“实用**生物信息技术**”**课程小组讨论总结报告**

组次：G5　组长：韩佳凝　执笔：赵宏芳

1. 时间

2026.3.25

2. 方式

线上腾讯会议讨论

3. 主题

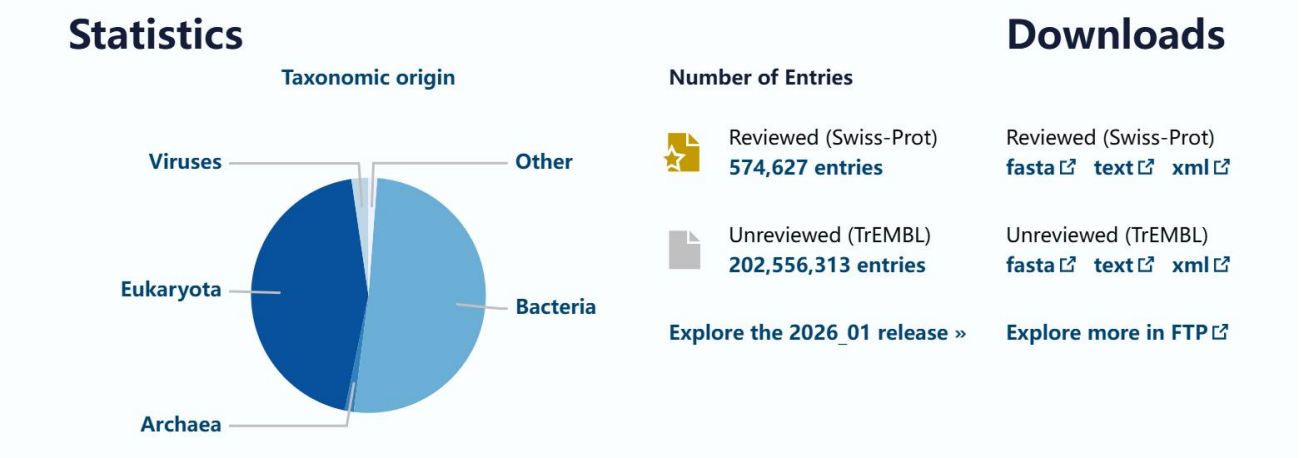
课后对 Uniprot 的复习总结

4. 总结内容

4.1Uniprot 的内容

Uniprot 共包含三部分

UniprotKB(蛋白质知识库)：UniprotKB 由两部分组成，一部分是包含经过人工标注的记录的区域，这些记录包含了 Swiss-Prot（从文献中提取的信息以及经过评估的计算分析结果）；另一部分是 TrEMBL(包含经过计算分析但尚待完成人工标注的记录区域）



UniParc(蛋白质序列归档库）

Databases

Cross-reference	Number of UniParc entries
EMBLWGS	665,739,677
RefSeq	509,606,349
UniProtKB	329,016,034
EMBL CDS	117,016,034
EnsemblBacteria	108,696,350

Searching the Database

UniParc provides text- and sequence-based searches. Performing a sequence similarity search against UniParc is equivalent to performing the same search against all databases cross-referenced in UniParc, as UniParc contains all proteins from its source databases.

Start a sequence similarity search in UniParc using BLAST »

Download

UniParc sequence archive

- **xml**
- **fasta**
- **README**

Explore more in FTP

UniRef(蛋白质序列参考集）

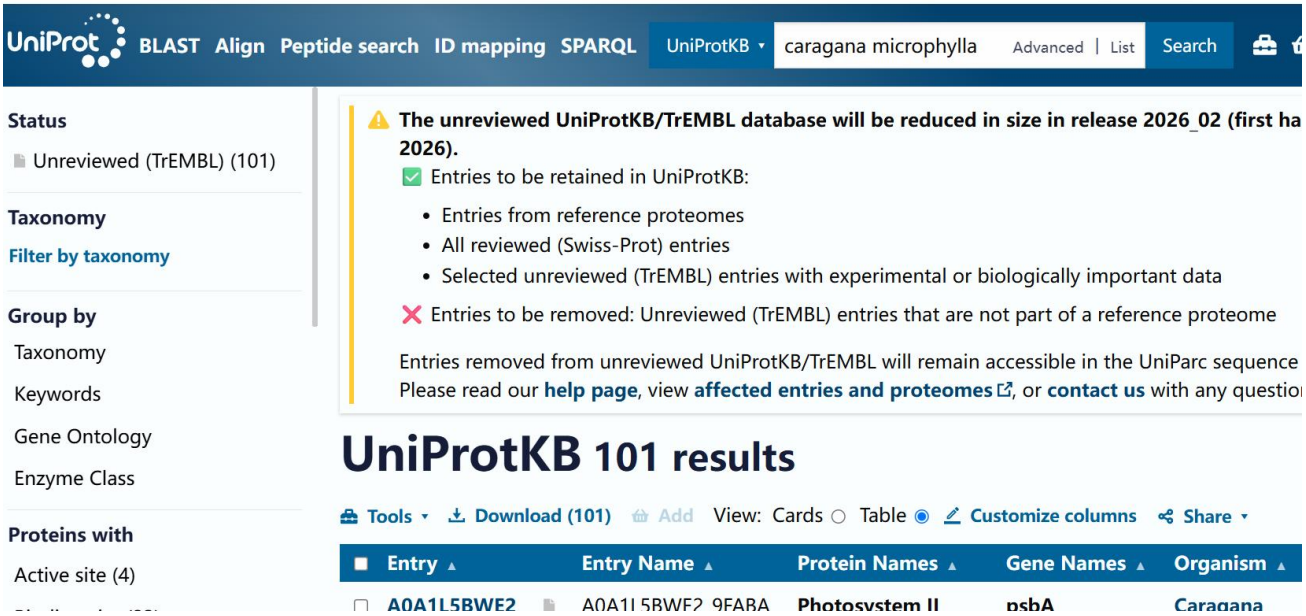
The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the **UniProt Knowledgebase** (including **isoforms**) and selected **UniParc** records in order to obtain complete coverage of the sequence space at several resolutions while hiding redundant sequences from view.

4.2 Uniprot 的使用

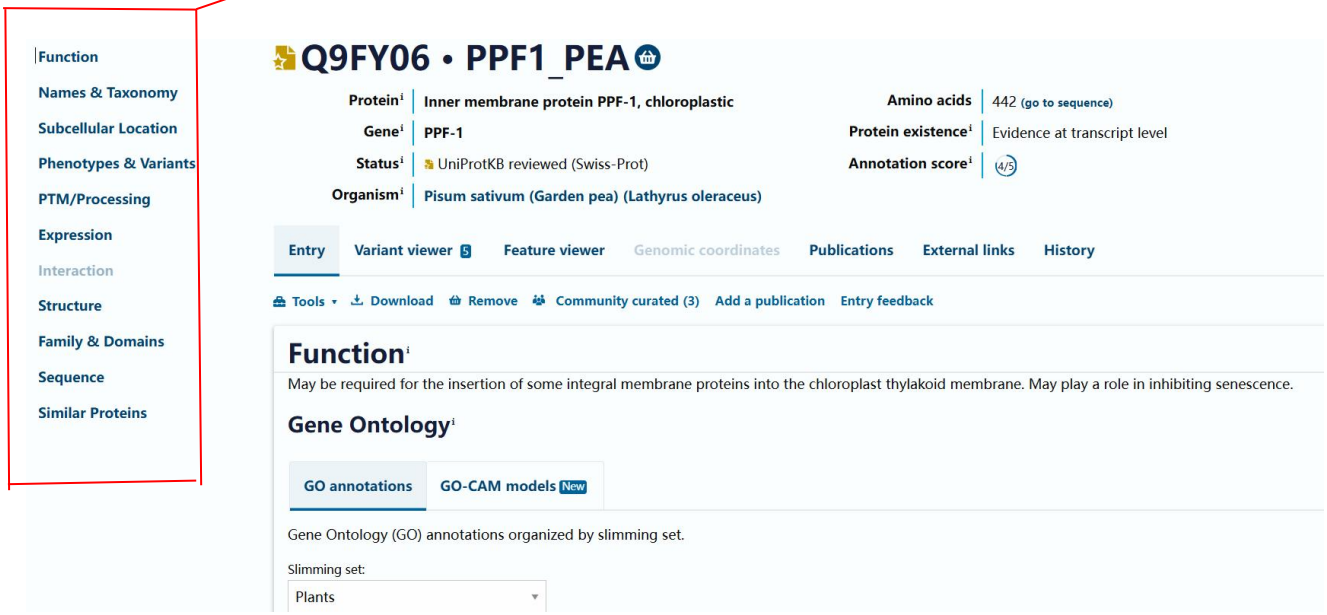
在 abc 英文主页可直接点击进入 Uniprot 主页

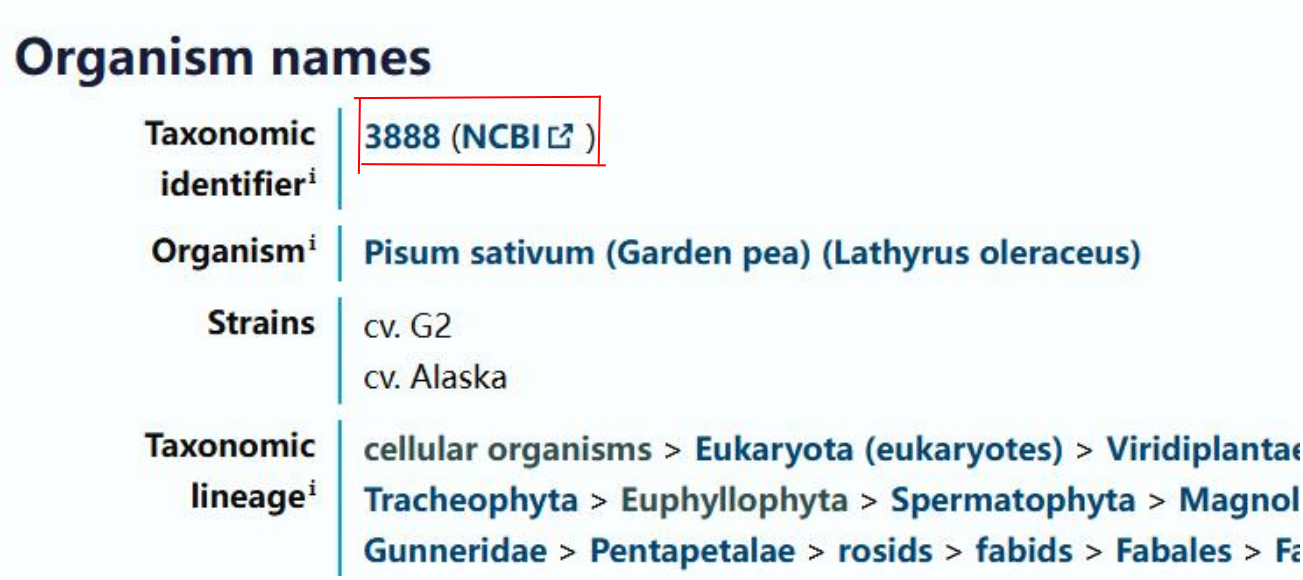
4.2.1 基础搜索

在搜索框中直接输入关键词即可，可输入的内容包括：基因名（如 tp53 或 ACE2）；UniProt 登录号（如 P04637）；关键词（apoptosis）；物种名（如 human、mouse）；蛋白名（如 insulin receptor）

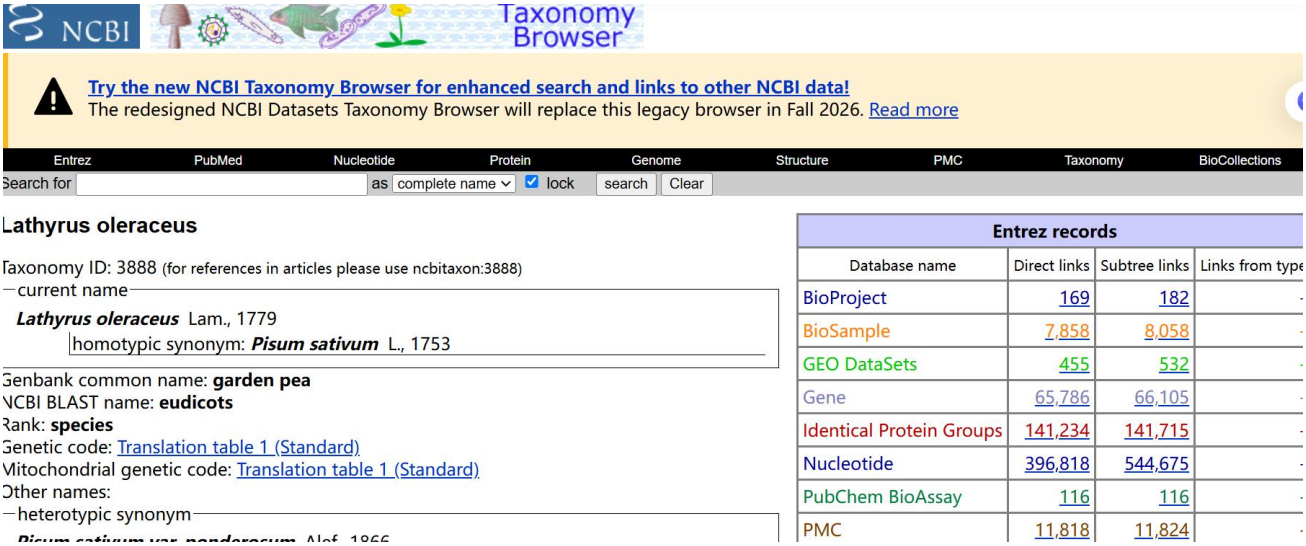


以内膜蛋白为例　　可点击检索出该蛋白的各种信息



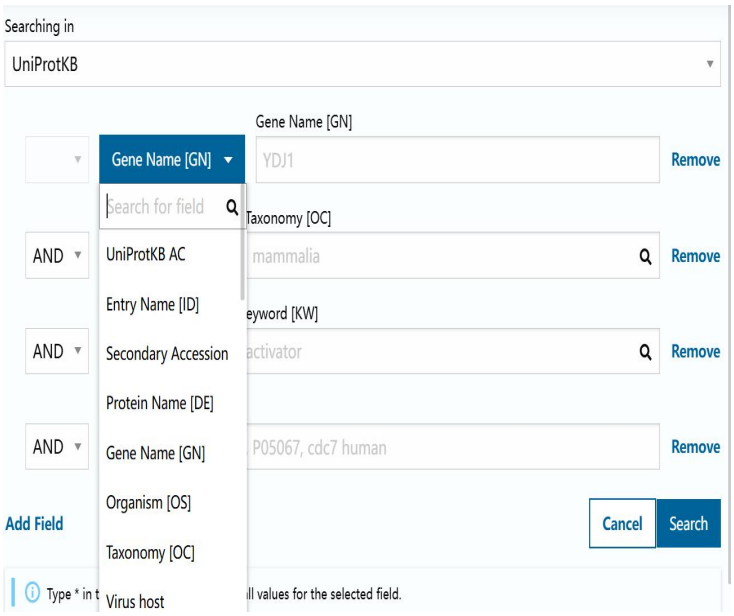


可直接点击 NCBI 进入，查阅到相关文章



4.2.2 高级检索

点击搜索框后的 Advanced 即可进入



按字段检索：从下拉菜单中选择检索字段（如 Organism、Gene name、Function、Disease 等），再输入关键词

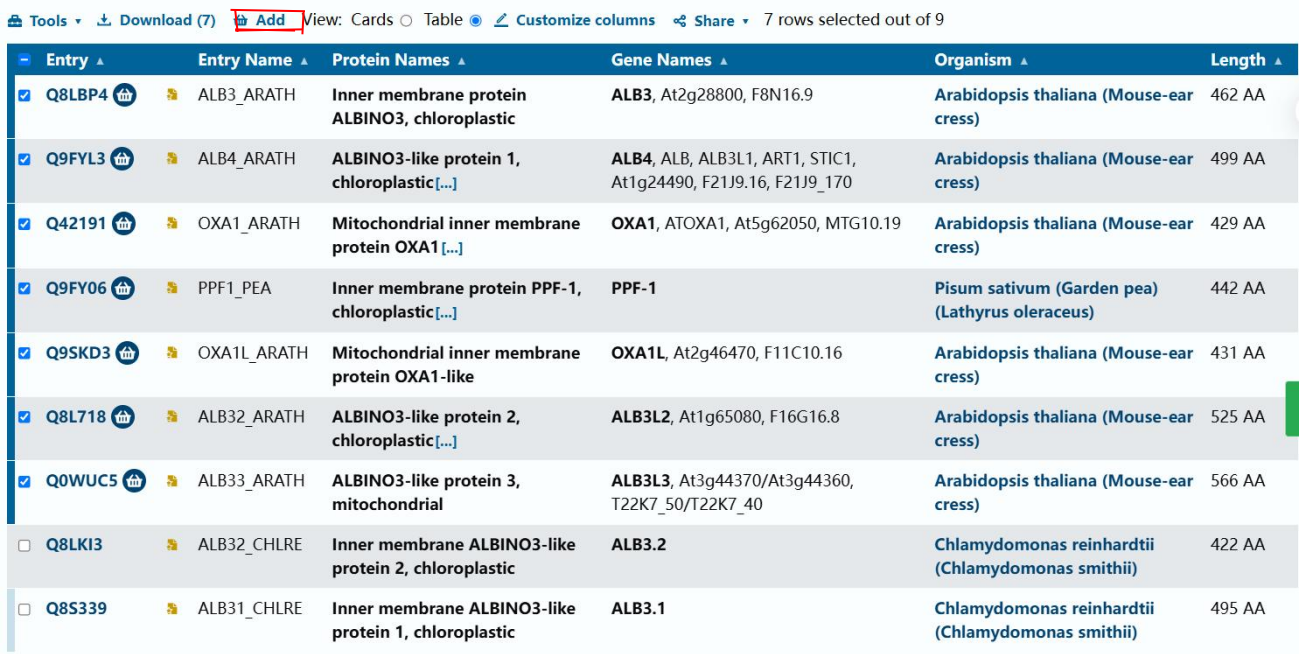
组合条件：通过“Add field”添加多个条件，并用 AND、OR、NOT 进行逻辑组合

4.2.3 批量检索与下载（可以一次性得到多个蛋白的信息）

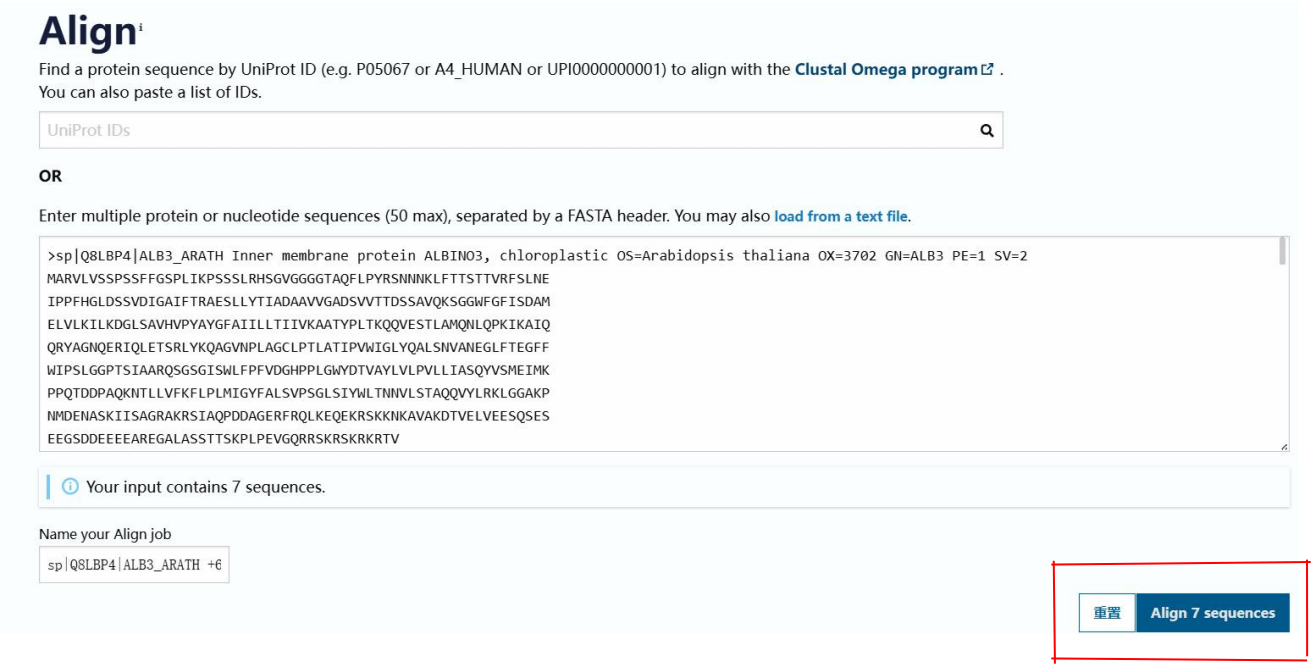
首先进入高级检索界面，输入所需要的所有信息，搜索后，对得到的结果进行 Download 下载，选择格式（TSV 或 Excel），即可下载。

4.2.4 序列比对 （以拟南芥为例）

首先，选种要比对的拟南芥蛋白，点击 add 添加

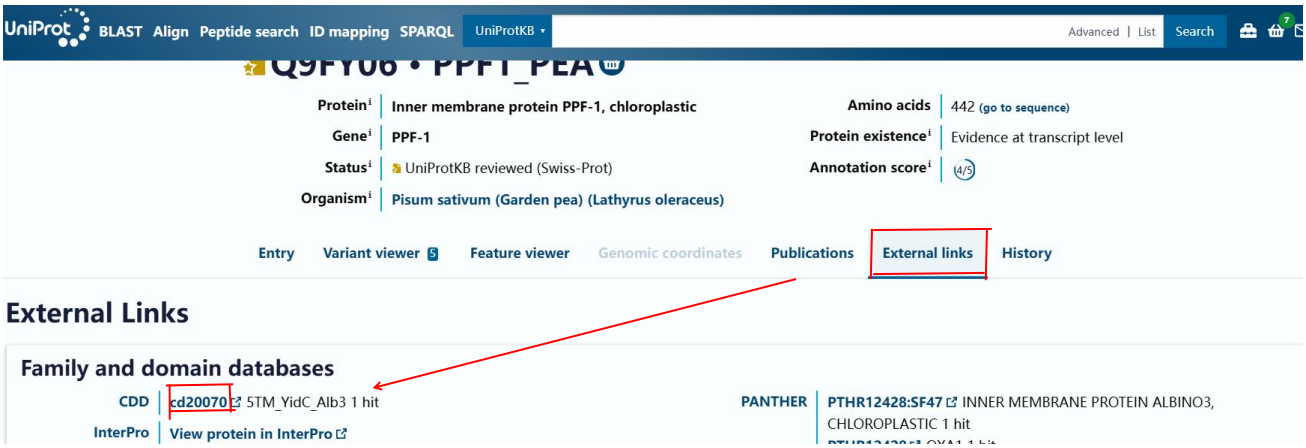


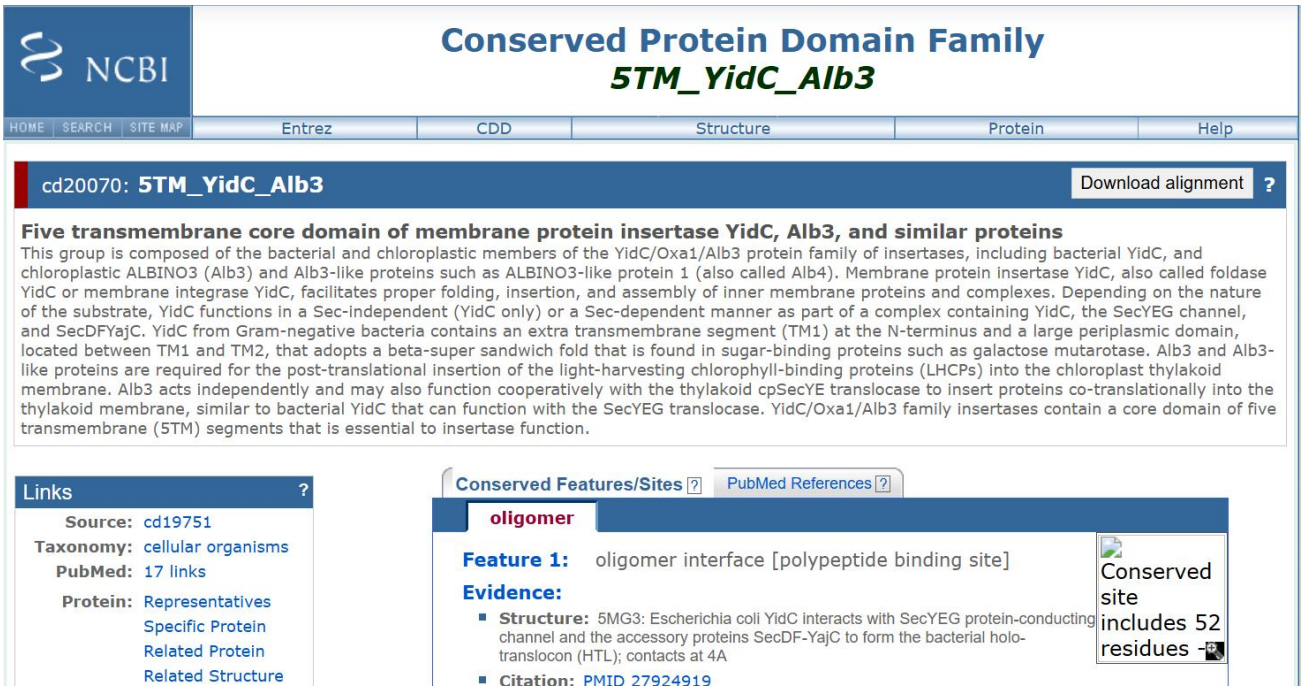
下拉 Tools,点击 Align

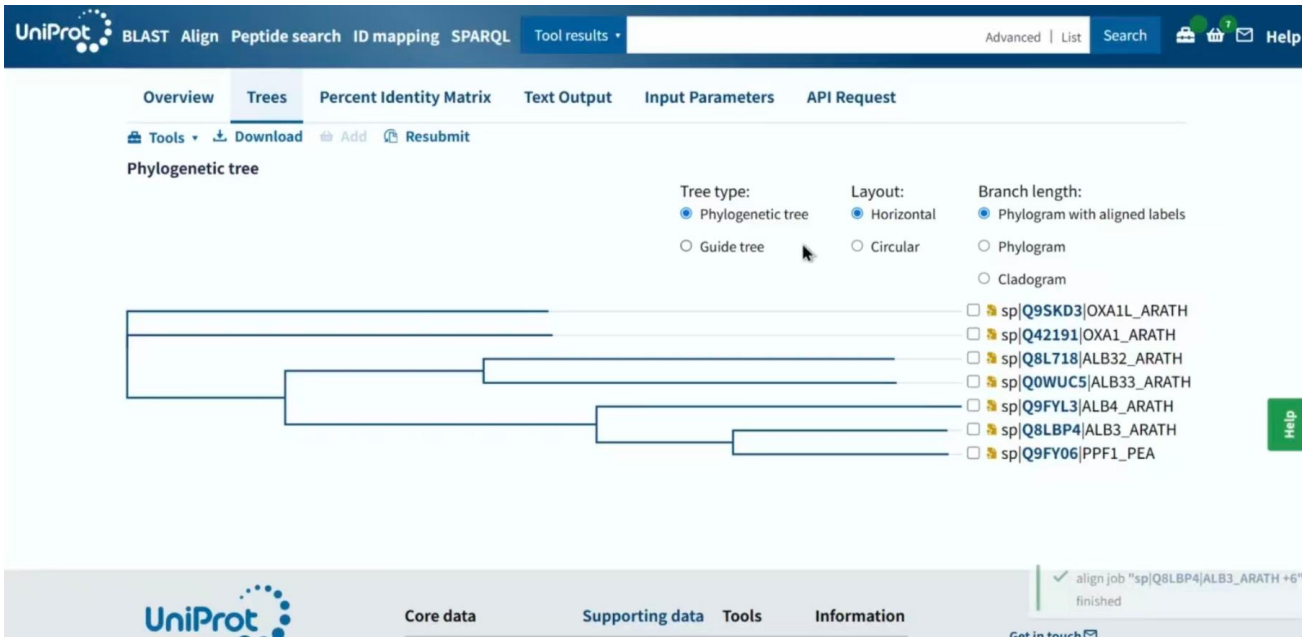


4.2.4（以 PPF1-PEA 为例）

External links（外部连接），可直接进入 NCBI 数据库，查找蛋白信息。







4.3 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	英文名	拉丁文名	Taxonomy 登录号	SW 蛋白水平（PE=1 条目数）	TrEMBL 蛋白水平（PE=1 条目数）
人	Human	*Homo sapiens*	9606	~20,400	~1,500
小鼠	Mouse	*Mus musculus*	10090	~17,200	~1,200
鸡	Chicken	*Gallus gallus*	9031	~6,500	~800
非洲爪蟾	African clawed frog	*Xenopus laevis*	8355	~2,800	~500
斑马鱼	Zebrafish	*Danio rerio*	7955	~5,200	~900
果蝇	Fruit fly	*Drosophila melanogaster*	7227	~10,800	~300
线虫	Nematode	*Caenorhabditis elegans*	6239	~7,900	~200
酿酒酵母	Baker's yeast	*Saccharomyces cerevisiae*	4932	~6,700	~100
大肠杆菌	E. coli	*Escherichia coli*	562	~4,300	~150
拟南芥	Thale cress	*Arabidopsis thaliana*	3702	~16,400	~600

4.4 个人总结

G5A 韩佳凝　　课题涉及物种：草地早熟禾

由于草地早熟禾非模式生物，且野外植株染色体倍数有极大差异，数据库收录数据有限。

（1）英文名 Kentucky bluegrass / Common meadow-grass

拉丁文学名　Poa pratensis L.

NCBI Taxonomy 登录号　4545

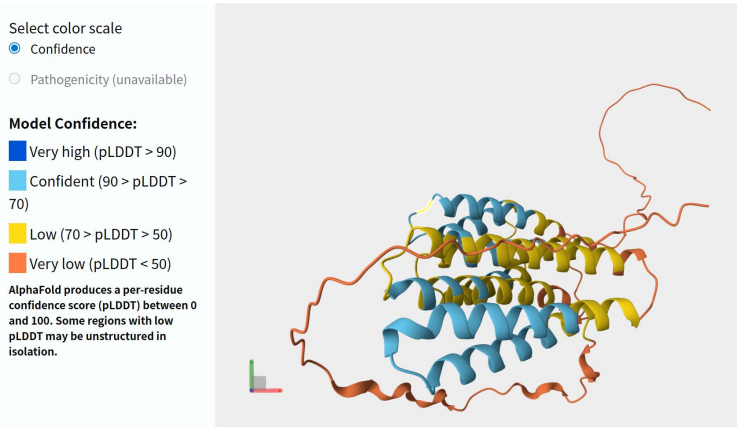
（2）

分类层级	名称
界	植物界（Plantae）
门	维管植物门（Tracheophyta）
纲	单子叶植物纲（Liliopsida）
目	禾本目（Poales）
科	禾本科（Poaceae）
属	早熟禾属（*Poa*）
种	草地早熟禾（*Poa pratensis*）

（3）草地早熟禾在 Swiss-Prot 和 TrEMBL 子库中序列条目数分别为 5 条和 378 条

（4）草地早熟禾在 Swiss-Prot 和 TrEMBL 子库中具有蛋白质水平证据的序列条目数分别为 5 条和 0 条

（5）Swiss-Prot 子库中具有三维空间结构的序列条目数严格意义上为 0 条，均没有实验解析，但全部 Swiss-Prot 条目均有 AlphaFold 预测的三维结构



（6）数据库交叉信息：PubMed 收录超 3000 篇关于其抗逆性、育种、生理生态的研究论文

（7）草地早熟禾已完成完整叶绿体基因组测序（长度 135,649 bp，含 131 个基因）；核基因组有初步组装数据（约 6.09 Gbp），但无高质量参考基因组发布。

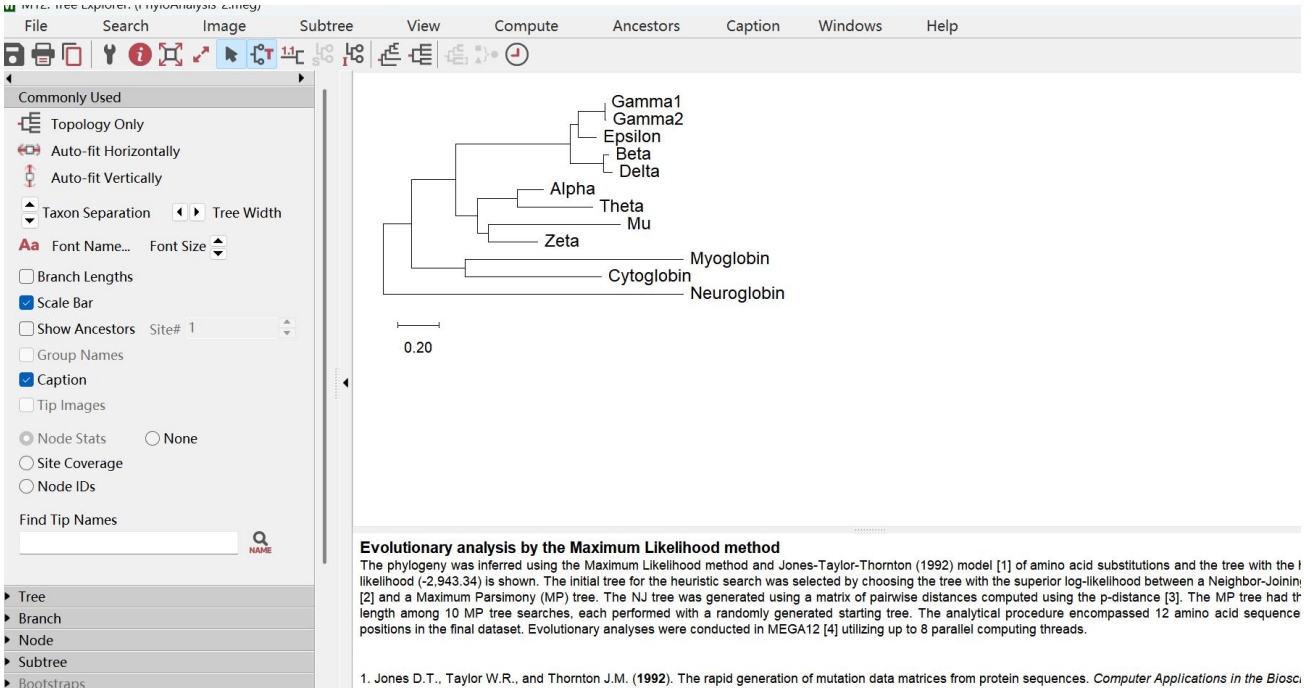
亲缘关系最近且有高质量基因组的物种为一年生早熟禾，其次为多年生黑麦草和二穗短柄草，而在抗旱性转评价中也常用水稻的数据作为参考。

5　预习内容

5.1 系统发生树构建 MEGA 的使用

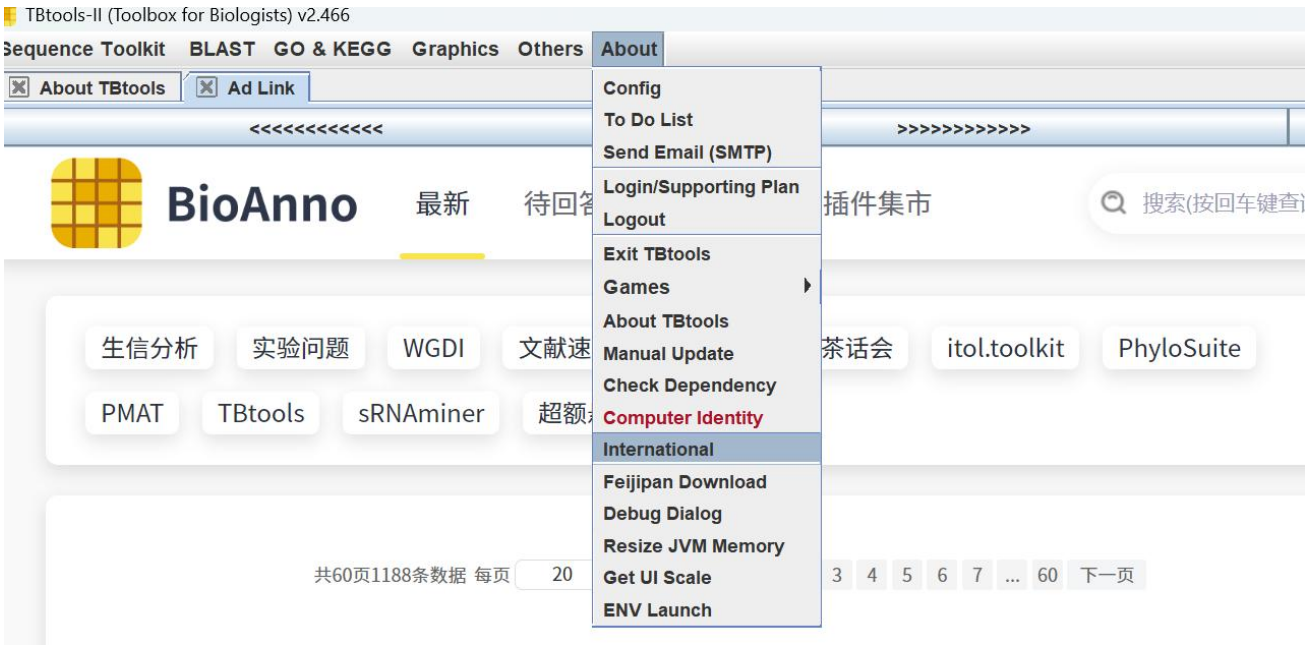
利用 MEGA12 构建系统发育树，可基于 DNA 或蛋白质序列差异，明确不同物种、菌株或基因间的亲缘关系远近与系统发育地位，判断其演化分支与分类合理性；同时能推断类群的起源、演化方向及分化时间，分析基因家族的复制与进化历程，并通过自举检验验证分支结果的可靠性，结合进化速率与选择压力分析，还可进一步揭示物种适应性进化特征，整体用于阐明生物类群的系统演化关系与进化历史。

以珠蛋白家族 12 个蛋白序列为例



5.2 TBtools 热图的构建

主页-About-international-可视化-热图-可制作出各种热图



热图能清晰呈现多个基因在不同组织、发育时期或胁迫处理下的表达高低，通过颜色深浅直观对比表达量差异；可快速识别共表达基因模块，判断哪些基因表达趋势一致，推测它们可能参与同一生物学通路；能筛选组织特异性或响应处理的关键基因，为后续功能验证提供靶点；同时便于比较不同分组样本的表达聚类情况，反映样本间重复性与生物学差异，辅助解释表型机制，是转录组分析、基因家族研究中展示表达模式、支撑科学结论的常用可视化手段。

6 问题

（1）UniProt 特定蛋白质页面左侧有许多板块，比如 Function 等等，有一些板块是褪色点击不了的状态，是因为相应的研究没有证明该蛋白质部分功能吗？

基本是这样的，意味着目前在该板块对应的领域内，没有足够的、经过验证的科学证据或注释信息。

（2）蛋白质结构模块中会有不同的来源和方法，有什么区别吗？为什么大多显示 AlphaFold 的结果？

SOURCE -- Select --	IDENTIFIER	METHOD -- Select --	RESOLUTION	CHAIN	POSITIONS	LINKS	
AlphaFold DB	AF-Q32880-F1	Predicted			1-702	**AlphaFold DB**	**Foldseek**
SWISS-MODEL	Q32880_1-701:9grx.1.P	Modeling		_, P	1-701	**SWISS-MODEL**	
AlphaFill	Q32880	Predicted		A, B, C, E, F, G, D, H	1-702	**AlphaFill**	

- AlphaFold 是一个革命性的人工智能系统，能够根据蛋白质的氨基酸序列，以前所未有的高精度预测其三维结构。Predicted (预测)表明这不是通过实验（如 X 射线晶体学）测出来的，而是通过 AI 算法算出来的。且 AlphaFold DB 是目前公认最准确的 AI 预测结果。
- SWISS-MODEL 是一个自动化的蛋白质同源建模服务器。它通常基于已知的实验结构（模板）来构建目标蛋白的模型。这里的“建模”通常指同源建模，即“找一个长得像的已知结构作为参考”来搭建模型，这与 AlphaFold 的深度学习预测略有不同，且只覆盖了 701 个氨基酸。
- AlphaFill 是一个基于 AlphaFold 模型的补充数据库。AlphaFill 的作用是把辅酶、金属离子或小分子药物重新填补（Fill）回 AlphaFold 的模型中，使其更接近真实的生物功能状态。