# Lab3 多重序列比对

需要下载MEGA X软件：https://www.megasoftware.net/

在本实验中，我们将学习如何比对来自不同生物的核苷酸序列，以识别它们的相似性和差异性。 多序列比对（MSA）是一种非常强大的方法，是许多生物信息学分析的基础。它可以与广泛应用的进化和功能分析结合使用，以识别从进化关系到功能图案和相互作用域的所有内容。 可靠的MSA可以识别相关序列之间的同源残基，这些数据对于系统发育分析和基序识别绝对必不可少。

有许多方法可以进行MSA。幸运的是，生成MSA非常容易。但不幸的是，产生不良的MSA通常与优质的MSA一样容易，不良的MSA可能削弱或完全损害您的分析。在本实验中，您将熟悉一些可用的MSA应用程序。 此外，您还应该了解，产生良好的MSA是一项有挑战性的技能，并且除简单的即插即用方法外，通常还需要更多。 尽管所有优秀的MSA应用程序都将为您提供“最佳”解决方案，但通常，一些手动调整和生物学直觉将产生更好的对齐方式。

MSA主要分为两类：全局对齐和局部对齐。你需要熟悉这两者之间的区别，并使用最适合你的特定需求的应用程序。 对MSA的最大误解是以为有一种工具对所有工作都同样有效。例如，大多数人只是将Clustal用于其MSA。尽管Clustal是功能强大的应用程序，但它对于大量的对齐问题完全不适合。

Box1. Global and Local Alignments
全局与局部比对

全局比对可以比对那些要比对的序列集在其整个长度上相似的序列。重要的是要认识到，尽管全局比对假定你可以对整个序列进行比对，但它们确实允许存在间隔和长度差异（即如当某些序列比其他序列长或短时）。

局部比对是比对仅在序列的特定子区域具有相似性的那些片段。 例如，在保留部分序列时，使用局部比对，而其余序列相差太大，以至于没有相似性。

下图（从MAFFT网站改编）很好地说明了全局和局部对齐方式之间的差异。 在每个图中，“ X”表示可对齐的残基，“O”表示不可对齐的残基，“N”表示间隙，在这种情况下，全局比对是合适的，因为序列的整个长度（从一端到另一端）是相似的：

XXXXXXXXXXNXXXXXXXXXXXXXXX
XXNXXXXXXXXXXXXXXNXXXXXXXX
XXXXXNNNNXXXXXXXXNNNXXXXXXX

当使用大多数良好的全局比对应用程序时，即使序列的长度非常不同，全局比对也将在以下情况下起作用：

OOOOOOOOOXXXXXXXXXXNXXXXXXXXXXXXXXNNNNNNNNN
NNNNNNNNNXXNXXXXXXXXXXXXXXXNXXXXXXXXOOOOONNNNN
NNNNNNNOOOXXXXXNNNNXXXXXXXXNNNXXXXXXXOOOOOOOONN

以下序列将需要局部MSA方法。 请注意，每个序列中都有可以对齐的子区域，但整个序列却不能。 此外，并非在所有序列中都找到这些保守区：

OOXXNNNXXXNNNNNNNNNXXXXXNXXXOOOO
NNXXXXXXXXOOONNNNNNNXXXXXXXNXOOOO
NOXXXXXNXXOOOOOOOONNXXNXXXNXXNNNN

**Clustal**

Clustal是迄今为止最受欢迎的MSA应用程序。这是一个全局对齐工具，几乎可以在任何计算机平台上运行。它也可以通过许多Web服务器获得，并且已集成到各种生物信息包中。我们将在MEGA X中使用Clustal实现，这是一个非常强大且易于使用的系统发育软件包。

1. 从Coursera网站下载包含未比对核苷酸序列的FASTA文件（使用“这些序列”链接从“生物信息学方法I”选项卡（与检索此实验文档的位置相同的选项卡）下载）...请注意，如果您使用的是Mac 用户，您将需要删除一个.txt扩展名，该扩展名将自动添加，以使文件名仅以“ ... Labs3,4_sequences.fas”结尾，否则Mega无法识别该文件）。 您始终可以使用简单的文本编辑器程序查看任何基于文本的文件。 我们建议对正在使用的任何文件执行此操作，以确认该文件包含正确格式的所需信息。 深红或Sublime编辑器或其他工具（请参阅附录1）在此类情况下非常方便，并且提供了记事本中未提供的一些强大功能。

2. 序列已经由我们在较早的实验中探索过的NCBI下载系统提供的相当长的标题改为了新的标题，格式为以下格式，以便标题信息足够紧凑，可以在MEGA的比对应用程序中完全显示：

   >Genus_species_GI
   Sequence…
   举例，我们把：
   >gi|42567417:17-484 Arabidopsis thaliana ribosomal protein L11 family…
   ACTGTCTA…
   改为
   >A_thaliana_42567417
   ACTGTCTA…

   这个看似耗时的步骤说明了具有一些基本编程能力的重要性。如果你打算进行任何严肃的生物信息学研究，则可以花一点时间来学习一种良好的脚本语言，例如Python。几分钟的简单脚本编写即可自动完成此任务，并在几分之一秒内完成工作。 另外，有时可以使用优质文本编辑器的搜索和替换功能使用正则表达式来完成此类任务：有关一些提示，请参见附录1

3. 打开MEGA X的对比编辑器
   a. 选择ALIGN键
   b. 选择**Edit/Build Alignment**，然后选择**Retrieve sequences from a file**去加载你的文件
   c. 滚动到末尾，探索序列。 请注意，你可以使用增加Excel或其他电子表格程序中表格列的大小的相同方法来增大名称字段（在窗口的左侧）。

d. 单击序列上方的“**Translated Protein Sequence**”选项卡（如果MEGA询问您是否要使用当前的标准遗传密码，请选中“是”）。同时将未比对的蛋白质序列保存在Data/Export > Fasta format下（保存前在.fas扩展名之前的名称中添加“ prot”之类的名称），以便稍后在实验室中用于MAFFT。

4. 回到DNA sequences并开始进行Clustal比对：
   a. 选择用来进行比对的序列，通过**Edit/Select All**全选。
   b. 点击**Alignment / Align by ClustalW**
   c. 接下来查看DNA比对参数，我们借此复习这些参数的含义：如何理解空位开启罚分(**Gap Opening Penalty**)和空位延伸罚分(**Gap Extension Penalty**)的含义?
   d. 暂时退出此菜单

5. 让我们回到**Translated Protein Sequences**进行实际比对，然后单击**Alignment / Align by ClustalW**进行比对。
   a. 通过**Edit / Select All**全选所有序列
   b. 选择**Alignment/Align by ClustalW**
   c. 将**Pairwise alignment**和**Multiple alignment**中的“**gap opening penalty**”全都修改为20
   d. 点击**OK**去运行
   e. 你可能需要再次单击对齐的序列才能看到颜色编码
   f. 你可以通过“**Display / Font**”查看更多的对齐方式，并减小字体大小。
   g. 请注意，一旦你比对了翻译的蛋白质序列，相应的DNA序列也会被比对，您可以通过单击序列上方的选项卡在这些视图之间来回切换。

滚动浏览DNA和翻译的蛋白质比对，注意一致的彩色条，它们表示保守的残基：
   a. 如果你查看比对的翻译蛋白序列，您会发现一些比对柱都是单色的，但是有许多不同的氨基酸。
   b. 通常，Clustal使用以下表示法标识比对列的保守性：* 和:和。 分别表示完美守恒，强守恒和弱守恒。 但应注意并非所有版本的MEGA都遵循此约定。

6. 让我们再次比对翻译后的蛋白质序列。 返回“**Alignment / Align by ClustalW**”菜单，并将gap open和extension penalties分别设置为100和0.1。 重新运行比对并观察它们与之前的区别。

7. 现在我们需要保存这些比对。 请注意，你将根据当前正在查看的选项卡保存DNA或蛋白质比对。 您可以FASTA或MEGA格式导出序列。对于简单数据，MEGA格式非常类似于FASTA格式，但是允许存储有关序列的更多信息。

a. 将我们最好的核酸比对结果导出为Lab3_dna_clustal.fas，同时将我们最好的蛋白质比对结果保存为Lab3_pro_clustal.fas。
b. 你也可以使用.meg文件扩展名以MEGA格式进行相同的操作。 以MEGA格式保存时，系统会提示您输入标题。 这可以是文件中数据的任何描述。

8. 关闭MEGA。将DNA序列翻译成蛋白质序列可能是增加比对能力的一种很好的方法。 通常，如果您想继续使用DNA序列，只需将其翻译为蛋白质，进行比对，然后再次反向翻译为DNA。 MEGA无缝执行此操作，因此功能非常强大。 但值得注意的是你不能总是这样做。
   a. 你注意到每次迭代中黄色新序列的变化了吗?
   b. 从理论上讲，高分的PSI-BLAST命中应该具有哪些品质?
   c. 你能猜出为什么你感兴趣的序列中有原始分数在以后的搜索中比阈值低的情况发生吗? 同样，如何移除那些阈值高但你不感兴趣的序列呢?

**DIALIGN**

DIALIGN是一种非常强大的局部比对MSA，它特别擅长识别嵌入在较长的不可比对区域中的短保守基序。 该程序具有许多不错的功能，包括DNA的自动翻译和反向翻译为蛋白质序列。 DIALIGN有许多不同的版本，在某些情况下，该程序将在每个比对列下方显示一个标准化的分数，代表该比对列的质量，并明确向您显示哪些残基已实际比对，以及哪些残基差异太大而无法可靠 对准。 您可以通过多种方式访问DIALIGN。 您可以在本地加载和运行它，也可以通过许多Web界面访问它。

1. 前往 http://dialign.gobics.de/ 并选择CHAOS-DIALIGN。

2. 从上方上传FASTA格式的未比对的核苷酸序列，输入电子邮件地址，然后单击**Run CHAOS + DIALIGN**。 （请注意，虽然Dialign可以处理蛋白质序列，但此特定服务器旨在有效地处理长DNA序列）。

3. 前往邮箱中点击链接，然后单击DIALIGN格式的“Full alignment”链接

4. 您可以以多种格式下载对齐文件，包括fasta。 DIALIGN格式具有两个重要且方便的功能：
   a. 大写字母表示比对的残基，即比对所组成的“片段”中至少一个所涉及的残基。 小写字母表示不属于这些所选“片段”中任何一个的残基。 DIALIGN不认为这些已对齐。 因此，如果一个小写字母与其他字母位于同一列中，则这是完全的机会； 这些残基不被认为是同源的。

b. 比对下方的数字大致反映了序列之间局部相似性的相对程度。 对数字进行归一化，以使每个位置在比对中具有最大相似度的区域得到一个介于0和9，9表示序列之间有最大的相似度。

**MAFFT**

MAFFT是一种功能强大，用途广泛且快速的应用程序，似乎可以解决几乎所有MSA问题。 与其他方法相比，它具有许多非常显着的优势。 MAFFT具有足够的通用性，能够比对序列非常，能够对比序列非常长和非常多的数据集。最值得注意的是，它能够自动调整其特定的算法选择，从而为每个数据集提供最合适的MSA。 如MAFFT网站上所述，MAFFT提供了多种多重比对策略，可将其分为三种类型：[1]渐进式（类似于Clustal），[2]使用客观评分功能评估MSA的质量的迭代细化，[3]使用更复杂的评分功能进行迭代优化，该评分功能特别擅长处理序列缺口（插入和缺失）。 通常我们要在速度和准确性之间要进行权衡。 速度的顺序为1> 2> 3，而精度的顺序相反。

MAFFT还可以选择包含同源序列，以提高比对的准确性。 这称为Mafft-homologs，在MAFFT网站的此图中对此进行了很好的描述。

1. 前往MAFFT的网站http://mafft.cbrc.jp/alignment/server/，你可以从实验室的第一部分上传未比对的蛋白质序列文件，也可以将fasta序列直接复制并粘贴到窗口中。
   a. 选择**Strategy**下的**Auto**。
   b. 在**Parameters**选项中将**Gap opening penalty**选项改为3，将**Offset value**改为0.2。
   c. 不要修改任何**Mafft-homologs**中的选项。
   d. 提交未对齐的蛋白质文件并运行。
2. 请注意，MAFFT为您提供了类似Clustal的输出以及底部的列保留指示符。对齐的FASTA格式显示在下面，然后在下面显示使用的特定方法。 您还可以通过页面顶部的链接以多种格式保存数据，以及构建系统发育树。 MSAViewer和Jalview是用于多个序列比对的出色查看工具。要查看MAFFT质量得分，请查看Jalview Desktop视图-您可能需要使用配置Java工具将http://mafft.cbrc.jp添加到“例外站点列表”中。我们将在下一个模块中构建发育树。
3. 在不关闭前一个窗口的情况下，打开一个新窗口或一个选项卡，然后像以前一样返回到MAFFT输入页面。
   a. 其他设置不变，将**Mafft-homologs**打开 (on)，并选择**show homologs**，将**number of homologs** 设为100。
   b. 运行
4. 在不关闭前一个窗口的情况下，打开一个新窗口或一个选项卡，然后像以前一样返回到MAFFT输入页面（复读）。

a. 其他步骤相同，但在**Strategy**选项中选择**FFT-NA-1**，同时在**Parameters**选项中将**Gap opening penalty**选项改为1，将**Offset value**改为0。
b. 关闭**show homologs**。
c. 运行

5. 使用参数设置进行更多对齐，以了解它们如何影响对齐。 请记住每次都打开一个新窗口或标签。
6. 解释不同的MAFFT结果，单击页面顶部附近的“查看”按钮以打开指向MSA查看器选项的链接。 MSAViewer可以可靠地工作，而不必担心Java设置。
7. 你可以通过MAFFT输出页面（页面顶部附近的选项）或通过Jalview的“文件”菜单以各种序列格式保存比对。 保存对齐方式时，最好使名称具有信息性，以使您知道使用了哪种算法，例如 将输出保存为“ Lab3_pro_MAFFT- FFT- NS-i.fas”之类的名称。

结束