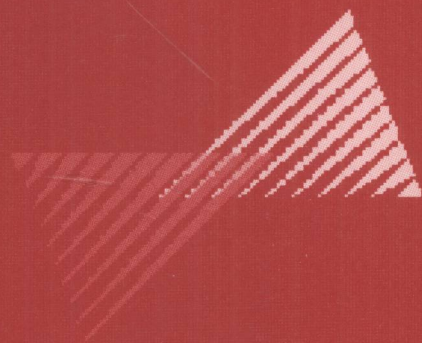


Bin Ma
Kaizhong Zhang (Eds.)

LNCS 4580

Combinatorial Pattern Matching

18th Annual Symposium, CPM 2007
London, Canada, July 2007
Proceedings



Springer

TP301.6-53

C731

2007

Bin Ma Kaizhong Zhang (Eds.)

Combinatorial Pattern Matching

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



 Springer



E2007003296

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

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

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 EasyChiar 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

VIII Organization

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

Lecture Notes in Computer Science

For information about Vols. 1–4492

please contact your bookseller or Springer

Vol. 4600: H. Comon-Lundh, C. Kirchner, H. Kirchner, *Rewriting, Computation and Proof*. XVI, 273 pages. 2007.

Vol. 4595: D. Bošnački, S. Edelkamp (Eds.), *Model Checking Software*. X, 285 pages. 2007.

Vol. 4592: Z. Kedad, N. Lammari, E. Métais, F. Meziane, Y. Rezgui (Eds.), *Natural Language Processing and Information Systems*. XIV, 442 pages. 2007.

Vol. 4591: J. Davies, J. Gibbons (Eds.), *Integrated Formal Methods*. IX, 660 pages. 2007.

Vol. 4590: W. Damm, H. Hermanns (Eds.), *Computer Aided Verification*. XV, 562 pages. 2007.

Vol. 4589: J. Münch, P. Abrahamsson (Eds.), *Product-Focused Software Process Improvement*. XII, 414 pages. 2007.

Vol. 4588: T. Harju, J. Karhumäki, A. Lepistö (Eds.), *Developments in Language Theory*. XI, 423 pages. 2007.

Vol. 4587: R. Cooper, J. Kennedy (Eds.), *Data Management*. XIII, 259 pages. 2007.

Vol. 4586: J. Pieprzyk, H. Ghodosi, E. Dawson (Eds.), *Information Security and Privacy*. XIV, 476 pages. 2007.

Vol. 4585: M. Kryszkiewicz, J.F. Peters, H. Rybinski, A. Skowron (Eds.), *Rough Sets and Intelligent Systems Paradigms*. XIX, 836 pages. 2007. (Sublibrary LNAI).

Vol. 4584: N. Karssemeijer, B. Lelieveldt (Eds.), *Information Processing in Medical Imaging*. XX, 777 pages. 2007.

Vol. 4583: S.R. Della Rocca (Ed.), *Typed Lambda Calculi and Applications*. X, 397 pages. 2007.

Vol. 4582: J. Lopez, P. Samarati, J.L. Ferrer (Eds.), *Public Key Infrastructure*. XI, 375 pages. 2007.

Vol. 4581: A. Petrenko, M. Veanes, J. Tretmans, W. Grieskamp (Eds.), *Testing of Software and Communicating Systems*. XII, 379 pages. 2007.

Vol. 4580: B. Ma, K. Zhang (Eds.), *Combinatorial Pattern Matching*. XII, 366 pages. 2007.

Vol. 4578: F. Masulli, S. Mitra, G. Pasi (Eds.), *Applications of Fuzzy Sets Theory*. XVIII, 693 pages. 2007. (Sublibrary LNAI).

Vol. 4577: N. Sebe, Y. Liu, Y. Zhuang (Eds.), *Multimedia Content Analysis and Mining*. XIII, 513 pages. 2007.

Vol. 4576: D. Leivant, R. de Queiroz (Eds.), *Logic, Language, Information, and Computation*. X, 363 pages. 2007.

Vol. 4574: J. Derrick, J. Vain (Eds.), *Formal Techniques for Networked and Distributed Systems – FORTE 2007*. XI, 375 pages. 2007.

Vol. 4573: M. Kauers, M. Kerber, R. Miner, W. Windsteiger (Eds.), *Towards Mechanized Mathematical Assistants*. XIII, 407 pages. 2007. (Sublibrary LNAI).

Vol. 4572: F. Stajano, C. Meadows, S. Capkun, T. Moore (Eds.), *Security and Privacy in Ad-hoc and Sensor Networks*. X, 247 pages. 2007.

Vol. 4570: H.G. Okuno, M. Ali (Eds.), *New Trends in Applied Artificial Intelligence*. XXI, 1194 pages. 2007. (Sublibrary LNAI).

Vol. 4569: A. Butz, B. Fisher, A. Krüger, P. Olivier, S. Owada (Eds.), *Smart Graphics*. IX, 237 pages. 2007.

Vol. 4566: M.J. Dainoff (Ed.), *Ergonomics and Health Aspects of Work with Computers*. XVIII, 390 pages. 2007.

Vol. 4565: D.D. Schmorow, L.M. Reeves (Eds.), *Foundations of Augmented Cognition*. XIX, 450 pages. 2007. (Sublibrary LNAI).

Vol. 4564: D. Schuler (Ed.), *Online Communities and Social Computing*. XVII, 520 pages. 2007.

Vol. 4563: R. Shumaker (Ed.), *Virtual Reality*. XXII, 762 pages. 2007.

Vol. 4561: V.G. Duffy (Ed.), *Digital Human Modeling*. XXIII, 1068 pages. 2007.

Vol. 4560: N. Aykin (Ed.), *Usability and Internationalization, Part II*. XVIII, 576 pages. 2007.

Vol. 4559: N. Aykin (Ed.), *Usability and Internationalization, Part I*. XVIII, 661 pages. 2007.

Vol. 4549: J. Aspnes, C. Scheideler, A. Arora, S. Madden (Eds.), *Distributed Computing in Sensor Systems*. XIII, 417 pages. 2007.

Vol. 4548: N. Olivetti (Ed.), *Automated Reasoning with Analytic Tableaux and Related Methods*. X, 245 pages. 2007. (Sublibrary LNAI).

Vol. 4547: C. Carlet, B. Sunar (Eds.), *Arithmetic of Finite Fields*. XI, 355 pages. 2007.

Vol. 4546: J. Kleijn, A. Yakovlev (Eds.), *Petri Nets and Other Models of Concurrency – ICATPN 2007*. XI, 515 pages. 2007.

Vol. 4545: H. Anai, K. Horimoto, T. Kutsia (Eds.), *Algebraic Biology*. XIII, 379 pages. 2007.

Vol. 4544: S. Cohen-Boulakia, V. Tannen (Eds.), *Data Integration in the Life Sciences*. XI, 282 pages. 2007. (Sublibrary LNBI).

Vol. 4543: A.K. Bandara, M. Burgess (Eds.), *Inter-Domain Management*. XII, 237 pages. 2007.

Vol. 4542: P. Sawyer, B. Paech, P. Heymans (Eds.), *Requirements Engineering: Foundation for Software Quality*. IX, 384 pages. 2007.

- Vol. 4541: T. Okadome, T. Yamazaki, M. Makhtari (Eds.), *Pervasive Computing for Quality of Life Enhancement*. IX, 248 pages. 2007.
- Vol. 4539: N.H. Bshouty, C. Gentile (Eds.), *Learning Theory*. XII, 634 pages. 2007. (Sublibrary LNAI).
- Vol. 4538: F. Escolano, M. Vento (Eds.), *Graph-Based Representations in Pattern Recognition*. XII, 416 pages. 2007.
- Vol. 4537: K.C.-C. Chang, W. Wang, L. Chen, C.A. Ellis, C.-H. Hsu, A.C. Tsoi, H. Wang (Eds.), *Advances in Web and Network Technologies, and Information Management*. XXIII, 707 pages. 2007.
- Vol. 4536: G. Concas, E. Damiani, M. Scotto, G. Succì (Eds.), *Agile Processes in Software Engineering and Extreme Programming*. XV, 276 pages. 2007.
- Vol. 4534: I. Tomkos, F. Neri, J. Solé Pareta, X. Masip Bruin, S. Sánchez Lopez (Eds.), *Optical Network Design and Modeling*. XI, 460 pages. 2007.
- Vol. 4533: F. Baader (Ed.), *Term Rewriting and Applications*. XII, 419 pages. 2007.
- Vol. 4531: J. Indulska, K. Raymond (Eds.), *Distributed Applications and Interoperable Systems*. XI, 337 pages. 2007.
- Vol. 4530: D.H. Akehurst, R. Vogel, R.F. Paige (Eds.), *Model Driven Architecture- Foundations and Applications*. X, 219 pages. 2007.
- Vol. 4529: P. Melin, O. Castillo, L.T. Aguilar, J. Kacprzyk, W. Pedrycz (Eds.), *Foundations of Fuzzy Logic and Soft Computing*. XIX, 830 pages. 2007. (Sublibrary LNAI).
- Vol. 4528: J. Mira, J.R. Álvarez (Eds.), *Nature Inspired Problem-Solving Methods in Knowledge Engineering*, Part II. XXII, 650 pages. 2007.
- Vol. 4527: J. Mira, J.R. Álvarez (Eds.), *Bio-inspired Modeling of Cognitive Tasks*, Part I. XXII, 630 pages. 2007.
- Vol. 4526: M. Malek, M. Reitenspieß, A. van Moorsel (Eds.), *Service Availability*. X, 155 pages. 2007.
- Vol. 4525: C. Demetrescu (Ed.), *Experimental Algorithms*. XIII, 448 pages. 2007.
- Vol. 4524: M. Marchiori, J.Z. Pan, C.d.S. Marie (Eds.), *Web Reasoning and Rule Systems*. XI, 382 pages. 2007.
- Vol. 4523: Y.-H. Lee, H.-N. Kim, J. Kim, Y. Park, L.T. Yang, S.W. Kim (Eds.), *Embedded Software and Systems*. XIX, 829 pages. 2007.
- Vol. 4522: B.K. Ersbøll, K.S. Pedersen (Eds.), *Image Analysis*. XVIII, 989 pages. 2007.
- Vol. 4521: J. Katz, M. Yung (Eds.), *Applied Cryptography and Network Security*. XIII, 498 pages. 2007.
- Vol. 4519: E. Franconi, M. Kifer, W. May (Eds.), *The Semantic Web: Research and Applications*. XVIII, 830 pages. 2007.
- Vol. 4517: F. Boavida, E. Monteiro, S. Mascolo, Y. Koucheryavy (Eds.), *Wired/Wireless Internet Communications*. XIV, 382 pages. 2007.
- Vol. 4516: L. Mason, T. Drwiega, J. Yan (Eds.), *Managing Traffic Performance in Converged Networks*. XXIII, 1191 pages. 2007.
- Vol. 4515: M. Naor (Ed.), *Advances in Cryptology - EUROCRYPT 2007*. XIII, 591 pages. 2007.
- Vol. 4514: S.N. Artemov, A. Nerode (Eds.), *Logical Foundations of Computer Science*. XI, 513 pages. 2007.
- Vol. 4513: M. Fischetti, D.P. Williamson (Eds.), *Integer Programming and Combinatorial Optimization*. IX, 500 pages. 2007.
- Vol. 4511: C. Conati, K. McCoy, G. Paliouras (Eds.), *User Modeling 2007*. XVI, 487 pages. 2007. (Sublibrary LNAI).
- Vol. 4510: P. Van Hentenryck, L. Wolsey (Eds.), *Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems*. X, 391 pages. 2007.
- Vol. 4509: Z. Kobti, D. Wu (Eds.), *Advances in Artificial Intelligence*. XII, 552 pages. 2007. (Sublibrary LNAI).
- Vol. 4508: M.-Y. Kao, X.-Y. Li (Eds.), *Algorithmic Aspects in Information and Management*. VIII, 428 pages. 2007.
- Vol. 4507: F. Sandoval, A. Prieto, J. Cabestany, M. Graña (Eds.), *Computational and Ambient Intelligence*. XXVI, 1167 pages. 2007.
- Vol. 4506: D. Zeng, I. Gotham, K. Komatsu, C. Lynch, M. Thurmond, D. Madigan, B. Lober, J. Kvach, H. Chen (Eds.), *Intelligence and Security Informatics: Biosurveillance*. XI, 234 pages. 2007.
- Vol. 4505: G. Dong, X. Lin, W. Wang, Y. Yang, J.X. Yu (Eds.), *Advances in Data and Web Management*. XXII, 896 pages. 2007.
- Vol. 4504: J. Huang, R. Kowalczyk, Z. Maamar, D. Martin, I. Müller, S. Stoutenburg, K.P. Sycara (Eds.), *Service-Oriented Computing: Agents, Semantics, and Engineering*. X, 175 pages. 2007.
- Vol. 4501: J. Marques-Silva, K.A. Sakallah (Eds.), *Theory and Applications of Satisfiability Testing - SAT 2007*. XI, 384 pages. 2007.
- Vol. 4500: N. Streitz, A. Kameas, I. Mavrommati (Eds.), *The Disappearing Computer*. XVIII, 304 pages. 2007.
- Vol. 4499: Y.Q. Shi (Ed.), *Transactions on Data Hiding and Multimedia Security II*. IX, 117 pages. 2007.
- Vol. 4498: N. Abdennahder, F. Kordon (Eds.), *Reliable Software Technologies - Ada Europe 2007*. XII, 247 pages. 2007.
- Vol. 4497: S.B. Cooper, B. Löwe, A. Sorbi (Eds.), *Computation and Logic in the Real World*. XVIII, 826 pages. 2007.
- Vol. 4496: N.T. Nguyen, A. Grzech, R.J. Howlett, L.C. Jain (Eds.), *Agent and Multi-Agent Systems: Technologies and Applications*. XXI, 1046 pages. 2007. (Sublibrary LNAI).
- Vol. 4495: J. Krogstie, A. Opdahl, G. Sindre (Eds.), *Advanced Information Systems Engineering*. XVI, 606 pages. 2007.
- Vol. 4494: H. Jin, O.F. Rana, Y. Pan, V.K. Prasanna (Eds.), *Algorithms and Architectures for Parallel Processing*. XIV, 508 pages. 2007.
- Vol. 4493: D. Liu, S. Fei, Z. Hou, H. Zhang, C. Sun (Eds.), *Advances in Neural Networks - ISNN 2007*, Part III. XXVI, 1215 pages. 2007.

¥562.00元

Table of Contents

Invited Talks (Abstracts)

A Combinatorial Approach to Genome-Wide Ortholog Assignment: Beyond Sequence Similarity Search	1
<i>Tao Jiang</i>	
Stringology: Some Classic and Some Modern Problems	2
<i>S. Muthukrishnan</i>	
Algorithmic Problems in Scheduling Jobs on Variable-Speed Processors	3
<i>Frances F. Yao</i>	

Session 1: Algorithmic Techniques I

Speeding Up HMM Decoding and Training by Exploiting Sequence Repetitions	4
<i>Shay Mozes, Oren Weimann, and Michal Ziv-Ukelson</i>	
On Demand String Sorting over Unbounded Alphabets	16
<i>Carmel Kent, Moshe Lewenstein, and Dafna Sheinwald</i>	

Session 2: Approximate Pattern Matching

Finding Witnesses by Peeling	28
<i>Yonatan Aumann, Moshe Lewenstein, Noa Lewenstein, and Dekel Tsur</i>	
Cache-Oblivious Index for Approximate String Matching	40
<i>Wing-Kai Hon, Tak-Wah Lam, Rahul Shah, Siu-Lung Tam, and Jeffrey Scott Vitter</i>	
Improved Approximate String Matching and Regular Expression Matching on Ziv-Lempel Compressed Texts	52
<i>Philip Bille, Rolf Fagerberg, and Inge Li Gørtz</i>	
Self-normalised Distance with Don't Cares	63
<i>Peter Clifford and Raphaël Clifford</i>	

Session 3: Data Compression I

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

A Lempel-Ziv Text Index on Secondary Storage	83
<i>Diego Arroyuelo and Gonzalo Navarro</i>	
Dynamic Rank-Select Structures with Applications to Run-Length Encoded Texts.....	95
<i>Sunho Lee and Kunsoo Park</i>	
Most Burrows-Wheeler Based Compressors Are Not Optimal	107
<i>Haim Kaplan and Elad Verbin</i>	
Session 4: Computational Biology I	
Non-breaking Similarity of Genomes with Gene Repetitions	119
<i>Zhixiang Chen, Bin Fu, Jinhui Xu, Boting Yang, Zhiyu Zhao, and Binhai Zhu</i>	
A New and Faster Method of Sorting by Transpositions.....	131
<i>Maxime Benoît-Gagné and Sylvie Hamel</i>	
Finding Compact Structural Motifs	142
<i>Jianbo Qian, Shuai Cheng Li, Dongbo Bu, Ming Li, and Jinbo Xu</i>	
Session 5: Computational Biology II	
Improved Algorithms for Inferring the Minimum Mosaic of a Set of Recombinants	150
<i>Yufeng Wu and Dan Gusfield</i>	
Computing Exact p-Value for Structured Motif	162
<i>Jing Zhang, Xi Chen, and Ming Li</i>	
Session 6: Algorithmic Techniques II	
Improved Sketching of Hamming Distance with Error Correcting.....	173
<i>Ely Porat and Ohad Lipsky</i>	
Deterministic Length Reduction: Fast Convolution in Sparse Data and Applications.....	183
<i>Amihud Amir, Oren Kapah, and Ely Porat</i>	
Guided Forest Edit Distance: Better Structure Comparisons by Using Domain-knowledge	195
<i>Zeshan Peng and Hing-fung Ting</i>	
Space-Efficient Algorithms for Document Retrieval	205
<i>Niko Välimäki and Veli Mäkinen</i>	

Session 7: Data Compression II

Compressed Text Indexes with Fast Locate	216
<i>Rodrigo González and Gonzalo Navarro</i>	
Processing Compressed Texts: A Tractability Border	228
<i>Yury Lifshits</i>	

Session 8: Computational Biology III

Common Structured Patterns in Linear Graphs: Approximation and Combinatorics	241
<i>Guillaume Fertin, Danny Hermelin, Romeo Rizzi, and Stéphane Vialette</i>	
Identification of Distinguishing Motifs	253
<i>WangSen Feng, Zhanyong Wang, and Lusheng Wang</i>	
Algorithms for Computing the Longest Parameterized Common Subsequence	265
<i>Costas S. Iliopoulos, Marcin Kubica, M. Sohel Rahman, and Tomasz Waleń</i>	
Fixed-Parameter Tractability of the Maximum Agreement Supertree Problem	274
<i>Sylvain Guillemot and Vincent Berry</i>	

Session 9: Pattern Analysis

Two-Dimensional Range Minimum Queries	286
<i>Amihoud Amir, Johannes Fischer, and Moshe Lewenstein</i>	
Tiling Periodicity	295
<i>Juhani Karhumäki, Yury Lifshits, and Wojciech Rytter</i>	
Fast and Practical Algorithms for Computing All the Runs in a String ..	307
<i>Gang Chen, Simon J. Puglisi, and W.F. Smyth</i>	
Longest Common Separable Pattern Among Permutations	316
<i>Mathilde Bouvel, Dominique Rossin, and Stéphane Vialette</i>	

Session 10: Suffix Arrays and Trees

Suffix Arrays on Words	328
<i>Paolo Ferragina and Johannes Fischer</i>	

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

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

Tao Jiang

Computer Science Department, University of California - Riverside
jiang@cs.ucr.edu

Abstract. The assignment of orthologous genes between a pair of genomes is a fundamental and challenging problem in comparative genomics. Existing methods that assign orthologs based on the similarity between DNA or protein sequences may make erroneous assignments when sequence similarity does not clearly delineate the evolutionary relationship among genes of the same families. In this paper, we present a new approach to ortholog assignment that takes into account both sequence similarity and evolutionary events at genome level, where orthologous genes are assumed to correspond to each other in the most parsimonious evolving scenario under genome rearrangement and gene duplication. It is then formulated as a problem of computing the signed reversal distance with duplicates between two genomes of interest, for which an efficient heuristic algorithm was constructed based on solutions to two new optimization problems, minimum common partition and maximum cycle decomposition. Following this approach, we have implemented a high-throughput system for assigning orthologs on a genome scale, called MSOAR, and tested it on both simulated data and real genome sequence data. Our predicted orthologs between the human and mouse genomes are strongly supported by ortholog and protein function information in authoritative databases, and predictions made by other key ortholog assignment methods such as Ensembl, Homologene, INPARANOID, and HGNC. The simulation results demonstrate that MSOAR in general performs better than the iterated exemplar algorithm of D. Sankoff's in terms of identifying true exemplar genes.

This is joint work with X. Chen (Nanyang Tech. Univ., Singapore), Z. Fu (UCR), J. Zheng (NCBI), V. Vacic (UCR), P. Nan (SCBIT), Y. Zhong (SCBIT), and S. Lonardi (UCR).

Stringology: Some Classic and Some Modern Problems

S. Muthukrishnan

Department of Computer Science, Rutgers University
and
Google Inc.
`muthu@cs.rutgers.edu`

Abstract. We examine some of the classic problems related to suffix trees from 70's and show some recent results on sorting suffixes with small space and suffix selection. Further, we introduce modern versions of suffix sorting and their application to XML processing. The study of combinatorial aspects of strings continues to flourish, and we present several open problems with modern applications.

Algorithmic Problems in Scheduling Jobs on Variable-Speed Processors

Frances F. Yao

Department of Computer Science,
City University of Hong Kong
Hong Kong SAR, China
`csfyao@cityu.edu.hk`

Abstract. Power and heat have become two of the major concerns for the computer industry, which struggles to cope with the energy and cooling costs for servers, as well as the short battery life of portable devices. Dynamic Voltage Scaling (DVS) has emerged as a useful technique: e.g. Intel's newest Foxton technology enables a chip to run at 64 different speed levels. Equipped with DVS technology, the operating system can then save CPU's energy consumption by scheduling tasks wisely. A schedule that finishes the given tasks within their timing constraints while using minimum total energy (among all feasible schedules) is called an optimal DVS schedule. A theoretical model for DVS scheduling was proposed in a paper by Yao, Demers and Shenker in 1995, along with a well-formed characterization of the optimum and an algorithm for computing it. This algorithm has remained as the most efficient known despite many investigations of this model. In this talk, we will first give an overview of the DVS scheduling problem, and then present the latest improved results for computing the optimal schedule in both the finite and the continuous (infinite speed levels) models. Related results on efficient on-line scheduling heuristics will also be discussed.

Speeding Up HMM Decoding and Training by Exploiting Sequence Repetitions

Shay Mozes^{1,*}, Oren Weimann¹, and Michal Ziv-Ukelson^{2,**}

¹ MIT Computer Science and Artificial Intelligence Laboratory,
32 Vassar Street, Cambridge, MA 02139, USA
shaymozes@gmail.com, oweimann@mit.edu

² School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel
michaluz@post.tau.ac.il

Abstract. We present a method to speed up the dynamic program algorithms used for solving the HMM decoding and training problems for discrete time-independent HMMs. We discuss the application of our method to Viterbi's decoding and training algorithms [21], as well as to the forward-backward and Baum-Welch [4] algorithms. Our approach is based on identifying repeated substrings in the observed input sequence. We describe three algorithms based alternatively on byte pair encoding (BPE) [19], run length encoding (RLE) and Lempel-Ziv (LZ78) parsing [22]. Compared to Viterbi's algorithm, we achieve a speedup of $\Omega(r)$ using BPE, a speedup of $\Omega(\frac{r}{\log r})$ using RLE, and a speedup of $\Omega(\frac{\log n}{k})$ using LZ78, where k is the number of hidden states, n is the length of the observed sequence and r is its compression ratio (under each compression scheme). Our experimental results demonstrate that our new algorithms are indeed faster in practice. Furthermore, unlike Viterbi's algorithm, our algorithms are highly parallelizable.

Keywords: HMM, Viterbi, dynamic programming, compression.

1 Introduction

Over the last few decades, Hidden Markov Models (HMMs) proved to be an extremely useful framework for modeling processes in diverse areas such as error-correction in communication links [21], speech recognition [6], optical character recognition [2], computational linguistics [17], and bioinformatics [12].

The core HMM-based applications fall in the domain of classification methods and are technically divided into two stages: a training stage and a decoding stage. During the *training* stage, the emission and transition probabilities of an HMM are estimated, based on an input set of observed sequences. This stage is usually executed once as a preprocessing stage and the generated ("trained") models are stored in a database. Then, a *decoding* stage is run, again and again, in order to

* Work conducted while visiting MIT.

** Work supported by an **Eshkol grant** of the Israeli Ministry of Science and Technology.