“实用生物信息技术”课程小组讨论总结报告

组：G1　　次：R2　　组长：薛林蕾　　执笔：薛林蕾，吴鸿珍，江浩燊

1. 时间

2026.4.10

2. 方式

线上会议讨论

3. 主题

课后复习与拓展

4. 内容

G1A：Uniprot 数据库、ENSEMBL 基因组数据库使用练习

A1:阅读分子月报、蛋白质精选以及维基百科等网站中有关血红蛋白的介绍，了解血红蛋白的生理功能、空间结构、亚基组成等基本知识。

血红蛋白是使血液呈现红色的蛋白质。它由四条蛋白质链组成，两条 α 链和两条 β 链，每条链都含有一个环状的血红素基团，其中包含一个铁原子。氧气可逆地与这些铁原子结合，并通过血液运输。

除了运输氧气外，血红蛋白还能结合并运输其他分子，例如一氧化氮和一氧化碳。一氧化氮作用于血管壁，使其舒张，从而降低血压。近期研究表明，一氧化氮可以与血红蛋白中的特定半胱氨酸残基以及血红素基团中的铁结合.因此，血红蛋白通过在血液中输送一氧化氮来调节血压。另一方面，一氧化碳是一种有毒气体。它很容易取代血红素基团上的氧以及其他许多物质，它们会形成难以去除的稳定复合物。这种对血红素基团的滥用会阻碍正常的氧气结合和运输，使周围细胞窒息而死。

血红蛋白是一种奇妙的分子机器，它利用运动和微小的结构变化来调节自身的功能。血红蛋白中四个血红素位点的氧结合并非同时发生。第一个血红素结合氧后，会引起相应蛋白质链结构发生微小变化。这些变化会促使相邻的蛋白质链改变形状，从而更容易结合氧。因此，第一个氧分子的结合较为困难，但第二个、第三个和第四个氧分子的结合则逐渐变得容易。这赋予了血红蛋白强大的功能。当血液流经氧气充足的肺部时，氧气很容易与第一个亚基结合，并迅速填充剩余的亚基。随后，随着血液在体内循环，氧气浓度下降，二氧化碳浓度上升。在这种环境下，血红蛋白会释放其结合的氧。第一个氧分子一旦脱离，蛋白质就开始改变其形状。这促使剩余的三个氧分子迅速释放。这样，血红蛋白就能在肺部吸收尽可能多的氧气，并在需要的时候将所有氧气输送到身体各处。

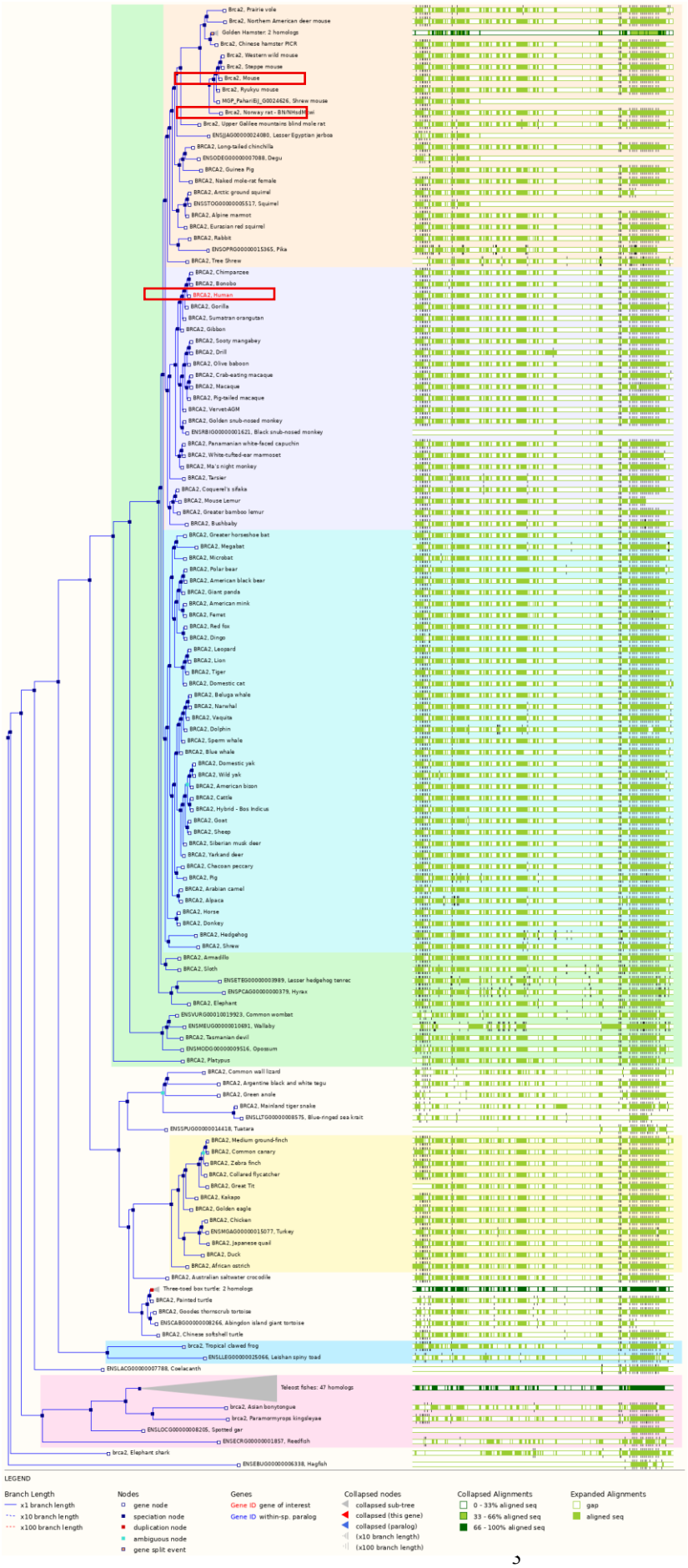
血红蛋白的结构为异源四聚体，有 4 个亚单位（α 亚基、β 亚基），即 2 对珠蛋白组成，每个珠蛋白结合 1 个血红素。

血红蛋白的每个亚基由一条肽链和一个血红素分子构成，肽链在生理条件下会盘绕折叠成球形，把血红素分子抱在里面，这条肽链盘绕成的球形结构又被称为珠蛋白。血红素分子是一个具有卟啉结构的小分子，在卟啉分子中心，由卟啉中四个吡咯环上的氮原子与一个亚铁离子配位结合，珠蛋白肽链中第 8 位的一个组氨酸残基中的咪唑侧链上的氮原子从卟啉分子平面的上方与亚铁离子配位结合，使中心离子铁(II)成为五配位，既是配位中心，又是活性中心。当血红蛋白不与氧结合的时候，有一个水分子从卟啉环下方与

亚铁离子配位结合，而当血红蛋白载氧的时候，就由氧分子顶替水的位置。

A2 查阅 ENSEMBL 基因组数据库中已经或正在进行基因组测序的物种树，了解人、小鼠、大鼠三个物种之间演化关系；检索物种分歧时间数据库 TimeTree，了解人和小鼠、小鼠和大鼠之间的分歧时间。

小鼠（*M. musculus*）和大鼠（*R. norvegicus*）亲缘关系极近，它们同属于啮齿目（Rodents），拥有一个较近的共同祖先。相比之下，人类属于灵长目（Primates）。在演化树上，灵长目和啮齿目在很久以前就分化了。在演化史上，大鼠与小鼠约在 13.1 MYA 分化；而人类与啮齿类的共同祖先需追溯至约 87.0 MYA 的白垩纪时期。


Brca2, Prairie vole
Brca2, Northern American deer mouse
Golden Hamster: 2 homologs
Brca2, Chinese hamster PICR
Brca2, Western wild mouse
Brca2, Steppe mouse
Brca2, Mouse
Brca2, Ryukyu mouse
MGP_PahariEJ_G0024626, Shrew mouse
Brca2, Norway rat - BN/NHsdMcwi
Brca2, Upper Galilee mountains blind mole rat
ENSGAG00000024080, Lesser Egyptian jerboa
BRCA2, Long-tailed chinchilla
ENSODEG00000007088, Degu
BRCA2, Guinea Pig
BRCA2, Naked mole-rat female
BRCA2, Arctic ground squirrel
ENSSTOG00000005517, Squirrel
BRCA2, Alpine marmot
BRCA2, Eurasian red squirrel
BRCA2, Rabbit
ENSOPRG00000015365, Pika
BRCA2, Tree Shrew
BRCA2, Chimpanzee
BRCA2, Bonobo
BRCA2, Human
BRCA2, Gorilla
BRCA2, Sumatran orangutan
BRCA2, Gibbon
BRCA2, Sooty mangabey
BRCA2, Drill
BRCA2, Olive baboon
BRCA2, Crab-eating macaque
BRCA2, Macaque
BRCA2, Pig-tailed macaque
BRCA2, Vervet-AGM
BRCA2, Golden snub-nosed monkey
ENSRBIG00000001621, Black snub-nosed monkey
BRCA2, Panamanian white-faced capuchin
BRCA2, White-tufted-ear marmoset
BRCA2, Ma's night monkey
BRCA2, Tarsier
BRCA2, Coquerel's sifaka
BRCA2, Mouse Lemur
BRCA2, Greater bamboo lemur
BRCA2, Bushbaby
BRCA2, Greater horseshoe bat
BRCA2, Megabat
BRCA2, Microbat
BRCA2, Polar bear
BRCA2, American black bear
BRCA2, Giant panda
BRCA2, American mink
BRCA2, Ferret
BRCA2, Red fox
BRCA2, Dingo
BRCA2, Leopard
BRCA2, Lion
BRCA2, Tiger
BRCA2, Domestic cat
BRCA2, Beluga whale
BRCA2, Narwhal
BRCA2, Vaquita
BRCA2, Dolphin
BRCA2, Sperm whale
BRCA2, Blue whale
BRCA2, Domestic yak
BRCA2, Wild yak
BRCA2, American bison
BRCA2, Cattle
BRCA2, Hybrid - Bos Indicus
BRCA2, Goat
BRCA2, Sheep
BRCA2, Siberian musk deer
BRCA2, Yarkand deer
BRCA2, Chacoan peccary
BRCA2, Pig
BRCA2, Arabian camel
BRCA2, Alpaca
BRCA2, Horse
BRCA2, Donkey
BRCA2, Hedgehog
BRCA2, Shrew
BRCA2, Armadillo
BRCA2, Sloth
ENSETEG00000003989, Lesser hedgehog tenrec
ENSPCAG00000000379, Hyrax
BRCA2, Elephant
ENSVURG00010019923, Common wombat
ENSMEUG00000010491, Wallaby
BRCA2, Tasmanian devil
ENSMODG00000009516, Opossum
BRCA2, Platypus
BRCA2, Common wall lizard
BRCA2, Argentine black and white tegu
BRCA2, Green anole
BRCA2, Mainland tiger snake
ENSLLTG00000008575, Blue-ringed sea krait
ENSSPUG00000014418, Tuatara
BRCA2, Medium ground-finch
BRCA2, Common canary
BRCA2, Zebra finch
BRCA2, Collared flycatcher
BRCA2, Great Tit
BRCA2, Kakapo
BRCA2, Golden eagle
BRCA2, Chicken
ENSMGAG00000015077, Turkey
BRCA2, Japanese quail
BRCA2, Duck
BRCA2, African ostrich
BRCA2, Australian saltwater crocodile
Three-toed box turtle: 2 homologs
BRCA2, Painted turtle
BRCA2, Goodes thornscrub tortoise
ENSCABG00000008266, Abingdon island giant tortoise
BRCA2, Chinese softshell turtle
brca2, Tropical clawed frog
ENSLLEG00000025066, Laishan spiny toad
ENSLACG00000007788, Coelacanth
Teleost fishes: 47 homologs
brca2, Asian bonytongue
brca2, Paramormyrops kingsleyae
ENSLOCG00000008205, Spotted gar
ENSECRG00000001857, Reedfish
brca2, Elephant shark
ENSEBUG00000006338, Hagfish
LEGEND
Branch Length
x1 branch length
x10 branch length
x100 branch length
Nodes
gene node
speciation node
duplication node
ambiguous node
gene split event
Genes
Gene ID gene of interest
Gene ID within-sp. paralog
Collapsed nodes
collapsed sub-tree
collapsed (this gene)
collapsed (paralog)
(x10 branch length)
(x100 branch length)
Collapsed Alignments
0 - 33% aligned seq
33 - 66% aligned seq
66 - 100% aligned seq
Expanded Alignments
gap
aligned seq

A3: 从 UniProt 数据库中提取人、小鼠、大鼠血红蛋白 alpha 亚基蛋白质序列，进行双序列全局比对。分析比对结果，说明所得结果和预期不符的原因和进一步分析思路。

运用 neddle 分别将人。小鼠。大鼠的血红蛋白 alpha 亚基蛋白质序列进行比对，设置 matrix files 为 EBLOSUM62，execution code 为 interactive，结果如下：

人和小鼠：

```
# Aligned_sequences: 2
# 1: HBA_HUMAN
# 2: HBA_MOUSE
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 142
# Identity:     122/142 (85.9%)
# Similarity:   131/142 (92.3%)
# Gaps:           0/142 ( 0.0%)
# Score: 648.0
```

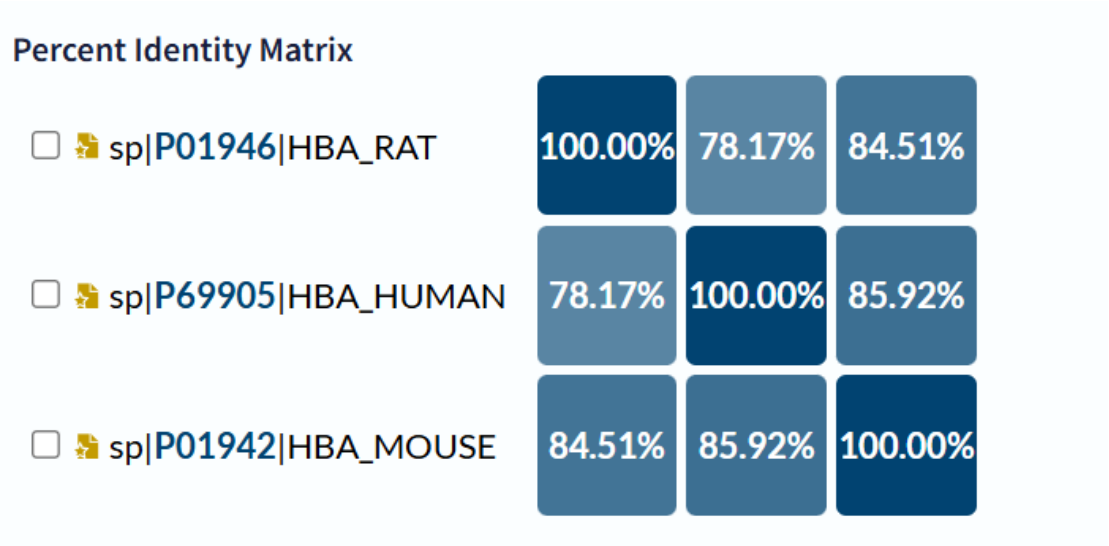
人和大鼠：

```
# Aligned_sequences: 2
# 1: HBA_HUMAN
# 2: HBA_RAT
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 142
# Identity:     111/142 (78.2%)
# Similarity:   120/142 (84.5%)
# Gaps:           0/142 ( 0.0%)
# Score: 587.0
```

小鼠和大鼠：

```
# Aligned_sequences: 2
# 1: HBA_MOUSE
# 2: HBA_RAT
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 142
# Identity:      120/142 (84.5%)
# Similarity:    127/142 (89.4%)
# Gaps:            0/142 ( 0.0%)
# Score: 632.0
```

用 uniprot 数据库中的序列比对功能得到的结果如下：

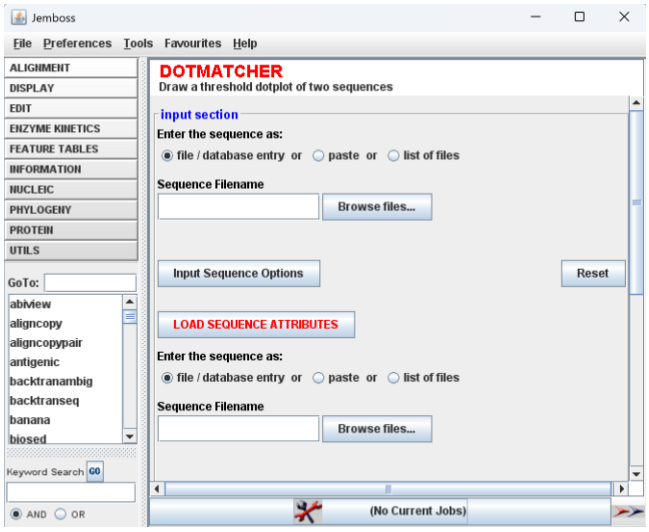


根据进化关系推测演化分化时间越长，由于积累了更多的突变，蛋白质序列相似度更低，所以双序列比对结果可能是小鼠与大鼠相似度最高，人与小鼠相似度次之，人和大鼠相似度与人和鼠相似，但是现在分析结果显示人和小鼠的相似度最高，其次是小鼠与大鼠，猜测可能是因为血红蛋白 alpha 亚基在不同谱系中进化速率并不完全一致，小鼠和大鼠虽然亲缘关系近，但两者的 HBA 基因都经历了各自独立的快速进化，啮齿类繁殖快、世代短，突变积累速度比灵长类快很多。有研究表明，某些啮齿类蛋白序列与人类的同源序列相似度，反而高于两种啮齿类之间的相似度，这在快速进化的基因家族中并不罕见。

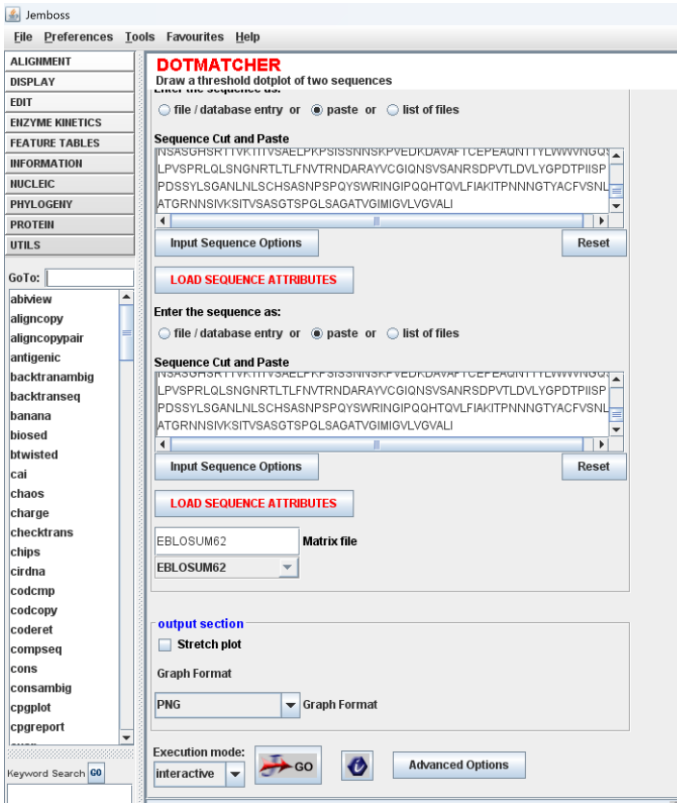
G1B： 序列比对点阵图（Dot Plot）方法应用实例

B1 以人癌胚抗原 CEAM5_HUMAN 为例，分别用 DotMatcher 和在线分析平台 Dotlet 进行点阵图分析，说明如何利用点阵图方法在寻找序列内部相似性区域。

DotMatcher：打开 EMBOSS 的 DotMatcher

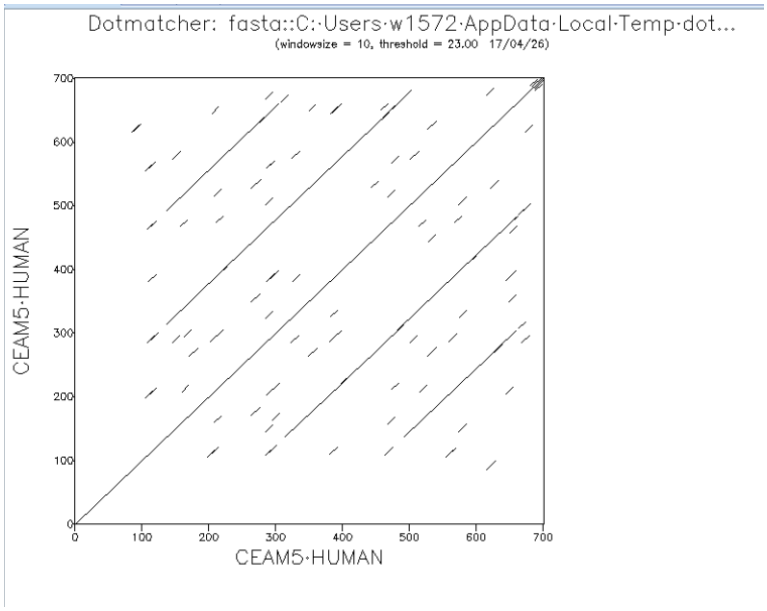


CEAM5_HUMAN 同时作为 Sequence1 和 Sequence2，输入序列，选择 BLOSUM62

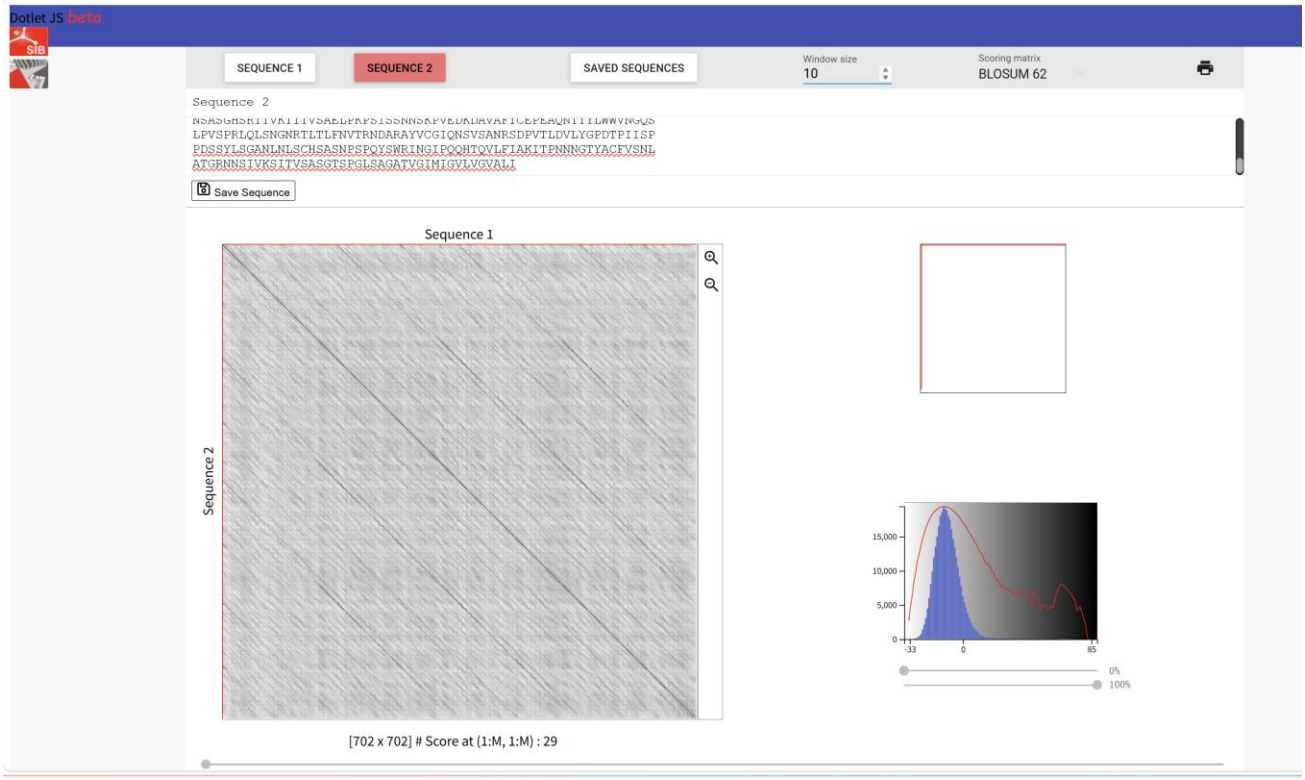


结果：主对角线是序列与自身的完全匹配，是所有自比对点阵图的基础。多条平行的短对角线簇是内部相似性 / 重复区域，与主对角线平行，说明这些序列片段在序列内部发生了正向重复。

主



Dotlet



主对角线：贯穿整个图，代表序列与自身的完全匹配，是自比对的基础信号。

多条平行的短对角线簇：是序列内部相似性 / 重复区域，它们的位置和 DotMatcher 里的信号完全对应。

右侧的得分分布直方图显示，高分信号集中，说明这些相似性是真实的同源信号，而非随机匹配。

2) 结合点阵图在线分析平台 Dotlet 中实例，说明点阵图方法的用途。

检测序列内部重复 / 结构域重复

Dotlet 实例表现：在 CEAM5_HUMAN 的自比对点阵图中，除了代表序列全长自身匹配的主对角线外，还出现了多条与主对角线平行的短对角线簇。快速定位 CEAM5 内部的高相似性重复区域。这些区域可以作为局部比对的重点，避免对全长序列进行无差别比对，提升后续分析的效率和准确性。

G1C：序列分析综合解答

C1 基本概念

1.1 序列相似性与同源性的差别和联系

同源性（Homology）是一个演化概念，指两条序列源于共同祖先（common ancestor），是"有或无"的定性关系，不能以百分比来表达。同源又分为直系同源（Orthologs）（由物种分化事件产生，如人与小鼠的 HBA1 基因）和旁系同源（Paralogs）（由基因重复事件产生，如人的 HBA1 与 HBA2 基因）。[^1]

相似性（Similarity）是序列字符串本身可计算的定量属性，用相同或相近氨基酸/核苷酸的比例来衡量，与演化历史无关。相似性是可以观察到的事实，而同源性是基于观察提出的假设。[2][3]

两者的联系：高相似性是推断同源性的主要证据，但两者不能相互替代。两条序列可以因随机或人工合成而相似，却不同源；反之，同源基因经历长期演化后可高度分化，序列相似性极低。经验上，统计学显著的序列相似性（尤其是长而复杂的序列）是共同演化起源的强有力证据。[3][4]

举例：

- 人 HBA1 与小鼠 Hba-a1 之间存在约 87% 的氨基酸相似性，两者同源（直系同源，ortholog）；
- 人 HBA1 与 HBA2 序列高度相似（编码区几乎相同），两者也同源，但属旁系同源（paralog，源于基因重复）；[^5]
- 两段随机相同的短序列（如 5 bp "ATGCA"）相似，但未必同源。

注意：在文献中不能写"序列有 85% 的同源性"，应写"序列相似性为 85%，推断两者同源"。

1.2 全局比对与局部比对的适用范围

比对方式	核心思想	典型算法	适用场景
全局比对	强制对齐两条序列的全长	Needleman-Wunsch（1970）	长度相近、整体同源的序列比较
局部比对	找出两条序列中相似性最高的子区域	Smith-Waterman（1981）	序列总体差异大，但含有保守结构域

全局比对（Global Alignment）要求比对结果覆盖两条序列的全部长度，适用于比较长度相近、亲缘关系较近的同源序列，例如人与小鼠的 HBA1 蛋白全长比对。也适合确认两条已知同源的序列之间的整体保守区域。[^6]

局部比对（Local Alignment）将负分区域截断为零，只寻找最优的局部匹配。适用于比对功能域位于序列不同位置的蛋白（例如含多个 Ig-like domain 的 CEACAM5 与其他 Ig 超家族成员），或用于数据库相似性搜索（BLAST 本质上是局部比对）。[^7]

举例：

- 全局比对：比较人、小鼠、大鼠血红蛋白 alpha 亚基全长蛋白，使用 Needleman-Wunsch 算法；
- 局部比对：在数据库中搜索未知蛋白的保守结构域，或比较 CEACAM5 中某个 Ig 重复域与免疫球蛋白的相似性，使用 Smith-Waterman 或 BLAST。

1.3 BLOSUM 与 PAM 计分矩阵的构建方法和特点

PAM 矩阵（Point Accepted Mutation）

PAM 矩阵由 Dayhoff 等人于 1978 年发表，构建方法如下：[^8]

1. 收集高度相似（>85% 序列相同）的近缘蛋白对进行全局比对；

2. 统计氨基酸替换频率，构建 PAM1 矩阵（每 100 个位置发生 1 次替换的概率）；

3. 通过矩阵乘方（PAM1 的 N 次方）外推到更远距离，得到 PAM120、PAM250 等。[^9]

特点：基于进化模型，假设所有氨基酸进化速率相同（不够准确）；数据集较老，信息量有限；PAM 数字越大，代表演化距离越远。[^8]

BLOSUM 矩阵（BLOcks SUbstitution Matrix）

BLOSUM 矩阵由 Henikoff 兄妹于 1992 年发表，构建方法如下：[^8]

1. 从 BLOCKS 数据库中收集无 gap 的保守局部序列块；

2. 对每个 BLOSUM-N 矩阵，仅使用序列相同度 ≥N% 的序列块（例如 BLOSUM62 使用 ≥62% 的序列对）；[^8]

3. 直接统计观察到的氨基酸替换频率对数比值（log-odds），无需矩阵外推。

特点：基于真实保守域观测数据，更能反映真实生物演化；BLOSUM 数字越大，序列越保守，适用于近源比对；BLOSUM 数字越小，适用于远源比对。[^10]

矩阵等价关系与选择建议

BLOSUM	PAM 等价	适用场景
BLOSUM80	PAM30	近源序列（>85% 相同）[^10]
BLOSUM62	PAM160	通用（BLAST 默认）[^9]
BLOSUM45	PAM250	远缘序列（<30% 相同）[^9]

对编码区核苷酸比对，可使用 DNA 替换矩阵（如 NUC4.4）；对蛋白质比对，BLOSUM62 是最常用默认选择。

1.4 空位罚分的意义和用法

生物学意义：序列比对中引入空位（gap）是为了模拟插入/缺失突变（indel）事件。真实的 indel 通常是单次突变事件（一次删除了若干连续核苷酸），因此把一段连续空位视为一个整体事件更符合生物学现实。[^11]

常见罚分方案：

- 线性罚分（Linear）：每个空位字符的代价相同，罚分 = gap 数 × 常数。简单但不够准确；
- 仿射罚分（Affine）：罚分 = gap 开放代价（d）+ gap 延伸代价（e）×（n-1），即 $\gamma(n) = d + (n-1) \cdot e$，其中 d > e > 0。这是目前最常用的方案，在 BLAST、Needleman-Wunsch 扩展版中均采用；[12][13]
- 凸型（Convex）：更接近实际，但计算复杂。

举例：在比对编码区核苷酸序列时，若设 gap open = -10，gap extend = -1（严格罚分），则只有在序列整体相似性较高时才会引入 gap，适合保守基因的精细比对；若设 gap open = -5，gap extend = -2（宽松罚分），则允许更多 gap，适合演化距离较远的序列。同一对序列在不同空位罚分下可能产生显著不同的比对结果，应结合生物学背景选择合理参数。

1.5 动态规划双序列比对算法基本思路

Needleman-Wunsch 全局比对

1. 初始化：构建 (m+1)×(n+1) 评分矩阵 F，其中 m、n 为两序列长度。第一行和第一列按 gap 罚分递增填写：$F(i,0) = i \cdot g$，$F(0,j) = j \cdot g$；[^14]

2. 矩阵填充：对每个格子 $F(i,j)$ 取三种来源的最大值：[^15]

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(a_i, b_j) & \text{（对角线，匹配/错配）} \\ F(i-1,j) + g & \text{（向上，序列 1 插入 gap）} \\ F(i,j-1) + g & \text{（向左，序列 2 插入 gap）} \end{cases}$$

其中 $s(a_i, b_j)$ 为替换得分（来自计分矩阵），g 为 gap 罚分；

3. 回溯：从矩阵右下角开始按最优路径回溯至左上角，还原最优比对方案。[^14]

Smith-Waterman 局部比对

核心差异：在矩阵填充时，将负值归零：[^7]

$$H(i,j) = \max\begin{cases} H(i-1,j-1) + s(a_i, b_j) \\ H(i-1,j) + g \\ H(i,j-1) + g \\ 0 \end{cases}$$

回溯从矩阵中得分最高的格子开始，到值为 0 的格子结束，得到最优局部比对。[^16]

C2 序列比对点阵图（Dot Plot）方法应用

2.1 CEACAM5 内部相似性分析

CEACAM5（癌胚抗原，CEAM5_HUMAN）是结直肠癌的经典肿瘤标志物，属于免疫球蛋白超家族。其成熟蛋白由 1 个 IgV 样 N 端结构域和 **6 个 C2 类 Ig 样结构域（A1-B1-A2-B2-A3-B3）**组成，总长 702 个氨基酸。这些重复结构域源于古老的基因重复事件。[17][18][^19]

点阵图操作步骤（以 DotMatcher 或 Dotlet 为例）：

1. 从 UniProt 下载 CEAM5_HUMAN 序列（UniProt ID: P06731）；
2. 在 DotMatcher（EMBOSS 套件）或 Dotlet（在线工具）中，将 CEAM5_HUMAN 序列与自身作图（Self Dot Plot）；
3. 适当调整窗口大小（window size，建议 10-20 aa）和阈值（threshold）以减少噪音；
4. 观察结果图像。

预期结果：

- 主对角线：表示序列与自身完全匹配；
- 平行于主对角线的次对角线：表示序列内部存在重复区域，即各 Ig 样结构域之间的相似性。CEACAM5 含有 6 个 C2 型 Ig 重复域，因此应看到多条等间距的次对角线，每条对应一对相似的重复单元。[^20]

2.2 Dotlet 点阵图方法用途总结

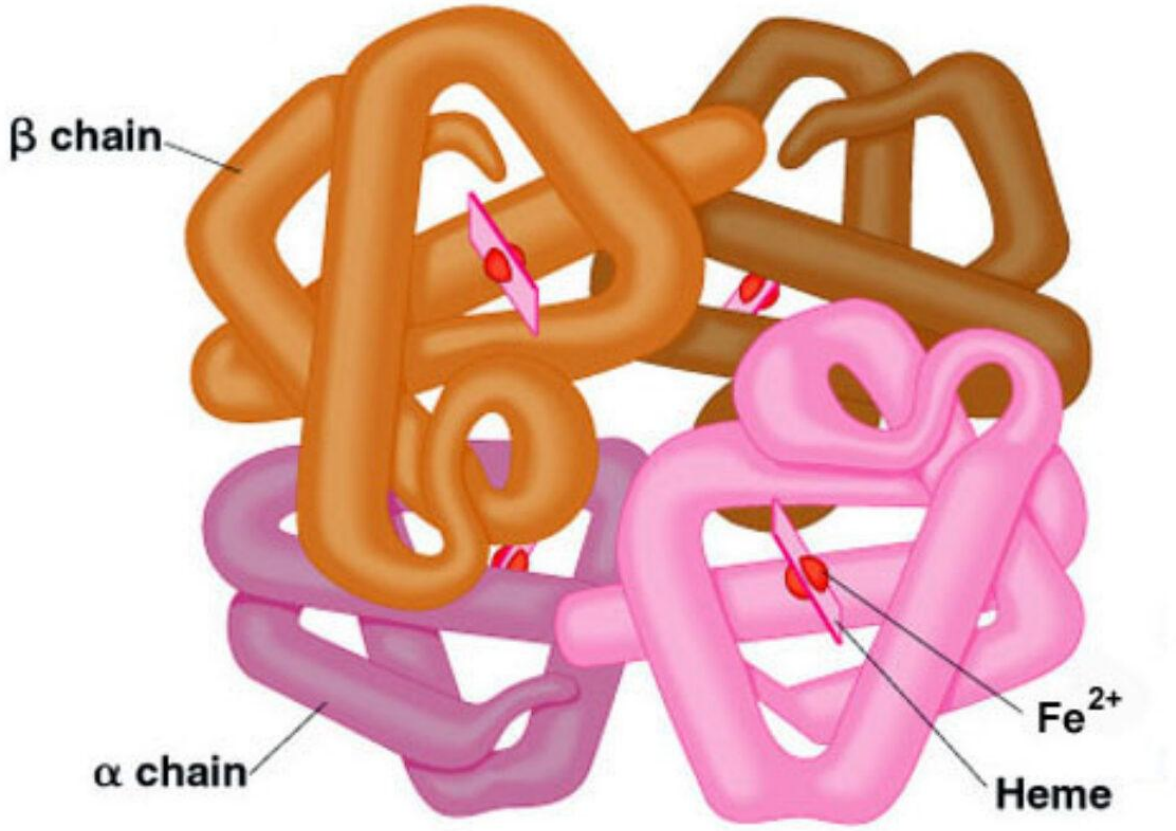
点阵图方法可直观揭示以下序列特征：[20][21]

图像特征	生物学含义
主对角线清晰连续	两序列（或自身）高度相似
平行于对角线的次对角线	序列内部存在直接重复（如 CEACAM5 的 Ig 重复域）
垂直于对角线的反向线	序列中存在反向互补区域或倒位（inversion）
对角线中断/跳跃	序列插入（垂直跳跃）或缺失（水平跳跃）
对角线在某段完全消失	高度分歧区域
多个分散的小点（噪声）	低复杂度序列或参数设置不当

Dotlet 支持调节窗口大小和匹配阈值，增大窗口可以降低背景噪声，更清晰地显示真正的重复区域。

C3 血红蛋白序列比对实例

3.1 血红蛋白基本知识



Hemoglobin structure diagram

血红蛋白（Hemoglobin, Hb）是脊椎动物红细胞中负责携带氧气的蛋白质。成人血红蛋白（HbA）是一个 $\alpha_2\beta_2$ 四聚体，由 2 条 alpha 亚基和 2 条 beta 亚基组成。每条亚基包含 1 个血红素（heme）基团，铁离子（Fe^{2+}）与 O_2 可逆结合。[22][23][^24]

血红蛋白的氧结合曲线呈 S 型（协同效应），当一个亚基结合 O_2 后，蛋白构象从 T 态（紧张态，低亲和力）向 R 态（松弛态，高亲和力）转变，促进其余亚基结合 O_2。2,3-DPG 等别构效应物可调节氧亲和力。[^25]

人 alpha 球蛋白基因座位于染色体 16p13.3，包含 7 个基因座：5'-HBZ-HBZP1-HBM-HBAP1-HBA2-HBA1-HBQ1-3'。HBA1 和 HBA2 两个基因编码区序列几乎完全相同，仅 5' UTR 和内含子略有差异，两者是旁系同源基因。[26][27][^5]

3.2 物种演化关系与分歧时间

根据 Ensembl 物种树，人（*Homo sapiens*）、小鼠（*Mus musculus*）和大鼠（*Rattus norvegicus*）的演化关系为：

- 人属灵长目（Primates），小鼠和大鼠属啮齿目（Rodentia）；
- 灵长目与啮齿目同属 Euarchontoglires 超目，分歧发生于约 87-90 百万年前（Mya）；[28][29]
- 小鼠与大鼠均属鼠科（Muridae），两者分歧时间约为 12-23 Mya（不同研究略有差异，化石校准建议为约 12-14 Mya）。[30][31]

物种对	分歧时间	数据来源
人 vs 小鼠	~87 Mya	TimeTree 文献汇总[28][29]
人 vs 大鼠	~87 Mya	（与人-小鼠相近，鼠科内部分歧远晚于此）
小鼠 vs 大鼠	~12-23 Mya	化石与分子估算[30][31]

3.3 人、小鼠、大鼠血红蛋白 alpha 亚基蛋白质比对结果分析

操作：从 UniProt 分别提取：

- 人 HBA1/HBA2：P69905（两者蛋白序列相同，均为 141 aa）
- 小鼠 Hba-a1/Hba-a2：P01942
- 大鼠 Hba1：P01946

进行双序列全局比对（Needleman-Wunsch，BLOSUM62 矩阵）。

预期结果：人与小鼠/大鼠 alpha 血红蛋白蛋白序列相似性约 87%；三者比对后大部分区域保守，尤其是铁结合的组氨酸残基和疏水核心高度保守。[32][33]

结果与预期不符的可能原因：

1. 旁系同源污染（Paralog contamination）：人有 HBA1 和 HBA2 两个旁系同源基因，小鼠有 Hba-a1 和 Hba-a2，大鼠同样有多个 alpha 球蛋白基因。若从数据库中未选择正确的直系同源基因，而误用了旁系同源基因比对，会导致比对结果偏差；[5][34]

2. 物种特异性 alpha 球蛋白基因家族差异：小鼠 alpha 球蛋白基因位于 11 号染色体，人位于 16 号染色体，染色体位置不同但功能同源。Hardison（2012）指出 alpha 样球蛋白基因簇相对稳定，但确有物种特异性的基因丢失和复制事件；[35][36]

3. 比对参数影响：若使用 PAM250 等远缘矩阵，可能引入过多 gap，使近缘比对结果显示不自然的差异；

4. 进一步分析思路：明确区分直系同源（用 ENSEMBL 物种树确认）与旁系同源；使用 BLOSUM62 进行蛋白比对；使用 Phylogenetic 分析验证进化关系。

3.4 编码区核苷酸序列比对分析

操作：从 RefSeq 提取编码区（CDS）序列：

- 人 HBA1：NM_000558（编码 141 aa，CDS 约 426 bp）
- 小鼠 Hba-a1：NM_008218（GenBank 标注为 Hba-a1）
- 大鼠 Hba1：NM_013096

进行双序列全局比对，推荐参数：

- 计分矩阵：NUC4.4（核苷酸替换矩阵）；
- 空位罚分：gap open = -10，gap extend = -0.5（仿射罚分，避免过度引入 gap）；

不同空位罚分的比对结果差异：

- 严格罚分（gap open = -15，extend = -1）：比对中几乎不引入 gap，相似区域清晰，但可能错过真实的 indel；
- 宽松罚分（gap open = -5，extend = -2）：引入较多 gap，可能揭示两物种间的 indel 事件，但也可能引入错误比对；
- 在比较人与啮齿类（分歧约 87 Mya）时，核苷酸序列同义位点替换已接近饱和，建议结合氨基酸比对和非同义位点分析。

3.5 alpha 血红蛋白基因家族演化（结合 Hardison 2012 和 Burmester 2002）

根据 Hardison（2012），alpha 样球蛋白基因演化历史如下：[35][37]

- 原始球蛋白基因经多次基因重复事件，产生 alpha 球蛋白、beta 球蛋白、肌红蛋白（myoglobin）等基因家族；
- **Burmester 等（2002）**发现了第四类脊椎动物球蛋白——细胞球蛋白（Cytoglobin, CYGB），在人体所有组织中广泛表达，190 个氨基酸，系统发育分析显示其与肌红蛋白存在共同祖先；[38][39]
- 人的 alpha 球蛋白基因簇位于染色体 16p13.3，包含功能基因 HBA2、HBA1（两者编码序列相同，可能经基因转换 gene conversion 事件均质化）及多个假基因；[5][40]
- 小鼠 alpha 球蛋白基因位于 11 号染色体（含 Hba-a1、Hba-a2 功能基因），同时在 17 号染色体上存在一组 alpha 样假基因簇，后者可能代表与人 16p13.3 的共线性断裂区域；[^41]
- 相比 beta 球蛋白基因簇（在不同物种间高度动态，存在频繁的扩张/缺失），alpha 球蛋白基因簇相对稳定，但仍存在物种特异性差异；[^35]
- 啮齿类（小鼠、大鼠）演化速率较高，氨基酸替换速率约为非啮齿类的两倍，这是导致人-啮齿类比对相似性低于预期的原因之一。[^30]

C4 课题相关蛋白质和编码基因

4.1 实验室及研究方向简介

本人所在实验室（或导师研究方向）从事肠道微生物组学与宿主健康相关研究，利用宏基因组学、宏转录组学等多组学方法，分析肠道菌群的结构变化与特定疾病（如结直肠癌、炎症性肠病）的关联，并通过 PanGenome 分析揭示菌株水平的基因组多态性。

4.2 研究课题背景

本人课题聚焦于结直肠癌相关肠道菌群的泛基因组结构变异分析，研究目标是利用 sgvFinder2 等工具鉴定肠道细菌（如 *Acinetobacter baumannii* 及其他致病菌）的结构变异（SV），揭示菌株水平的功能基因多态性及其与宿主疾病状态的关联。这一工作对于理解肿瘤微环境中菌群的功能适应性具有重要意义。

4.3 CEACAM5 蛋白序列比对分析

选取课题背景相关的人类癌胚抗原 CEACAM5（P06731, CEAM5_HUMAN） 进行跨物种同源蛋白比对。CEACAM5 是结直肠癌的经典肿瘤标志物，在肿瘤细胞中高度表达，参与细胞黏附、分化抑制及凋亡调控。[42][43]

UniProt 序列提取（推荐物种及 ID）：

- 人 CEACAM5：P06731（702 aa）
- 小鼠 Ceacam5：Q9R0E2
- 恒河猴 CEACAM5：UniProt 搜索 CEACAM5 Macaca

比对参数建议：BLOSUM62，gap open = -11，gap extend = -1（BLAST 默认）。

预期比对结果分析：

- 人与小鼠 CEACAM5 蛋白序列相似性约在 70% 左右（Ig 结构域保守，连接区变异较大）；
- N 端 IgV 结构域（负责同型/异型结合）在各物种间相对保守；
- 6 个 C2 型 Ig 重复域存在内部相似性（点阵图可见次对角线），提示其起源于串联基因重复；[^19]
- 与微生物某些蛋白的局部相似性可能揭示宿主-病原体共进化关系。

4.4 CEACAM5 编码区核苷酸比对

从 RefSeq 提取 CDS：

- 人 CEACAM5 CDS：NM_004363（约 2109 bp）
- 小鼠 Ceacam5 CDS：相应 RefSeq accession

比对分析要点：

- 同义位点（synonymous sites）替换率 dS 高于非同义位点（non-synonymous sites）替换率 dN（dN/dS < 1），说明 CEACAM5 整体处于净化选择（purifying selection）；
- 某些外显子（对应高度可变的 Ig 结构域）可能存在正向选择，反映病原体-宿主共进化压力；
- 核苷酸比对的空位罚分应严格设置（gap open ≥ -10），避免不合理的外显子边界跨越 gap。

4.5 序列比对对课题研究的参考价值

1. 确认保守结构域：通过人-小鼠 CEACAM5 比对，明确蛋白质功能保守区域，为后续功能实验设计提供依据；

2. 识别多态性位点：蛋白序列比对揭示的物种间差异位点，可对应到疾病相关的 SNP 或 SV 位点，辅助解读 VCF 文件中的变异功能意义；

3. 验证同源性：在宏基因组泛基因组分析中，通过序列比对可区分同一基因家族内的旁系同源与直系同源拷贝，减少 SV 分析中的假阳性；

4. 功能预测：对菌株中发现的结构变异基因，通过序列相似性搜索（BLAST）可初步推断其功能，为宿主-微生物互作分析提供分子层面的证据。

5. 问题

Q1：序列比对的结果与预期不一致时怎么办？

首先排除技术问题，看序列是不是选错了，参数设置是否合理，然后考虑生物学解释，最后可以通过多序列比对或者构建系统发育树来分析结果。