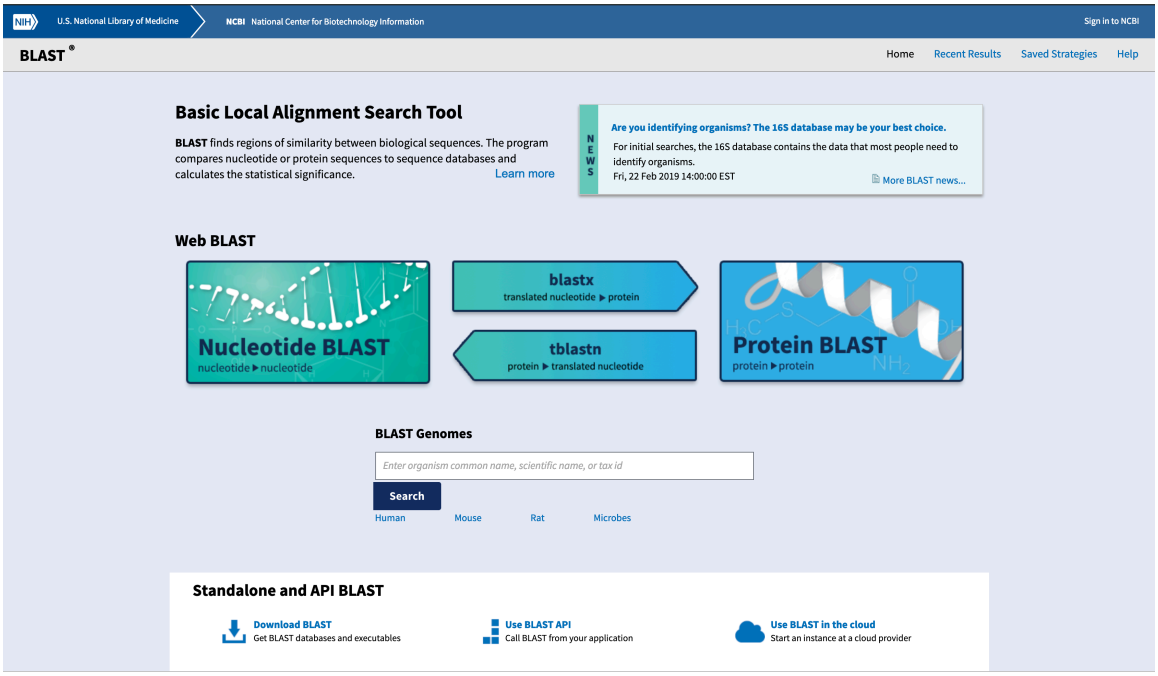
BLAST 的基本原理

什么是BLAST

Basic Local Alignment Search Tool (BLAST) 是一套在蛋白质数据库或DNA数据库中进行相似性比较的分析工具。BLAST程序能迅速与公开数据库进行相似性序列比较。BLAST结果中的得分是对一种对相似性的统计说明。



在NCBI上分为多种，常用的有N-BLAST(核苷酸序列比对检索)和P-BLAST(蛋白比对检索)。此外还有Primer BLAST等用于搜索和设计特殊核酸序列的工具。BLAST对一条或多条序列(可以是任何形式的序列)在一个或多个核酸或蛋白序列库中进行比对。

BLAST还能发现具有缺口的能比对上的序列。BLAST可处理任何数量的序列，包括蛋白序列和核算序列；也可选择多个数据库但数据库必须是同一类型的，即要么都是蛋白数据库要么都是核酸数据库。所查询的序列和调用的数据库则可以是任何形式的组合，既可以是核酸序列到蛋白库中作查询，也可以是蛋白序列到蛋白库中作查询，反之亦然。

BLAST是一个被广泛使用于分析生物信息的程序，因为它可以兼顾我们在做搜索时的速度以及搜索结果的精确度。因为当我们所要搜索的目标数据库非常庞大的时候，速度就变成一项很需要考量的因素。在像BLAST和FASTA这些快速算法被开发之前，我们是使用动态规划算法来作数据库的序列搜索，这真的非常的耗时。BLAST使用启发式搜索来找出相关的序列，在速度上比完全只使用动态规划大约快上50倍左右，不过

它不像动态规划能够保证搜索到的序列（Database sequence）和所要找的序列（Query sequence）之间的相关性，BLAST的工作就是尽可能找出数据库中和所要查询的序列相关的信息而已，精确度稍微低一点。

BLAST的基本工作原理

1. 移除Query序列中之低复杂度以及有串接重复现象的区域

低复杂度是指由很少种类的元素（如氨基酸）所组成的一个区域；而串接重复现象是指在一个Query序列中，有两段串连的区域它们组成的方式一模一样。这两种在序列中的区域可能会让BLAST找出一些虽然分数够高，但是其实和Query序列并不相关之序列，所以在我们执行搜索之前，要先把Query序列中的这两种区域滤掉。BLAST的实际作法是，它会把这些区域用符号代替，并且在搜索的时候忽略这些符号。蛋白质序列中，就用X符号标示；而DNA序列中，则用N符号标示。低复杂度区域的部分，BLAST是用一个叫做SEG的程序来处理蛋白质序列，而用叫做DUST的程序来处理DNA序列。另一方面，蛋白质序列中之串接重复现象的区域则是使用XNU来处理。

2. 将Query序列中每k个字的组合做成一个表

以k=3为例（DNA序列中，我们则常以k=11为例），我们"依序"将Query序列中每3个字的组合视为一个字组，并将这些字组列在一张字组表上，直到Query序列中最后一个字也被收入进表上为止。

3. 列出我们所关心的所有可能之字组

这个步骤就是BLAST和FASTA之间很重要的一点不同处。FASTA关心所有在第二步中所找出的字组表上的每一个字组，它会去搜索数据库中的序列，看看这些序列是否含有这些字组；然而，BLAST只对高分的一些字组有兴趣，而字组的分数是由依序比较字组间的每个字，再配合得分矩阵（substitution matrix或scoring matrix）所产生的。因此，对于每一个字组而言，可能有20^3 个BLAST可能关心的字组，当然这些字组经过一个门槛分数的筛选后，只有少数的字组会留下，而这些就是BLAST真正所关心的字组。

4. 将这些经筛选后剩下的高分字组组织成快速搜索树的结构

这是为了要让程序能快速的比对这些高分字组和数据库中的序列是否有完全匹配（exact match）的情形。

5. 对每个Query序列中的字组重复步骤1到4，并得到所有相应的高分字组

6. 扫描数据库中的序列，看看是否有跟剩下的高分字组完全匹配的情形

BLAST会搜索数据库序列中是否有高分字组出现，像是在第三步找出来的PEG。若扫描到有完全匹配的情形发生，那这个完全匹配的位置就会是我们之后要对Query序列和数据库序列做无间隙的区域排比（ungapped local alignment）的起点。

7. 将这些完全匹配的地方扩展为高分序对（high-scoring segment pair, HSP）

8. 将所有分数够高的HSP列出来，并评估这些留下来的HSP它们的分数是否具有意义

所有分数高于某个由经验法则推测出来的门槛值S之HSP都将被列出。这个门槛值S是由检视两个不相关的序列去作大量无间隙的区域排比得来的分数之分布，进而推测出S该怎么制定以保证被留下来的HSP都具有一定程度的意义。

BLAST会利用Gumbel extreme value distribution (EVD)这个极值的分布，来评估每个HSP分数的统计意义（已经有人证明两个不相关的序列去作区域排比时，不考虑gap的使用，做出来的分数都会呈现Gumbel EVD的情况；考虑gap的使用时，该结论尚未被证明）。