“实用生物信息技术”课程小组讨论总结报告

组：G1　　次：R3　　组长：薛林蕾　　执笔：吴鸿珍，江浩燊，薛林蕾

1. 时间

2026.5.8

2. 方式

线上讨论

3. 主题

序列比对及 BLAST 程序复习总结

4. 内容

1.序列比对的基本概念

1. 打分系统

得分矩阵：

DNA 得分矩阵：DNA 只有 4 种碱基（A, T, C, G），结构相对简单。

等价矩阵：匹配给正分，不匹配（错配）给负分。转换/颠换矩阵： 生物学上，相同类型的碱基替换（如 A→G，嘌呤变嘌呤）比不同类型的替换（如 A→C，嘌呤变嘧啶）更频繁。因此，转换的扣分通常比颠换少。

蛋白质得分矩阵：蛋白质由 20 种氨基酸组成。因为氨基酸具有不同的理化性质（大小、电荷、亲水性），某些替换在进化中很容易被接受（如异亮氨酸替换为缬氨酸，两者都是小分子疏水性），而某些替换则会导致蛋白质功能丧失。

PAM 矩阵：基于进化模型。它计算在一定进化时间内，一个氨基酸被另一个氨基酸替换的概率。PAM 后的数字代表进化距离。PAM1： 代表 1% 的氨基酸发生了突变。PAM250： 适用于比对亲缘关系较远（序列一致性低）的序列。

BLOSUM 矩阵：基于实验观察。通过对保守蛋白质家族数据库（Blocks 数据库）进行直接统计得出。BLOSUM 后的数字代表序列一致性的阈值。BLOSUM62 是最常用的通用矩阵。它是基于一致性在 62% 以上的序列簇统计得出的。 BLOSUM 数字越小，越适合比对差异大的序列（与 PAM 相反）。

空位罚分：生物进化过程中，DNA 序列不仅会发生碱基替换，还会发生片段的插入 (Insertion) 或 缺失 (Deletion)（合称 Indel）。为了让两个序列对齐，我们必须在其中一个序列中插入“空位”（通常用横杠 - 表示）。空位开启罚分为开启第一个空位时的惩罚，分值通常很高（比如 -10）。空位延伸罚分为在已有空位后面继续增加空位的惩罚，分值通常较低。在进行全局比对时，两个序列长度可能不一。如果序列 A 比序列 B 长很多，那么 B 的开头和结尾必然会出现大量空位。很多算法（如在拼接测序片段时）会选择不惩罚末端空位（Terminal Gaps），因为这通常是由于实验测序长度不同造成的，而非进化上的变异。

2. 全局比对和局部比对

全局比对：全局比对尝试将两个序列从第一个残基对齐到最后一个残基。它强制覆盖序列的全长，即使序列的某些部分差异很大。所用算法为 Needleman-Wunsch 这是 1970 年提出的经典动态规划算法。它计算从矩阵左上角到右下角的最优路径。 每个单元格的值取决于左、上、对角线三个方向的得分加上当前位置的

匹配/空位分值。适用于同源性极高，长度接近序列比较和进化分析。

局部比对 ：局部比对不强求全长对齐，而是在两个长序列中寻找相似度最高的片段（子序列）。它会忽略掉那些匹配度低的前缀和后缀。所用算法为 Smith-Waterman 这是 1981 年对全局算法的改进。引入了“0”分底线。如果计算出的得分变成负数，则直接记为 0。这意味着如果这一段比得太烂，算法就“翻篇”，从下一个位置重新开始寻找。每个单元格取（左、上、对角线得分，以及 0）之中的最大值。从矩阵中得分最高的单元格开始回溯，直到遇到得分为 0 的单元格为止。适用场景为功能结构域 (Domain) 搜索，长度差异巨大序列和数据库搜索

2.BLAST 基本概念

1.核心原理：

BLAST 的计算过程主要分为三个阶段：算法首先将查询序列切分为长度为 W（核酸通常为 11-28，蛋白为 3）的短字符串，称为“单词（Words）”。随后，算法不仅寻找数据库中完全匹配的单词，还会根据得分矩阵寻找那些虽然不完全相同但得分超过阈值 T 的“邻近词”。算法快速扫描数据库，定位这些单词出现的位置。一旦在数据库中发现匹配点（种子），算法便尝试向两个方向延伸。在延伸过程中计算累计得分，如果得分开始下降且跌幅超过预设的阈值 X（X-dropoff），延伸则停止。最终保留下来的高分片段被称为 HSP

2.常用的 BLAST 程序变体

根据查询序列和数据库类型的不同，BLAST 衍生出了多种针对性程序：

blastn： 核酸序列比对核酸数据库。常用于物种鉴定或寻找非编码区的同源性。

blastp： 蛋白质序列比对蛋白质数据库。这是寻找同源基因功能最准确的方法，因为氨基酸序列比核酸序列更保守。

blastx： 将核酸序列按 6 个框架翻译成蛋白后再比对蛋白库。常用于分析没有预测出基因的基因组片段或 EST 序列。

tblastn： 用蛋白质序列去比对翻译后的核酸数据库。常用于在全基因组中寻找潜伏的同源基因。

tblastx： 核酸对核酸，但双方都先翻译成蛋白。计算量极大，用于探测亲缘关系极远的核酸序列间的相似性。

3.评价参数

得分 (Bit Score)： 经过归一化处理的分数。分值越高，代表序列间的相似性越显著，且该分数不随数据库大小改变，具有横向可比性。

期望值 (E-value)：代表在随机情况下，从当前规模的数据库中搜索到同样高分的序列的期望次数。E < 10^-5 时通常认为具有极高的显著性，存在同源关系。

4.检索策略与优化

在实际应用中，灵活调整参数和策略能显著提高搜索效率和准确度，可以对以下几个参数或者程序进行调整：

单词长度 (Word Size)： 增加 W 值可以大幅提高速度，但会降低灵敏度；对于高度发散的序列，减小 W 值能更容易发现微弱信号。

低复杂度过滤 (Low-complexity Filter)： 很多序列包含重复的简单片段（如 AAAAA），这些片段会导致大

量无意义的匹配。开启过滤可以减少这类假阳性。

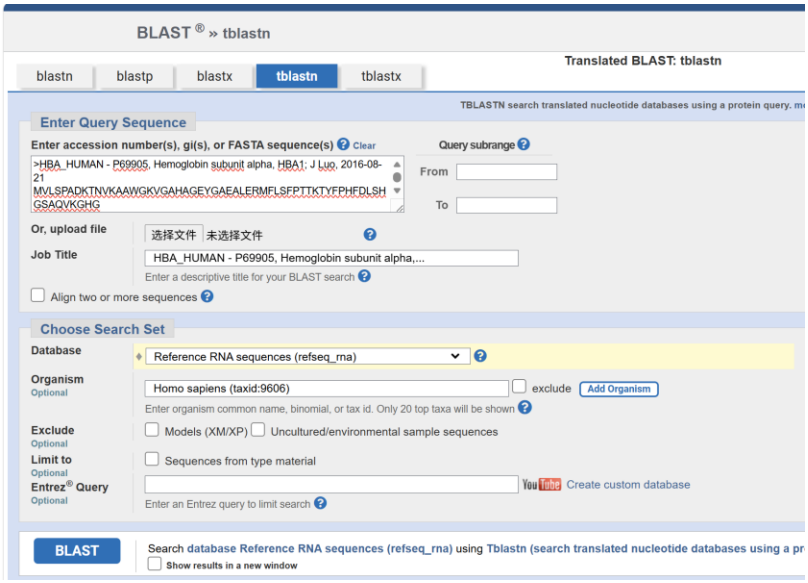
PSI-BLAST 策略： 如果普通搜索找不到同源蛋白，可以使用 PSI-BLAST。它通过多次迭代建立位置特异性矩阵（PSSM），能灵敏地捕捉到序列差异巨大但功能模体（Motif）一致的“远方亲戚”。

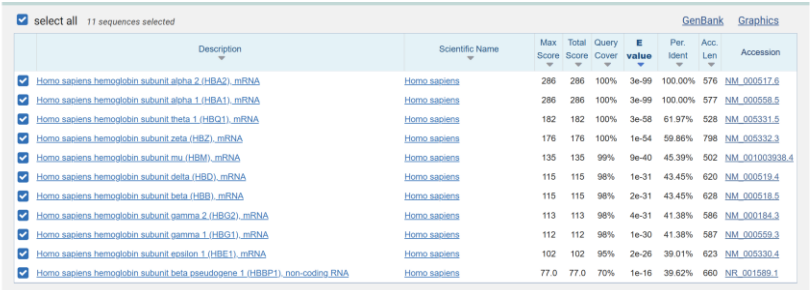
Megablast 策略： 当比对非常相似的核酸序列（如近缘物种的基因组比对）时，Megablast 使用超大的单词长度，能以极快的速度完成任务。

3. BLAST 练习实例

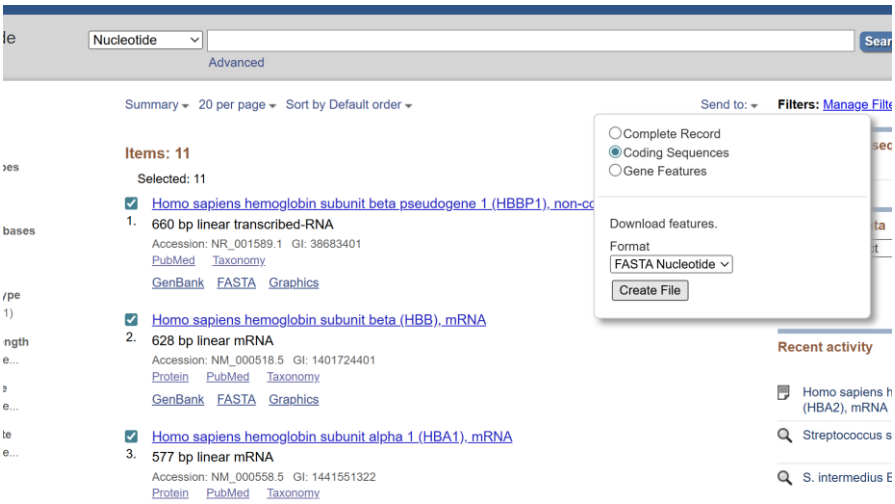
a 以人血红蛋白 alpha 亚基（HBA_HUMAN）为检测序列，用 tBlastN 搜索 RefSeq 数据库中人珠蛋白家族 mRNA 序列，提取其编码区序列， 进行多序列比对，分析结果。

首先进入 NCBI 的 tBlastn。输入 HBA_HUMAN 序列，选择 RefSeq RNA 数据库，Organism 选择 Homo sapiens (taxid:9606)，点击 blast，结果得到 11 个人珠蛋白家族 mRNA。





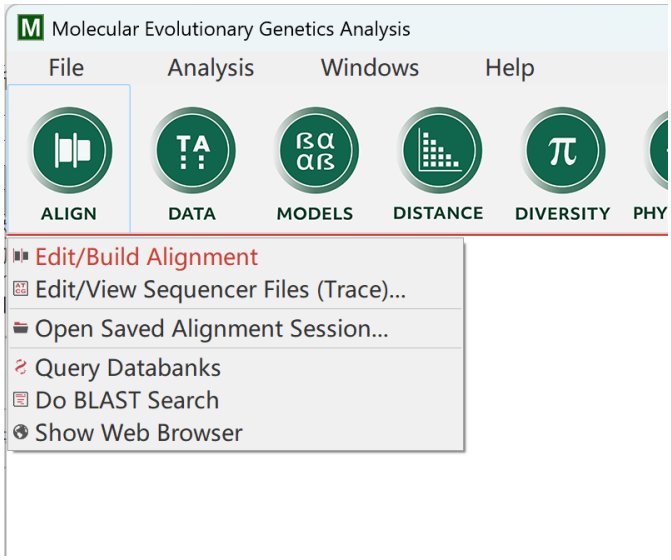
下载 mRNA 序列并提取 CDS，打开 NCBI Batch Entrez，Database 选择 Nucleotide，把这 11 条 Accession 号粘贴进去，下载 11 条 CDS 序列

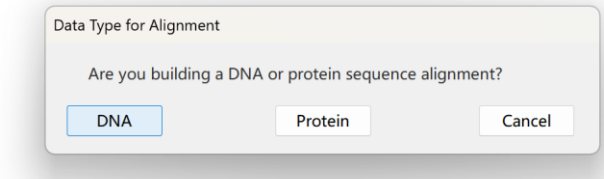



```
>lcl|NM_000559.3_cds_NP_000550.2_1 [gene=HBG1] [db_xref=CCDS:CCDS7754.1] [protein=hemoglobin subunit gamma-1] [protein_id=NP_000550.2] [location=54..497] [gbkey=CDS]
ATGGGTCATTTCACAGAGGAGGACAAGGCTACTATCACAAGCCTGTGGGGCAAGGTGAATGTGGAAGATG
CTGGAGGAGAAACCCTGGGAAGGCTCCTGGTTGTCTACCCATGGACCCAGAGGTTCTTTGACAGCTTTGG
CAACCTGTCCTCTGCCTCTGCCATCATGGGCAACCCCAAAGTCAAGGCACATGGCAAGAAGGTGCTGACT
TCCTTGGGAGATGCCACAAAGCACCTGGATGATCTCAAGGGCACCTTTGCCCAGCTGAGTGAACTGCACT
GTGACAAGCTGCATGTGGATCCTGAGAACTTCAAGCTCCTGGGAAATGTGCTGGTGACCGTTTTGGCAAT
CCATTTCGGCAAAGAATTCACCCCTGAGGTGCAGGCTTCCTGGCAGAAGATGGTGACTGCAGTGGCCAGT
GCCCTGTCCTCCAGATACCACTGA

>lcl|NM_005332.3_cds_NP_005323.1_1 [gene=HBZ] [db_xref=CCDS:CCDS10397.1] [protein=hemoglobin subunit zeta] [protein_id=NP_005323.1] [location=267..695] [gbkey=CDS]
ATGTCTCTGACCAAGACTGAGAGGACCATCATTGTGTCCATGTGGGCCAAGATCTCCACGCAGGCCGACA
CCATCGGCACCGAGACTCTGGAGAGGCTCTTCCTCAGCCACCCGCAGACCAAGACCTACTTCCCGCACTT
CGACCTGCACCCGGGGTCCGCGCAGTTGCGCGCGCACGGCTCCAAGGTGGTGGCCGCCGTGGGCGACGCG
GTGAAGAGCATCGACGACATCGGCGGCGCCCTGTCCAAGCTGAGCGAGCTGCACGCCTACATCCTGCGCG
TGGACCCGGTCAACTTCAAGCTCCTGTCCCACTGCCTGCTGGTCACCCTGGCCGCGCGCTTCCCCGCCGA
CTTCACGGCCGAGGCCCACGCCGCCTGGGACAAGTTCCTATCGGTCGTATCCTCTGTCCTGACCGAGAAG
TACCGCTGA

>lcl|NM_000519.4_cds_NP_000510.1_1 [gene=HBD] [db_xref=CCDS:CCDS31376.1] [protein=hemoglobin subunit delta] [protein_id=NP_000510.1] [location=51..494] [gbkey=CDS]
ATGGTGCATCTGACTCCTGAGGAGAAGACTGCTGTCAATGCCCTGTGGGGCAAAGTGAACGTGGATGCAG
TTGGTGGTGAGGCCCTGGGCAGATTACTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGG
GGATCTGTCCTCTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAGGTGCTAGGT
GCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACTTTTTCTCAGCTGAGTGAGCTGCACT
GTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCTTGGGCAATGTGCTGGTGTGTGTGCTGGCCCG
CAACTTTGGCAAGGAATTCACCCCACAAATGCAGGCTGCCTATCAGAAGGTGGTGGCTGGTGTGGCTAAT
GCCCTGGCTCACAAGTACCATTGA

>lcl|NM_005330.4_cds_NP_005321.1_1 [gene=HBE1] [db_xref=CCDS:CCDS7756.1] [protein=hemoglobin subunit epsilon] [protein_id=NP_005321.1] [location=56..499] [gbkey=CDS]
ATGGTGCATTTTACTGCTGAGGAGAAGGCTGCCGTCACTAGCCTGTGGAGCAAGATGAATGTGGAAGAGG
CTGGAGGTGAAGCCTTGGGCAGACTCCTCGTTGTTTACCCCTGGACCCAGAGATTTTTTGACAGCTTTGG
AAACCTGTCGTCTCCCTCTGCCATCCTGGGCAACCCCAAGGTCAAGGCCCATGGCAAGAAGGTGCTGACT
```

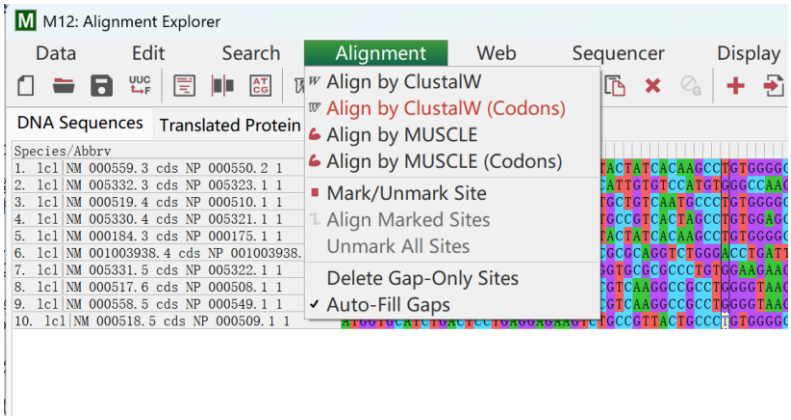
MEGA 多序列比对

打开软件。点击 Edit/Build Alignment,选择 DNA，粘贴 11 条 CDS 序列。





点击 Alignment,选择 Align by ClustalW(Codons)



结果如图，标*是高度保守区域

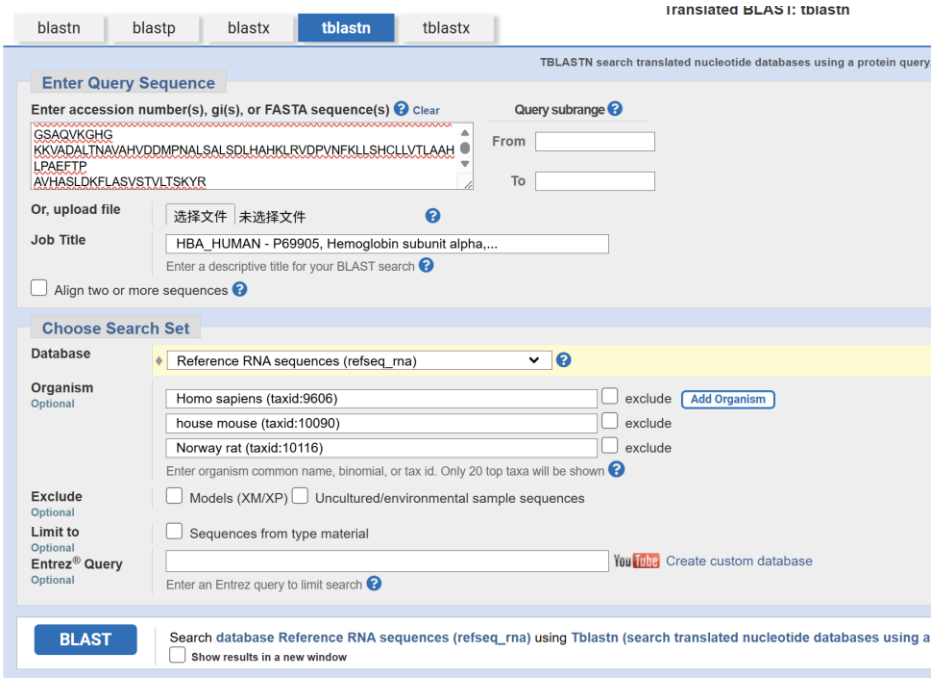
b 以人血红蛋白 alpha 亚基（HBA_HUMAN）为检测序列，搜索 RefSeq 数据库中人、小鼠和大鼠三个物种珠蛋白家族 mRNA 序列，提取其编码区序列， 进行多序列比对，分析结果。

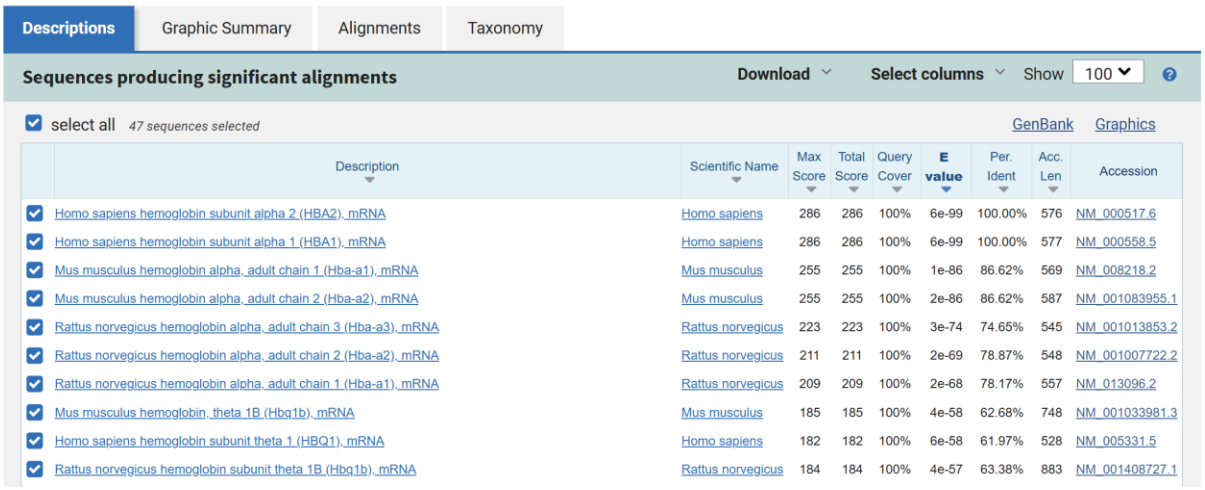
同上一个问题类似，只在 Organism 选择有修改

Homo sapiens (human)（taxid:9606）

Mus musculus (house mouse)（taxid:10090）

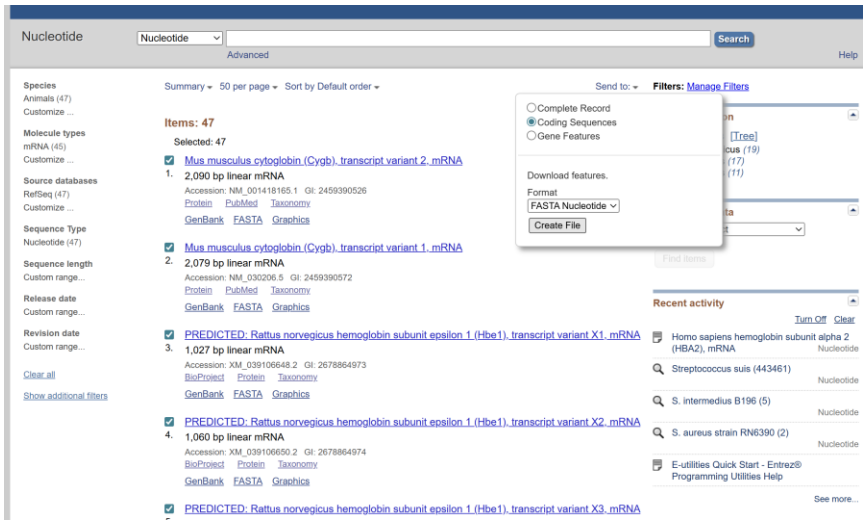
Rattus norvegicus (Norway rat)（taxid:10116）





Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Homo sapiens hemoglobin subunit alpha 2 (HBA2), mRNA	Homo sapiens	286	286	100%	6e-99	100.00%	576	NM_000517.6
Homo sapiens hemoglobin subunit alpha 1 (HBA1), mRNA	Homo sapiens	286	286	100%	6e-99	100.00%	577	NM_000558.5
Mus musculus hemoglobin alpha, adult chain 1 (Hba-a1), mRNA	Mus musculus	255	255	100%	1e-86	86.62%	569	NM_008218.2
Mus musculus hemoglobin alpha, adult chain 2 (Hba-a2), mRNA	Mus musculus	255	255	100%	2e-86	86.62%	587	NM_001083955.1
Rattus norvegicus hemoglobin alpha, adult chain 3 (Hba-a3), mRNA	Rattus norvegicus	223	223	100%	3e-74	74.65%	545	NM_001013853.2
Rattus norvegicus hemoglobin alpha, adult chain 2 (Hba-a2), mRNA	Rattus norvegicus	211	211	100%	2e-69	78.87%	548	NM_001007722.2
Rattus norvegicus hemoglobin alpha, adult chain 1 (Hba-a1), mRNA	Rattus norvegicus	209	209	100%	2e-68	78.17%	557	NM_013096.2
Mus musculus hemoglobin, theta 1B (Hbq1b), mRNA	Mus musculus	185	185	100%	4e-58	62.68%	748	NM_001033981.3
Homo sapiens hemoglobin subunit theta 1 (HBQ1), mRNA	Homo sapiens	182	182	100%	6e-58	61.97%	528	NM_005331.5
Rattus norvegicus hemoglobin subunit theta 1B (Hbq1b), mRNA	Rattus norvegicus	184	184	100%	4e-57	63.38%	883	NM_001408727.1

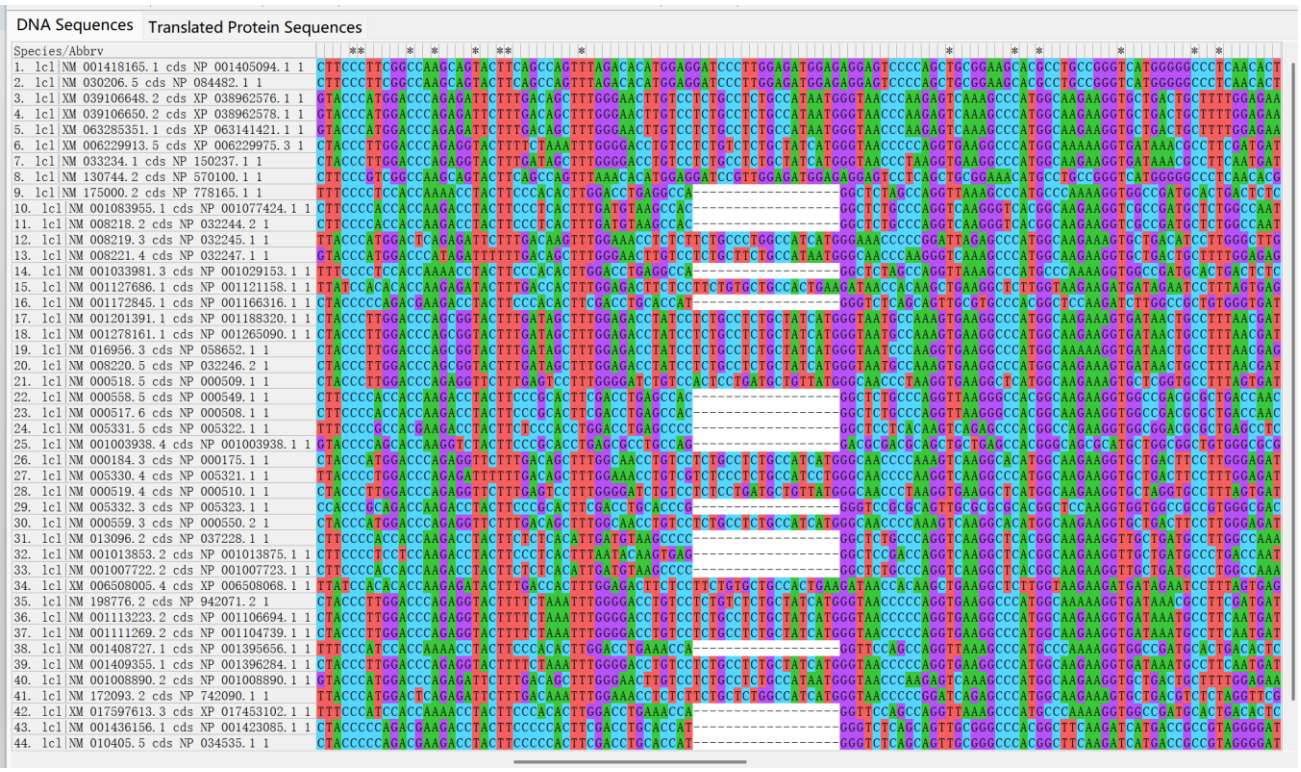
获取 CDS 序列

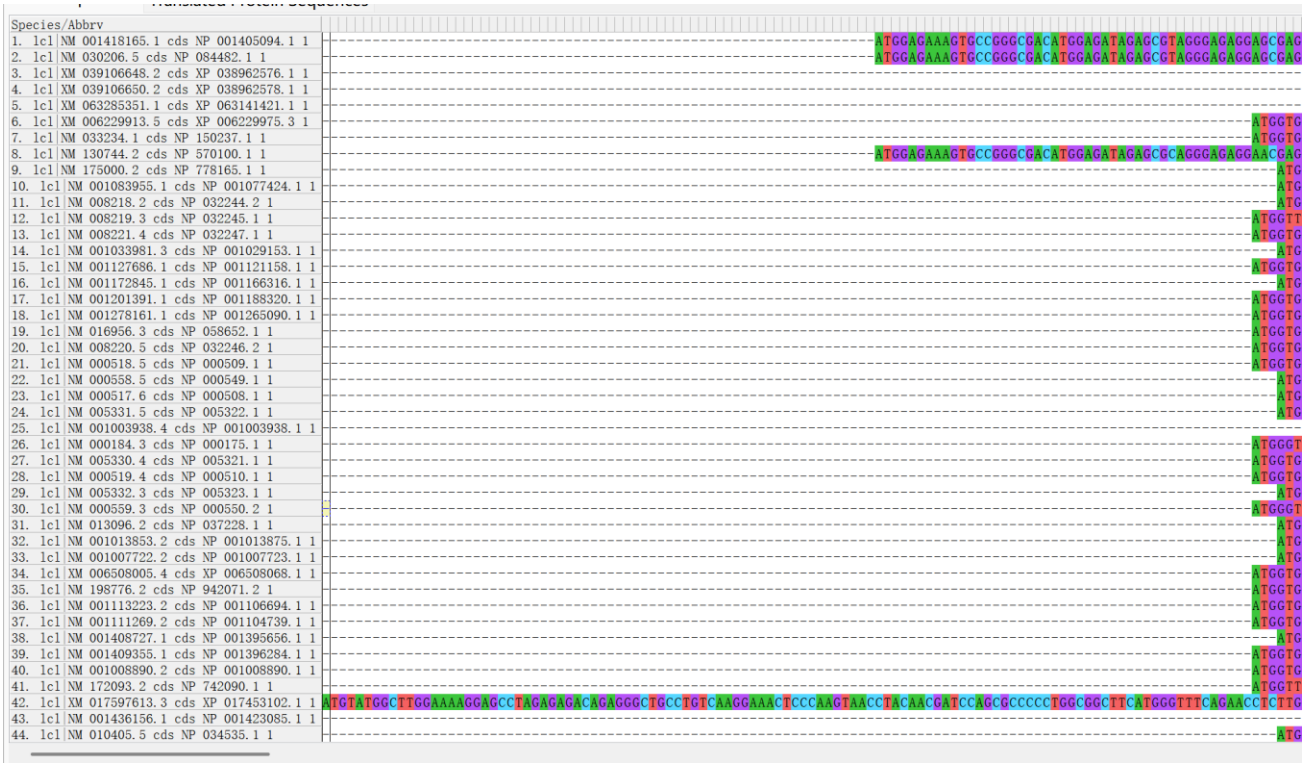


>lcl|NM_001418165.1_cds_NP_001405094.1_1 [gene=Cygb] [db_xref=GeneID:114886,MGI:MGI:2149481] [protein=cytoglobin isoform 2] [protein_id=NP_001405094.1] [location=143..685] [gbkey=CDS]
ATGGAGAAAGTGCCGGGCGACATGGAGATAGAGCGTAGGGAGAGGAGCGAGGAGCTGTCCGAGGCGGAGA
GGAAGGCGGTTCAGGCTACGTGGGCCCGGCTGTATGCCAACTGCGAGGACGTGGGGGTGGCCATCCTGGT
GAGGTTCTTTGTGAACTTCCCTTCGGCCAAGCAGTACTTCAGCCAGTTTAGACACATGGAGGATCCCTTG
GAGATGGAGAGGAGTCCCCAGCTGCGGAAGCACGCCTGCCGGGTCATGGGGGCCCTCAACACTGTCGTGG
AGAACCTGCATGACCCAGACAAGGTATCCTCTGTGCTCGCCCTGGTCGGCAAGGCCCACGCCCTCAAGCA
CAAGGTGGAACCTATGTACTTTAAGATTCTCTCTGGGGTCATTCTGGAGGTGATCGCCGAGGAATTTGCC
AATGACTTCCCTGTGGAGACGCAGAAAGCCTGGGCCAAGCTGCGGGGTCTCATCTACAGCCACGTGACCG
CAGCCTACAAGGAAGTGGGCTGGGTACAGCAGGTCCCCAACACCACCACGTGA

>lcl|NM_030206.5_cds_NP_084482.1_1 [gene=Cygb] [db_xref=CCDS:CCDS25673.1] [protein=cytoglobin isoform 1] [protein_id=NP_084482.1] [location=143..715] [gbkey=CDS]
ATGGAGAAAGTGCCGGGCGACATGGAGATAGAGCGTAGGGAGAGGAGCGAGGAGCTGTCCGAGGCGGAGA
GGAAGGCGGTTCAGGCTACGTGGGCCCGGCTGTATGCCAACTGCGAGGACGTGGGGGTGGCCATCCTGGT
GAGGTTCTTTGTGAACTTCCCTTCGGCCAAGCAGTACTTCAGCCAGTTTAGACACATGGAGGATCCCTTG
GAGATGGAGAGGAGTCCCCAGCTGCGGAAGCACGCCTGCCGGGTCATGGGGGCCCTCAACACTGTCGTGG
AGAACCTGCATGACCCAGACAAGGTATCCTCTGTGCTCGCCCTGGTCGGCAAGGCCCACGCCCTCAAGCA
CAAGGTGGAACCTATGTACTTTAAGATTCTCTCTGGGGTCATTCTGGAGGTGATCGCCGAGGAATTTGCC
AATGACTTCCCTGTGGAGACGCAGAAAGCCTGGGCCAAGCTGCGGGGTCTCATCTACAGCCACGTGACCG
CAGCCTACAAGGAAGTGGGCTGGGTACAGCAGGTCCCCAACACCACCACCCCACCAGCCACTCTGCCCTC
TTCAGGGCCGTAA

MEGA 多序列比对分析





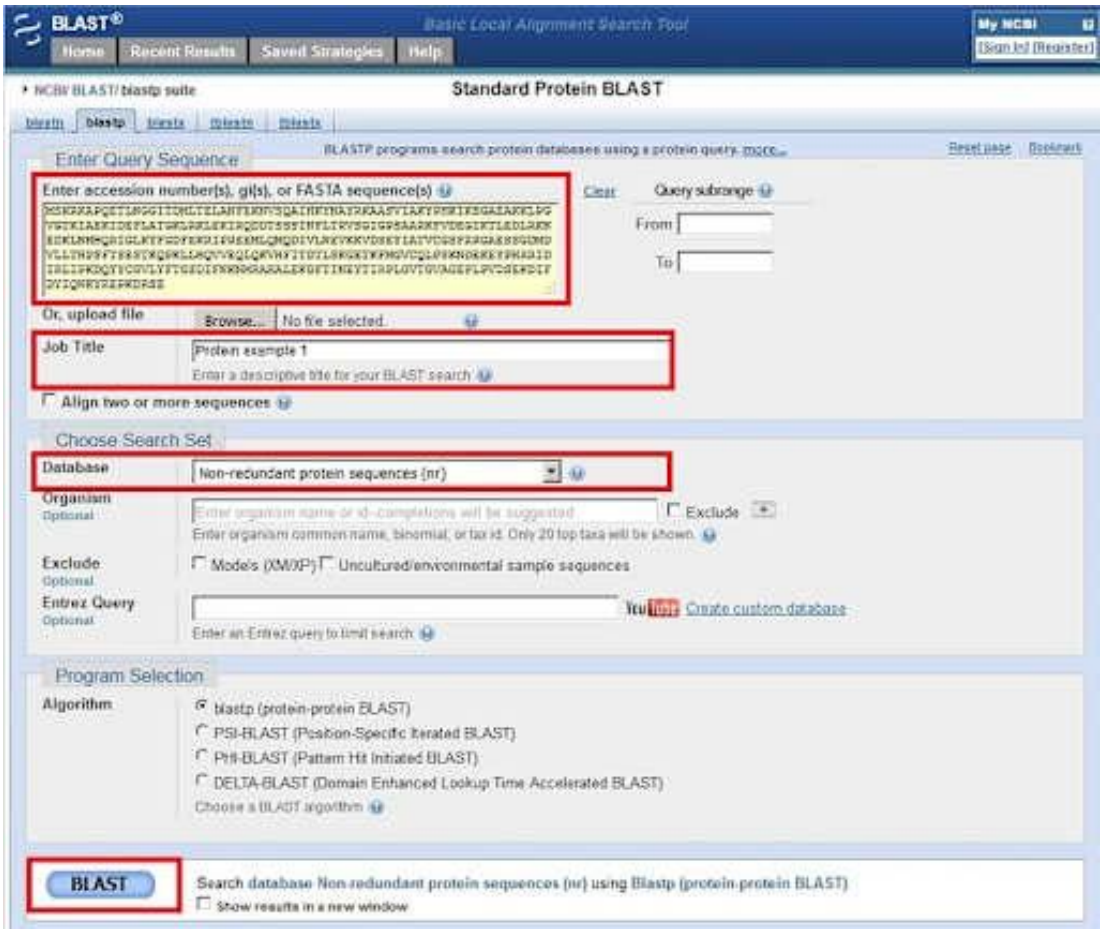
两个分析结果对比：人珠蛋白家族 mRNA 序列同物种家族内部保守性极高。人、小鼠和大鼠三个物种珠蛋白家族 mRNA 序列功能核心区依然高度保守，两端序列跨物种变异明显增大。

7. 以人癌胚抗原 CEAM1_HUMAN 的恒定结构域 Ig-like C2-type 1（145–232）搜索 Swiss-Prot 中人的 CEA 家族成员，分析搜索结果，并说明 Max Score 和 Total Score 的含义

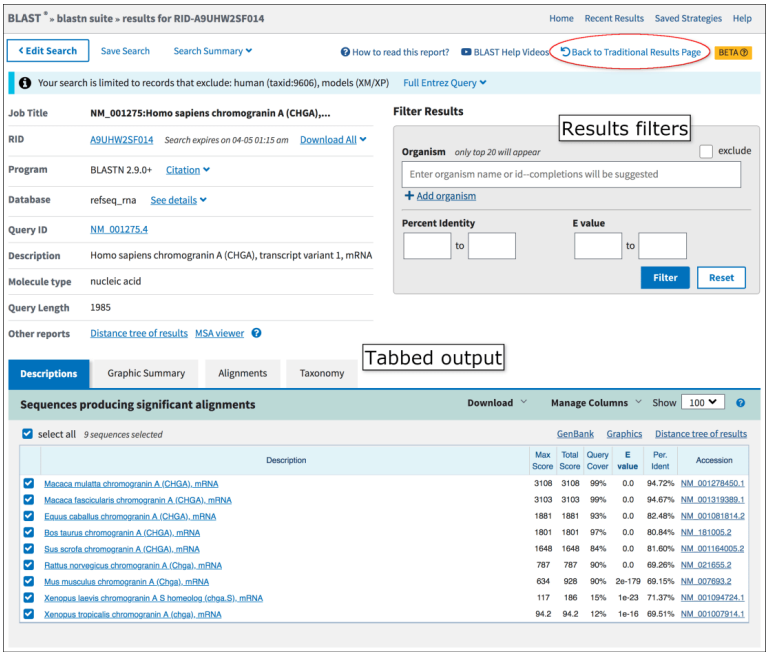
本实践是利用人 CEAM1_HUMAN 的一个恒定结构域片段，在 Swiss-Prot 蛋白数据库中搜索同源蛋白，并观察人 CEA/CEACAM 家族成员的匹配情况。 CEACAM 家族属于免疫球蛋白超家族，其胞外区域通常由 1 个 N 端 IgV 样结构域和若干个 Ig-like C2-type 结构域组成，因此用 C2 型结构域片段搜索时，往往可以命中多个 CEACAM 家族成员。

首先进入 UniProt，检索 CEACAM1 Homo sapiens，可以找到 CEAM1_HUMAN 对应的人源蛋白条目，并在条目中看到其结构域注释信息。 根据题目要求，选择其恒定结构域 Ig-like C2-type 1（145–232） 作为查询序列，将该片段对应的氨基酸序列提取出来，保存为 FASTA 格式，作为后续 BLAST 的输入序列。

在进行同源搜索时，可进入 Protein BLAST 页面，将上述 C2 结构域序列粘贴到 Query 框中，并将数据库限定为经过人工校注的 Swiss-Prot 蛋白集，这样可以减少冗余序列干扰，使结果更适合课程作业分析。如果只想看人源蛋白，还可以将 Organism 限定为 Homo sapiens (taxid:9606)，从而聚焦于人的 CEA 家族成员



提交检索后，结果页通常会先出现结果概览和命中列表，其中常见字段包括 Description、Max Score、Total Score、Query Cover、E-value、Per. Ident 等。 在这种结果表里，若查询片段与某条目标蛋白只有一段显著局部同源区域，则该蛋白的 Max Score 与 Total Score 往往相同；若目标蛋白中存在多个相似结构域并形成多个 HSP，则 Total Score 会大于 Max Score。



根据 CEACAM 家族的结构特点，用 CEAM1_HUMAN 的 C2 型结构域搜索时，通常可以检出多个人的 CEACAM 家族成员，如 CEACAM1、CEACAM3、CEACAM5、CEACAM6、CEACAM7、CEACAM8 等。其中 CEACAM5 是经典癌胚抗原成员，具有多个 Ig-like C2-type 结构域，因此它与该查询片段之间不一定只有一个局部匹配，而可能形成多个 HSP。

BLAST 中的 Max Score 是指某一条命中序列与查询序列之间，所有高分片段对（HSP）中得分最高的那一个片段的比特分值。 也就是说，Max Score 反映的是“这条蛋白和查询序列之间最好的一段局部比对到底有多强”。 如果把一条命中蛋白看成“若干局部片段的集合”，那么 Max Score 只看其中最好的一段。

而 Total Score 是该命中蛋白与查询序列之间所有 HSP 比特分值的总和，因此它反映的是“整条蛋白上有多少处局部区域都能和查询序列匹配”。 对于只形成一个 HSP 的序列，Total Score 与 Max Score 相等；对于像 CEACAM5 这种含多个相似 C2 结构域的蛋白，Total Score 往往大于 Max Score，因为多个局部匹配区域的得分被累加了起来。

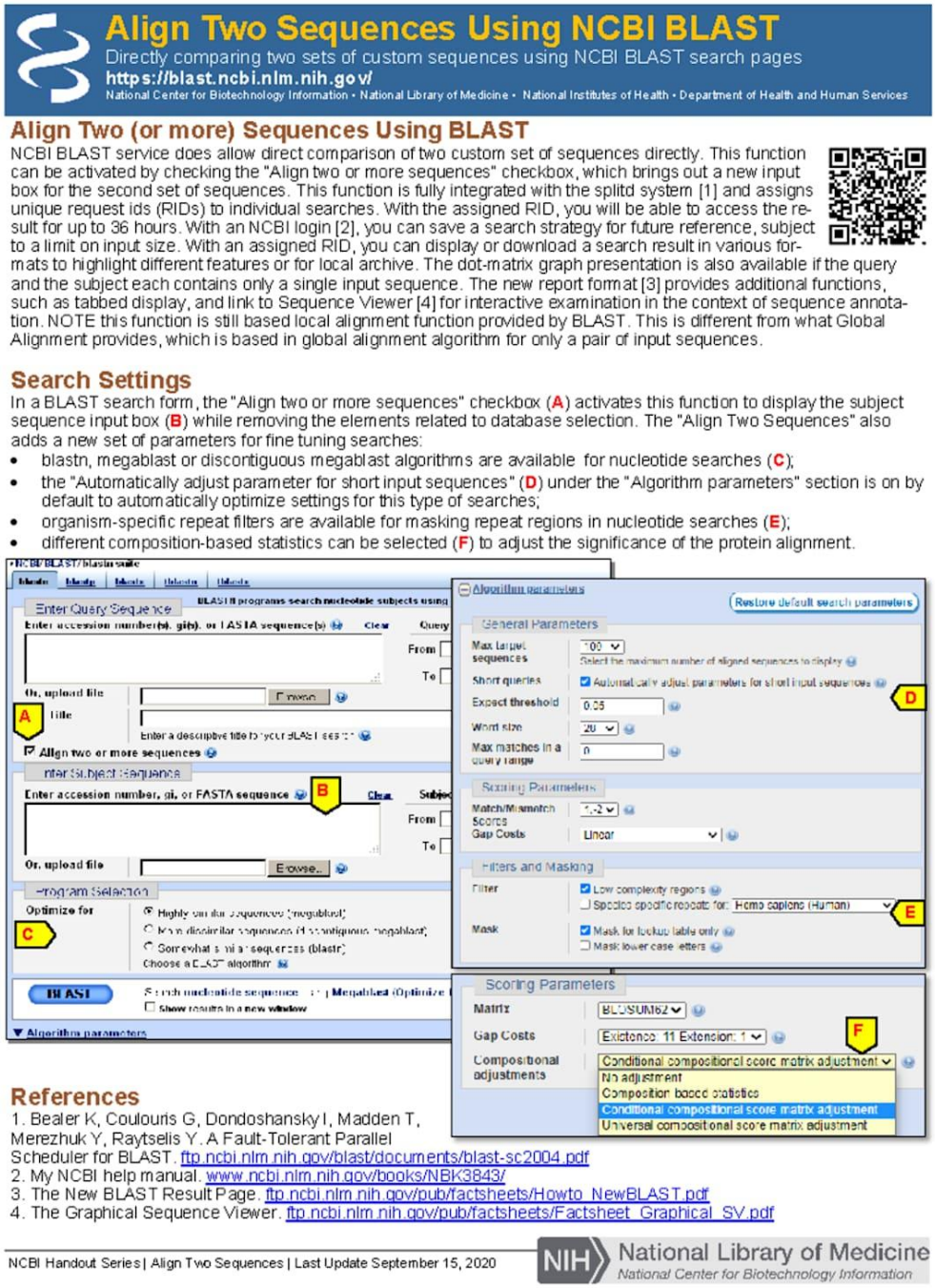
因此，在解释结果时可以这样写：如果某条命中蛋白 Max Score 高而 Total Score 与其接近，说明这条蛋白主要只有一段与查询序列高度相似的区域；如果某条命中蛋白 Max Score 高且 Total Score 明显更高，则提示该蛋白可能含有多个与查询片段相似的重复结构域。 对 CEA 家族来说，这种现象正好与家族成员存在多个 Ig-like C2-type 结构域的特点相一致。

本题结果分析可以总结为：用 CEAM1_HUMAN 的 Ig-like C2-type 1 片段搜索 Swiss-Prot 中的人源蛋白，能够检出多种 CEACAM 家族成员，说明该恒定结构域在家族内部具有明显保守性。 同时，Max Score 用于衡量单个最佳局部比对片段的质量，而 Total Score 用于反映整条命中蛋白与查询片段在多个局部区域上的总体相似程度。

8. 以人癌胚抗原 CEA21_HUMAN 的恒定结构域 Ig-like C2-type（147–231）与 CEAM5_HUMAN 进行双序列比对，分析比对结果，并说明 BLAST 与 EMBOSS 软件包中程序 water 运行结果的差别

重点不是再做数据库搜索，而是拿两条已经确定的蛋白序列直接进行双序列比对，观察局部同源区域的具体对齐情况。 其中 BLAST 的 “Align two or more sequences” 页面属于基于 BLAST 的双序列局部比对界面，而 EMBOSS 的 water 程序则是基于 Smith-Waterman 算法的严格局部比对程序。

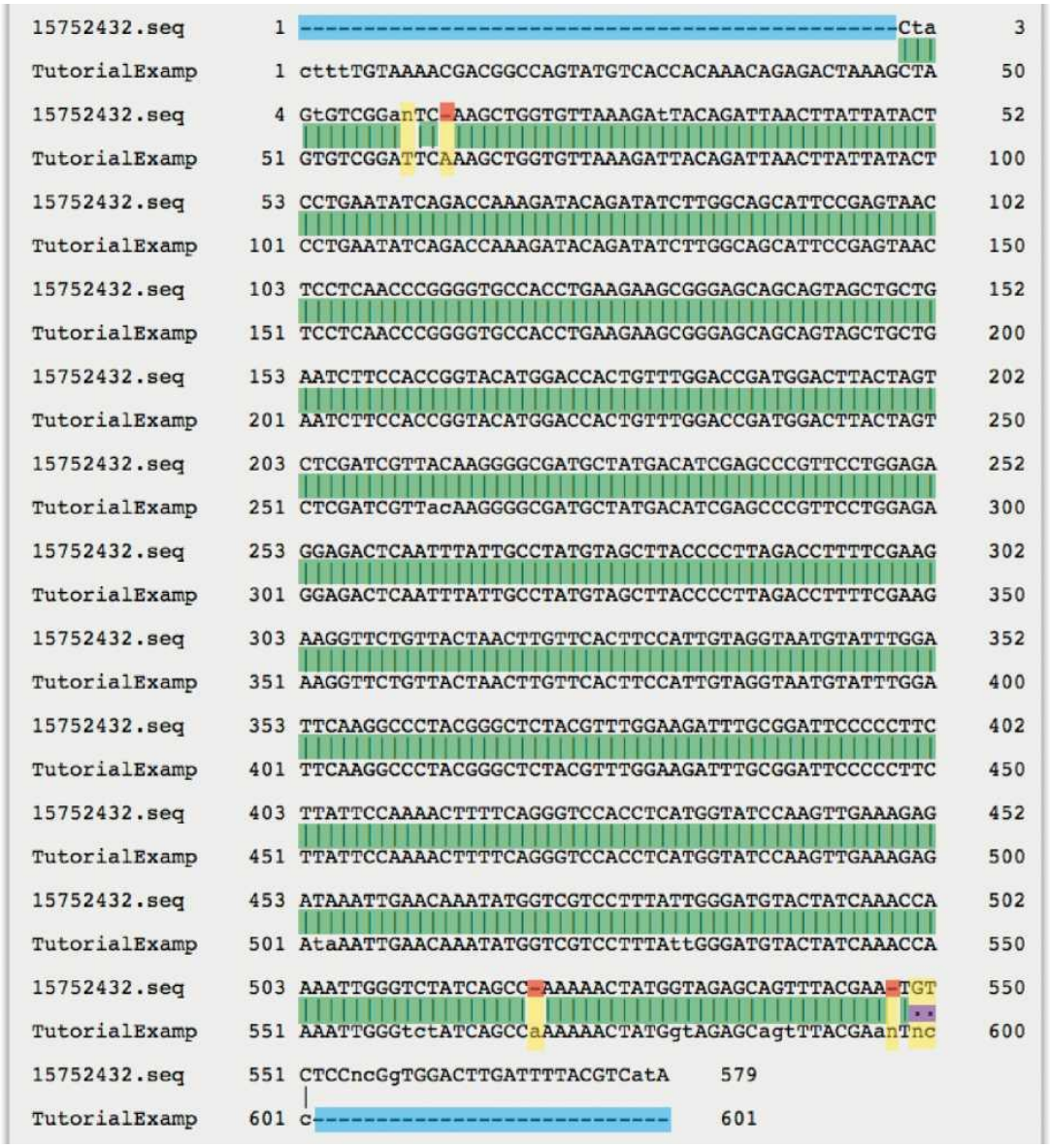
第一步，先准备两条序列：一条是 CEA21_HUMAN 的 Ig-like C2-type（147–231）片段，另一条是 CEAM5_HUMAN 的蛋白序列。 然后进入 NCBI Protein BLAST 页面，选择 Align two or more sequences，将 CEA21_HUMAN 的 C2 结构域序列粘贴在 Query 中，把 CEAM5_HUMAN 的序列粘贴在 Subject 中，再使用默认参数运行比对。



BLAST 的核心思想是先把查询序列切分成较短的词（word），先找到种子匹配，再向两侧延伸形成高分片段对 HSP，因此它属于一种启发式局部比对方法，速度快，但理论上不保证一定得到严格最优的局部比对结果。 在本题中，由于 CEAM5_HUMAN 含有多个 Ig-like C2-type 相关结构域，因此 BLAST 结果中有可能出现 1 个以上的 HSP，其中得分最高的那个 HSP 往往对应于 CEAM5 上与查询片段最相似的那个 C2 样区域。

BLAST 输出通常会给出 Score、E-value、Identities、Positives、Gaps 等信息。 其中 Identities 表示完全相同氨基酸所占比例，Positives 表示理化性质相近、在替换矩阵中得分为正的保守替换比例，而 E-value 用于评估该比对结果在数据库背景下随机出现的可能性，数值越小，说明显著性越高。

第二步，再使用 EMBOSS 的 water 程序进行同样的双序列比对。water 使用 Smith-Waterman 算法，通过动态规划在全部可能的局部对齐方案中寻找得分最高者，因此它输出的是两条序列之间严格最优的局部比对结果。



water 的输出通常包括 Alignment length、Identity、Similarity、Gaps 和 Score 等内容。 与 BLAST 不同，water 重点展示一段最优局部比对本身，而不是从数据库检索意义上去评估该结果，因此它通常不提供像 BLAST 那样的 E-value 统计显著性指标。

在本题中，如果查询片段与 CEAM5_HUMAN 的某一个 C2 型恒定结构域高度相似，那么 BLAST 结果中的最佳 HSP 与 water 给出的最优局部比对区域通常会大体一致。 但两者输出形式不同：BLAST 可能给出多个 HSP 并带有 E-value，而 water 只给出一段局部最优比对，同时更直接地列出 Identity、Similarity 与 Gaps 百分比。

从软件原理上看，BLAST 更适合“先找像谁”这一类任务，因为它本质上是快速搜索工具，适合数据库检索和同源序列筛选。water 更适合“这两条序列到底怎样对齐”这一类任务，因为它强调的是两条指定序列之间最优局部比对的精细展示。

比较项目	BLAST	EMBOSS water
算法基础	启发式局部比对，先找种子再延伸 chanzuckerberg.zendesk+1	Smith-Waterman 动态规划局部比对 bioinformatics+1
结果形式	可能出现多个 HSPbigcat-um.github	给出单个最优局部比对 bioinformatics
是否保证最优	不保证严格最优 biostars	保证最优局部比对 bioinformatics
是否有 E-value	有 chanzuckerberg.zendesk	无 bioinformatics
适用场景	数据库检索、快速筛选同源序列 blast.ncbi.nlm.nih	两条指定序列的精细局部对齐分析 ebi.ac

5. 问题

1. 为什么有些序列的 score 低但是 e-value 值比较高？哪个参数更可信？

Score 衡量的是两条序列匹配的质量，而 E-value 则是将该得分置于数据库背景下评估其为巧合的概率；由于 E-value 受序列长度和数据库规模的影响，一个低 Score 在大数据背景下极易随机出现，从而导致其 E-value 升高（显著性变差）。在判定同源性时，E-value 比 Score 更具统计学参考价值

2. 为什么用 tBLASTN 以人 HBA 蛋白为检索序列，能筛选出人、小鼠、大鼠的珠蛋白 mRNA？跨物种比对相比单物种比对，多了哪些生物学分析价值？

tBLASTN 是蛋白检索核酸库，珠蛋白在哺乳动物进化中蛋白功能域高度保守，氨基酸序列相似度高；即便人、鼠核酸序列有差异，保守的蛋白序列仍能匹配到 RefSeq 中同源珠蛋白 mRNA，且 E 值显著，可准确筛选家族序列。

单物种比对只能分析物种内基因家族保守性与亚型进化关系；跨物种比对可进一步比较灵长类与啮齿类珠蛋白的保守程度、物种特有变异；能验证基因分化早于物种分化的进化特征，更全面揭示珠蛋白基因家族的演化。