“十二五”普通高等教育本科国家级规划教材配套教材
国家卫生和计划生育委员会“十二五”规划教材配套教材
全国高等医药教材建设研究会“十二五”规划教材配套教材

全国高等学校配套教材

供8年制及7年制（“5+3”一体化）临床医学等专业用

生物信息学学习指导及习题集

第2版

主　编　李　霞　李亦学
副主编　张　岩　徐良德

人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

“十二五”普通高等教育本科国家级规划教材配套教材
国家卫生和计划生育委员会“十二五”规划教材配套教材
全国高等医药教材建设研究会“十二五”规划教材配套教材
全国高等学校配套教材

供8年制及7年制(“5+3”一体化)临床医学等专业用

生物信息学学习指导及习题集

第2版

主　编　李　霞　李亦学

副主编　张　岩　徐良德

编　者　(以姓氏笔画排序)

王　宏(哈尔滨医科大学)　　李永生(哈尔滨医科大学)
王　举(天津医科大学)　　肖　云(哈尔滨医科大学)
宁尚伟(哈尔滨医科大学)　　邹凌云(第三军医大学)
许丽艳(汕头大学)　　沈百荣(苏州大学)
许超汉(哈尔滨医科大学)　　张　岩(哈尔滨医科大学)
朱　浩(南方医科大学)　　张云鹏(哈尔滨医科大学)
刘洪波(哈尔滨医科大学)　　张绍军(哈尔滨医科大学)
李　瑛(吉林大学)　　徐　沁(上海交通大学)
李　霞(哈尔滨医科大学)　　徐　娟(哈尔滨医科大学)
李　曦(中南大学)　　徐良德(哈尔滨医科大学)
李冬果(首都医科大学)　　崔庆华(北京大学)
李亦学(同济大学)　　智　慧(哈尔滨医科大学)
李学荣(中山大学)

人民卫生出版社

图书在版编目（CIP）数据

生物信息学学习指导及习题集 / 李霞，李亦学主编 . —2 版 . —北京：人民卫生出版社，2016

ISBN 978-7-117-22238-9

Ⅰ. ①生… Ⅱ. ①李… ②李… Ⅲ. ①生物信息论 - 医学院校 - 教学参考资料 Ⅳ. ①Q811.4

中国版本图书馆 CIP 数据核字（2016）第 045455 号

生物信息学学习指导及习题集

第 2 版

主　　编：李　霞　李亦学
出版发行：人民卫生出版社（中继线 010-59780011）
地　　址：北京市朝阳区潘家园南里 19 号
邮　　编：100021
E - mail：pmph @ pmph.com
购书热线：010-59787592　010-59787584　010-65264830
印　　刷：北京市艺辉印刷有限公司
经　　销：新华书店
开　　本：787 × 1092　1/16　　**印张：**9
字　　数：230 千字
版　　次：2011 年 6 月第 1 版　　2016 年 2 月第 2 版
2016 年 2 月第 2 版第 1 次印刷（总第 2 次印刷）
标准书号：ISBN 978-7-117-22238-9/R · 22239
定　　价：24.00 元
打击盗版举报电话：010-59787491　E-mail：WQ @ pmph.com
（凡属印装质量问题请与本社市场营销中心联系退换）

前　言

生物信息学作为现代生命医学领域最前沿的交叉学科，利用计算机科学与信息技术，进行海量生物医药数据信息的提取、整合与分析，高通量、系统性开展生物医学大数据研究，指导生物医学科技进展与产业开发。伴随着新一代测序技术的深入发展，“大数据”、“组学”研究成为当今生命科学领域的热点。经全国高等医药教材建设研究会及国家卫生计生委论证后，决定启动《生物信息学》第2版及配套教材编写。本套教材坚持“三基”“五性”原则，并力求在内容和形式上有所创新，主要培养长学制学生运用生物信息学方法解决临床问题、进行科研设计的能力。

《生物信息学学习指导及习题集》第2版是“十二五”普通高等教育本科国家级规划教材《生物信息学》第2版的配套教材。各章与主教材保持一致，合并第一版“双序列比对”、“多序列比对”两章为“序列比对”；合并第1版“序列特征分析”、“表达序列分析”两章为“序列特征分析”；合并第1版“蛋白质分析与蛋白质组学”、“蛋白质结构分析”两章为“蛋白质组与蛋白质结构分析”；新增非编码RNA与复杂疾病的生物信息学研究、新一代测序、“组学”研究等前沿热点。每一章包含学习目标、知识要点、复习思考题和参考答案与题解，内容全面，重点突出。旨在于帮助学生在理论学习基础上，增强实践运用能力，自测学习效果。本书适合学生同步复习、准备课程考试及准备执业医师考试等。

本书是在第1版基础上修订完成的，修订过程中借鉴了第1版作者的论著和成果，在此致以谢意！第2版编委是来自于全国15所高校相关研究方向的专家学者和青年教师，每一章都凝聚了独特的学术思想、研究心得和研究成果。他们在百忙之中精心组织素材，斟字酌句编写，付出了大量心血。在此对全体编委的无私奉献深表谢意！

第2版学习指导及习题集得到国家863项目、973课题和黑龙江省生物医学工程重点学科经费资助，特此鸣谢！

本书修订过程中，尽管学者们努力跟踪学科新发展、新技术，并尽力纳入到教材中来，以保持先进性和实用性，但时间紧迫，直至完稿，仍觉得有许多不足之处，希望学术同仁不吝赐教，以便再版时改正。

李　霞　李亦学

2015年6月

目　录

第一章　生物序列资源

【学习目标】

1. **掌握**　获取 DNA、RNA 和蛋白质序列数据资源的方法和工具。

2. **熟悉**　NCBI 数据库、UCSC 基因组浏览器、EMBL-EBI 数据库及重要的 RNA 数据库。

3. **了解**　核酸、蛋白质数据库发展历史；生物信息学数据库发展趋势及其对生物医药科学领域的影响。

【知识要点】

一、引言

随着各种高通量技术，特别是大分子序列测序技术的快速发展，为适应众多物种海量基因组序列数据存储、维护的需要，形成了数以千计的生物信息学数据库和网络分析平台。理清和认识世界各重要数据平台维护的 DNA、RNA、蛋白质序列数据资源和分析工具，掌握数据存储、整理、分析的一般规律，将为生命医学研究提供丰富的知识借鉴，为基因组、转录组、蛋白质组等功能基因组学、生物医药研究和科技转化提供重大的资源和技术支持。

二、NCBI 数据库与数据资源

(一) NCBI 序列数据库概述

NCBI 为大规模生物医药数据存储、分类与管理，生物分子序列、结构与功能分析，分子生物软件开发、发布与维护，生物医学文献收集与整理，全球范围数据提交与专家注释于一体的世界最大规模的生物医学信息和技术资源数据库。

1992 年，NCBI 建立了 GenBank 核酸序列数据库，并与 EMBL、DDBJ 实现数据资源的交换和共享。除 GenBank 外，NCBI 为医学和生命科学研发提供多种数据信息支持，包括生物医学文献公共检索与分析平台(PUBMED)、人类孟德尔遗传在线(OMIM)、3D 蛋白结构分子模型数据库(MMDB)、特有人类基因序列集(UniGene)、全物种基因组图谱、癌症基因组剖析计划(CGAP)等。为用户提供友好的信息查询和批量下载方式，并向用户提供 BLAST 序列相似性比对、ORF Finder 开放读码框搜索等软件工具，为功能基因组研究提供了便利的条件。

(二) NCBI 中的重要子库介绍

NCBI 整合了科学文献(PUBMED)、序列数据、基因组、结构数据、表达数据、种群研究数据集和分类学信息等各个子库，形成紧密链接的系统和高效集约的查询平台。重要的数据子库包括：

1. **GenBank 与 RefSeq**　GenBank 收录的核酸序列数据根据其不同的研究属性，分属于 Nucleotide、GSS 和 EST 三个子库。Nucleotide 收录绝大多数常规的核酸序列；GSS 收录测序起始阶段用来进行序列或基因示踪、重复序列或基因数量预判等的各种短读长(reads)序列；EST

收录 cDNA 及 cDNA 特征序列信息。NCBI 在 GenBank 数据基础上针对每个基因不同的数据类型提取一个可靠的注释条目作为参考条目，组成 RefSeq 数据库。

2. Gene 基因数据库收录全部已测序物种的基因注释信息，包括基因的名称、染色体定位、基因序列、基因功能和相关文献信息等，是目前最权威的基因注解数据库。Gene 数据库标识符为 Entrez gene ID。

3. Genome NCBI 收录了超过 1000 种已经完成测序的生物体全部基因组序列和定位数据，及正在进行测序的物种阶段性发布的基因组信息。Genome 涉及的物种包括所有的生物领域：细菌、古细菌、真核生物，以及许多病毒、噬菌体、类病毒、质粒和含遗传物质的细胞器。

4. **遗传多态数据库** dbSNP、dbVar、dbGaP 和 ClinVar 四个子库涉及 DNA 多态或变异信息。dbSNP 收录了所有物种中发现的短序列多态和突变信息，包括单核苷酸多态、微卫星、小片断插入 / 删除多态等定位、侧翼序列和功能、频率信息；dbVar 收录较大规模的基因组变异，包括大片断的插入、缺失、易位、倒置和拷贝数多态等信息资源；dbGaP 收录大量以遗传多态为分子标记物的基因型和表型（疾病）关联性研究数据；ClinVar 收录临床中发现或报道的有证据支持的与人类疾病或健康状态有关的变异位点。

5. GEO GEO 数据库接收和管理各研究机构提交的基因芯片或测序技术获得的不同生理、病理状态个体或细胞系基因（包括非编码基因）表达数据。其数据类型包括：GPL 是特定的芯片或测序平台类型；GSM 参与基因表达测序的样本或个体信息；GSE 是一组相关样本实验测定的基因表达数据谱；GDS 是由 GEO 数据库维护团队综合多组实验产生的整合的表达数据集。

6. **蛋白质数据库** NCBI Protein 数据库收录来源于 GenPept、RefSeq、Swiss-Prot、PIR、PRF 及 PDB 等蛋白质数据资源的蛋白质序列和注释数据。Protein Cluster 数据库提供存在一定联系的蛋白质集合信息。Structure 数据库是由蛋白质三维结构数据库 PDB 衍生而来的大分子模建数据库，提供蛋白质三维结构信息及相关的可视化和结构比对工具。

7. Epigenomics 表观基因组数据查询和浏览相结合的数据库。提供 DNA 甲基化、组蛋白修饰等表观遗传学数据集下载，基因序列、表观遗传状态的定位比较和可视化等。

8. Unigene Unigene 数据库针对每一个基因建立一个独立的数据体系，分别将不同来源的基因序列、蛋白质相似性、基因表达、染色体定位、cDNA 序列、mRNA 序列、EST 序列等进行罗列和比较，旨在为研究者提供全面、丰富的信息资源，更好地对基因的功能和注释信息的可靠性进行梳理。

9. **与生物医学相关的重要数据库** OMIM 数据库阐述遗传变异介导的疾病相关基因情况，及变异介导的基因参与不同疾病情况。dbMHC 收录人类主要组织相容性复合体数据及其相关的分子标记物信息。

10. **NCBI 提供的重要支持工具** BLSAT 是序列相似性搜索程序，检索速度快，有助于识别基因和基因特征。Primer-BLAST 可用于多方面生物医学研究过程的核酸引物设计。

三、UCSC 基因组浏览器与数据资源

（一）UCSC 概述

UCSC 是应用广泛的网络存储、分析和基因组可视化工具，可在任何尺度快速查询和显示基因组内容，同时伴有一系列序列比对注释“通道”。UCSC 基因组浏览器包含大量收集的基因组参考序列和拼接数据信息，提供了多种便利的基因组查询和注释工具。

(二) UCSC 基因组浏览器

UCSC 基因组浏览器能够实现已知人类基因或疾病相关基因检索，多物种基因组中同源基因显示，定位修复酶、STS 标签以及 BAC 末端配对，参考基因组中的 SNPs 和其他变异分布，通道元件逐个碱基比对图谱上的基因组详细信息，微阵列芯片基因表达数据，多物种 mRNA 和 ESTs 与用户拼接序列的对应图谱显示，及生成适用于学术出版的基因组注释图像等。

(三) UCSC 中的数据资源和常用工具

1. UCSC 中的数据资源　UCSC 收录了来自全世界研究机构提供的包括人类基因组在内的 48 种哺乳动物(mammal)、19 种其他脊椎动物(vertibrate)、3 种后口动物(deuterostome)、20 种昆虫(insect)、线虫(nematode)等众多动物，及病毒(virus)、酵母等微生物全基因组数据。

2. view 中的图像输出和 DNA 序列检索功能

(1) 基因组浏览器图像输出：UCSC 支持生成适于文献出版和打印的高质量图像。

(2) DNA 序列检索：导航栏 view 按钮中的 DNA 选项能够实现浏览器中显示的染色体区段的 DNA 序列提取和下载。

3. Table Browser 下载数据　Table Browser 工具可以完整地获取 UCSC 的后台数据：①获取 DNA 序列、全基因组、指定的坐标区段或一组注册号的隐含注释通道数据；②应用过滤器设置约束条件，确定输出结果类型和格式；③生成在基因组浏览器中图形显示的查询通道；实现数据结构和任意格式 SQL 检索；④整合多表格或查询通道交叉或统一检索，以及生成单一的数据输出集；⑤显示指定数据集碱基统计计算结果；⑥显示表格概要并且查看数据库中所有与查询表格相关的其他表格清单；⑦将输出数据整理成几种不同的格式用于电子表格、数据库或查询通道等不同用途。

4. BLAT 序列比对工具　BLAT 是一种常用序列比对工具，支持目标序列与参考基因组进行 DNA 或蛋白序列的比对。BLAT 进行 DNA 比对时，可快速寻找 95% 或更高的匹配度的 40 碱基以上相似序列。BLAT 进行蛋白序列比对模式时，快速搜索比对长度在 20 氨基酸以上、相似性超过 80% 的序列。

四、EMBL-EBI 数据库与数据资源

(一) EMBL-EBI 数据库概况

目前 EBI 是协调搜集和传播生物学数据的欧洲节点，维护着世界上最广泛的生物分子数据资源。包括：EMBL-Bank(DNA 和 RNA 序列)、Ensembl(基因组)、ArrayExpress(微阵列基因表达)、UniProt(蛋白质序列和注释)、interPro(蛋白质家族、结构域和基序)、Reactome(细胞通路)和 ChEBI(小分子)等。EMBL-EBI 数据资源遵循严格的规模化管理：①可访问性，所有数据和工具完全开放访问；②兼容性，数据达到世界最高层次的标准化规范，有利于推动数据共享；③数据集综合性，与各研究、出版机构和各大数据库达到数据提交、共享协议，保障数据来源和交叉引用；④便携性，EBI 所有数据库均可下载，全部软件系统可以下载并本地安装；⑤保证质量，EBI 具有专家注释系统，大量数据资源通过生物医学专家注释保障数据质量。

(二) EMBL 基因组和核酸序列资源

1. Ensembl 基因组序列数据资源　EMBL-EBI 中有 Ensembl 和 Ensembl Genomes 基因组序列资源数据库。Ensembl 数据库提供高质量、综合注释的脊椎动物基因组数据，Ensembl Genomes 数据库提供非脊椎动物全基因组数据。

EMBL 目前收录了 72 个物种基因组数据信息，并在主页中提供 ENCODE 数据访问、基因

表达的组织差异性分析、基因序列提取、变异位点效应预测、基因多态性定位、跨物种基因比较、用户数据分析、疾病与表型分析 8 个功能研究模块。

2. EMBL ENA 核酸测序数据资源 EMBL-EBI 维护的欧洲核苷酸数据库 ENA 提供世界范围的核酸测序原始数据、序列拼装和功能注释信息的维护和下载，并记录和存储数据集测序全过程的技术应用情况。ENA 数据包括机构或个人提交的原始数据，序列拼接和小规模测序注释数据，欧洲各大测序中心提供的测序数据，国际核酸序列数据库协作组织（INSDC）的合作伙伴的定期交换数据等。向 ENA 或其 INSDC 合作伙伴提供核苷酸序列数据，已成为学术界发表研究成果必不可少的步骤，ENA 与学术机构以及出版商合作为发表文献提供序列提交系统和数据检索工具。

（三）UniProt 蛋白质数据资源

UniProt 是目前世界上最权威的蛋白质序列与注释数据综合资源数据库。UniProt 数据库包括 UniProt Knowledgebase（UniProtKB）、UniProt Reference Clusters（UniRef）、UniProt Archive（UniParc）三个主要部分，以及用于专门存放元基因组和环境基因组数据信息。

1. UniProtKB UniProtKB 是 UniProt 的核心资源，收录非冗余的、高质量的专家手工注释数据。收录的蛋白质序列信息包括：①DDBJ/ENA/GenBank 来源的编码序列（CDS）翻译；②PDB 中存储了结构信息的蛋白质序列；③Ensembl 和 RefSeq 提供的序列；④直接提交到 UniProtKB 或文献检索到的氨基酸序列。

2. UniProt 中的其他数据资源 UniRef 根据蛋白质序列在不同物种中的序列相似性进行分簇（cluster），它包括 3 个子库：UniRef100、UniRef90 和 UniRef50，分别表示跨物种 100%、90% 以上和 50% 以上相似性的蛋白质序列集合。

UniParc 是蛋白质序列仓库，收集全部公开发布或文献发表的蛋白质序列。不考虑物种、功能等信息，存储蛋白质序列时采用纯序列文本格式，只要是序列 100% 相同，均列为同一条目。

（四）Biomart 数据检索平台

Biomart 是由 EBI 开发和维护的最经典的生物数据库检索、处理和下载平台，可以便捷的将储存在不同数据库中的基因、蛋白等序列和注释信息进行整合，查询不同数据库来源的基因 ID、基因组定位、表达、结构等信息，进行不同数据资源条目代码的转换、功能富集，并可以批量获取相关数据，方便地得到一个物种全部基因组或局部区域的核酸、蛋白序列及各种注释信息。

五、重要的非编码基因数据库

伴随新一代测序技术的广泛应用，从序列层面探索非编码基因功能和潜在的作用方式至关重要。除 GenBank、Ensembl 等收录了 ncRNA 的相关信息外，ENCODE 和 mirBase 等数据库存储 ncRNA 序列和功能信息。

（一）ENCODE 数据库与数据资源

ENCODE 被认为是“人类基因组计划”之后国际科学界在基因研究领域取得的又一重大进展。人类基因组中约 80% 的 DNA 序列是具有某种特定功能的，ENCODE 为深入研究基因组作用模式提供了第一手资料。

（二）microRNA 数据资源 miRBase

miRBase 是存储、维护和命名微小 RNA 的主要数据库，主要数据资源为 microRNA 序列和

注释信息；提供 miRNA 前体和成熟序列下载；允许用户使用关键词或序列检索数据库，通过关联链接到 miRNA 的原始参考文献，分析基因组中的定位和挖掘 miRNA 序列间的关系。

【复习思考题】

一、名词解释

1. ENCODE
2. miRBase

二、问答题

1. 生物数据库根据其存储的数据类型可以分为几类？
2. DDBJ 和另外哪两个数据库并称为世界三大核酸数据库，并通过网络查询 DDBJ 数据库的信息存储情况？
3. Entrez Gene 数据库从哪些方面对基因进行注释？
4. dbSNP 数据库维护的数据类型有哪些？
5. UCSC 基因组浏览器显示的数据资源如何以可出版的图片形式输出？
6. 如何利用 UCSC 模块实现序列数据的批量下载？
7. EMBL-EBI 维护数据的规模化标准有哪些？
8. 如何利用 Ensembl-BioMart 平台实现核酸序列数据的查询和下载？
9. 简述 UniProt 数据的基本构建。
10. 试列举 2 个非编码基因序列维护数据库。

【答案与题解】

一、名词解释

1. ENCODE　ENCODE 全称为 DNA 元件百科全书计划，对基因组功能元件进行解析。ENCODE 被认为是“人类基因组计划”之后国际科学界在基因研究领域取得的又一重大进展。人类基因组中约 80% 的 DNA 序列是具有某种特定功能的，ENCODE 为深入研究基因组作用模式提供了第一手资料。

2. miRBase　miRBase 是存储、维护和命名微小 RNA（microRNA）的主要数据库，主要数据资源为 microRNA 序列和注释信息；为用户提供 miRNA 前体和成熟序列下载；允许用户使用关键词或序列检索数据库，通过关联链接到 miRNA 的原始参考文献，分析基因组中的定位和挖掘 miRNA 序列间的关系。

二、问答题

1. 答：从数据库功能和收录数据类型层面进行细化，主要包括：DNA 序列（DNA sequence）、RNA 序列（RNA sequence）、微阵列数据和基因表达（microarray data and gene expression）、蛋白质序列（protein sequence）、分子结构（structure）、蛋白质组学与蛋白质互作（proteomics and interaction）、代谢与信号通路（metabolic and signaling pathways）、人类基因与疾病（human genes and diseases）、生理与病理（physiology and pathology）、药物与药物靶标（drug and drug targets）、

细胞器与细胞生物学(organelle and cell biology)、人类及其他脊椎动物基因组(human and other vertebrate genomes)、非脊椎动物基因组(non-vertebrate genomes)、植物基因组(plant genomes),以及其他分子生物学数据库等。

2. 答:1992 年,NCBI 建立了 GenBank 核酸序列数据库,将美国专利商标局存储的专利序列并入 GenBank 管理,并与 EMBL、DDBJ(与 GenBank 并称世界三大生物序列信息数据库)实现数据资源的交换和共享。此后,NCBI 成为世界范围公认的序列相关知识产权申报或研究成果发表时,数据信息指定提交和保管机构,存储数据规模急剧增长。

3. 答:Entrez gene 数据库收录全部已测序物种的基因注释信息,包括基因的名称、染色体定位、基因序列和产物编(mRNA、蛋白质)情况、基因功能和相关文献信息等,并与 GenBank、OMIM、遗传多态数据库(如 dbSNP、dbVar)等 NCBI 子库,及 KEGG、Gene Ontology 等外源性数据库进行交叉引用。基因数据库是目前最权威的基因注解数据库。Gene 数据库标识符(即 Entrez gene ID)依据基因的发现顺序由 1 到多位数字组成,如 IL10 基因的标识符为 3586。以 IL10 基因为例介绍 Gene 数据的注释信息,其注释内容包括基因概况、基因组结构、基因组定位、参考书目、表现型、基因变异、HIV-1 互作、通路注释、互作、基因功能、同源性、编码蛋白质情况、序列信息,及交叉引用链接。

4. 答:dbSNP 收录了所有物种中发现的短序列多态和突变信息,包括单核苷酸多态(single nucleotide polymorphysm,SNP)、微卫星(microsatlite)、小片断插入 / 删除多态(in/del)等定位、侧翼序列和功能、频率信息,收录的 SNP 条目一般以“rs+ 数字”的形式表示。

5. 答:UCSC 基因组浏览器支持生成适于文献出版和打印的高质量图像。打印前用户可以在序列碱基栏左端标签处点击鼠标右键选择配置管理(configure ruler)按钮,打开设置页面,可在标题栏中添加通道输出图片标题,还可以选择增加组合名称和染色体位置方式将标题加入到通道中。鼠标左键拖拽各通道对应的灰色工具条还可以根据输出需要改变各通道的位置。用户完成通道图像配置后,点击导航栏中 view 按钮下拉菜单中的 PDF/PS 选项,选择所需的文件输出格式保存图像。

6. 答:如需批量获取多段 DNA 序列,可使用“Table Browser”工具下载数据。基因组浏览器中显示的注释信息具备后台数据的支撑,这些数据保存在一个或多个数据表格中,使用 Table Browser(表格浏览器)工具可以完整地获取 UCSC 的后台数据。点击 UCSC 首页导航栏中的 Tables 按钮或基因组浏览器上部蓝色导航栏中 Tools 中的 Table Browser 选项即可打开表格浏览器。使用表格浏览器可以:①获取 DNA 序列、全基因组、指定的坐标区段或一组注册号的隐含注释通道数据;②应用过滤器设置约束条件,确定输出结果类型和格式;③生成在基因组浏览器中图形显示的查询通道;实现数据结构和任意格式 SQL 检索;④整合多表格或查询通道交叉或统一检索,以及生成单一的数据输出集;⑤显示指定数据集碱基统计计算结果;⑥显示表格概要并且查看数据库中所有与查询表格相关的其他表格清单;⑦将输出数据整理成几种不同的格式用于电子表格、数据库或查询通道等不同用途。

7. 答:EMBL-EBI 数据资源遵循严格的规模化管理:①可访问性,所有数据和工具完全开放访问;②兼容性,数据达到世界最高层次的标准化规范,有利于推动数据共享;③数据集综合性,与各研究、出版机构和各大数据库达到数据提交、共享协议,保障数据来源和交叉引用;④便携性,EBI 所有数据库均可下载,全部软件系统可以下载并本地安装;⑤保证质量,EBI 具有专家注释系统,大量数据资源通过生物医学专家注释保障数据质量。

8. 答:BioMart 可实现数据检索和下载功能:①选择数据集,打开 BioMart 后首先出现数据

集选择页面，右侧的第一个下拉框中有 Ensembl 第 78 发布版本的 Genes、Variation、Regulation 选项，及第 58 版本的 Vega 和 EBI 的 PRIDE 选项；选择 Genes 选项，并在下面的物种下拉框中选择人类基因组 Homo sapiens genes（GRCh38）版本，选择后的信息出现在左侧工具栏中；②数据筛选，点击左侧工具栏的 Filters 选项，对要下载的数据类型进行限定，分别可以从染色体定位、指定基因名、表型相关、基因功能类、直系同源、蛋白质结构域或家族性、变异类型 7 个方面进行限定；③数据类别和属性设定，点击左侧工具栏的 Attributes 选项进行下载数据类别和属性的设定，可以从特征（features）、变异（variation）、结构（structures）、序列（sequences）、同源（homologs）五种数据类型中选择一种；④结果预览与输出，点击左侧工具栏上方的 Count 按钮可以查看本次检索的数据量，点击 Results 按钮即可预览查询结果。点击结果预览页面右上角的 GO 按钮，检索到的序列数据将以 FASTA 格式下载到本地电脑。

9. 答：UniProt 数据库包括 UniProt Knowledgebase（UniProtKB）、UniProt Reference Clusters（UniRef）、UniProt Archive（UniParc）三个主要部分，及用于专门存放元基因组和环境基因组数据信息的 UniProt Metagenomic 和 Environmental Sequences（UniMES）数据库。

10. 答：(1) ENCODE 数据库与数据资源：ENCODE 被认为是“人类基因组计划”之后国际科学界在基因研究领域取得的又一重大进展。人类基因组中约 80% 的 DNA 序列是具有某种特定功能的，ENCODE 为深入研究基因组作用模式提供了第一手资料。

(2) microRNA 数据资源 miRBase：miRBase 是存储、维护和命名微小 RNA 的主要数据库，主要数据资源为 microRNA 序列和注释信息；提供 miRNA 前体和成熟序列下载；允许用户使用关键词或序列检索数据库，通过关联链接到 miRNA 的原始参考文献，分析基因组中的定位和挖掘 miRNA 序列间的关系。

（王　宏　张云鹏）