Applied Bioinformatics

Paul M. Selzer
Richard J. Marhöfer
Oliver Koch

Applied Bioinformatics

An Introduction

Second Edition

Springer

Paul M. Selzer
Boehringer Ingelheim
Animal Health
Ingelheim am Rhein, Germany

Richard J. Marhöfer
MSD Animal Health
Innovation GmbH
Schwabenheim, Germany

Oliver Koch
TU Dortmund University
Faculty of Chemistry
and Chemical Biology
Dortmund, Germany

The first edition of this textbook was written by Paul M. Selzer, Richard J. Marhöfer, and Andreas Rohwer

Originally published in German with the title:
Angewandte Bioinformatik 2018
ISBN 978-3-319-68299-0 ISBN 978-3-319-68301-0 (eBook)
https://doi.org/10.1007/978-3-319-68301-0

Library of Congress Control Number: 2018930594

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Though a relatively young discipline, bioinformatics is finding increasing importance in many life science disciplines, including biology, biochemistry, medicine, and chemistry. Since its beginnings in the late 1980s, the success of bioinformatics has been associated with rapid developments in computer science, not least in the relevant hardware and software. In addition, biotechnological advances, such as have been witnessed in the fields of genome sequencing, microarrays, and proteomics, have contributed enormously to the bioinformatics boom. Finally, the simultaneous breakthrough and success of the World Wide Web has facilitated the worldwide distribution of and easy access to bioinformatics tools.

Today, bioinformatics techniques, such as the Basic Local Alignment Search Tool (BLAST) algorithm, pairwise and multiple sequence comparisons, queries of biological databases, and phylogenetic analyses, have become familiar tools to the natural scientist. Many of the software products that were initially unintuitive and cryptic have matured into relatively simple and user-friendly products that are easily accessible over the Internet. One no longer needs to be a computer scientist to proficiently operate bioinformatics tools with respect to complex scientific questions. Nevertheless, what remains important is an understanding of fundamental biological principles, together with a knowledge of the appropriate bioinformatics tools available and how to access them. Also and not least important is the confidence to apply these tools correctly in order to generate meaningful results.

The present, comprehensively revised second English edition of this book is based on a lecture series of Paul M. Selzer, professor of biochemistry at the Interfaculty Institute for Biochemistry, Eberhard-Karls-University, Tübingen, Germany, as well as on multiple international teaching events within the frameworks of the *EU FP7* and *Horizon 2020* programs. The book is unique in that it includes both exercises and their solutions, thereby making it suitable for classroom use. Based on both the huge national success of the first German edition from 2004 and the subsequently overwhelming international success of the first English edition from 2008, the authors decided to produce a second German and English edition in close proximity to each other. Working on the same team, each of the three authors had many years of accumulated expertise in research and development within the pharmaceutical industry, specifically in the area of bioinformatics and cheminformatics, before they moved to different career opportunities to widen their individual industrial and academic scientific areas of expertise. The aim of this book is both to introduce the daily application of a variety of bioinformatics tools and provide an overview of a complex field. However, the intent is neither to describe nor even derive formulas or algorithms, but rather to facilitate rapid and structured access to applied bioinformatics by interested students and scientists. Therefore, detailed knowledge in computer programming is not required to understand or apply this book's contents.

Each of the seven chapters describes important fields in applied bioinformatics and provides both references and Internet links. Detailed exercises and solutions are meant to encourage the reader to practice and learn the topic and become proficient in the relevant software. If possible, the exercises are chosen in such a way that examples, such as protein or nucleotide sequences, are interchangeable. This allows readers to choose examples that are closer to their scientific interests based on a sound understanding of the underlying principles. Direct input required by the user, either through text or by pressing buttons, is indicated in `Courier font` and *italics*, respectively. Finally, the book concludes with a detailed glossary of common definitions and terminology used in applied bioinformatics.

We would like to thank our former colleague and coauthor of the first edition, Dr. Andreas Rohwer, for his contributions, which are still of great importance in the second edition. We are very grateful to Ms. Christiane Ehrt and Ms. Lina Humbeck – TU Dortmund, Germany – for mindfully reading the book and actively verifying all exercises and solutions. We wish to thank Dr. Sandra Noack for her constructive contributions. Finally, we wish to thank Ms. Stefanie Wolf and Ms. Sabine Schwarz from the publisher Springer for their continuous support in producing the second edition.

Paul M. Selzer
Ingelheim am Rhein, Germany

Richard J. Marhöfer
Worms, Germany

Oliver Koch
Dortmund, Germany
May 2018

The Circulation of Genetic Information

Genetic information is encoded by a 4-letter alphabet, which in turn is translated into proteins using a 20-letter alphabet. Proteins fold into three-dimensional structures that perform essential functions in single-celled or multicellular organisms. These organisms are under constant selection pressure, which in turn leads to changes in their genetic information.

Cover Image

The three-dimensional molecular structure of a protein-DNA complex is depicted. The transcription activator Gal4 from *Saccharomyces cerevisiae* is shown bound to a DNA oligomer (PDB-ID: 1D66). Gal4 is represented by a ribbon model in which α-helices and loops are drawn in red and yellow, respectively. The side chains of the amino acids in the loops are not shown. For the DNA oligomer, local bending of the molecular surface is color-coded where darker colors represent increased bending [Brickmann J, Exner TE, Keil M, Marhöfer RJ (2000) Molecular graphics - trends and perspectives. J Mol Mod 6:328-340]. The structure was produced on a Silicon Graphics Octane 2 workstation using the software package MOLCAD/Sybyl (Tripos Inc.) [Brickmann J, Goetze T, Heiden W, Moeckel G, Reiling S, Vollhardt H, Zachmann CD (1995) Interactive visualization of molecular scenarios with MOLCAD/Sybyl. In: Bowie JE (Hrsg) Data visualization in molecular science - tools for insight and innovation. Addison-Wesley Publishing Company Inc, Reading, Massachusetts, USA, S 83-97].

A Short History of Bioinformatics

The first algorithm for comparing protein or DNA sequences was published by Needleman and Wunsch in 1970 (▶ Chap. 3). Bioinformatics is thus only 1 year younger than the Internet progenitor ARPANET and 1 year older than e-mail, which was invented by Ray Thomlinson in 1971. However, the term *bioinformatics* was only coined in 1978 (Hogeweg 1978) and was defined as the "study of informatic processes in biotic systems." The Brookhaven Protein Data Bank (PDB) was also founded in 1971. The PDB is a database for the storage of crystallographic data of proteins (▶ Chap. 2). The development of bioinformatics proceeded very slowly at first until the complete gene sequence of the bacteriophage virus ϕX174 was published in 1977 (Sanger et al. 1977). Shortly after, the IntelliGenetics Suite, the first software package for the analysis of DNA and protein sequences, was used (1980). In the following year, Smith and Waterman published another algorithm for sequence comparison, and IBM marketed the first personal computer (▶ Chap. 3). In 1982, a spin-off of the University of Wisconsin – the Genetics Computer Group – marketed a software package for molecular biology, the Wisconsin Suite. At first, both the IntelliGenetics and the Wisconsin Suite were packages of single, relatively small programs that were controlled via the command line. A graphical user interface was later developed for the Wisconsin Suite, which made for more convenient operation of the programs. The IntelliGenetics suite has since disappeared from the market, but the Wisconsin Suite was available under the name GCG until the 2000s.

The publication of the polymerase chain reaction (PCR) process by Mullis and colleagues in 1986 represented a milestone in molecular biology and, concurrently, bioinformatics (Mullis et al. 1986). In the same year, the SWISS-PROT database was founded, and Thomas Roderick coined the term *genomics*, describing the scientific discipline of sequencing and description of whole genomes (Kuska 1998). Two years later, the National Center for Biotechnology Information (NCBI) was established; today, it operates one of the most important primary databases (◘ Fig. 1; see ▶ Chap. 2). The same year also saw the start of the Human Genome Initiative and the publication of the FASTA algorithm (▶ Chap. 3). In 1991, CERN released the protocols that made possible the World Wide Web (▶ https://home.cern/topics/birth-web; ▶ https://timeline.web.cern.ch/timelines/The-birth-of-the-World-Wide-Web). The Web made it possible, for the first time, to provide easy access to bioinformatics tools. However, it took a few years until such tools actually became available. Also, in 1991 Greg Venter published the use of Expressed Sequence Tags (ESTs) (▶ Chap. 4). By the next year, Venter and his wife, Claire Fraser, had founded The Institute for Genomics Research (TIGR). With the publication of GeneQuiz in 1994, a fully integrated sequence analysis tool appeared that, in 1996, was used in the GeneCrunch project for the first automatic analysis of the over 6000 proteins of baker's yeast, *Saccharomyces cerevisiae* (Goffeau et al. 1996). In the same year,

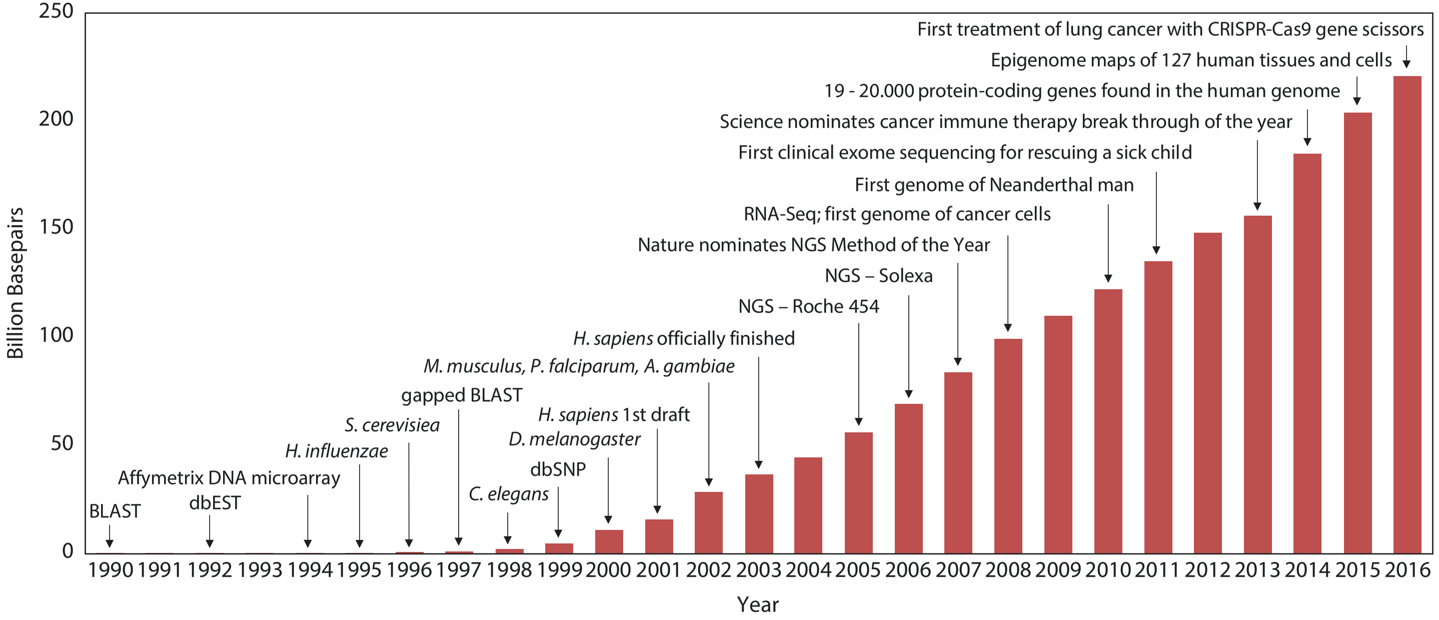

Fig. 1 Development of NCBI's GenBank database in connection with some milestones of bioinformatics. Coauthored by Dr. Quang Hon Tran

the launch of the Prosite database (► Chap. 2) was announced. One year after the successful implementation of the GeneQuiz package for automatic sequence analysis, LION Biosciences AG was founded in Heidelberg, Germany. The basis for one of LION's main products, the integrated sequence analysis package, termed bioSCOUT, was GeneQuiz. Together with other products of the Sequence-Retrieval System (SRS) package, LION Biosciences AG quickly became a very successful bioinformatics company with a worldwide presence. This did not last for long, however, and in 2006 the bioinformatics division was sold to BioWisdom, which continued to modify and sell SRS. At this time, SRS was certainly one of the most important systems for the indexing and managing of flat file databases. The importance of SRS has steadily declined in recent years; nevertheless, a few installations can still be found on the Web.

Twenty years after the term *bioinformatics* had been coined, another term, *chemoinformatics*, was published (Brown 1998). Up till that time, the terms *chemometrics*, *computer chemistry*, and *computational chemistry* were common and are still in use today. The term chemoinformatics, sometimes also cheminformatics, is used as an umbrella term that sometimes even includes additional terms like *molecular modeling*. Note that : traditionalists still use the term only for the representation and handling of chemical structures in databases.

The 1990s saw additional milestones in bioinformatics and molecular biology. The genomes of three important model organisms were published: *Haemophilus influenzae* (Fleischmann et al. 1995), *S. cerevisiae* (1996), and *Caenorhabditis elegans* (*C. elegans Sequencing Consortium* 1998). Also, in 1998, Greg Ventor founded his company Celera, and in 2000 the genomes of two additional model organisms followed, *Arabidopsis thaliana* and *Drosophila melanogaster*. The next year saw the publication of the first draft of the human genome, which officially was declared to be completed in 2003. In 2002 three important institutes, the European Bioinformatics Institute (EMB-EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR), founded the UniProt Consortium and combined their databases Swiss-Prto, TrEMBL, and PIR-PSD in the UniProt database (► Chap. 2). The same year saw the publication of the mouse (*mus musculus*) genome, the genome of the causative agent of human malaria, *Plasmodium falciparum*, and its vector, the mosquito *Anopheles gambiae*. Shortly after, in 2004, the genome of the brown rat (*Rattus norvegicus*) was published, followed by the genome of the chimpanzee (*Pan troglodytes*) in 2005. The sequencing of other genomes is an ongoing process, and to list them all would go beyond the scope of this short survey. An overview of the completed and ongoing genome projects can be found in the Genomes OnLine Database GOLD: ► http://www.genomesonline.org/.

In 2005, 454 sequencing – the first technique of the Next-Generation Sequencing (NGS, see ► Chap. 4) – was presented, followed shortly – in 2006 – by Solexa sequencing. NGS was nominated method of the year by the journal *Nature Methods* already 1 year later. Another year later, in 2008, RNA-Seq, which is based on NGS, was introduced and led to a number of new disciplines, for example, pharmacogenetics and proteogenomics (► Chap. 4). NGS has also taken on an important role in medical practice, where it is extensively used in the field of personalized medicine. As a matter of course, new Web services and new databases are developed and published constantly, in part for highly specialized purposes. It would go far beyond the scope of this book to list all of those purposes. A comprehensive list of databases, however, can be found once a year in the January issue of the journal *Nucleic Acids Research* (database issue), and a listing of Web services is published also ones a year in the July issue (software issue): NAR: https://nar.oxfordjournals.org/.

References

Brown (1998) Chemoinformatics: what is it and how does it impact drug discovery. Annu Rep Med Chem 33:375–384

C. elegans Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans:* a platform for investigating biology. Science 282:2012–2018

Fleischmann et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd*. Science 269:496–512

Goffeau et al. (1996) Life with 6000 genes. Science 274:546–567

Hogeweg (1978) Simulation of cellular forms. In: Zeigler BP (ed) Frontiers in system modelling. Simulation Councils, Inc., pp 90–95

Kuska (1998) Beer, Bethesda, and biology: how "genomics" came into being. J Nat Cancer Inst 90:93

Mullis et al. (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol 51(Pt 1):263–273

Sanger et al. (1977) Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265:687–695

Contents

Supplementary Information

About the Authors

Paul M. Selzer

works as a researcher and scientific manager at Boehringer Ingelheim Animal Health, Germany. He is also visiting professor at the Interfaculty Institute of Biochemistry at the University of Tübingen, Germany, and honorary professor at the Department of Infection, Immunity, and Inflammation at the University of Glasgow, Scotland.

Richard J. Marhöfer

is chemoinformatics researcher at MSD Animal Health, Germany.

Oliver Koch

is independent group leader for medicinal chemistry at the TU Dortmund University, Germany.