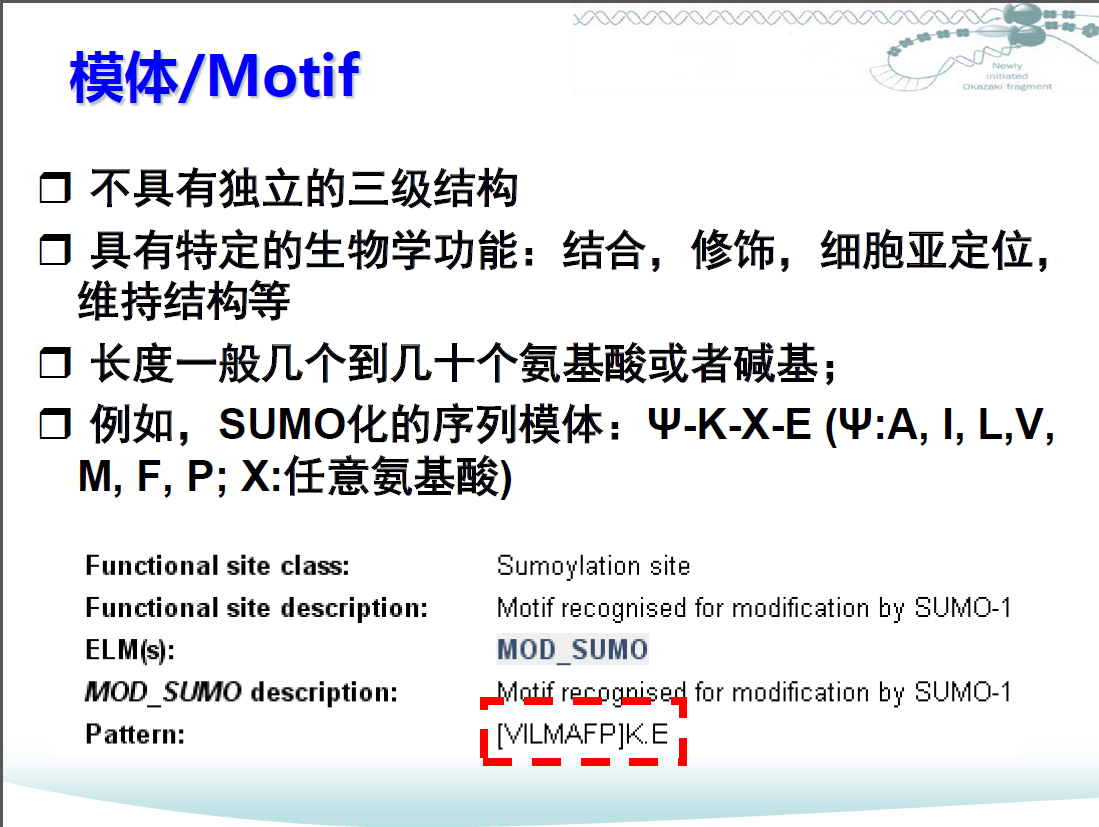
# 1.什么是模体



# 2.模体识别的算法

①序列比对算法

**动态规划算法**

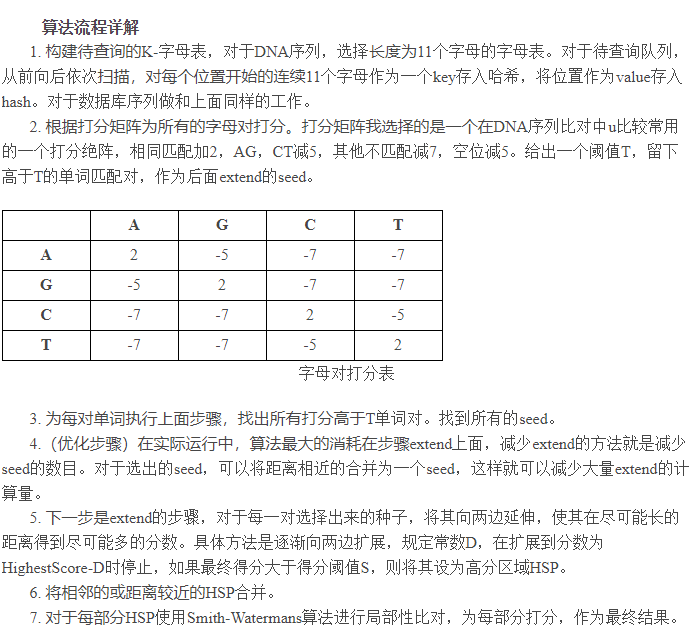
给定两条序列，两序列联配就是在某个打分方案下，试图得到一个最优的联配。Needleman Wunsch算法是一个动态规划算法，它计算了一个的二维矩阵[Needleman1970]，矩阵中的元素表示子串和之间的最优联配打分，动态规划的递推表达式如下：

其中表示氨基酸的联配打分，最常用的打分矩阵是BLOSUM62和PAM250；为插入和删除罚分。在填完二维矩阵以后，就是最优全局联配的分数。为了得到具体的联配情况，我们还需要一个回溯矩阵来记录每一个位点的联配情况，然后从最后一个单元格开始回溯，就可以得到最优联配。

对很多蛋白质来说，常常只有部分区域相似，而其他部分差异很大。此时全局最优联配并不是我们想要的结果，而局部最优联配是一个更合理的选择。获得最优局部联配同样是填写一个二维矩阵，具体递推式如下：

二维矩阵中分数最大者即为最优局部联配打分，而最优联配可从该单元回溯到某个分数为0的单元获取。

**blast算法**

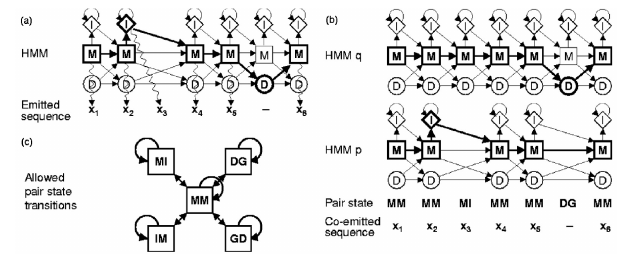


以PSI-BLAST为例，序列型是采用迭代策略从同源序列之间的联配建立。由于序列型包含了更多的进化信息，因此能够有效提升联配敏感度 。

序列型-序列型联配方法非常类似于序列-序列联配方法。唯一的区别在于序列型之间的替代打分矩阵有所不同。两列序列型的相似度打分通常采用如下的log-odds表示：

其中分别为两个序列型中残基出现的概率，为背景分布中残基出现的概率。若log-odds取正值，则表示相应的两个序列型比较相似；否则表示序列型之间不相似。

②隐马尔科夫模型



其中表示联配中的第k列的状态(M/I/D)；表示两列之间的状态的转移概率；而具体形式如下：

其中表示对应两个序列型中氨基酸a出现的概率，表示在背景分布中氨基酸a出现的概率。除以背景分布的作用在于增大那些不常出现氨基酸的权重。对于两列相似的profile来说打分为正，否则为负，这个性质有助于在动态规划联配过程中获得局部最优联配。

图2-7显示一个HMM-HMM联配的例子，其中 HMM每一列都有一个隐含状态，即匹配 (M)、插入 (I)和删除 (D)。只有匹配状态和插入状态可以发射残基，所以一个匹配状态只能与另一个HMM的匹配状态或插入状态联配；而删除状态只能与空位联配。因此，HMM-HMM联配中有五种状态MM，MI, IM, DG和GD，联配的动态规划递推表达式如下：

类似的，我们可以计算和，其中分别在两个HMM上位点和上的各个状态的转移概率。

③gibbs采样法

