新一代信息科学与技术

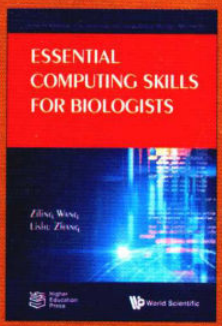
Essential Computing Skills for Biologists

Ziling Wang
Lishu Zhang

生物信息计算

（英文版）

高等教育出版社



高等教育出版社与 World Scientific 公司合作出版
海外发行 ISBN: 978-1-84816-924-1

生物信息计算是在生命科学的研究中，以计算机为工具对生物信息进行储存、检索和分析的科学，是当今生命科学和自然科学的重大前沿领域之一。本书重点集中在基因组学和蛋白质组学两方面，分 12 章介绍：序列数据资源及其检索、蛋白质序列数据库、序列比对与比对搜索、真核生物基因结构的预测分析、分子进化分析、蛋白质结构数据库和结构可视化、蛋白质结构与功能预测、微阵列数据分析、系统生物学网络结构分析等内容，并对 Gene Ontology 数据库、KEGG 数据库及基因功能注释、R 语言及 Bioconductor 进行了简述和应用示例。

本书是计算机科学、生物信息学以及统计学的知识综合，可供相关领域的专业人员参考使用，也可用于生物学、计算机科学专业的教学。

ISBN 978-7-04-049018-3
9 787040 490183 >
定价 89.00 元

学科分类：生物电子学与生物信息处理
http:// academic.hep.com.cn

Essential Computing Skills for Biologists

生物信息计算

（英文版）

Ziling Wang
Lishu Zhang

SHENGWU XINXI JISUAN 高等教育出版社·北京

Ziling Wang
Beijing Jiaotong University
Beijing, China
zlw@bjtu.edu.cn

Lishu Zhang
Beijing Jiaotong University
Beijing, China

图书在版编目（CIP）数据

生物信息计算基础 = Essential Computing Skills for Biologists : 英文 / 王字玲，张丽姝主编. -- 北京 : 高等教育出版社，2018.4
ISBN 978-7-04-049018-3

Ⅰ. ①生… Ⅱ. ①王… ②张… Ⅲ. ①生物信息论－英文 Ⅳ. ① Q811.4

中国版本图书馆 CIP 数据核字（2018）第 008093 号

策划编辑	冯 英	责任编辑	冯 英	封面设计	张 楠	版式设计	童 丹
插图绘制	杜晓丹	责任校对	高 歌	责任印制	尤 静		

出版发行	高等教育出版社	网 址	http://www.hep.edu.cn
社 址	北京市西城区德外大街4号		http://www.hep.com.cn
邮政编码	100120	网上订购	http://www.hepmall.com.cn
印 刷	北京佳信达欣艺术印刷有限公司		http://www.hepmall.com
开 本	787mm×1092mm 1/16		http://www.hepmall.cn
印 张	19		
字 数	480 千字	版 次	2018 年 4 月第 1 版
购书热线	010-58581118	印 次	2018 年 4 月第 1 次印刷
咨询电话	400-810-0598	定 价	89.00 元

物 料 号 49018-00

Preface

Over the past few decades, developments in genomic, molecular research, as well as information technologies have produced a tremendous amount of information related to molecular biology. This made computing skills become essential in molecular biology research. Computational methods have evolved to analyze and interpret various types of data, such as nucleotide and amino acid sequences, protein domains, and protein structures, etc. Bioinformatics is the name given to these mathematical and computing approaches used for understanding biological processes and becomes an important part of many areas of biology. Bioinformatics not only provides tools we can use to understand the basic aspect of biology, including development, metabolism, adaptation to the environment, genetic variations of individual, and evolution, but also facilitates our understanding of the diseases processes through the analysis of molecular sequence data.

This book is a handbook of methods and protocols for biologists. It aims at undergraduate, graduate students and researchers originally trained in biological or medical sciences who need to know how to access the data archives of genomes, proteins, metabolites, gene expression profiles and the questions these data and tools can answer, for example, how to draw inferences from data archives and how to make connections among them to deduce useful predictions. For each chapter, the conceptual and experimental background is provided, together with specific guidelines for handling raw data, including preprocessing and analysis.

The book is structured in three parts. Part I introduces the basic knowledge about popular bioinformatics tools, databases and web resources, including online sequence databases, sequence alignment, predicting DNA and protein function from sequence, protein structure prediction and analysis, molecular phylogeny and evolution. Part II presents examples of Omics bioinformatics applications, including genetic variation and human disease, gene expression profile and data management, qualitative and quantitative proteomics, bioinformatics for metabolomics, and integrating Omics data for pathways and interaction networks. Part III provides basic statistical analysis skills and programming skills needed to handle and analyze Omic datasets.

Bioinformatics is an interdisciplinary field involving molecular biology and genetics, computer science, mathematics, and statistics. It is too far broad to be understood by one person. Thus, this book is written by multiple authors, each of whom brings a deeper knowledge of the subject. I wish to express my gratitude to all authors for their dedication in providing excellent chapters, some of the data and examples presented in the book are the results of their own research. As for any omissions or errors, the responsibility is mine.

In preparing the book, we read many textbooks and published papers and viewed many websites, we sincerely apologize to those authors and researchers whose work we did not cite.

Enjoy reading.

Wang Ziling
College of Life Sciences and Bioengineering,
Beijing Jiaotong University, Beijing
October, 2017

Contents

PART I DATABASES AND BIOINFORMATICS TOOLS

PART I

DATABASES AND BIOINFORMATICS TOOLS

Chapter 1 Online Sequence Database

Yong Liu[1], Lishu Zhang[2]

In recent years, with the rapid development of computer and network technology, a large number of biological information resources can be retrieved through the Internet. Such a large number of biological databases, software resources and Internet connection are making life science research more convenient and efficient.

The major objectives of biological databases are not only to store, organize and share data in a structured and searchable manner with the aim to facilitate data retrieval and visualization for humans, but also to provide web application programming interfaces (APIs) for computers to exchange and integrate data from various database resources in an automated manner.

According to the report of 2016 database issue of *Nucleic Acids Research*, there are 1685 databases that are publicly accessible online. Various databases cover all areas of life sciences, and in this chapter, we will focus on some of the commonly used biology databases or online resources, including, ① nucleic acid sequence database, such as GenBank, EMBL, DDBJ, etc.; ② protein database uniprot; ③ protein three-dimensional structure of the database PDB; ④ online human genome resources: UCSC genome browser, Ensembl genome browser and NCBI Map Viewer.

1.1 Nucleic Acid Sequence Database

There are three well-known large-scale nucleic acid sequence databases. They are GenBank (maintained by National Center for Biotechnology Information, NCBI), EMBL (maintained by The European Bioinformatics Institute, EBI) and DDBJ (maintained by National Institute of Genetics, NIG). In 2005, GenBank, EMBL and DDBJ announced the International Nucleotide Se-

1. Liu Yong, College of Life Sciences and Bioengineering, School of Science, Beijing Jiao Tong University, Beijing, China, 100044.

2. Zhang Lishu, College of Life Sciences and Bioengineering, School of Science, Beijing Jiao Tong University, Beijing, China, 100044.

quence Database Collaboration (INSDC). According to the agreement, these three databases each collects the nucleic acid sequence data published around the world, and shared their sequence data daily, to ensure that a uniform and comprehensive collection of sequence information is available worldwide. That means the sequence information underlying DDBJ, EMBL-Bank, and GenBank is equivalent. Let's begin with GenBank.

1.1.1 GenBank

GenBank is a comprehensive public database of nucleotide sequences and supporting bibliographic and biological annotation. GenBank is built and distributed by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM), located on the campus of the US National Institutes of Health (NIH) in Bethesda, MD, USA. NCBI builds GenBank primarily from the submission of sequence data from authors and from the bulk submission of expressed sequence tag (EST), genome survey sequence (GSS), whole-genome shotgun (WGS) and other high-throughput data from sequencing centers. The US Patent and Trademark Office also contributes sequences from issued patents. GenBank data is available at no cost over the Internet, through FTP and a wide range of Web-based retrieval and analysis services. The website of GenBank is https://www.ncbi.nlm.nih.gov/genbank/, and its homepage is shown in Figure 1.1.

1. The source in GenBank

Virtually all records enter GenBank as direct electronic submissions (http://www.ncbi.nlm.nih.gov/genbank/), with the majority of authors using the BankIt or Sequin programs. Many journals require authors with sequence data to submit the data to a public sequence database as a condition of publication. GenBank staff can usually assign an accession number to a sequence submission within two working days of receipt, and do so at a rate of ~3500 per day. The accession number serves as confirmation that the sequence has been submitted and provides a means for readers of articles in which the sequence is cited to retrieve the data. Direct submissions receive a quality assurance review that includes checks for vector contamination, proper translation of coding regions, correct taxonomy and correct bibliographic citations. A draft of the GenBank record is passed back to the author for review before it enters the database.

Authors may ask that their sequences be kept confidential until the time of publication. Since GenBank policy requires that the deposited sequence data be made public when the sequence or accession number is published, authors are instructed to inform GenBank staff of the publication date of the

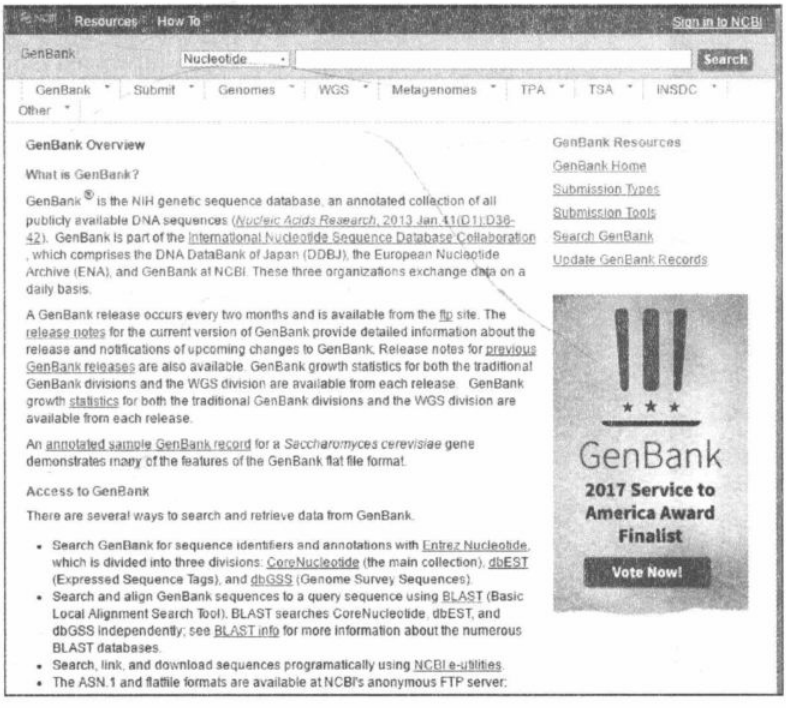


Fig. 1.1 The homepage of GenBank

article in which the sequence is cited in order to ensure a timely release of the data. Although only the submitter is permitted to modify sequence data or annotations, all users are encouraged to report lags in releasing data or possible errors or omissions to GenBank at update@ncbi.nlm.nih.gov.

NCBI works closely with sequencing centers to ensure timely incorporation of bulk data into GenBank for public release. GenBank offers special batch procedures for large scale sequencing groups to facilitate data submission, including the program tbl2asn, described at http://www.ncbi.nlm.nih.gov/genbank/tbl2asn2.html.

2. Organization of the database

(1) Sequence-based taxonomy

Database sequences are classified and can be queried using a comprehensive sequence-based taxonomy (http://www.ncbi.nlm.nih.gov/taxonomy/). Over 300 000 formally described species are represented in GenBank, and the top species in the non-WGS GenBank divisions are listed in Table 1.1.

(2) GenBank divisions

GenBank assigns sequence records to various divisions based either on the source taxonomy or the sequencing strategy used to obtain the data. There

are 12 taxonomic divisions (BCT, ENV, INV, MAM, PHG, PLN, PRI, ROD, SYN, UNA, VRL, VRT) and 5 high-throughput divisions: expressed sequence tags (EST), genome survey sequences (GSS), high-throughput cDNA (HTC), high-throughput genomic (HTG), and the sequence-tagged site (STS). Finally, the PAT division contains records supplied by patent offices, the transcriptome shotgun assembly (TSA) division contains sequences from TSA projects and the WGS division contains sequences from WGS projects.

Table 1.1 Top organisms in GenBank

Organism	Non-WGS base pairs
Homo sapiens	17 575 474 103
Mus musculus	9 993 232 725
Rattus norvegicus	6 525 559 108
Bos taurus	5 391 699 711
Zea mays	5 079 812 801
Sus scrofa	4 894 315 374
Danio rerio	3 128 000 237
Triticum aestivum	1 925 428 081
Solanum lycopersicum	1 764 995 265
Hordeum vulgare	1 617 554 059
Strongylocentrotus purpuratus	1 435 261 003
Macaca mulatta	1 297 237 624
Oryza sativa Japonica Group	1 265 215 013
Xenopus tropicalis	1 249 788 384
Nicotiana tabacum	1 200 025 462
Arabidopsis thaliana	1 165 816 533
Drosophila melanogaster	1 155 228 906
Vitis vinifera	1 071 458 039
Glycine max	1 020 646 789
Pan troglodytes	1 010 316 029

Expressed sequence tags.

ESTs is a major source of sequence records and gene sequences. The top organisms represented in the EST division are *H. sapiens*, *M. musculus*, *S. scrofa*, *Arabidopsis thaliana*, *B. Taurus*, *Z. mays* and *D. rerio*. As part of its daily processing of GenBank EST data, NCBI identifies through BLAST searches all homologies for new EST sequences and incorporates that information into the companion database, dbEST (www.ncbi.nlm.nih.gov/dbEST/index.html). The data in dbEST are processed further to produce the UniGene database (www.ncbi.nlm.nih.gov/ sites/entrez?db=unigene) of millions of gene-oriented sequence clusters.

Sequence-tagged sites, GSSs and ENV.

The sequence-tagged site (STS) division of GenBank (www.ncbi.nlm.nih.gov/dbSTS/index.html) contains anonymous STSs based on genomic sequences as well as gene-based STSs derived from the 3'-ends of genes and ESTs. These STS records usually include mapping information.

The GSS division of GenBank (www.ncbi.nlm.nih.gov/dbGSS/index.html) grew rapidly over the past years. GSS sequences are the products of as many as 80 different experimental techniques, including meta-genomic surveys of sequences arising from biological communities. However, more than one-quarter of all GSS records are single reads from bacterial artificial chromosomes ('BAC-ends') used in a variety of genome sequencing projects. The most highly represented species in the GSS division, including meta-genomic surveys, are marine meta-genome, *M. musculus*, *Z. mays* and *H. sapiens*. The human data have been used (www.ncbi.nlm.nih.gov/projects/genome/clone/) along with the STS records in tiling the BACs for the Human Genome Project.

The ENV division of GenBank accommodates non-WGS sequences obtained via environmental sampling methods in which the source organism is unknown. Many ENV sequences arise from meta-genome samples derived from various animal tissues, such as the gut or skin, or from particular environments, such as freshwater sediment, hot springs or areas of mine drainage. Records in the ENV division contain 'ENV' in the keyword field and use an '/environmental_sample' qualifier in the source feature.

HTG and HTC sequences.

The HTG division of GenBank (www.ncbi.nlm.nih.gov/HTGS/) contains unfinished large-scale genomic records, which are in transition to a finished state. These records are designated as Phase 0–3 depending on the quality of the data, with Phase 3 being the finished state. Upon reaching Phase 3, HTG records are moved into the appropriate organism division of GenBank. The HTC division of GenBank accommodates HTC sequences, which are of draft quality but may contain 5'-UTRs and 3'-UTRs, partial-coding regions and introns. HTC sequences which are finished and of high quality are moved to the appropriate organism division of GenBank.

WGS sequences.

WGS sequences appear in GenBank as a set of WGS contigs, many of them bearing annotations originating from a single sequencing project. These sequences are issued accession numbers consisting of a four-letter project ID, followed by a two-digit version number and a six-digit contig ID. Hence,

the WGS accession number 'AAAA01072744' is assigned to contig number '072744' of the first version of the project 'AAAA'. WGS project contigs for *H. sapiens*, *Pan trodlodytes*, *Macacca mulatta*, *Equus caballus*, *Canis familiaris*, *Drosophila*, *Saccharomyces* and 800 other organisms and environmental samples are available. For a complete list of WGS projects with links to the data, see www.ncbi.nlm.nih.gov/projects/WGS/WGSprojectlist.cgi. Although WGS project sequences may be annotated, many low-coverage genome projects do not contain annotation. Because these sequence projects are ongoing and incomplete, these annotations may not be tracked from one assembly version to the next and should be considered preliminary.

TSA sequences.

In recent years, a growing number of sequencing traces have been deposited in the NCBI Trace Archive (TA). Given the advent of next-generation sequencing technologies, including those from Roche-454 Life Sciences, Illumina Solexa and Applied Biosystems SOLiD, NCBI deployed a Short Read Archive (SRA) in 2007. Neither of these archives is a part of GenBank, but beginning with release 166, GenBank added a new TSA division for TSA sequences, which are shotgun assemblies of sequences deposited in TA, SRA and the EST division of GenBank. TSA records (e.g. EZ000001) have 'TSA' as their keyword and a primary block that provides the base ranges and identifiers of the sequences used in the TSA assembly.

(3) Special record types

Third-party annotation (TPA).

TPA records are sequence annotations published by someone other than the original submitter of the primary sequence record in DDBJ/ENA/GenBank (http://www.ncbi.nlm.nih.gov/genbank/TPA). Each TPA record falls into one of three categories: experimental, in which case there is direct experimental evidence for the existence of the annotated molecule; inferential, in which case the experimental evidence is indirect; and assembly, where the focus is on providing a better assembly of the raw reads. TPA sequences may be created by assembling a number of primary sequences. The format of a TPA record (e.g. BK000016) is similar to that of a conventional GenBank record but includes the label 'TPA exp:', 'TPA inf:' or 'TPA asm:' at the beginning of each definition line as well as corresponding keywords. TPA experimental and inferential records also contain a primary block that provides the base ranges and identifier for the sequences used to build the TPA. TPA sequences are not released to the public until their accession numbers or sequence data and annotation appear in a peer reviewed biological journal. TPA submissions to GenBank may be made using either BankIt or Sequin.