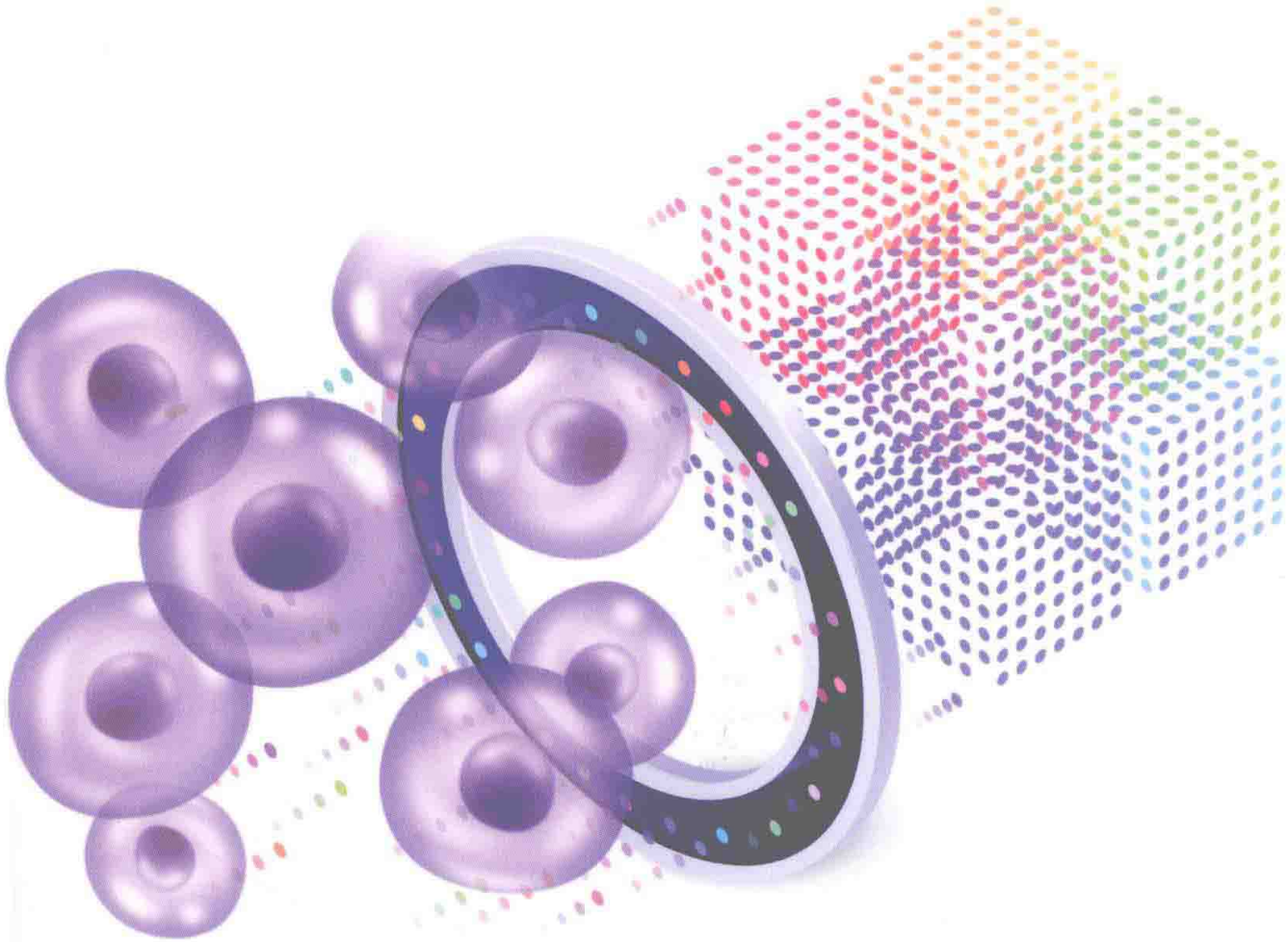
网络生物分子数据库的全面探索

资源与应用

WANGLUO SHENGWU FENZI SHUJUKU
DE QUANMIAN TANSUO
ZIYUAN YU YINGYONG

王宏久 王珍珍 编著



陕西新华出版
陕西科学技术出版社
Shaanxi Science and Technology Press

网络生物分子数据库的全面探索：资源与应用

王宏久　王珍珍　编著

陕西新华出版
陕西科学技术出版社
Shaanxi Science and Technology Press
西安

图书在版编目(CIP)数据

网络生物分子数据库的全面探索：资源与应用 / 王宏久，王珍珍编著. — 西安 ：陕西科学技术出版社，2023.8

ISBN 978-7-5369-8759-3

Ⅰ. ①网… Ⅱ. ①王… ②王… Ⅲ. ①生物信息论-数据库系统-研究 Ⅳ. ①Q811.4②TP311.13

中国国家版本馆 CIP 数据核字(2023)第 121511 号

WANGLUO SHENGWU FENZI SHUJUKU DE QUANMIAN TANSUO: ZIYUAN YU YINGYONG

网络生物分子数据库的全面探索:资源与应用

王宏久　王珍珍　编著

责任编辑　高　曼

封面设计　曾　珂

出 版 者　陕西科学技术出版社

西安市曲江新区登高路 1388 号 陕西新华出版传媒产业大厦 B 座

电话 (029)81205187　传真 (029) 81205155　邮编 710061

http://www.snstp.com

发 行 者　陕西科学技术出版社

电话(029)81205180　81206809

印　　刷　广东虎彩云印刷有限公司

规　　格　710mm×1000mm　16 开

印　　张　12.75

字　　数　220 千字

版　　次　2023 年 8 月第 1 版

2023 年 8 月第 1 次印刷

书　　号　ISBN 978-7-5369-8759-3

定　　价　88.00 元

内容简介

本书是一本介绍生物信息学中常用的数据库资源的详细指南。这本书旨在提供有关基因组、基因、蛋白质、代谢途径和信号通路等多个层次的数据库资源的信息。

本书分为八章，分别介绍了核苷酸序列数据库、基因组数据库、基因信息数据库、基因功能注释数据库、基因组突变数据库、高通量组学数据资源数据库以及生物分子网络数据库等常用数据库资源。每一章节都包含了该数据库的来源、组织结构、数据类型、查询方式和使用示例等方面的详细信息，同时还涵盖了该数据库在生物信息学研究中的应用和未来发展方向等内容。第一章由王珍珍编写，第二章至第八章由王宏久编写。

在核苷酸序列数据库部分，本书详细介绍了 NCBI、DDBJ 和 EBI 数据库，这些数据库包含了来自不同生物界的 DNA 和 RNA 序列，是基因组、转录组和表观基因组等研究的重要数据来源。在基因组数据库部分，着重介绍了 Ensembl、UCSC 和 NCBI Genome 数据库，它们提供了各种生物物种的基因组序列和注释信息，是进行基因组学研究和比较基因组学研究的重要资源。在基因信息数据库部分，介绍了 GeneCards 和 UniGene 两个主要的基因信息数据库，提供了与基因相关的信息。在基因功能注释数据库部分，介绍了 Gene Ontology、KEGG 和 Reactome 基因功能注释数据库，提供了基因的生物学功能、代谢途径和信号通路等方面的信息，有助于理解基因功能和生物过程的调节机制。在基因组突变数据库部分，本书介绍了 COSMIC、dbSNP 和 1000 Genomes 突变数据库，提供了基因突变和人类遗传多态性的信息，是进行人类基因组学和疾病研究的重要资源。在高通量组学数据资源数据库部分，介绍了 GEO、TCGA 和 ArrayExpress 数据库，提供了基因表达、蛋白质组、代谢组和表观基因组等，关键数据资源。最后，在生物分子网络数据库部分，介绍了 STRING、BioGRID 和 HPRD 网络数据

库，提供了蛋白质相互作用、代谢网络和信号通路等方面的信息，有助于理解生物分子之间的相互作用和网络调控机制。

本书全面介绍了生物信息学中常用的数据库资源，有助于读者了解各个数据库的特点、使用方法和应用领域，进而在研究中合理选择和应用这些数据库资源，提高研究效率和质量。

前　言

随着生物学研究的不断深入和技术的飞速发展，越来越多的生物数据被产生和积累。这些数据是研究生命科学的基石，但也给生物信息学的研究者带来了挑战。如何有效地管理、存储和分析这些大量的生物数据，成为生物信息学领域的一个重要问题。

本书介绍了常见的生物数据库，这些数据库包括核苷酸序列数据库、基因组数据库、基因信息数据库、基因功能注释数据库、基因组突变数据库、高通量组学数据资源数据库和生物分子网络数据库。对于每个数据库，我们都提供了详细的介绍，以帮助读者更好地理解这些数据库的功能和应用。

本书旨在帮助生物信息学领域的研究者和生命科学领域的从业者了解和使用这些数据库，以加速生物学研究的进程。我们相信，通过本书的学习和实践，读者将能够更好地利用这些数据库来探索生物学的奥秘，为人类的健康和福祉做出更大的贡献。

祝愿读者学有所获，收获满满！

编者

2023 年 5 月

目　录

第一章　概述

随着生命科学研究的深入和高通量技术的不断发展，生物分子之间的相互作用信息已经成为了解细胞调控机制、发现新的药物靶标、预测疾病风险等方面的重要资源。生物分子网络数据库是存储和管理生物分子相互作用信息的重要工具，通过对这些数据库的综合利用，可以加速生命科学研究的进展，推动基础研究和转化医学的发展。

生物分子网络数据库在生命科学研究中的重要性主要体现在以下几个方面：

(1)揭示生物分子之间的相互作用关系

生物分子网络数据库是存储生物分子之间相互作用信息的集合，可以帮助科学家快速查找和获取分子之间的相互作用信息，从而加深对分子之间相互作用关系的理解。通过分析这些相互作用关系，可以揭示分子在细胞内的功能、调控机制以及参与生理和病理状态的变化和相互作用，有助于揭示复杂生物系统的运作机理。

(2)建立生物分子网络模型

通过将生物分子之间的相互作用建模为网络，可以更直观地展示分子之间的相互关系，有助于研究者更深入地了解分子网络的特性。生物分子网络模型可以用于预测分子之间的相互作用关系、探索分子之间的调控机制和识别生物过程中的关键节点等。这些信息对于研究生物系统的结构和功能至关重要，因此生物分子网络数据库也被广泛用于生物分子网络模型的构建。

(3)探索疾病的发病机制

疾病是生物系统中的复杂现象，多个分子之间的相互作用导致疾病的发生

和发展。利用生物分子网络数据库可以更好地了解疾病发生的机制，对疾病的预测、诊断和治疗提供帮助。例如，通过分析疾病相关的分子网络，可以预测疾病风险、筛选潜在的药物靶标和治疗方法等。

(4)推动新药研发

生物分子网络数据库可以用于筛选潜在的药物靶标，加速新药的研发。利用分子网络数据库可以挖掘出与疾病相关的关键分子，对这些关键分子进行针对性的药物筛选，可以提高药物研发的成功率。

(5)解析生物进化

生物分子网络数据库也可以用于生物进化研究。通过比较不同物种的分子网络，可以了解不同物种的生物功能和生物进化的历史。例如，通过比较不同物种的基因互作网络，可以了解这些物种的遗传特性和进化过程。

(6)支持系统生物学研究

系统生物学研究强调从整体上理解生物系统的结构和功能，网络生物分子数据库可以提供大量的分子相互作用信息，为系统生物学研究提供了丰富的数据基础和理论依据。通过分析分子相互作用网络，可以揭示生物系统的层次结构、功能模块和调控机制等。

(7)支持生物信息学研究

生物信息学是一门快速发展的交叉学科，其研究对象往往是大规模、高维度的生物数据。网络生物分子数据库是生物信息学研究中不可或缺的资源之一，可以提供生物分子的基本信息、相互作用关系等大量数据，有助于研究者对生物数据进行整合、分析和挖掘。

(8)支持医学研究和转化

生物分子网络数据库可以帮助医学研究者更好地了解疾病的发生和发展机制，从而开展更加有效的治疗和预防措施。此外，生物分子网络数据库也可以支持转化医学研究，即将基础研究成果转化为可应用于临床的医疗技术和产品。

随着科技的不断发展和生命科学的深入研究，越来越多的生物分子数据库

被建立起来。这些数据库中包含了大量的生物分子信息，包括基因、蛋白质、代谢产物等。这些生物分子之间的相互作用关系是构成生物系统的基础，也是生命科学研究的重要内容之一。因此，基于网络生物分子数据库的全面探索与应用成为必要的研究方向。编写基于网络生物分子数据库的全面探索与应用的书籍，可以将不同数据库之间的联系和区别进行系统化的总结和介绍。同时，这本书还可以为研究者提供基于不同数据库进行数据挖掘和分析的方法和技巧。这对于加深对分子之间相互作用关系的理解和揭示复杂生物系统的运作机理至关重要。

此外，随着生物分子网络数据库的不断更新和扩展，需要对其进行整合和归纳，以方便研究者进行更高效的数据挖掘和分析。这本书可以帮助研究者更好地利用不同数据库之间的信息和关系，快速获取所需的数据和分析结果。

最后，基于网络生物分子数据库的全面探索与应用不仅对生命科学研究有着重要的意义，也对医学、药物研发、环境保护等领域有着广泛的应用价值。因此，编写这样一本书是有必要的，这有助于促进生命科学研究和相关领域的发展。

第二章　核苷酸序列数据库

国际核苷酸序列数据库合作组织(International Nucleotide Sequence Database Collaboration,简称 INSDC)是由美国国家生物技术信息中心(National Center for Biotechnology Information,NCBI)、欧洲生物信息研究所(European Bioinformatics Institute,EBI)和日本 DNA 数据银行(DNA Data Bank of Japan,DDBJ)三个数据库组成的国际性合作组织。INSDC 的目的是收集、存储和分发全球范围内的核苷酸序列数据,以促进生命科学研究的发展。

INSDC 的成员数据库相互协作,通过共享技术和经验,确保数据的一致性和可靠性。INSDC 成员数据库共享同一种数据格式,即序列数据提交格式(Sequence Submission Format),这使得用户可以通过任意一个成员数据库访问所有其他成员数据库的数据。这也意味着,不管是哪一个国家或地区的研究机构提交了数据,只要符合 INSDC 的数据质量标准,就能在全球范围内被广泛利用。

INSDC 的主要任务是收集和储存全球范围内的核苷酸序列数据,并为研究人员和公众提供免费访问和下载服务。这些数据包括基因组序列、转录组序列、插入片段序列等,可以帮助研究人员了解生物系统的基本结构和功能,识别和研究新的基因和调控元件,开发新的诊断和治疗方法等。

INSDC 成员数据库共同管理着全球范围内的核苷酸序列数据,每个成员数据库都有其独特的功能和优势。NCBI 是 INSDC 成员中最大的数据库,其数据资源丰富,包括基因组序列、蛋白质序列、化合物结构数据、文献数据库等,同时 NCBI 还提供了许多在线工具和数据库,如 BLAST、GenBank、PubMed 等,使得用户可以方便地对数据进行分析和查询。EBI 是欧洲最大的生物信息学中心,其强项是生物信息学的分析和注释。EBI 的主要数据库包括 EMBL、ENA、UniProt 等,其中 ENA 是 INSDC 的成员之一,负责核苷酸序列数据的收集和储存。DDBJ是日本最大的生物信息学中心,其主要任务是管理日本的核苷酸序列数

据,同时也是 INSDC 成员之一,承担着重要的全球性任务。

2.1 NCBI

NCBI(National Center for Biotechnology Information)是美国国家生物技术信息中心,成立于 1988 年,是美国国立卫生研究院下属的一个科研机构。NCBI 旨在提供有关生物信息学和生物技术的科学信息和数据,促进生物医学研究和发展。NCBI 的数据库和工具被广泛应用于生命科学研究和医学领域。

(1)GenBank

GenBank 是 NCBI 维护的全球最大的核苷酸序列数据库之一,包含了大量的 DNA 和 RNA 序列信息。目前 GenBank 数据库中存储了数百万条核苷酸序列数据,其中包括了已知的基因、基因组、EST(表达序列标签)等数据。GenBank的数据来源包括科研机构、学术出版物、专利文献等。研究者可以在 GenBank 中查询和下载序列数据,以便进一步地分析和研究。

(2)PubMed

PubMed 是 NCBI 维护的生命科学文献检索系统,收录了包括生物医学、生命科学、生物技术等领域的众多学术期刊文章和会议论文等文献信息。PubMed 的检索系统支持关键词检索、文献类型筛选、文献来源筛选等多种功能,帮助研究者快速地找到所需的文献资料。

(3)BLAST

BLAST(Basic Local Alignment Search Tool)是 NCBI 开发的一款常用的序列比对工具,可用于比对核苷酸序列和蛋白质序列。BLAST 提供了一种快速、高效的方式来比对和分析不同来源的生物序列数据,支持多种比对方式和参数设置,适用于各种生物信息学分析任务。

(4)dbSNP

dbSNP 是 NCBI 维护的一个人类单核苷酸多态性(SNP)数据库,包含了大量的人类基因组中的 SNP 信息。dbSNP 数据库可以帮助研究者了解不同个体之间的基因差异和表达差异,对于人类基因组的研究和临床应用具有重要

意义。

(5) RefSeq

RefSeq 是 NCBI 维护的一个基因序列和注释数据库，包含了多种生物物种的基因序列信息和注释信息。RefSeq 数据库中的基因序列和注释信息都是由 NCBI 的专家团队进行严格的筛选和审核的，可以帮助研究者快速了解和分析基因信息，为基因功能和调控机制的研究提供可靠的基础数据。

(6) ClinVar

ClinVar 是 NCBI 维护的一个与人类疾病相关的遗传变异数据库，收集了大量的致病和良性变异信息。这些变异信息是由临床实验室、医疗保健机构和个人研究者提交的，可以帮助研究者了解基因变异与人类疾病的关系，为疾病的预测、诊断和治疗提供支持。

(7) Gene

Gene 是 NCBI 维护的一个基因信息数据库，包含了各种生物物种的基因信息和注释信息。Gene 数据库中的基因信息和注释信息都是由 NCBI 的专家团队进行严格的筛选和审核的，可以帮助研究者快速了解和分析基因信息，为基因功能和调控机制的研究提供可靠的基础数据。

(8) PubChem

PubChem 是 NCBI 维护的一个化合物信息数据库，包含了大量的化合物结构信息和相关的生物活性信息。这些化合物信息和生物活性信息是由 NCBI 的专家团队从各种文献和实验数据中收集整理而来的，可以帮助研究者了解不同化合物的结构和生物活性特性，为新药物的研发和设计提供支持。

(9) SRA

SRA 是 NCBI 维护的一个高通量测序数据存储库，包含了大量的各种生物物种的测序数据。这些测序数据是由各种研究机构和个人研究者提交的，可以帮助研究者快速获得各种生物物种的测序数据，为基因组学、转录组学和表观遗传学的研究提供支持。

2.1.1 GenBank

GenBank 是一个由美国国家生物技术信息中心(National Center for Biotechnology Information, NCBI)维护的数据库,包含了全球各地提交的基因序列和相关生物信息,是全球最大的基因序列数据库之一。该数据库的目的是为全球科学家提供一个开放的平台,方便他们分享基因序列、注释信息、序列分析工具等方面的知识。GenBank 数据库中的数据包含了多种类型的生物序列信息,包括基因序列、mRNA 序列、蛋白质序列、基因组序列等。这些数据可以用于生物学研究、基因工程、生物医学研究、进化生物学研究等领域。

在 GenBank 数据库中,每个序列都有一个唯一的标识符,称为 GenBank Accession Number,它可以用于快速访问特定的序列。此外,每个序列都附有一些元数据,例如序列的来源、作者、发布日期、实验方法等,这些元数据可以帮助研究人员更好地理解序列信息和数据的来源。GenBank 数据库是一个公共数据库,任何人都可以免费访问和使用其中的数据。NCBI 还提供了一系列的工具和服务,帮助研究人员使用 GenBank 中的数据,例如 BLAST(Basic Local Alignment Search Tool)搜索工具,用于比对新序列与 GenBank 中已知序列的相似性;GenBank 数据下载服务,可用于下载特定序列、特定类型的序列或整个数据库的数据等。

总之,GenBank 数据库是一个极为重要的基因序列数据库,为全球生物学和医学研究提供了重要的数据和资源,对于加速科学研究和推进生物医学领域的进展具有重要的意义。

2.1.2 GenBank 数据库存储数据的类型

GenBank 包含了各种类型的生物序列信息,包括基因序列、mRNA 序列、蛋白质序列、基因组序列等。这些序列数据在 GenBank 数据库中被存储为文本格式,可通过各种工具和方法进行访问和分析。下面将详细介绍 GenBank 数据库存储的数据类型。

(1)基因序列(Gene Sequence)

基因序列是指 DNA 中编码蛋白质的基因区域,通常由 ATCG 四个核苷酸组成。在 GenBank 数据库中,基因序列通常以 FASTA 格式存储,每个序列都由

一个唯一的标识符(GenBank Accession Number)和一个描述性标题组成。此外,基因序列还包括了序列的长度、来源、作者、参考文献等元数据。

(2)mRNA 序列(mRNA Sequence)

mRNA 是基因转录过程中产生的一种 RNA 分子,其序列与编码该基因的 DNA 序列是一一对应的。在 GenBank 数据库中,mRNA 序列也通常以 FASTA 格式存储,每个序列都由一个唯一的标识符和描述性标题组成。与基因序列不同的是,mRNA 序列通常包含了开放阅读框(Open Reading Frame, ORF)信息,用于指示该序列中可能存在的蛋白质编码序列。

(3)蛋白质序列(Protein Sequence)

蛋白质序列是指由氨基酸组成的蛋白质链,在 GenBank 数据库中通常以 FASTA 格式存储。每个蛋白质序列都有一个唯一的标识符和一个描述性标题,包括序列的长度、来源、作者、参考文献等元数据。此外,蛋白质序列通常还包括了一些附加信息,例如分子量、等电点、亲水性、氨基酸序列中的保守区域等信息。

(4)基因组序列(Genome Sequence)

基因组序列是指一个生物个体的所有基因序列组成的序列,通常包括了大量的非编码序列。在 GenBank 数据库中,基因组序列通常以 FASTA 格式存储,每个序列都有一个唯一的标识符和描述性标题,包括了序列的长度、来源、作者、参考文献等元数据。与基因序列、mRNA 序列和蛋白质序列不同的是,基因组序列通常会包含一些附加信息,例如基因组大小、GC 含量、基因密度、反转录转座子数量等信息。

(5)核糖体 RNA 序列(Ribosomal RNA Sequence)

核糖体 RNA(rRNA)是一种 RNA 分子,存在于所有细胞中,并在蛋白质合成过程中起着关键作用。在 GenBank 数据库中,核糖体 RNA 序列也通常以 FASTA 格式存储,每个序列都有一个唯一的标识符和描述性标题,包括序列的长度、来源、作者、参考文献等元数据。此外,核糖体 RNA 序列还会包含 rRNA 结构的注释信息,如 16S、18S、28S 等。

(6)miRNA 序列(MicroRNA Sequence)

miRNA 是一种非编码 RNA 分子,其长度通常为 20 ~ 24 个核苷酸,可以参与调控基因表达。在 GenBank 数据库中,miRNA 序列通常以 FASTA 格式存储,每个序列都有一个唯一的标识符和描述性标题,包括序列的长度、来源、作者、参考文献等元数据。此外,miRNA 序列还会包含 miRNA 结构的注释信息,如成熟 miRNA、前体 miRNA 等。

(7)EST 序列(Expressed Sequence Tag Sequence)

EST 是由转录后修饰(如剪切)的 mRNA 产生的短序列,通常长度为 200 ~ 500 个核苷酸。在 GenBank 数据库中,EST 序列通常以 FASTA 格式存储,每个序列都有一个唯一的标识符和描述性标题,包括序列的长度、来源、作者、参考文献等元数据。EST 序列可以用于鉴定基因、构建转录组、寻找新的基因等。

(8)SSR 序列(Simple Sequence Repeat Sequence)

SSR 是指由重复单元(通常是 1 ~ 6 个核苷酸)组成的短序列,也被称为微卫星(Microsatellite)。在 GenBank 数据库中,SSR 序列通常以 FASTA 格式存储,每个序列都有一个唯一的标识符和描述性标题,包括序列的长度、来源、作者、参考文献等元数据。SSR 序列可以用于遗传多样性研究、物种识别等方面。

GenBank 数据库存储的数据类型非常丰富,包括基因序列、mRNA 序列、蛋白质序列、基因组序列、核糖体 RNA 序列、miRNA 序列、EST 序列、SSR 序列等。每种数据类型都有其独特的特点和应用场景,为生物学研究提供了极大的便利。同时,GenBank 数据库的持续更新和完善,也为生物学研究提供了强大的数据支持和资源共享平台。

2.1.3 GenBank 数据库存储数据的格式

GenBank 数据库存储的数据以一种特定的格式进行组织和表示,以便于数据的管理、存储、检索和共享。这里将介绍 GenBank 数据库存储数据的格式,包括序列记录的各个部分、格式标记的含义以及序列的版本控制。

(1)序列记录的各个部分

GenBank 数据库存储的每个序列记录都包含若干个部分,其中最重要的部