Lab1b 基础BLAST（blastn）

用于基因和基因组功能注释的最重要的生物信息学策略之一是比对基于未表征的基因或蛋白质与具有更好功能注释的序列的相似性来预测未表征的基因或蛋质的功能。 BLAST可能是查找与目标查询序列相似的数据库序列的最重要的工具之一。

Box3. BLAST and Homology
BLAST与同源性

基本本地比对和搜索工具（BLAST）是一种非常强大的方法，用于识别与查询序列具有本地相似性的数据库序列（有关定义，请参见下文）。 使用BLAST时，通常会遵循生物学研究中使用的非常重要的假设：

同源基因具有相似序列

> 直系同源基因在多个物种中具有最高的相似性

>> 直系同源基因很可能具有相似的功能

>>> 因此，多个物种之间最相似的序列具有相似的功能

请注意，了解这些只是假设是非常重要的，并且有许多原因和实例证明这些假设是错误的。 但是，它们是一个合理的起点。

定义：

相似序列（Similar sequences）:共有大量残基（核苷酸或氨基酸）的序列。 由于同源性或偶然性，序列可能相似。 序列之间的相似性越高，它们越可能同源。

同源序列（Homologous sequences）:通过共同祖先相关的序列。 同源性是定性的-两个序列在共同祖先之间是相关的，也可能是不相关的。 同源序列的相似性水平差异很大，从100%到0%。

直系同源序列（Orthologous sequences）:通过过去物种形成事件关联的序列，假定直系同源序列共享共同的功能。

旁系同源序列（Paralogous sequences）:通过过去的基因复制从而产生相关的序列。 基因在复制后通常在功能上有所不同。 因此，不假定旁系同源序列具有共同的功能。

查询顺序（Query sequence）:关于你感兴趣的序列的顺序。

高得分片段（High Scoring Segment Pair）（HSP）:数据库的“命中”。用于描述 你的查询序列和BLAST返回的数据库序列之间的子序列匹配程度。

局部比对（Local alignment）:仅在部分序列上延伸的序列比对。

全局比对（Global alignment）:在全序列上延伸的序列比对。

我们将在本章学习如何使用BLASTn。

1. 首先，我们需要一个查询序列进行搜索。 让我们再次从给定的基因开始，但是这次我们将对应于蛋白质序列的核苷酸序列，而不是蛋白质序列。 首先，尝试再次使用GQuery查找基因的DNA序列。
    a. 在search NCBI（GQuery）功能界面，通过登记号并搜索“All Database”来获得我们想要的蛋白质序列。我们将搜索**NP_001318308**以便使用本实验第一部分中的蛋白质。
    b. 出现的第一页是摘要页面。 进入此页面后，你可以移至感兴趣的数据库。 在这种情况下，你可能没有太多数据库中的匹配，因为将进行非常具体的搜索。

c. 尝试单击“**Gene**”链接。 基因页面是否仅给你基因序列? 你得到了什么呢? 请注意，当你用鼠标指针悬停在基因的图形上时，将弹出上下文相关的链接菜单。 您可以单击表示基因外显子的绿色框，以获取与该基因相关的各种序列和分析的链接。 请注意，绿色轨道是mRNA和CDS轨道的组合-单击NM_或NP_号以查看绿色轨道的展开。

d. 单击右侧“Related Information”面板中的RefSeq RNA链接。 这将带您到编码您一直在寻找的蛋白质的mRNA。 请注意记录中的功能列表。 GenBank记录中的一个特征是基因（Gene），它对应于该记录的基本位置在1 – 1949行。 另一个功能是编码序列（CDS），它对应于基本位置的33 – 1781行。

e. 在Sequence Viewer面板上方（即主界面/步骤c中的Genomic regions, transcripts, and products)中，单击“Go to nucleotide：Genbank”链接。 您将被带到编码您正在寻找的mRNA的基因组区域。 请注意，基因特征如何对应于位置1 – 2201，而mRNA特征如何对应于位置1至1296，1383至1832，1916至2032和2116至2201，而CDS特征对应于核苷酸位置169至1296、1383至1832、1916至2032和2116至2169。(tips: 为什么这d中的位置不同?)

f. 让我们返回以前使用的mRNA记录（NM_001336190）。 单击CDS链接。 现在，你正在查看编码序列的信息，而不是整个基因或蛋白质（以棕色突出显示）。

g. 使用页面底部灰色栏中的“Display：FASTA”选项可生成FASTA格式的CDS版本。

h. 现在，你将拥有最基本且易于管理的格式（FASTA格式）的序列。 FASTA格式只是标题行，以“>”开头，后跟描述序列的文本，然后实际序列从下一行开始。 该序列可以是DNA或蛋白质，并且可以是连续的（从页面上滚动下来），或切成通常在60-80个残基之间的更易控制的长度。

2. 让我们做一些BLAST！ 使用网页“Analyze This Sequence”部分中的“Run BLAST”链接。可以或在浏览器中打开一个新的选项卡或窗口，然后返回NCBI主页，从顶部的Resources下拉列表中的DNA＆RNA小节中选择BLAST。 由于我们的序列是核苷酸序列，因此我们想进行核苷酸BLAST，即BLASTn（Nucleotide BLAST）。(www.ncbi.nlm.nih.gov)
这里有很多选择。 我们将在下一个实验中讨论其中一些，但现在让我们以最简单的方式进行工作。 由于我们的序列是核苷酸序列，因此我们想进行核苷酸BLAST。

3. 在“ BLAST”页面上，请注意，在“输入查询序列”部分下，NCBI系统已自动载入了登录号（但您也可以输入一个GI号或FASTA序列）和子范围（我们将仅使用编码序列进行搜索mRNA序列的一部分）。 您也可以将1-h中找到的FASTA格式的CDS序列复制并粘贴到查询框中，而无需定义子范围-此时，您应该清楚mRNA序列与编码序列之间的区别……

a. 浏览页面的各个部分。 您对算法的运行方式有很多控制权（尤其是单击底部附近的[+]**Algorithm parameters**时，尤其如此）。

b. 我们要查询完整的NCBI数据库； 由于我们的序列不是人类，因此NCBI链接系统已自动将默认的**Database**（人类）更改为**Other**和**Nucleotide collection (**nr / nt)。nr数据库是GenBank中序列的非冗余集合。

c. 将“**Program Selected > Optimized for**”更改为“**Somewhat similar sequences (blastn)**”。

d. 注意页面周围的所有小问号图标。 单击其中任何一个，以查找有关关联参数的更多信息。 例如，通过单击“**Program Selection**”部分中的问号，您将获得有关不同方法的非常简短的摘要。 通过单击更多，您可以跳到包含该算法的完整文档的新页面。

e. 打开底部附近的算法参数界面(**Algorithm Parameters**)：

   i. 什么是期望阈值（**Except threshold**）？

   ii. 如果减少它会发生什么? 增加呢?

   iii. 增加子序列长度（**Word size**）大小会造成什么影响?

   iv. 为什么会有低复杂度区域（**Low Complexity regions**）过滤器? 我们应该启用这个功能吗?

f. 确保在输入框中输入了查询顺序，然后选中靠近“BLAST” 按钮的“Show results in a new”旁边靠近“BLAST” 按框。 最后单击“BLAST”按钮。

g. 当BLAST运行时或搜索完成后，您可以选择通过单击“**Format options**”链接来调整搜索结果的格式。 我们现在不会执行此操作，因为默认设置通常可以正常工作。

Box4. How Good is My Hit?
我的匹配率好吗?
BLAST HSP的质量可以通过多种不同方式进行量化。 了解这些指标之间的差异并使用适当的指标很重要。
Identity（同一性）:两个序列不变的程度。这是一个非常差的衡量标准，因为它没有考虑到序列关系的微妙之处（例如，两个序列中高度保守的域的一小块区域之间的差别）。
Bit Score（位分数）:对齐分数(S)。 在所采用的特定评分系统上标准化的非常精确的度量。但依赖于查询长度的缺点。
请注意，了解这些只是假设是非常重要的，并且有许多原因和实例证明这些假设是错误的。 但是，它们是一个合理的起点。
E value（期望值）:是指在随机情况下，获得比当前比对分数相等或更高分数的可能比对条数。E值越低，分数越显着，匹配效果越好。这是最好的度量标准，因为可以轻松比较同一数据库中不同搜索的结果。请注意，E值取决于数据库的大小（n）和查询序列的长度（m）。包含相同命中序列的不同数据库中搜索的相同序列将导致报告不同的E值。

4. 结果页面分为几个部分：

   a. 最上面是结果摘要，它仅显示有关查询和搜索的数据库的详细信息。 您可以通过单击“**Search Summary**”找到有关搜索的更多详细信息。

b. 接下来是四个并列的区域，首先是描述（**Descriptions**）界面，这里有介绍（**Description**）区，是各种相关片段比对的链接；然后是最大得分（**Max Score**）区，这是二进制得分得出的原始比对分数S；接着是总分（**Total Score**），如果堆在其他比对分数，那么总分可能和最大得分有所不同；查询范围（**Query Cover**）是查询片段与命中片段相似度的百分比；**E-value**是匹配质量的衡量准则，如果为0那么匹配最好；登记号（**Accession**）用于链接至NCBI的指定序列。

c. 然后是图形区域（**Graphic Summary**），可以通过在图上点击不同颜色的片段来查看他们的E-value等。

d. 接着是HSP对齐区域（**Alignments**）。将针对首次HSP对齐提供的信息与图形摘要和HSP摘要中的第一项进行比较，向下滚动路线时，您会看到匹配质量下降，也就是E-value增加，这里的**Stand = plue/plus**指序列是同向匹配

e. 最后是分类学区域（**Taxonomy**），可以在这里观察到我们的遗传片段在分类学上的位置。

f. 返回HSP对齐区域，在**Alignment view**区域将“**Pairwise**”改为“**Query-anchored with dots for identities**”（用点做特征进行固定查询）。观察发生了哪些变化，之后再回到图形区域，点击得分较低的蓝色代码块，看看他们的特征。