Lab5 选择分析

上一个实验中，我们对同源细菌基因进行了比对，并创建了系统进化树，这使我们可以推断它们之间的关系以及它们可能如何进化。现在我们要看看自然选择对这些基因的影响。在本实验中，我们将尝试确定选择是否起作用，选择的种类以及序列如何因选择而改变。

我们为什么要关心自然选择的作用? 重要的是要了解你是否对研究基因功能有严格的兴趣? 自然选择是所有适应性变化的基础机制。基因，蛋白质及其相互作用物的有效功能在很大程度上是由于自然选择的作用。因此，我们不仅可以通过自然选择的研究获得关于基因和蛋白质如何进化的深刻见解，而且还可以通过识别基因组上自然选择留下的模式来识别导致特定适应的遗传变化。

自然选择是非常简单的过程，由于遗传变异对携带它们的生物体适应性的影响，遗传变异的频率会改变。自然选择的关键是变异：自然选择只能在存在可遗传的遗传变异时起作用。其中一些变化可能是“好”，因此可能会增加频率，而其他变化可能是“不好”，因此可能会降低频率。“好”和“坏”不应被过分理解，例如，它可能仅意味着蛋白质的特定变体在特定环境中能更有效地发挥作用。实际上，相同的变体在不同的环境中可能无法有效发挥作用，因此自然选择不仅取决于可遗传遗传变异的存在，而且还取决于发现该变异的特定环境。自然选择也不应被视为某种行为实体。很简单，这就是遗传变异体向后代的不同生存和传播。

Box1. The Major Types of Natural Selection
自然选择的主要方式

自然选择可以多种不同的方式发挥作用。 下面列出了与本课程最简单和最相关的内容：

1. 当在种群中发生有益突变并增加频率时，就会发生正向选择。 当然，如果此突变的频率增加，则其他变化必须同时降低频率。 正向选择的经典例子是抗生素抗性等位基因通过暴露于该抗生素的细菌群体的传播。
2. 负向（纯化）选择与正向选择相反。当从种群中选择有害突变时，就会发生这种情况。由于大多数蛋白质已经经历了数百万或数十亿年的进化，因此大多数导致蛋白质编码序列发生变化的突变被认为是有害的。这些有害的突变将通过阴性选择从种群中清除。 对于任何蛋白质，有害的突变比例都与该蛋白质的“进化保守性”直接相关。
3. 平衡选择或多样化选择是有利于保持基因座遗传变异的选择。 正选择和负选择都清除变异（选择变异还是对抗变异），平衡选择实际上是通过选择多个遗传变异来维持变异。 想象这种情况发生的最简单方法是考虑选择不同蛋白质等位基因形式的多种环境的情况。 例如，一种受体蛋白在人群中存在病原体时非常有益，但在没有病原体时只会对系统造成负担。

我们如何检测和衡量自然选择? 对我们来说幸运的是，所有进化过程（包括自然选择，突变，重组，基因流动和遗传漂移）在基因组上都留下了它们的特征标记或足迹。 这些足迹中有一些是明显且持久的，而另一些则是模糊的和 / 或非常短暂的。 如果你知道寻找的内容以及寻找的位置，那么你通常可以重建遗传区域的进化历史。

有很多方法可以识别和表征自然选择留下的足迹。 为了我们的目的，我们将专注于最常见的测试以识别正向和负向选择。如前所述，正选择是针对增加适应性（有益于有机体）的遗传变化的选择，而负选择通常与进化保护相关，以保留基本功能。

dN / dS Ratio Test 选择压力计算

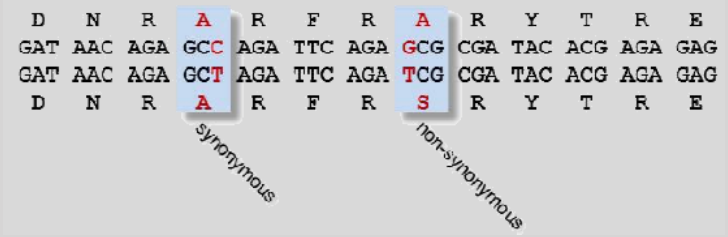
dN / dS比率测试可能是从核苷酸序列数据中检测自然选择模式的最广泛使用的方法。该测试特别有用，因为它可以推断选择一直作用到密码子级别。

Box2. Using dN and dS to Infer Selection

使用dN和dS的推断选择

dN / dS（也称为Ka / Ks或ω）测试计算非同义替换率（dN，每个非同义位点的非同义替换数）与同义替换率（ dS，每个同义位点的同义替换数）。 非同义取代是那些导致蛋白质序列改变的突变，而同义取代是那些由于遗传密码的简并性而改变DNA序列而不是蛋白质序列的突变。 请注意，我们对这些替换的速率感兴趣，而不是它们的绝对数量。

同义替换通常不会受到强烈的选择压力，因为它们不会导致蛋白质序列发生变化（尽管有证据表



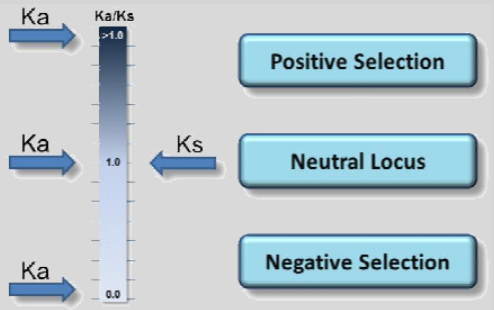
明，某些密码子比其他密码子更易发生有效翻译）； 因此，它们倾向于以大致恒定的速率积累。将此速率视为基线，通过该基线我们可以比较改变蛋白质序列的替代率（非同义替代）。

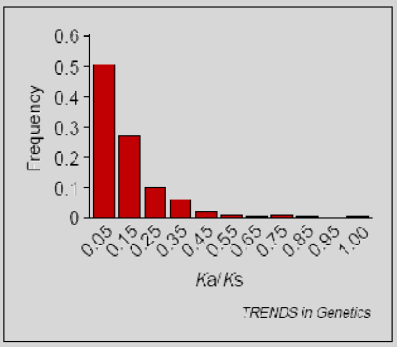
对于完全中性的序列（可以无限制自由更改的序列），你会期望dN与dS相同，或者dN / dS = 1。当序列存在选择性限制（负选择）时，可以预期会有更少的可改变蛋白质序列的取代或较低的dN。 因此，dN / dS <1。

在正向选择的情况下，您会期望在您的群体中看到更高比例的氨基酸替换（因为正选择会增加氨基酸替换），因此dN会更高； 因此，dN / dS> 1。

我们可以通过测量dN / dS比来确定基因是处于正选择还是负选择。 dN / dS> 1是积极选择的有力指标。 dN / dS <1是否定选择的有力指标。 理论上，中性序列的dN / dS =1。考虑这一点的最简单方法是，同义（无声）替换的速率应保持恒定，因为它不受选择的影响，而非选择的速率的上升或下降取决于分别选择改变氨基酸序列还是保持氨基酸序列相同。

大多数功能基因都处于一定水平的选择约束之下。 因此，大多数编码区域的dN / dS比通常远低于1.0：





我们还可以使用dN / dS测试来确定在单个氨基酸水平上起作用的选择。这对于识别功能上重要的区域可能很有用–你可以想象，如果宿主与病原体之间通过宿主蛋白质和病原体蛋白质相互作用介导的“军备竞赛”，则两者相互作用的表面可能处于正选择状态，从而其余的蛋白质可能处于负选择状态，以维持蛋白质的整体生化功能。 为了使这样的测试起作用，必须有一个合适的进化似然模型，该模型可以允许序列中的某些位点处于正选择状态，而大多数位点处于负选择状态。 Datamonkey和其他选择分析工具（例如PAML）使用了不同类型的模型。

我们将从一组植物病原体丁香假单胞菌的HrpZ基因的一组比对序列中进行研究。 HrpZ编码这种细菌的III型分泌系统的一个成分，这对于提供可破坏植物免疫系统的蛋白质非常重要。 在github的本文件中下载Lab5_Psy_hrpZ.fas获得本次课程的数据。

显然，使用根据氨基酸比对进行比对的DNA序列非常重要。为什么？如果不清楚，请参见Box2中的第一句。使用EBI的EMBOSS工具套件的TRANSEQ应用程序检查比对的DNA序列通常很有用。 转到http://bar.utoronto.ca/EMBOSS（本地实例），然后使用左上角的链接按字母顺序对功能进行排序，以找到transeq函数(在nucletic translation区内)。 将Lab5_Psy_hrpZ.fas文件中的序列粘贴到框中，然后选择第一个阅读框，以确认所有序列都在框架中并且不包含终止密码子。

我们将使用在线工具Datamonkey（http://www.datamonkey.org/）在序列中寻找选择。Datamonkey是一个非常直接且功能强大的工具，可用于进行复杂而复杂的进化分析。 这些分析是在远程服务器上进行的，因此您无需访问功能强大的工作站。Datamonkey使用HyPhy程序包，这是一个非常强大的多平台程序包，用于执行基于似然性的序列进化速率和模式分析。

可以在其网站上找到有关Datamonkey的大量信息，以及非常完整的帮助页面（http://www.datamonkey.org/help）。请注意，本实验参照了许多原始Datamonkey教程。

我们将分析丁香假单胞菌HrpZ基因的45个序列。 Datamonkey提供了一系列不同的分析，可以使用“方法”下拉菜单进行选择。我们将专注于本实验室的两个最直接的分析，即SLAC和FEL（如果您从Datamonkey主页开始，则进入FEL分析过程如下：Selection；Detect selection at: Sites；Detect: Pervasive selection；Dataset: Small ——你可以看到我们只是在本实验中略读了Datamonkey的功能），但我们还是鼓励你仔细阅读网站上提供的参考文献 / 引用来获得更复杂分析的信息。 SLAC（单似然祖先计数）和FEL（固定效应似然）方法显示出大致相同的效果，但建议对较大数据集（> 40个序列）使用SLAC。 FEL适用于较小的数据集（20-40个序列）。 SLAC在计算上效率更高，但统计能力略低。

1. 下载数据Lab5_Psy_hrpZ.fas

2. 前往Datamonkey并使用SLAC分析：http://datamonkey.org/slac

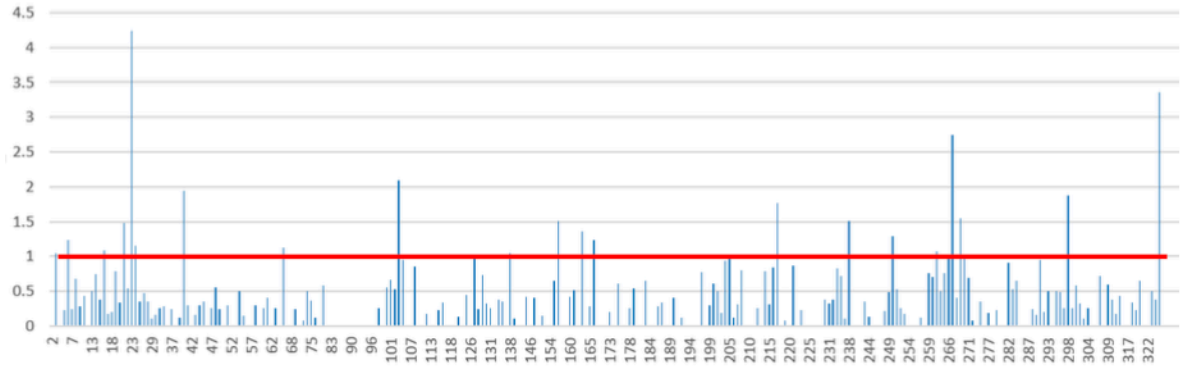
3. 上传多序列对比文件：
    a. 选择你电脑上的序列文件
    b. 选择遗传密码，在这种情况下为通用密码(**Universal code**)。
    c. 输入你的电子邮件地址

4. 点击**Run Analysis**运行。

5. SLAC结果：
    a. 提交数据后，Datamonkey将显示一个页面，显示你的数据已提交。 随着分析的进行，您将看到显示的作业日志。 您的“ticket” / job number在页面的URL中（例如，http://datamonkey.org/slac/5ddc8b943b682c361e77af12），你可以将其添加为书签以在限定时间后返回（如果您输入了电子邮件地址，则此URL会发送给你说明你的工作已经完成）。我们将会进入结果界面（页面标题为results sumary）
    b. 在结果页面的左侧有许多更为详细分析：
        i. **Table**: 表格（可以以逗号分隔的值文件形式下载）提供了序列中每个密码子分析的详细信息。CSV输出可以轻松导入到电子表格程序（如Microsoft Excel）中：我们将在实验结束时进行此操作，因此请确保您能够再次找到SLAC结果页面！
        ii. **Graph**: 图形将带您到一个页面，你可以在其中查看分析的图形输出。 对于SLAC，您可以跨站点绘制dN-dS。 请注意，Datamonkey不提供更典型的dN / dS图，而是绘制这两者之间的差异。 这样做是因为某些站点的dS可以为0，从而导致比值无穷大。
        iii. **Tree**: 树提供了针对使用的替代模型进行调整的比对数据的树（SLAC算法会对你的数据进行检查，并为数据提供最佳选择）。
    c. 在标记为“**Partition information**”的部分中提供了重要的数据，该数据提供了数据适合测试模型的可能性（对数转换: 数值越小越好）。
    d. 在此页面的下一部分，Datamonkey呈现所有正向（dN / dS> 1）和负向（dN / dS <1）选定密码子。 请记住，dN是每个可能的非同义位点的非同义（氨基酸变化）取代的比率，而dS是每个可能的同义位点的同义取代的比率； 因此，dN-dS值显着正值表示选择正，而负值显着负值表示选择负。 通过在顶部输入新的级别，可以动态地为不同的重要性级别重新生成此部分（这还将在顶部选择正负选择的站点数目进行更新）。 Datamonkey报告估计的dN-dS，该值由树的总长度缩放（以促进不同数据集之间的直接比较），该密码子的测试dN≠dS的p值等。
    e. 单击页面左面板上的“**Graph**”，或向下滚动到该部分。
    f. 结果图显示了沿序列（x轴）每个密码子的dN和dS（y轴）之间的差异。

6. FEL分析
    a. FEL（固定效应似然性）比SLAC强大一点，但速度较慢，因为它的计算量更大。
    b. 转到顶部的“**Methods and tools**”选项卡下的FEL分析。
    c. 加载文件并运行，选择所有的演化分支并“**Save Branch Selection**”。
    d. 注意点位的数量和重要性级别。

7. dN / dS图：如果你装了Excel可以试试这个
    a. 现在，我们通过导出Datamonkey数据并在MS Excel中执行一些简单的分析来制作更标准的dN / dS图。
    b. 从SLAC分析结果的“**Table**”部分中，选择**[Export Table to CSV]**。
    c. 将此页面另存为data.csv或类似文件的本地文件（不过请保留.csv扩展名）。
    d. 打开Excel并将你的文件导入。
    e. 你可以在Excel中看到你的dN、dS和dN-dS数据。
    f. 通过选择所有列（ctrl + A），以dS排序数据：在“**数据**”选项卡中，选择“**排序**...”。在弹出的对话框中，确保选中“**我的列表**”具有“**标题行**”，然后选择 在“**排序依据**”中选择“ **dS**”列。按最小到最大排序（即按增加值排序）。
    g. 删除dS值为空(null)的数值的行，同时按密码子位点进行重新排序。
    h. 计算一个新的dN / dS列（在所有站点上将dN列值除以dS列值）并通过插入柱形图进行绘制：标记要绘制的两组值（Site和dN / dS） 突出显示第一组，然后按住Ctrl键，同时突出显示第二组。 然后执行Insert => Chart，选择Column type，然后选择Clustered column子类型（左上方）。 在弹出的向导中，从数据系列中删除包含位置信息的系列（系列1），并将其添加到X轴标签上。 在向导中继续。 您应该得到一个如下图所示的图表。



    i. 回顾一下，正向选择的残基的dN / dS> 1，而负向选择的残基的dN / dS <1。知道哪些位置处于正选择之下可能对例如设计可以在病原体的情况下在该部位起作用的药物或疫苗产生帮助。
    j. 最后，我们可以用excel计算平均的dN / dS分数。

Datamonkey使用简介

为了进行选择分析，Datamonkey需要至少三个同源编码核苷酸序列的上传比对。 估计dN和dS的基于密码子的方法可以应用于任何序列比对，但要牢记一些注意事项：

理想情况下，比对应代表单个基因或其一部分（例如，一个亚基），在多个分类单元或不同的种群样本中进行采样。 比对中的序列数量很重要：太少的序列将包含太少的信息以进行有意义的推断，而太多的序列可能需要很长时间才能运行。 在撰写本文时，Datamonkey允许多达150个序列进行SLAC分析，并允许100个序列进行FEL / IFEL分析。 根据经验，至少需要10个序列才能以任何可靠性检测单个位点的选择。

目视检查数据以确保序列正确对齐是一个好习惯。 当然，永远不能确定一个比对在客观上是“正确的”，但是总的不比对（例如，超出序列的序列）很容易通过提供比对的图形化可视化的软件来发现，例如：MEGA，HyPhy ，Se-Al或BioEdit。**使用的数据应该验证对齐方式是否在框架内，即它不包含终止密码子，包括过早的终止密码子，表示帧偏移，例如由于未对齐或编码序列无效，以及终止密码子**。你的比对应排除核苷酸序列的任何非编码区域，例如内含子或启动子区域，而现有密码子替换模型将不适用这些区域。 当编码核苷酸序列直接比对时，可插入移码（即，不是3的倍数）缺口，因为比对程序通常不考虑序列的编码性质。 因此，通常比对翻译的蛋白质序列，然后将其映射回组成核苷酸是一个好主意。 Datamonkey收到编码序列并报告遇到的所有问题时，将执行许多检查。

如果比对包含相同序列，则Datamonkey将丢弃重复序列中的所有序列，但重复序列之一除外。 这样做是为了加快分析速度，因为相同的序列不会对似然推断过程提供任何信息（通过基本频率除外），但是系统发育分析的计算复杂度会随着序列的数量而增加。

如果比对包含相同序列，则Datamonkey将丢弃重复序列中的所有序列，但重复序列之一除外。 这样做是为了加快分析速度，因为相同的序列不会对似然推断过程提供任何信息（通过基本频率除外），但是系统发育分析的计算复杂度会随着序列的数量而增加。

Copyright Attribution