Lecture Notes in Bioinformatics 9838

Subseries of Lecture Notes in Computer Science

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

Springer

Editors
Martin Frith
AIST and University of Tokyo
Tokyo
Japan

Christian Nørgaard Storm Pedersen
Aarhus University
Aarhus
Denmark

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Bioinformatics
ISBN 978-3-319-43680-7 ISBN 978-3-319-43681-4 (eBook)
DOI 10.1007/978-3-319-43681-4

Library of Congress Control Number: 2016945963

LNCS Sublibrary: SL8 – Bioinformatics

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)
Marianne Dammand Iversen
Trine Ji Holmgaard
Katrine Østerlund Rasmussen

Additional Reviewers

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

Donna Henderson
Farhad Hormozdiari
Jeff Howbert
Alex Hu
Laurent Jacob
Mateusz Krzysztof Łącki
Manuel Lafond
Cong Ma
Guillaume Marcais
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 Kou[1], Si Wu[2], Nikola Tolić[3], Yunlong Liu[4,5], Ljiljana Paša-Tolić[3] 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/.

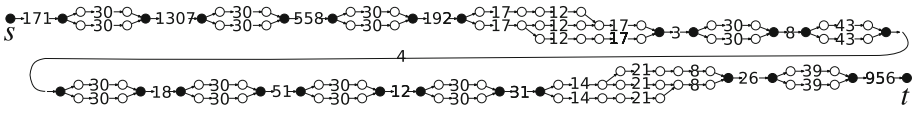

Fig. 1. A de Bruijn graph ($k = 31$) built on *S.aureus* data. We represent unary paths by numbers indicating their length. The estimated gap length is $d = 3774$, and there are 9216 different *s*-*t* paths of length d. The safe sub-paths (in black) have length 3337 and the precision of our method on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

Contents