《Linux 生物信息基础》小组讨论总结报告

第 4 组，第 6 次讨论

组长：陈奕晗

执笔：陈奕晗、朱瑾煜、高培翔

1 时间

2021 年 5 月 3 日，19:00 ~ 21:30

2021 年 5 月 7 日，18:30 ~ 19:30

2 地点

泊星地咖啡厅

3 人员

陈奕晗、邹济平、朱瑾煜、高培翔

> **报告分工：**
>
> 陈奕晗：6.1.3
>
> 朱瑾煜：6.1.1、6.1.2、6.2.2、6.2.3、6.2.4
>
> 高培翔：6.1.4、6.2.1、6.3

4 方式

线下讨论

5 主题

5.1 隐马氏模型软件包 HMMER 复习

5.2 网站建设复习

5.3 关于本组课题

6 内容

6.1 隐马氏模型软件包 HMMER 复习

6.1.1 软件包安装

1. 下载软件包：`wget http://eddylab.org/software/hmmer/hmmer-3.3.2.tar.gz`

   关于 `wget` 命令：

   1) `wget` 后加网址，即从网址上下载文件；

   2) 使用 `^z` 可以暂停下载；

   3) 使用 `bg` 命令可以恢复下载。

2. 安装：

   1) 解压缩：`tar xf hmmer-3.3.2.tar.gz`

   2) 打开解压缩后的目录 (`cd`)，使用 `configure` 命令配置：`./configure --prefix=[ABSOLUTE PATH]`

   注意：所谓绝对路径 (absolute path)，即不能使用 ~ 表示路径，而应以 `/rd1/home/lebXX` 代替。

   3) 编译：`make`

   4) 安装程序和帮助文件：`make install`

6.1.2 配置环境变量

配置环境变量的作用是使得可执行文件可以直接通过命令被调用。

这里，为了使 hmmer 软件包的可执行文件能被调用，我们需要按以下方式配置环境变量，即在 `.profile` 文件加入以下内容。

```
if [ -d "$HOME/install/path/bin" ] ; then
    PATH="$HOME/install/path/bin:$PATH"
fi
```

> 解释：如果路径 `$HOME/install/path/bin` 存在，并且描述的是一个目录，则将这个目录加入 `PATH` 中。
>
> 编辑完成后，键入命令 `source` 使改动生效，此时便配置完成了。
>
> > `hmmbuild -h | less` 查看 `HMMER` 的版本

```
# hmmbuild :: profile HMM construction from multiple sequence alignments
# HMMER 3.3.2 (Nov 2020); http://hmmer.org/
# Copyright (C) 2020 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# input alignment file:             globins4.sto
# output HMM file:                  globins4.hmm
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

# idx name                  nseq  alen  mlen eff_nseq re/pos description
#---- -------------------- ----- ----- ----- -------- ------ -----------
1     globins4                 4   171   149     0.96  0.589

# CPU time: 0.10u 0.00s 00:00:00.10 Elapsed: 00:00:00.10
(END)
```

显示 `HMMER 3.3.2` 表示环境变量设置成功

6.1.3 初步使用：`hmmbuild`，`hmmsearch`，`hmmalign`

1. `hmmbuild`

   作用：　多重序列比对

   输入：多重序列比对的文件（`.sto`）

   输出：建立的这些多重序列比对的隐马尔可科夫模型

   范例：　`hmmbuild globins4.hmm globins4.sto`

```
# hmmbuild :: profile HMM construction from multiple sequence alignments
# HMMER 3.3.2 (Nov 2020); http://hmmer.org/
# Copyright (C) 2020 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Usage: hmmbuild [-options] <hmmfile_out> <msafile>
```

> `alen` (aligned length): 多重序列比对时对齐后的列长度
>
> `mlen` (model length): 隐马尔可科夫模型长度，`HMMER`将多重序列比对后，生成了`mlen`个一致位置的轮廓 (profile, 或模型 model)，剩余的`alen`-`mlen`个包含间隙的对齐列作为相对于模型的插入
>
> `eff_nseq` (effective number of sequence): 计数为有效的总序列数占比
>
> `re/pos` (relative entropy): 每个位置的相对熵

2. `hmmsearch`

   作用：　以已经建立好的模体在目标数据库中搜索序列

   输入：建立好的参考数据模型、要搜索的数据库

   输出：目标序列

   范例：　`hmmsearch globins4.hmm uniprot_sprot.fasta >`

```
--- full sequence ---   --- best 1 domain ---    -#dom-
 E-value  score  bias    E-value  score  bias    exp  N  Sequence             Description
 ------- ------ -----    ------- ------ -----   ---- --  --------             -----------
 4.9e-65  223.2   0.1    5.4e-65  223.0   0.1    1.0  1  sp|P02024|HBB_GORGO    Hemoglobin subunit beta OS=Gorilla gor
 6.8e-65  222.7   0.1    7.6e-65  222.6   0.1    1.0  1  sp|P68871|HBB_HUMAN    Hemoglobin subunit beta OS=Homo sapien
 6.8e-65  222.7   0.1    7.6e-65  222.6   0.1    1.0  1  sp|P68872|HBB_PANPA    Hemoglobin subunit beta OS=Pan paniscu
 6.8e-65  222.7   0.1    7.6e-65  222.6   0.1    1.0  1  sp|P68873|HBB_PANTR    Hemoglobin subunit beta OS=Pan troglod
 1.2e-64  222.0   0.1    1.3e-64  221.8   0.1    1.0  1  sp|P02025|HBB_HYLLA    Hemoglobin subunit beta OS=Hylobates l
 2.1e-64  221.2   0.2    2.3e-64  221.0   0.2    1.0  1  sp|P02033|HBB_PILBA    Hemoglobin subunit beta OS=Piliocolobu
```

> 第一列是整条序列的 E-value，score，bias（偏差值，校正调整），第二列表示该条序列中得分最好区域的相应情况，最后两列是每个目标的序列名称和一些描述。

3. `hmmalign`

   作用：　多重序列比对

   输入：输入是 `fasta` 文件和建立的模型文件

   输出：`sto.` 格式的多重序列比对文件

6.1.4 拓展：用 hmmbuild 读取其他格式的多序列比对文件（Clustal 工具的使用）

`hmmbuild` 可以接受多种格式的多序列比对文件，包括 stockholm，aligned FASTA，clustal，a2m，phylip 等

Clustal 是一类常用的多序列比对工具，包括经典工具 ClustalW 和新开发的 Clustal Omega。ClustalW 可以生成 clustal 格式的多序列比对文件，Clustal Omega 可以生成 stockholm 或者 clustal 格式的多序列比对文件，这些格式可被 `hmmbuild` 识别并用于 profile 构建。

6.2 网站建设复习

6.2.1 相关原理

当我们输入域名访问某个网站时，输入的字符串被发送给域名服务器 (Domain Name Server, DNS)，域名服务器解析这一字符串，将其映射到网站所在的服务器的 IP 地址，然后从网站服务器请求资源，最后发送回本地计算机并在浏览器中显示出网页。

通过改写 hosts 文件，可以绕过域名服务器，直接给出从某个特定的域名字符串到某个特定的网站服务器 IP 地址的映射。也就是说，hosts 文件指定的映射的优先级高于域名服务器。但这种更改只对当前计算机生效。

修改 hosts 文件可以实现屏蔽网站，暂时解决 DNS 污染，对开发中的网站进行虚拟调试等功能（本课程即使用此种功能）。

6.2.2 HTML 和 Markdown

1. **HTML**

超文本标记语言 (hypertext markup language, HTML) 是网页的标准格式。

可以使用 HTML 来建立自己的 Web 站点。HTML 运行在浏览器上，由浏览器来解析。

HTML 的内容较多，具体使用时会不时参考菜鸟教程。一些简单的语法，我们在 Markdown 中也会提到。

2. **Markdown**

Markdown 是一种轻量级标记语言，它允许人们使用易读易写的纯文本格式编写文档，然后转换成有效的 XHTML（或者HTML）文档。由于 Markdown 的轻量化、易读易写特性，它掌握起来比 HTML 更快、更容易；且 **Markdown 本身包含很多 HTML 的语法**，因此作为 HTML 的入门也是很合适的。

基于以上理由，本组考虑主要使用 Markdown 制作网页的大致框架，而以 HTML 添加一些必要的细节。

许多网站，如 GitHub，都支持直接呈现 Markdown 界面。例如在 GitHub 的 repository 中直接上传 index.md，是可以直接作为主页显示的，这里 GitHub 实际执行了将 index.md 自动转化为 HTML 格式的功能。大部分 Markdown 编辑器，如我们使用的 Typora（本报告就是用 Typora 写成的），都支持直接由 Markdown 导出 HTML。我们只要把这个导出的 HTML 根据需要稍加修改，然后就可以直接上传到服务器上去了。

6.2.3 Markdown 的常用语法

Typora 的可视化效果很好，然而一些常用的语法也是必须要会的。

1. **粗体** `**[TEXT]**` 、*斜体* `*[TEXT]*` 、<u>下划线</u> `<u>[TEXT]</u>` (同 HTML)。

   代码块（行内代码使用两个反引号 (`) 包围，行间代码使用上下两行的三个反引号包围）

   公式块（行内公式使用两个美元符号 ($) 包围，行间公式使用上下两行的两个美元符号包围。LaTeX 格式，例如 $ \frac{1}{2} $ 显示为 $\frac{1}{2}$）

2. **标题**前加 # 号，几级标题就加几个 #。

   区块引用前加 >。

   分割线可用三个 * 号实现。

3. **图片**的格式为 `![alt 属性文本](图片地址 "可选标题")`。

   表格可用以下格式实现：

   ```
   |  表头   | 表头  |
   |  ----  | ----  |
   | 单元格  | 单元格 |
   | 单元格  | 单元格 |
   ```

4. **链接**可以使用 `[链接文本](链接内容)` 实现。

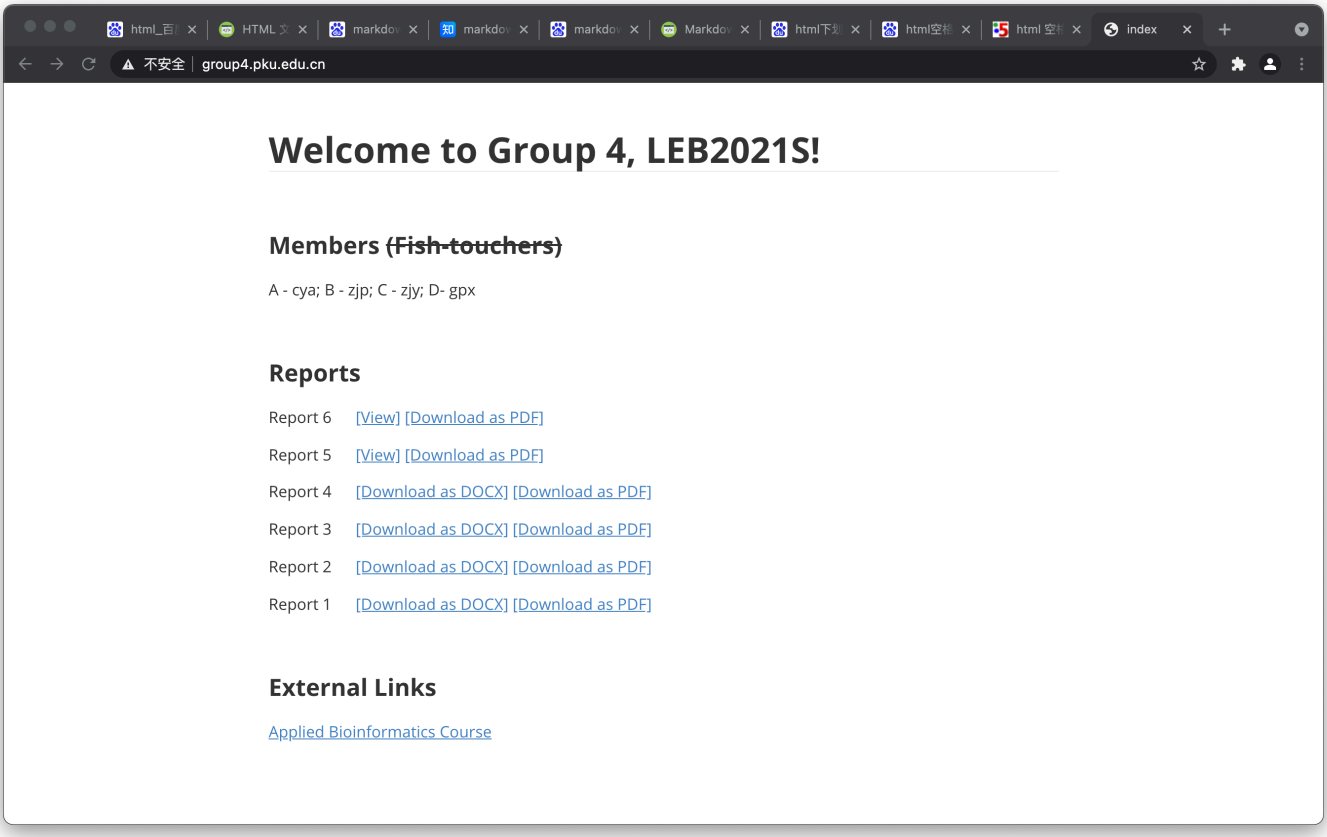
不在 Markdown 涵盖范围之内的标签，都可以直接在文档里面用 HTML 撰写。

目前支持的 HTML 元素有：`<kbd> <b> <i> <em> <sup> <sub> <br>` 等，如：

例如，可以用 `<br>` 实现换行，用 ` ` 实现空格等。

6.2.4 实例：Group4「门户」网站的建立

将 index.html (由 index.md 生成)、上次课的报告 report_5.html 及 report_5.pdf （包括其中的图片）通过 scp 命令或 FTP 上传至服务器的 `/rd1/www/group4`，在本地配置 hosts 文件后打开网页 http://group4.pku.edu.cn，可以看到带有 index.html 的界面，此即为主页，如下图。



其中 Reports 部分，因为有的文件还没上传，点开一些链接会显示 404 报错。但是 Report 5 的 View 链接（http://group4.pku.edu.cn/report_5.html）和 Download as PDF（http://group4.pku.edu.cn/report_5.pdf）是可以打开的。这里我们点击 View，就能看到上一次的报告内容。再点击 Return Home，就能回到主页。

7 小组课题

7.1 项目概述：一个小规模的植物转录因子数据库。

7.2 目标范围

1) 总体范围：MADS-BOX 家族转录因子。
2) 物种范围：视情况而定，上限为已测得基因组的全部植物，下限为藻类、苔藓、蕨类、裸子植物、植物的一些代表性物种。

7.3 收录内容

项目以基因作为收录条目，包括物种、基因序列、在染色体上的位置、转录本、蛋白质序列等基本信息，以及功能注释、结构、亚细胞定位等附加信息。附加信息的多少视情况而定。

7.4 项目功能

项目将在网页上提供一个交互页面，提供按条件检索，库内序列比对，系统发生树构建等功能。其他功能视情况而定。

7.5 项目流程

1) 数据收集。从已有的数据库收集我们所希望包含的信息。
2) 库的再构成。将收集到的数据以一定的格式重新构建为数据库。
3) 交互界面的构建。在网页上构建可以实现各种功能的前端页面，连接到已构建的库，提供各种服务。

8 收获

1. 基本确定了本组将要进行的课题；
2. 复习了上节课讲到的 HMMER 软件的安装和使用，并根据 Userguide 进行了一定的拓展；
3. 初步学习了使用 HTML 和 Markdown 制作网页的方法。