“实用生物信息技术”课程小组讨论总结报告

组：G4　次：3　组长：高嘉禾　执笔：张俊鑫

1.时间：2026 年 4 月 22 日 18：00

2.方式：线上（腾讯会议，微信）

3.主题：深入了解序列比对与 BLAST，完成课外练习等。

4.内容：

A. 阅读《双序列比对基础和应用实例》复习序列比对基本概念并浏览序列比对常用网站;

B. 完成序列比对部分实例;

C. 阅读《 序列数据库搜索系统 BLAST 简介》复习 BLAST 基本概念并浏览常用的 BLAST 网站;

D. 复习 BLAST 课堂练习；

A.序列比对基本概念

序列相似性（Similarity）和序列同源性（Homology）

序列相似性（Similarity）：核酸序列的相似性高低，是指通过序列比对所得结果中相同核苷酸残基所占比例，通常用百分比表示。而蛋白质序列比对结果中，除了用相同氨基酸残基所占比例作为相似性指标外，也经常用相同氨基酸加上相似氨基酸作为相似性指标。不论是核酸序列还是蛋白质序列，序列相似性是指相同和相似残基所占全长序列的比例，比例越高，相似性越高。

序列同源性（Homology）：指所比较的两个序列是否具有共同的祖先序列，序列同源性只有是非之别

一般说来，同源序列特别是亲缘关系较近的序列，相似性通常较高; 反之，相似性较高的两条序列，很有可能具有共同祖先。也就是说，序列相似性的高低经常用来推断其是否同源。，没有高低之分

直系同源（Ortholog）和旁系同源（Paralog）

直系同源（Ortholog）：随着物种分化，共同祖先分化为不同物种，所形成的新物种通过遗传机制获得祖先物种基因组中的基因。产生的不同物种中的同一种蛋白称为直系同源蛋白，其编码基因则为直系同源基因。

旁系同源（Paralog）：蛋白基因在物种形成后，由基因复制产生了多个基因，编码多个蛋白。这两个由同一个基因复制产生的多个基因编码的蛋白则称为并系同源(有时也译作旁系同源)蛋白，其编码基因则称为并系同源基因。

动态规划算法（Dynamic Programing）和启发式算法（Heuristic Programing）

动态规划算法（Dynamic Programing）：在给定计分矩阵和空位罚分的条件下，通过插入适当空位，使比对结果的总分值最高，即找到最优解。核心思想，是把一个复杂问题分解为若干子问题，并通过寻找子问题的解，最终找到初始复杂问题的解。

启发式算法（Heuristic Programing）：BLAST 采用启发式算法。首先，将检测序列按一定字长(Word Size)拆分成种子(Seed)序列，并按给定计分矩阵和设定阈值，找到与种子序列相似性较高的近邻(Neighbor)序列；接着，逐个找到各近邻序列在数据库中匹配序列，并按分值增加原则向两边延伸，得到高分对(High Scoring Pair)，将所得主对角线方向距离较近的高分对连接起来，并用 Smith-Waterman 方法进行比对；最后，对搜索到的靶标序列进行统计检验，输出期望值(Expect Value)低于设定阈值的靶标序列，即搜索结果。

计分矩阵（Scoring Matrix）和空位罚分（Gap Penalty）

计分矩阵（Scoring Matrix）：是序列比对的基础，指比对过程中相同或不同核苷酸或氨基酸之间的匹配或错配分值，不同计分矩阵具有不同匹配分值和错配分值。用于核酸序列比对的计分矩阵通常有两种，即 DNAMatrix 和 DNAFull

空位罚分（Gap Penalty）：是指比对过程中在适当位置插入空位，使比对总分值更高、比对结果更好。空位罚分大小设置通常采用经验值，起始空位罚分较大，而延伸空位罚分较小。所谓起始空位，是指插入的第一个空位，而延伸空位则是当插入多个空位时，第二个空位开始的其它空位。此外，当两个长度差别较大的序列进行整体比对时，往往需要考虑是否对末端空位也进行罚分。

PAM (Point Accept Mutation) 计分矩阵和 BLOSUM (Block Substitution)计分矩阵

BLOSUM 系列计分矩阵于上世纪九十年代基于蛋白质序列模块数据库 BLOCKS 构建。BLOSUM62 计分矩阵主对角线的 20 个矩阵单元为相同氨基酸之间的分值，即匹配分值。除主对角线外的其它矩阵单元为不同氨基酸之间的替换分值，即错配分值。错配分值有正有负，范围在 3 到-4 之间，其中大部分为零或负值。一般说来，BLOSUM100、BLOSUM90 等用于相似性高的近缘物种之间的序列比对，而 BLOSUM30、BLOSUM35 等则用于相似性较低的远缘物种之间的序列比对

PAM 系列计分矩阵的英文全称为 Point Accepted Mutation，即位点可接受突变矩阵，于上世纪七十年代构建，包括 PAM10、PAM20、PAM30，一直到 PAM500，共五十个矩阵。与 BLOSUM62 类似，PAM250 的匹配分值均为正值，而错配分值绝大部分为零或负值。PAM10、PAM20 等数字较小的矩阵，适用于相似性较高的序列之间的比对，而 PAM250 及其以上的矩阵，适用于相似性较低的序列之间的比对。

全局比对（Global Alignment）和局部比对（Local Alignment）

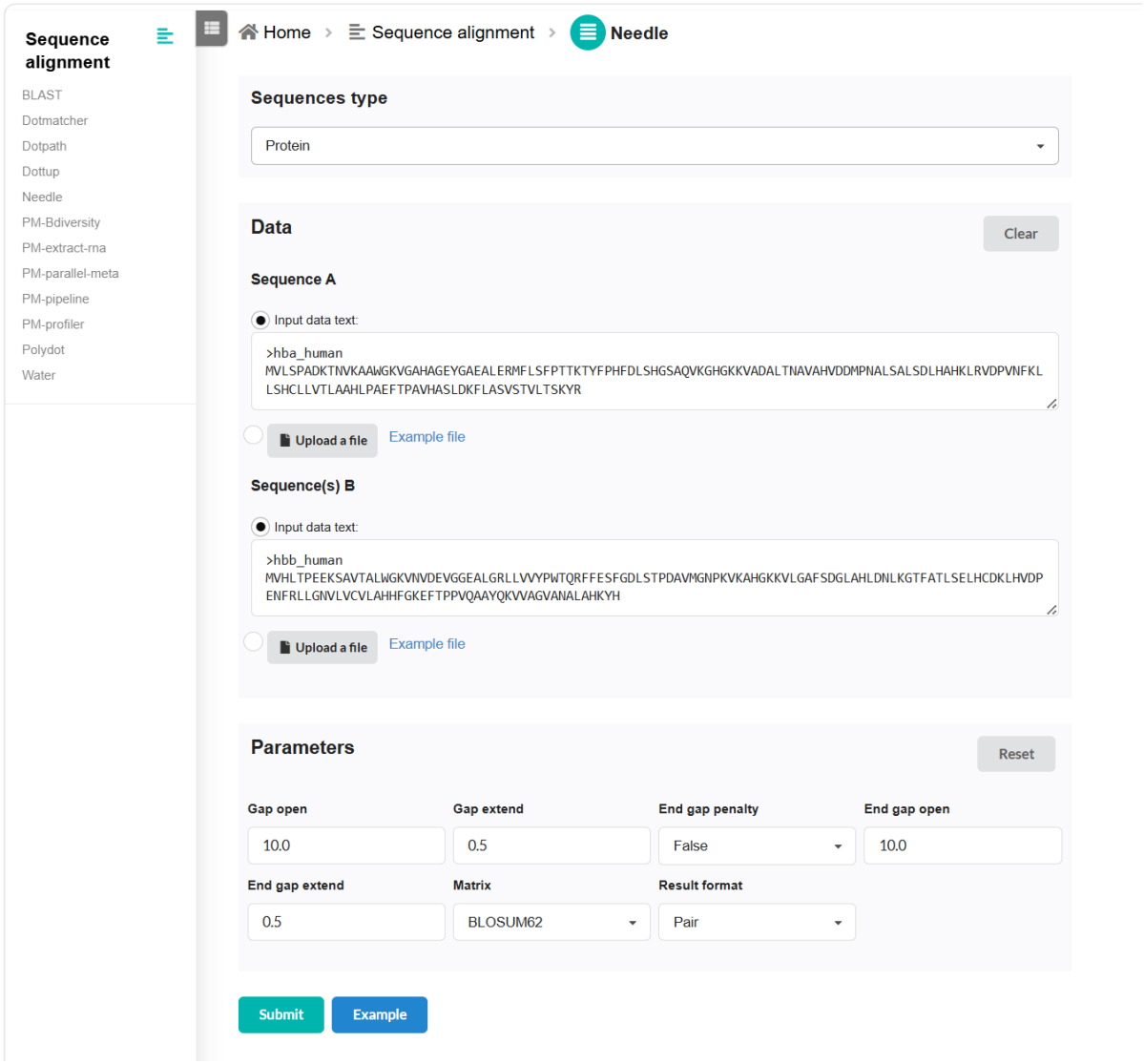
从全长序列出发，考虑所比对的两条序列的整体相似性，即整体比对(Global Alignment)，整体比对算法：Needleman-Wunsch 算法; 仅考虑所比对序列部分区域的相似性，即局部比对(Local Alignment)，局部比对算法：Smith-Waterman 算法。整体比对常用来考察两条序列是否在整体上具有较大相似性，并由此推测它们是否具有同源性。而局部比对则可以找出两个序列中的保守序列片段，如蛋白质序列中某个结构域或功能位点，基因上游启动子区域核酸序列调控元件等。

双序列比对（Pairwise Sequence Alignment）和多序列比对（Multiple Sequence Alignment）

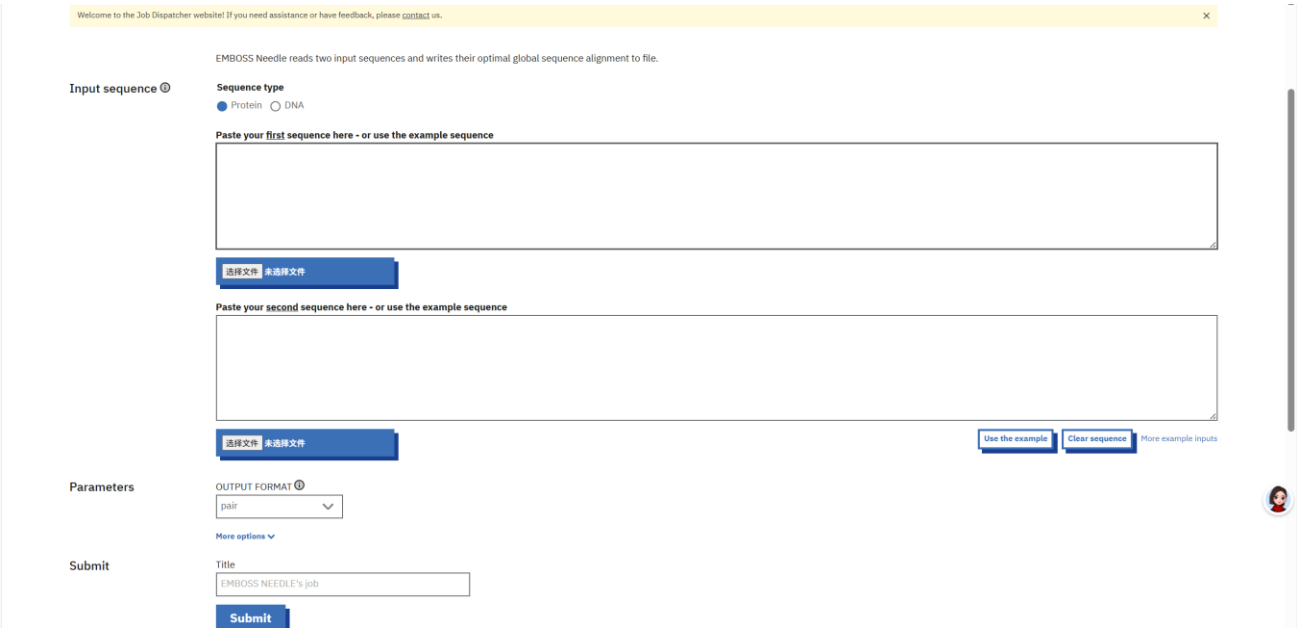
双序列比对（Pairwise Sequence Alignment）：对两条 DNA/RNA/ 蛋白质序列进行对齐，通过插入空位（gap）使相同或相似字符尽可能匹配，量化相似性并推断同源关系。

多序列比对（Multiple Sequence Alignment, MSA）：同时对齐三条及以上序列，逐列比较字符保守性，识别保守位点、推断进化关系与功能约束。

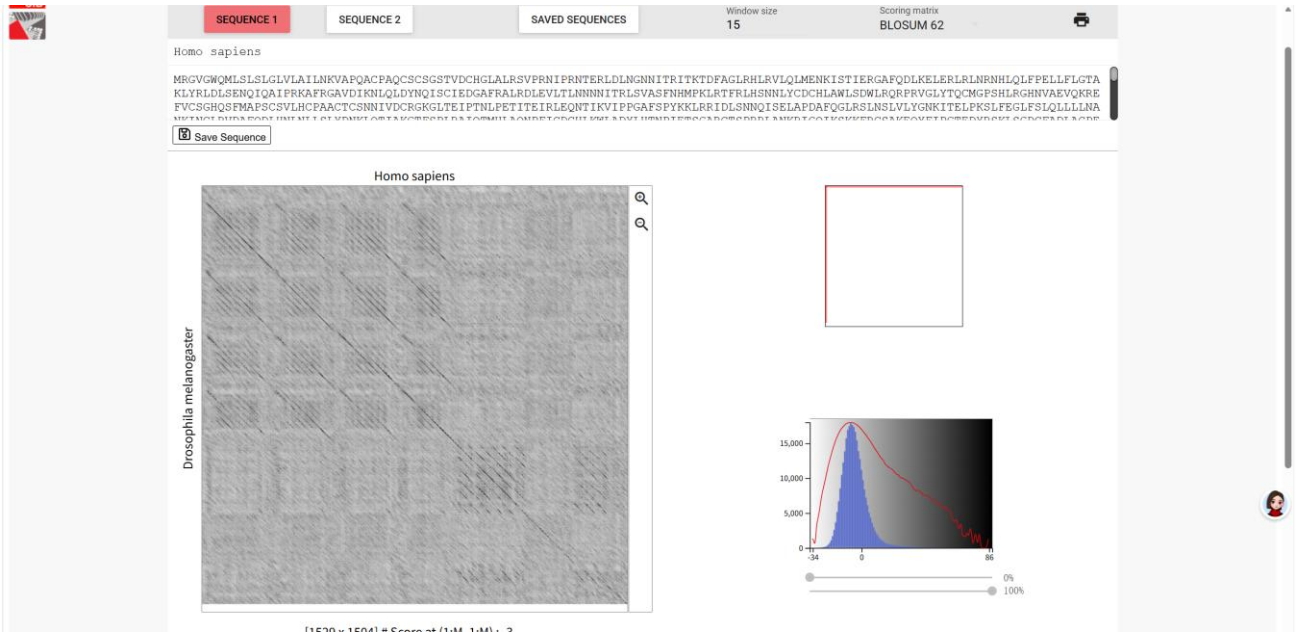
BIT NEEDLE



EBI EMBOSS NEEDLE



Dotlot



B.序列比对部分实例

人源血红蛋白 alpha 亚基 和 beta 亚基 全局比对

在 UniProt 数据库检索框中分别检索人源血红蛋白 alpha 亚基 HBA_HUMAN 与 beta 亚基 HBB_HUMAN，在 dowenload 页面中点击 Format 按钮，选择 FASTA 格式，将序列分别拷贝粘贴到 needle 程序的两个输入框中。

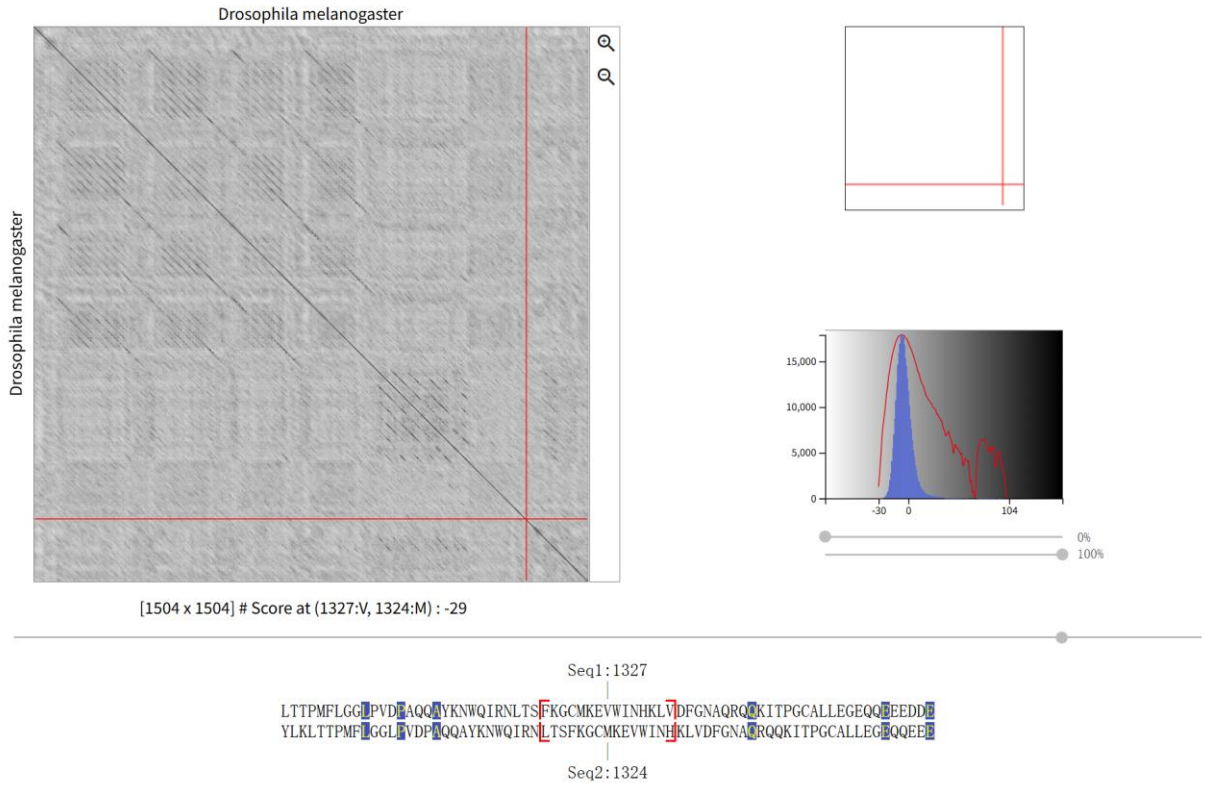
选择默认计分矩阵BLOSUM62、默认起始空位(GAP OPEN)罚分10和延伸空位(GAP EXTEND)罚分0.5，序列末端空位(END GAP PENALTY)不予罚分(No)。点击 Submit 按钮，几秒钟后，输出运行结果。

Open PAIR

```
########################################
# Program: needle
# Rundate: Thu  7 May 2026 01:23:08
# Commandline: needle
#    -asequence /blastdisk2/blastdata/toolkitv3/task/04L8JUJE3CV/sequence_a.faa
#    -bsequence /blastdisk2/blastdata/toolkitv3/task/04L8JUJE3CV/sequence_b.faa
#    -gapopen 10.0
#    -gapextend 0.5
#    -endweight false
#    -endopen 10.0
#    -endextend 0.5
#    -datafile EBLOSUM62
#    -aformat pair
#    -outfile output_protein.needle
# Align_format: pair
# Report_file: output_protein.needle
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: HBA_HUMAN
# 2: HBB_HUMAN
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 149
# Identity:      65/149 (43.6%)
# Similarity:    90/149 (60.4%)
# Gaps:           9/149 ( 6.0%)
# Score: 292.5
#
#
#=======================================

HBA_HUMAN          1 MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-D     48
                     || |:|.:|:.|.|.||||  :..|.|.|||.|:.:.:|.|:.:|..| |
HBB_HUMAN          1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD     48

HBA_HUMAN         49 LS-----HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLR     93
                     ||     .|:.:||.|||||..|.::.:||:|::....:.||:||..||.
HBB_HUMAN         49 LSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLH     98

HBA_HUMAN         94 VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR    142
                     |||.||:||.:.|:..||.|...||||.|.|:..|.:|.|:..|..||.
HBB_HUMAN         99 VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH    147
```

果蝇体节发育相关基因编码蛋白自身比对展示重复序列 SLIT_DROME 和 SLIT_DROME Dotlet 点阵图

分别将序列 **SLIT_DROME 和 SLIT_DROME** 拷贝粘贴到 Dotlet 程序的两个 sequence 输入框中。



C.BLAST 基本概念

数据库搜索和数据库检索

数据库检索，实质上是文本检索，即通过关键词匹配，从某个数据库中找出需要的条目。例如，输入作者姓名、期刊名称、文章标题等特定关键词，从美国国家生物信息技术中心（National center for biotechnology information，NCBI）生物医学文献摘要数据库 PubMed 中检索相关文献，就是数据库检索的典型应用。

与基于关键词的数据库检索不同，数据库相似性搜索所输入的不是文本信息，而是蛋白质或核酸一级结构序列信息；搜索对象也不是文本信息，而是数据库中的核酸或蛋白质序列信息。

序列比对和数据库搜索

数据库搜索的基础是序列比对，即序列相似性比对。序列比对通常分为两种，一种是两个序列之间的比对，找出它们的相同位点或相同区域，即双序列比对（Pairwise sequence alignment）；另一种是多个序列同时进行比对（Multiple sequence alignment），找出它们之间的保守位点或保守区域。数据库相似性搜索则是从数据库中找出与查询序列具有一定相似性的目标序列，逐个进行比对。因此，数据库相似性搜索的基础是双序列比对。数据库搜索方法和软件有多种，BLAST 是较为常用的一种。可以理解为“基于局部比对的序列数据库搜索工具”。具体说来，BLAST 是蛋白质和核酸序列数据库搜索程序，即用一个或多个蛋白质或核酸序列为查询序列，搜索蛋白质或核酸序列数据库，找出与查询序列具有较高相似性的目标序列，也称匹配序列和命中序列。

灵敏度和特异度

判断数据库相似性搜索结果优劣的指标通常有两个，即灵敏度和特异度。所谓灵敏度，是指从数据库中搜索到目标序列的多少；所谓特异度，是指搜索到的目标序列与查询序列是否具有较高相似性。理想状况当然希望搜索结果既有较高的灵敏度，也有较好的特异度。

BLAST 数据库搜索基本策略

BLAST 数据库搜索采用启发式算法以提高搜索速度，该算法的基本策略是找出序列中的保守片段并向两侧延伸。

分割种子串

首先按给定字长（Word size），将查询序列分割成一定长度的字符串，通常称种子串（Seed string）。如字长为 3，则称三字串。

确定近邻串

分割查询序列得到种子串后，下一步则需要确定与种子序列相似性较高的其它三字串，称近邻串。确定近邻串，基于计分矩阵计算匹配分值，选定计分矩阵，确定计分阈值，就可找出查询序列中所有种子串及其近邻串

搜索高分对

获得种子串和近邻串后，可构建一个搜索列表，逐个搜索数据库中每个序列，确定它们在每个序列中的位置。

延伸高分对

通过上述分割种子串、确定近邻串和定位高分对（High scoring pair，HSP），逐个扫描数据库中每一个序列，找到与查询序列具有高分匹配的高分对及其在序列中的位置。接下来则是将找到的高分对向两侧延伸，以增加高分对的长度。延伸的方法分为无空位延伸和有空位延伸两种。

计算期望值

通过统计检验的方法，给出比对结果的可信度；换言之，对于某个查询序列，数据库中搜索到的每个目标序列，都有一个期望值 E。 期望值 E 可以用公式 $E = kmN / e^{\lambda S}$ 计算。这里，m 为查询序列长度，N 为数据库中所有序列长度总和，e 为自然对数底 2.718，S 为归一化后的比对分值，K 和 λ 为常数，其大小与计分矩阵有关。当期望值 E 小于 0.05 时，搜索结果具有显著可靠性；E 值越小，显著性也越高。

BLAST 主要参数

字长

第一步分割种子串中种子串的大小与搜索速度和灵敏度有关。种子串长度越小，灵敏度越高，搜索速度越慢。字长即种子串的长度

计分矩阵

NCBI 基于浏览器的 BLAST 系统用于蛋白质序列比对的计分矩阵包括 PAM 系列和 BLOSUM 系列，默认矩阵为 BLOSUM62，可选矩阵为 BLOSUM90/80/50/45 和 PAM30/70/250。若需要提高搜索灵敏度，可选择 BLOSUM45 或 PAM250；反之，若需要提高搜索的特异度，可选择 BLOSUM90 或 PAM30。

空位罚分

空位罚分是指序列比对过程中为达到最佳匹配，插入一个或连续几个空位。空位的生物学意义可理解为，来自同一祖先的两条同源序列在演化过程中，其中一条在某些部位有单个或多个碱基的插入或删除。

可信度值

BLAST 采用期望作为评判搜索结果可靠性指标。E 值由公式 E = m×n×P 计算得到，其中，m 表示数据库中所有残基总数，n 表示查询序列残基数，P 表示随机匹配概率。E 值越低，说明随机匹配的可能性越小，所得结果越具统计显著性。

低复杂度区域屏蔽

所谓低复杂度序列，使指核酸或蛋白质序列中的重复区域，如基因组序列中的 Alu 序列、mRNA 序列中的多聚腺苷酸、蛋白质序列中简单重复序列等，为避免低复杂度重复序列对搜索结果的影响，可以对查询序列中的低复杂度区域加以过滤或屏蔽。

BLAST 程序

BLASTP

蛋白质序列搜索经常用于寻找同源序列，以推断查询序列的功能和演化。BLASTP 也经常用于搜索查询序列中的保守结构域、重复序列片段、序列模体（Motif）在数据库序列中的相似序列。

BLASTN

BLASTN 用于核酸序列搜索，即用核酸序列作为查询序列，搜索核酸序列数据库；搜索步骤和 BLASTP 基本相同，只是种子串长度、计分矩阵和空位罚分等参数不同。若查询序列为蛋白质编码序列，且已经知道编码区和起始密码子，则可将其翻译成蛋白质序列，再用 BLASTP 搜索蛋白质序列数据库。由于蛋白质序列由二十种不同氨基酸组成，而核酸序列由四种不同核苷酸组成，一般情况下，BLASTP 搜索结果的特异度较高，假阳性率较低。

BLASTX

BLASTX 用于以核酸序列为查询序列，搜索蛋白质序列数据库。查询序列为蛋白质编码序列，如转录组测序（RNA-Seq）或表达序列标签（Expressed sequence tag，EST）测序所得结果。

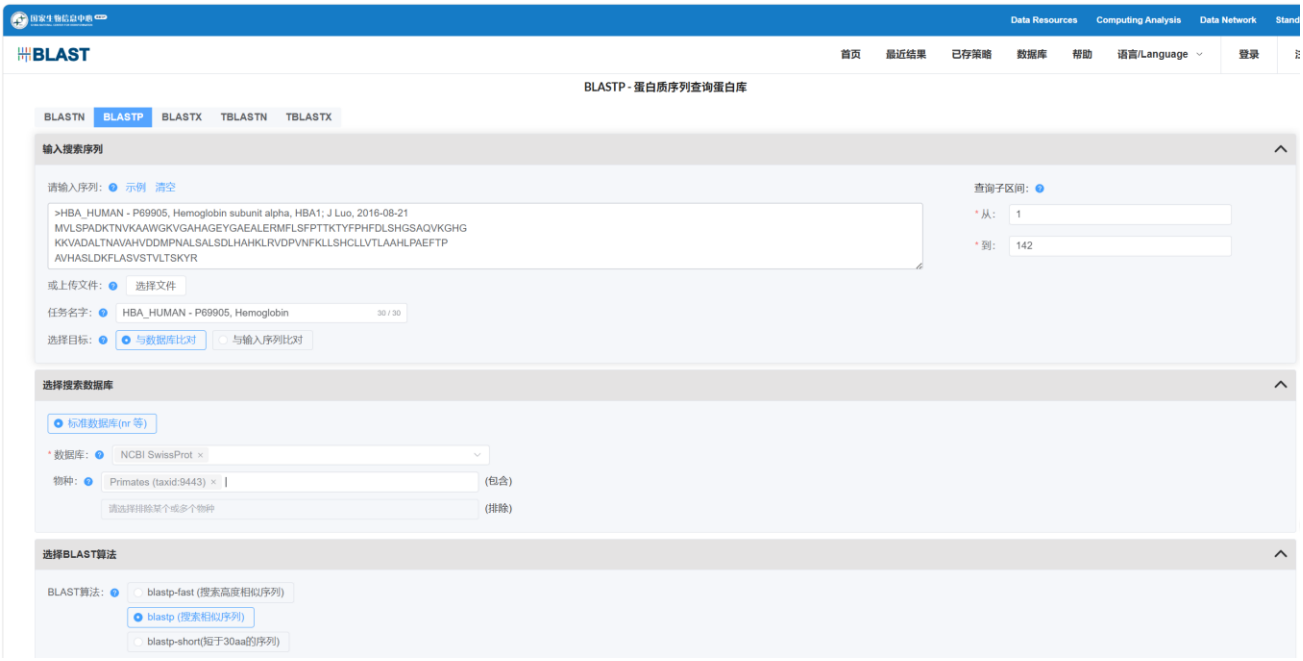
TBLASTN

与 BLASTX 相反，TBLASTN 以蛋白质为查询序列，搜索核酸序列数据库，如 NCBI 核酸参考序列数据库（Reference RNA，RefSeq_RNA）。也就是说，TBLASTN 以蛋白质序列为查询序列，搜索核酸序列数据库中的核酸序列翻译所得的蛋白质序列，实际上还是蛋白质序列数据库搜索。

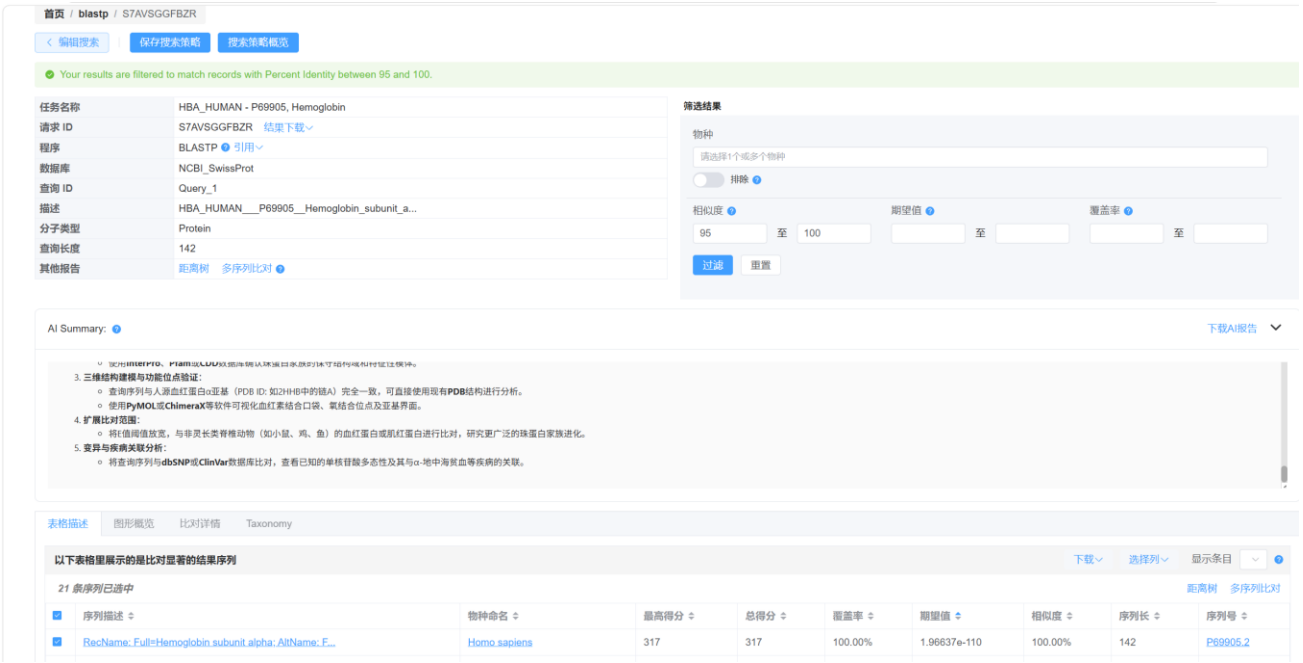
D.复习 BLAST 课堂练习

灵长目 alpha 血红蛋白搜索

以人血红蛋白 alpha 亚基 HBA_HUMAN 为搜索序列，采用默认参数利用中国国家生物信息中心 CNCB 蛋白质数据库搜索程序 BlastP，搜索 SwissProt 数据库，找出灵长目动物 Primates（分类学数据库登录号 taxid 9443）中 alpha 血红蛋白。

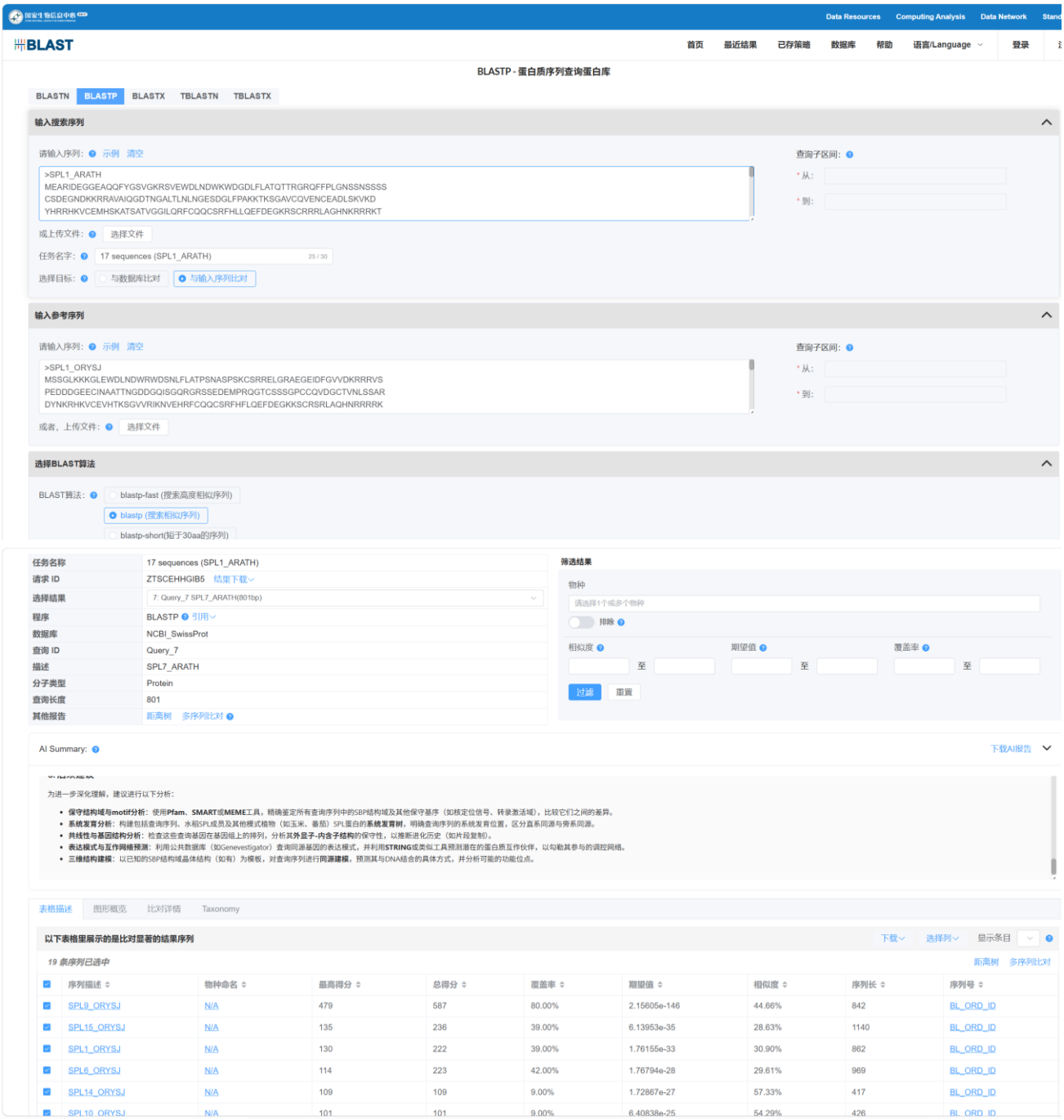


利用筛选结果功能，找出与 HBA_HUMAN 相同位点高于 95%的物种



SBP 转录因子家族搜索

以 17 个拟南芥 SPL7 转录因子 17 AT_SPL 和 19 个粳稻转录因子 19 OJ_SPL 进行 BLAST 比对，找出水稻中与拟南芥中 SPL7_ARATH 最可能的直系同源蛋白。



个人总结

G4C-张俊鑫：

这段时间学作物种质资源学，重点琢磨了序列比对和 BLAST 这两块内容，内容上不算太难，但确实得多上手练才行。刚开始接触的时候，对序列比对一点概念都没有，看着那些 DNA、RNA 序列就头疼，不知道怎么下手。后来经过老师系统性的教学与联系才慢慢搞明白，序列比对其实就是把不同作物的 DNA、RNA 序列摆在一起，找出它们相似的地方，这样就能推断这些序列的功能，还有作物之间的进化关系，这对研究作物种质的多样性特别有帮助。BLAST 工具是重点学习内容，根据需求选合适的比对类型以及调整参数尤为重要，比如比对核苷酸序列就用

blastn。通过这段时间的学习，我也清楚了这两个工具在作物种质资源研究里的用处，比如发掘新基因、鉴定种质资源的遗传差异。现在我操作还不够熟练，对一些算法原理也没摸透，但基本能用来解决简单的问题，在后续正式接触课题后我会结合自己的实验数据多做实操练习，争取把这部分知识融会贯通。

G4B-赵雪倩

序列比对是生物信息学的核心方法之一，在两个或多个核酸或蛋白质序列之间寻找相似区域，从而推断进化关系、预测结构和功能。在抗菌肽研究中，序列比对发挥着不可替代的作用，贯穿于新抗菌肽的发现、结构特征分析、功能位点识别等多个环节。

首先，抗菌肽中有一大类富含半胱氨酸的成员，其生物学活性高度依赖于特定的二硫键配对模式。通过序列比对，可以将新发现的抗菌肽与已知结构的抗菌肽或多肽毒素进行比对，快速判断其是否含有保守的半胱氨酸骨架。同时，通过优化比对参数提高结果准确性，抗菌肽序列往往较短（通常为 20～50 个氨基酸），且富含半胱氨酸，常规的默认参数（如 BLOSUM62 矩阵、起始空位罚分 10）并不总是适用，我们可以调整计分矩阵和空位罚分使研究序列之间的对比更加精确。

其次，也可以预测结构与功能的关系；序列比对的核心逻辑是“序列决定结构、结构决定功能”。通过将未知抗菌肽与已知功能的参考序列进行比对，可以推测其可能的三维结构和生物学功能。如果比对结果显示两者在关键区域具有高度相似性（如带正电的氨基酸分布、疏水区域的保守性），则可以初步推断该抗菌肽可能采用类似的膜作用机制，如形成 α 螺旋或 β 折叠结构，插入细菌细胞膜导致膜通透性改变。在实际研究中，我们可以利用局部比对工具快速定位抗菌肽与已知功能蛋白之间的保守结构域，从而聚焦于最有可能决定活性的核心区域，避免对全长序列的无效分析。

同时，分析序列差异对功能的影响序列比对不仅可以找出保守区域，还能揭示差异位点，这些差异往往与功能的分化或物种特异性相关。在抗菌肽研究中，通过比对同一家族不同成员之间的序列，可以识别出某些位点的替换、插入或缺失，进而推测这些变化对抗菌谱、抗菌活性和稳定性的影响。结合定点突变实验，可以验证关键残基的功能，为抗菌肽的理性改造提供依据。同时通过整体比对和点阵图分析，可以识别家族内部的重复结构域和保守基序，推断基因家族的演化历史。例如，点阵图能够直观显示某一抗菌肽序列内部是否存在串联重复单元，这对于分析抗菌肽的起源和功能多样化具有重要意义。

综上所述，序列比对在抗菌肽研究中发挥着从序列到结构、从结构到功能的多层次作用。它不仅帮助研究者识别保守的半胱氨酸骨架和二硫键模式，预测空间结构和作用机制，还能揭示功能相关的关键残基，指导抗菌肽的理性设计与改造。掌握整体比对与局部比对的方法，灵活选择计分矩阵和空位罚分参数，是开展抗菌肽生物信息学分析的基本功。

G4A-高嘉禾

通过对序列比对和 BLAST 的系统学习，我掌握了点阵图可视化、双序列比对、多序列比对以及数据库搜索的完整流程，以下是我的个人总结：

序列比对按序列数量可分为双序列比对和多序列比对。双序列比对是将两条序列中的残基进行配对，以揭示它们的相似性或同源性。根据比对范围的不同，双序列比对分为全局比对和局部比对。全局比对将两条序列完整对齐，适合比较整体结构相似度较高的序列；局部比对只提取序列中相似度最高的局部区域，适合检测共同的保守序列。上课时学习了使用 EMBOSS 软件包，通过 Needle 程序执行全局比对，Water 程序执行局部比对。

通过点阵图可以直观感受序列的相似分布。在瑞士生物信息学研究所点阵图网站 Dotlet 可以绘制点阵图，另外在 Jemboss 中也可以使用 Dottup 和 Dotmatch 两种方式绘制点阵图。根据序列类型选择对应的程序，Dottup 适用于核苷酸序列，Dotmatch 适用于氨基酸序列。对于点阵图，有连续的对角线说明两条序列在该区域高度相似，如果对角线出现断裂或平移说明存在插入或缺失，另外单独出现的短线表示存在局部相似的短片段。

对于序列比对的结果，以双序列比对为例，结果会显示两行序列，中间用符号表示匹配程度。“|”表示该位点残基完全相同；“:”表示高度保守的替换（通常是有相同化学性质的氨基酸或碱基）；“.”表示较保守的替换；空格表示不匹配。并且附带有统计结果：一致性（identity）是指完全相同残基数占总对齐长度的百分比；相似性（similarity）则包括了相同和保守替换的比例。并且明确了同源性和相似性的不同，相似性是可以量化的比例值；同源性是定性概念，只能判断“是”或“否”，表示序列是否源自共同的祖先。相似性高的通常是同源的，但是两者不能直接等同。

还可以使用 NCBI 的在线 BLAST 功能，包括 BLASTP（蛋白质序列比对蛋白质数据库）、BLASTN（核酸序列比对核酸数据库）、BLASTX（核酸序列翻译成蛋白质后比对蛋白质数据库）、TBLASTN（蛋白质序列比对翻译后的核酸数据库）和 TBLASTX（核酸序列比对核酸数据库，双向翻译）。其中，TBLASTN 适用于在基因组 DNA 中寻找已知蛋白的同源序列，TBLASTX 适合比较两个核酸序列可能存在的编码区相似性。另外，CNCB 中也有在线的 BLAST 功能，并且访问更稳定快捷。在处理大批量数据时还可以使用 Linux 命令行的 NCBI BLAST+，运行速度更快。