第二章 分子数据库

生物信息学涉及的数据库可大致分为二种：初级数据库和二级数据库。初级数据库贮存原始的生物数据，如 DNA 序列，由晶体衍射(Crystallography)获得的蛋白质结构等。二级数据是在初级数据库的基础上经加工和增加相关信息，使它们更便于特定专业人员的使用，如真核生物启动子序列库 EPD 和蛋白质一般结构或功能模体(motif)数据库 PROSITE。

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和有应用价值，这一点值得注意。在基因组规模上的测序过程便产生了注释问题。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细胞类型，而其它方面却知之甚少。如果你在确定一段未知蛋白质序列的功能，发现一个与之匹配的序列，但该序列却没有任何有关功能的信息时，你的研究工作便很难为继了。

不同的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，但这必定会影响序列的注释；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，新的发现不断被补充进去，所以，本书中用到的一些注释信息可能很快便被更新了。在所有的生物信息数据库中总会有一小部分的记录(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。

第一节 初级数据库

一．DNA 数据库

DNA 序列构成了初级数据库的主体部分。目前国际上有 3 个主要的 DNA 序列公共数据库(表 2.1)：欧洲分子生物学实验室(European Molecular Biology Laboratory, EMBL)(位于英国剑桥)，GenBank[美国国家生物技术信息中心(National Center for Biotechnology Information, NCBI)，该中心隶属于美国家医学图书馆，位于美国国家卫生研究院(NIH)内]和日本 DNA 数据库(DNA Databank of Japan,DDBJ)。这 3 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 DNA 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(EMBL 负责欧洲，GenBank 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，3 个数据库共同享有并向世界开放，故这 3 个数据库又被称为公共序列数据库 (Public Sequence Database)。所以从理论上说，这 3 个数据库所拥有的 DNA 序列数据是完全相同的。你可以从中选择一个你喜欢的数据库；但是如果你的研究需要实时(24 小时以内)的，则要注意这些数据库间的记录是会有差异的。

表 2.1 三个主要 DNA 序列数据库网址

数据库 (Database)	网址 (Address)
EMBL	www.ebi.ac.uk/ebi_docs/embl_db/ebi/topembl.html
GenBank	www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html
DDBJ	www.ddbj.nig.ac.jp

DNA 序列数据库的增长是飞速的。EMBL 核酸数据近 20 年的增长情况(表 2.2)充分说明了这一点。从历史看，每 22 个月，数据库的数据规模将翻一翻，而且随着表达序列标签(EST)数据的迅猛增长，这一速率已快速递增。EMBL 数据翻一翻的周期最近已缩短到 9 个月左右。数据库的膨胀对于数据库的搜索非常有好处。也许你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，当进行生物信息学分析时，分析结果中务必要注明你当时所使用序列数据库的数据状况。2004 年 12 月 EMBL (Release81)的 DNA 碱基对数已接近 800 亿，序列数超过 4 亿条。为了有效地管理如此庞大的数据，数据库数据根据物种(species)分为几类，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 2.3)，例如 EMBL 数据库最近的分类为人、真菌等等，同时每一类的数据文件往往又分成一定的亚类，例如 EST 类数据文件 (表 2.4)。这些分类有助于我们便捷地进入数据库的相关部分。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，这主要是数据库规模快速扩大的需要。如新加入的高通量测序数据(HTG)等。EMBL 和 GenBank 等数据库的使用手册均可在相应的网址上找到，这些手册提供了详尽的数据库组成、分类等细节，不妨到那些网站(见表 2.1)上看看。

表 2.2 EMBL 数据库 DNA 序列数据库增长情况

数据库报告 (Release)	释放日期 (Month)	记录数 (Entries)	核苷酸数 (Nucleotides)
Release 1	1982 年 6 月	568	585433
Release 7	1985 年 12 月	5789	5622638
Release 25	1990 年 11 月	41580	52900354
Release 29	1991 年 12 月	57655	75400487
Release 33	1992 年 12 月	89100	111413979
Release 37	1993 年 12 月	146576	158171400
Release 41	1994 年 12 月	230950	226259607
Release 45	1995 年 12 月	622566	427620278
Release 49	1996 年 12 月	1047263	696183789
Release 53	1997 年 12 月	1917868	1281391651
Release 57	1998 年 12 月	3046471	2164718256
Release 61	1999 年 12 月	5303436	4508169737
Release 65	2000 年 12 月	9549328	10710321435
Release 69	2001 年 12 月	14366182	15383451165
Release 73	2002 年 12 月	20857746	27903283528
Release 77	2003 年 12 月	30351263	36042464651
Release 81	2004 年 12 月	46105397	79271300840

表 2.3 EMBL 数据库 2004 年 12 月数据状况 (Release81)

类别 (Division)		代码 (Code)	记录数 (Entry)	核苷酸数 (Nucleotide)
表达序列标签	ESTs	EST	24481418	12837493911
真菌	Fungi	FUN	110405	221397562
基因组检测序列	Genome Survey Sequences	GSS	10726912	6608825736
高通量基因组	High Throughput Genome	HTG	68564	11613533555
人	Human	HUM	292205	4126190851
无脊椎动物	Invertebrates	INY	175545	677544114
其它哺乳动物	Other Mammals	MAM	70355	341455910
细胞器	Organelles	ORG	314215	270405172
专利	Patents		2276431	1332968224
噬菌体	Bacteriophage	PHG	2625	12989224
植物	Plants	PLN	287510	1084488061
原核生物	Prokaryotes	PRO	282227	993811176
啮齿类动物	Rodents	ROD	31538	110601526
序列标签位点	STSs	STS	380660	168545968
合成	Synthetic	SYN	14240	22721647
未分类	Unclassified	UNC	2869	2823924
病毒	Viruses	VRL	262346	241496438
其它脊椎动物	Other Vertebrates	VRT	113601	879447919
总和	Total		39893666	41546740918

表 2.4 EMBL 数据库 EST 类数据文件分类情况(Release 81)

亚类数据文件名 (Subdivision)		说明 (Comments)
est_fun.dat	est_fun05.dat	真菌 EST
est_hum.dat	est_hum57.dat	人 EST
est_inv.dat	est_inv31.dat	无脊椎动物 EST
est_mam.dat	est_mam11.dat	哺乳动物 EST
est_pln.dat	est_pln55.dat	植物 EST
est_pro.dat	est_pro01.dat	原核生物 EST
est_rod.dat	est_rod07.dat	啮齿类动物 EST
Est_vrt.dat	est_vrt27.dat	脊椎动物 EST

二．基因组数据库

第二个主要的初级数据源来自各种基因组计划。一些基因组计划已经完成，如真核生物酵母(*Saccharomyces cerevisiae)*，原肠生物(*Methanococcus janeschii*) 和 3 个原核生物流感嗜血杆菌(*Haemophilus influenzae*)、(*Mycoplasma genitaliam*)和大肠杆菌(*Escherichia coli*)等。这些计划的大部分信息在 EMBL 中均可找到。很多基因组计划正在进行中，表 2.5 列出了一些基因组计划的网址。

表 2.5 部分生物基因组计划网址

生物种类	Organism	网址(Address)
曲霉菌	Aspergillus	http://www.ncbi.nlm.nih.gov/genome/guide/aspergillus
蜜蜂	Bee	http://www.ncbi.nlm.nih.gov/genome/guide/bee
猫	Cat	http://www.ncbi.nlm.nih.gov/genome/guide/cat
青蛙	Frog	http://www.ncbi.nlm.nih.gov/genome/guide/frog
老鼠	Mouse	http://www.ncbi.nlm.nih.gov/genome/guide/mouse
小鼠	Rat	http://www.ncbi.nlm.nih.gov/genome/guide/rat/index.html
狗	Dog	http://www.ncbi.nlm.nih.gov/genome/guide/dog
牛	Cow	http://www.ncbi.nlm.nih.gov/genome/guide/cow
猪	Pig	http://www.ncbi.nlm.nih.gov/genome/guide/pig
羊	Sheep	http://www.ncbi.nlm.nih.gov/genome/guide/sheep
鸡	Chicken	http://www.ncbi.nlm.nih.gov/genome/guide/chicken
斑马鱼	Zebra fish	http://www.ncbi.nlm.nih.gov/genome/guide/zebrafish/index.html
海胆	Sea urchin	http://www.ncbi.nlm.nih.gov/genome/guide/sea_urchin
线虫	Caenorhabditis elegans	http://www.ncbi.nlm.nih.gov/genome/guide/nematode
	Dictyostelium discoideum	http://www.ncbi.nlm.nih.gov/genome/guide/dicty
果蝇	Drosophila	http://www.ncbi.nlm.nih.gov/genome/guide/fly
蚊子	Mosquito	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?chr=agambiae.inf
黑猩猩	Chimp	http://www.ncbi.nlm.nih.gov/genome/guide/chimp
人	Human	http://www.ncbi.nlm.nih.gov/genome/guide/human
拟南芥	Arabidopsis	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=3702
棉花	Cotton	http://algodon.tamu.edu
玉米	Maize	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=4577
水稻	Rice	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=4530
小麦	Wheat	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=4565
大麦	Barley	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=4513
大豆	Soybean	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=3847
西红柿	Tomato	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=4081
高粱	Sorghum	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=4557

生物基因组差异极大，这种差异不仅表现在基因组大小(表 2.6)而且在于单链或双链 DNA、RNA 的遗传信息存储特性。另外，一些基因组是线性的(如哺乳动物)，而另一些则上封闭的环状(如绝大多数细菌)。

人类最早(1977)获得的生物基因组全序列是噬菌体（53kb)，1987 年自动测序仪问世，随后第一个病毒基因组序列（1990）在自动测序仪上完成；后来是第一个细菌基因组(1995)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，最后是人类自身(2000)的遗传密码被解开。最早完成的噬菌体、病毒和细胞器的基因组数据在 80 年代早期就存入了 EMBL 数据库。从那以后，随着测序技术的革命性改进，大量的基因组数据被存入该数据库，涉及的物种种类不断增多，如最近又增加了（Chimp）和（Fruit Fly)。表 2.6 列举了一些生物基因组测序的进展状况。欧洲生物信息学研究所(European Bioinformatics Institute，EBI)的基因组网站提供了已完成的基因组序列数据，可以自由访问(www.ebi.ac.uk/genomes/)（图 2.1)。一些网站提供世界范围内基因组测序进展的最新情况，如 Genome MOT（图 2.2)，通过它们可以了解基因组测序的发展动态。

2000 年 6 月 26 日，人类基因组草图被宣告完成。原始的序列数据可在 EMBL 等数据库的 HTG 和 HUM 部分找到。人类基因组的大小为 3.2 兆亿碱基对(Gigabases)，而由于冗余原因，实际获得的碱基数超过了这一数字。总的估计，冗余的碱基比例为 30-40%。

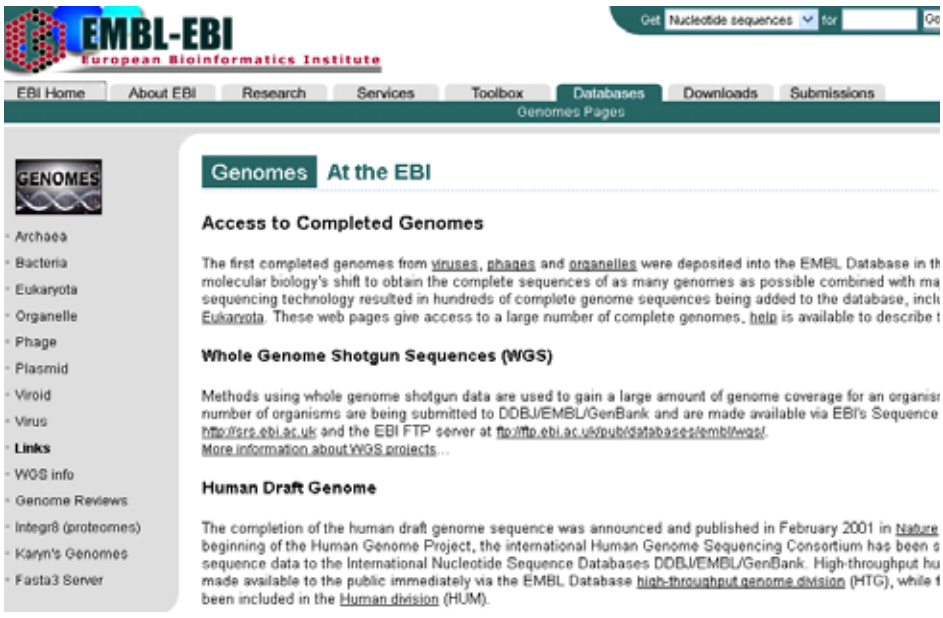

图 2.1 欧洲生物信息学研究所(EBI)的基因组网站主页。该网站提供了已完成的各类生物（真核生物、细菌、病毒等，见图中左列）基因组情况。

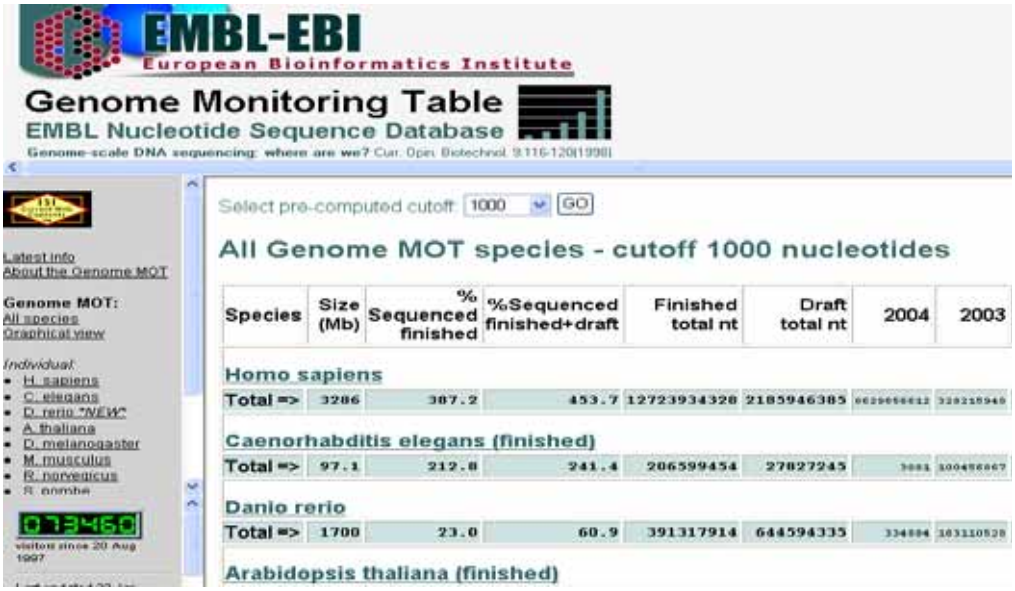

图 2.2 基因组测序进展状况服务器 Genome MOT 主页。该网站每日报告基因组测序进展情况，图中为 2004 年 2 月 29 日的进展报告，具体数据可参见表 2.6。

表 2.6 部分基因组测序情况。根据 Genome MOT，截止 2004/2/29。

物种 Species	基因组大小 Size (Mb)	完成比率 %Sequenced finished	完成比率（包括草图）% Sequenced finished+draft	完成核苷酸总数 Finished total nt	完成(草图)核苷酸总数 Draft total nt
人类 *Homo sapiens*	3286	387.2	453.7	12723934328	2185946385
线虫 *Caenorhabditis elegans* (finished)	97.1	212.8	241.4	206599454	27827245
拟南芥 *Arabidopsis thaliana* (finished)	118	205.0	207.6	241889120	3060859
果蝇 *Drosophila melanogaster* (finished)	135.6	324.0	432.3	439303114	146899025
Danio rerio	1700	23.0	60.9	391317914	644594335
老鼠 *Mus musculus*	3059	134.0	197.4	4099660644	1938348462
小鼠 *Rattus norvegicus*	3000	1.8	170.6	52581865	5066240322
Schizosaccharomyces pombe (finished)	13.8	205.7	205.7	28386200	0
酵母 *Saccharomyces cerevisiae* (finished)	12.1	274.9	274.9.7	33258957	0

三．蛋白质序列数据库

SWISS-PROT 和 PIR 是国际上二个主要的蛋白质序列数据库，目前这二个数据库在 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。PIR 数据库的数据由美国家生物技术信息中心(NCBI)翻译自 GenBank 的 DNA 序列。PIR 根据注释程度(质量)分为 4 个等级(表 2.7)。

表 2.7 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)	20685
PIR2	注释 (Annotated)	262300
PIR3	未核实 (Unverified)	24
PIR4	未翻译 (Unencoded or untranslated)	407

表 2.8 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 2.8 主要蛋白质序列数据库网址

数据库 (Database)	网址 (Address)
SWISS-PROT	http://www.ebi.ac.uk/swissprot/
TREMBL	http://www.ebi.ac.uk/trembl/
PIR	http://pir.georgetown.edu/

4．蛋白质结构数据库

实验获得的三维蛋白质结构均贮存在蛋白质数据库 PDB 中。PDB 是国际上主要的蛋白质结构数据库，虽然它没有蛋白质序列数据库那么庞大，但其增长速度很快。PDB 贮存有由 X 射线和核磁共振(NMR)确定的结构数据。NRL-3D 数据库提供了贮存在 PDB 库中蛋白质的序列，它可以进行与已知结构的蛋白质序列的比较。对来自 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，被贮存在 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。

表 2.9 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

第二节 初级序列数据的注释

到目前为此，尚没有一个统一的序列注释格式，各数据库间均存在差异。但总的来说，各数据库所提供的注释内容还是相同的。现在比较使人不放心的是针对一个相同基因的 DNA 和蛋白质序列注释之间的差异。以下给出了一个 EMBL 数据库记录的注释例子(图 2.2)。表 2.10 对注释中的代码及内容进行了说明。

```
ID   LISOD      standard; DNA; PRO; 756 BP.
XX
AC   X64011; S78972;
XX
SV   X64011.1
XX
DT   28-APR-1992 (Rel. 31, Created)
DT   30-JUN-1993 (Rel. 36, Last updated, Version 6)
XX
DE   L.ivanovii sod gene for superoxide dismutase
XX
KW   sod gene; superoxide dismutase.
XX
OS   Listeria ivanovii
OC   Bacteria; Firmicutes; Bacillus/Clostridium group;
OC   Bacillus/Staphylococcus group; Listeria.
XX
RN   [1]
RX   MEDLINE; 92140371.
RA   Haas A., Goebel W.;
RT   "Cloning of a superoxide dismutase gene from Listeria ivanovii by
RT   functional complementation in Escherichia coli and characterization of the
RT   gene product.";
RL   Mol. Gen. Genet. 231:313-322(1992).
XX
RN   [2]
RP   1-756
RA   Kreft J.;
RT   ;
RL   Submitted (21-APR-1992) to the EMBL/GenBank/DDBJ databases.
RL   J. Kreft, Institut f. Mikrobiologie, Universitaet Wuerzburg, Biozentrum Am
RL   Hubland, 8700 Wuerzburg, FRG
XX
DR   SWISS-PROT; P28763; SODM_LISIV.
XX
FH   Key             Location/Qualifiers
FH

FT   source          1..756
FT                   /db_xref="taxon:1638"
FT                   /organism="Listeria ivanovii"
FT                   /strain="ATCC 19119"
FT   RBS             95..100
FT                   /gene="sod"
FT   terminator      723..746
FT                   /gene="sod"
FT   CDS             109..717
FT                   /db_xref="SWISS-PROT:P28763"
FT                   /transl_table=11
FT                   /gene="sod"
FT                   /EC_number="1.15.1.1"
FT                   /product="superoxide dismutase"
FT                   /protein_id="CAA45406.1"
FT                   /translation="MTYELPKLPYTYDALEPNFDKETMEIHYTKHHNIYVTKLNEAVSG
FT                   HAELASKPGEELVANLDSVPEEIRGAVRNHGGGHANHTLFWSSLSPNGGGAPTGNLKAA
FT                   IESEFGTFDEFKEKFNAAAAAARFGSGWAWLVVNNGKLEIVSTANQDSPLSEGKTPVLGL
FT                   DVWEHAYYLKFQNRRPEYIDTFWNVINWDERNKRFDAAK"
XX
SQ   Sequence 756 BP; 247 A; 136 C; 151 G; 222 T; 0 other;
     cgttatttaa ggtgttacat agttctatgg aaatagggtc tatacctttc gccttacaat        60
     gtaatttctt ttcacataaa taataaacaa tccgaggagg aatttttaat gacttacgaa       120
     ttaccaaaat taccttatac ttatgatgct ttggagccga attttgataa agaaacaatg       180
     gaaattcact atacaaagca ccacaatatt tatgtaacaa aactaaatga agcagtctca       240
     ggacacgcag aacttgcaag taaacctggg gaagaattag ttgctaatct agatagcgtt       300
     cctgaagaaa ttcgtggcgc agtacgtaac cacggtggtg gacatgctaa ccatacttta       360
     ttctggtcta gtcttagccc aaatggtggt ggtgctccaa ctggtaactt aaaagcagca       420
     atcgaaagcg aattcggcac atttgatgaa ttcaaagaaa aattcaatgc ggcagctgcg       480
     gctcgttttg gttcaggatg ggcatggcta gtagtgaaca atggtaaact agaaattgtt       540
     tccactgcta accaagattc tccacttagc gaaggtaaaa ctccagttct tggcttagat       600
     gtttgggaac atgcttatta tcttaaattc caaaaccgtc gtcctgaata cattgacaca       660
     ttttggaatg taattaactg ggatgaacga aataaacgct ttgacgcagc aaaataatta       720
     tcgaaaggct cacttaggtg ggtctttta tttcta                                  756
```

图 2.3 EMBL 数据库记录(记录是 X64011)注释例举。有关说明见表 2.10

表 2.10 EMBL 数据库记录注释代码和内容说明

代码 (Code)	全 称 (Full meaning)		说 明 (Comments)
ID	identifier	(身份号)	该行的第一项内容是该数据库记录的名称，该名称是唯一的，是由 EMBL 数据库给定的。其它内容注明了该记录的一些状况(如是否已经被核实－本例中为已核实，即 standard；记录的碱基数等)
AC	accession number	(记录号)	每个记录号均是唯一的，并从不更改，是由 GenBank 给定的。如果两个记录被合并成一个记录，原始上着 2 个记录号均会被注明
DT	data	(日期)	2 个日期被注出，一个是该数据第一次被记录时间，另一个是最后一次的时间。
DE	description	(描述)	对该基因的文字描述
KW	keywords	(关键词)	描述该基因的关键词
OS	organism(species)	(物种)	物种名称
OC	organism(classification)	(分类)	物种的一个简单分类，该分类并不一定准确，应谨慎从事
OG	Organelle	(细胞器)	该基因是否在某一个特殊的细胞器中
RN	reference number	(文献编号)	
RC	reference comment	(文献说明)	
RP	reference positions	(文献大小)	与该记录研究相关的文献信息
RX	cross-reference	(相关文献)	
RA	reference authors	(文献作者)	
RT	reference title	(文献题目)	
RL	reference location	(文献出处)	
DR	database cross-reference	(相关文献数据库)	见文中说明
FH	feature header	(主表头)	该记录主要内容列表表头
FT	feature table data	(主表数据)	见文中说明
CC	comments	(说明)	对记录的文字说明
XX	spacer line	(空白行)	
SQ	sequence header	(序列头)	有关该序列大小和组成的信息
blank	sequence data	(空白)	
//	termination line	(终止行)	一个记录的终止符号

相关文献数据库 (database cross-reference，DR)需要做进一步的说明。许多二级数据库内容来自初始数据库，例如 OMIM(Online Mendelian Inheritance in Man)数据库是有关人类遗传疾病的数据，如果 OMIM 中的一个记

录与 EMBL 中一个已知序列的基因有关，则该基因将与该记录建立联系，则 EMBL 库中该序列的 DR 栏中将包括 OMIM 和 OMIM 中相关记录的名称。上述例子(图 2.3)的 DR 栏中有该 DNA 序列翻译成蛋白质序列的 SWISS-PROT 记录号等。从此可见，DR 栏内容非常重要，它有助于了解与该原始 DNA 序列相关信息的状况和存贮站点。与 DR 栏可能有关的一些数据库包括 SWISS-PROT、 EMBL、OMIM、PROSITE(保守蛋白质模序数据库，见下文)、HSSP、PDB、MEDLNE(与 RL 栏相关的文献摘要数据库)、PIR 等。注释中另一个需要说明的重要内容是主表数据(feature table data，FT)栏。主表试图将尽可能多的序列信息囊括其中，并以计算机可以阅读的格式编排。3 个主要 DNA 数据库(EMBL、GenBank 和 DDBJ)已经对该表的表述格式达成了一致。具体表述格式内容说明可在 www.ebi.ac.uk/ebi_docs/embl_db/ft/feature_table.html 找到。

大量的 DNA 序列记录包含有一个以上的开放读框(ORF)。主表中的 PID 编号被用于唯一地指定每一个 ORF。这一编号是一个非常重要的注释信息，因为它可以使许多不同的 SWISS-PROT 记录与一个相同的 EMBL 序列相链接，可以精确地知道 EMBL 序列中的 ORF 所对应的 SWISS-PROT 蛋白质记录。

第三节 数据库信息检索系统

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(由美国建立)和 SRS(Sequence retrieval System)(由 EMBL Theore Etzold 建立)。

SRS 检索系统在欧洲的许多网站被广泛使用。SRS 是一个具有弹性的系统，可应用于大量不同的数据库。这意味着使用 SRS 的数据库在各个站点可能略有差异，而这种差异是由数据库管理者所决定的。例如，OWL 数据库是一个非冗余蛋白质序列库，它的数据来源主要是从其它主要蛋白质数据库中收集而来的，在 SEQNET 服务器(www.seqnet.dl.ac.uk/srs/srsc)可通过 SRS 搜索而进入 OWL，但在 EBI 网站通过 SRS 则不能进入 OWL。

序列一般可通过记录号(如来自 1 篇发表的论文)或是该序列注释中的一些信息进行检索。SRS 的优势是可以使你通过普通的终端去检索大范围的数据库，并通过 DR 栏链接到在其它数据库。

SRS 的使用非常直观。图 2.4 所示是 EBI 网站 SRS 的主网。如果想检索序列数据可选择第一按钮“Search sequence libraries”。这时出现一个新图面(图 2.5)，然后选定你想搜索的数据库并在文字框(text)中键入正确的检索词。检索可建立逻辑关系(and,or,not)进行。按下“DO-QVERY”按钮便开始检索。检索结果的输出格式等也可设定，选定的记录内容可通过网络浏览器上的保存功能存入你的计算机中。

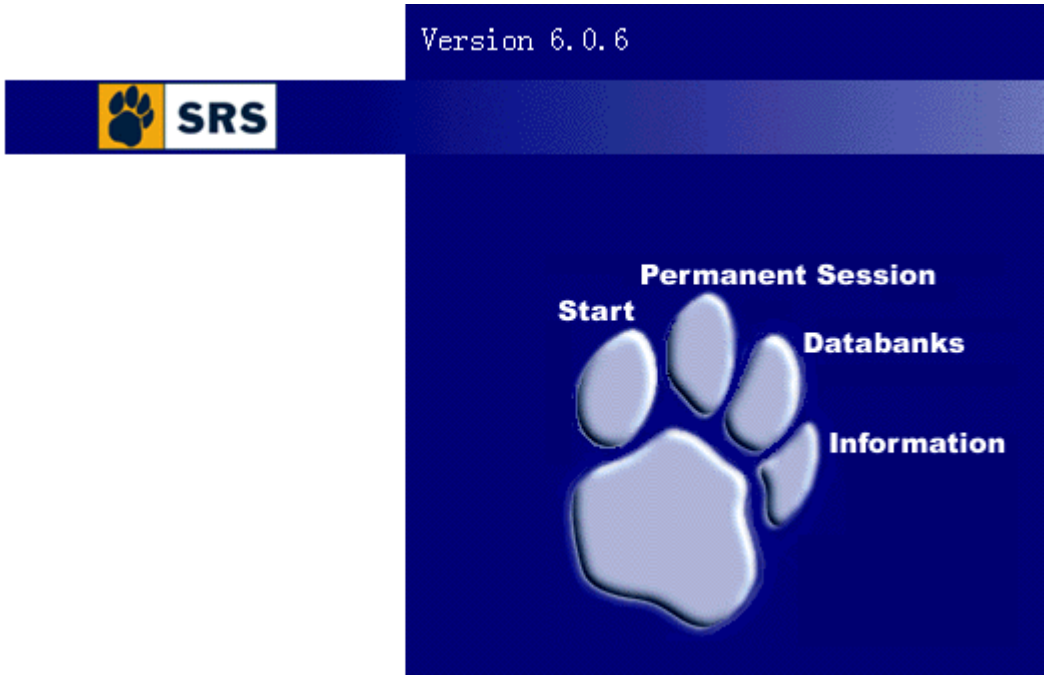

图 2.4 数据库搜索系统 SRS 主页

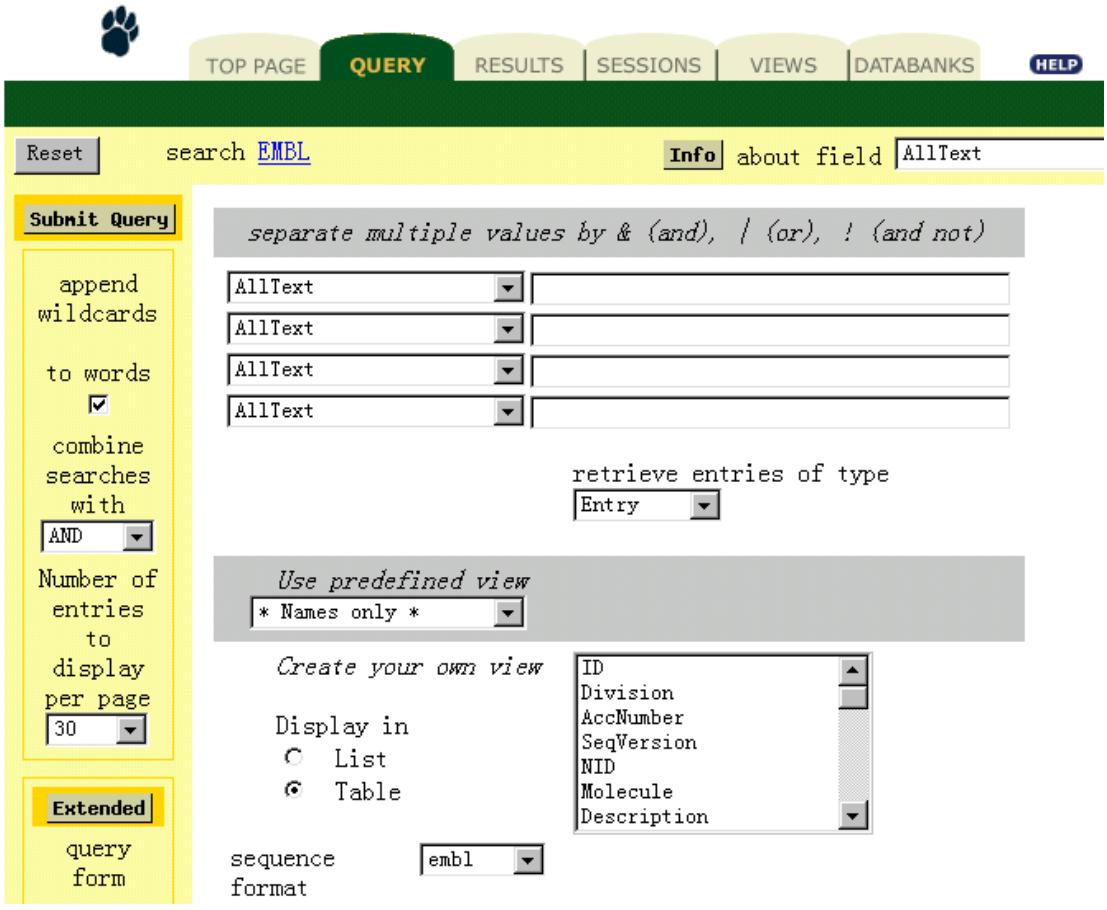

图 2.5 数据库搜索系统 SRS 进行序列搜索设置的页面

第四节 数据库的冗余与偏差[1]

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发送了相同的序列数据，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，可以被认定为它们是相同序列，但一些显著的差异可能是由于基因组多样性的结果。

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—如人工智能方法－发生偏离。

基于以上原因，有必须避免在数据库中存在太过于相似的序列，很多数据库也是这样做了，努力使他们的数据库为非冗余(non-redundant,nr)。但是，生物数据非常复杂，它远非“冗余”二字可以准备描述，例如，同一位点上的 2 个等位基因是不是冗余的？同一生物体内的 2 个同功酶是否冗余？因此，过于苛刻地去除“太过于相似的序列”可能导致一些有价值的信息被删除，应在数据规模和非冗余之间找到一个合理的平衡点。“太过于相似”的准确界定应主要依据所要研究的问题。实际研究中，试验数据往往“随机”地从数据库中抽取而不考虑减少冗余问题；即使考虑到冗余问题，也或多或少地存在随意性，即随意地进行一些同源性分析，确定一些蛋白质或 DNA 簇，然后从各簇中选取一个数据样本来组合所谓的“代表性”数据样本。

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 cDNA 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。

对付以上这些偏差，一个聪明的策略是用可能污染数据记录的序列(如载体)去估计误差程度。同时，一些去除污染的专门软件系统已被研制出来，如 EBI 网站便提供了去除载体污染的在线服务，网址为 http://www.ebi.ac.uk/blastall/vectors.html。EMBL 研制了基于 BLAST 的载体扫描服务和一个特殊的序列数据库 EMVEC。EMVEC 的序列来自 EMBL 的 SYN(synthetic division)类 2000 余条一般用于克隆和测序实验的序列，该库随着 EMBL 的扩充而实时更新。

[1]本部分内容主要取自F．奥斯伯，R．E．金斯顿等．精编分子生物学实验指南，北京：科学出版社，1998

第五节 向数据库发送序列数据及其它[2]

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、全面表述生物信息学研究的“材料与方法”和普通用户可利用的数据库服务内容。

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贮到某一数据库中。如果该序列是在欧洲完成的，则应储存到 EMBL，如来自美洲，则存到 GenBank，其它地区则应发送到日本的 DDBJ。这些数据库的主页上均有详细的发送说明。数据库往往特别要求发送者要注意去除载体污染，例如 EMBL 提供了 EBI 的相关服务(网址见上节)。序列的发送可以通过网上进行。EMBL 的发送系统为 WEBIN(http://www.ebi.ac.uk/embl/Submission/webin.html)，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)。GenBank 的发送系统 Sequin(http://www.ncbi.nlm.nih.gov/Sequin/index.html)是由 NCB1 开发的多平台(Mac/pc/unix)工具，适用于 EMBL、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)将产生并送给发送者，该记录号可用于论文发表。但发送的序列在公共数据库中出现可能会有一个滞后期，因为注释和核查将颇费一番周折。

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、EMBL、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)：数据库在快速变化，它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。

(3)所使用的计算机：这可能是不重要的一项说明，因为算法等不论在何种计算机上均应相同，但如果在使用异地(off-site)计算机系统(如 E-mail 和 Internet)服务，那么，科学的态度应是注明其服务器及其管理者。

如果进行序列的比较研究，还应包括以下内容（具体内容可参见第三章第 2 节）：

(4)替换矩阵(substitution matrix)：所有的现代搜索程序均使用替换矩阵，选用不同的矩阵会产生完全不同的结果，所以必须注明在搜索和列阵(aligning)中使用何种矩阵。

(5)空位罚值(gap penalty)：很多算法使用空位罚值(如 FASTA)。

一般用户可利用的分子数据库服务内容可分为几种：E-mail 服务、匿名 FTP 服务、www 服务和序列相似性搜索服务等。通过 E-mail 可向数据库发送相关要求来获取有关数据和服务。例如，可发一个服务指令到 EBI 的 mail to：netserv.ebi.ac.uk 地址，服务指令中应以一个指令开头，比如你想获得记录号为 X55652 的 DNA 序列，则应在指令栏中键入“GET NUC：X55652”，这样 EBI 服

[2]本部分内容主要取自F．奥斯伯，R．E．金斯顿等．精编分子生物学实验指南，北京：科学出版社，1998

务器便会将该序列的信息发到你的信箱中。匿名 FTP 服务是另外一种进入数据库获取信息的方法，研究者可利用本地的 FTP(file transfer protocol)程序连接到相应的数据库主机上，以“ anonymous ”(匿名)为用户名和自己的 E-mail 地址为口令进入。www 服务是通过网络直接进入相关数据库网址，进行数据检索、数据传送等。同时各数据库均提供序列相似性检索等序列分析的服务，如 FASTA、BLAST 和 BLITS 等服务(具体说明见第三章第 3 节)，分析结果通过 E-mail 发送返回或直接显示在浏览器上。