BIOINFORMATICS

FOURTH EDITION

EDITED BY

ANDREAS D. BAXEVANIS

GARY D. BADER

DAVID S. WISHART

with website

WILEY

Bioinformatics

Bioinformatics

Edited by

Andreas D. Baxevanis, Gary D. Bader, and David S. Wishart

Fourth Edition

WILEY

This fourth edition first published 2020
© 2020 John Wiley & Sons, Inc.

Edition History
Wiley-Blackwell (1e, 2000), Wiley-Blackwell (2e, 2001), Wiley-Blackwell (3e, 2005)

The right of Andreas D. Baxevanis, Gary D. Bader, and David S. Wishart to be identified as the authors of the editorial material in this work has been asserted in accordance with law.

Registered Office
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty
While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data

Names: Baxevanis, Andreas D., editor. | Bader, Gary D., editor. | Wishart,
David S., editor.
Title: Bioinformatics / edited by Andreas D. Baxevanis, Gary D. Bader, David S.
Wishart.
Other titles: Bioinformatics (Baxevanis)
Description: Fourth edition. | Hoboken, NJ : Wiley, 2020. | Includes
bibliographical references and index.
Identifiers: LCCN 2019030489 (print) | ISBN 9781119335580 (cloth) | ISBN
9781119335962 (adobe pdf) | ISBN 9781119335955 (epub)
Subjects: MESH: Computational Biology–methods | Sequence Analysis–methods
| Base Sequence | Databases, Nucleic Acid | Databases, Protein
Classification: LCC QH324.2 (print) | LCC QH324.2 (ebook) | NLM QU
550.5.S4 | DDC 570.285–dc23
LC record available at https://lccn.loc.gov/2019030489
LC ebook record available at https://lccn.loc.gov/2019030490

Cover Design: Wiley
Cover Images: © David Wishart, background © Suebsiri/Getty Images

Set in 9.5/12.5pt STIXTwoText by SPi Global, Chennai, India

10 9 8 7 6 5 4 3 2 1

Contents

Foreword

As I review the material presented in the fourth edition of *Bioinformatics* I am moved in two ways, related to both the past and the future.

Looking to the past, I am moved by the amazing evolution that has occurred in our field since the first edition of this book appeared in 1998. Twenty-one years is a long, long time in any scientific field, but especially so in the agile field of bioinformatics. To use the well-trodden metaphor of the "biology moonshot," the launchpad at the beginning of the twenty-first century was the determination of the human genome. Discovery is not the right word for what transpired – we knew it was there and what was needed. Synergy is perhaps a better word; synergy of technological development, experiment, computation, and policy. A truly collaborative effort to continuously share, in a reusable way, the collective efforts of many scientists. Bioinformatics was born from this synergy and has continued to grow and flourish based on these principles.

That growth is reflected in both the scope and depth of what is covered in these pages. These attributes are a reflection of the increased complexity of the biological systems that we study (moving from "simple" model organisms to the human condition) and the scales at which those studies take place. As a community we have professed multiscale modeling without much to show for it, but it would seem to be finally here. We now have the ability to connect the dots from molecular interactions, through the pathways to which those molecules belong to the cells they affect, to the interactions between those cells through to the effects they have on individuals within a population. Tools and methodologies that were novel in earlier editions of this book are now routine or obsolete, and newer, faster, and more accurate procedures are now with us. This will continue, and as such this book provides a valuable snapshot of the scope and depth of the field as it exists today.

Looking to the future, this book provides a foundation for what is to come. For me this is a field more aptly referred to (and perhaps a new subtitle for the next edition) as Biomedical Data Science. Sitting as I do now, as Dean of a School of Data Science which collaborates openly across all disciplines, I see rapid change akin to what happened to birth bioinformatics 20 or more years ago. It will not take 20 years for other disciplines to catch up; I predict it will take 2! The accomplishments outlined in this book can help define what other disciplines will accomplish with their own data in the years to come. Statistical methods, cloud computing, data analytics, notably deep learning, the management of large data, visualization, ethics policy, and the law surrounding data are generic. Bioinformatics has so much to offer, yet it will also be influenced by other fields in a way that has not happened before. Forty-five years in academia tells me that there is nothing to compare across campuses to what is happening today. This is both an opportunity and a threat. The editors and authors of this edition should be complimented for setting the stage for what is to come.

Philip E. Bourne, University of Virginia

Preface

In putting together this textbook, we hope that students from a range of fields – including biology, computer science, engineering, physics, mathematics, and statistics – benefit by having a convenient starting point for learning most of the core concepts and many useful practical skills in the field of bioinformatics, also known as computational biology.

Students interested in bioinformatics often ask about how should they acquire training in such an interdisciplinary field as this one. In an ideal world, students would become experts in all the fields mentioned above, but this is actually not necessary and realistically too much to ask. All that is required is to combine their scientific interests with a foundation in biology and any single quantitative field of their choosing. While the most common combination is to mix biology with computer science, incredible discoveries have been made through finding creative intersections with any number of quantitative fields. Indeed, many of these quantitative fields typically overlap a great deal, especially given their foundational use of mathematics and computer programming. These natural relationships between fields provide the foundation for integrating diverse expertise and insights, especially when in the context of performing bioinformatic analyses.

While bioinformatics is often considered an independent subfield of biology, it is likely that the next generation of biologists will not consider bioinformatics as being separate and will instead consider gaining bioinformatics and data science skills as naturally as they learn how to use a pipette. They will learn how to program a computer, likely starting in elementary school. Other data science knowledge areas, such as math, statistics, machine learning, data processing, and data visualization will also be part of any core curriculum. Indeed, the children of one of the editors recently learned how to construct bar plots and other data charts in kindergarten! The same editor is teaching programming in R (an important data science programming language) to all incoming biology graduate students at his university starting this year.

As bioinformatics and data science become more naturally integrated in biology, it is worth noting that these fields actively espouse a culture of open science. This culture is motivated by thinking about why we do science in the first place. We may be curious or like problem solving. We could also be motivated by the benefits to humanity that scientific advances bring, such as tangible health and economic benefits. Whatever the motivating factor, it is clear that the most efficient way to solve hard problems is to work together as a team, in a complementary fashion and without duplication of effort. The only way to make sure this works effectively is to efficiently share knowledge and coordinate work across disciplines and research groups. Presenting scientific results in a reproducible way, such as freely sharing the code and data underlying the results, is also critical. Fortunately, there are an increasing number of resources that can help facilitate these goals, including the bioRxiv preprint server, where papers can be shared before the very long process of peer review is completed; GitHub, for sharing computer code; and data science notebook technology that helps combine code, figures, and text in a way that makes it easier to share reproducible and reusable results.

We hope this textbook helps catalyze this transition of biology to a quantitative, data science-intensive field. As biological research advances become ever more built on interdisciplinary, open, and team science, progress will dramatically speed up, laying the groundwork for fantastic new discoveries in the future.

We also deeply thank all of the chapter authors for contributing their knowledge and time to help the many future readers of this book learn how to apply the myriad bioinformatic techniques covered within these pages to their own research questions.

Andreas D. Baxevanis
Gary D. Bader
David S. Wishart

Contributors

Gary D. Bader, PhD is a Professor at The Donnelly Centre at the University of Toronto, Toronto, Canada, and a leader in the field of Network Biology. Gary completed his postdoctoral work in Chris Sander's group in the Computational Biology Center (cBio) at Memorial Sloan-Kettering Cancer Center in New York. Gary completed his PhD in the laboratory of Christopher Hogue in the Department of Biochemistry at the University of Toronto and a BSc in Biochemistry at McGill University in Montreal. Dr. Bader uses molecular interaction, pathway, and -omics data to gain a "causal" mechanistic understanding of normal and disease phenotypes. His laboratory develops novel computational approaches that combine molecular interaction and pathway information with -omics data to develop clinically predictive models and identify therapeutically targetable pathways. He also helps lead the Cytoscape, GeneMANIA, and Pathway Commons pathway and network analysis projects.

Geoffrey J. Barton, PhD is Professor of Bioinformatics and Head of the Division of Computational Biology at the University of Dundee School of Life Sciences, Dundee, UK. Before moving to Dundee in 2001, he was Head of the Protein Data Bank in Europe and the leader of the Research and Development Team at the EMBL European Bioinformatics Institute (EBI). Prior to joining EMBL-EBI, he was Head of Genome Informatics at the Wellcome Trust Centre for Human Genetics, University of Oxford, a position he held concurrently with a Royal Society University Research Fellowship in the Department of Biochemistry. Geoff's longest running research interest is using computational methods to study the relationship between a protein's sequence, its structure, and its function. His group has contributed many tools and techniques in the field of protein sequence and structure analysis and structure prediction. Two of the best known are the Jalview multiple alignment visualization and analysis workbench, which is in use by over 70 000 groups for research and teaching, and the JPred multi-neural net protein secondary structure prediction algorithm, which performs predictions on up to 500 000 proteins/month for users worldwide. In addition to his work related to protein sequence and structure, Geoff has collaborated on many projects that probe biological processes using proteomic and high-throughput sequencing approaches. Geoff's group has deep expertise in RNA-seq methods and has recently published a two-condition 48-replicate RNA-seq study that is now a key reference work for users of this technology.

Andreas D. Baxevanis, PhD is the Director of Computational Biology for the National Institutes of Health's (NIH) Intramural Research Program. He is also a Senior Scientist leading the Computational Genomics Unit at the NIH's National Human Genome Research Institute, Bethesda, MD, USA. His research program is centered on probing the interface between genomics and developmental biology, focusing on the sequencing and analysis of invertebrate genomes that can yield insights of relevance to human health, particularly in the areas of regeneration, allorecognition, and stem cell biology. His accomplishments have been recognized by the Bodossaki Foundation's Academic Prize in Medicine and Biology in 2000, Greece's highest award for young scientists of Greek heritage. In 2014, he was elected to the Johns Hopkins Society of Scholars, recognizing alumni who have achieved marked distinction in their field of study. He was the recipient of the NIH's Ruth L. Kirschstein Mentoring Award in 2015, in recognition of his commitment to scientific training, education, and mentoring. In 2016, Dr. Baxevanis was elected as a Senior Member of the International Society for Computational Biology for his sustained contributions to the field and, in 2018, he was elected as a Fellow of the American Association for the Advancement of Science for his distinguished contributions to the field of comparative genomics.

Robert G. Beiko, PhD is a Professor and Associate Dean for Research in the Faculty of Computer Science at Dalhousie University, Halifax, Nova Scotia, Canada. He is a former Tier II Canada Research Chair in Bioinformatics (2007–2017), an Associate Editor at mSystems and BMC Bioinformatics, and a founding organizer of the Canadian Bioinformatics Workshops in Metagenomics and Genomic Epidemiology. He is also the lead editor of the recently published book *Microbiome Analysis* in the Methods in Molecular Biology series. His research focuses on microbial genomics, evolution, and ecology, with concentrations in the area of lateral gene transfer and microbial community analysis.

Fiona S.L. Brinkman, PhD, FRSC is a Professor in Bioinformatics and Genomics in the Department of Molecular Biology and Biochemistry at Simon Fraser University, Vancouver, British Columbia, Canada, with cross-appointments in Computing Science and the Faculty of Health Sciences. She is most known for her research and development of widely used computer software that aids both microbe (PSORTb, IslandViewer) and human genomic (InnateDB) evolutionary/genomics analyses, along with her insights into pathogen evolution. She is currently co-leading a national effort – the Integrated Rapid Infectious Disease Analysis Project – the goal of which is to use microbial genomes as a fingerprint to better track and understand the spread and evolution of infectious diseases. She has also been leading development into an approach to integrate very diverse data for the Canadian CHILD Study birth cohort, including microbiome, genomic, epigenetic, environmental, and social data. She coordinates community-based genome annotation and database development for resources such as the Pseudomonas Genome Database. She also has a strong interest in bioinformatics education, including developing the first undergraduate curricula used as the basis for the first White Paper on Canadian Bioinformatics Training in 2002. She is on several committees and advisory boards, including the Board of Directors for Genome Canada; she chairs the Scientific Advisory Board for the European Nucleotide Archive (EMBL-EBI). She has received a number of awards, including a TR100 award from MIT, and, most recently, was named as a Fellow of the Royal Society of Canada.

Andrew Emili, PhD is a Professor in the Departments of Biochemistry (Medical School) and Biology (Arts and Sciences) at Boston University (BU), Boston, MA, USA, and the inaugural Director of the BU Center for Network Systems Biology (CNSB). Prior to Boston, Dr. Emili was a founding member and Principal Investigator for 18 years at the Donnelly Center for Cellular and Biomolecular Research at the University of Toronto, one of the premier research centers in integrative molecular biology. Dr. Emili is an internationally recognized leader in functional proteomics, systems biology, and precision mass spectrometry. His group develops and applies innovative technologies to systematically map protein interaction networks and macromolecular complexes of cells and tissues on a global scale, publishing "interactome" maps of unprecedented quality, scope, and resolution.

Tatyana Goldberg, PhD is a postdoctoral scientist at the Technical University of Munich, Germany. She obtained her PhD in Bioinformatics under the supervision of Dr. Burkhard Rost. Her research focuses on developing models that can predict the localization of proteins within cells. The results of her study contribute to a variety of applications, including the development of pharmaceuticals for the treatment of Alzheimer disease and cancer.

Emma J. Griffiths, PhD is a research associate in the Department of Pathology and Laboratory Medicine at the University of British Columbia in Vancouver, Canada, working with Dr. William Hsiao. Dr. Griffiths received her PhD from the Department of Biochemistry and Biomedical Sciences at McMaster University in Hamilton, Canada, with her doctoral work focusing on the evolutionary relationships between different groups of bacteria. She has since pursued postdoctoral training in the fields of chemical and fungal genetics and microbial genomics with Dr. Fiona Brinkman in the Department of Biochemistry and Molecular Biology at Simon Fraser University in Vancouver, Canada. Her current work focuses on the development of ontology-driven applications designed to improve pathogen genomics contextual data ("metadata") exchange during public health investigations.

Desmond G. Higgins, PhD is Professor of Bioinformatics in University College Dublin, Ireland, where his laboratory works on genomic data analysis and sequence alignment algorithms. He earned his doctoral degree in zoology from Trinity College Dublin, Ireland, and has worked in the field of bioinformatics since 1985. His group maintains and develops the Clustal package for multiple sequence alignment in collaboration with groups in France, Germany, and the United Kingdom. Dr. Higgins wrote the first version of Clustal in Dublin in 1988. He then moved to the EMBL Data Library group located in Heidelberg in 1990 and later to EMBL-EBI in Hinxton. This coincided with the release of ClustalW and, later, ClustalX, which has been extremely widely used and cited. Currently, he has run out of version letters so is working on Clustal Omega, specifically designed for making extremely large protein alignments.

Lynn B. Jorde, PhD has been on the faculty of the University of Utah School of Medicine, Salt Lake City, UT, USA, since 1979 and holds the Mark and Kathie Miller Presidential Endowed Chair in Human Genetics. He was appointed Chair of the Department of Human Genetics in September 2009. Dr. Jorde's laboratory has published scientific articles on human genetic variation, high-altitude adaptation, the genetic basis of human limb malformations, and the genetics of common diseases such as hypertension, juvenile idiopathic arthritis, and inflammatory bowel disease. Dr. Jorde is the lead author of *Medical Genetics*, a textbook that is now in its fifth edition and translated into multiple foreign languages. He is the co-recipient of the 2008 Award for Excellence in Education from the American Society of Human Genetics (ASHG). He served two 3-year terms on the Board of Directors of ASHG and, in 2011, he was elected as president of ASHG. In 2012, he was elected as a Fellow of the American Association for the Advancement of Science.

Marieke L. Kuijjer, PhD is a Group Leader at the Centre for Molecular Medicine Norway (NCMM, a Nordic EMBL partner), University of Oslo, Norway, where she runs the Computational Biology and Systems Medicine group. She obtained her doctorate in the laboratory of Dr. Pancras Hogendoorn in the Department of Pathology at the Leiden University Medical Center in the Netherlands. After this, she continued her scientific training as a postdoctoral researcher in the laboratory of Dr. John Quackenbush at the Dana-Farber Cancer Institute and Harvard T.H. Chan School of Public Health, during which she won a career development award and a postdoctoral fellowship. Dr. Kuijjer's research focuses on solving fundamental biological questions through the development of new methods in computational and systems biology and on implementing these techniques to better understand gene regulation in cancer. Dr. Kuijjer serves on the editorial board of *Cancer Research*.

David H. Mathews, MD, PhD is a professor of Biochemistry and Biophysics and also of Biostatistics and Computational Biology at the University of Rochester Medical Center, Rochester, NY, USA. He also serves as the Associate Director of the University of Rochester's Center for RNA Biology. His involvement in education includes directing the Biophysics PhD program and teaching a course in Python programming and algorithms for doctoral students without a programming background. His group studies RNA biology and develops methods for RNA secondary structure prediction and molecular modeling of three-dimensional structure. His group developed and maintains RNAstructure, a widely used software package for RNA structure prediction and analysis.

Sean D. Mooney, PhD has spent his career as a researcher and group leader in biomedical informatics. He now leads Research IT for UW Medicine and is leading efforts to support and build clinical research informatic platforms as its first Chief Research Information Officer (CRIO) and as a Professor in the Department of Biomedical Informatics and Medical Education at the University of Washington, Seattle, WA, USA. Previous to being appointed as CRIO, he was an Associate Professor and Director of Bioinformatics at the Buck Institute for Research on Aging. As an Assistant Professor, he was appointed in Medical and Molecular Genetics at Indiana University School of Medicine and was the founding Director of the Indiana University School of Medicine Bioinformatics Core. In 1997, he received his BS with Distinction in Biochemistry and Molecular Biology from the University of Wisconsin at Madison. He received his PhD from the University of California in San Francisco in 2001, then pursued his postdoctoral studies under an American Cancer Society John Peter Hoffman Fellowship at Stanford University.

Stephen J. Mooney, PhD is an Acting Assistant Professor in the Department of Epidemiology at the University of Washington, Seattle, WA, USA. He developed the CANVAS system for collecting data from Google Street View imagery as a graduate student, and his research focuses on contextual influences on physical activity and transport-related injury. He’s a methods geek at heart.

Hunter N.B. Moseley, PhD is an Associate Professor in the Department of Molecular and Cellular Biochemistry at the University of Kentucky, Lexington, KY, USA. He is also the Informatics Core Director within the Resource Center for Stable Isotope Resolved Metabolomics, Associate Director for the Institute for Biomedical Informatics, and a member of the Markey Cancer Center. His research interests include developing computational methods, tools, and models for analyzing and interpreting many types of biological and biophysical data that enable new understanding of biological systems and related disease processes. His formal education spans multiple disciplines including chemistry, mathematics, computer science, and biochemistry, with expertise in algorithm development, mathematical modeling, structural bioinformatics, and systems biochemistry, particularly in the development of automated analyses of nuclear magnetic resonance and mass spectrometry data as well as knowledge–data integration.

Yanay Ofran, PhD is a Professor and head of the Laboratory of Functional Genomics and Systems Biology at Bar Ilan University in Tel Aviv, Israel. His research focuses on biomolecular recognition and its role in health and disease. Professor Ofran is also the founder of Biolojic Design, a biopharmaceutical company that uses artificial intelligence approaches to design epitope-specific antibodies. He is also the co-founder of Ukko, a biotechnology company that uses computational tools to design safe proteins for the food and agriculture sectors.

Joseph N. Paulson, PhD is a Statistical Scientist within Genentech’s Department of Biostatistics, San Francisco, CA, USA, working on designing clinical trials and biomarker discovery. Previously, he was a Research Fellow in the Department of Biostatistics and Computational Biology at the Dana-Farber Cancer Institute and Department of Biostatistics at the Harvard T.H. Chan School of Public Health. He graduated with a PhD in Applied Mathematics, Statistics, and Scientific Computation from the University of Maryland, College Park where he was a National Science Foundation Graduate Fellow. As a statistician and computational biologist, his interests include clinical trial design, biomarker discovery, development of computational methods for the analysis of high-throughput sequencing data while accounting for technical artifacts, and the microbiome.

Sadhna Phanse, MSc is a Bioinformatics Analyst at the Donnelly Centre for Cellular and Biomolecular Research at the University of Toronto, Toronto, Canada. She has been active in the field of proteomics since 2006 as a member of the Emili research group. Her current work involves the use of bioinformatics methods to investigate biological systems and molecular association networks in human cells and model organisms.

John Quackenbush, PhD is Professor of Computational Biology and Bioinformatics and Chair of the Department of Biostatistics at the Harvard T.H. Chan School of Public Health, Boston, MA, USA. He also holds appointments in the Channing Division of Network Medicine of Brigham and Women's Hospital and at the Dana-Farber Cancer Institute. He is a recognized expert in computational and systems biology and its applications to the study of a wide range of human diseases and the factors that drive those diseases and their responses to therapy. Dr. Quackenbush has long been an advocate for open science and reproducible research. As a founding member and past president of the Functional Genomics Data Society (FGED), he was a developer of the Minimal Information About a Microarray Experiment (MIAME) and other data-reporting standards. Dr. Quackenbush was honored by President Barack Obama in 2013 as a White House Open Science Champion of Change.

Jonas Reeb, MSc is a PhD student in the laboratory of Burkhard Rost at the Technical University of Munich, Germany (TUM). During his studies at TUM, he has worked on predictive methods for the analysis and evaluation of transmembrane proteins; he has also worked on the NYCOMPS structural genomics pipeline. His doctoral thesis focuses on the effect of sequence variants and their prediction.

Burkhard Rost, PhD is a professor and Alexander von Humboldt Award recipient at the Technical University of Munich, Germany (TUM). He was the first to combine machine learning with evolutionary information, using this combination to accurately predict secondary structure. Since that time, his group has repeated this success in developing many other tools that are actively used to predict and understand aspects of protein structure and function. All tools developed by his research group are available through the first internet server in the field of protein structure prediction (PredictProtein), a resource that has been online for over 25 years. Over the last several years, his research group has been shifting its focus to the development of methods that predict and annotate the effect of sequence variation and their implications for precision medicine and personalized health.

Fabian Sievers, PhD is currently a postdoctoral research fellow in the laboratory of Des Higgins at University College Dublin, Ireland. He works on multiple sequence alignment algorithms and, in particular, on the development of Clustal Omega. He received his PhD in mathematics from Trinity College, Dublin and has worked in industry in the fields of algorithm development and high-performance computing.

Michael F. Sloma, PhD is a data scientist at Xometry, Gaithersburg, MD, USA. He received his BA degree in Chemistry from Wells College. He earned his doctoral degree in Biochemistry in the laboratory of David Mathews at the University of Rochester, where his research focused on computational methods to predict RNA structure from sequence.

W. Scott Watkins, MS is a researcher and laboratory manager in the Department of Human Genetics at the University of Utah, Salt Lake City, UT, USA. He has a long-standing interest in human population genetics and evolution. His current interests include the development and application of high-throughput computational methods to mobile element biology, congenital heart disease, and personalized medicine.

David S. Wishart, PhD is a Distinguished University Professor in the Departments of Biological Sciences and Computing Science at the University of Alberta, Edmonton, Alberta, Canada. Dr. Wishart has been developing bioinformatics programs and databases since the early 1980s and has made bioinformatics an integral part of his research program for nearly four decades. His interest in bioinformatics led to the development of a number of widely used bioinformatics tools for structural biology, bacterial genomics, pharmaceutical research, and metabolomics. Some of Dr. Wishart's most widely known bioinformatics contributions include the Chemical Shift Index (CSI) for protein secondary structure identification by nuclear magnetic resonance spectroscopy, PHAST for bacterial genome annotation, the DrugBank database for drug research, and MetaboAnalyst for metabolomic data analysis. Over the course of his academic career, Dr. Wishart has published more than 400 research papers, with many being in the field of bioinformatics. In addition to his long-standing interest in bioinformatics research, Dr. Wishart has been a passionate advocate for bioinformatics education and outreach. He is one of the founding members of the Canadian Bioinformatics Workshops (CBW) – a national bioinformatics training program that has taught more than 3000 students over the past two decades. In 2002 he established Canada's first undergraduate bioinformatics degree program at the University of Alberta and has personally mentored nearly 130 undergraduate and graduate students, many of whom have gone on to establish successful careers in bioinformatics.

Tyra G. Wolfsberg, PhD is the Associate Director of the Bioinformatics and Scientific Programming Core at the National Human Genome Research Institute (NHGRI), National Institutes of Health (NIH), Bethesda, MD, USA. Her research program focuses on developing methodologies to integrate sequence, annotation, and experimentally generated data so that bench biologists can quickly and easily obtain results for their large-scale experiments. She maintains a long-standing commitment to bioinformatics education and outreach. *S*he has authored a chapter on genomic databases for previous editions of this textbook, as well as a chapter on the NCBI MapViewer for *Current Protocols in Bioinformatics* and *Current Protocols in Human Genetics*. She serves as the co-chair of the NIH lecture series Current Topics in Genome Analysis; these lectures are archived online and have been viewed over 1 million times to date. In addition to teaching bioinformatics courses at NHGRI, she served for 13 years as a faculty member in bioinformatics at the annual AACR Workshop on Molecular Biology in Clinical Oncology.

Michael Zuker, PhD retired as a Professor of Mathematical Sciences at Rensselaer Polytechnic Institute, Troy, NY, USA, in 2016. He was an Adjunct Professor in the RNA Institute at the University of Albany and remains affiliated with the RNA Institute. He works on the development of algorithms to predict folding, hybridization, and melting profiles in nucleic acids. His nucleic acid folding and hybridization web servers have been running at the University of Albany since 2010. His educational activities include developing and teaching his own bioinformatics course at Rensselaer and participating in both a Chautauqua short course in bioinformatics for college teachers and an intensive bioinformatics course at the University of Michigan. He currently serves on the Scientific Advisory Board of Expansion Therapeutics, Inc. at the Scripps Research Institute in Jupiter, Florida.

About the Companion Website

This book is accompanied by a companion website:

www.wiley.com/go/baxevanis/Bioinformatics_4e

The website includes:

- Test Samples
- Word Samples

Scan this QR code to visit the companion website.

1

Biological Sequence Databases

Andreas D. Baxevanis

Introduction

Over the past several decades, there has been a feverish push to understand, at the most elementary of levels, what constitutes the basic "book of life." Biologists (and scientists in general) are driven to understand how the millions or billions of bases in an organism's genome contain all of the information needed for the cell to conduct the myriad metabolic processes necessary for the organism's survival – information that is propagated from generation to generation. To have a basic understanding of how the collection of individual nucleotide bases drives the engine of life, large amounts of sequence data must be collected and stored in a way that these data can be searched and analyzed easily. To this end, much effort has gone into the design and maintenance of biological sequence databases. These databases have had a significant impact on the advancement of our understanding of biology not just from a computational standpoint but also through their integrated use alongside studies being performed at the bench.

The history of sequence databases began in the early 1960s, when Margaret Dayhoff and colleagues (1965) at the National Biomedical Research Foundation (NBRF) collected all of the protein sequences known at that time – all 65 of them – and published them in a book called the *Atlas of Protein Sequence and Structure*. It is important to remember that, at this point in the history of biology, the focus was on sequencing proteins through traditional techniques such as the Edman degradation rather than on sequencing DNA, hence the overall small number of available sequences. By the late 1970s, when a significant number of nucleotide sequences became available, those were also included in later editions of the *Atlas*. As this collection evolved, it included text-based descriptions to accompany the protein sequences, as well as information regarding the evolution of many protein families. This work, in essence, was the first annotated sequence database, even though it was in printed form. Over time, the amount of data contained in the *Atlas* became unwieldy and the need for it to be available in electronic form became obvious. From the early 1970s to the late 1980s, the contents of the *Atlas* were distributed electronically by NBRF (and later by the Protein Information Resource, or PIR) on magnetic tape, and the distribution included some basic programs that could be used to search and evaluate distant evolutionary relationships.

The next phase in the history of sequence databases was precipitated by the veritable explosion in the amount of nucleotide sequence data available to researchers by the end of the 1970s. To address the need for more robust public sequence databases, the Los Alamos National Laboratory (LANL) created the Los Alamos DNA Sequence Database in 1979, which became known as GenBank in 1982 (Benson et al. 2018). Meanwhile, the European Molecular Biology Laboratory (EMBL) created the EMBL Nucleotide Sequence Data Library in 1980. Throughout the 1980s, EMBL (then based in Heidelberg, Germany), LANL, and (later) the National Center for Biotechnology Information (NCBI, part of the National Library of Medicine at the National Institutes of Health) jointly contributed DNA sequence data to these databases. This was done

Bioinformatics, Fourth Edition. Edited by Andreas D. Baxevanis, Gary D. Bader, and David S. Wishart.

Companion Website: www.wiley.com/go/baxevanis/Bioinformatics_4e

by having teams of curators manually transcribing and interpreting what was published in print journals to an electronic format more appropriate for computational analyses. The DNA Databank of Japan (DDBJ; Kodama et al. 2018) joined this DNA data-collecting collaboration a few years later. By the late 1980s, the quantity of DNA sequence data being produced was so overwhelming that print journals began asking scientists to electronically submit their DNA sequences directly to these databases, rather than publishing them in printed journals or papers. In 1988, after a meeting of these three groups (now referred to as the International Nucleotide Sequence Database Collaboration, or INSDC; Karsch-Mizrachi et al. 2018), there was an agreement to use a common data exchange format and to have each database update only the records that were directly submitted to it. Thanks to this agreement, all three centers (EMBL, DDBJ, and NCBI) now collect direct DNA sequence submissions and distribute them so that each center has copies of all of the sequences, with each center acting as a primary distribution center for these sequences. DDBJ/EMBL/GenBank records are updated automatically every 24 hours at all three sites, meaning that all sequences can be found within DDBJ, the European Nucleotide Archive (ENA; Silvester et al. 2018), and GenBank in short order. That said, each database within the INSDC has the freedom to display and annotate the sequence data as it sees fit.

In parallel with the early work being done on DNA sequence databases, the foundations for the Swiss-Prot protein sequence database were also being laid in the early 1980s by Amos Bairoch, recounting its history from an engaging perspective in a first-person review (Bairoch 2000). Bairoch converted PIR's *Atlas* to a format similar to that used by EMBL for its nucleotide database. In this initial release, called PIR+, additional information about each of the proteins was added, increasing its value as a curated, well-annotated source of information on proteins. In the summer of 1986, Bairoch began distributing PIR+ on the US BIONET (a precursor to the Internet), renaming it Swiss-Prot. At that time, it contained the grand sum of 3900 protein sequences. This was seen as an overwhelming amount of data, in stark contrast to today's standards. As Swiss-Prot and EMBL followed similar formats, a natural collaboration developed between these two groups, and these collaborative efforts strengthened when both EMBL's and Swiss-Prot's operations were moved to EMBL's European Bioinformatics Institute (EBI; Cook et al. 2018) in Hinxton, UK. One of the first collaborative projects undertaken by the Swiss-Prot and EMBL teams was to create a new and much larger protein sequence database supplement to Swiss-Prot. As maintaining the high quality of Swiss-Prot entries was a time-consuming process involving extensive sequence analysis and detailed curation by expert annotators (Apweiler 2001), and to allow the quick release of protein data not yet annotated to Swiss-Prot's stringent standards, a new database called TrEMBL (for "translation of EMBL nucleotide sequences") was created. This supplement to Swiss-Prot initially consisted of computationally annotated sequence entries derived from the translation of all coding sequences (CDSs) found in INSDC databases. In 2002, a new effort involving the Swiss Institute of Bioinformatics, EMBL-EBI, and PIR was launched, called the UniProt consortium (UniProt Consortium 2017). This effort gave rise to the UniProt Knowledgebase (UniProtKB), consisting of Swiss-Prot, TrEMBL, and PIR. A similar effort also gave rise to the NCBI Protein Database, bringing together data from numerous sources and described more fully in the text that follows.

The completion of human genome sequencing and the sequencing of numerous model genomes, as well as the existence of a gargantuan number of sequences in general, provides a golden opportunity for biological scientists, owing to the inherent value of these data. At the same time, the sheer magnitude of data also presents a conundrum to the inexperienced user, resulting not just from the size of the "sequence information space" but from the fact that the information space continues to get larger by leaps and bounds. Indeed, the sequencing landscape has changed significantly in recent years with the development of new high-throughput technologies that generate more and more sequence data in a way that is best described as "better, cheaper, faster," with these advances feeding into the "insatiable appetite" that scientists have for more and more sequence data (Green et al. 2017). Given the inherent value of the data contained within these sequence databases, this chapter will focus on providing the reader with a solid understanding of these major public sequence databases, as a first step toward being able to perform robust and accurate bioinformatic analyses.

Nucleotide Sequence Databases

As described above, the major sources of nucleotide sequence data are the databases involved in INSDC – DDBJ, ENA, and GenBank – with new or updated data being shared between these three entities once every 24 hours. This transfer is facilitated by the use of common data formats for the kinds of information described in detail below.

The elementary format underlying the information held in sequence databases is a text file called the *flatfile*. The correspondence between individual flatfile formats greatly facilitates the daily exchange of data between each of these databases. In most cases, fields can be mapped on a one-to-one basis from one flatfile format to the other. Over time, various file formats have been adopted and have found continued widespread use; others have fallen to the wayside for a variety of reasons. The success of a given format depends on its usefulness in a variety of contexts, as well as its power in effectively containing and representing the types of biological data that need to be archived and communicated to scientists.

In its simplest form, a sequence record can be represented as a string of nucleotides with some basic tag or identifier. The most widely used of these simple formats is FASTA, originally introduced as part of the FASTA software suite developed by Lipman and Pearson (1985) that is described in detail in Chapter 3. This inherently simple format provides an easy way of handling primary data for both humans and computers, taking the following form.

```
>U54469.1
CGGTTGCTTGGGTTTTATAACATCAGTCAGTGACAGGCATTTCCAGAGTTGCCCTGTTCAACAATCGATA
GCTGCCTTTGGCCACCAAAATCCCAAACTTAATTAAAGAATTAAATAATTCGAATAATAATTAAGCCCAG
TAACCTACGCAGCTTGAGTGCGTAACCGATATCTAGTATACATTTCGATACATCGAAATCATGGTAGTGT
TGGAGACGGAGAAGGTAAGACGATGATAGACGGCGAGCCGCATGGGTTCGATTTGCGCTGAGCCGTGGCA
GGGAACAACAAAAACAGGGTTGTTGCACAAGAGGGGAGGCGATAGTCGAGCGGAAAAGAGTGCAGTTGGC
```

For brevity, only the first few lines of the sequence are shown. In the simplest incarnation of the FASTA format, the "greater than" character (>) designates the beginning of a new sequence record; this line is referred to as the *definition line* (commonly called the "def line"). A unique identifier – in this case, the *accession.version number* (U54469.1) – is followed by the nucleotide sequence, in either uppercase or lowercase letters, usually with 60 characters per line. The accession number is the number that is always associated with this sequence (and should be cited in publications), while the version number suffix allows users to easily determine whether they are looking at the most up-to-date record for a particular sequence. The version number suffix is incremented by one each time the sequence is updated.

Additional information can be included on the definition line to make this simple format a bit more informative, as follows.

```
>ENA|U54469|U54469.1 Drosophila melanogaster eukaryotic initiation factor 4E (eIF4E)
gene, complete cds, alternatively spliced.
```

This modified FASTA definition line now has information on the source database (ENA), its accession.version number (U54469.1), and a short description of what biological entity is represented by the sequence.

Nucleotide Sequence Flatfiles: A Dissection

As flatfiles represent the elementary unit of information within sequence databases and facilitate the interchange of information between these databases, it is important to understand what each individual field within the flatfile represents and what kinds of information can be found in varying parts of the record. While there are minor differences in flatfile formats, they can all be separated into three major parts: the *header*, containing information and descriptors pertaining to the entire record; the *feature table*, which provides relevant annotations to the sequence; and the sequence itself.

The Header

The header is the most database-specific part of the record. Here, we will use the ENA version of the record for discussion (shown in its entirety in Appendix 1.1), with the corresponding DDBJ and GenBank versions of the header appearing in Appendix 1.2. The first line of the record provides basic identifying information about the sequence contained in the record, appropriately named the ID line; this corresponds to the LOCUS line in DDBJ/GenBank.

```
ID   U54469; SV 1; linear; genomic DNA; STD; INV; 2881 BP.
```

The accession number is shown on the ID line, followed by its sequence version (here, the first version, or SV 1). As this is SV 1, this is equivalent to writing U54469.1, as described above. This is then followed by the topology of the DNA molecule (linear) and the molecule type (genomic DNA). The next element represents the ENA data class for this sequence (STD, denoting a “standard” annotated and assembled sequence). Data classes are used to group sequence records within functional divisions, enabling users to query specific subsets of the database. A description of these functional divisions can be found in Box 1.1. Finally, the ID line presents the taxonomic division for the sequence of interest (INV, for invertebrate; see Internet Resources) and its length (2881 base pairs). The accession number will also be shown separately on the AC line that immediately follows the ID lines.

Box 1.1 Functional Divisions in Nucleotide Databases

The organization of nucleotide sequence records into discrete functional types provides a way for users to query specific subsets of the records within these databases. In addition, knowledge that a particular sequence is from a given technique-oriented database allows users to interpret the data from the proper biological point of view. Several of these divisions are described below, and examples of each of these functional divisions (called “data classes” by ENA) can be found by following the example links listed on the ENA Data Formats page listed in the Internet Resources section of this chapter.

CON Constructed (or “contigged”) records of chromosomes, genomes, and other long DNA sequences resulting from whole -genome sequencing efforts. The records in this division do not contain sequence data; rather, they contain instructions for the assembly of sequence data found within multiple database records.

EST Expressed Sequence Tags. These records contain short (300–500 bp) single reads from mRNA (cDNA) that are usually produced in large numbers. ESTs represent a snapshot of what is expressed in a given tissue or at a given developmental stage. They represent tags – some coding, some not – of expression for a given cDNA library.

GSS Genome Survey Sequences. Similar to the EST division, except that the sequences are genomic in origin. The GSS division contains (but is not limited to) single-pass read genome survey sequences, bacterial artificial chromosome (BAC) or yeast artificial chromosome (YAC) ends, exon-trapped genomic sequences, and Alu polymerase chain reaction (PCR) sequences.

HTG High-Throughput Genome sequences. Unfinished DNA sequences generated by high-throughput sequencing centers, made available in an expedited fashion to the scientific community for homology and similarity searches. Entries in this division contain keywords indicating its phase within the sequencing process. Once finished, HTG sequences are moved into the appropriate database taxonomic division.

STD A record containing a standard, annotated, and assembled sequence.

STS Sequence-Tagged Sites. Short (200–500 bp) operationally unique sequences that identify a combination of primer pairs used in a PCR assay, generating a reagent that maps to a single position within the genome. The STS division is intended to facilitate cross-comparison of STSs with sequences in other divisions for the purpose of correlating map positions of anonymous sequences with known genes.

WGS Whole-Genome Shotgun sequences. Sequence data from projects using shotgun approaches that generate large numbers of short sequence reads that can then be assembled by computer algorithms into sequence contigs, higher -order scaffolds, and sometimes into near-chromosome- or chromosome-length sequences.

Following the ID line are one or more date lines (denoted by DT), indicating when the entry was first created or last updated. For our sequence of interest, the entry was originally created on May 19, 1996 and was last updated in ENA on June 23, 2017:

```
DT   19-MAY-1996 (Rel. 47, Created)
DT   23-JUN-2017 (Rel. 133, Last updated, Version 5)
```

The release number in each line indicates the first quarterly release made *after* the entry was created or last updated. The version number for the entry appears on the second line and allows the user to determine easily whether they are looking at the most up-to-date record for a particular sequence. Please note that this is different from the accession.version format described above – while some element of the record may have changed, the sequence may have remained the same, so these two different types of version numbers may not always correspond to one another.

The next part of the header contains the definition lines, providing a succinct description of the kinds of biological information contained within the record. The definition line (DE in ENA, DEFINITION in DDBJ/GenBank) takes the following form.

```
DE   Drosophila melanogaster eukaryotic initiation factor 4E (eIF4E) gene,
DE   complete cds, alternatively spliced.
```

Much care is taken in the generation of these definition lines and, although many of them can be generated automatically from other parts of the record, they are reviewed to ensure that consistency and richness of information are maintained. Obviously, it is quite impossible to capture all of the biology underlying a sequence in a single line of text, but that wealth of information will follow soon enough in downstream parts of the same record.

Continuing down the flatfile record, one finds the full taxonomic information on the sequence of interest. The OS line (or SOURCE line in DDBJ/GenBank) provides the preferred scientific name from which the sequence was derived, followed by the common name of the organism in parentheses. The OC lines (or ORGANISM lines in DDBJ/GenBank) contain the complete taxonomic classification of the source organism. The classification is listed top-down, as nodes in a taxonomic tree, with the most general grouping (Eukaryota) given first.

```
OS   Drosophila melanogaster (fruit fly)
OC   Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota;
OC   Neoptera; Holometabola; Diptera; Brachycera; Muscomorpha; Ephydroidea;
OC   Drosophilidae; Drosophila; Sophophora.
```

Each record must have at least one reference or citation, noted within what are called *reference blocks*. These reference blocks offer scientific credit and set a context explaining why this particular sequence was determined. The reference blocks take the following form.

```
RN   [1]
RP   1-2881
RX   DOI; .1074/jbc.271.27.16393.
RX   PUBMED; 8663200.
RA   Lavoie C.A., Lachance P.E., Sonenberg N., Lasko P.;
RT   "Alternatively spliced transcripts from the Drosophila eIF4E gene produce
RT   two different Cap-binding proteins";
RL   J Biol Chem 271(27):16393-16398(1996).
XX
RN   [2]
RP   1-2881
RA   Lasko P.F.;
RT   ;
RL   Submitted (09-APR-1996) to the INSDC.
RL   Paul F. Lasko, Biology, McGill University, 1205 Avenue Docteur Penfield,
RL   Montreal, QC H3A 1B1, Canada
```

In this case, two references are shown, one referring to a published paper and the other referring to the submission of the sequence record itself. In the example above, the second block provides information on the senior author of the paper listed in the first block, as well as the author's postal address. While the date shown in the second block indicates when the sequence (and accompanying information) was submitted to the database, it does not indicate when the record was first made public, so no inferences or claims based on first public release can be made based on this date. Additional submitter blocks may be added to the record each time the sequence is updated.

Some headers may contain COMMENT (DDBJ/GenBank) or CC (ENA) lines. These lines can include a great variety of notes and comments (*descriptors*) that refer to the entire record. Often, genome centers will use these lines to provide contact information and to confer acknowledgments. Comments also may include the history of the sequence. If the sequence of a particular record is updated, the comment will contain a pointer to the previous versions of the record. Alternatively, if an earlier version of the record is retrieved, the comment will point forward to the newer version, as well as backwards, if there was a still earlier version. Finally, there are database cross-reference lines (marked DR) that provide links to allied databases containing information related to the sequence of interest. Here, a cross-reference to FlyBase can be seen in the complete header for this record in Appendix 1.1. Note that the corresponding DDBJ/GenBank header in Appendix 1.2 does not contain these cross-references.

The Feature Table

Early on in the collaboration between INSDC partner organizations, an effort was made to come up with a common way to represent the biological information found within a given database record. This common representation is called the *feature table*, consisting of *feature keys* (a single word or abbreviation indicating the described biological property), *location* information denoting where the feature is located within the sequence, and additional *qualifiers* providing additional descriptive information about the feature. The online INSDC feature table documentation is extensive and describes in great detail what features are allowed and what qualifiers can be used with each individual feature. Wording within the feature table uses common biological research terminology wherever possible and is consistent between DDBJ, ENA, and GenBank entries.

Here, we will dissect the feature table for the eukaryotic transcription factor 4E gene from *Drosophila melanogaster*, shown in its entirety in both Appendices 1.3 (in ENA format) and 1.4 (in DDBJ/GenBank format). This particular sequence is alternatively spliced, producing two distinct gene products, 4E-I and 4E-II. The first block of information in the feature table is always the source feature, indicating the biological source of the sequence and additional information relating to the entire sequence. This feature must be present in all INSDC entries, as all DNA or RNA sequences derive from some specific biological source, including synthetic DNA.

```
FT   source          1..2881
FT                   /organism="Drosophila melanogaster"
FT                   /chromosome="3"
FT                   /map="67A8-B2"
FT                   /mol_type="genomic DNA"
FT                   /db_xref="taxon:7227"
FT   gene            80..2881
FT                   /gene="eIF4E"
```

In the first line of the source key, notice that the numbering scheme shows the range of positions covered by this feature key as two numbers separated by two dots (1..2881). As the source key pertains to the entire sequence, we can infer that the sequence described in this entry is 2881 nucleotides in length. The various ways in which the location of any given feature can be indicated are shown in Table 1.1, accounting for a wide range of biological scenarios. The qualifiers then follow, each preceded by a slash. The full scientific name of the organism is provided, as are specific mapping coordinates, indicating that this sequence is at map location 67A8-B2 on chromosome 3. Also indicated is the type of molecule that was sequenced (genomic DNA). Finally, the last line indicates a database cross-reference (abbreviated as db_xref) to the NCBI taxonomy database, where taxon 7227 corresponds to *D. melanogaster*. In general, these cross-references are controlled qualifiers that allow entries to be connected to an external database, using an identifier that is unique to that external database. Following the source block above is the gene feature, indicating that the gene itself is a subset of the entire sequence in this entry, starting at position 80 and ending at position 2881.

```
FT   mRNA            join(80..224,892..1458,1550..1920,1986..2085,2317..2404,
FT                   2466..2881)
FT                   /gene="eIF4E"
FT                   /product="eukaryotic initiation factor 4E-I"
FT   mRNA            join(80..224,1550..1920,1986..2085,2317..2404,2466..2881)
FT                   /gene="eIF4E"
FT                   /product="eukaryotic initiation factor 4E-II"
```

Table 1.1 Indicating locations within the feature table.

`345`	Single position within the sequence
`345..500`	A continuous range of positions bounded by and including the indicated positions
`<345..500`	A continuous range of positions, where the exact lower boundary is not known; the feature begins somewhere prior to position 345 but ends at position 500
`345..>500`	A continuous range of positions, where the exact upper boundary is not known; the feature begins at position 345 but ends somewhere after position 500
`<1..888`	The feature starts before the first sequenced base and continues to position 888
`(102.110)`	Indicates that the exact location is unknown, but that it is one of the positions between 102 and 110, inclusive
`123ˆ124`	Points to a site *between* positions 123 and 124
`123ˆ177`	Points to a site *between* two adjacent nucleotides or amino acids anywhere between positions 123 and 177
`join(12..78,134..202)`	Regions 12–78 and 134–202 are joined to form one contiguous sequence
`complement(4918..5126)`	The sequence complementary to that found from 4918 to 5126 in the sequence record
`J00194:100..202`	Positions 100–202, inclusive, in the entry in this database having accession number J00194

The next feature in this example indicates which regions form the two mRNA transcripts for this gene, the first for eukaryotic initiation factor 4E-I and the second for eukaryotic initiation factor 4E-II. In the first case (shown above), the `join` line indicates that six distinct DNA segments are transcribed to form the mature RNA transcript while, in the second case, the second region is missing, with only five distinct DNA segments transcribed into the mature RNA transcript – hence the two splice variants that are ultimately encoded by this molecule.

```
FT   CDS             join(201..224,1550..1920,1986..2085,2317..2404,2466..2629)
FT                   /codon_start=1
FT                   /gene="eIF4E"
FT                   /product="eukaryotic initiation factor 4E-II"
FT                   /note="Method: conceptual translation with partial peptide
FT                   sequencing"
FT                   /db_xref="GOA:P48598"
FT                   /db_xref="InterPro:IPR001040"
FT                   /db_xref="InterPro:IPR019770"
FT                   /db_xref="InterPro:IPR023398"
FT                   /db_xref="PDB:4AXG"
FT                   /db_xref="PDB:4UE8"
FT                   /db_xref="PDB:4UE9"
FT                   /db_xref="PDB:4UEA"
FT                   /db_xref="PDB:4UEB"
FT                   /db_xref="PDB:4UEC"
FT                   /db_xref="PDB:5ABU"
FT                   /db_xref="PDB:5ABV"
FT                   /db_xref="PDB:5T47"
FT                   /db_xref="PDB:5T48"
FT                   /db_xref="UniProtKB/Swiss-Prot:P48598"
FT                   /protein_id="AAC03524.1"
FT                   /translation="MVVLETEKTSAPSTEQGRPEPPTSAAAPAEAKDVKPKEDPQETGE
FT                   PAGNTATTTAPAGDDAVRTEHLYKHPLMNVWTLWYLENDRSKSWEDMQNEITSFDTVED
FT                   FWSLYNHIKPPSEIKLGSDYSLFKKNIRPMWEDAANKQGGRWVITLNKSSKTDLDNLWL
FT                   DVLLCLIGEAFDHSDQICGAVINIRGKSNKISIWTADGNNEEAALEIGHKLRDALRLGR
FT                   NNSLQYQLHKDTMVKQGSNVKSIYTL"
```

Following the mRNA feature is the CDS feature shown above, describing the region that ultimately encodes the protein product. Focusing just on eukaryotic initiation factor 4E-II, the CDS feature also shows a `join` line with coordinates that are slightly different from those shown in the mRNA feature, specifically at the beginning and end positions. The difference lies in the fact that the 5′ and 3′ untranslated regions (UTRs) are included in the mRNA feature but not in the CDS feature. The CDS feature corresponds to the sequence of amino acids found in the translated protein product whose sequence is shown in the `/translation` qualifier above. The `/codon_start` qualifier indicates that the amino acid translation of the first codon begins at the first position of this joined region, with no offset.

The `/protein_id` qualifier shows the accession number for the corresponding entry in the protein databases (AAC03524.1) and is hyperlinked, enabling the user to go directly to that entry. These unique identifiers use a "3 + 5" format – three letters, followed by five numbers. Versions are indicated by the decimal that follows; when the protein sequence in the record changes, the version is incremented by one. The assignment of a gene product or protein name (via the `/protein` qualifier) often is subjective, sometimes being assigned via weak similarities to other (and sometimes poorly annotated) sequences. Given the potential for the transitive propagation of poor annotations (that is, bad data tend to beget more bad data), users are advised to consult *curated* nucleotide and protein sequence databases for the most up-to-date, accurate information regarding the putative function of a given sequence. Finally, notice the extensive cross-referencing via the `/db_xref` qualifier to entries in InterPro, the

Protein Data Bank (PDB), and UniProtKB/Swiss-Prot, as well as to a Gene Ontology annotation (GOA; Gene Ontology Consortium 2017).

Implicit in the source feature and the organism that is assigned to it is the genetic code used to translate the nucleic acid sequence into a protein sequence when a CDS feature is present in the record. Also, the DNA-centric nature of these feature tables means that all features are mapped through a DNA coordinate system, not that of amino acid reference points, as shown in the examples in Appendices 1.3 and 1.4.

```
SQ   Sequence 2881 BP; 849 A; 699 C; 585 G; 748 T; 0 other;
     cggttgcttg ggttttataa catcagtcag tgacaggcat ttccagagtt gccctgttca         60
     acaatcgata gctgcctttg gccaccaaaa tcccaaactt aattaaagaa ttaaataatt        120
     cgaataataa ttaagcccag taacctacgc agcttgagtg cgtaaccgat atctagtata        180
     .
     . <truncated for brevity>
     .
     aaacggaacc ccctttgtta tcaaaaatcg gcataatata aaatctatcc gctttttgta       2820
     gtcactgtca ataatggatt agacggaaaa gtatattaat aaaaacctac attaaaaccg       2880
     g                                                                        2881
//
```

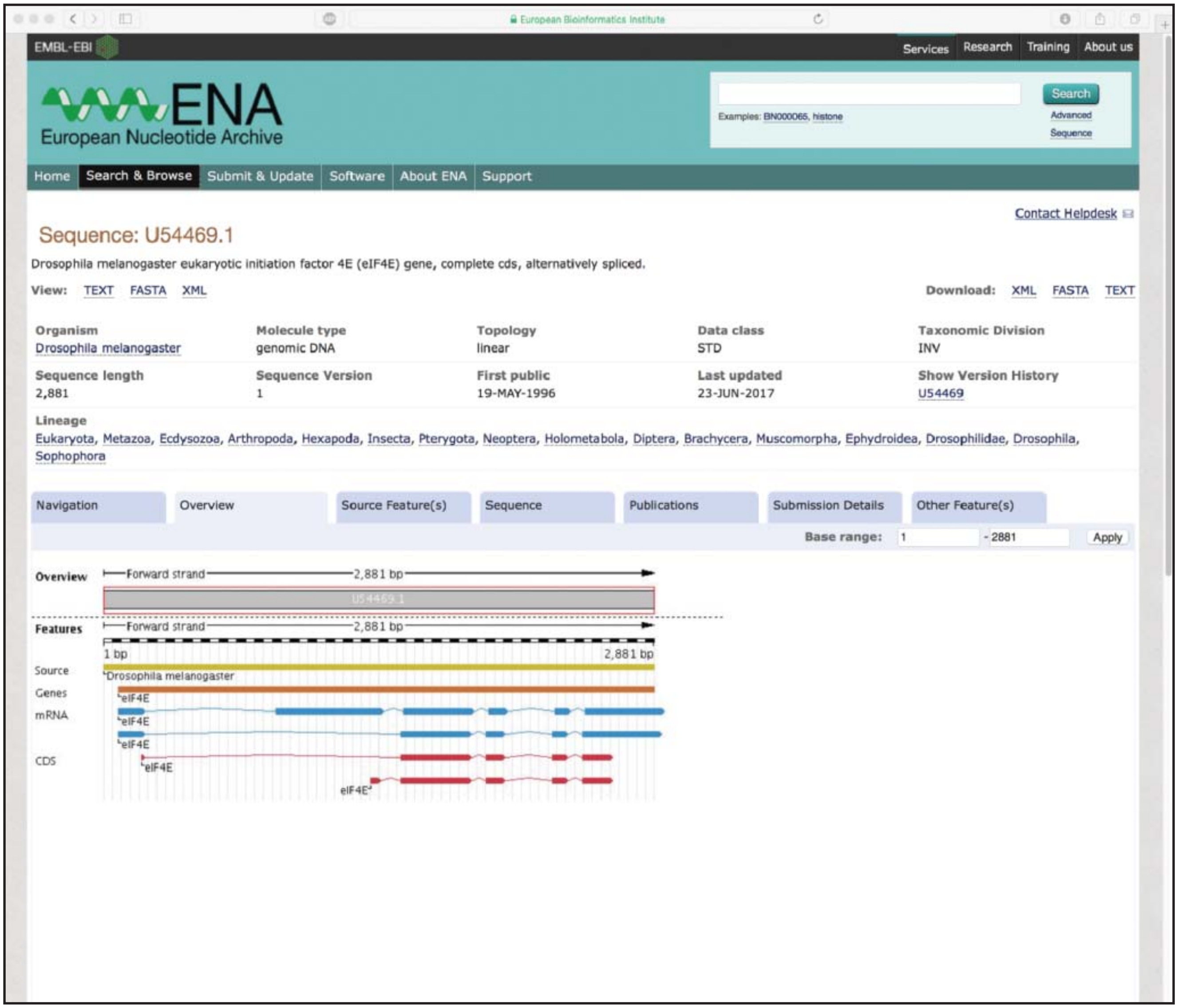
Finally, at the end of every nucleotide sequence record, one finds the actual nucleotide sequence, with 60 bases per row. Note that, in the SQ line signaling the beginning of this section of the record, not only is the overall length of the sequence provided, but a count of how many of each individual type of nucleotide base is also provided, making it quite easy to compute the GC content of this sequence.

Graphical Interfaces

Graphical interfaces have been developed to facilitate the interpretation of the data found within text-based flatfiles, with an example of the graphical view of the ENA record for our sequence of interest (U54469.1) shown in Figure 1.1. These graphical views are particularly useful when there is a long list of documented biological features within the feature table, enabling the user to visualize potential interactions or relationships between biological features. An additional example of the use of graphical views to assist in the interpretation of the information found within a database record is provided in the discussion of the NCBI Entrez discovery pathway in Chapter 2, as well as later in this chapter.

RefSeq

As one might expect, especially given the breakneck speed at which DNA sequence data are currently being produced, there is a significant amount of redundancy within the major sequence databases, with a good number of sequences being represented more than once. This is often problematic for the end user, who may find themselves confused as to which sequence to use after performing a search that returns numerous results. To address this issue, NCBI developed RefSeq, the goal of which is to provide a single reference sequence for each molecule of the central dogma – DNA, RNA, and protein. The distinguishing features of RefSeq go beyond its non-redundant nature, with individual entries including the biological attributes of the gene, gene transcript, or protein. RefSeq entries encompass a wide taxonomic range, and entries are updated and curated on an ongoing basis to reflect current knowledge about the individual entries. Additional information on RefSeq can be found in Box 1.2.



Sequence: U54469.1

Drosophila melanogaster eukaryotic initiation factor 4E (eIF4E) gene, complete cds, alternatively spliced.

View: TEXT FASTA XML | Download: XML FASTA TEXT

Organism	Molecule type	Topology	Data class	Taxonomic Division
Drosophila melanogaster	genomic DNA	linear	STD	INV
Sequence length	**Sequence Version**	**First public**	**Last updated**	**Show Version History**
2,881	1	19-MAY-1996	23-JUN-2017	U54469

Lineage
Eukaryota, Metazoa, Ecdysozoa, Arthropoda, Hexapoda, Insecta, Pterygota, Neoptera, Holometabola, Diptera, Brachycera, Muscomorpha, Ephydroidea, Drosophilidae, Drosophila, Sophophora

Figure 1.1 The landing page for ENA record U54469.1, providing a graphical view of biological features found within the sequence of the *Drosophila melanogaster* eukaryotic initiation factor 4E (*eIF4E*) gene. The tracks within the graphical view show the position of the gene, mRNAs, and coding regions (marked CDS) within the 2881 bp sequence reported in this record.

Box 1.2 RefSeq

The first several chapters of this book describe a variety of ways in which sequence data and sequence annotations find their way into public databases. While the combination of data derived from systematic sequencing projects and individual investigators' laboratories yields a rich and highly valuable set of sequence data, some problems are apparent. The most important issue is that a single biological entity may be represented by many different entries in various databases. It also may not be clear whether a given sequence has been experimentally determined or is simply the result of a computational prediction.

To address these issues, NCBI developed the RefSeq project, the major goal of which is to provide a reference sequence for each molecule in the central dogma (DNA, mRNA, and protein). As each biological entity is represented only once, RefSeq is, by definition, non-redundant. Nucleotide and protein sequences in RefSeq are explicitly linked to one