Lab2 高级BLAST与比较基因组学

本实验分为4个部分：BlastP，PSI-Blast，Translated Blast和Comparative Genomics（比较基因组学）。上一次，我们使用BLAST针对NCBI nr 数据库查询核苷酸序列。现在，我们使用蛋白质序列进行搜索：

**PSI-BLAST**

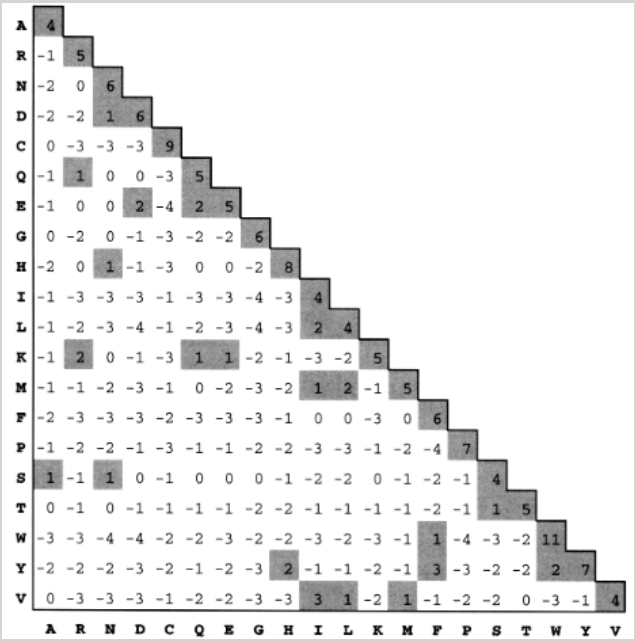
现在，我们将使用一种新的蛋白质搜索算法：特定位置迭代算法，即PSI-BLAST。PSI-BLAST是高度敏感的BLAST程序，对于查找远距离相关的蛋白质或蛋白质家族的新成员非常有用。 除了跟踪蛋白质家族成员之外，当您的standard protein-protein BLAST搜索未能找到重要的匹配结果或返回描述为“hypothetical protein”或“similar ti…”的匹配结果时，你还可以使用PSI-BLAST。

在一般意义上，PSI-BLAST以standard protein-protein BLAST开头，然后使用这些结果来建立更精细的搜索，该搜索通过连续的搜索迭代针对您的查询量身定制。 它通过建立特定位置的评分矩阵（PSSM）来做到这一点，该矩阵识别在查询序列和类似数据库序列之间最有可能出现的特定氨基酸变化。特定于位置的评分矩阵实际上是为你感兴趣的查询量身定制的替代矩阵。

Box1. Substitution Matrices

替换矩阵

替换矩阵描述了残基（无论是核苷酸还是氨基酸）随着进化时间变化的可能性。 它们是用于比较核苷酸或蛋白质序列的评分系统，该系统考虑了对序列进化的限制。 最著名的蛋白质替代矩阵是PAM和BLOSUM矩阵，其开发目的是确定在特定的进化时间内更可能发生哪些氨基酸变化。 下面的矩阵是BLOSUM62。 沿X和Y轴的字母代表20个氨基酸。 对角线以外的正数表示一个氨基酸变为另一种氨基酸的可能性更高，而低数字表示一种氨基酸的可能性较低。 您会注意到对角线上的数字都是非常强的正数，这表明氨基酸残基最可能要做的是保持不变。



所有替换矩阵均以数字索引表示（例如PAM120, PAM250, BLOSUM62, BLOSUM80）。数字表示不同的大小等级，具体取决于特定的矩阵。 对于PAM矩阵，较高数量的矩阵更适合于更趋同的对齐方式，而BLOSUM矩阵则相反。
位置特定评分矩阵（PSSM）本质上是专门针对目标蛋白质家族开发的替换矩阵，与PAM和BLOSUM矩阵相反，PAM和BLOSUM矩阵已被开发用于广泛的蛋白质序列。
PAM-n中，n 越小，表示氨基酸变异的可能性越小；相似的序列之间比较应该选用n值小的矩阵，不太相似的序列之间比较应该选用n值大的矩阵。PAM-250用于约20%相同序列之间的比较。BLOSUM-n中，n越小，表示氨基酸相似的可能性越小；相似的序列之间比较应该选用 n 值大的矩阵，不太相似的序列之间比较应该选 用n值小的矩阵。BLOSUM-62用来比较62%相似度的序列。

1. 从**blastp**页面进行BlastP计算，加载与Lab1相同的序列（NP_001318308）和参数（Blosum 80 Matrix和1e-15作为Expect Threshold，low complexity filter）。 在”Database" 下，选择“ UniProtKB / Swiss-Prot”。 Swiss-Prot数据库是精心策划的数据库，仅包含注释最好的（表征的）蛋白质序列。 使用此数据库的弊端在于，它不如默认的nr选项那么全面。

2. 在程序选择下，选择“ **PSI-BLAST**”。 在页面底部靠近**PSI / PHI BLAST**选项的位置，请注意原来的**PSI-BLAST Threshold**为0.005，让我们将其降低到1e-40。

3. 运行BLAST！

4. 检查PSI-BLAST输出。你会注意到一个非常明显的差异。HSP摘要部分中现在有两个部分：一个具有**E-values比[PSI-BLAST]Threshold更好**的序列区，另一个具有**E-values比[PSI-BLAST]threshold更差**的序列区。

5. 第一轮PSI-BLAST只是一个标准的BLASTP。 现在，我们将运行另一个迭代以优化搜索。 连续迭代那些有更好的截止基因的序列，以创建新的位置特定替换矩阵，该矩阵替换了原始搜索中使用的BLOSUM80矩阵。 该矩阵评分基于BLAST执行的每个成对比对中每个位置上发生的残基保守性的非随机模式：例如，当您从实验室的第一部分检查blastp比对时，您可能已经注意到这些模式在第16个小节。

6. 点击第一区和第二区之间的“**Run PSI-BLAST interation 2**”。向下滚动序列列表进行查看。 请注意，在某些序列旁边有一个绿色的复选打勾标记或以黄色突出显示的新序列，这些新序列在上一次迭代中并不重要，但在经过精炼的PSSM上得分很高。

7. 反复使用point6中的功能5次。
    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?

Box2. Potential PSI-BLAST issues
PSI-BLAST的潜在问题
尽管PSI-BLAST非常有效，但它也包含了一些潜在问题：
1. 我们必须假设数据库序列是独立的，并且样本空间足够大，足以代表该家族的真正基础多样性。 如果它们不是，那么PSSM将受到影响。例如，如果你正在搜索仅包含来自变形杆菌的蛋白质的数据库，则你的PSSM仅适用于该分类组。
2. 如果您的数据库包含大量紧密相关的蛋白质，则可能会看到错误的保留。在这种情况下，某些残基在功能上是保守的，但实际上它们之间的联系非常紧密，以至于它们没有时间分化。

## Translated BLAST

除了核苷酸与蛋白质的blast，还有其他三种blast，分别称为**blastx**，**tblastn**和**tblastx**。可以将BLAST的这些类型归为已翻译的BLAST搜索。 翻译的搜索可以使你可以在核苷酸和蛋白质水平之间来回移动。它们通常用于将蛋白质和核苷酸查询链接到无注释数据库中的同源DNA序列和蛋白质输出。 因为它们使用沿所有六个框架转换的查询和/或数据库，所以即使在存在测序错误和移码突变的情况下，它们也可以保持鲁棒性。

表1描述了基本的和翻译后的BLAST程序。 表格中的对齐（Alignment）列描述了将要查询并和数据库序列进行比较的级别。例如BLASTX会将您的DNA查询转换为蛋白质，并将其与蛋白质数据库对齐。 您会注意到，翻译后的BLAST程序执行多次搜索：对查询和/或数据库的每个阅读框进行一次搜索。

**Table 1**: Basic and Translated BLAST Programs

| Program | Query | Database | Alignment | N searches | Uses |
|---|---|---|---|---|---|
| blastn | DNA | DNA | DNA | 1 | find homologous DNA sequences |
| tblastx | DNA | DNA | protein | 36 | find homologous proteins from unannotated query and db sequences |
| blastx | DNA | protein | protein | 6 | identify coding sequences in query DNA sequence |
| tblastn | protein | DNA | protein | 6 | find homologous proteins in unannotated DNA db |
| blastp | protein | protein | protein | 1 | find homologous proteins |

举例：如何使用blatx

假设您下面有一个神秘的原核核苷酸序列，是通过对细菌基因组中的随机基因组文库克隆进行测序而获得的。 您想知道它是否编码蛋白质，如果是，则推断该蛋白质的功能。

>mystery_sequence

GTCACGTTACCGGTGGCCGAACAGGCCCGTCATGAAGTGTTCGATGTCGCGTCGGTCAGCGCGGCTGCCGCCCCAGTAAACA

CCCTGCCGGTGACGACGCCGCAGAATTTGCAGACCGCCACTTACGGCAGCACGTTGAGTGGCGACAATCACAGTCGTCTGAT
TGCCGGTTATGGCAGTAACGAGACCGCTGGCAACCACAGTGATCTAATTGCCGGTTATGGAAGTACAGGCACCGCCGGCTAC
GGCAGTACCCAGACTTCCGGAGAAGACAGCTCGCTCACAGCGGGTTACGGCAGCACGCAAACGGCTCAGGAAGGCAGCAATC
TCACCGCTGGGTATGGCAGCACCGGCACGGCAGGCTCGGACAGCTCGTTGATCGCCGGTTATGGCAGTACACAAACCTCGGG
AGGCGACAGTTCGCTGACCGCGGGCTACGGCAGTACGCAGACGGCCCAGGAGGGCAGCAATCTGACGGCGGGGTACGGCAGC
ACGGGTACAGCAGGTGTCGACAGCTCTCTGATCGCGGGATACGGCAGCACGCAGACCTCGGGAAGTGACAGCGCCCTGACCG
CAGGCTATGGCAGCACGCAAACGGCCCAGGAAGGCAGCAATCTCACTGCTGGGTATGGCAGCACCGGCACGGCAGGTTCCGA
CAGCTCGCTGATCGCCGGTTACGGCAGCACGCAAACCTCGGGCAGTGACAGCTCGCTCACGGCGGGGTACGGCAGTACGCAG
ACGGCTCAGGAAGGCAGCAATCTGACGGCGGGGTACGGCAGCACGGGTACAGCAGGTGTCGACAGTTCGTTGATCGCCGGAT
ATGGCAGCACGCAGACCTCGGGAAGTGACAGTGCGCTGACAGCGGGTTACGGCAGCACGCAAACGGCCCAGGAAGGCAGCAA
CCTGACGGCGGGCTACGGCAGCACTGGCACGGCAGGTGCCGACAGTTCGTTGATCGCCGGATATGGCAGCACGCAGACGTCA
GGCAGCGAAAGTTCGCTTACCGCAGGCTATGGCAGTACCCAGACTGCCCGTGAGGGCAGCACCCTGACGGCCGGATATGGCA
GTACCGGAACAGCTGGCGCTGACAGCTCGCTGATCGCCGGTTACGGCAGCACGCAAACCTCGGGCAGTGAAAGCTCGCTCAC
GGCAGGTTATGGCAGTACCCAGACCGCACAGC

1. 前往BLAST主页

2. 选择**blastx**，然后将序列复制并粘贴到搜索框中。确保将非冗余蛋白序列**Nonredundant protein sequences（nr）**选择为数据库。你可以使用其余的默认参数。

3. 让我们做一些BLAST！

4. 滚动查看结果：
   a. 得分最高的蛋白质序列属于什么物种?
   b. 你认为该基因的功能是什么?
   c. 你认为该基因来自什么组织器官?

5. 前往第一个Alignment区域：你可能会注意到，一些查询残差使用小写的灰色字母（与大写的黑色相对）。你能猜出这些可能意味着什么吗? 尝试返回查询页面并更改“Filters”和“Masking”选项。 这如何影响你的结果呢?

## 比较基因组学

高通量测序计划，蛋白质组学，转录组学和其他高通量基因组技术，再加上分子表征和文献整理，已经产生了大数据收集。除了人类基因组本身之外，还存在蠕虫，果蝇，小鼠，拟南芥和其他生物的模式生物的存储库。然而，由于这些数据的范围，复杂性和数量都相当巨大，因此基因组数据的表示具有挑战性。有效显示这种复杂数据的工具和手段正在不断发展和完善。我们将研究试图解决该问题的几种应用程序，这些应用程序允许比较相关物种的基因组区域。重要的是，可以假定在一个物种中研究的直系同源基因在另一物种中具有相似的功能，而在直系同源物中的高度保守的残基可以认为对该蛋白质的功能至关重要。其中存在比较基因组学的力量。

Box3. Comparative Genomics
比较基因组学
正如我们在lab1中看到的那样，为了能够进行比较基因组学，我们需要能够确定直系同源物和旁系同源物。从概念上讲，最简单的方法是将一个基因组的区域（基因）与另一个基因组进行blast，并确定第二个基因组中e值最低的区域（基因）。 这是有问题的，因为如果第二个基因组中识别出的区域实际上在第一个基因组中的其他位置具有更好的匹配会发生什么呢？ 此方法的一种变体涉及在两个方向上进行blast，即从一个基因组上的基因到另一个基因组进行blast，然后在相反的方向上进行爆炸，并确定“相互最佳匹配（reciprocal best hit）”或RBH。 这减少了假阳性直系同源物的数目，但是增加了假阴性的数目。
已经开发出多种方法来解决该问题，并涉及使用系统发育方法（在计算上更“昂贵”）或BLASTP，然后使用聚类方法来鉴定直系同源基因。 这些方法在第一种情况下是RIO和Orthostrapper，在后一种情况下是InParanoid和OrthoMCL。

使用基因组浏览器探索基因组

每个模式生物（之所以如此命名是因为其易于操作，空间需求，良好的遗传学以及其他原因，因此在医学或农业领域进行了长期的研究）拥有自己的基因组数据库，可用于探索基因组区域。 这样的区域通常具有与之相关的其他分子：同源基因，EST等。 这些基因组区域可以使用基因组浏览器进行探索。在本实验的这一部分中，我们将只研究Mouse基因组浏览器，但是这里有些其他门户可能会在您的未来研究中使用：

**Flybase**

FlyBase是基因组信息库，可为模型生物果蝇（Drosophila melanogaster）和许多其他相关的果蝇物种提供信息。在http://flybase.org上连接到FlyBase，然后单击GBrowse图标以访问Genome Browser。

**WormBase**

WormBase是秀丽隐杆线虫和其他蠕虫模型物种基因组数据的存储库。 通过http://www.wormbase.org连接到wormbase。 单击页面顶部“工具”部分中的“ GBrowse”链接，以访问秀丽隐杆线虫的基因组浏览器。

**The Arabidopsis Information Resource**

如我们所见，TAIR是拟南芥基因组信息的存储库。 在http://www.arabidopsis.org上连接到TAIR，然后在“工具”选项卡下，单击GBrowse。 拟南芥信息门户网站是一项不同的计划，旨在整合来自不同来源的拟南芥的所有数据：http://araport.org。

**NCBI**

也可以使用NCBI Map Viewer应用程序以比较方式检查人类基因组。连接到NCBI网站的基因组部分，网址为：http://www.ncbi.nlm.nih.gov/Genomes/，然后在“自定义资源”下选择“人类基因组”链接，然后在染色体图标上选择最多的最新的Map Viewer版本。在“Maps and Options”下，可以从其他物种中选择序列以显示在人类Map Viewer上。

**小鼠基因组信息学–比较基因组学实例:**

连接到Mouse Genome Informatics网站，网址为：http://www.informatics.jax.org/。 在左上方的“快速搜索”框中输入Pax6。 Pax6是对眼睛正常发育很重要的基因。 单击结果列表中第一个链接，标记为Pax6。 您将被带到该基因的“Gene Detail”页面。 在第二行中，称为“**Location & Maps**”，单击“更多”以展开该部分，然后单击“ Ensembl Genome Browser”链接。 Ensembl由NCBI的欧洲对口机构欧洲生物信息学研究所管理。在某些方面，该浏览器比NCBI地图查看器功能更强大，但是由于它包含了大量信息和众多选项，因此更加令人困惑。我们可以使用它来检查相似的人类和小鼠基因组区域，并查看直系同源物。

我们在页面左侧Location-based displays中找到Region Comparison，接着在右侧齿轮图标处Select species or regions选择人类即可对比，同时我们可以通过滑动图片右上角的滑块来放大缩小片段，结果中绿色线条相连的部分即位人类和老师基因片段中相似的序列。接着可以查看围绕PAX6的其他人类直系同源物及其染色体位置。

Box4. Comparative Genomics and Synteny
比较基因组学与同义性
基因顺序通常在密切相关的物种之间，甚至在不紧密相关的物种之间，例如人类和小鼠之间，都保持保守。可以通过使用一些已开发的可视化工具来观察这种同义性。例如Artemis Comparison Tool, ACT。
这种可视化工具可用于识别插入: 连接水平显示的基因组的两个区块的斜率差异越大，在一个或另一个基因组中的插入越大；易位和倒置: 这些表现为交叉的区块 其他块，以及“X”形图形。

**WebACT**

你可以使用Sanger研究所的http://www.webact.org/WebACT/prebuilt工具检查几个预先计算的基因组比较的异同。例如，检查根癌农杆菌菌株C58 / ATCC 33970，Cereon圆亚型菌株（AE007869）与根癌农杆菌菌株C58 / ATCC 33970，杜邦亚种（AE008688）基因组。 我们全程使用默认值。如果浏览器中未激活Java Web Start，则可能需要下载WebACT提供的.jnlp和/或.zip文件，然后在使用Java的计算机上运行它（并添加http://www.webact.org:80 /（位于“配置Java”设置中的“例外站点列表”中）。