《Linux 生物信息基础》小组讨论总结报告

第 4 组，第 10 次讨论

组长：陈奕晗

执笔：朱瑾煜、邹济平、陈奕晗、高培翔

1 时间

2021 年 6 月 4 日，19:00 ~ 22:00

2 地点

泊星地咖啡厅

3 人员

陈奕晗、邹济平、朱瑾煜、高培翔

4 方式

线下讨论

5 主题

5.1 数据库网站介绍

5.2 本组课题进展

6 内容

6.1 数据库网站介绍

6.1.1 `uniprot` 最常用的数据库

`Uniprot` 是包含蛋白质序列，功能信息，研究论文索引的蛋白质数据库，整合了包括`EBI`，`SIB`，`PIR`三大数据库的资源。

`EBI`：欧洲生物信息学研究所（`EMBL-EBI`）是欧洲生命科学旗舰实验室`EMBL`的一部分。
`SIB`：瑞士日内瓦的SIB维护着`ExPASy`（专家蛋白质分析系统）服务器，这里包含有蛋白质组学工具和数据库的主要资源。
`PIR`：`PIR`由美国国家生物医学研究基金会（`NBRF`）成立，旨在协助研究人员识别和解释蛋白质序列

信息。

主要包括以下子库：

`UniProtKB/Swiss-Prot` 高质量、手动注释、非冗余的数据库。可筛选查询下载可靠的蛋白序列，提供blast查询和align多序列比对功能

`UniProtKB/TrEMBL` 自动翻译蛋白质序列，预测序列，是未经过验证的数据库

`UniParc` 一个非冗余蛋白质序列数据库

`UniRef` 聚类序列减少数据库，加快搜索的速度

`Proteomes` 为全测序基因组物种提供蛋白质组信息

6.1.2 `neXTProt` 人类蛋白质库

`neXTProt` 是一个以人类为中心的蛋白质数据库。

从蛋白质组学、`microarray`、抗体、`siRNAs`、`interactomics` 等多种高通量方法中导入 `neXtProt` 数据。

提供有关人类蛋白质的信息，如功能、疾病参与、`mRNA`/蛋白质表达、蛋白质/蛋白质相互作用、`PTM`、蛋白质变异及其表型效应。

6.1.3 `GenBank` DNA序列库

`GenBank` 是美国国家生物技术信息中心建立的DNA序列数据库，有丰富的核苷酸数据，部分数据来源于大规模基因组测序计划。

`Genbank` 包括了基因组DNA数据库、对应于表达基因的 `cDNA` 数据库、表达序列标签、`Unigene` 多部分。对某一个蛋白，从DNA或者RNA多个水平上进行描述。

6.1.4 `PDB` 蛋白质结构数据库

`PDB` 数据库由结构生物信息研究合作组织维护，数据主要来源于通过实验(X射线晶体衍射，核磁共振，电子显微镜方法等)测定的生物大分子三维结构（主要是蛋白质三维结构，还包括核酸、糖类、蛋白质与核酸复合物）

提供信息包括：蛋白质原子的空间坐标，形成α螺旋和β折叠的模式，双硫键连接模式，与蛋白质结合的配体，参与生化功能的残基。

6.2 本组课题

6.2.1 目标功能

本组期望建成一个mads-box转录因子的数据库网站，包括序列查询、blast比对以及下载功能。

共有7个板块，包括：

1）Home：

网站主页，展现网站基本信息，也提供序列搜索功能

2）Browse：

数据库浏览，可选择查看网站数据库中所有的mads-box转录因子序列及其基本信息

3） Blast：

提供mads-box转录因子的blast比对查询功能

4） Download：

可供用户下载mads-box转录因子序列

5）Help：

介绍本网站的功能

6）about

7） Link

6.2.2 进展

现已建好网站的基本框架，实现了大部分功能。

1) 通过uniprot下载了165条mads-box家族转录因子序列信息，上传到数据库中，可在Browse板块浏览或通过名称查询，包括序列描述，gene ontology，结构域等信息。

2) 已初步完成BLAST板块的工作，将用户输入序列和搜索条件转入到本地BLAST完成，并上传到网站呈现。

3) 正在编写help板块，介绍本网站的功能和操作。

6.2.3 待完善

1）Download和Link模块：

尚未开始Download和link模块的编写

2）页面美化

改善Browse板块数据信息的呈现效果，丰富home界面的内容

3) BLAST

已经初步完成BLAST部分的功能，能够比对查询输入序列并呈现。正在改善结果信息呈现效果，并解决BLAST运行过程可能出现的其他问题。

6.2.4 未来工作

若时间允许，可以添加其他家族的转录因子数据，并提供更完善的注释信息

7 问题与建议

无。