Genomes, Browsers, and Databases

The recent explosive growth of biological data has led to a rapid increase in the number of molecular biology databases. These databases are located in many different locations and often use varying interfaces and non-standard data formats. Consequently, integrating and comparing data from them can be difficult and time-consuming. This book provides an overview of the key tools currently available for large-scale comparisons of gene sequences and annotations, focusing on the databases and tools from the University of California, Santa Cruz (UCSC), Ensembl, and the National Center for Biotechnology Information (NCBI). Written specifically for biology and bioinformatics students and researchers, it aims to give an appreciation for the methods by which the browsers and their databases are constructed, enabling readers to determine which tool is the most appropriate for their requirements. Each chapter contains a summary and exercises to aid understanding and promote effective use of these important tools.

Peter Schattner is a research associate in computational biology at the University of California, Santa Cruz. His principal research interests are in the genome-wide identification and characterization of non-protein-coding RNA genes and cis-regulatory mRNA motifs. Dr. Schattner has taught bioinformatics courses at the University of California and California State University and has worked in the research and development of medical ultrasound and magnetic resonance instrumentation at SRI International (Stanford Research Institute) and Diasonics, Inc. He has been a Woodrow Wilson Fellow and was leader of the team that received the 1990 Matzuk Award for technical innovation of the American Institute of Ultrasound in Medicine.

Genomes, Browsers, and Databases

PETER SCHATTNER
University of California, Santa Cruz

Data-Mining Tools for Integrated Genomic Databases

CAMBRIDGE UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi

Cambridge University Press
32 Avenue of the Americas, New York, NY 10013–2473, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521711326

First published 2008

Printed in the United States of America

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication Data

Schattner, Peter, 1948–
Genomes, browsers, and databases : data-mining tools for integrated genomic databases / Peter Schattner.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-0-521-88443-3 (hardback) – ISBN 978-0-521-71132-6 (pbk.)
1. Gene libraries. 2. Genomics – Data processing. 3. Databases. 4. Browsers (Computer programs) I. Title.
[DNLM: 1. Databases, Genetic. 2. Genomics. 3. Gene Library. 4. Molecular Biology. 5. Software. 6. User-Computer Interface. QU 470 S312g 2008]
QH442.4.S33 2008
572.8′602856312 – dc22 2008003901

ISBN 978-0-521-88443-3 hardback
ISBN 978-0-521-71132-6 paperback

To my wife, Sue

Contents

Preface

The idea behind this book developed in late 2004–early 2005 while I was working on two unrelated projects in computational genomics. The first project involved the computational detection of small nucleolar RNAs (snoRNAs) in genome sequences. In the course of this work, I noticed – as others had, as well – that, in mammals, snoRNA genes are located within introns of protein-coding genes (so-called snoRNA host genes), which are often genes that code for ribosomal proteins. This observation led to speculation as to whether there were additional common features of the introns and genes that contain snoRNAs. For example, are the host genes of homologous mammalian snoRNAs themselves homologous? Do those host genes have other shared functions beyond the fact that several of them code for ribosomal proteins? Are the introns that contain the snoRNAs consistently longer (or shorter) than the average introns found in these genes? Are the snoRNAs found at any characteristic distance from the nearest exon-intron junctions in their host gene? To answer these questions would require accessing sequence and annotation data for both the human and mouse genomes and performing some simple calculations and statistics on that data. Moreover, because there were some 200 human snoRNAs already known (and a similar number of mouse snoRNAs), performing this data acquisition and manipulation would require computer processing.

The second project involved regions of the mammalian genome exhibiting "extreme codon conservation," that is, extended regions (typically 150 nt or longer) in which homologous protein-coding genes from several species not only have identical amino acid sequences (which is not unusual), but also use identical codons. Although such extreme codon conservation is unusual, many such regions do exist in mammalian genes. One hypothesis for the existence of these regions is that they not only code for proteins but also contain motifs for post-transcriptional processing, such as RNA binding sites or secondary structures. To assess this hypothesis, we needed to determine whether the conserved regions overlapped known conserved alternative splice sites, whether they were enriched for known exonic splicing enhancers or RNA editing sites, and so forth. Answering these questions again required accessing

sequence and annotation data from numerous genomic regions and performing statistical analyses on these data.

Of course, others were also addressing biological questions that required these kinds of bioinformatic analyses. For example, E. Levanon and colleagues (Levanon et al., 2004) and others discovered RNA editing sites by screening for genomic locations where DNA adenine bases align with mRNA/EST guanines. Similarly, S. Brenner and colleagues detected instances of "nonsense mediated decay" (NMD) via genome-wide screens for mRNAs with "premature stop codons" (Green et al., 2003). What was apparent was that although the biological phenomena in these examples were unrelated – snoRNA host genes, codon conservation in mRNAs, RNA editing sites, and NMD – the types of bioinformatics tasks required for addressing them were strikingly similar: identify a set of genomic regions, obtain the sequence and a set of annotations corresponding to each region, and perform some comparisons or manipulations on the resulting data to enable some inference to be made about the region.

It was also clear that integrated genome databases, such as those used by the Ensembl, MapViewer, and the UCSC Genome Browser, were excellent data sources for obtaining the required sequences and annotations. Although the genome browser interfaces were principally designed to examine one genomic region at a time, their underlying databases could access data from multiple genomic regions simultaneously – that is, in "batch" mode – as required by these types of bioinformatics analyses. Moreover, much of the software required to perform these analyses already existed in the genome database computer code, as this code was needed to create the genome browser displays.

However, I found very few papers in which this approach to genomic analysis was actually adopted. My impression was that this was in part because of limited documentation describing how one could use the genome browser databases for general genomic data mining. In addition, because at this time tools such as Ensembl's BioMart, the UCSC Table Browser, and Galaxy had only recently become available, this sort of strategy was still largely restricted to the biologist with programming skills.

As a result, the molecular biology and bioinformatics research communities generally took advantage of only a fraction of the capabilities provided by genome browsers and databases. Admittedly, there is a considerable learning curve to using these tools. My goal for this book is to ease that learning curve so that more researchers – including those with limited computer skills – are able to use these remarkable tools to address a much wider range of biological questions than they have previously.

This book is intended for graduate or advanced undergraduate students in bioinformatics or biology or for self-study by researchers or students who want to more fully exploit the power of the genome databases. I envision two distinct audiences: biologists with little or no programming experience and bioinformaticists and biologists with programming backgrounds. The first five chapters of the book, as well as Chapter 12, should be accessible to both groups. There are no formal programming prerequisites for these chapters – although, in some places, a general sense of how

computer databases and data files are set up will be helpful. Chapters 6 through 11 do assume some programming background. Nevertheless, non-programmer biologists are encouraged to read these chapters as well. Even if they are unable to follow the programming details, they should get a sense of the types of biological questions that their more computationally oriented collaborators will be able to address with these tools. The biological background assumed is that of an introductory molecular biology course. With this background, the descriptions of the biological examples used should be reasonably self-contained. In addition, the book contains a glossary including both biology and computer science terms that may be unfamiliar to the reader.

I have focused the book on the techniques that I have found most useful in addressing practical biological queries. In particular, the book is primarily focused on tools that integrate data from multiple primary biological data repositories in a standardized and unified manner. As such, the book is not intended to be an "egalitarian" treatment of the three genome browser databases, in the sense of giving equal time to each one. Rather it is one researcher's perspective on useful tools for various bioinformatics tasks. That said, I realize that sometimes my emphasis may be biased by my having more experience with certain tools than with others. In this regard, I should note that (notwithstanding my affiliations with UCSC and the BioPerl project) I have not been involved in the design or development of any of the genome databases or browsers. Similarly, the opinions I express here are solely mine and not those of any of the browser teams.

It is also worth emphasizing that the genome databases and browsers are still rapidly evolving. Consequently, the book is primarily focused on the basic architecture and concepts of the genome databases and tools and where to obtain information on them, rather than on creating a comprehensive catalog of browser features. Such features are continually being added and enhanced, and their online documentation is generally quite good. In addition, the reader should not be surprised if some interface or display found on a browser has changed from the way it is described in the book. Moreover, some features described here will undoubtedly disappear, whereas in other cases statements like "such-and-such database system does not currently support such-and-such capability" will no longer be true. For brevity, I have not always included the word "currently" when describing features that a given browser or database does or does not have, but the reader should realize that I do imply this in every such statement of genome database features.

I am thankful for the help of many people, without whose assistance this book would never have been written. I am particularly indebted to Lincoln Stein for his thorough and insightful review of the entire book. I am also grateful to Deanna Church, Hiram Clawson, Sean Eddy, Xose Fernandez, Jim Kent, and Anton Nekrutenko for reading and commenting on parts of the manuscript. Without their input, there would certainly be far more errors in this book. Of course, any mistakes that remain are solely mine.

I would like to thank Ewan Birney, Mathieu Blanchette, Eli Hatchwell, Fan Hsu, Arek Kasprzyk, Bob Kuhn, Heikki Lehvaslaiho, Todd Lowe, Jason Stajich, Daryl Thomas, Heather Trumbower, David Wheeler, Ann Zweig, and especially Mark Diekhans for numerous discussions on genomes, browsers, and databases, and Katrina Halliday, Alison Evans, Katie Greczylo, and the entire team at Cambridge University Press for their assistance in the preparation of the manuscript. Last, but not least, I want to express my heartfelt gratitude to my wife for her support during this project and for her tolerating my sometimes extreme mood swings and work habits during the course of preparing the manuscript.