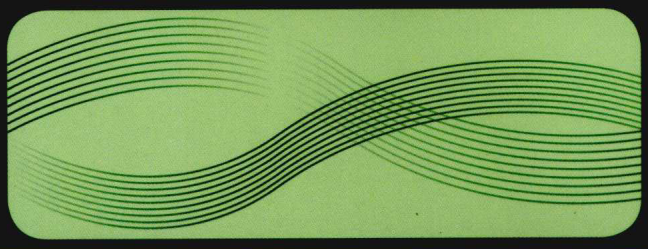
全国高等学校配套教材

供8年制及7年制临床医学等专业用



生物信息学学习指导与习题集

主编 李霞 李亦学

副主编 廖飞 张岩

人民卫生出版社

PEOPLE'S MEDICAL PUBLISHING HOUSE

生物信息学学习指导与习题集

全国高等学校配套教材

供8年制及7年制临床医学等专业用

生物信息学学习指导与习题集

主　编　李　霞　李亦学

副主编　廖　飞　张　岩

编　委　（以姓氏笔画为序）

田　心（天津医科大学）　　张　岩（哈尔滨医科大学）

朱　浩（南方医科大学）　　张绍军（哈尔滨医科大学）

刘建国（河北大学）　　茆灿泉（西南交通大学）

许丽艳（汕头医科大学）　　胡福泉（第三军医大学）

李　霞（哈尔滨医科大学）　　宫滨生（哈尔滨医科大学）

李亦学（上海交通大学）　　徐良德（哈尔滨医科大学）

肖　云（哈尔滨医科大学）　　高　磊（首都医科大学）

吴忠道（中山大学）　　童隆正（首都医科大学）

汪强虎（哈尔滨医科大学）　　廖　飞（重庆医科大学）

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与习题集/李霞等主编. —北京：人民卫生出版社，2011.6

ISBN 978-7-117-14280-9

Ⅰ.①生… Ⅱ.①李… Ⅲ.①生物信息论-高等学校-教学参考资料 Ⅳ.①Q811.4

中国版本图书馆 CIP 数据核字（2011）第 052857 号

门户网：www.pmph.com 出版物查询、网上书店
卫人网：www.ipmph.com 护士、医师、药师、中医师、卫生资格考试培训

版权所有，侵权必究！

生物信息学学习指导与习题集

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：010-67605754　010-65264830
　　010-59787586　010-59787592
印　　刷：北京市燕鑫印刷有限公司(万通)
经　　销：新华书店
开　　本：787×1092　1/16　　印张：10
字　　数：233 千字
版　　次：2011 年 6 月第 1 版　　2011 年 6 月第 1 版第 1 次印刷
标准书号：ISBN 978-7-117-14280-9/R·14281
定　　价：17.00 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销售中心联系退换）

前言

生物信息学是伴随着人类基因组计划发展起来的一门前沿交叉学科。近十年来，生物信息学方法和结论正在悄悄改变人们研究生物医学的传统方式。几乎所有生物医学领域的研究人员都会赞同，我们研究生物的方法、探索生命奥秘的途径已经发生了巨大的、深刻的变化。现在的研究人员也许很难想象，如果不用计算机、不使用很多设备中集成的软件、不使用遍布网络的关于人类和其他模式生物大量已有的实验数据和结论，我们该如何从事生物医学的相关研究。因此，掌握生物信息学的基本原理、基本方法和基本结论对临床医学专业长学制学生全面提高科研素质有着重要意义。

《生物信息学学习指导与习题集》是卫生部“十一五”全国高等学校8年制及7年制临床医学等专业规划教材《生物信息学》的配套读物。各章的章名和顺序与主教材保持一致。每一章首先按照掌握内容、熟悉内容和了解内容三个层次提出本章的学习目标，然后概述本章知识要点，最后给出复习思考题和参考答案，其内容既突出重点，又照顾全面。编写本书的目的是帮助学生在学习教材的基础上加深对生物信息学基本理论的理解，以进一步强化生物信息学知识，自测学习效果，让学生运用生物信息学的理论去分析和处理实际问题。为了提高学习生物信息学的兴趣，并从学习生物信息学重大发现的过程中获得启迪，本书还注意选编了一些富有启迪性的内容供学生阅读。本书可供临床医学、基础医学、口腔医学、麻醉学等医学专业(7)8年制学生使用，也可供其他相近专业的各类学生使用。

本书共收集各类习题364题。采用选择题、名词解释和问答题三种题型。大部分章节根据培养目标的需要还设计了上机实践题，目的是使学生在掌握理论的同时，加深对知识的理解与运用，既巩固了本章内容，又使得学生的实践能力得以提高和强化。

本书的编者大多是卫生部“十一五”规划教材《生物信息学》相应章节的原编者，因此对教程比较熟悉，并具有相当丰富的教学经验。另外，每章的习题都由熟悉相关领域的研究生进行演练，以确保习题和答案的正确性。在编写的过程中，各位编委参考了大量国内外的参考资料，充分融入了各自的教学心得和科研成果，倾注了他们大量心血，在此完稿之际，特向各位编委表示衷心的感谢！

在本书的编写过程中，尽管我们努力跟踪学科的新发展、新技术，并尽力把它们纳入到教材中来，以保持本习题集的先进性和实用性，但由于时间紧迫、能力有限，直至完稿，仍觉有许多不足之处，希望学术同仁和广大读者不吝赐教，以便再版时改正。

李　霞　李亦学

2011年5月

此为试读，需要完整PDF请访问：www.ertongbook.com

目 录

第一章

DNA、RNA和蛋白质序列信息资源

学习目标

掌握内容 通过国际互联网获取DNA、RNA和蛋白质序列信息的方法和技巧。

熟悉内容 三大核酸(GenBank、EMBL、DDBJ)数据库和蛋白质(SWISS-PROT)数据库的登录网址和数据库内容。

了解内容 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学的影响。

知识要点

一、核酸序列数据库

自从20世纪80年代第一个核酸数据库建立以来,核酸数据库迅速发展,在互联网上不仅有核酸序列数据库,而且出现了基因组相关数据库、核酸三维结构数据库、基因表达数据库、人类基因突变及疾病相关数据库、进化相关数据库及其他与核酸有关的数据库。

(一)GenBank数据库

GenBank(http://www.ncbi.nlm.nih.gov/Genbank/)是一个综合数据库,该数据库中包含了已经公开的30万余种不同物种生物的核酸序列,这些数据主要来源于全世界不同实验室和大规模测序计划项目。多数序列信息通过网络版的BankIt程序或独立的Sequin程序提交给GenBank数据库,GenBank工作人员在接收序列数据后,赋予该序列特定的数据库记录号(登录号)。GenBank数据库每天与欧洲分子生物学实验室(European Molecular Biology Laboratory Nucleotide Sequence Database,EMBL)的核酸序列数据库和日本的DNA数据库(DNA Data Bank of Japan,DDBJ)进行数据交换,以保证数据库内容在全世界范围的同步性。通过NCBI的检索系统(Entrez)可以进入GenBank,Entrez检索程序整合了主要的DNA和蛋白质序列数据的分类学、基因组、图谱、蛋白质结构和结构(功能)域信息,还包括相关的PubMed的生物医学文献信息。BLAST程序提供GenBank和其他序列数据库中序列相似性搜索服务。通过FTP站点可以获得完整的2个月一次的发布和每天更新的GenBank数据。在NCBI(http://www.ncbi.nlm.nih.gov/)主页提供了进入GenBank的路径、相关检索和分析服务。

(二)EMBL数据库

EMBL核酸序列数据库(http://www.edi.ac.uk/embl/)是欧洲主要的核酸序列收集单位。欧洲生物信息中心EBI〔欧洲分子生物学实验室(EMBL)在德国海德堡的站点〕维护这个数据库。

核酸数据来自基因组测序中心、世界各地的科学家、欧洲专利局,以及与合作伙伴DDBJ

此为试读,需要完整PDF请访问: www.ertongbook.com

和 GenBank 交换的数据。为了达到最佳的同步性，每天 DDBJ/EMBL/GenBank 之间都要交换最新的数据。在 2009 年 11 月 21 日发布的公告中宣布，EMBL 数据库已保存了包含 265 969 305 274 个核苷酸的 164 218 403 条记录。通过网址（http://www.ebi.ac.uk/Services/DBstats/）可以看到数据库统计信息。

（三）DDBJ 数据库

日本 DNA 数据库（DDBJ）是在亚洲唯一的核酸序列数据库，是公认的搜集研究者测定核酸序列的数据库，并且发放给数据提交者国际认证的核酸序列编号。由于 DDBJ 每天将搜集的数据与 EMBL-Bank/EBI 和 GenBank/NCBI 进行交换，使得三个核酸数据库几乎在任何时候都享有相同数据。这种几乎一致的数据库被称作“国际核酸序列数据库（INSD）”。DDBJ 主要收集来自日本研究者获得的序列数据，但也收集数据和发放编号给任何其他国家的研究者。

（四）其他重要的核酸序列数据库：dbEST、ncRNAdb、miRBase

1. dbEST　dbEST 是 GenBank 中的一个子数据库，包含来源于不同物种的表达序列数据和表达序列标签序列的其他信息。人类表达序列标签（ESTs）是由随机选择的 600 多个人脑互补 DNA（complementary DNA，cDNA）自动生成的部分 DNA 序列。ESTs 已被用于人类新基因的发现、人类基因组图谱绘制和基因组序列编码区识别。

2. ncRNAdb　非编码 RNA（non-coding RNA，ncRNA）数据库旨在提供非编码 RNA 的序列和功能信息。非编码转录物不编码蛋白质，但在细胞中起调节作用。目前，该数据库包含来源于 99 种细菌、古生菌和真核生物的 3 万多条序列。近年来，为了避免在特殊数据库中过度冗余，在以前版本中的 microRNAs 或 snoRNAs，以及其他管家（基础）RNAs（如：rRNA、tRNA、snRNA、SRP RNA）不包含在 ncRNA 数据库中。

3. miRBase　miRBase 序列数据库主要存放已发表的微小 RNA（microRNA，miRNA）序列和注释的数据库。miRBase 使用友好的网络界面，为用户提供 miRNA 数据，允许用户使用关键词或序列检索数据库，通过关联信息链接到 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、蛋白质序列数据库

随着分子生物学的发展，人们获得了越来越多关于蛋白质序列、结构和功能的信息。世界各国的生物学家和计算机科学家合作利用这些信息构建了蛋白质序列数据库、蛋白质三维结构数据库、蛋白质组数据库（二维凝胶电泳数据库）、信号传导及蛋白质-蛋白质相互作用数据库、DNA 和蛋白质相互作用数据库等。常用的蛋白质序列数据库主要有 PIR、MIPS 和 Swiss-Prot 数据库。

（一）PIR 数据库

蛋白质信息库（PIR）（http://pir.georgetown.edu/pirwww/）是一个支持基因组学、蛋白质组学和系统生物学检索和科学研究的综合公共生物信息学资源。

PIR 是由美国国家生物医学基金会（NBRF）于 1984 年建立，帮助研究者确认和解释蛋白质序列信息的数据库。在此之前，NBRF 首次广泛收集了 1965～1978 年，Margaret O. Dayhoff 和她的研究小组编辑的大分子蛋白质序列和结构图谱。这个研究小组率先用计算机方法比较蛋白质序列，检测远源序列间的相关性和序列内部的重复，依据蛋白质序列相似性比较，推论蛋白质分子进化历史。

PIR 主要提供如下数据库：

1. UniProt(通用蛋白质资源库) UniProt(http://www.uniprot.org/)是存储和链接其他蛋白质数据库的资源库，并且是蛋白质序列和具有综合功能注释目录的中心资源库。使用 UniprotKB 可以检索准确、可靠的蛋白质综合信息。使用 UniRef 可以减少冗余，加速序列相似性搜索。使用 UniParc 可以检索存档序列和它们来源的数据库。

2. iProClass(蛋白质知识整合数据库) iProClass(http://pir.georgetown.edu/iproclass/)提供来自 90 多个生物学数据库的大量整合数据。使用 iProClass 可以检索最新的蛋白质综合信息。

3. PIRSF(蛋白质家族分类系统) PIRSF(http://pir.georgetown.edu/pirsf/)蛋白质家族分类系统是根据超家族到亚家族序列分歧构建的多级网络分类系统，序列分歧反映了全序列蛋白质和功能域进化的关系。

4. iProLINK(蛋白质文献、信息和知识整合数据库) iProLINK(http://pir.georgetown.edu/iprolink/)提供有关注释内容的文献、蛋白质名称词典和其他有助于文献挖掘的人文语言处理技术开发的信息、数据库校正、蛋白质名称标记和功能注释标准体系(ontology)。

(二)MIPS 数据库

生物信息学和系统生物学研究所(IBIS)是慕尼黑亥姆霍兹中心-德国环境卫生研究中心的一部分并主办了慕尼黑蛋白质序列信息中心(MIPS)(http://www.helmholtz-muenchen.de/en/mips)。MIPS 的工作重点是基因组生物信息学，特别注重基因组信息系统分析，包括应用生物信息学方法注释基因组、表达分析和蛋白质组学研究。该站点提供基因组分析工具、数据库检索系统、表达分析、蛋白质相互作用等网络服务。

(三)其他重要的蛋白质序列数据库：PRINTS、Pfam

1. PRINTS PRINTS(http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php)是蛋白质基序指纹图综合数据库，每个指纹图都是使用数据扫描程序 ADSP 或 VISTAS 序列分析软件包反复优化后定义的。数据库中有两种类型指纹图，根据指纹图的复杂性分为简单和复合指纹图：简单指纹图基本上是单一的基序，而复合指纹图包含多个基序。

2. Pfam Pfam 数据库(http://pfam.sanger.ac.uk/)是一个大的蛋白质域家族集合，每个家族是用多序列比对和隐马模型(HMMs)分析结果的代表。Pfam 记录使用四种方法中的一种进行分类：

(1)家族：相关蛋白质集合；

(2)域：在多种蛋白质中被发现的结构单元；

(3)重复：多拷贝出现可形成稳定结构的，孤立不稳定的短单元；

(4)基序：在球形域(globular domain)以外发现的短单元。

相关的 Pfam 记录用序列、结构或隐马模型谱的相似性定义被分为不同家族。

三、NCBI 与 EBI

(一)NCBI 简介

NCBI 作为美国国家分子生物学信息资源，它的使命是开发新的信息技术，帮助理解控制健康和疾病的基本分子和遗传过程。它履行的职责包括：①使用数学和计算方法在分子

水平进行基本生物医学问题的研究；②与 NIH 研究所、学术界、产业界和其他政府机构保持合作；③通过召集会议、研讨会和系列讲座促进学术交流；④通过院内研究项目支持计算生物学基础和应用研究的博士后培训；⑤通过学术访问项目吸引国际科学界成员参与信息学研究和培训；⑥开发、传播、提供和协调多样化数据库和软件以便于自然科学和医学界使用；⑦开发和推进数据库、数据存储和交换以及生物学系统命名。

1992 年，NCBI 承担了建立 GenBank-DNA 序列数据库的责任，受过高级分子生物学培训的 NCBI 工作人员将来自独立实验室提交的序列和与 EMBL、DDBJ 数据交换的序列建成数据库，与美国专利商标局协商将专利中的序列数据并入 GenBank 数据库。

(二)EBI 简介

欧洲生物信息学研究所(EBI)，是欧洲分子生物学实验室(EMBL)的一部分，是世界上少有的生物信息资源，并且拥有履行这种重要任务的专业知识的部门之一。

EMBL-EBI 起源于 EMBL 核酸序列数据库(现称为 EMBL-Bank)，1980 年成立于德国海德堡的 EMBL 实验室，曾经是世界上第一个核酸序列数据库。

基于 20 多年生物信息学经验，EMBL-EBI 维护世界上最广泛的分子数据库。EMBL-EBI 是在全球范围内，努力协调搜集和传播生物学数据的欧洲节点，EMBL-EBI 的许多数据库是生物学家们熟知的，包括：EMBL-Bank(DNA 和 RNA 序列)、Ensemble(基因组)、ArrayExpress(基于微阵列的基因表达数据)、UniProt(蛋白质序列)、interPro(蛋白质家族、域和基序)、Reactome(传导通路)和 ChEBI(小分子)等。

(三)通过 Entrez Gene 从 NCBI 获取序列信息

Entrez 主要是用于 NCBI 数据库综合的、基于文本的搜索和检索系统。Entrez 综合了科学文献、DNA 和蛋白质序列数据、3D 蛋白质结构和蛋白质域数据、种群研究数据集、表达数据、完整基因组组装和分类学信息，形成一个紧密链接的系统。

Entrez Gene 检索到的记录提供关键链接，将图谱、序列、表达、结构、功能、索引文献和同源数据链接在一起构成关键链接。用定义序列、已知的图谱定位和从表型信息推测的基因，为基因分配特有标识符。这些标识符在 NCBI 的数据库中通用，可以用于注释、更新和相关信息跟踪。检索 Entrez Gene 最简捷的方法是登录到 NCBI(http://www.ncbi.nlm.nih.gov/)的首页，在检索窗口的选择数据库栏(Search)的下拉菜单中选择 Gene 选项，然后在检索栏(for)输入欲查询的检索词或词组，进行检索。

(四)通过 SRS 从 EBI 中获取蛋白质序列信息

SRS 是世界上主要的生物信息学、基因组和相关数据整合、分析和显示工具。SRS 检索系统是个开放的系统，可以根据不同的需要安装不同的数据库，现在安装在 EBI 的数据库有 300 多个。SRS 有三种检索方式：快速检索、标准检索和批量检索。

我们可以通过网址(http://srs.ebi.ac.uk/)进入 SRS 开始页面。在这个页面中，可以开始一个永久项目，在该项目中，允许用户在 SRS 系统中安装用户自己的相关数据库。该页面就是快速检索页面，在页面下部的“List Search”窗口允许用户进行批量检索。当用户开始检索或选择数据库时，用户不必特意操作，服务器将自动为用户生成一个临时项目。在开始页面点击 Quick 快捷方式按钮，就可以开始快速检索 EBI 数据库操作。如果用户想进一步查看详细完整的记录内容，可以点击超链接查看。由于快速检索是在 SRS 系统的所有数据库中检索，检索到的记录比较多，很多记录不是用户需要查询的内容，因此，可以使用标

准检索，能够较快检索到用户需要的记录。点击“Query Form”按钮，进入标准检索页面，在开始标准检索前，用户必须先点击“Library Page”按钮，选择需要查询的数据库。SRS 系统允许用户保存检索结果，已备用户后期使用。

复习思考题

一、选 择 题

1. 下列数据库不属于核酸数据库的是

A. ncRNA　B. dbEST　C. dbSTS　D. tRNAdb　E. PIR

2. 下列数据库不属于蛋白质数据库的是

A. PROSITE　B. NDB　C. SCOP　D. SMART　E. MIPS

3. 世界上第一个核酸数据库是何时建立的？

A. 20 世纪 50 年代　B. 20 世纪 60 年代　C. 20 世纪 70 年代
D. 20 世纪 80 年代　E. 20 世纪 90 年代

4. NCBI 为医学和学术界提供的数据库包括

A. GenBank　B. interPro　C. OMIM　D. UniGene　E. MMDB

5. EMBL-EBI 研究团队的研究领域包括

A. 进化途径的基因组分析　B. 序列数据进化分析
C. 神经信号计算系统生物学　D. 蛋白质组学：结构、功能和进化
E. 基因组规模调节系统分析和功能基因组学

6. 三大核酸数据库包括

A. GenBank　B. GEO　C. EMBL　D. GOLD　E. DDBJ

二、名 词 解 释

1. dbEST　2. Entrez　3. PRINTS　4. GenBank　5. miRBase

三、问 答 题

1. 除了 GenBank、EMBL 和 DDBJ 三大核酸序列数据库外，还有哪些特殊类型核酸序列数据库？

2. 为了保证核酸数据库的内容在全世界范围的同步性，GenBank 数据库每天与哪些数据库进行数据交换？

3. 常用的蛋白质序列数据库有哪些？

4. PIR 主要提供哪四个数据库？

5. NCBI 作为美国国家分子生物学信息资源，其履行的职责包括哪些主要内容？

6. Entrez 的主要功能有哪些？

7. EBI 的含义是什么？主要包含哪些数据库？

8. Entrez Gene 记录的全文报告包括哪些主要内容？

9. 什么是 SRS？

10. SRS 有哪三种检索方式？

答案与题解

一、选 择 题

1. E　2. B　3. D　4. ACDE　5. ABCDE　6. ACE

二、名 词 解 释

1. dbEST：dbEST 是 GenBank 中的一个子数据库，包含来源于不同物种的表达序列数据和表达序列标签序列的其他信息。

2. Entrez：Entrez 是 NCBI 的搜索和检索系统，为用户提供序列、图谱、分类学和结构数据整合的访问程序。

3. PRINTS：PRINTS 是蛋白质基序指纹图综合数据库。

4. GenBank：GenBank 是具有目录和生物学注释的核酸序列综合数据库，由美国国家医学图书馆（the National Library of Medicine，NLM）的国家生物技术信息中心（the National Center for Biotechnology Information，NCBI）构建、维护和管理。

5. miRBase：miRBase 序列数据库主要存放已发表的微小 RNA（microRNA，miRNA）序列和注释的数据库。miRBase 使用友好的网络界面，为用户提供 miRNA 数据，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。

三、问 答 题

1. 答：除了 GenBank、EMBL 和 DDBJ 三大核酸序列数据库外，还有特殊类型核酸序列数据库：非编码 RNA 数据库（ncRNA）、表达序列标签数据库（dbEST）、序列标签位点数据库（dbSTS）等，其他还有 miRBase、tRNAdb 等。基因组相关数据库：人类基因组数据库（HGD）；基因组序列数据库（GSDB）；基因组在线数据库（GOLD）等。核酸三维结构数据库：核苷酸三维结构数据库（NDB）；普纳大学核酸结构数据库（BNASDB）等。基因表达数据库：基因表达库（GEO）；斯坦福微阵列数据库（SMD）；以及 ArrayExpress、CGED、GXD、BodyMap 等。人类基因突变及疾病相关数据库：人类基因变异数据库（HMGD）、人类遗传双等位基因序列数据库（HGBASE）、人类孟德尔遗传在线（OMIM）、国际单体型计划（HapMap）、人类单核苷酸多态性数据库（dbSNP）、肿瘤基因数据库（TGDB）、疾病关联数据库（GAD）、癌症基因数据库（CGAP）、人类表观遗传数据库（HEP）、人类 DNA 甲基化与癌症数据库（MethylCancer）等。

2. 答：GenBank 数据库每天与欧洲分子生物学实验室的核酸序列数据库（European Molecular Biology Laboratory Nucleotide Sequence Database，EMBL）和日本的 DNA 数据库（DNA Data Bank of Japan，DDBJ）进行数据交换，以保证数据库内容在全世界范围的同步性。

3. 答：常用的蛋白质序列数据库主要有 PIR、MIPS 和 Swiss-Prot 数据库。另外还有与蛋白质功能、结构域和蛋白质家族有关的数据库：PROSITE、InterPro、Pfam、ProDom、SMART 等，蛋白质三维结构相关数据库：PDB、BioMagResBank、SWISS-MODEL Reposi-

tory、ModBase、CATH、SCOP、ReLiBase、TOPS、SWISS-3DIMAGE 和 BioImage 等，蛋白质二维凝胶电泳数据库：WORLD-2DPAGE、Phoretix links，信号传导及蛋白质-蛋白质相互作用相关数据库：DIP、INTERACT、ProNet、KEGG、CANSITE、SPAD、CSNDB 等，DNA 和蛋白质相互作用数据库：DPInteract，蛋白质翻译后修饰相关数据库：O-GlycBase、PhosphoBase、RESID 等。

4. 答：PIR 主要提供如下四个数据库资源：①UniProt（通用蛋白质资源库）（http://www.uniprot.org/）是存储和链接其他蛋白质数据库的资源库，并且是蛋白质序列和具有综合功能注释目录的中心资源库。使用 UniprotKB 可以检索准确、可靠的蛋白质综合信息。使用 UniRef 可以减少冗余，加速序列相似性搜索。使用 UniParc 可以检索存档序列和它们来源的数据库。②iProClass（蛋白质知识整合数据库）（http://pir.georgetown.edu/iproclass/）提供来自 90 多个生物学数据库的大量整合数据，包括蛋白质 ID 图谱服务、UniProtKB 编注蛋白质摘要描述和筛选 UnParc 数据库的蛋白质序列。使用 iProClass 可以检索最新的蛋白质综合信息，包括：功能、转导通路、相互作用、家族分类、基因和基因组、功能注释标准体系（ontology）、文献和分类学信息。使用 iProClass 还可以检索 ID 图谱、蛋白质词典和相关序列。③PIRSF（蛋白质家族分类系统）（http://pir.georgetown.edu/pirsf/）蛋白质家族分类系统是根据超家族到亚家族序列分歧构建的多级网络分类系统，序列分歧反映了全序列蛋白质和功能域进化的关系。主要的 PIRSF 分类单位是同源家族，家族成员具有同源性（进化来自共同的祖先）和同拓扑性（具有全长序列相似性和共同的功能域结构）。PIRSF 人工排列家族成员，注释特殊的生物学功能、生物化学活动和序列特征。另外，制定功能位点和蛋白质命名规则，协助传播和规范蛋白质注释标准，系统地检测注释错误。PIRSF 的报告提供研究蛋白质进化相关的独立平台。它概要论述家族的特征，如家族名称、分类分布、分级和功能域结构，以及家族成员，包括功能、结构、传导通路、功能注释标准体系和家族分类，还具有广泛的相关数据库的链接。利用这些信息可以获得你所关注蛋白质的准确功能或预测的功能和该蛋白质所属家族成员共有的其他特征。④iProLINK（蛋白质文献、信息和知识整合数据库）（http://pir.georgetown.edu/iprolink/）提供有关注释内容的文献、蛋白质名称词典和其他有助于文献挖掘的人文语言处理技术开发的信息、数据库校正、蛋白质名称标记和功能注释标准体系。使用 iProLINK 可以获得描述蛋白质记录的文本文献资源，在 UniProtKB 记录（生物词典）中加入蛋白质/基因命名的图谱，获得用于开发文本挖掘算法的注释数据集、挖掘蛋白质磷酸化（RLIMS-P）文献和获得蛋白质功能注释标准体系（PRO）信息。

5. 答：NCBI 履行的主要职责包括：①使用数学和计算方法在分子水平进行基本生物医学问题的研究；②与 NIH 研究所、学术界、产业界和其他政府机构保持合作；③通过召集会议、研讨会和系列讲座促进学术交流；④通过院内研究项目支持计算生物学基础和应用研究的博士后培训；⑤通过学术访问项目吸引国际科学界成员参与信息学研究和培训；⑥开发、传播、提供和协调多样化数据库和软件以便于自然科学和医学界使用；⑦开发和推进数据库、数据存储和交换以及生物学系统命名。

6. 答：Entrez 是 NCBI 的搜索和检索系统，为用户提供序列、图谱、分类学和结构数据整合的访问程序。Entrez 还提供序列和结构数据图像视图，Entrez 强大和独特的特征是能够检索相关序列、结构和参考文献。通过 PubMed，提供检索超过 1000 万以上期刊索引内

此为试读，需要完整PDF请访问：www.ertongbook.com