“实用生物信息技术”课程小组讨论总结报告

组：G6　　次：2　　组长：冯睿桥　　执笔：　冯睿桥、崔硕、凌子涵

1. 时间

2026 年 3 月 30 日、4 月 1 日

2. 方式

线上线下交流学习

3. 主题

Uniprot 数据库使用

4. 内容

A. 问题讨论

UniProt 数据库概况

UniProt（Universal Protein Resource）是国际上权威的综合性蛋白质序列与功能信息数据库，由欧洲生物信息研究所（EBI）、瑞士生物信息学研究所（SIB）和美国蛋白质信息资源（PIR）共同维护，旨在为全球科研界提供全面、高质量且免费的蛋白质数据资源。

(1) UniProt 数据库由哪三部分组成？

UniProt 核心由三大数据库构成：

UniProtKB（UniProt Knowledgebase，UniProt 知识库）：核心数据库，收录带详细功能注释的蛋白质序列与信息。

UniRef（UniProt Reference Clusters，UniProt 参考序列簇）：按序列一致性（100%、90%、50%）对蛋白质序列聚类，减少冗余、方便检索。

UniParc（UniProt Archive，UniProt 归档库）：非冗余的全球公开蛋白质序列总归档库，收录所有来源的完整序列历史记录。

(2) UniProtKB 知识库由哪两部分组成？

UniProtKB 分为两个互补部分 UniProt：

UniProtKB/Swiss-Prot（Reviewed，已审阅）：人工注释、高质量、非冗余数据集，信息经专家文献审核，含明确功能、结构、互作等注释。

UniProtKB/TrEMBL（Unreviewed，未审阅）：自动注释、高覆盖数据集，由核酸编码序列翻译而来，未经人工审核，用于快速收录基因组数据。

(3) UniProt 统计报表（Statistics）包括哪些主要信息？

主要统计内容 UniProt：

版本与总量：当前发布版本、总条目数、Swiss-Prot 与 TrEMBL 条目数、更新条目数。

分类统计：物种分布（界、门、种）、序列长度分布、蛋白质存在证据（PE）等级分布。

注释统计：功能、结构域、翻译后修饰、亚细胞定位等注释项的覆盖度。

引用统计：文献来源、高引期刊、数据引用频次。

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

常规注释涵盖蛋白质核心背景信息：

名称与标识：蛋白质名、基因名、UniProt 登录号、别名。

物种信息：来源物种、分类学归属（OC 行）、物种分类 ID（OX）。

功能描述：生物学功能、催化活性、分子功能、参与通路。

亚细胞定位：胞内 / 胞外 / 膜定位、组织特异性表达。

相互作用：蛋白质 - 蛋白质、配体、核酸互作。

疾病关联：缺陷 / 突变相关疾病、表型。

文献与证据：引用文献、注释证据代码（ECO）。

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

聚焦序列结构与修饰特征 UniProt：

序列信息：氨基酸序列、长度、分子量、等电点、序列冲突 / 变异。

加工处理：信号肽、前体切割、成熟肽区域。

结构特征：结构域、跨膜区、螺旋 / 折叠、二硫键、结合位点（ATP / 金属等）。

翻译后修饰（PTM）：磷酸化、糖基化、乙酰化、泛素化等位点。

拓扑与特征：膜拓扑、重复序列、低复杂度区。

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

与超 180 个外部库交叉引用，主要类别 UniProt：

序列数据库：ENA/GenBank/DDBJ、RefSeq、PIR

3D 结构数据库：PDB、PDBe、PDBj、SWISS-MODEL

家族与结构域：Pfam、SMART、InterPro、PROSITE

酶与通路：KEGG、ENZYME、Reactome

基因表达：GEO、ArrayExpress

变异与疾病：OMIM、ClinVar、HGMD

互作数据库：IntAct、STRING、DIP

本体与分类：GO（Gene Ontology）

物种特异库：FlyBase、SGD、WormBase、TAIR

(7) UniProt 中的帮助文档包括哪些信息？

帮助文档（Help）覆盖使用与数据规范 UniProt：

入门指南：检索、浏览、下载、高级搜索。

注释规范：人工 / 自动注释规则、命名标准、证据代码。

数据结构：字段含义、条目结构、交叉引用说明。

工具使用：BLAST、批量检索、API 编程访问。

常见问题：数据更新、标识符、序列质量、统计解释。

发布说明：版本更新、新增内容、格式变更。

B. 课堂内容复习与互助

Uniprot 蛋白质数据库检索

(1) 基础检索和高级检索

在官网首页（https://www.uniprot.org/）的搜索框中输入关键词，点击搜索，可以输入的内容：基因名，蛋白名，物种名或拉丁名，登录号，关键词等。高级检索：点击搜索框右侧 Advanced，可以通过搜索字段和条件组合进行搜索，常用字段（Field）：Protein name，Gene，Organism。多条件组合：每行一个条件，用 AND/OR/NOT，点击 Add field 增加条件。

(2) 结果页筛选与导出

左侧筛选面板（Filters）：

Reviewed：勾选 Yes 只看高质量 Swiss-Prot

Organism：快速选人、小鼠、酵母等

Sequence、Annotation、Taxonomy 等

排序与下载

排序：按名称、长度、日期等

Download 支持：FASTA（序列），TSV/Excel（表格：ID、基因、物种、功能等），XML/RDF（完整注释）

批量下载：支持全结果或选中条目。

(3) 搜索结果

Protein

左侧小图标：

黄色对勾 = Reviewed（Swiss-Prot，人工注释，高质量）

无图标 = Unreviewed（TrEMBL，自动注释）

第一行：Entry name（条目名），格式如 INS_HUMAN

第二行：Protein name（蛋白质名称）

Organism

来源物种，如 *Homo sapiens*（人）、*Mus musculus*（小鼠）

①. Status

Reviewed：已审（Swiss-Prot）

Unreviewed：未审（TrEMBL）

②. Gene

对应的基因名，如 INS、BRCA1

③. Length

氨基酸序列长度

(4) 条目详情页（Entry Page）

点开任意蛋白（如 P69905），页面分为多个模块：

①. Entry Information（条目基本信息）

Entry name：条目名，稳定但可更新

Primary accession number：主登录号（Accession），终身不变，最常用标识

Secondary accession：旧 / 合并后的登录号

Status：Reviewed / Unreviewed

Taxonomic ID：物种分类号

Created / Last modified：创建与最后更新日期

②. Names & Taxonomy（名称与物种）

Protein names：推荐名、别名、功能名

Gene names：基因符号、同义词

Organism：物种学名

Taxonomic lineage：界门纲目科属种

③. Function（功能注释）

生物学功能、催化活性、参与通路

带 Evidence code 证据代码（实验 / 预测）

④. Subcellular location（亚细胞定位）

蛋白在细胞中的位置：细胞膜、细胞核、线粒体、分泌等

可能标注："Secreted""Membrane""Nucleus"

⑤. Pathology & Biotech（疾病与生物技术）

与哪些疾病相关

突变导致的表型

药物靶点、应用等

⑥. Sequence（序列信息）

Length：氨基酸数目

Mass：分子量

Sequence：完整氨基酸序列

可直接复制 FASTA

⑦. Features（序列特征）

用位置标注序列上的功能区域：

Signal peptide 信号肽

Transmembrane 跨膜区

Domain 结构域

Site 活性位点、结合位点

Helix / Strand / Turn 二级结构

Glycosylation / Phosphorylation 翻译后修饰位点

Disulfide bond 二硫键

Variant 自然变异 / 突变

⑧. Cross-references（交叉引用）

链接到其他数据库：

PDB：蛋白质三维结构

Pfam / InterPro：结构域

GO：基因本体（功能 / 组分 / 过程）

KEGG / Reactome：代谢 / 信号通路

OMIM / ClinVar：疾病突变

Ensembl / RefSeq：基因 ID

STRING / IntAct：蛋白互作

⑨. Literature（文献）

支持注释的原始研究论文

PMID 号可直接跳转 PubMed

⑩. Family & Domains（家族与结构域）

属于哪个蛋白家族

保守结构域信息

5. 问题

Q1:怎么在数据库中获得一个物种的全基因组数据。