第 1-2 章 分子数据库

第一节 分子数据库概述

一、分子数据库概念

分子生物信息数据库是生命科学数据信息库的集合,种类繁多(表 1-2.1),主要有基因组、核酸和蛋白质三类一级数据库,生物大分子三维空间结构数据库,以及以上述三类一级数据库和文献为基础的二级数据库。初级数据库储存原始的基础生物数据资源,如 DNA 序列、由晶体衍射(crystallography)获得的蛋白质结构等。二级数据库则是在初级数据库和相关文献等数据基础上经加工和增加相关信息,构建具有特殊生物学意义和专门用途的数据库,如真核生物启动子序列库 EPD 和蛋白质一般结构或功能域数据库 PROSITE。

表 1-2.1 分子生物信息数据库总表

数据库类型	主要数据库
核酸数据库	ENA、GenBank、DDBJ
蛋白数据库	SWISS-PROT、PIR、PDB
基因组数据库	GDB、GenBank、Ensembl

一个数据库记录(entry)一般由两部分组成:原始序列数据和描述这些数据生物学信息的注释(annotation)。注释中包含的信息与相应的序列数据同样重要,对于那些从自动测序仪中出来的序列,我们往往只知道它们来自何种物种何种细胞类型,而其它方面却知之甚少。试想,如果你在确定一段未知蛋白质序列的功能时,发现了不了与该序列有关的任何有关功能的信息时,你的研究工作便会变得更加困难。

不同数据库的注释质量差异很大,因为一个数据库往往要在数据的完整性和注释工作量之间寻找一个平衡点。一些数据库提供的序列数据很广,但其提供的序列注释信息不大;相反,一些数据库数据面较窄,但它提供了非常全面的注释。数据库记录的注释工作是一个动态过程,新的研究成果被不断补充进去。另外需要注意的是,不是数据库所有的信息都是正确的,生物信息数据库中总会有一定的错误率,即一小部分的记录(包括原始序列数据和注释)是不正确的,这是一个无法避免的事实,尤其不同数据库,不同基因组版本对应注释的起始与终止,以及位置起始偏移量,有的数据库以 1 为起始,有的数据库以 0 为起始,有时需要转换。

二、数据库记录格式

所谓格式是对信息描述的统一规范,规范的格式为数据的收集、整理、交流和应用提供了方便。分子生物信息数据库的格式很多,较为常见包括 FASTA、FASTQ、GBFF、GFF 格式,下面主要对这几种格式进行介绍。

FASTA 格式主要分两部分(图 1-2.1 为 FASTA 格式实例),第一部分即首行,为描述行,

以“>”为起始,后跟这段序列的描述信息;首行之后即为原始序列。

```
>gi|332309159|ref|NM_181652.2| Homo sapiens peroxiredoxin 5 (PRDX5), nuclear gene encoding
mitochondrial protein, transcript variant 3, mRNA
CGCGCCTGCGCAGTGGAGGCGGCCCAGGCCCGCCTTCCGCAGGGTGTCGCCGCTGTGCCGCTAGCGGTGC
CCCGCCTGCTGCGGTGGCACCAGCCAGGAGGCGGAGTGGAAGTGGCCGTGGGGCGGGTATGGGACTAGCT
GGCGTGTGCGCCCTGAGACGCTCAGCGGGCTATATACTCGTCGGTGGGGCCGGCGGTCAGTCTGCGGCAG
CGGCAGCAAGACGGTGCAGTGAAGGAGAGTGGGCGTCTGGCGGGGTCCGCAGTTTCAGCAGAGCCGCTGC
AGCCATGGCCCCAATCAAGGTTCGGCTCCTGGCTGATCCCACTGGGGCCTTTGGGAAGGAGACAGACTTA
TTACTAGATGATTCGCTGGTGTCCATCTTTGGGAATCGACGTCTCAAGAGGTTCTCCATGGTGGTACAGG
ATGGCATAGTGAAGGCCCTGAATGTGGAACCAGATGGCACAGGCCTCACCTGCAGCCTGGCACCCAATAT
CATCTCACAGCTCTGAGGCCCTGGGCCAGATTACTTCCTCCACCCCTCCCTATCTCACCTGCCCAGCCCT
GTGCTGGGGCCCTGCAATTGGAATGTTGGCCAGATTTCTGCAATAAACACTTGTGGTTTGCGGCCATCTC
CTTGGTTAAAAAAAAA
```

图 1-2.1　FASTA 格式

FASTQ 格式(图 1-2.2)与 FASTA 格式类似,但是 FASTQ 比 FASTA 多了序列的质量信息。一般情况下,FASTQ 的一条序列有四部分信息(即四行一序列):

第一行一般包含序列的名称等其他描述信息,以“@”开头。

第二行即序列的具体碱基信息。

第三行的内容与第一行的内容相同,但是以“+”开头;有时候“+”后面的内容可以省略,但是“+”一定不能省略。

第四行是序列的质量信息(quality value,即测序的质量评价),与第二行的碱基序列一一对应。

```
@ST-E00169:117:HG2T3CCXX:4:1101:23794:1432 1:N:0:NTGTCA
NTCTAAGAATTCAATGGCTTTAAGAAGGCTTAGATAGCGAGTATTCTCAGAATTCAACGATATCTTCGCATTATCGGTTACTAGACG
+
#AA<FFFKKFKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKFKKKKFFFA<FKKKKKKFFAKKFKKFKKK
```

图 1-2.2　FASTQ 格式

FASTQ 质量编码格式有如下几种:Sanger 格式、Illumina1.8+格式以及 Solexa 格式(即 Illumina1.0+、Illumina1.3+和 Illumina1.5+格式),注意不同格式编码表示范围不同(详见 https://en.wikipedia.org/wiki/FASTQ_format)。图 1-2.2 列举为 Sanger 编码格式,第四行中每个字符对应的 ASCII 值减去 33,即为对应第二行碱基的测序质量值。对于 Illumina1.8+之前版本,则需要每个字符对应的 ASCII 值减去 64,即为对应第二行碱基的测序质量值。如果测序错误率用 e 表示,测序碱基质量值用 Qphred 表示,则有下列关系:

$$\mathrm{Qphred} = -10\log10(e)$$

测序质量与错误率对应关系如表 1-2.2 所示。

表 1-2.2　Illumina 测序质量与错误率对照表

测序错误率	测序质量值	对应字符
5%	13	.
1%	20	5
0.1%	30	?
0.01%	40	I

GFF(general feature format)格式是 Sanger 研究所定义的一种简单方便的数据格式,对 DNA、RNA 以及蛋白质序列的特征进行描述。图 1-2.3 为 GFF 格式实例,具体每列的详细解释如下:

ZD39 生物信息学　兴邦二校　成品 185mm×260mm　39 行×39 字　10.5SS

```
##gff-version 3
##sequence-region ctg123 1 1497228
seqid  source  type start  end score strand phase attributes
ctg123 . gene 1000 9000 . + . ID=gene00001;Name=EDEN
ctg123 . TF_binding_site 1000 1012 . + . Parent=gene00001
ctg123 . mRNA 1050 9000 . + . ID=mRNA00001;Parent=gene00001
ctg123 . mRNA 1050 9000 . + . ID=mRNA00002;Parent=gene00001
ctg123 . mRNA 1300 9000 . + . ID=mRNA00003;Parent=gene00001
ctg123 . exon 1300 1500 . + . Parent=mRNA00003
ctg123 . exon 1050 1500 . + . Parent=mRNA00001,mRNA00002
```

图 1-2.3 GFF 格式

列 1:序列的 ID(seqid),一般是序列的名称,如 scaffold 编号或者是染色体号。

列 2:注释软件来源(source),若没有则用点代替。

列 3:注释的类型(type),mRNA、CDS、gene 等。

列 4 和 5:对应序列的起始位置和终止位置(start & end)。

列 6:得分(score),序列相似性比对时的 E-value 值或者基因预测时的 P-value 值,“.”表示为空。

列 7:序列方向(strand),问号表示未知,正负号代表正反链。

列 8:相位(phase),表明 CDS 或可编码的外显子的相位。

列 9:群(attributes),表明附属关系,也可用作注释用途。

GBFF 格式(GenBank file format)为 GenBank 数据库使用的记录格式。其他相关数据库,如 ENA 数据库记录格式类似,就不再做介绍。GBFF 格式整体分为三个部分,分别为描述符部分、注释部分和序列部分如表 1-2.4 所示。

描述符部分包括了整个记录的相关信息,比如位置(LOCUS)、定义(DEFINITION)、检索号(ACCESSION)、版本(VERSION)、关键词(KEYWORDS)、来源(SOURCE)、参考文献(REFERENCE)等。

注释部分(FEATURES)描述基因和基因的产物以及与序列相关的生物学特性。对该序列的 mRNA、CDS 等进行描述,通过点击带有下划线的 mRNA、CDS、gene 等字符,可以在序列部分查看到相关的注释信息。

第三部分为序列部分(ORIGIN),即核苷酸序列本身。在 GBFF 文件的最后,以类似于 FASTA 格式的方式给出了所记录的序列。末尾的“//”是结束符,所有 GBFF 格式序列数据库记录中最后一行都以“//”结尾。

三、数据库冗余、序列递交和检索

1. 数据库的冗余

在进行 DNA 和蛋白质序列分析时,碰到的一个棘手问题是数据库的冗余(redundancy)。DNA 和蛋白质数据库中的很多记录是属于同一基因和蛋白质家族,或在不同生物体上发现的同源基因。不同的研究机构可能向数据库递交了相同的序列数据,如果没有被检查出来,则这些记录或多或少地紧密相关。当然,这些记录如果的确非常相近,可以认定它们是相同序列,但一些显著的差异可能是由于基因组多样性的结果,导致看似序列不同实则相同。

冗余数据至少可能导致以下 3 个潜在的错误:一是如果一组 DNA 或氨基酸序列包含了大量非常相关序列族,则相应的统计分析将偏向这些族,在分析结果中,这些族的特性被夸

图 1-2.4　GBFF 格式

大;二是序列间不同部分的显著相关,可能是在数据样本抽样时是有偏的和不正确的;最后是如果这些数据是被用于预测,则这些序列将使预测方法发生偏离。

基于以上原因,必须避免在数据库中存在太过于相似的序列。很多数据库通过全局序列联配以及人工复查等方式,使数据库为非冗余(non-redundant, nr)。例如,应用比较广泛且数据比较齐全的 NCBI 的 NR 数据库,包括 GenBank 的 CDS 翻译序列、RefSeq、PDB 等等内容(详见表 1-2.3)。这些数据库去除了其中多数冗余序列,但要真正做到百分之百无冗余是困难的,而且一点点的冗余度对于我们大多数使用者的查询来说并不会带来太多影响,尤其在数据库相对庞大时。

表 1-2.3　GenBank NR 数据库数据来源情况

非冗余数据库(NR)	非冗余的 GenBank CDS 翻译序列+RefSeq+PDB+SwissProt+PIR+PRF, 同时包括 PAT, TSA 和 env_nr 等来源序列
参考序列(RefSeq)	序列来自 NCBI 参考序列库
SWISS-PROT	蛋白序列来自 UniProtKB/SWISS-PROT 最新版本的蛋白数据库
PAT	蛋白序列来自 NCBI 的专利蛋白数据库
PDB	蛋白序列来自 PDB 三维结构记录
env_nr	蛋白序列翻译自环境基因组核苷酸序列中的 CDS 注释
tsa_nr	蛋白序列翻译自转录组拼接序列中的 CDS 注释
PIR	蛋白序列来自已经注释蛋白序列数据库
PRF	蛋白序列来自 PRF(Protein Research Foundation)最新版本数据库

对于生命科学研究而言,初始序列是待挖掘的金矿,所以序列的质量关系到生命科学研究者是否能够挖到金矿。而初始序列数据的偏差或错误(artifacts)主要来自实验过程,这与其它科学数据的情况相同。这些错误主要来自以下几个方面:

①载体序列污染:在测序列等实验过程中,载体序列可能造成污染,致使序列记录数据中包含了载体序列。

②异源(heterologous)序列污染:有研究表明一些人类转录组测序结果在实验过程中被酵母和细菌序列污染。

③序列的重排和缺失。

④重复序列污染:cDNA 克隆方法有时会受到逆转录因子(如 Alus)的影响。

⑤测序误差和自然多态性:测序过程存在一定的误差概率。

去除载体污染是获得准确干净序列最关键的一步。有一些去除污染的专门软件和工具,如 NCBI 的 VecScreen 网站便提供了去除载体污染的在线服务(http://www.ncbi.nlm.nih.gov/tools/vecscreen/)。VecScreen 能够快速地发现核苷酸序列中可能的载体片段,这能够帮助科研工作者在分析前或者上传序列前快速鉴定和移除载体污染片段。VecScreen 基于 UniVec(非冗余载体数据库),UniVec 数据库也随着 NCBI 的扩充而不断更新,紧跟科研工作者的需求。

2. 向数据库递交序列数据及其说明

本节将简单介绍如何向相关数据库发送自己的序列数据,如何准确、全面地表述生物信息学研究的"材料与方法"。

许多学术期刊在发表含有序列数据的论文时,均要求作者先将该序列发送并存储到相应数据库中。这些数据库的主页上均有详细的发送说明,按照要求操作即可。数据库往往特别要求发送者要注意去除载体污染,例如 NCBI 提供了 VecScreen 的相关服务(网址见上节)。序列的发送可以通过网上进行。发送序列前都需要在上传网站进行注册。GenBank 有多种可以选择的发送系统,如 BankIt、Sequin、tbl2asn、Submission Portal、Barcode Submission Tool 等。其中 BankIt、Submission Portal 和 Barcode Submission Tool 是自动向 NCBI 发送序列的,而 Sequin 和 tbl2asnb 必须向 gbsub@ ncbi.nlm.nih. gov 发送邮件进行说明,如果序列文件过大超过邮件可上传界限,应该直接上传至 Sequin MacroSend。Sequin(http://www.ncbi.nlm.nih.gov/Sequin/index.html)工具适用于多平台(Mac/pc/unix),由 NCBI 独立开发,适用于 ENA、GenBank 和 DDBJ 数据库的发送服务。具体发送格式和要求可到这些网站上查获。一旦数据被接收,一个记录号(对应于发送的数据)将产生并发送给发送者,该记录号即本数据在数据库中的索引号,可用于论文发表(论文中需注明记录号)与查询。其中 ENA 的序列优先上传系统为 WEBIN(http://www.ebi.ac.uk/ena/submit),它除了可进行一般大小的序列数据发送外,还可进行大批量的数据发送(Bulk submission)。

试验结果的可重复性是科学研究的一个重要特征。为了保证生物信息学研究结果的可重复性,准确、全面的"材料与方法"说明比其它学科显得更为重要和严格。一份清楚、准确的"材料与方法"说明应包括:

①数据库的名称:SWISSPROT、PIR、GenBank、ENA、dbEST 等等,不应是以类别(蛋白、核酸、序列等)说明。

②数据库的版本(Version):数据库的更新速度远快于期刊的发行速度,所以严格注明所

用数据库的版本;如果你的检索是实时的,则注明最后检索的日期等。

3. 数据库检索与序列搜索

许多系统可以为使用者提供简便的序列库信息查寻服务,其中最著名和操作性最强的 2 个系统是 Entrez(由 NCBI 创建)和 SRS(sequence retrieval system,由 EMBL TheoreEtzold 建立)。下面即以 Entrez 为例讲解如何在数据库中进行检索和序列搜索。

Entrez 是一个基于 Web 界面的综合生物信息数据库检索系统。用户不仅可以方便地检索 GenBank 的核酸数据,还可以检索来自其它数据库的蛋白质序列数据、基因组图谱数据、分子模型数据库(MMDB)的蛋白质三维结构数据、种群序列数据集以及由 PubMed 获得 Medline 文献数据。

在 NCBI 主页默认 All Databases 时点击搜索框右边的 Search 进入。如图 1-2.5 所示,在搜索栏输入你要查找的关键词,点击“GO”即可开始搜索。如果输入多个关键词,它们之间默认的是“与”(AND)的关系。搜索的关键词可以是一个单词、短语、句子、数据库的识别号、基因名字等等,但必须明确,不能是“gene”、“protein”等没有明确指向的词语。

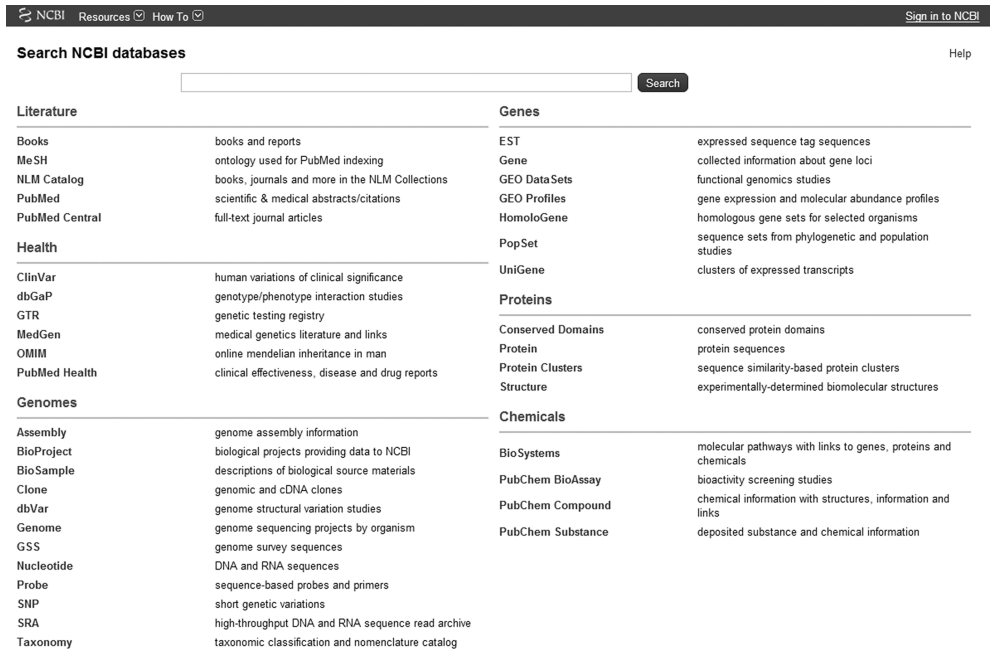

图 1-2.5　NCBI 数据库 Entrez 搜索首页

输入关键词,点击“Search”之后,每个数据库图标前方出现了数字,代表在相对应数据库里搜索到的条目数。点击进入对应的数据库,可以查看搜索到的条目。如果数据库前面显示“0”,说明在对应的数据库里没有搜索到任何结果。也可以在 NCBI 任一页面上的搜索栏里输入关键字,点击搜索框前面的下拉菜单,选择数据库,点击“Search”即可(图1-2.6)。

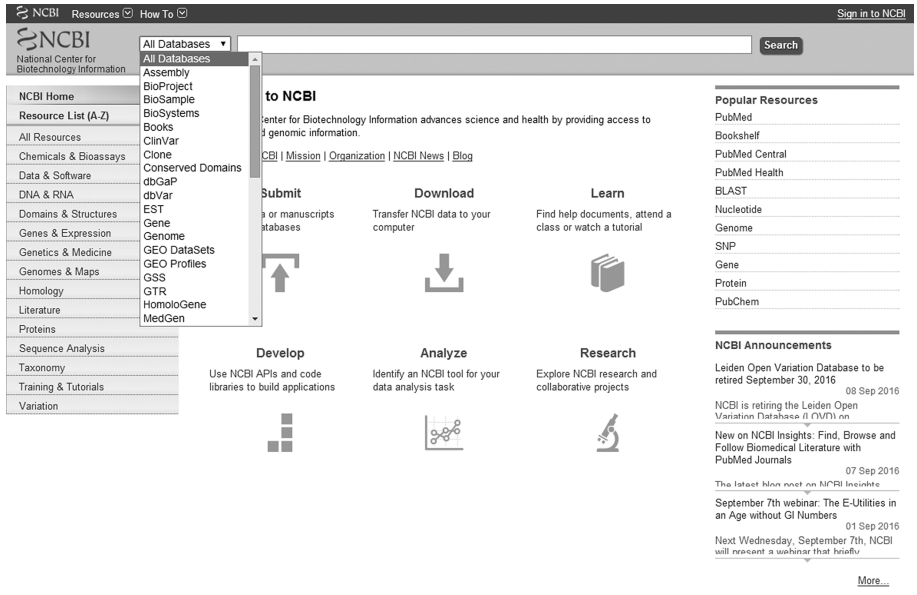

图 1-2.6 NCBI 数据库选项下拉界面

但是这种简单搜索会产生大量的结果,其中很多信息不是我们所需要的。为此,NCBI 提供了“Limits”、“Advanced Search”等辅助功能,只有充分理解并熟练运用这些工具进行检索,才能充分发挥 Entrez 的强大功能,实现精确高效的检索。限制性(Limits)搜索和高级(Advanced)搜索结构可以根据该数据库结构,将输入的关键词的查询范围限制在某个范围内,如领域、物种、分子类型等等。不同的数据库,其限定内容略有不同(图 1-2.7)。

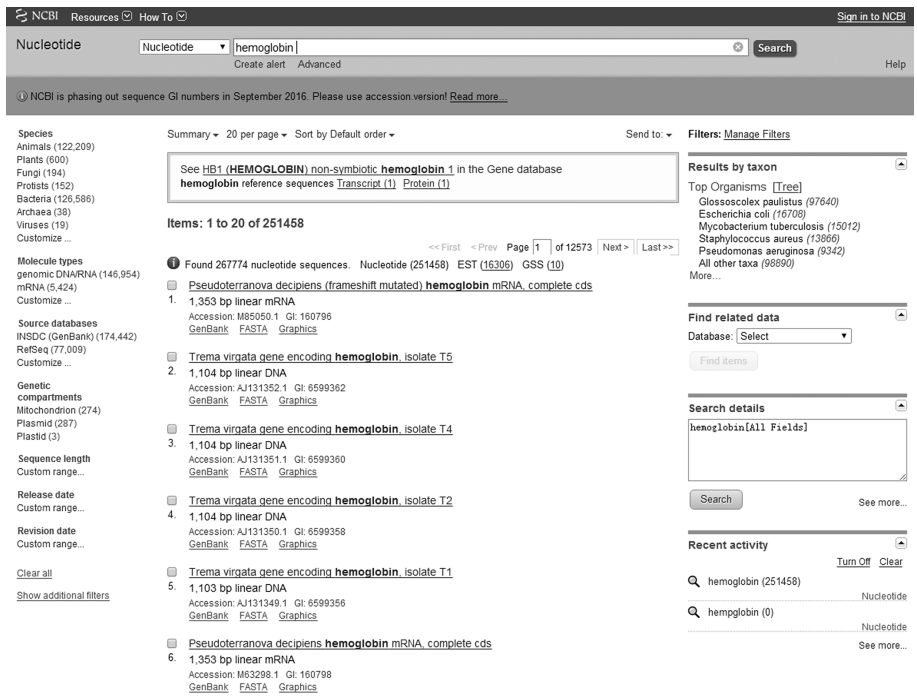

图 1-2.7 GenBank 数据库搜索结果界面

第二节　核苷酸及其相关数据库

一、DNA/RNA 序列数据库

DNA/RNA 序列构成了初级数据库的主体部分。目前国际上有三个主要核苷酸序列公共数据库(表 1-2.4):位于英国剑桥的欧洲分子生物学实验室(EMBL)下的欧洲核苷酸档案库(European Nucleotide Archive, ENA);位于美国国家卫生研究院(NIH)的美国国家生物技术信息中心(NCBI)下的 GenBank 数据库;日本 DNA 数据库(DNA Databank of Japan, DDBJ)。这三个大型数据库于 1988 年达成协议,组成合作联合体。它们每天交换信息,并对数据库序列记录的统一标准达成一致。每个机构负责收集来自不同地理分布的数据(ENA 负责欧洲,NCBI 负责美洲,DDBJ 负责亚洲等)。然后来自各地的所有信息汇总在一起,三个数据库共享并向世界开放,故这三个数据库又被称为国际公共序列数据库(public sequence database)。所以从理论上说,这三个数据库所拥有的序列数据是完全相同的,但是由于同步时间的关系,这些数据库之间的记录可能有一定差异。

表 1-2.4　国际公共核苷酸序列数据库网址

数据库(Database)	网址(Address)
GenBank	http://www.ncbi.nlm.nih.gov/Genbank
ENA	http://www.ebi.ac.uk/ena
DDBJ	http://www.ddbj.nig.ac.jp

核苷酸序列数据库的增长呈爆炸式。ENA 核酸数据近 35 年的增长情况表(表 1-2.5)充分说明了这一点。2015 年 12 月 ENA(Release 126)的 DNA 碱基对数已接近 15,000 亿,序列数超过 6 亿条,均为 2005 年的 10 倍以上,而 1995 年的数据仅是其一个零头了。可见近 20 年的生物分子大数据增长非常迅速。从历史看,每 22 个月,数据库的数据规模翻一翻(图 1-2.8a),其中以全基因组鸟枪法测序所占比重最大,超过一半(图 1-2.8b)。数据库的膨胀对于我们进行数据库搜索非常有好处。也许这个月还找不到一个匹配序列,但可能在下一次更新的数据中寻获。所以,当描述生物信息学分析时,务必要注明当时所使用序列数据库的数据状况及时间。

表 1-2.5　ENA 数据库序列数据库每十年增长情况

数据库报告(Release)	释放日期(Month)	记录数(Entries)	核苷酸数(Nucleotides)
1	06/1982	568	585 433
7	12/1985	5 789	5 622 638
43	06/1995	420 111	315 840 053
85	12/2005	64 739 883	116 106 677 726
126	12/2015	668 347 471	1 496 520 157 048

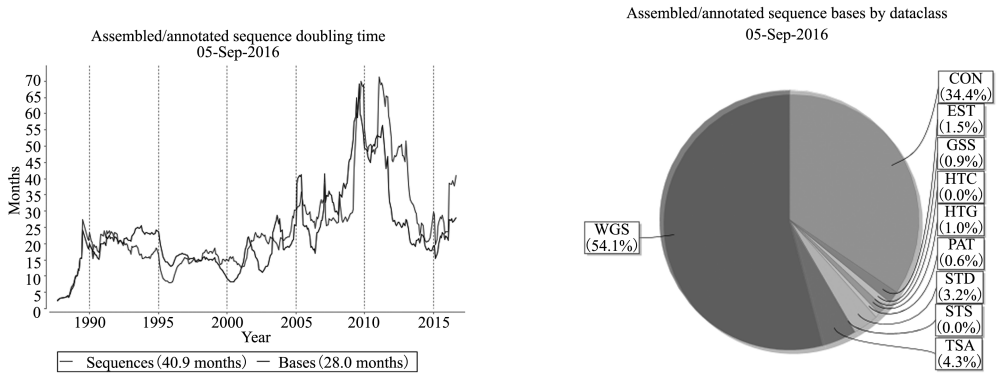

图 1-2.8 ENA 已经注释或用于拼接序列总量翻番时间趋势(a)和数据类别分布情况(b)

为了有效地管理如此庞大的数据,数据库数据分别以物种(taxonomic division)和数据类别(data class)进行分类,每个记录都被严格地归入某一类中。每一类用了 3 个字母代码表示(表 1-2.6,表 1-2.7)。这些分类并非一成不变,随着时间的推移可能进行一定的修正,如新加入的高通量测序数据(HTG)等。ENA 和 GenBank 等数据库的使用手册均可在相应的网址(表 1-2.4)上找到,这些手册提供了详尽的数据库组成、分类等细节。

表 1-2.6 ENA 数据库 2015 年 12 月以物种划分数据情况(Release126)

类别(Division)		代码(Code)	记录数(Entry)	核苷酸数(Nucleotide)
环境样品	Environmenta Samples	ENV	113 124 175	106 004 477 077
真菌	Fungi	FUN	17 554 414	41 693 568 682
人	Human	HUM	30 701 139	91 755 265 972
无脊椎动物	Invertebrates	INV	122 715 096	193 220 407 457
其他哺乳动物	Other Mammals	MAM	61 602 613	306 977 883 053
小鼠	Musmusculus	MUS	13 334 120	23 231 575 691
噬菌体	Bacteriophage	PHG	12 257	227 103 197
植物	Plants	PLN	196 451 384	257 969 981 990
原核生物	Prokaryotes	PRO	10 912 565	212 102 974 767
啮齿类动物	Rodents	ROD	10 860 305	60 621 535 857
合成的	Synthetic	SYN	6 088 260	1 733 688 566
未分类的	Unclassified	UNC	10 875 727	6 084 119 311
病毒	Viruses	VRL	2 021 770	2 733 902 891
其他脊椎动物	Other Vertebrates	VRT	71 808 213	191 304 650 548
总和	Total		668 347 471	1 496 520 157 048

表 1-2.7　ENA 数据库 2015 年 12 月以数据种类划分数据情况(Release126)

数据种类(Data class)		代码(Code)	记录数(Entry)	核苷酸数(Nucleotide)
构建的序列	Constructed Sequence (contig upnards)	CON	37 522 080	902 376 693 011
表达序列标签	Expressed Sequence Tag	EST	76 263 441	42 566 131 380
基因组调查测序	Genome Sequence Survey	GSS	39 683 735	25 607 333 433
交通量 cDNA 测序	High Throughput cDNA sequencing	HTC	540 403	630 397 561
交通量基因组测定	High Throughput Genome sequencing	HTG	169 918	26 987 558 587
专利	Patents	PAT	31 911 675	15 797 983 859
标准注释拼接序列	Annotated Assembled Sequence	STD	24 665 513	79 145 914 632
序列标签位点	Sequence Tagged Site	STS	1 346 989	640 842 065
转录组拼接	Transcriptome Shotgun Assembly	TSA	104 353 227	85 543 036 327
全基因组鸟枪法	Whole Genome Shotgun	WGS	351 890 490	1 219 600 959 204
合计	Total		668 347 471	1 496 520 157 048

二、基因组数据库

除了核苷酸序列数据库,另外一个主要的初级数据源来自各种基因组测序计划。基因组数据库的主要内容为收集基因组序列、注释结果并且展示这些序列。目前许多基因组已经测序完成,这些基因组的大部分信息在 ENA、GenBank 等数据库中均可找到。表 1-2.8 是部分已经测序的生物基因组列表网址。

表 1-2.8　已测序生物基因组列表网址

生物分类(Division)	网址(Address)
动物	http://en.wikipedia.org/wiki/List_of_sequenced_animal_genome
植物	http://en.wikipedia.org/wiki/List_of_sequenced_plant_genomes
细菌	http://en.wikipedia.org/wiki/List_of_sequenced_bacterial_genomes
真菌	http://en.wikipedia.org/wiki/List_of_sequenced_fungi_genomes
原生生物	http://en.wikipedia.org/wiki/List_of_sequenced_protist_genomes
古生生物	http://en.wikipedia.org/wiki/List_of_sequenced_archaeal_genomes

人类最早(1977)获得基因组全序列的物种是噬菌体(53Kb)。1987 年自动测序仪问世,第一个病毒基因组序列(1990)在自动测序仪上完成。1995 年一个细菌基因组被完全测序,紧接着是酵母(1996)、多细胞线虫(1998)和果蝇(1999)基因组。2001 年,人类基因组计划(Human Genome Project,HGP)宣告完成。这项由美国、英国、法国、德国、日本和中国六个国家共同参与,历经十年、耗资数十亿美元的人类基因组计划,成为了人类基因研究史上一个

重要的里程碑。基于大规模平行测序(massively parallel sequencing)的第二代测序技术的出现,从此测序速度实现大飞跃,二代测序帮助人们以更低廉的价格,更全面、更深入的分析基因组。近些年,第三代单分子测序技术的出现,其更长的读长倍受青睐。

有一些基因组数据库(表 1-2.9)值得关注。如一些模式生物基因组数据库,如 RGD、EcoCyc、Tair 等,不仅仅只有基因组数据,同样也有其它许多重要的信息,对于对应模式生物的相关研究来说有非常巨大的参考价值。

表 1-2.9 部分基因组数据库网址

数据库	网址	备注
Ensembl	http://www.ensembl.org/index.html	人类,鼠,脊椎动物和真核生物基因组自动注释数据库
Ensembl Genomes	http://ensemblgenomes.org/	细菌,原生生物,真菌,植物以及无脊椎动物后生动物基因组数据库
NCBI genome	https://www.ncbi.nlm.nih.gov/genome	NCBI 整合基因组各类信息包括序列,图谱,染色体,拼装,注释等信息
UCSC Genome Browser	http://genome.ucsc.edu/index.html	脊椎动物模式生物拼装注释以及基因组可视化分析数据库
CAMERA	http://camera.calit2.net/index.php/	微生物基因组和宏基因组资源
The 1000 Genomes Project	http://www.1000genomes.org/	来自不同族群的超过 1 000 个匿名参与者基因组数据库
Personal Genome Project	http://www.personalgenomes.org/	人类个体基因组分享数据库
GDB	http://www.gdb.org	人类基因组原始数据库
RGD	http://rgd.mcw.edu/	鼠表型及基因组数据库
EcoCyc	http://ecocyc.org/	大肠杆菌基因组及转录调控数据库
Flybase	http://flybase.org/	果蝇基因及基因组数据库
ZFIN	http://zfin.org/	斑马鱼信息网络数据库
TAIR	http://www.arabidopsis.org/	拟南芥信息资源数据库
maizegdb	http://www.maizegdb.org/	玉米基因组数据库
BRAD	http://brassicadb.org/brad/	芸苔属基因组数据库
plantGDB	http://www.plantgdb.org/	部分植物基因组数据库

在表 1-2.9 中的基因组数据库中,Ensembl Genomes 是较为常用的数据库之一。下面以 Ensembl Genomes 中的粳稻(*Oryza sativa* ssp. *japonica*)基因组为例,讲解一般基因组数据库的格式。在图 1-2.9 中我们可以看到基因组的基本信息界面,其中包括物种信息、基因组版本以及基因组可视化工具 GBrowse 链接;下载界面包括基因组序列以及注释等相关信息的下载链接。基本信息界面和下载界面是一般基因组数据库都具备的,同时基因组数据库还会提供 BLAST 以及相关在线分析服务。除此之外,Ensembl Genomes 还提供了进行比较基因组学和查询基因组变异的相关服务。

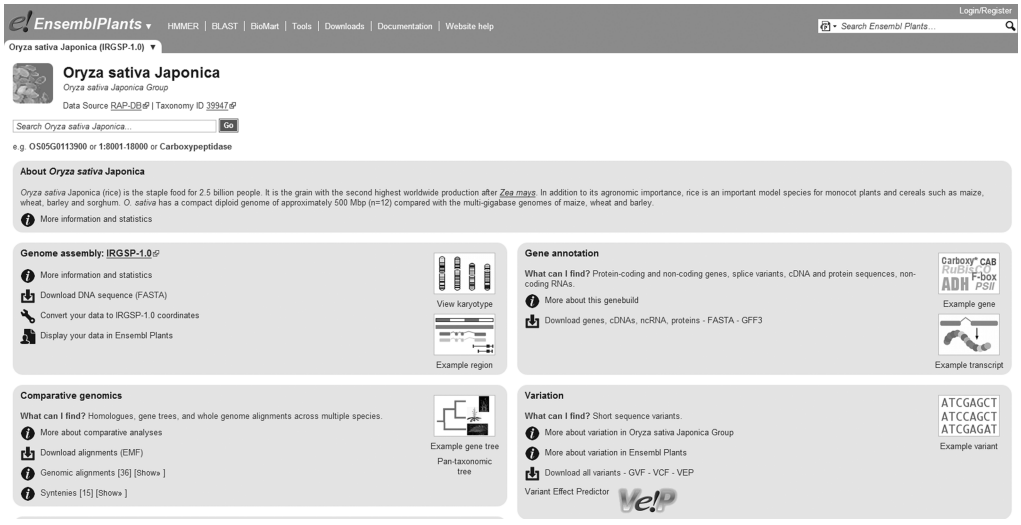

图 1-2.9　Ensembl Genomes 粳稻基因组数据库主页界面

三、非编码 RNA 数据库

非编码 RNA(non-coding RNA),包括 rRNA、tRNA、snRNA、snoRNA 和 microRNA 等它们的共同特点是都能转录但是不翻译成蛋白,在 RNA 水平上就能行使各自的生物学功能。非编码 RNA 从长度上来划分可以分为 3 类:小于 50nt,包括 miRNA、siRNA、piRNA;50nt 到 500nt,包括 rRNA、tRNA、snRNA、snoRNA、SLRNA、SRPRNA 等;大于 500nt,包括 miRNA-like 的长非编码 RNA,以及不带 polyA 尾巴的长非编码 RNA 等。数据库 DASHR(一个人类非编码小 RNA 数据库)将非编码 RNA 分为长非编码(lncRNA)和非编码小 RNA(sncRNA)两类。下面将对非编码小 RNA、长非编码 RNA 以及 RNA 家族等其他 RNA 数据库进行概述。

1. 非编码小 RNA 数据库

miRBase(http://www.mirbase.org/)是一个收录已发表 microRNA 序列及相关注释的数据库。2014 年 6 月发布版本 21,共有 28 645 个条目。数据库中的每个条目均包含 miRNA 的前体序列和成熟序列以及它们的位置,条目均能够通过名字、关键词、文献和注释进行搜索。

piRNAbank(http://pirnabank.ibab.ac.in/)是一个 piRNA 数据库,其中收集的物种只包括人、老鼠以及果蝇。该数据库在基因组上汇总展示了所有可能的 piRNA 簇以及 piRNA 在基因组上相关的元件如基因和重复片段等。

GtRNAdb(Genomic tRNA Database,http://gtrnadb.ucsc.edu/)是一个转运 RNA(tRNA)数据库,数据均通过软件 tRNAscan-SE 在完整基因组或接近完整基因组 tRNA 基因预测自动获得。该数据库收录物种包括真核生物、古生菌和细菌。

SILVA(http://www.arb-silva.de/)是一个核糖体 RNA(rRNA)数据库,同样收录真核生物、细菌和古生菌。RDP 同样也是核糖体 RNA 数据库,但只提供微生物(16S 细菌和古生菌 rRNA 序列以及 28S 真菌 rRNA 序列)。

2. 长非编码 RNA 数据库

lncRNAdb(http://www.lncrnadb.org/)是一个长非编码 RNA 数据库,为真核生物已注释功能的长非编码 RNA。2015 年 11 月更新到版本 2.0,数据库中的条目均手工取自参考文献。在该数据库可以通过检索 lncRNA 的名字或者进行序列 BLAST 搜索,有必要的情况下可以在检索中设置物种。

LncRNAWiki(http://lncrna.big.ac.cn/index.php/Main_Page)数据库为人类长非编码 RNA 数据库。截至 2014 年 7 月数据库共收录了 105 257 个人类长非编码 RNA,来源于 GENCODE(包含 23 898 个人类 lncRNA 转录本)、NONCODE(包含 95 135 个人类 lncRNA 转录本)和 LNCipedia(包含 32 181 个人类 lncRNA 转录本)这三大数据库,经过错误和冗余去除后,得到 105 257 个人类 lncRNA。

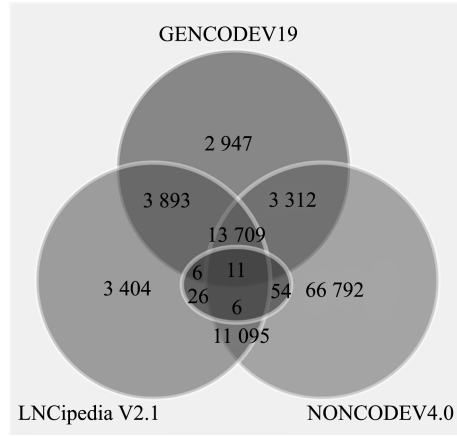

图 1-2.10 LncRNAWiki 数据库收集的人类长非编码 RNA 数据(LncRNAWiki2014)

3. RNA 家族等其他 RNA 数据库

Rfam(http://rfam.xfam.org/)是一个包含非编码 RNA(ncRNA)家族以及其他一些 RNA 元素家族的数据库,该数据库目前由 EMBL-EBI 维护。Rfam 数据库类似于著名的 Pfam 数据库(Pfam 数据库旨在注释蛋白家族),是为了 RNA 家族注释。但是与蛋白质结构上所不同的是,同一家族的 RNA 通常是相似的二级结构而不是相似的初级序列。RNA 家族在 Rfam 进行搜索的方式与 Pfam 有所不同,主要方式以联配、搜索识别二级结构的一致性以及协方差模型对 ncRNA 家族进行匹配搜索。目前 Rfam 的版本为 12.1,发布于 2016 年 4 月,共有 2 474 个 RNA 家族。

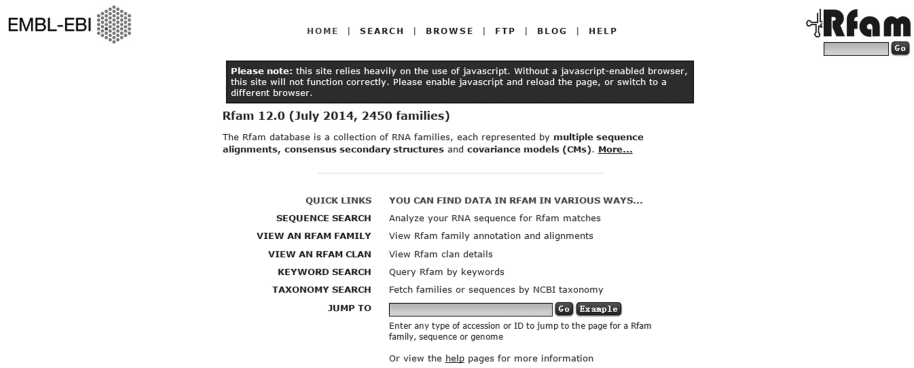

图 1-2.11 RNA 家族数据库 Rfam 主页界面

第三节　蛋白质及其相关数据库

1. 蛋白质序列数据库

SWISS-PROT 和 PIR 是国际上两个主要的蛋白质序列数据库,目前这两个数据库在 EMBL 和 GenBank 数据库上均建立了镜像(mirror)站点。Swiss-Prot 数据库包括了从 EMBL 翻译而来的蛋白质序列,这些序列经过人工检验和注释。该数据库主要由日内瓦大学医学生物化学系和欧洲生物信息学研究所(EBI)合作维护。Swiss-Prot 的序列数量呈直线增长。Swiss-Prot 的数据存在一个滞后问题,即把 EMBL 的 DNA 序列准确地翻译成蛋白质序列并进行注释需要时间,一大批含有开放阅读框(ORF)的 DNA 序列尚未列入 SWISS-PROT。为了解决这一问题,TrEMBL(Translated EMBL)被建立了起来。TrEMBL 也是一个蛋白质数据库,它包括了所有 EMBL 库中的蛋白质编码区序列,提供了一个非常全面的蛋白质序列数据源,但这势必导致其注释质量的下降。目前,Swiss-Prot 和 TrEMBL 已经合并为 UniProtKB 数据库(Universal Protein knowledgebase),2016 年 10 月释放的版本包括手工注释经审核的 Swiss-Prot 条目 552 884 条,自动注释未经审核的 TrEMBL 条目 70 656 157 条。PIR 数据库的数据由美国国家生物技术信息中心(NCBI)翻译自 GenBank 的 DNA 序列。PIR 根据序列注释程度(质量)分为 4 个等级(表 1-2.10)。

表 1-2.10　PIR 数据库的分类情况(Release 80)

分类名称(Name)	说明(Comment)	记录数(Number of entries)
PIR1	分类并注释(Classified and annotated)	20 685
PIR2	注释(Annotated)	262 300
PIR3	未核实(Unverified)	24
PIR4	未翻译(Unencoded or untranslated)	407

表 1-2.11 列出了以上主要蛋白质序列数据库的网址,有关详情可到这些网站上获得。

表 1-2.11　主要蛋白质序列数据库网址

数据库(Database)	网址(Address)
UniProt	http://www.uniprot.org
PIR	http://pir.georgetown.edu/

2. 蛋白质结构数据库

实验获得的三维蛋白质结构均储存在蛋白质结构数据库中。PDB 是国际上主要的蛋白质结构数据库(图 1-2.12),虽然它没有蛋白质序列数据库那么庞大,但其增长速度也很快。该数据库储存有由 X 射线和核磁共振(NMR)确定的结构数据。NRL-3D 数据库提供了储存在 PDB 库中蛋白质的序列,它可以进行与已知结构蛋白质序列的比较。对 PDB 中每个已知三维结构的蛋白质序列进行多序列同源性比较(multiple sequence alignment)的结果,被储存在 HSSP(homology-derived structures of proteins)数据库中。被列为同源的蛋白质序列很有可能具有相同的三维结构,HSSP 因此根据同源性给出了 SWISS-PROT 数据库中所有蛋白质序列最有可能的三维结构。要想了解对已知结构蛋白质进行等级分类的情况可利用

SCOP(structural classification of proteins)数据库,在该库中可以比较某一蛋白质与已知结构蛋白的结构相似性。CATH 是与 SCOP 类似的一个数据库。上述数据库网址见表 1-2.12。

表 1-2.12 主要蛋白质结构数据库网址

数据库(Database)	网址(Address)
PDB	http://www.rcsb.org/pdb
NRL-3D	http://pir.georgetown.edu/pirwww/search/textnrl3d.html
HSSP	http://www.sander.embl-heidelberg.de/hssp
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop
CATH	http://www.biochem.ucl.ac.uk/bsm/cath

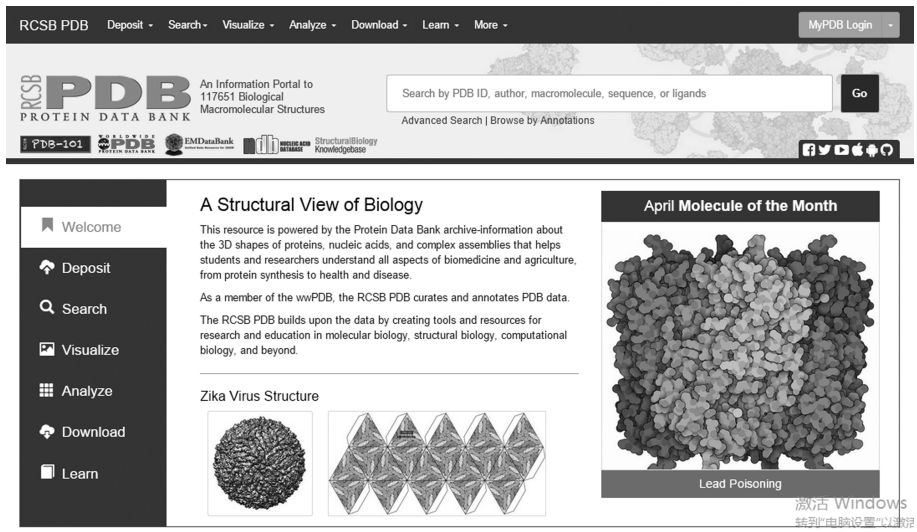

图 1-2.12 蛋白质结构数据库 PDB 主页界面

3. 蛋白质组数据库

蛋白质组(proteome)一词,意指“一个基因组所表达的全套蛋白质”,即包括一个细胞乃至一种生物所表达的全部蛋白质。1995 年 Marc Wikins 首次提出蛋白质组的概念。1997 年,Peter James 又在此基础上率先提出蛋白质组学的概念,基因组学和蛋白质组学的概念又进一步催生了各种各样的组学(omics)。

蛋白质组鉴定数据库(proteomics identification database,PRIDE,http://www.ebi.ac.uk/pride/archive/)是欧洲生物信息研究所建立的主要基于质谱数据的蛋白质组学数据库。PRIDE 允许研究者们存储、分享并比较他们的结果。这个免费使用的数据库目的就在于通过集合不同来源的蛋白质识别资料,让研究者们能方便地搜索已经公开发表的蛋白质数据。

4. 蛋白质功能域数据库

蛋白质功能域一般是指一条蛋白质序列中一段保守的区域,该区域能够独立行使功能、进化等。许多蛋白质序列包含若干结构功能域。在分子进化上,不同功能域可以作为一个单元被重组,产生新的蛋白质序列,行使不同的功能,因此,一个功能域可能在许多不同蛋白质序列中存在。功能域长度不一,例如可以从 25 到 500 氨基酸不等。目前蛋白质功能域数据库国际上主要包括 PROSITE、Pfam、ProDom、PRINTS、SMART 等,它们均属于 InterPro 功能

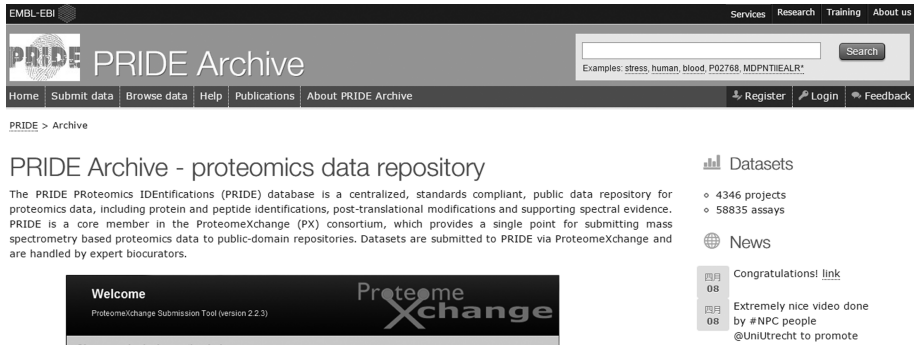

图 1-2.13　蛋白质组数据库 PRIDE 主页界面

域联盟。另外还有 BLOCKS、CDD 等(有关内容详见第 1-4 章多序列联配有关介绍)。

PROSITE(http://prosite.expasy.org/)数据库收集了生物学有显著意义的蛋白质位点和序列模式,并能根据这些位点和模式快速可靠地鉴别一个未知功能的蛋白质序列应该属于哪一个蛋白质家族。有的情况下,某个蛋白质与已知功能蛋白质的整体序列相似性很低,但由于功能的需要保留了与功能密切相关的序列模式,这样就可能通过 PROSITE 的搜索找到隐含的功能基序(motif,以正则表达式 pattern 方式储存),因此是序列分析的有效工具。PROSITE 中涉及的序列模式包括酶的催化位点、配体结合位点、与金属离子结合的残基、二硫键的半胱氨酸、与小分子或其它蛋白质结合的区域等;除了序列模式之外,PROSITE 还包括由多序列比对构建的概型(profile),能更敏感地发现序列与概型的相似性。PROSITE 的主页上提供各种相关检索服务。

Pfam(http://pfam.xfam.org/)是一个蛋白质家族数据库,其中包括蛋白质家族的注释以及通过隐马尔可夫模型产生的多序列联配结果。直至 2015 年 12 月 22 日,Pfam 已经发布了版本 29.0,其中包含了 16 295 条目(entry),序列、结构或者 HMM 模型相似相关的条目汇聚为更高等级的 559 个族(clan)。Pfam 蛋白家族又被分为质量高低的两类:Pfam-A 和 Pfam-B。Pfam-A 是高质量的人工注释的蛋白质家族,其中条目来自 Pfamseq(Pfam 序列数据库),这个数据库基于最新发布的 UniProtKB。Pfam-B 是未经注释的,从最新发布的 ADDA 中非冗余聚类中自动生成的低质量蛋白质家族。ADDA(automatic domain decomposition algorithm)是一个用于对所有蛋白质家族进行结构域分解和聚类的自动算法,专门用于建立 Pfam-B,虽然 Pfam-B 的质量不高,但是在功能保守性区域且在 Pfam-A 中找不到结果的时候就可以发挥作用。

5. 蛋白分子互作数据库

BioGRID(http://thebiogrid.org/)是一个包含了蛋白之间互作、遗传互作、化学物质互作以及翻译后修饰的专业生物数据库。2016 年发布的版本 3.4.135 共收录 56 300 篇已经发表文献中主要模式生物共 1 060 041 个蛋白以及遗传互作,27 501 个化学物质关联,38 559 个翻译后修饰。

DIP(Database of Interacting Proteins)(http://dip.doe-mbi.ucla.edu/dip/Main.cgi)收录蛋白之间的相互作用。目前共收录来自 7 937 篇论文的 803 个物种的 28 384 个蛋白以及 80 715 种相互作用。

IntAct molecular interaction database(http://www.ebi.ac.uk/intact/)是 EBI 数据库分子互

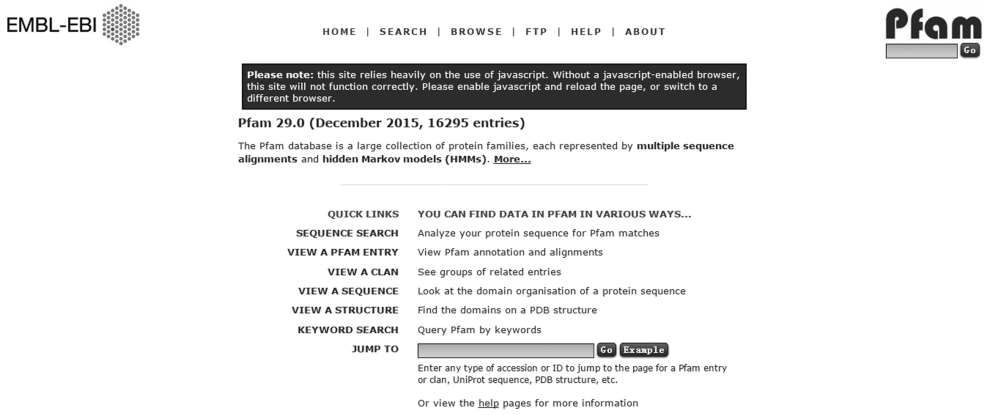

图 1-2.14 功能域数据库 Pfam 主页界面

作的一个分数据库,其中包括蛋白互作、蛋白小分子互作、蛋白核酸互作。

STRING(http://string-db.org/)数据库收纳了已知蛋白之间相互作用并能够预测蛋白互作。目前共收录了 2 031 个物种的 9 643 763 种蛋白互作。

第四节 代谢途径等专业数据库

一、代谢途径数据库

生物体内基因经由转录并翻译成蛋白质后,参与的各种复杂的生化反应,使物质 A 到物质 X 的酶反应常规程序(A→B→C→……X),称为 A 至 X 的代谢途径(metabolic pathway)。代谢途径数据库中较为常用也是较为知名的数据库是 KEGG。一些常用的数据库网址见表 1-2.13。

表 1-2.13 部分代谢途径数据库网址

数据库(Database)	网址(Address)
KEGG	http://www.genome.jp/kegg/
IMP	http://imp.princeton.edu/
PlantCyc	http://www.plantcyc.org/
GO	http://geneontology.org/
HPD	http://discern.uits.iu.edu:8340/HPD/
NCBI BioSystems	http://www.ncbi.nlm.nih.gov/biosystems
MANET	http://www.manet.uiuc.edu/
MetaNetX	http://metanetx.org/mnxdoc/cite.html
MetaCyc Database	http://www.metacyc.org/
MapMan	http://mapman.gabipd.org/web/guest/mapmanweb

KEGG(Kyoto Encyclopedia of Genes and Genomes,http://www.kegg.jp/)由日本京都大学和东京大学联合开发的数据库,是现在常用的查询代谢途径的数据库,也可用来查询酶(或

编码酶的基因)、产物等,也可通过 BLAST 比对查询未知序列的代谢途径信息。KEGG 主要通过 Web 界面进行访问,也可通过本地运行的 perl 或 java 等程序进行访问。

MANET(Molecular Ancestry Network,http://www.manet.uiuc.edu/)是一个蛋白结构演化关系直接映射到生物分子网络上的数据库。MANET 数据库的主旨是以生物信息、进化以及数据统计的方式来研究调查代谢酶个体的祖先以及代谢的演化问题。MANET 数据库目前将 SCOP(Structural Classification of Protein)、KEGG 利用系统发生关系重建的方式在全局的角度来阐释蛋白折叠结构的演化问题。

MetaNetX(http://metanetx.org/)是一个能够在基因组水平对代谢网络以及生化通路进行收集分析操作的在线数据库。该数据库提供了直观可视化的在线生物信息工具,为通路的基础研究、基因组分析、系统生物的发展和教育提供可能。目前最新更新版本为版本 56,于 2016 年 3 月更新。

Mapman WebTools (http://mapman.gabipd.org/web/guest/mapmanweb) 为 Mapman 的在线使用数据库。包括了大麦拟南芥在内的三个物种的表达数据集。Mapman 是一个用户为主导的将大量代谢组表达数据通路以图像形式表现的软件,目前已经更新至版本 2013.06。Maman WebTools 仅能提供包括了大麦拟南芥在内的三个物种的表达数据集做为测试参考。

代谢途径数据库仅是通路(pathway)数据库中的一员(表 1-2.14)。通路数据库总汇网站 Pathguide(http://www.pathguide.org/)对通路数据库进行了详细的总结,所有通路相关数据库都能在 Pathguide 上找到,包括一些历史上已经不可用的网站。根据 Pathguide2013 年 8 月发布的版本,目前共有 547 个生物通路相关和分子间相互作用相关资源。各类通路数据库在 Pathguide 均有链接和详细介绍,这里就不做详细介绍。

表 1-2.14　通路数据库总汇(基于 Pathguide 数据库)

数据库分类(Category)	数据库数量
蛋白间相互作用	257
代谢途径或通路	133
信号通路	95
通路图	81
转录因子/基因调控网络	80
蛋白成分间相互作用	61
遗传互作网络	28
Protein sequence focused	25
其他	17
无重复或冗余总数	547

二、代谢组学等数据库

1. 代谢组学数据库

代谢组学数据库是收录在代谢组学通路中的酶、化合物以及基因等成分的信息的数据

库。其中 MetaboLights(http://www.ebi.ac.uk/metabolights/)为 EMBL 下属的代谢组学数据库(图 1-2.15),主要内容包含代谢组学实验数据以及相关联的衍生信息。该数据库的信息物种交叉、技术交叉,覆盖了包括代谢组结构以及参考光谱、生物作用、位置、着重点、实验数据等一系列信息。模式生物人类,酵母,大肠杆菌有各自独立的代谢组学数据库(表 1-2.15)。

表 1-2.15 部分代谢组学数据库网址

数据库(Database)	网址(Address)
MetaboLights	http://www.ebi.ac.uk/metabolights/
HMDB	http://www.hmdb.ca/
YMDB	http://www.ymdb.ca/
ECMDB	http://ecmdb.ca/

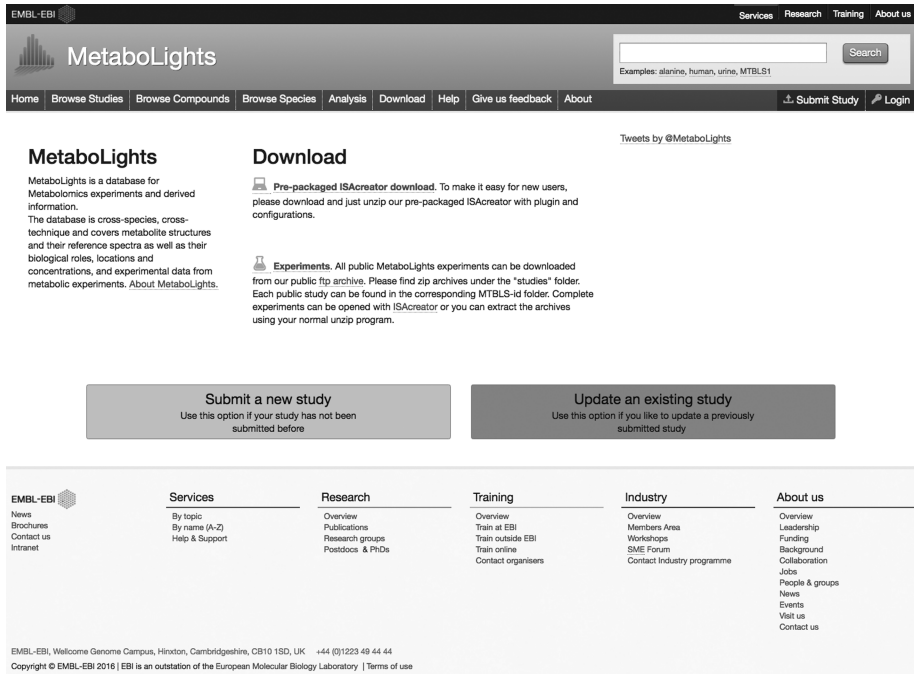

图 1-2.15 MetaboLights 主页界面

2. 表型数据库

PhenCode(http://phencode.bx.psu.edu/)是一个为了更好地理解人类表型和基因型之间关系的合作项目(数据库)。它将人类表型、各类人种特异性位点变异数据库(LSDBS)的临床数据和人类基因组数据、进化历史以及功能联系起来并共同展示在 UCSC 基因组浏览器上。

PhenomicDB(http://www.phenomicdb.de/)是一个多种类的表型基因型数据库,其中包括了人类、老鼠、果蝇、秀丽线虫以及其他一些模式生物。其中的基因目录(来自 NCBI gene)以及直系同源基因目录(来自 HomoloGene)允许一个给定基因同时在多个物种进行分析鉴定。PhenomicDB 数据主要来源于初级公共免费数据库。

PHI-base(pathogen-host interaction database,http://www.phi-base.org/),即病原体寄主互作数据库,也是病原体寄主表型数据库。该数据库主要是将寄主身上由微生物引起的病原体基因信息和表型信息进行相关联。该数据库的信息来源于公开发表的文献。

习　题

1. 什么是一级数据库和二级数据库,它们有什么异同?
2. 简述 Fasta 和 Fastq 格式,并比较它们的异同。
3. 如何向 NCBI 递交序列? 列举三种方法。如果序列文件数据很大或序列条数很多应该如何解决?
4. 如何下载水稻基因组的特定区段序列或注释?
5. 如何对 PDB 数据库记录进行三维结构的可视化?
6. 如何确定未知基因/序列属于哪个 KEGG 代谢途径?