“Linux 生物信息基础”课程

小组集体练习、讨论、交流

总 结 报 告

组：4　　次：3　　组长：陈奕晗　　执笔：高培翔

1. 时间：2021 年 4 月 9 日，15:00 ~ 18:00

2. 地点：王克桢 311

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：

1) EMBOSS 软件包功能学习与复习

2) 本小组综合课题选题方向

6. 内容

6.1 EMBOSS 软件包功能学习与复习

6.1.1 运行模式

交互式：程序给定大部分默认参数，提示用户输入个别参数。

菜单式：利用参数`-options`，使用户在交互式的操作方式下可以决定所有可选参数。

参数式：在输入命令的同时指定所有所需参数，不进行交互。

6.1.2 帮助程序

`tfm`：详细的帮助文档，the file manual。

`wossname`：输入关键词查找相关的程序。

`seealso`：列出与给定程序相关的其他程序。

6.1.3 格式转换

`seqret`：将各种格式的核酸/序列文件转换为 fasta 格式。可被转换的文件格式包括 GenBank 和 UniProt 等。

6.1.4 序列提取

seqretsplit：将一个包含多段序列的 fasta 文件拆分为多个单序列 fasta 文件。

extractseq：根据用户指定的区间从 fasta 文件中提取特定的序列。

coderet：从 GenBank 等包含注释信息的文件中提取各种 feature 的序列，包括编码区、mRNA、翻译产物、非编码区序列。

extractfeat：从 GenBank 等包含注释信息的文件中提取特定的 feature，使用-type 参数指定所需 feature。由于 UniProt 格式有调整，目前使用 extractfeat 处理 UniProt/Swiss-prot 数据会出错。

6.1.5 序列变换

revseq：输出给定序列的反向互补序列。

msbar：对给定序列进行点突变，片段突变或密码子突变，包括替换、插入和删除。用户可决定突变类型与次数。

shuffleseq：将给定序列的碱基打乱重排。

6.1.6 序列显示

infoseq：显示序列名称、登录号、长度、GC 含量等信息，可用于 GenBank，Swiss-Prot 以及 fasta 格式。

showseq：以指定的方式显示序列，包括指定位置，指定读码框并翻译等。

showfeat：显示 GenBank，Swiss-Prot 等格式的文件的特征信息。

6.1.7 序列比对

needle：进行双序列的全局比对。使用 Needleman-Wunsch 算法对两个给定序列进行比对，使得全局上的得分最高。比对结果将显示两个序列的相似性。

stretcher：needle 的改进版本，仍使用 Needleman-Wunsch 算法，降低了运行所需的内存但增加了所需时长。

water：进行双序列的局部比对。比对结果用于寻找两个序列中相似的部分。

emma：进行多个序列之间的比对。比对结果通过树形分支显示输入序列之间的“亲缘远近”。

edialign：进行多个序列之间的比对。比对结果用于寻找多个序列之中共同存在的相似部分(保守的 motif)。

6.1.8 点阵图

用于直观地显示两个序列（更广义地情形下可扩展到两个任意的字符串）的相似片段的位置和数量信息。如果两个序列相同则可以显示自身多次重复的相似片段的位置和数量信息。

二维点阵图绘制所需要的输入是两个序列（更一般地，两个字符串），两个关键参数是滑动窗口大小（wordsize）和相似性阈值（threshold）。对于双色无灰度点阵图，黑色点代表相似性超过阈值的滑动窗口，黑色点连成的线表示高相似性的连续区间，从两个坐标轴上可以分别读出这个区间在两个输入序列上的位置。

当两个输入序列相同时，点阵图可用于寻找序列内的重复片段。由于输入了相同的片段，直线 *y=x* 必定在点阵图中出现，平行于该直线的线段则指示重复的部分，从横纵坐标轴上可读出该片段两次重复出现的位置。当某一片段连续重复多次时，将在点阵图上呈现近似正方形阴影的图像。

dottup：给定两个序列及滑动窗口长度，输出点阵图。多用于核酸序列。

dotmatcher：给定两个序列及滑动窗口长度及相似性阈值，输出点阵图。多用于氨基酸序列。

6.1.9 序列组分统计

compseq：统计一定长度字串的出现频率。用于发现特别高频和特别低频的字串。

freak：给定初始序列、滑动窗口大小与步长，以图形方式输出初始序列各位置的 GC 含量信息。

6.1.10 开放读码框分析

getorf：给定核酸序列，从中提取可能的开放阅读框及其对应的氨基酸序列。

sixpack：给定核酸序列，输出全部 6 种读码方式及其对应氨基酸序列。

showorf：给定核酸序列，指定一种或多种读码方式，输出翻译结果。

6.1.11 CpG 岛识别

cpgplot：给定核酸序列，以图形方式显示预测的 CpG 岛。

6.1.12 密码子分析

cusp：给定 CDS 序列，统计密码子使用频率。

chips：给定 CDS 序列，统计有效密码子数（ENC）。

6.1.13 重复序列寻找

palindrome：给定核酸序列，给定重复长度范围，给定重复序列允许间隔，给定允许错配数，寻找给定序列中的反向重复序列。用于寻找较短的回文重复。

einverted：给定核酸序列，给定罚分参数，寻找倒转重复。用于寻找较长的倒转重复。

6.1.14 蛋白质序列分析

pepstats：给定核酸序列，统计各氨基酸出现频率。

wordcount：给定核酸序列，统计特定字长的短肽的出现频率。（类比 compseq）

6.1.15 序列特征位点识别

antigenic：给定蛋白质氨基酸序列，寻找可能的抗原决定簇。

fuzzprot：给定若干蛋白质氨基酸序列，寻找共有的相似片段（motif）在各序列中的位置。

6.1.16 二级结构分析

tmap：给定蛋白质氨基酸序列，寻找可能的跨膜螺旋。

pepwheel：给定蛋白质氨基酸序列，寻找可能的 alpha 螺旋。

garnier：给定蛋白质氨基酸序列，预测整体的二级结构，包括 alpha 螺旋，beta 折叠，转角，无规则卷曲等。

6.2 本小组综合课题选题方向

我组倾向于选择网站建设。尚未确定明确的内容。

7. 问题：无

8. 建议：

1) 希望老师提供之前学生所做的综合项目范例。

2) 小组成员目前均没有可直接与本课程产生关联的实验室课题（尚未参与特定课题或课题偏向湿实验），难以确定综合课题的选题内容。希望老师明确本课程的“综合课题”项目对我们的要求，并告知我们期待我们做到何种程度，这将有助于我们选择内容合适，难度合适的综合课题。