

Subseries of Lecture Notes in Computer Science

LNBI Series Editors

Sorin Istrail

Brown University, Providence, RI, USA

Pavel Pevzner

University of California, San Diego, CA, USA

Michael Waterman

University of Southern California, Los Angeles, CA, USA

LNBI Editorial Board

Søren Brunak

Technical University of Denmark, Kongens Lyngby, Denmark

Mikhail S. Gelfand

IITP, Research and Training Center on Bioinformatics, Moscow, Russia

Thomas Lengauer

Max Planck Institute for Informatics, Saarbrücken, Germany

Satoru Miyano

University of Tokyo, Tokyo, Japan

Eugene Myers

Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany

Marie-France Sagot

Université Lyon 1, Villeurbanne, France

David Sankoff

University of Ottawa, Ottawa, Canada

Ron Shamir

Tel Aviv University, Ramat Aviv, Tel Aviv, Israel

Terry Speed

Walter and Eliza Hall Institute of Medical Research, Melbourne, VIC, Australia

Martin Vingron

Max Planck Institute for Molecular Genetics, Berlin, Germany

W. Eric Wong

University of Texas at Dallas, Richardson, TX, USA

More information about this series at <http://www.springer.com/series/5381>

Martin Frith
Christian Nørgaard Storm Pedersen (Eds.)

Algorithms in Bioinformatics

16th International Workshop, WABI 2016
Aarhus, Denmark, August 22–24, 2016
Proceedings

Editors

Martin Frith
AIST and University of Tokyo
Tokyo
Japan

Christian Nørgaard Storm Pedersen
Aarhus University
Aarhus
Denmark

ISSN 0302-9743

Lecture Notes in Bioinformatics

ISBN 978-3-319-43680-7

DOI 10.1007/978-3-319-43681-4

ISSN 1611-3349 (electronic)

ISBN 978-3-319-43681-4 (eBook)

Library of Congress Control Number: 2016945963

LNCS Sublibrary: SL8 – Bioinformatics

© Springer International Publishing Switzerland 2016

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, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature

The registered company is Springer International Publishing AG Switzerland

Preface

This proceedings volume contains papers presented at the 16th Workshop on Algorithms in Bioinformatics (WABI 2016) that was held at Aarhus University, Aarhus, Denmark, August 22–24, 2016. WABI 2016 was one of eight conferences that were organized as part of ALGO 2016. The Workshop on Algorithms in Bioinformatics was established in 2001, and is an annual conference on all aspects of algorithmic work in bioinformatics, computational biology, and systems biology. The emphasis is on discrete algorithms and machine-learning methods that address important problems in molecular biology, that are founded on sound models, that are computationally efficient, and that have been implemented and tested in simulations and on real datasets. The goal is to present recent research results, including significant work-in-progress, and to identify and explore directions of future research. WABI 2016 was sponsored by the European Association for Theoretical Computer Science (EATCS).

In 2016, a total of 56 manuscripts were submitted to WABI from which 27 were selected for presentation at the conference. Among them, 25 are included in this proceedings volume as full papers presenting novel results not previously published in journals, and two are included as short abstracts of papers that are in the process of being published simultaneously in journals. The 27 papers were selected based on thorough reviewing, usually involving three independent reviewers per submitted paper, followed by discussions in the WABI Program Committee. The selected papers cover a wide range of topics from networks, to phylogenetic studies, sequence and genome analysis, comparative genomics, and mass spectrometry data analysis. Extended versions of selected papers will be published in a thematic series in the journal *Algorithms for Molecular Biology* (AMB), published by BioMed Central.

We thank all the authors of submitted papers and the members of the WABI Program Committee and their reviewers for their efforts that made this conference possible, and the WABI Steering Committee for their help and advice. We also thank all the conference participants and speakers. In particular, we are indebted to the keynote speaker of the conference, Kiyoshi Asai, for his presentation. Finally, we thank Gerth Stølting Brodal and the local ALGO Organizing Committee for their hard work.

June 2016

Martin Frith
Christian N. S. Pedersen

Organization

Program Chairs

Martin Frith AIST and University of Tokyo, Japan
Christian N.S. Pedersen Aarhus University, Denmark

Program Committee

Tatsuya Akutsu Kyoto University, Japan
Timothy L. Bailey University of Queensland, Australia
Jan Baumbach University of Southern Denmark, Denmark
Anne Bergeron Université du Québec à Montréal, Canada
Paola Bonizzoni Università di Milano-Bicocca, Italy
Alessandra Carbone Université Pierre et Marie Curie, France
Rita Casadio UNIBO, Italy
Nadia El-Mabrouk University of Montreal, Canada
Anna Gambin Warsaw University, Poland
Raffaele Giancarlo Università di Palermo, Italy
Michiaki Hamada Waseda University, Japan
Thomas Hamelryck University of Copenhagen, Denmark
Fereydoun Hormozdiari University of Washington, USA
Katharina Huber University of East Anglia, UK
Carl Kingsford Carnegie Mellon University, USA
Hisanori Kiryu University of Tokyo, Japan
Gregory Kucherov CNRS/LIGM, France
Timo Lassmann Telethon Kids, Australia
Ming Li University of Waterloo, Canada
Zsuzsanna Liptak University of Verona, Italy
Stefano Lonardi University of California at Riverside, USA
Gerton Lunter University of Oxford, UK
Thomas Mailund Aarhus University, Denmark
Paul Medvedev Pennsylvania State University, USA
Daniel Merkle University of Southern Denmark, Denmark
István Miklós Rényi Institute, Hungary
Bernard Moret EPFL, Switzerland
Burkhard Morgenstern University of Göttingen, Germany
Vincent Moulton University of East Anglia, UK
Veli Mäkinen University of Helsinki, Finland
Luay Nakhleh Rice University, USA
William Noble University of Washington, USA

Nadia Pisanti	Università di Pisa, Italy, and Inria, France
Mihai Pop	University of Maryland, USA
Teresa Przytycka	NIH, USA
Sven Rahmann	University of Duisburg-Essen, Germany
Marie-France Sagot	Inria, France
Kengo Sato	Keio University, Japan
Michael Schatz	Cold Spring Harbor Laboratory, USA
Russell Schwartz	Carnegie Mellon University, USA
Kana Shimizu	Waseda University, Japan
Anish Man Singh Shrestha	University of Tokyo, Japan
Peter F. Stadler	University of Leipzig, Germany
Jens Stoye	Bielefeld University, Germany
Krister Swenson	CNRS, Université de Montpellier, France
Hélène Touzet	CNRS, University of Lille and Inria, France
Lusheng Wang	City University of Hong Kong, China
Siu Ming Yiu	University of Hong Kong, SAR China
Louxin Zhang	National University of Singapore, Singapore
Michal Ziv-Ukelson	Ben-Gurion University of the Negev, Israel

WABI Steering Committee

Bernard Moret	EPFL, Switzerland
Vincent Moulton	University of East Anglia, UK
Jens Stoye	Bielefeld University, Germany
Tandy Warnow	University of Illinois at Urbana-Champaign, USA

ALGO Organizing Committee

Gerth Stølting Brodal (Chair)	Trine Ji Holmgaard
Marianne Dammand Iversen	Katrine Østerlund Rasmussen

Additional Reviewers

Nicolas Alcaraz	Pietro Di Lena
Hind Alhakami	Daniel Doerr
Eloi Araujo	Norbert Dojer
Matthias Bernt	Mikhail Dubov
Karel Brinda	Oliver Eulenstein
Laurent Bulteau	Pedro Feijao
Victoria Cepeda	Jay Ghurye
Daniel Cooke	Krzysztof Gogolewski
Phuong Dao	Roberto Grossi
Gianluca Della Vedova	Laurent Gueguen
Alex Di Genova	Marc Hellmuth

Donna Henderson
Farhad Hormozdiari
Jeff Howbert
Alex Hu
Laurent Jacob
Mateusz Krzysztof Łącki
Manuel Lafond
Cong Ma
Guillaume Marçais
Damon May
Blazej Miasojedow
Ibrahim Numanagic
Rachid Ounit
Solon Pissis

Raffaella Rizzi
Abbas Roayaei Ardakany
Giovanna Rosone
Yutaka Saito
Guillaume Scholz
Marcel Schulz
Celine Scornavacca
Mingfu Shao
Grzegorz Skoraczyński
Bianca Stöcker
Peng Sun
Hao Wang
Lusheng Wang
Martin Weigt

Abstracts

Mass Graphs and Their Applications in Top-Down Proteomics

Qiang Kou¹, Si Wu², Nikola Tolić³, Yunlong Liu^{4,5},
Ljiljana Paša-Tolić³ and Xiaowen Liu^{1,5}(✉)

¹ Department of BioHealth Informatics,
Indiana University-Purdue University Indianapolis, Indianapolis, Indiana

² Department of Chemistry and Biochemistry,
University of Oklahoma, Norman, Oklahoma

³ Biological Science Division, Pacific Northwest National Laboratory,
Richland, USA

⁴ Department of Medical and Molecular Genetics,
Indiana University School of Medicine, Indianapolis, Indiana

⁵ Center for Computational Biology and Bioinformatics,
Indiana University School of Medicine, Indianapolis, Indiana
xwliu@iupui.edu

Abstract. Although proteomics has rapidly developed in the past decade, researchers are still in the early stage of exploring the world of complex proteoforms, which are protein products with various primary structure alterations resulting from gene mutations, alternative splicing, post-translational modifications, and other biological processes. Proteoform identification is essential to mapping proteoforms to their biological functions as well as discovering novel proteoforms and new protein functions. Top-down mass spectrometry is the method of choice for identifying complex proteoforms because it provides a “bird’s eye view” of intact proteoforms. Fragment ion series in top-down tandem mass spectra provide essential information for identifying primary sequence alterations in proteoforms. Extended proteoform databases and spectral alignment are the two main approaches for proteoform identification. However, due to the combinatorial explosion of various alterations on a protein and the limitations of available spectral alignment algorithms, the proteoform identification problem has still not been fully solved.

We propose a new data structure, called the mass graph, for efficient representation of proteoforms of a protein with variable post-translational modifications and/or terminal truncations. The proteoform identification problem is transformed to the mass graph alignment problem, and dynamic programming algorithms are proposed for a restricted version of the problem. The proposed algorithms were tested on two top-down tandem mass spectrometry data sets. Experimental results showed that the proposed algorithms were efficient in identifying proteoforms with variable post-translational modifications and outperformed MS-Align-E in running time and sensitivity for identifying complex proteoforms, especially those with terminal truncations.

Acknowledgement. The research was supported by the National Institute of General Medical Sciences, National Institutes of Health (NIH) through Grant R01GM118470.

Safely Filling Gaps with Partial Solutions Common to all Solutions

Leena Salmela and Alexandru I. Tomescu

Department of Computer Science, Helsinki Institute for Information
Technology HIIT, University of Helsinki, Helsinki, 00014, Finland
{leena.salmela,tomescu}@cs.helsinki.fi

Abstract. Gap filling has emerged as a natural sub-problem of many *de novo* genome assembly projects (e.g., filling gaps in scaffolds). Several methods have addressed it, but only few have focused on strategies for dealing with multiple gap filling solutions and for guaranteeing reliable results. Such strategies include reporting only unique solutions, or exhaustively enumerating all filling solutions and heuristically creating their consensus.

The gap filling problem is usually formulated as finding an s - t path in the assembly graph, whose length matches the gap length estimate. In this paper we address it with the “safe and complete” framework proposed in [Tomescu and Medvedev, RECOMB 2016] for the contig assembly problem. In terms of gap filling, a *safe solution* is a path of the assembly graph that is a sub-path of all possible s - t paths whose length matches the gap length estimate.

We give an efficient safe algorithm for the gap filling problem, running in time $O(dm)$, where d is the gap length estimate and m is the number of edges of the assembly graph. To transform the safe paths into a single filling sequence usable in downstream analysis, we fill the gap with an arbitrary filling path, in which we mark the safe subsequences. Experiments on the GAGE bacterial datasets show that our method retrieves over 90 % more safe and correct bases as compared to previous methods differentiating between ambiguous and unambiguous positions, with a precision similar to the one of previous methods.

We implemented this method as version 2.0 of our gap filler of scaffolds, Gap2Seq, available at www.cs.helsinki.fi/u/lmsalmel/Gap2Seq/.

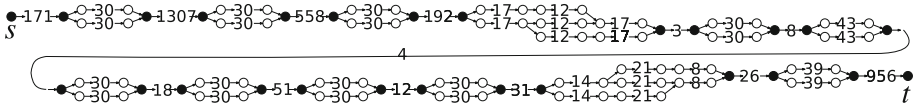


Fig. 1. A de Bruijn graph ($k = 31$) built on *S.aureus* data. We represent unary paths by numbers indicating their length. The estimated gap length is $d = 3774$, and there are 9216 different s - t paths of length d . The safe sub-paths (in black) have length 3337 and the precision of our method on these sub-paths is 99.9 %. Notice that most of the bubbles of this graph are caused by SNPs.

Contents

Optimal Computation of Avoided Words	1
<i>Yannis Almirantis, Panagiotis Charalampopoulos, Jia Gao, Costas S. Iliopoulos, Manal Mohamed, Solon P. Pissis, and Dimitris Polychronopoulos</i>	
A Biclique Approach to Reference Anchored Gene Blocks and Its Applications to Pathogenicity Islands	14
<i>Arnon Benshahar, Vered Chalifa-Caspi, Danny Hermelin, and Michal Ziv-Ukelson</i>	
An Efficient Branch and Cut Algorithm to Find Frequently Mutated Subnetworks in Cancer	27
<i>Anna Bomersbach, Marco Chiarandini, and Fabio Vandin</i>	
Isometric Gene Tree Reconciliation Revisited	40
<i>Broňa Brejová, Askar Gafurov, Dana Pardubská, Michal Sabo, and Tomáš Vinař</i>	
Further Improvement in Approximating the Maximum Duo-Preservation String Mapping Problem	52
<i>Brian Brubach</i>	
SpecTrees: An Efficient Without a Prior Data Structure for MS/MS Spectra Identification	65
<i>Matthieu David, Guillaume Fertin, and Dominique Tessier</i>	
Predicting Core Columns of Protein Multiple Sequence Alignments for Improved Parameter Advising	77
<i>Dan DeBlasio and John Kececioglu</i>	
Fast Compatibility Testing for Phylogenies with Nested Taxa	90
<i>Yun Deng and David Fernández-Baca</i>	
The Gene Family-Free Median of Three	102
<i>Daniel Doerr, Pedro Feijão, Metin Balaban, and Cedric Chauve</i>	
Correction of Weighted Orthology and Paralogy Relations - Complexity and Algorithmic Results.	121
<i>Riccardo Dondi, Nadia El-Mabrouk, and Manuel Lafond</i>	

Copy-Number Evolution Problems: Complexity and Algorithms	137
<i>Mohammed El-Kebir, Benjamin J. Raphael, Ron Shamir, Roded Sharan, Simone Zaccaria, Meirav Zehavi, and Ron Zeira</i>	
<i>Gerbil</i> : A Fast and Memory-Efficient k -mer Counter with GPU-Support	150
<i>Marius Erbert, Steffen Rechner, and Matthias Müller-Hannemann</i>	
Genome Rearrangements on Both Gene Order and Intergenic Regions.	162
<i>Guillaume Fertin, Géraldine Jean, and Eric Tannier</i>	
Better Identification of Repeats in Metagenomic Scaffolding	174
<i>Jay Ghurye and Mihai Pop</i>	
A Better Scoring Model for De Novo Peptide Sequencing: The Symmetric Difference Between Explained and Measured Masses	185
<i>Ludovic Gillet, Simon Rösch, Thomas Tschager, and Peter Widmayer</i>	
<i>StreAM-T_g</i> : Algorithms for Analyzing Coarse Grained RNA Dynamics Based on Markov Models of Connectivity-Graphs.	197
<i>Sven Jager, Benjamin Schiller, Thorsten Strufe, and Kay Hamacher</i>	
Solving Generalized Maximum-Weight Connected Subgraph Problem for Network Enrichment Analysis	210
<i>Alexander A. Loboda, Maxim N. Artyomov, and Alexey A. Sergushichev</i>	
A Natural Encoding of Genetic Variation in a Burrows-Wheeler Transform to Enable Mapping and Genome Inference.	222
<i>Sorina Maciuca, Carlos del Ojo Elias, Gil McVean, and Zamin Iqbal</i>	
Inferring Population Genetic Parameters: Particle Filtering, HMM, Ripley’s K -Function or Runs of Homozygosity?	234
<i>Svend V. Nielsen, Simon Simonsen, and Asger Hobolth</i>	
A Graph Extension of the Positional Burrows-Wheeler Transform and Its Applications	246
<i>Adam M. Novak, Erik Garrison, and Benedict Paten</i>	
Compact Universal k -mer Hitting Sets	257
<i>Yaron Orenstein, David Pellow, Guillaume Marçais, Ron Shamir, and Carl Kingsford</i>	
A New Approximation Algorithm for Unsigned Translocation Sorting	269
<i>Lianrong Pu, Daming Zhu, and Haitao Jiang</i>	
Independent Component Analysis to Remove Batch Effects from Merged Microarray Datasets.	281
<i>Emilie Renard, Samuel Branders, and P.-A. Absil</i>	

A Linear Time Approximation Algorithm for the DCJ Distance
for Genomes with Bounded Number of Duplicates 293
*Diego P. Rubert, Pedro Feijão, Marília D.V. Braga, Jens Stoye,
and Fábio V. Martinez*

A Hybrid Parameter Estimation Algorithm for Beta Mixtures
and Applications to Methylation State Classification 307
Christopher Schröder and Sven Rahmann

Author Index 321