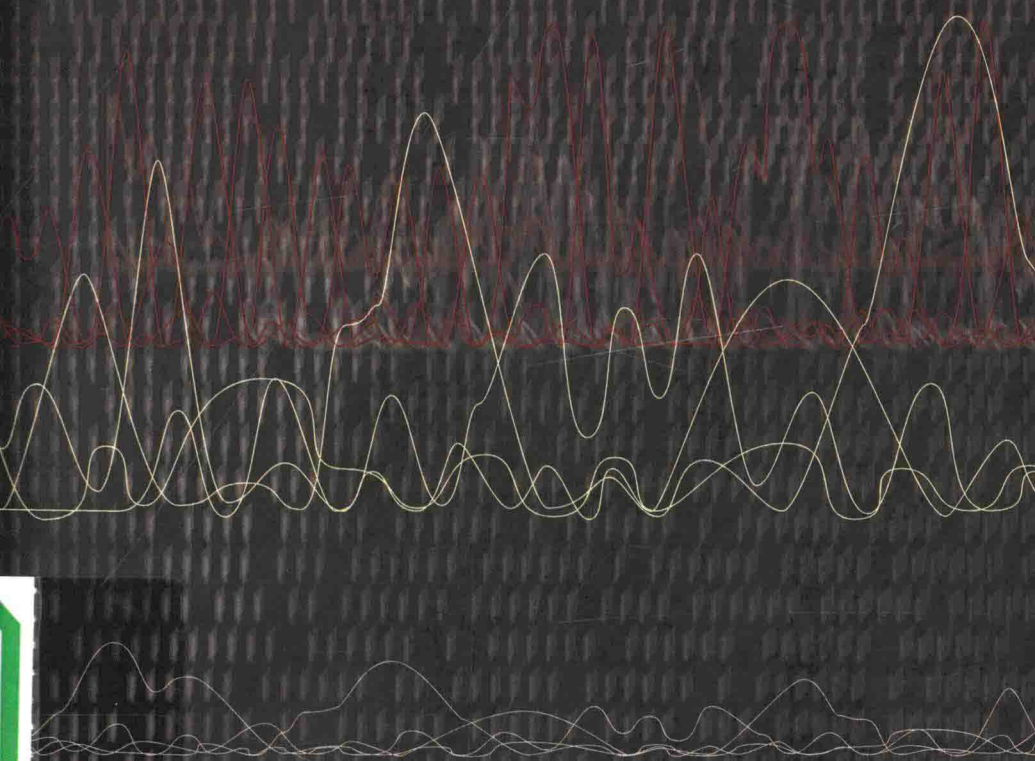
SEQUENCE ALIGNMENT

Methods, Models, Concepts, and Strategies



Edited by Michael S. Rosenberg

Sequence Alignment

Methods, Models, Concepts, and Strategies

Edited by Michael S. Rosenberg

UNIVERSITY OF CALIFORNIA PRESS
Berkeley · Los Angeles · London

University of California Press, one of the most distinguished university presses in the United States, enriches lives around the world by advancing scholarship in the humanities, social sciences, and natural sciences. Its activities are supported by the UC Press Foundation and by philanthropic contributions from individuals and institutions. For more information, visit www.ucpress.edu.

University of California Press
Berkeley and Los Angeles, California

University of California Press, Ltd.
London, England

Library of Congress Cataloging-in-Publication Data

Sequence alignment: methods, models, concepts, and strategies/edited by Michael S. Rosenberg.
p.; cm.
Includes bibliographical references and index.
ISBN 978-0-520-25697-2 (cloth: alk. paper)
1. Bioinformatics. 2. Computational biology.
I. Rosenberg, Michael S., 1972-
[DNLM: 1. Sequence Alignment—methods.
QU 450 S479 2009]
QH324.2.S47 2009
572.80285—dc22

2008029222

Manufactured in the United States

16 15 14 13 12 11 10 09 08
10 9 8 7 6 5 4 3 2 1

The paper used in this publication meets the minimum requirements of ANSI/NISO Z39.48-1992 (R 1997) (*Permanence of Paper*).

Sequence Alignment

Contributors

Roland Fleißner
Center for Integrative Bioinformatics, Vienna, Austria
Roland.Fleissner@campus.lmu.de

Anne Friedrich
Institut de Génétique et de Biologie Moléculaire et Cellulaire, France
friedric@igbmc.fr

Joseph J. Gillespie
University of Maryland, Baltimore County
Virginia Bioinformatics Institute, Virginia Tech
jgille@vbi.vt.edu

Gonzalo Giribet
Harvard University
ggiribet@oeb.harvard.edu

Karl Kjer
Rutgers University
kjer@aesop.rutgers.edu

Liam J. McGuffin
University of Reading, United Kingdom
l.j.mcguffin@reading.ac.uk

Dirk Metzler
Ludwig-Maximilians-Universität, Munich, Germany
metzler@bio.lmu.de

Burkhard Morgenstern
University of Göttingen, Germany
bmorgen@gwdg.de

Luc Moulinier
Institut de Génétique et de Biologie Moléculaire et Cellulaire, France
luc.moulinier@igbmc.fr

Cédric Notredame
Centre for Genomic Regulation, Spain
cedric.notredame@crg.es

T. Heath Ogden
Idaho State University
ogdet@isu.edu

Olivier Poch
Institut de Génétique et de Biologie Moléculaire et Cellulaire, France
poch@igbmc.fr

Benjamin Redelings
North Carolina State University
benjamin_redelings@ncsu.edu

Michael S. Rosenberg
Arizona State University
msr@asu.edu

Usman Roshan
New Jersey Institute of Technology
usman@oak.njit.edu

Marc A. Suchard
University of California, Los Angeles
msuchard@ucla.edu

Julie D. Thompson
Institut de Génétique et de Biologie Moléculaire et Cellulaire, France
julie@igbmc.fr

Ward Wheeler
American Museum of Natural History
wheeler@amnh.org

Preface

Alignment is a vastly underappreciated aspect of comparative genomics and bioinformatics, in part because alignment tools have become so good. As a biologist who occasionally writes software (although, so far, no alignment software), I have become very aware of the tradeoff between ease of use and potential for misuse and abuse. If software is difficult to use, the implemented algorithms may not be widely applied, but those who do apply them will often have a greater understanding of what they are actually doing. When software is easy, many more people will use the algorithms, but the average level of understanding will drop, and the potential for users to use the algorithms in ways which they should not increases. For the most part (there are certainly exceptions), alignment software is easy and fast and has become an almost trivial part of bioinformatics. The danger is the assumption of triviality. Aligned sequences are used as the "raw" input for a wide array of genome studies, and people thus forget that the alignment is itself a hypothesis of homology—a hypothesis that can be wrong. More precisely, each pair of aligned sites is itself a hypothesis. Thus the alignment is actually a set of hypotheses, some of which may be correct and some incorrect—meaning that the alignment as a whole may be neither right nor wrong, but somewhere in the middle.

My own interest in alignment started when I casually tried aligning the upstream regions of a large number of genes from a single species just to see what would happen. While examining the results, I started

to wonder what the expectation would be if the data were completely random (which, given the expected homology of the upstream region of a large set of unrelated genes, is more or less what I had). All of my work on alignment stems from my horror upon discovering the answer (when a pair of completely random DNA sequences, which should have 25% identity due to simple random chance, are aligned using common algorithms and parameters, the resulting aligned sequences are identical at over 40% of the sites) and wondering what the consequence of this and similar issues was on pretty much everything we do in bioinformatics.

This volume had its genesis over an encounter at the 2005 joint meeting of the Society for the Study of Evolution, the Society of Systematic Biologists, and the American Society of Naturalists in Fairbanks, Alaska. Having just finished a presentation about some underappreciated aspects of sequence alignment, I was approached by Chuck Crumly from the University of California Press, who had become interested in finding someone to put together a broad volume on just those sorts of topics. As planned, this volume contains a range of opinions and input from alignment researchers and users in a wide variety of disciplines, including biology, genomics, bioinformatics, computer science, and mathematics. There are two general underlying themes: First, sequence alignments should not be taken for granted; one way or another, they are extremely important for comparative sequence analysis in evolutionary and functional genomics and bioinformatics. Second, the sequence alignment problem is not solved; there are still many challenges and issues that need to be overcome. This book is a dialectic meant to encourage discussion addressing these challenges.

The eleven chapters roughly fall into four (unlabeled) sections: introduction (Chapter 1), biological mechanisms (Chapter 2), algorithms (Chapters 3–5), and broader issues (Chapters 6–11), although in some sense we never escape from either algorithms or broader issues, which are discussed in greater or lesser detail through most of the book.

I begin the book with an introduction to the concepts and history of sequence alignment by describing the dynamic programming approach and the basic algorithms that have been fundamental to the development of sequence alignment software. Emphasis is on the biological concept of homology and the contrast between the biological motivation for aligning data and the computational goals for which algorithms are generally designed.

Liam McGuffin next summarizes the current state of knowledge about the molecular mechanisms leading to indel (insertion and deletion)

mutations and explores the root causes of indel events that allow for better approaches to context-dependent alignment.

There follows Burkhard Morgenstern's in-depth comparison of global and local alignment, including alignment tools that combine both local and global procedures into single algorithms. His discussion of benchmarks highlights methods designed to test the sensitivity of alignment (correctly aligning homologous regions), which should also be evaluated on specificity (not aligning nonhomologous regions). He ends with a discussion of the challenges and tools developed for full genome alignment.

In the next chapter, Cédric Notredame examines the state of the art in multiple sequence alignment by summarizing the major approaches for multiple sequence alignment, including matrix- and consistency-based approaches. He discusses recent methods for combining alternate alignments into a single meta-alignment, and he concludes by examining the importance of using additional data sources (such as structural information) in guiding multiple sequence alignments using template-based approaches.

Dirk Metzler and Roland Fleißner follow with a review of statistical approaches for simultaneously estimating alignments and phylogeny. They focus especially on the modeling of insertion and deletion events in a phylogenetic framework and how advances in maximum-likelihood and Bayesian approaches enable these advanced statistical procedures to be used to align sequences.

In the next contribution, Ward Wheeler and Gonzalo Giribet view alignments as inferential objects rather than data and maintain that alignments should be treated thus in phylogenetic analysis. They then criticize the traditional approach to alignment and phylogenetic analysis (Ogden and Rosenberg 2007a) with a specific implementation of simultaneous phylogeny and alignment construction (De Laet and Wheeler 2003). This highlights the difference between a computational goal (finding the phylogeny by minimizing the number of steps necessary to create an observed set of sequences) and a biological goal (constructing an alignment that best represents the true positional homologies of the sequences or finding the phylogeny that best represents the true evolutionary history of the sequences). The chapters in this book and the two references cited in this paragraph are intended to help readers to draw their own conclusions.

Karl Kjer and colleagues, in Chapter 7, delve into the biological motivations for aligning data and explore the limitations of strict algorithmic

approaches. They emphasize, with simple examples, the importance of including both structural and evolutionary information in postalgorithmic manual curation of alignments. They include a detailed "how-to" explanation for the manual structural alignment of rRNA sequences.

Next, Julie Thompson follows by discussing the current state of benchmark databases for evaluating alignment algorithms. These databases are critical to algorithm development because most new algorithms use such benchmark databases to evaluate performance. There may be a danger of algorithms being overoptimized for the specific characteristics of these databases. She concludes by examining recent benchmark tests for a variety of alignment programs, focusing on recent algorithmic advances that have generally improved alignment quality, but also identifying areas where there is still room for improvement.

Heath Ogden and I team up to examine the increasing role of computer simulation in alignment evaluation, not just for the benchmarking of alignment algorithms but also for exploring the consequences of alignment errors (or use of alternate alignments) in bioinformatic analysis. We describe, compare, and contrast the strengths and weaknesses of a number of approaches for comparing alternate alignments.

In the penultimate chapter, Benjamin Redelings and Marc Suchard take a general look at certainty and uncertainty in sequence alignment, including the root causes for ambiguity, and they explore a variety of approaches for including (or excluding) ambiguously aligned sites in an analysis. They detail how recent advances in Bayesian statistical approaches permit the estimation of alignment uncertainty and how uncertainty can be used in other bioinformatic analyses, including, in particular, phylogeny construction.

In the concluding chapter, Anne Friedrich and colleagues put alignments to use in further evolutionary, structural, functional, and mutational studies of proteins. They also summarize many of the programs and packages currently available.

"WHAT ALIGNMENT PROGRAM SHOULD I USE?"

The question of what software to use is the most common question anyone who works on alignment receives. A definitive answer will not be found in this book. The best alignment program may depend on the specific circumstances of the data being aligned, including the nature of the sequences (e.g., DNA, RNA, protein), the number of sequences to be aligned, the lengths of the sequences, the evolutionary divergence of

the sequences, whether structural information is available, the type of structures (e.g., globular or disordered), and perhaps even the specific purpose for constructing the alignment (i.e., what will the alignment be used for). For large-scale bioinformatic studies, the speed of an algorithm is very important, although I personally feel it can be an overrated factor for smaller comparative studies of the kind encountered in the average molecular lab. If waiting a few extra hours (or even days) will produce a better result, one should take the time to get the best possible answer (in large-scale bioinformatics, where thousands to millions of alignments may be produced as part of a single study, the time difference may scale to months or even years, at which point speed becomes of greater concern).

For a long time, ClustalW/ClustalX (Thompson et al. 1997; Thompson et al. 1994) has been the alignment program of choice for many users because of the general quality of its alignments, its wide implementation, and its ease of use (as of this writing, ClustalW and ClustalX have been cited over 33,000 times [data from ISI Web of Knowledge]). Over the past few years, recent programs such as MUSCLE (Edgar 2004a, b), MAFFT (Katoh et al. 2005), and ProbCons (Do et al. 2005) have consistently received high marks in a variety of benchmark tests (e.g., Blackshields et al. 2006; Pollard et al. 2004; Rosenberg, unpublished data). Given the constant and continued development of alignment programs (see Chapter 1), one would expect that the answer to the question of which is the best program for a specific circumstance may change through time. So, although this book will not provide a definitive answer to the question, it will hopefully help guide your decision as to the factors and issues to consider when thinking about how best to align and interpret your data.

I wish to thank all of the people who contributed to the production of this volume. First and foremost, all of the authors who responded to requests for a contribution and followed through with a manuscript: without you, this all would have come to nothing. As previously mentioned, Chuck Crumly at UC Press deserves much of the credit for getting this volume started; he also deserves thanks for his patience in dealing with the delays and foibles of the editor. Sudhir Kumar offered support and encouragement throughout the creation of this book, including critical commentary on much of my own work. The postdocs and students in my lab deserve thanks for helping with analyses as well as patience and understanding when the needs of the book may have occasionally taken priority over their own: Heath Ogden, Corey Anderson, Ahmet

Kurdoglu, Loretta Goldberg, Meraj Aziz, Virginia Earl-Mirowski, and Amy Harris. During the preparation of this volume I received financial support from the National Library of Medicine of the National Institutes of Health, the National Science Foundation, and the Center for Evolutionary Functional Genomics of the Biodesign Institute and the School of Life Sciences at Arizona State University. Finally, I must thank my wife, Maureen Olmsted, for her continual support, love, and understanding.

Michael S. Rosenberg
January 2008

Contents

CHAPTER 1

Sequence Alignment

Concepts and History

MICHAEL S. ROSENBERG
Arizona State University

Sequence alignment is a fundamental procedure (implicitly or explicitly) conducted in any biological study that compares two or more biological sequences (whether DNA, RNA, or protein). It is the procedure by which one attempts to infer which positions (sites) within sequences are homologous, that is, which sites share a common evolutionary history (see the section "Homology" in this chapter for more detail). For the majority of scientists, alignment is a task whose automated solution was solved years ago; the alignment is of little direct interest but is rather a necessary step that allows one to study deeper questions, such as the identification and quantification of conserved regions or functional motifs (Kirkness et al. 2003; Thomas et al. 2003), profiling of genetic disease (Miller and Kumar 2001; Miller et al. 2003), phylogenetic analysis (Felsenstein 2004), and ancestral sequence profiling

and prediction (Cai et al. 2004; Hall 2006). For other scientists, alignment is an active area of research, where basic questions on how one should construct and evaluate an alignment are under heavy scrutiny and debate. Because alignment is the first step in many complex, high-throughput studies (Lecompte et al. 2001), it is important to remember that alignment algorithms produce a hypothesis of homology (just as a phylogenetic tree is a hypothesis of evolutionary history). Like other hypotheses, these alignments may contain more or less error depending on the nature of the data, some of which may have huge downstream effects on other analyses (Kumar and Filipski 2007; Ogden and Rosenberg 2006; Rosenberg 2005a, b).

In a casual survey, most researchers guess that two or three dozen alignment programs and algorithms have been published. The true number is actually in the hundreds, with the numbers increasing each year. Figure 1.1 shows a summary of the number of named alignment programs released over the 20-year period from 1986 to 2005 (alignment algorithms go back to 1970, but prior to the mid-1980s and the advent of the personal computer, most were simply published as logical descriptions or as source code rather than as compiled executables). Just counting named programs and not papers describing algorithmic

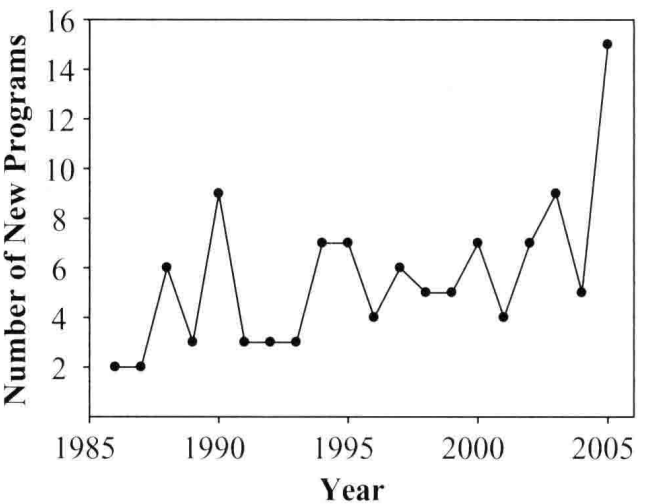


Figure 1.1. Number of new named alignment programs released each year from 1986 to 2005.