“实用生物信息技术”课程小组讨论总结报告

组：G4 次：2 组长：高嘉禾 执笔：赵雪倩

1.时间：2026 年 4 月 1 日 18：00

2.方式：线上（腾讯会议，微信）

3.主题：深入了解 UniProt 蛋白质数据库网站，完成课外练习等。

4.内容：

A. 熟悉 UniProt 蛋白质数据库网站;

B. 阅读《UniProt 蛋白质数据库简介》等参考文献;

C. 小组成员分别说明课外练习 2 中的相应部分;

D. 小组成员分别介绍与自己课程相关的课题内容;

E. 小组成员分别说明使用时的问题并探讨解决办法。

课外练习 2：

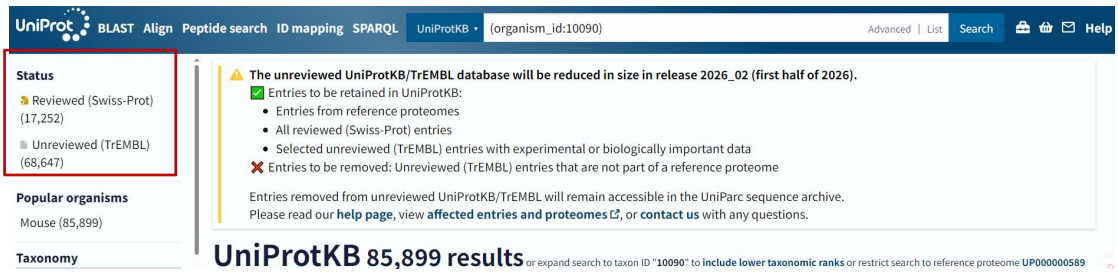
（一）UniProt 蛋白质数据库

1. UniProt 数据库概况

(1) Swiss-Prot：对序列条目进行人工审阅和注释，包括物种分类学来源、功能、定位、表达等，同时也包括与其它数据库的链接。

(2) TrEMBL：由核酸序列通过计算机程序翻译得到的蛋白质序列。[瑞士生物信息研究所（Swiss Institute of Bioinformatics, SIB）成立，主要负责管理、维护、发布和进一步开发 Swiss-Prot 数据库，而 EBI 主要负责管理、维护和发布 TrEMBL 数据库。]

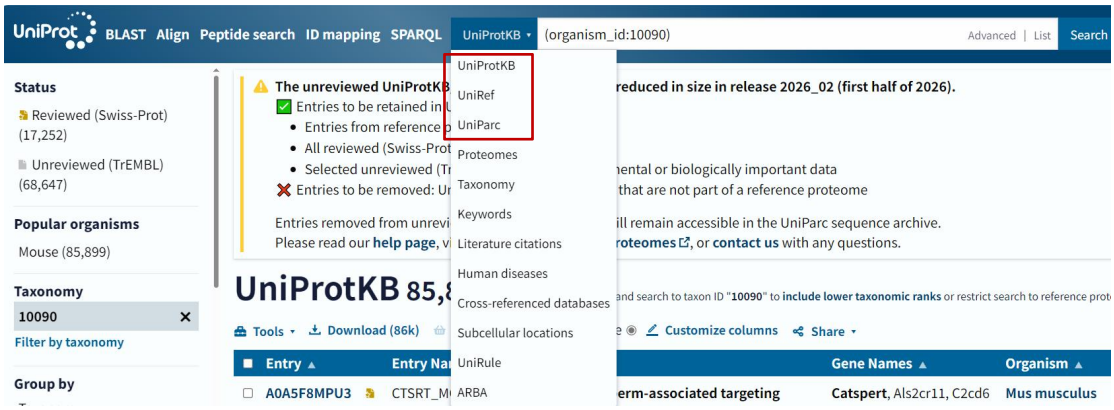
蛋白质数据库 UniProt，核心数据是蛋白质序列，因此也常被称为蛋白质序列数据库，或简称蛋白质数据库。UniProt 从创建至今,一直遵循人类基因组计划实施时国际科学界达成的共识，即基因组、蛋白组等生物信息数据资源应该为全人类共享，为世界各国公众提供无偿服务。



1)UniProt 数据库由哪三部分组成？

UniProt 包括三个主要部分，即蛋白质知识库(UniProt Knowledgebase， UniProtKB)、蛋白质序列归档库（UniProt Sequence Archive, UniPace）和蛋白质序列参考集(UniProt Reference Clusters,UniRef)。为适应蛋白组学研究的需要，UniProt 数据库还新增了蛋白组(Proteome)和参考蛋白组数据。

此外，UniProt 数据库还包括文献引用(Literature Citations)、物种分类学来源(Taxonomy)、亚细胞定位（Subcellular Locations)、数据库交叉链接(Crossreference Databases)、相关疾病(Diseases)和关键词(Keywords)等辅助数据。



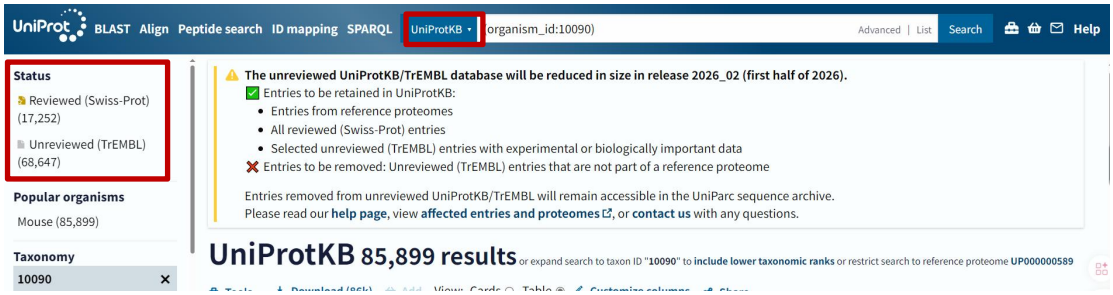
蛋白质知识库 UniProtKB 分为 Swiss-Prot 和 TrEMBL 两个子库。

蛋白质序列归档库 UniPace：目前数据最为齐全的非冗余蛋白质序列数据库。对相同序列归并到同一个记录中，并赋予特定标识符 UPI。不论这些序列条目源自何处，具有同一标识符的所有条目序列完全相同。若源数据库已经不复存在或源数据库中该序列条目已经不复存在，则标注为无效条目。

蛋白质序列参考集 UniRef：分为三个数据集（Sequence Cluster），分别为 UniRef100、UniRef90 和 UniRef50。第一步是把不同物种中长度不小于 11 个氨基酸的相同序列和序列片段合并在一起，得到 UniRef100 数据集。第二步是按相同位点所占序列全长比例 90%为阈值，将 UniRef100 数据集中高度相似序列合并在一起，产生 UniRef90 数据集。第三步则是按相同位点所占序列全长比例 50%为阈值，将 UniRef90 数据集中具有一定相似性的序列合并在一起，所得数据集即 UniRef50。

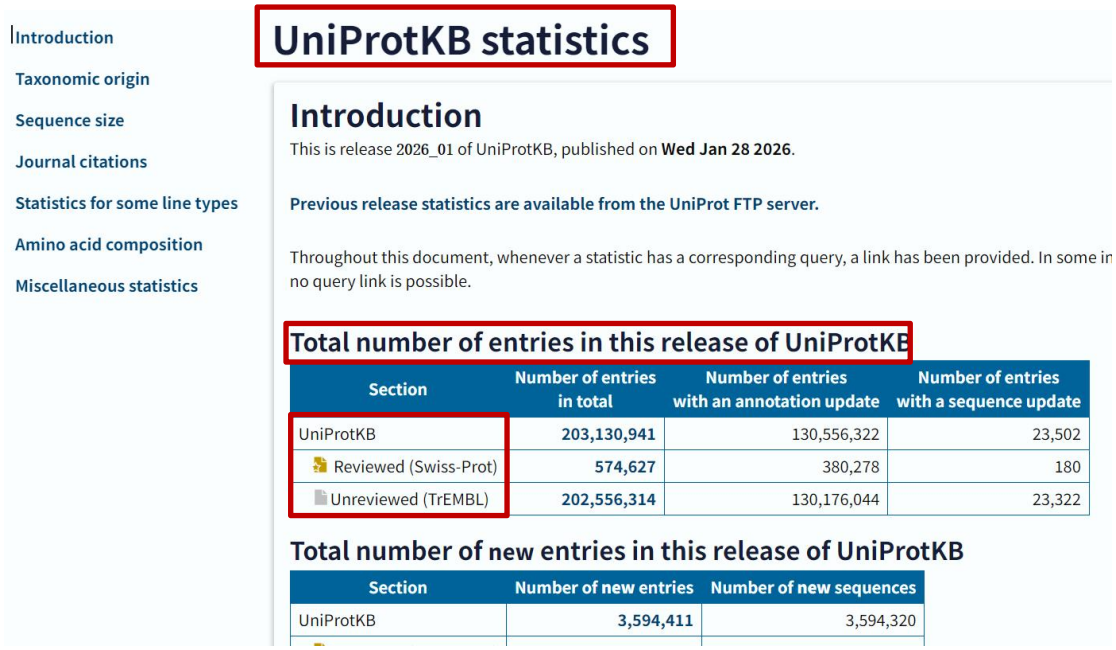
2)UniProtKB 知识库由哪两部分组成？

UniProtKB 分为 Swiss-Prot 和 TrEMBL 两个子库。内容包括蛋白质功能基因本体（Geneontology，GO）注释、物种名及分类、亚细胞定位、蛋白质加工修饰、表达等信息。此外，UniProtKB 还提供与基因组、核酸序列、蛋白质结构、蛋白质家族、蛋白质功能位点、蛋白质相互作用等其它数据库的交叉链接。需要说明的是，TrEMBL 子库中的序列未经手工注释，也未经人工审阅（UnReviewed），可靠性远不及 Swiss-Prot 子库中的序列，使用时需谨慎。



3)UniProt 统计报表（Statistics）包括哪些主要信息？

UniProtKB 知识库通常每四周更新发布一次。每次发布新版时，同时发布 Swiss-Prot 和 TrEMBL 两个子库的统计报表，除数据总量、更新情况、数据类别、物种分布等基本信息外，还列出所有注释信息更新情况，包括常规注释信息、序列特征注释信息和数据库交叉链接等。熟悉这些注释信息，不仅有助于了解 UniProtKB 知识库主要内容，而且有助于通过高级检索从数据库中快速高效地获取所需信息，有助于利用数据库条目中丰富的注释信息和数据库交叉链接，深入了解研究课题相关或感兴趣的蛋白质。



Introduction
Taxonomic origin
Sequence size
Journal citations
Statistics for some line types
Amino acid composition
Miscellaneous statistics

UniProtKB statistics

Introduction

This is release 2026_01 of UniProtKB, published on **Wed Jan 28 2026**.

Previous release statistics are available from the UniProt FTP server.

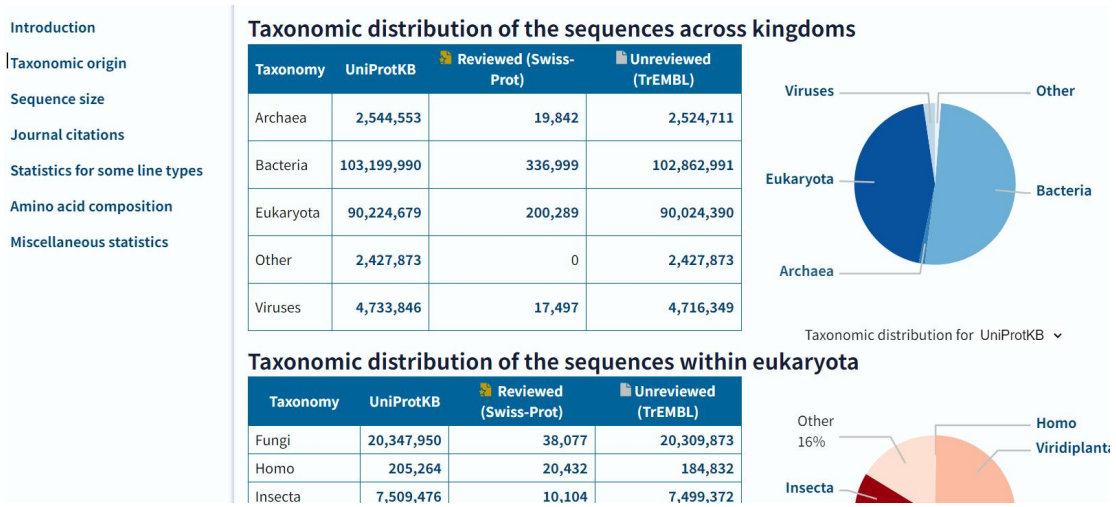
Throughout this document, whenever a statistic has a corresponding query, a link has been provided. In some ir
no query link is possible.

Total number of entries in this release of UniProtKB

Section	Number of entries in total	Number of entries with an annotation update	Number of entries with a sequence update
UniProtKB	**203,130,941**	130,556,322	23,502
Reviewed (Swiss-Prot)	**574,627**	380,278	180
Unreviewed (TrEMBL)	**202,556,314**	130,176,044	23,322

Total number of new entries in this release of UniProtKB

Section	Number of new entries	Number of new sequences
UniProtKB	**3,594,411**	3,594,320



Introduction
Taxonomic origin
Sequence size
Journal citations
Statistics for some line types
Amino acid composition
Miscellaneous statistics

Taxonomic distribution of the sequences across kingdoms

Taxonomy	UniProtKB	Reviewed (Swiss-Prot)	Unreviewed (TrEMBL)
Archaea	**2,544,553**	**19,842**	**2,524,711**
Bacteria	**103,199,990**	**336,999**	**102,862,991**
Eukaryota	**90,224,679**	**200,289**	**90,024,390**
Other	**2,427,873**	0	**2,427,873**
Viruses	**4,733,846**	**17,497**	**4,716,349**

Taxonomic distribution for UniProtKB

Taxonomic distribution of the sequences within eukaryota

Taxonomy	UniProtKB	Reviewed (Swiss-Prot)	Unreviewed (TrEMBL)
Fungi	**20,347,950**	**38,077**	**20,309,873**
Homo	**205,264**	**20,432**	**184,832**
Insecta	**7,509,476**	**10,104**	**7,499,372**

4)UniProt 常规注释信息（General Annotation）包括哪些主要部分？

基于整条序列的常规 注释信息，如功能、表达、亚细胞定位等。

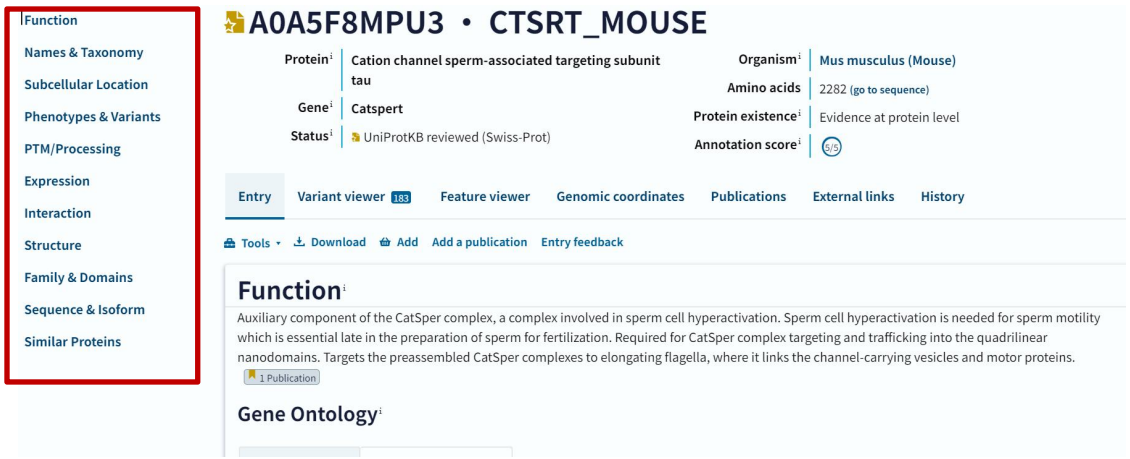


表 5　UniProtKB 知识库 Swiss-Prot 子库中常规注释信息统计表

Table 5　Statistics of general annotation in UniProt/Swiss-Prot

排序	注释信息(英文)	数量	条目数	%
1	序列相似性（Similarity）	507 948	503 800	91
2	功能（Function）	466 212	445 565	83
3	亚细胞定位（Subcellular Location）	349 160	341 774	62
4	催化活性（Catalytic Activity）	282 182	237 510	50
5	亚基（Subunit）	278 602	277 779	50
6	代谢通路（Pathway）	138 188	125 357	25
7	辅助因子（Cofactor）	123 171	112 770	22
8	翻译后修饰（Post-translational Modification，PTM）	56 591	41 747	10
9	结构域（Domain）	48 832	42 035	9
10	组织特异性（Tissue Specificity）	45 806	45 805	8
11	序列警示（Sequence Caution）	44 055	44 055	8
12	杂类（Miscellaneous）	38 582	35 515	7
13	选择性剪接产物（Alternative Products）	25 265	25 265	5
14	诱导（Induction）	20 976	20 964	4
15	相互作用（Interaction）	19 170	19 170	3
16	活性调节（Activity Regulation）	14 807	14 807	3
17	中断表型（Disruption Phenotype）	14 316	14 314	3
18	警示（Caution）	12 858	12 612	2
19	发育阶段（Developmental Stage）	12 154	12 153	2
20	理化特性（Biophysicochemical Properties）	8 251	8 251	1
21	疾病（Disease）	7 031	4 716	1
22	质谱（Mass Spectrometry）	6 938	5 229	1
23	网站资源（Web Resource）	6 749	5 589	1
24	单核苷酸多态性（Polymorphism）	1 324	1 268	<1
25	生物技术（Biotechnology）	974	960	<1
26	过敏原（Allergen）	764	764	<1
27	毒性剂量（Toxic Dose）	668	610	<1
28	RNA 编辑（RNA Editing）	627	627	<1
29	药物（Pharmaceutical）	117	113	<1

5)UniProt 序列特征注释信息（SequenceAnnotation）包括哪些主要部分？

基于序列特定区域或特定位点，因此也称序列特征注释信息。序列特征注释信息共分以下七大类。

（1）分子加工

（2）序列区域

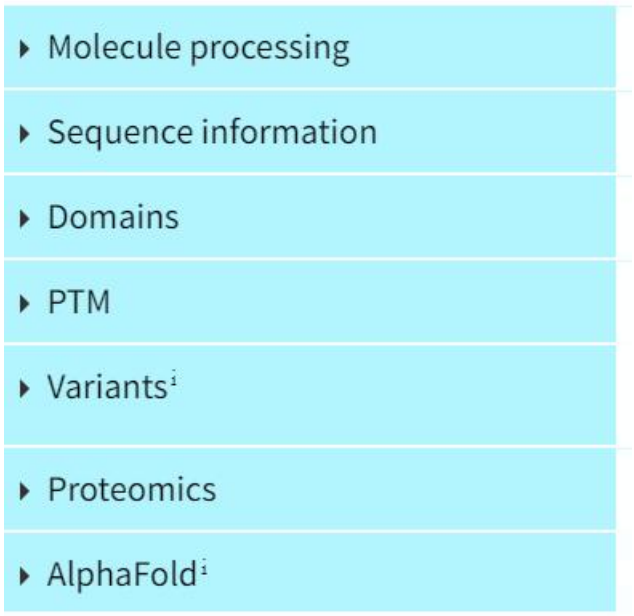
（3）序列位点

（4）氨基酸修饰

（5）天然变异

（6）实验信息

（7）二级结构

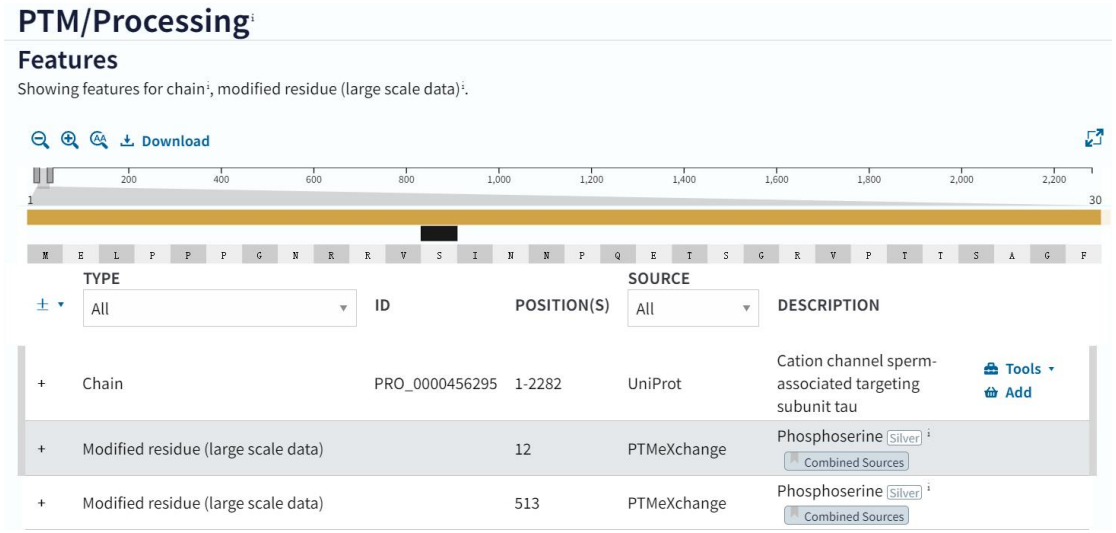


步骤：以一个具体的蛋白质条目（如 BMPR2_HUMAN）为例，有两种主要方式：

方式一：直接在蛋白质条目页面查看（最常用）

当你在搜索结果页面或 ID 映射结果页面得到一个蛋白质列表时，可以高效地同时查看多个蛋白质的特定序列特征：

1. 在 UniProt 网站搜索一个蛋白质（如 BMPR2_HUMAN 或 Q13873）并进入其页面。
2. 向下滚动，找到"PTM / Processing" (翻译后修饰/加工) 部分。
3. 你会看到一个名为 "Feature" (特征) 或 的表格。这个表格就是所有序列特征注释的集合，按氨基酸位置排列，清楚地标明了每种特征的类型（如 DOMAIN, TRANSMEM, MOD_RES）、位置和描述。



方式二：自定义结果显示列（适合批量查看）

当你在搜索结果页面或 ID 映射结果页面得到一个蛋白质列表时，可以高效地同时查看多个蛋白质的特定序列特征：

1. 在搜索结果或 ID 映射结果页面，点击右上角的 "Columns" (列) 按钮。
2. 在弹出的窗口中找到 "Features" (特征) 部分，展开后你会看到所有可选的注释类型（如 DOMAIN, TRANSMEM, MOD_RES 等），与上面表格中的内容完全对应。
3. 勾选你感兴趣的类型（例如勾选 DOMAIN 和 MOD_RES），点击 "Save" (保存)。
4. 页面表格就会新增你选择的列，直接显示每个蛋白质在这些特定特征上的注释息。

6)UniProt 与哪几大类数据库建立了交叉链接（CrossReference）？

（１）序列数据库

（２）蛋白质三维结构数据库

（３）蛋白质相互作用数据库

（４）化学小分子数据库

（５） 特殊类别蛋白质数据库

（６）翻译后修饰数据库

（７）多态性和突变体数据库

（８）双向凝胶电泳数据库

（９）蛋白组数据库

（１０）基因组注释数据库

（１１）特殊物种数据库

7)UniProt 中的帮助文档包括哪些信息？

无论是用户指南中给出的文本检索实例 （text Search），还是有关 UniProt 数据库的基本介绍（About UniProt），或者是常见问题解答（FAQ），以及 UniProtKB 用户手册（User Manual），都提供了大量数据库使用的帮助信息。

2. 基本统计数据

1) 列表说明 Swiss-Prot 和 TrEMBL 子库中人和主要模式生物英文名、拉丁名、分类学数据库（Taxonomy）登录号、序列条目总数、 具有蛋白证据的序列条目数。

物种	英文	拉丁文	登录号	SW 蛋白水平	TrEMBL 蛋白水平
人	Human	*Homo sapiens*	9606	20,431	184,724
小鼠	mouse	*Mus musculus*	10090	17,252	68,647
鸡	chicken	*Gallus gallus*	9031	2,314	49,059
非洲爪蟾	African clawed frog	*xenopus laevis*	8355	3,514	108,387

斑马鱼	zebrafish	*Danio rerio*	7955	3,369	87,308
果蝇	Fruit fly	*Drosophila melanogaster*	7227	3,868	38,774
线虫	Roundworm	*caenorhabditis elegans*	6239	4,499	23,013
酿酒酵母	saccharomyces cerevisiae	*Baker's yeast*	559292	6,733	10
大肠杆菌	escherichia coli	*E.coli*	562	739	483,596
拟南芥	arabidopsis	*Arabidopsis thaliana*	3702	16,418	119,891
粳稻	japonica rice	*Oryza sativa subsp. japonica*	39947	4,197	144,676
课题相关物种（猪）	Pig	*Sus scrofa*	9823	1,462	299,460

3. 人珠蛋白家族检索

1) 写出从 UniProt 数据库中检索已审阅的人珠蛋白（globin）家族 12 个亚基的步骤。

i. 打开 UniProt 官方网站（https://www.uniprot.org/），点击 “Advanced” ，进行高级搜索界面。

ii. 在下拉菜单中选择 "Organism"，在输入框中填写 Human 系统会自动提示条目。

iii. 点击 "Add a new field"，添加一个新的筛选条件，在下拉菜单中选择 "Reviewed"，保持选择状态为 Yes，将结果限定在手动注释的 Swiss-Prot 条目中。

iv. 再次点击 "Add a new field"，选择"Protein name" /”Protein family”，在输入框中输入 globin。

v. 点击 search 查找。

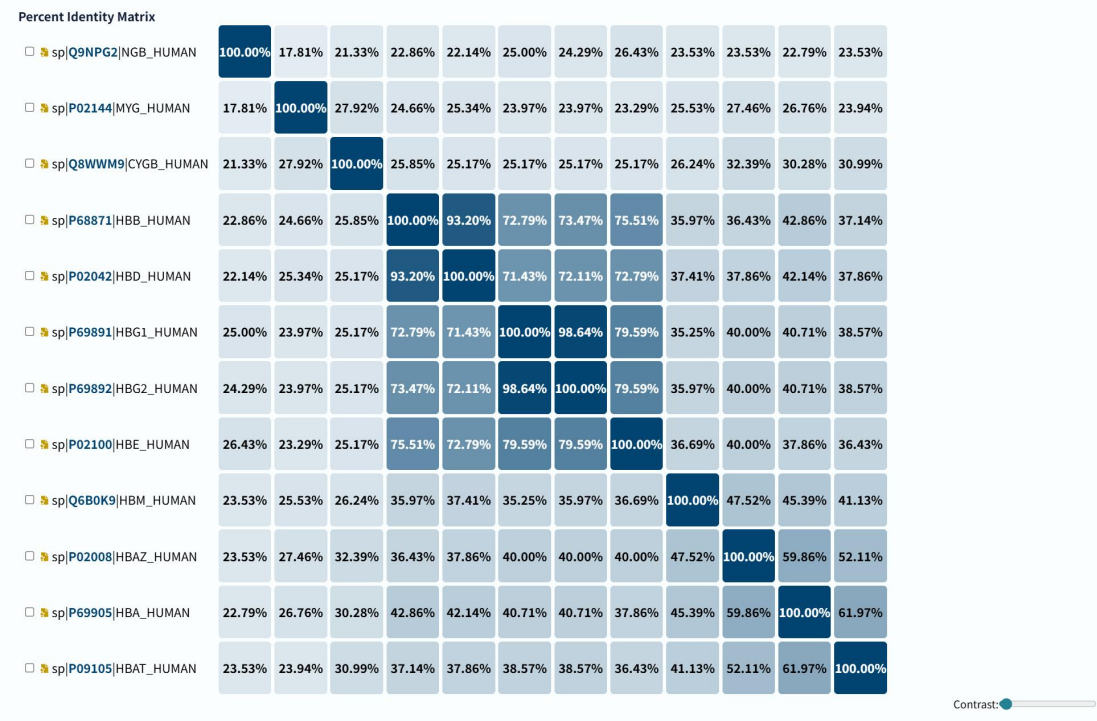
2) 列表说明这 12 个珠蛋白的登录号、蛋白质名称、和序列长度。

珠蛋白名	登录号	蛋白质名称	序列长度
HBE_HUMAN	P02100	Hemoglobin subunit epsilon	147AA
HBD_HUMAN	P02042	Hemoglobin subunit delta	147AA
HBAZ_HUMAN	P02008	Hemoglobin subunit zeta	142AA
HBA_HUMAN	P69905	Hemoglobin subunit alpha	142AA
TFCP2_HUMAN	Q12800	Alpha-globin transcription factor CP2	502AA
HBB_HUMAN	P68871	Hemoglobin subunit beta	147AA
HBG1_HUMAN	P69891	Hemoglobin subunit gamma-1	147AA
HBG2_HUMAN	P69892	Hemoglobin subunit gamma-2	147AA
NPRL3_HUMAN	Q12980	GATOR1 complex protein NPRL3	569AA
HBM_HUMAN	Q6B0K9	Hemoglobin subunit mu	141AA
ANR49_HUMAN	Q8WVL7	Ankyrin repeat domain-containing protein 49	239AA
HBAT_HUMAN	P09105	Hemoglobin subunit theta-1	142AA
NGB_HUMAN	Q9NPG2	Neuroglobin	151AA
CYGB_HUMAN	Q8WWM9	Cytoglobin	190AA
MYG_HUMAN	P02144	Myoglobin	154AA
ADGB_HUMAN	Q8N7X0	Androglobin	1,667AA

*蓝色底纹是通过 protein name 搜索出来的，且不包含 NGB, CYGB, MYG；灰色底纹是通过 protein family 搜索出来的，包括了 12 种珠蛋白，且多出了一种 ADGB（雄红蛋白）。

**雄红蛋白：C 端区域确实含有一个经典的珠蛋白结构域（globin domain），能够结合血红素和可逆结合氧气。这一结构域与血红蛋白、肌红蛋白的珠蛋白折叠高度相似。因此，从结构分类上，它被归入珠蛋白超家族。

3) 与血红蛋白 alpha 亚基差异最大的序列是哪个？相同位点百分比？



答：由图可见与血红蛋白 alpha 亚基差异最大的序列是神经珠蛋白（NGB），相同位点百分比为 22.79%。

4) 与血红蛋白 beta 亚基差异最小的序列是哪个？差异位点共多少个？

答：由图可见与血红蛋白 beta 亚基差异最小的序列是血红蛋白δ亚基（HBD），差异位点 = 147 × (100% - 93.20%) = 147 × 6.8% =9.996， 约 10 个差异位点。

5. 问题:

A. UniProt 中蛋白质序列参考集 UniRef 中有三个数据集，UniRef100、UniRef90 和 UniRef50，其中 UniRef100 的数据集最多，UniRef50 中的数据集最少，但在文献《UniProt 蛋白质数据库简介》示例中，标识符为 UniRef100__ P01308 的记录中含 4 条胰岛素序列，UniRef90__ P01308 记录中包含 11 条序列，UniRef50__ P01308 记录共含 79 条序列。呈现的是 UniRef50 中的序列最多，UniRef100__ P01308 的最少，对此有点疑问。（已理解，以下是 deepseek 的解释）

1. 为什么单个 UniRef50 条目包含的序列反而最多？

核心在于三个数据集“由严到松”的聚类逻辑。

UniRef100 100% (完全相同) 极致严格：只有完全一致的序列才能聚在一起。成员最少。对于胰岛素，它只包含氨基酸序列完全相同的几个成员（如可能来自不同物种但序列恰好完全

相同的亚型）。

UniRef90 ≥ 90% 相对宽松：将相似度超过 90%的序列聚类。 成员增多。它将 UniRef100 的多个簇合并。与 P01308 序列相似度超 90%的序列（如来自不同物种、有少量突变的胰岛素），都会被归入这个更大的家族。

UniRef50 ≥ 50% 非常宽松：将相似度超过 50%的序列聚类。 成员最多。它在此基础上进一步合并，将更广泛的胰岛素/胰岛素样生长因子家族都包含进来。这 79 条序列可能涵盖了多种物种的胰岛素及其同源蛋白。

总结来说：每个 UniRef50 条目就像一个“大家族”，而 UniRef100 条目只是家族中“血缘最近的小家庭”。标准越宽松，聚在一起的“亲戚”就越多，单个条目的成员数量自然就越多。

2. 为什么整体数据库规模却是 UniRef50 最小？

这与“家族”数量有关。宽松的聚类标准，使得大量原本在 UniRef100 中独立的“家庭”，被合并成了数量少得多的“大家族”。

UniRef100：因为是按 100%相同聚类，几乎每个独特的序列都是一个独立的簇。所以，它的代表条目（家族）数量最多，数据库规模最大。

UniRef50：因为标准宽松，可能几百万个 UniRef100 的簇被合并成了几十万个 UniRef50 的簇。所以，它的代表条目（家族）数量最少，数据库规模最小。

6. 组员个人总结

A. G4A 高嘉禾个人总结：

通过对 UniProt 数据库的系统学习，我认识到 Uniprot 的功能非常强大，需要自己探索的地方仍有很多。以下是我在学习过程中觉得比较实用操作体会：

首先，我发现了登录号在 Uniprot 中查询蛋白质非常方便。每个 UniProt 条目都有一个唯一的登录号（Accession Number），例如猪的肌肉抑制素 MSTN 对应 P0C670。当我从文献中读到某个关键蛋白及其登录号后，可以直接在 UniProt 主页的搜索框中输入登录号，这样可以直接查询到这个蛋白，这种方式可以提高信息的获取效率。

另外，Uniprot 的高级检索能够精准定位我们需要查找的内容。在高级检索第一个选项的下拉列表中选择“Organism”，然后输入“Sus scrofa”就可以直接搜索出猪的相关蛋白，后者直接使用登录号 9823 搜索猪的相关蛋白。还可以进行多个条件的搜索，如课外练习第 3 题，限定物种为“Organism”-“Human”，勾选“Reviewed”-“True”（这样会筛选出 Swiss-Prot，即经过人工注释的条目），再勾选“Protein family”-“globin”筛选出人珠蛋白家族的 12 个

亚基。通过这种组合查询，可以快速缩小我们需要的蛋白的范围。

还学习了使用结果页面上的 “Customize columns” 功能设置显示的信息列。默认的结果表格只显示条目名称、登录号、基因名、物种等基础信息，但通过点击“Customize columns”按钮，可以添加我们需要的其他选项或取消不需要的信息选项。例如增加“Protein existence”、“Length”、“Mass”、“3D structure”等。

进入一个具体的蛋白页面后，在左边菜单栏有很多选项，如“Function”、“Names & Taxonomy”、“Subcellular Location”、“Phenotypes & Variants”、“PTM/Processing”、“Expression”、“Interaction”、“Structure”、“Family & Domains”、“Sequence”和“Similar Proteins”。可以根据需要阅读相应的信息。在“Names & Taxonomy”中可以找到蛋白的名称和分类，并且可以在“Taxonomic”中找到 NCBI 的链接跳转到 NCBI 的分类学网站，然后看到与其他数据库交叉的数目和链接。需要序列比对时，可以在“Sequence”中下载蛋白序列。

UniProt 中有 BLAST、Align 等在线工具。BLAST 功能可以对两个蛋白序列进行全局或局部比对。Align 功能则可以用来比较不同蛋白的氨基酸差异，定位可能影响性状的位点。如课外练习的第 3 题需要比对 12 个人珠蛋白亚基，可以得到差异序列和相同位点百分比和差异位点。另外，也可以直接选择某几个蛋白序列，加入到购物篮，并在右上角的购物篮图标中找到这些序列进行 BLAST、Align 和 Map IDs 进行操作。

通过对第四部分的学习，我认识到 UniProt 不仅仅是一个蛋白质序列的数据库，更是一个可以查询、筛选、交叉链接的知识系统，让我对生物信息数据库的实用价值有了全新的认识。以下是课外练习中关于我本人的课题的回答：

*课外练习：

7. 课题相关物种信息

1) 该物种的中文名、英文名、拉丁文学名、分类学登录号。

答：猪，pig, Sus scrofa, P9823

2) 该物种的分类学地位（界、门、纲、目、科、属、种）。

答：Metazoa (动物界) > Chordata (脊索动物门) > Mammalia (哺乳纲) > Artiodactyla (偶蹄目) > Suidae (猪科) > Sus (猪属) > Sus scrofa (野猪种)

3) 该物种在 Swiss-Prot 和 TrEMBL 子库中序列条目数。

答 ：Swiss-Prot：1,462 条；TrEMBL：299,460 条

4) 该物种在 Swiss-Prot 和 TrEMBL 子库中具有蛋白质水平证据的序列条目数。

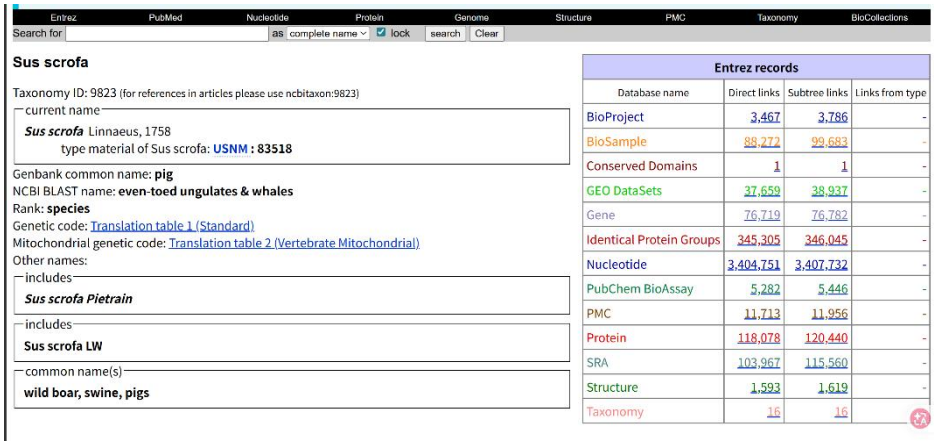
答：点击左侧菜单栏“Protein existence”中的“protein level”查看：

Swiss-Prot：618 条；TrEMBL：15,076 条

5) 该物种在 Swiss-Prot 子库中具有三维空间结构的序列条目数。

答：先选择“reviewed（swiss-prot）”，再点击左侧菜单栏 “Protein with” — “3D structural” 查看：一共有 214 条具有三维空间结构的序列。

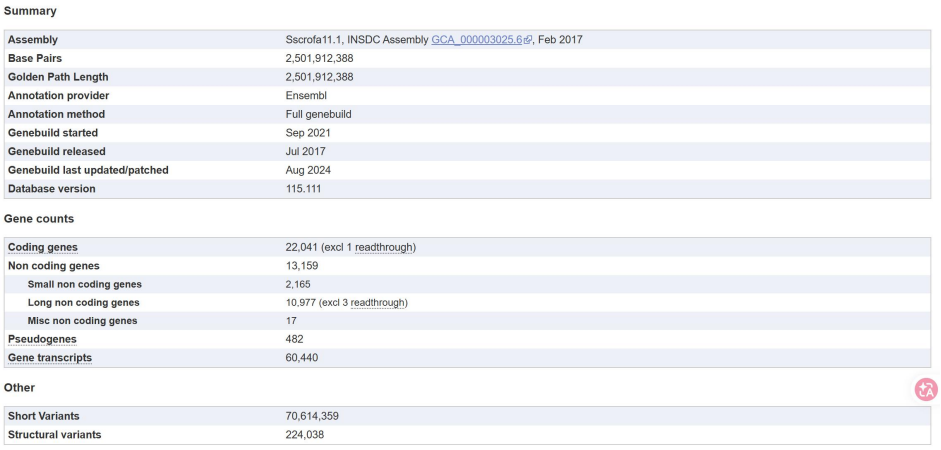
6) 查阅该物种在 NCBI 分类学网站中与其它数据库的交叉链接，列表说明其基本信息。



Database	Information	Direct links
BioProject	是一个可搜索的完整和不完整（正在进行）的大型分子项目的集合，包括基因组测序和组装、转录组、宏基因组、注释、表达和制图项目。可以链接到 NCBI 分子和文献数据库中与项目相关的所有数据。	3,467
BioSample	包含在其他 NCBI 分子数据库（如 Nucleotide 和 SRA）中有数据的研究中使用的生物来源材料的描述。	88,272
Conserved Domains	是一个蛋白质结构域数据库，主要内容是序列比对和分子进化中保守的蛋白质结构域。它还包括在 MMDB 数据库中对已知三维蛋白质结构域的比对。保守域的源数据库包括 Pfam、Smart 和 COG。	1

GEO DataSets	存储了 NCBI 从基因表达 Omnibus （GEO）微阵列数据库中收集的基因表达和分子丰度数据集。	37,659
Gene	是一个可搜索的基因数据库，专注于已经完全测序的基因组，并且有一个活跃的研究社区来贡献基因特异性数据。	76,719
Identical Protein Groups	包含 NCBI 多个来源中每种蛋白质翻译的单个条目，包括 GenBank 和 RefSeq 中的注释编码区，以及 SwissProt 和 PDB 的记录。	345,305
Nucleotide	包含来自国际核苷酸序列数据库协作组成员 GenBank、EMBL 和 DDBJ 的所有序列数据。Nucleotide 还包括 ncbi 策划的参考序列（RefSeqs），来自第三方注释（TPA）数据库的提交程序集和注释，以及从蛋白质数据库（PDB）的结构记录中提取的核苷酸序列。	3,404,751
PubChem BioAssay	包含 PubChem Substance 中描述的化学物质的生物活性筛选。它提供了每种生物测定的可搜索描述，包括特定于筛选程序的条件和读数的描述。	5,282
PMC	PMC（PubMed Central）是美国国家医学图书馆生命科学期刊文献的数字档案。PMC 包含作者和出版商提供的文章的全文手稿。	11,713
Protein	包含从国际核苷酸序列数据库协作（INSDC）成员的 GenBank、EMBL 和 DDBJ 的核苷酸记录中提供的编码区翻译创建的氨基酸序列，以及 NCBI 参考序列和第三方注释（TPA）数据库记录中的编码区翻译的氨基酸序列。	118,078
SRA	SRA （Sequence Read Archive）包含来自下一代测序平台的测序数据。SRA 接受并提供来自所有当前下一代测序平台的数据。	103,967
Structure	包含来自晶体学和核磁共振结构测定的实验数据。MMDB 的数据来源于 Protein data Bank（PDB	1,593
Taxonomy	包含在 NCBI 数据库中有分子数据的生物体的名称和系统发育谱系。当为新分类群存储数据时，它们会添加到 Taxonomy 数据库中。	16

7) 查阅 Ensembl 或 Phytozome 等基因组数据库，若该物种已经完成基因组测序，熟悉其基因组基本信息；若该物种尚未测序， 找出与其亲缘关系最近的物种，熟悉其基本信息。



Summary

Assembly	Sscrofa11.1, INSDC Assembly GCA_000003025.6, Feb 2017
Base Pairs	2,501,912,388
Golden Path Length	2,501,912,388
Annotation provider	Ensembl
Annotation method	Full genebuild
Genebuild started	Sep 2021
Genebuild released	Jul 2017
Genebuild last updated/patched	Aug 2024
Database version	115.111

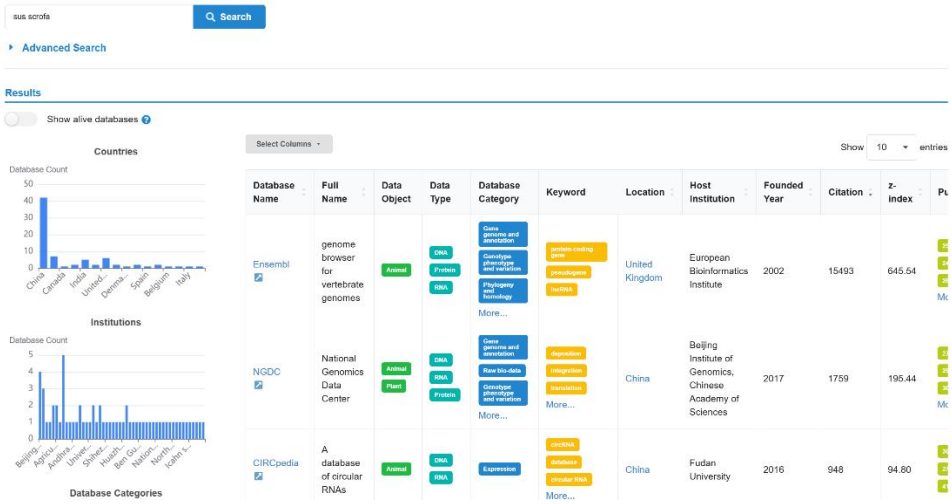
Gene counts

Coding genes	22,041 (excl 1 readthrough)
Non coding genes	13,159
Small non coding genes	2,165
Long non coding genes	10,977 (excl 3 readthrough)
Misc non coding genes	17
Pseudogenes	482
Gene transcripts	60,440

Other

Short Variants	70,614,359
Structural variants	224,038

8) 搜索 Database Common，找出该物种相关数据库，熟悉相关数据库的基本信息和已发表论文。



B. G4B 赵雪倩个人总结：

通过本次学习，掌握了 UniProt 数据库的基本使用流程，明确了在抗菌肽研究中如何利用 Swiss-Prot 获取高质量注释信息，利用序列特征注释分析肽类功能区域，利用交叉链接拓展结构及相互作用信息。以下是具体的内容:

一、UniProt 数据库结构及其在研究中的应用价值

UniProt 由三大核心部分组成："UniProtKB（知识库）"、"UniParc（序列归档库）" 和 "UniRef（序列参考集）"。其中，UniProtKB 分为两个子库：

"Swiss-Prot"：经人工审阅，注释信息全面、可靠，包括功能、亚细胞定位、表达、修

饰等。对于抗菌肽研究，可利用该子库获取经过验证的肽类序列及其功能注释。

- "TrEMBL"：由核酸序列自动翻译生成，未经人工审阅，使用时需谨慎验证。

UniProt 提供的"序列特征注释"（如信号肽、跨膜区、翻译后修饰、天然变异等）对抗菌肽的结构与功能分析尤为重要，有助于判断肽段的活性区域、修饰状态及潜在作用机制。

二、课题相关物种与蛋白信息检索

学习材料中以小鼠（*Mus musculus*）为例，演示了如何查询物种的分类地位、序列条目数、蛋白质水平证据及三维结构信息。对于抗菌肽研究，可借鉴该方法检索目标物种中的抗菌肽家族成员。

此外，UniProt 与多种数据库建立了交叉链接，包括：

- "蛋白质结构数据库"（如 PDB）：用于查看抗菌肽的三维结构；
- "翻译后修饰数据库"：了解肽类是否发生磷酸化、糖基化等修饰；
- "多态性与突变数据库"：分析天然变异对抗菌活性的影响。

三、课外练习：

课题相关物种信息：

1) 该物种的中文名、英文名、拉丁文学名、分类学登录号。

中文名：小鼠；

英文名：mouse；

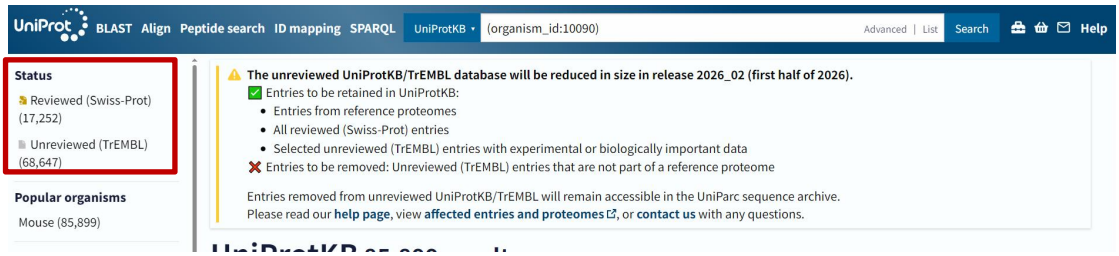
拉丁文学名：Mus musculus；

分类学登录号：10090（NCBI）)

2) 该物种的分类学地位（界、门、纲、目、科、属、种）。

Metazoa (动物界) > Chordata (脊索动物门) > Mammalia (哺乳纲) > Rodentia (啮齿目) > Muridae (鼠科) > Mus (小鼠属) > Mus musculus (家鼠种)

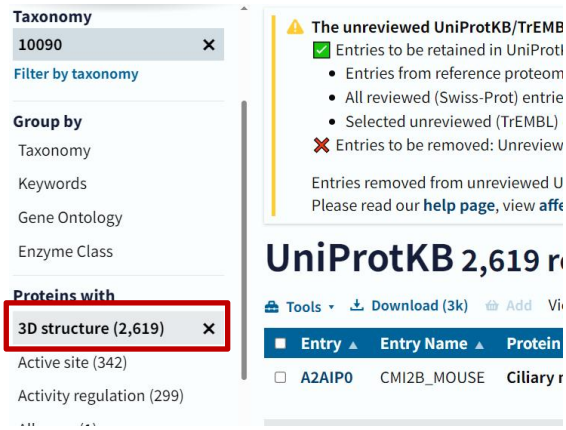
3) 该物种在 Swiss-Prot 和 TrEMBL 子库中序列条目数。



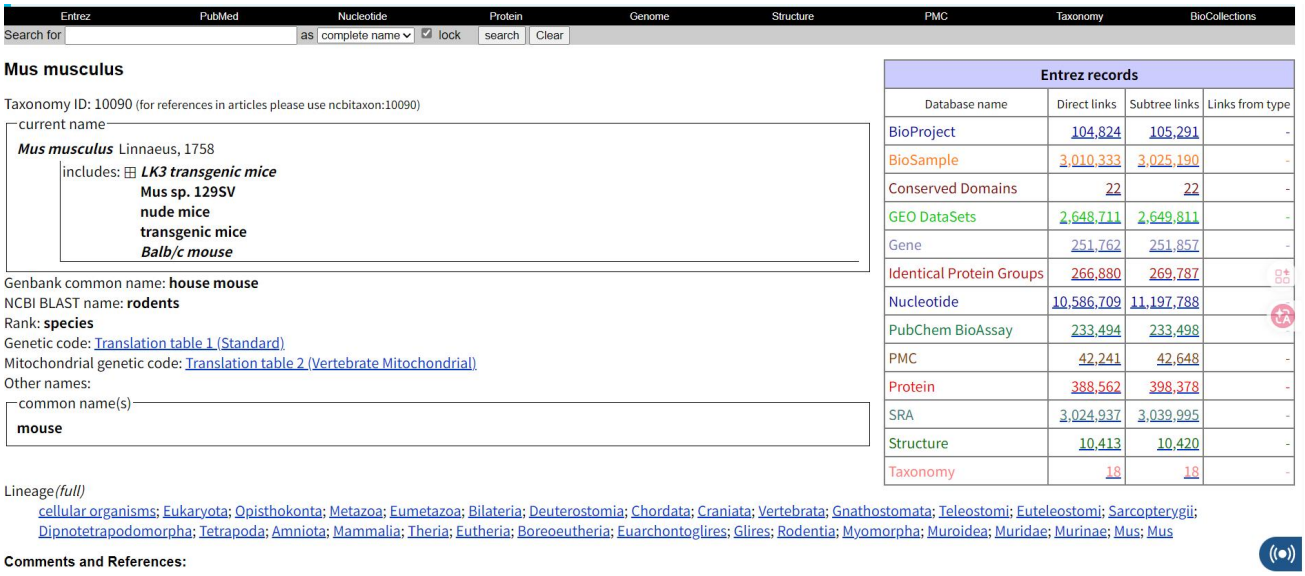
4) 该物种在 Swiss-Prot 和 TrEMBL 子库中具有蛋白质水平证据的序列条目数。

Swiss-Prot：13,322 条；TrEMBL：26,514 条

5) 该物种在 Swiss-Prot 子库中具有三维空间结构的序列条目数。



6) 查阅该物种在 NCBI 分类学网站中与其它数据库的交叉链接，列表说明其基本信息。



Database	Information	Direct links
BioProject	是一个可搜索的完整和不完整（正在进行）的大型分子项目的集合，包括基因组测序和组装、转录组、宏基因组、注释、表达和制图项目。可以链接到 NCBI 分子和文献数据库中与项目相关的所有数据。	104,824
BioSample	包含在其他 NCBI 分子数据库（如 Nucleotide 和 SRA）中有数据的研究中使用的生物来源材料的描述。	3,010,333
Conserved Domains	是一个蛋白质结构域数据库，主要内容是序列比对和分子进化中保守的蛋白质结构域。它还包括在 MMDB 数据库中对已知三维蛋白质结构域的比对。保守域的源数据库包括 Pfam、Smart 和 COG。	22
GEO DataSets	存储了 NCBI 从基因表达 Omnibus （GEO）微阵列数据库中收集的基因表达和分子丰度数据集。	2,648,711
Gene	是一个可搜索的基因数据库，专注于已经完全测序的基因组，并且有一个活跃的研究社区来贡献基因特异性数据。	251,762
Identical Protein Groups	包含 NCBI 多个来源中每种蛋白质翻译的单个条目，包括 GenBank 和 RefSeq 中的注释编码区，以及 SwissProt 和 PDB 的记录。	266,880
Nucleotide	包含来自国际核苷酸序列数据库协作组成员 GenBank、EMBL 和 DDBJ 的所有序列数据。Nucleotide 还包括 ncbi 策划的参考序列（RefSeqs），来自第三方注释（TPA）数据库的提交程序集和注释，以及从蛋白质数据库（PDB）的结构记录中提取的核苷酸序列。	3,024,937
PubChem BioAssay	包含 PubChem Substance 中描述的化学物质的生物活性筛选。它提供了每种生物测定的可搜索描述，包括特定于筛选程序的条件和读数的描述。	10,413
PMC	PMC（PubMed Central）是美国国家医学图书馆生命科学期刊文献的数字档案。PMC 包含作者	42,214

	和出版商提供的文章的全文手稿。	
Protein	包含从国际核苷酸序列数据库协作（INSDC）成员的 GenBank、EMBL 和 DDBJ 的核苷酸记录中提供的编码区翻译创建的氨基酸序列，以及 NCBI 参考序列和第三方注释（TPA）数据库记录中的编码区翻译的氨基酸序列。	388,562
SRA	SRA （Sequence Read Archive）包含来自下一代测序平台的测序数据。SRA 接受并提供来自所有当前下一代测序平台的数据。	3,024,937
Structure	包含来自晶体学和核磁共振结构测定的实验数据。MMDB 的数据来源于 Protein data Bank（PDB	10,413
Taxonomy	包含在 NCBI 数据库中有分子数据的生物体的名称和系统发育谱系。当为新分类群存储数据时，它们会添加到 Taxonomy 数据库中。	18

C. G4C 张俊鑫个人总结：

Uniprot 数据库在水稻作物种质资源学专业中的应用个人思考总结：

UniProt 作为全球权威蛋白质数据库，为水稻种质资源学研究提供了核心蛋白信息支撑，是连接基因序列与功能表型的关键枢纽，在种质功能解析、优异基因挖掘与分子育种中作用显著。

在种质资源功能注释方面，UniProt 收录了籼稻、粳稻及野生稻的高质量蛋白序列与功能注释。通过检索目标基因对应的蛋白条目，可快速获取亚细胞定位、结构域、翻译后修饰位点等信息，为解析不同水稻种质的性状差异提供分子依据。

在优异种质基因挖掘中，利用 BLAST 比对水稻种质的蛋白序列与 UniProt 参考序列，能精准定位抗逆、高产、优质等关键蛋白的变异位点，判断变异对功能的影响。同时可查找同源蛋白，预测未知基因功能，大幅提升种质资源中优异等位基因的筛选效率。

在种质演化与多样性研究中，通过多序列比对 UniProt 中不同水稻种质的保守蛋白，可构建系统发育树，阐明籼粳分化与野生稻亲缘关系，为种质分类与保护提供蛋白层面证据。

作为种质资源研究的核心工具，熟练运用 UniProt，能有效推动水稻种质从序列到功能的高效挖掘与利用。