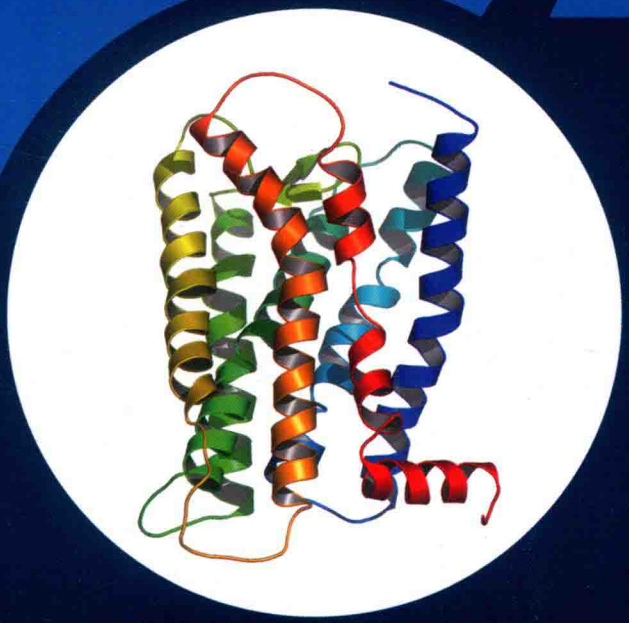
DANBAIZHI JIEGOU SHENGWU XINXIXUE

蛋白质结构生物信息学

鄢仁祥 王晓锋 陈 震 蔡伟文 林 娟◎编著



海峡出版发行集团 | 福建科学技术出版社
THE STRAITS PUBLISHING & DISTRIBUTING GROUP | FUJIAN SCIENCE & TECHNOLOGY PUBLISHING HOUSE

DANBAIZHI JIEGOU SHENGWU XINXIXUE

蛋白质结构生物信息学

鄢仁祥 王晓锋 陈 震 蔡伟文 林 娟◎编著

海峡出版发行集团 THE STRAITS PUBLISHING & DISTRIBUTING GROUP | 福建科学技术出版社 FUJIAN SCIENCE & TECHNOLOGY PUBLISHING HOUSE

图书在版编目（CIP）数据

蛋白质结构生物信息学 / 鄢仁祥等编著. —福州 ：福建科学技术出版社，2017.5
ISBN 978-7-5335-5096-7

Ⅰ.①蛋… Ⅱ.①鄢… Ⅲ.①蛋白质－生物结构－生物信息论 Ⅳ.①Q510.1

中国版本图书馆 CIP 数据核字（2016）第 155356 号

书　　名　蛋白质结构生物信息学
编　　著　鄢仁祥　王晓锋　陈　震　蔡伟文　林　娟
出版发行　海峡出版发行集团
　　福建科学技术出版社
社　　址　福州市东水路 76 号（邮编 350001）
网　　址　www.fjstp.com
经　　销　福建新华发行（集团）有限责任公司
印　　刷　福州万紫千红印刷有限公司
开　　本　889 毫米×1194 毫米　1/16
印　　张　10
字　　数　300 千字
版　　次　2017 年 5 月第 1 版
印　　次　2017 年 5 月第 1 次印刷
书　　号　ISBN 978-7-5335-5096-7
定　　价　35.00 元

前 言

Sanger团队在1977年利用双脱氧链终止法对噬菌体Phi X 174进行了完整测序(Sanger等 Nature 1977)，之后关于测序技术的研究就广泛展开了。在1985年，美国科学家基于该技术率先提出了被誉为与“人造卫星登月计划”相媲美的人类基因组测序的构想，旨在获得人类基因组中30亿个碱基对序列信息，探索基因在人类染色体上的准确位置，并希望破译人类全部遗传信息。该计划在当时预计需耗时十多年并要求至少约30亿美元的科研经费投入，测序结果可以在分子水平上全面地认识人类自身结构。该计划最终由美国、英国、法国、日本、德国以及中国共同参与完成，并于2001年公布了人类第一份基因组草图，这个测序草图的完成标志着生物学研究达到了一个新的里程碑。随着人类基因组计划的实施，相应的测序技术得到快速的发展；水稻、花生、玉米、小麦、果蝇、真菌、细菌、小鼠等生物基因组草图也相继顺利获得。利用基因组数据，人们可以精准地定位相应特性所在的基因区段，可以对水果和蔬菜品种进行改良，进而为提高农作物的品质提供更有效的研究线索。因此，更多的转基因植物和动物，以及相关食品将出现。另外，通过基因组数据研发新药物，调节人体的生化特性，人类将可能恢复或修复人体细胞和器官的功能，甚至可能改变人类的进化历程。各种具有代表性物种基因组测序的完成，标志着后基因组时代(post-genomic era)已经来临。

随着基因组计划的成功完成，结构基因组计划正在开展，研究人员将会积累越来越多的核酸序列、蛋白质序列、结构与功能数据。另外，蛋白质序列与结构数据之间的数量差异不断加大。目前，如何对这些数据进行分析，挖掘其中潜在的生物学知识成为研究人员亟需面对的科学问题之一。仅仅通过实验方法进行研究存在一定的局限性，例如实验方案一般较复杂且周期长，而生物信息学(Bioinformatics)可能是另外一种解决方案。生物信息学是利用数学、信息学、统计学和计算机科学等计算方法从理论层面上研究生物学问题的一门学科。生物信息学通过对海量的生物学数据进行分析，有助于全面而深入地了解基因组和蛋白质组中蕴含的生物学功能，同时可以极大地节约大量的人力、物力和财力。蛋白质是细胞活性及功能的执行者，蛋白质复杂的结构决定着生物体系的复杂程度。蛋白质结构生物信息学(Protein Structural Bioinformatics)，生物信息学中的一个分支，是利用计算机、统计学等方法来处理和分析蛋白质结构和功能的一门学科。蛋白质结构生物信息学主要研究以下问题：(1)蛋白质结构数据的存储；(2)根据结构与进化信息对蛋白质进行分类；(3)蛋白质结构与功能位点预测；(4)基于结构的功能位点处理与分析。蛋白质的序列、结构与功能的关系是一个复杂的科学问题。虽然各种蛋白质预测算法被相继提出，而且这些方法的预测精度也在持续提高，然而关于蛋白质是如何折叠的生物学本质并没有得到更深入的认识。如果

蛋白质的折叠过程被很好地认识，那么从序列到结构，即蛋白质结构的预测问题也许就能顺利地解决。

近几十年来蛋白质结构生物信息学研究一直是生物学研究中的热点之一，特别是蛋白质结构预测研究。如果能够进一步提高蛋白质结构预测的精度，那么一方面可以让该方法有更加广阔的应用空间；另一方面也可以让人们对蛋白质序列与结构的关系有更加深入的理解。蛋白质结构预测虽然已经得到了广泛的应用，但是在膜蛋白质结构预测上的预测性能还不太理想。主要原因是已经通过实验手段解析出的膜蛋白的结构数量还非常少，在这种情况下蛋白质结构预测方法难以找到合适的模板。在直接预测出膜蛋白的三维结构比较困难的情况下，研究者们一般通过预测膜蛋白的一些拓扑特性进行相关的科学研究。对水溶性的球蛋白而言，已经有观点认为：Protein Data Bank 数据库中已经包含所有单结构域蛋白所需的模板；相关数据库中已经积累了足够多的数据，可以对蛋白质起重要功能作用的位点和结合口袋进行统计建模。所以目前可能是发展和应用蛋白质结构与功能分析方法的最佳时机。

编写本书的初衷是笔者计划在教学中为学生们提供一本实用而又简易的结构生物信息学教材，内容需要涵盖一些关键核心技术，例如结构相关数据库、序列比对、二级和三级结构预测、结构模拟以及重要功能位点预测等。本书的主要目标读者是高年级本科生、研究生以及一线的科研人员。该书可以作为一本蛋白质结构计算的简要读本。虽然本书的几位作者已经在该领域从事多年的科研工作，但是在编写过程中仍然遇到不少问题。因此，我们没有打算让本书成为覆盖全面的大百科全书式的著作。该书在写作过程中引用较多经典的文献，我们希望以点带面，引导读者了解相应的生物信息学知识。

在本书完成之际，我们要感谢中国农业大学的张子丁教授，他为本书的写作提出了非常中肯的建议，同时也是本书前三位作者共同的博士生导师。同时，感谢福州大学为本书的作者之一鄢仁祥博士提供本科生和研究生《生物信息学》课程的授课机会，本书雏形即在该课程讲义的基础上不断修订完善而成。本书的写作及出版过程得到了国家自然科学基金青年项目《G 蛋白偶联受体结构及与药物配体结合的计算研究》(项目编号：31500673)、福建省教育厅科技项目《膜蛋白质序列、结构与功能关系的挖掘》(项目编号：JA14049)、福州大学人才基金项目《与疾病相关的生物信息学平台的构建》(项目编号：XRC-1336)的资助。在此表示衷心感谢。

编者

目 录

第一章　蛋白质相关数据库简介

生物信息学，一个传统生物学和现代信息技术相结合的产物，是采用数学和计算机技术来存储和分析生物学数据的一门综合性学科。该学科发展的动力主要源于不断积累的生物学数据，包括核酸序列、蛋白质序列、蛋白质三维结构信息、重要功能位点和表观遗传等在内的一系列数据。其中，蛋白质数据对解释相关生物学功能作用尤其重要。蛋白质(Protein)这个词是由希腊语“Proteios”转化而来的，意思是“头等重要的”。在 1839 年，荷兰医药化学家 Mulder 推导出了一个基本的化学式：$C_{40}H_{62}O_{12}N_{10}$，其中含有碳、氢、氧和氮元素。Mulder 认为只要在这个基本化学式中加入含硫或含磷的基团，就可以形成各种蛋白质化合物。从那时起蛋白质就被认为是一种大的分子化合物。生物体中绝大多数的细胞功能都是通过蛋白质介导调控的。蛋白质由 20 种氨基酸分子组成，相邻氨基酸残基的羧基和氨基通过肽键连接在一起，并折叠形成各式各样的结构类型。蛋白质通过折叠成特定的三维结构来行使其生物学功能，例如一些蛋白质就是细胞生化反应过程中需要的酶[1]。生物体内蛋白质的三维结构基本是稳定的，但有时蛋白质中的一些氨基酸还可以被修饰从而引起蛋白质结构的变化，并通过这个过程起到一定的生物调控作用。一些膜蛋白还具有信号传导作用和机体免疫作用等[2, 3]。蛋白质也是人们日常饮食中必需的营养物质，人体自身无法合成某些必需氨基酸，但可以通过消化所摄入的食物(例如牛奶)中包含的蛋白质，将蛋白质降解为氨基酸，再将吸收的氨基酸用于自身蛋白质的合成。因此，蛋白质在生命科学研究中的重要性是不言而喻的。在 1989 年发起的人类基因组计划于 2000 年初步完成之后，大批物种的基因组及蛋白质组相继被成功测定，大量的生物学原始数据在这个过程中产生。如何解读这些生物学数据，特别是蛋白质序列、结构与功能数据，成为研究者们面临的生物学问题之一。同时，这也是现代生物学家和计算科学家们面对的发展机遇之一。犹如 Pevsner[4] 所说：“我们正面对生物学史上一个不平凡的时刻，类似于在第十九世纪元素周期表刚刚完成时，那时元素周期表已清晰地排列成行和列，但我们仍然花了一个世纪来掌握元素的意义。今天我们虽已测得了数以千计的生物基因组数据，而要寻找一个逻辑来解释这些数据的作用和功能，这一过程可能需要另外一个一百年”。

数据库(Database)是指按照一定的数据结构来组织、存储和管理数据的文件或者软件。在数据库系统中信息可以非常方便地进行检索。相对成规模的生物序列数据收集和整理也许最早可以追溯到 1965 年由 Dayhoff 开发的蛋白质数据库(Atlas of Protein Sequence and Structure)。在 1970 年左右，Brookhaven 国家实验室建立了 Protein Data Bank (PDB)数据库。十年之后(1980 年左右) GenBank 数据库才出现。这些数据库建立之初数据量都非常少，但是近年来生物学数据呈指数级增长，促使科研人员构建新的数据库系统并开发新的分析工具来存储和处理这些数据。NCBI、EMBL 和 DDBJ 为世界三大生物数据信息中心。这三个世界主流中心提供了非常多样和重要的研究数据。下面将简要地介绍这三个生物信息学

中心，并重点和详细地介绍一些与本书后续内容直接相关的一些蛋白质数据库。另外，表 1-1 中列出了三大生物信息中心和一些常用的生物学数据库。

第一节　三大生物信息研究中心

一、NCBI

NCBI (National Center for Biotechnology Information)是指美国国立生物技术信息中心，创建于 1988 年，可以通过网址 http://www.ncbi.nlm.nih.gov/访问。NCBI 是 NIH(National Institutes of Health)下属的国立医学图书馆(NLM)的一个分支。从 1992 年起，NCBI 承担起维护 GenBank DNA 序列数据库的责任。NCBI 网站的数据库包含大部分已知的核酸序列和蛋白质序列，以及与它们相关的文献著作和生物学功能注释。NCBI 中的核酸数据来源包括三个方面：直接来源于测序工作者提交的序列；由测序中心提交的大量 EST 序列和其他测序数据；以及与其他数据机构协作交换数据而来。NCBI 的文献著作数据库 PubMed 存储着大量与核酸及蛋白质序列相关的文献，以及 PubChem 数据库存储小分子结构。NCBI 也提供多种生物序列分析工具，例如 BLAST。总体来说，NCBI 是目前使用最为广泛的集生物数据库、分析工具和文献在内的一个综合性网站。

二、EMBL

EMBL(The European Molecular Biology Laboratory)是指欧洲分子生物学实验室(即欧洲生物信息学中心)，创建于 1974 年，可以通过网址 http://www.ebi.ac.uk 访问。EMBL 是一个综合性和国际化的科学研究中心，由欧洲 14 个国家和亚洲的以色列等国共同发起建立。EMBL 提供核酸与蛋白相关的各种数据库，也包含各种与疾病相关的数据。同时，EMBL 也提供一系列可灵活使用的序列分析工具。由于具有开放和创新的良好学术氛围，该研究中心目前已发展成欧洲最重要和核心的分子生物学基础研究和教育培训机构。例如，著名的结构生物信息学家 Rost 和 Sander 就曾经在 EMBL 研究中心工作过。

三、DDBJ

DDBJ(DNA Data Bank of Japan) 是指日本构建的 DNA 数据库，现也称为日本生物信息研究中心，创建于 1984 年，可以通过网址 http://www.ddbj.nig.ac.jp 访问。DDBJ 创建之初主要收集及存储核酸数据，后期也逐渐收集蛋白及其他与生物序列相关的数据。DDBJ 也提供不少序列分析工具，例如 SQmateh，可以用来搜索基因或蛋白质中短的碱基或氨基酸序列区域。

四、三个研究中心关系

NCBI、EMBL 和 DDBJ 这三个研究中心的数据库每天都在更新数据和交换信息，而且这三个研究中心同时主持两个国际年会：国际 DNA 数据库咨询会议

和国际 DNA 数据库协作会议。因为定期相互验证与交换数据信息，所以三个信息中心中的数据在理论上应该是相同或者相近的，但这三个中心在分析工具上各有一些特点。

表 1-1　三大生物信息中心及主流蛋白质相关数据库

数据库	主要内容	网址
	三大生物信息中心	
NCBI	美国国立生物技术信息中心	http://www.ncbi.nlm.nih.gov
EMBL	欧洲生物信息数据中心	http://www.ebi.ac.uk
DDBJ	日本生物信息数据中心	http://www.ddbj.nig.ac.jp
	主流蛋白质相关数据库	
PDB	PDB 数据库是储存通过实验测定的蛋白质结构的数据库，作为一级数据库，为其他数据库提供原始数据	http://www.rcsb.org/pdb
SCOP	SCOP 数据库主要依赖于结构生物学专家对 PDB 数据库中的结构进行分类。与其他数据库相比，SCOP 数据库更新速度较慢	http://scop.mrc-lmb.cam.ac.uk
CATH	CATH 数据库的构建既使用计算机程序(SSAP 结构比对软件)，也进行人工检查以验证结果正确与否	http://www.cathdb.info
FSSP	FSSP 数据库采用 Dali 程序把蛋白质结构分成不同的家族	http://swift.cmbi.kun.nl/swift/fssp
HOMSTRAD	HOMSTRAD 数据库采用结构比对软件 COMPARER 对蛋白质结构进行比对及分类，数据库中提供相应的多序列比对结果	http://tardis.nibio.go.jp/homstrad
PubChem	美国国家健康研究院(US National Institutes of Health)构建的有机小分子生物活性数据库	http://www.ncbi.nlm.nih.gov/pccompound

```
HEADER    TRANSFERASE/TRANSFERASE INHIBITOR       14-AUG-97   10GS
TITLE     HUMAN GLUTATHIONE S-TRANSFERASE P1-1, COMPLEX WITH TER117
COMPND   3 CHAIN: A, B;
COMPND   4 SYNONYM: GSTP1-1;
COMPND   5 EC: 2.5.1.18;
COMPND   6 ENGINEERED: YES
SOURCE    MOL_ID: 1;
REMARK   1
REMARK   1 REFERENCE 1
REMARK   1  AUTH   P.REINEMER,H.W.DIRR,R.LADENSTEIN,R.HUBER,M.LO BELLO,
REMARK   1  AUTH 2 G.FEDERICI,M.W.PARKER
REMARK   1  TITL   THREE-DIMENSIONAL STRUCTURE OF CLASS PI GLUTATHIONE
REMARK   1  TITL 2 S-TRANSFERASE FROM HUMAN PLACENTA IN COMPLEX WITH
REMARK   1  TITL 3 S-HEXYLGLUTATHIONE AT 2.8 A RESOLUTION
REMARK   1  REF    J.MOL.BIOL.                   V. 227   214 1992
REMARK   1  REFN                   ISSN 0022-2836
REMARK   2
REMARK   2 RESOLUTION.    2.20 ANGSTROMS.
REMARK 800 SITE_IDENTIFIER: AC4
REMARK 800 EVIDENCE_CODE: SOFTWARE
REMARK 800 SITE_DESCRIPTION: BINDING SITE FOR RESIDUE MES B 211
HET    VWW  A 210      33
HET    MES  A 211      12
HET    VWW  B 210      33
HET    MES  B 211      12
HETNAM     VWW L-GAMMA-GLUTAMYL-S-BENZYL-N-[(S)-CARBOXY(PHENYL)
HETNAM   2 VWW  METHYL]-L-CYSTEINAMIDE
HETNAM     MES 2-(N-MORPHOLINO)-ETHANESULFONIC ACID
FORMUL   3  VWW    2(C23 H27 N3 O6 S)
FORMUL   4  MES    2(C6 H13 N O4 S)
FORMUL   7  HOH   *169(H2 O)
ATOM      1  N   PRO A   2      31.242   3.064  39.284  1.00 39.90           N
ATOM      2  CA  PRO A   2      31.195   2.392  37.963  1.00 31.96           C
ATOM      3  C   PRO A   2      29.975   2.923  37.197  1.00 30.23           C
ATOM      4  O   PRO A   2      29.727   4.132  37.181  1.00 27.03           O
ATOM      5  CB  PRO A   2      31.063   0.905  38.251  1.00 36.57           C
ATOM      6  CG  PRO A   2      30.276   0.947  39.549  1.00 35.11           C
ATOM      7  CD  PRO A   2      30.829   2.121  40.343  1.00 42.06           C
ATOM      8  N   TYR A   3      29.189   2.020  36.613  1.00 22.83           N
ATOM      9  CA  TYR A   3      28.011   2.405  35.850  1.00 18.42           C
ATOM     10  C   TYR A   3      26.711   1.995  36.517  1.00 19.46           C
ATOM     11  O   TYR A   3      26.629   0.949  37.161  1.00 24.89           O
ATOM     12  CB  TYR A   3      28.055   1.772  34.459  1.00 17.73           C
ATOM     13  CG  TYR A   3      29.318   2.059  33.684  1.00 17.23           C
ATOM     14  CD1 TYR A   3      29.678   3.366  33.355  1.00 19.19           C
ATOM     15  CD2 TYR A   3      30.149   1.023  33.267  1.00 16.84           C
ATOM     16  CE1 TYR A   3      30.835   3.633  32.629  1.00 18.79           C
ATOM     17  CE2 TYR A   3      31.308   1.279  32.539  1.00 20.85           C
ATOM     18  CZ  TYR A   3      31.645   2.584  32.226  1.00 20.77           C
ATOM     19  OH  TYR A   3      32.803   2.828  31.528  1.00 22.24           O
ATOM     20  N   THR A   4      25.683   2.804  36.297  1.00 13.22           N
ATOM     21  CA  THR A   4      24.361   2.547  36.835  1.00 19.34           C
ATOM     22  C   THR A   4      23.337   2.985  35.797  1.00 20.48           C
ATOM     23  O   THR A   4      23.374   4.124  35.335  1.00 22.30           O
ATOM     24  CB  THR A   4      24.097   3.357  38.132  1.00 18.80           C
ATOM     25  OG1 THR A   4      25.094   3.042  39.110  1.00 19.60           O
ATOM     26  CG2 THR A   4      22.713   3.031  38.699  1.00 12.79           C
ATOM     27  N   VAL A   5      22.463   2.071  35.385  1.00 18.58           N
ATOM     28  CA  VAL A   5      21.420   2.437  34.438  1.00 17.96           C
ATOM     29  C   VAL A   5      20.078   2.369  35.160  1.00 19.40           C
ATOM     30  O   VAL A   5      19.743   1.367  35.790  1.00 22.04           O
ATOM     31  CB  VAL A   5      21.436   1.582  33.113  1.00 17.57           C
ATOM     32  CG1 VAL A   5      22.717   0.779  32.989  1.00 12.49           C
ATOM     33  CG2 VAL A   5      20.204   0.704  32.988  1.00 22.38           C
ATOM     34  N   VAL A   6      19.366   3.488  35.149  1.00 19.86           N
```

图 1-1　一个 PDB 文件的部分信息

表 1-2　氨基酸三字母和单字母符号

#	中文名称	英文名称	三字母	单字母
1	丙氨酸	Alanine	Ala	A
2	半胱氨酸	Cystine	Cys	C
3	天冬氨酸	Asparticacid	Asp	D
4	谷氨酸	Glutamicacid	Glu	E
5	苯丙氨酸	Phenylalanine	Phe	F
6	甘氨酸	Glycine	Gly	G
7	组氨酸	Histidine	His	H
8	异亮氨酸	Isoleucine	Ile	I
9	赖氨酸	Lysine	Lys	K
10	亮氨酸	Leucine	Leu	L
11	甲硫氨酸	Methionine	Met	M
12	天冬酰胺	Asparagine	Asn	N
13	脯氨酸	Proline	Pro	P
14	谷氨酰胺	Glutarnine	Gln	Q
15	精氨酸	Arginine	Arg	R
16	丝氨酸	Serine	Ser	S
17	苏氨酸	Threonine	Thr	T
18	缬氨酸	Valine	Val	V
19	色氨酸	Tryptophan	Trp	W
20	酪氨酸	Tyrosine	Tyr	Y

第二节　蛋白质序列和结构相关数据库

一、PDB 数据库

PDB[5] 数据库始建于 1971 年，由美国布鲁海克海文国家实验室开发及维护。PDB 数据库收集通过 X 射线衍射和核磁共振等实验方法测定的蛋白质及其他生物分子结构，该数据库建立之初仅包含 7 个蛋白质结构数据。早期版本的 PDB 数据库也接收通过计算方法得到的理论结构模型，但目前 PDB 数据库中仅包含也仅接收通过实验手段测定的蛋白质结构。PDB 数据库以文本文件的格式存放蛋白质结构数据，可以通过该网站直接检索单个蛋白的序列及其对应的三维结构数据，或者通过该网站提供的 FTP 地址批量下载。PDB 数据库中一般用 4 个字符表示一个蛋白，例如 1F88[6]。若是复合物蛋白，通常用第 5 个字母表示其链，例如 1F88A 和 1F88B 分别表示该蛋白中的 A 链和 B 链结构。蛋白质三维结构文件内容中除了包含原子坐标的信息外，还包含物种来源、化合物名称、结构分辨率、结构因子、温度系数和蛋白质主链数目等数据。图 1-1 是一个 PDB 文件(10gs 蛋白)的部分数据，其中包含酶分类号、与之相互作用的小分子以及原子坐标等信息。PDB 文件中相应的氨基酸是以三个字母的形式表示的，而一般序列文件中氨基酸是以单字母形式表示的。表 1-2 列出了 20 种氨基酸单字母与三字母符号。需要注意的是一些小分子(例如甘油、乙二醇等)有时会被用来作为添加剂来解析蛋白质晶体结构，所以并不是所有在 PDB 文件中存在的小分子都与该蛋白

有生物相关性，这时就需要进一步了解 PDB 文件中蛋白及小分子的结构，或者挖掘相关文献加以判断。PDB 数据库中对小分子一般以 SDF 格式存储。另外，PDB 数据库中同时存储经过 X 晶体衍射和 NMR 方法获得的结构，但两个方法存储的数据是不太相同的。一般经过 X 晶体衍射法获得的蛋白就只有一种结构类型，而通过 NMR 法获得的是蛋白质结构的集合(ensemble)。两者不同的原因是 X 晶体衍射法通过蛋白质晶体获得数据，而 NMR 是扫描蛋白质在溶液中的结构(即动态的过程)获得数据。PDB 结构需要通过特定的结构工具才能直观地显示其图形。学术界目前已经开发不少蛋白质三维结构的可视化工具，例如 Pymol[7] 和 Rasmol [8] 等。

二、PDBsum 数据库

PDBsum 是 PDB 的一个拓展数据库。PDB 数据库属于一级数据库，即收集最原始的蛋白质序列及结构数据。PDB 数据库中包含大量的有用信息，但同时可能含有许多冗余信息，甚至错误信息。因此，有些研究组对 PDB 数据库中的数据进行了正确性检验，并把分析结果以数据库的形式存储。PDBsum 就是这样的数据库，它由英国伦敦大学开发与维护。总体来讲，PDBsum 数据库是一个基于 PDB 注释信息的综合型数据库，该数据库是对 PDB 数据库中的数据进行分析和总结，可以通过网址 http://www.ebi.ac.uk/pdbsum 访问。随着 PDB 数据库中蛋白质结构数据量的增长，不少研究组开发了基于 PDB 数据库的蛋白质结构分类数据库，例如 SCOP、CATH、FSSP 和 HOMSTRAD 等。

```
>1F88A
MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIMLGFPINFLTLY
VTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLG
GEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGVAFTWVMALACAAPPLVGWSRYIP
EGMQCSCGIDYYTPHEETNNESFVIYMFVVHFIIPLIVIFFCYGQLVFTVKEAAAAXSATT
QKAEKEVTRMVIIMVIAFLICWLPYAGVAFYIFTHQGSDFGPIFMTIPAFFAKTSAVYNP
VIYIMMNKQFRNCMVTTLCCGKNPXSTTVSKTETSQVAPA
```

图 1-2　一个 fasta 格式的序列文件

三、SCOP 数据库

蛋白质三维结构信息可以在一定程度上揭示其功能和进化历程。蛋白质结构分类数据库 SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/)是对 PDB 数据库中已知三维结构的蛋白质进行分类，并描述蛋白质结构和进化之间关系一个非常具有代表性的数据库，分成家族(family)、超家族(super family)、折叠(fold)和类型(class)四个层次。SCOP 数据库中的蛋白质序列及三维结构信息可以通过其网站中的 ASTRAL 页面 (http://astral.berkeley.edu/) [9] 链接下载，其中既提供蛋白的 PDB 结构文件，同时也提供 fasta 格式的序列文件。图 1-2 就是一个 fasta 格式序列的例

子。但是，SCOP 数据库中的不同层次之间的区分界限并不十分严格，通常层次越高，越能清晰地反映结构及进化的相关性。下面将简要介绍这四个层次。根据 SCOP 数据库网站上的介绍：属于 SCOP 数据库同一家族的蛋白质成员序列的相似性程度在 30%以上，而且同一家族的蛋白质之间有比较明确的进化关系。但在某些情况下，尽管序列的相似性很低，例如某些球蛋白之间的序列全同率(sequence identity)虽然只有 10%，也可以从结构和功能相似性上推断它们来自共同祖先，这些序列相似性低但又同源的序列一般用来分析蛋白质弱同源性。超家族中的蛋白一般是结构和功能上都有一定的相似性。无论有无共同的进化起源，只要二级结构单元具有相同的排列和拓扑结构，即认为这些蛋白质具有相同的折叠方式。在这些情况下，结构的相似性主要依赖于二级结构单元的排列方式或拓扑结构。SCOP 数据库定义蛋白质结构类型（classes），主要包括 α-螺旋蛋白、β-折叠蛋白、α/β 结构域（主要由平行的 β-折叠片层和 α-螺旋构成）、α+β 结构域（主要由反平行的 β-片层结构和 α-螺旋构成）。SCOP 数据库由英国医学研究委员会(Medical Research Council)的分子生物学实验室和蛋白质工程研究中心维护。SCOP 数据库的分类主要依赖于结构生物学家的专业人工判断。由于蛋白质结构种类繁多，所以人工构建蛋白质结构分类数据库是一项十分复杂的工作，因此 SCOP 数据库的版本更新速度比较慢。同时，SCOP 数据库提供了根据不同的序列全同率和 *E-value* 阈值筛选子数据集的功能。这项功能常用于构建数据集，评价折叠识别算法识别弱同源蛋白的性能，寻找目标蛋白的序列不相似但为同源蛋白的算法的性能。

四、CATH 数据库

CATH[10] 数据库是与 SCOP 数据库相提并论的另一个著名的蛋白质结构分类数据库，有许多研究是基于 SCOP 及 CATH 两个数据库之间的差异展开的。CATH 数据库名称来自英文拼写(Class, Architecture, Topology and Homologous)，其含义为类型(Class)、构架(Architecture)、拓扑结构(Topology)和同源性(Homology)。CATH 数据库由英国伦敦大学开发和维护。与 SCOP 数据库不同，CATH 数据库的构建在使用计算机程序 SAP 结构比对软件的同时，也进行专家人工手动检查。CATH 数据库的分类标准之一是由 α-螺旋和 β-折叠形成的超二级结构排列方式，如同建筑物的立柱、横梁等主要部件，这一层次的分类主要依靠人工方法。CATH 的分类同时也考虑到序列的相似性。目前，CATH 数据库可以通过网址 http://www.cathdb.info 访问。

五、FSSP 数据库

FSSP 是基于蛋白质结构相似家族构建的一个数据库 (Families of Structurally Similar Proteins)[11, 12]。FSSP 数据库最早由 Holm 和 Sander 开发。目前，该数据库由欧洲生物信息学研究所 EMBL-EBI 的研究人员进行维护和进一

步开发(http://www.sander.ebi.ac.uk/dali/fssp/)。该数据库中的序列比对数据是基于蛋白质结构比对软件 Dali 计算得到的，其中的多序列比对结果可以用于分析不同蛋白质家族的结构保守性。用户可以从 FSSP 数据库中查询到不同蛋白的结构邻居(structural neighbours)以及基于邻居蛋白的多序列比对结果。

六、HOMSTRAD 数据库

HOMSTRAD[13] 数据库(http://tardis.nibio.go.jp/homstrad)是一个蛋白同源家族数据库，该数据库同时提供了基于结构的比对数据。HOMSTRAD 数据库建立之初仅包含 130 个蛋白家族的结构比对数据，现在该数据库数据量已经远远超过这个数目。HOMSTRAD 数据库的结构比对数据已经被用在蛋白质折叠方法 FUGUE[14] 程序中。HOMSTRAD 数据库中结构比对过程是采用结构比对软件 COMPARER 进行的，而且该数据库还根据蛋白质序列及结构特点把蛋白聚成不同的类别，并采用 JOY[15] 程序对其建立的序列比对结果进行可视化的展示。

七、SwissProt 数据库

SwissProt 数据库由瑞士日内瓦大学的研究人员于 1986 年开发和构建。该数据库有专门的专家团队支持，负责从科学文献中搜集、整理、分析蛋白质序列的功能信息，并注释和发布经过整理的数据。该数据库同时包含与 EMBL、NCBI、DDBJ、PDB、Prosite 和 PRINTTS 在内的多个数据库的交叉引用信息。从这个数据库中可以获得关于已知功能蛋白的详细信息。例如，G 蛋白偶联受体(GPCR)的重要功能位点和跨膜区位置信息就可以从 SwissProt 数据库中准确获得。目前，SwissProt 数据库由瑞士生物信息学研究所 SIB 和欧洲生物信息学研究所 EBI 共同维护和更新。

八、NR 数据库

NR 数据库一般是特指由 NCBI 提供的非冗余蛋白质序列数据库。NR 取自英文 Non-Redundant，即非冗余的意思。NR 数据库由包括 RefSeq、PDB、SwissProt、PIRH 和 PRF 在内现有已存在序列数据库中所有不相同的序列组成。该数据库一般以 fasta 格式存储序列，可以从网址 ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz 中下载得到其全部序列。虽然 NR 是非冗余数据库的英文缩写，但是该数据库仍然存在不少相同和高相似性的序列。因此，在实际研究中会采用一些过滤手段，例如使用 CD-HIT[16] 程序以一定的阈值去除掉高相似性的冗余序列。通常研究人员会使用 PSI-BLAST 搜索 NR 数据库来构建序列谱(Profile)，该数据可以直观反映一个蛋白及其家族在氨基酸分布频率以及序列进化上的信息。