Matching under Rotation and Scaling . . . . .	127
<i>Christian Hundt and Maciej Liśkiewicz</i>	

Online Approximate Matching with Non-local Distances . . . . .	142
<i>Raphaël Clifford and Benjamin Sach</i>	
Faster and Space-Optimal Edit Distance “1” Dictionary . . . . .	154
<i>Djamal Belazzougui</i>	
Approximate Matching for Run-Length Encoded Strings Is 3SUM-Hard . . . . .	168
<i>Kuan-Yu Chen, Ping-Hui Hsu, and Kun-Mao Chao</i>	
Modeling and Algorithmic Challenges in Online Social Networks (Invited Talk) . . . . .	180
<i>Ravi Kumar</i>	
Permuted Longest-Common-Prefix Array . . . . .	181
<i>Juha Kärkkäinen, Giovanni Manzini, and Simon J. Puglisi</i>	
Periodic String Comparison . . . . .	193
<i>Alexander Tiskin</i>	
Deconstructing Intractability: A Case Study for Interval Constrained Coloring . . . . .	207
<i>Christian Komusiewicz, Rolf Niedermeier, and Johannes Uhlmann</i>	
Maximum Motif Problem in Vertex-Colored Graphs . . . . .	221
<i>Riccardo Dondi, Guillaume Fertin, and Stéphane Vialette</i>	
Fast RNA Structure Alignment for Crossing Input Structures . . . . .	236
<i>Rolf Backofen, Gad M. Landau, Mathias Möhl, Dekel Tsur, and Oren Weimann</i>	
Sparse RNA Folding: Time and Space Efficient Algorithms . . . . .	249
<i>Rolf Backofen, Dekel Tsur, Shay Zakov, and Michal Ziv-Ukelson</i>	
Multiple Alignment of Biological Networks: A Flexible Approach . . . . .	263
<i>Yves-Pol Deniélou, Frédéric Boyer, Alain Viari, and Marie-France Sagot</i>	
Graph Mining: Patterns, Generators and Tools (Invited Talk) . . . . .	274
<i>Christos Faloutsos</i>	
Level- $k$ Phylogenetic Networks Are Constructable from a Dense Triplet Set in Polynomial Time . . . . .	275
<i>Thu-Hien To and Michel Habib</i>	
The Structure of Level- $k$ Phylogenetic Networks . . . . .	289
<i>Philippe Gambette, Vincent Berry, and Christophe Paul</i>	
Finding All Sorting Tandem Duplication Random Loss Operations . . . . .	301
<i>Matthias Bernt, Ming-Chiang Chen, Daniel Merkle, Hung-Lung Wang, Kun-Mao Chao, and Martin Middendorf</i>	

Average-Case Analysis of Perfect Sorting by Reversals . . . . .	314
<i>Mathilde Bowvel, Cedric Chauve, Marni Mishna, and Dominique Rossin</i>	
Statistical Properties of Factor Oracles . . . . .	326
<i>J�r�mie Bourdon and Irena Rusu</i>	
Haplotype Inference Constrained by Plausible Haplotype Data . . . . .	339
<i>Michael R. Fellows, Tzvika Hartman, Danny Hermelin, Gad M. Landau, Frances Rosamond, and Liat Rozenberg</i>	
Efficient Inference of Haplotypes from Genotypes on a Pedigree with Mutations and Missing Alleles (Extended Abstract) . . . . .	353
<i>Wei-Bung Wang and Tao Jiang</i>	
<b>Author Index</b> . . . . .	369