一、 ClustalW 多重序列比对

问题：试利用 ClustalW，对附件给出的球蛋白家族的 13 条氨基酸序列 globins.fasta 进行多重序列比对；

(1) 实验过程：

访问 Kyoto University Bioinformatics Center 的在线多重序列比对网站，选择 ClustalW 工具：https://www.genome.jp/tools-bin/clustalw

上传 fasta 文件，多序列比对的参数为：Weight Matrix：BLOSUM，Gap Open Penalty:10, Gap Extension Penalty:0.5，进行多序列比对。

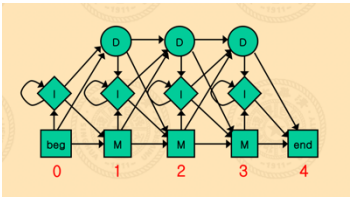
(2) 实验结果：得到多序列比对结果的 clustalw.aln 文件，以及 clustalw.dnd 文件，dnd 文件可以通过 http://itol.embl.de/来在线建树

(3) 实验结果分析：多序列比对结果中双点：表示强相似，表示该位置的突变是个保守性突变；单点. 表示弱相似，半保守性突变；星号*表示该位点所有氨基酸一致；从比对结果中可以发现 globins 蛋白家族有 2 个高度保守位置。

二、 profile HMM 模型建立

问题：基于上述多重序列比对结果，建立相应的 profile HMM 模型；

(1) 实验过程：



profile HMM 模型如图，在 begin 和 end 之间的每一个位点都有三种状态：匹配状态 M，插入状态 I，缺失状态 D；转态定义规则：

每个位置（列）中总插入/缺失的氨基酸残基（-）大于θ，则认为这列的氨基酸残基为插入状态 I；否则，则认为这列的（-）为缺失状态 I；

匹配状态 M 和插入状态 I 的观测值各有 20 种氨基酸残基，缺失状态 D 没有观测值；

通过统计比对结果中，位置 j 到位置 j+1 的各种状态转移（M->M, M->I, M->D, I->M, I->I, I->D, D->M, D->D, D->I）的次数，得到位置 j 的状态转移概率；统计位点 j 的 20 种氨基酸出现的频率，得到位置 j 的发射概率。

(2) 实验结果：

本次作业的利用 HmmerBuild 建立 profile HMM 模型，运行命令为

./hmmbuild pHMM.out clustalw.aln。

结果在 pHMM.out 文件中：A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y 对应的列为发射概率，m->m,m->i,m->d,i->m,i->i,d->m,d->d 对应的列为转移概率；共 150 行（150 个位点）

三、 计算氨基酸序列的概率和最优路径

问题： 试利用上述 profile HMM 模型，分别计算生成下列 2 个氨基酸序列的概率和最优路径

(1) 实验过程：利用建立好的 profile HMM 模型，

前向算法评估观察序列概率：

$$
\begin{aligned}
F_j^{\mathrm{M}}(i) &= \log\frac{e_{\mathrm{M}_j}(x_i)}{q_{x_i}} + \log[a_{\mathrm{M}_{j-1}\mathrm{M}_j}\exp(F_{j-1}^{\mathrm{M}}(i-1)) \\
&\quad + a_{\mathrm{I}_{j-1}\mathrm{M}_j}\exp(F_{j-1}^{\mathrm{I}}(i-1)) + a_{\mathrm{D}_{j-1}\mathrm{M}_j}\exp(F_{j-1}^{\mathrm{D}}(i-1))] \\
F_j^{\mathrm{I}}(i) &= \log\frac{e_{\mathrm{I}_j}(x_i)}{q_{x_i}} + \log[a_{\mathrm{M}_j\mathrm{I}_j}\exp(F_j^{\mathrm{M}}(i-1)) \\
&\quad + \log a_{\mathrm{I}_j\mathrm{I}_j}\exp(F_j^{\mathrm{I}}(i-1)) + a_{\mathrm{D}_j\mathrm{I}_j}\exp(F_j^{\mathrm{D}}(i-1))] \\
F_j^{\mathrm{D}}(i) &= \log[a_{\mathrm{M}_{j-1}\mathrm{D}_j}\exp(F_{j-1}^{\mathrm{M}}(i)) + \log a_{\mathrm{I}_{j-1}\mathrm{D}_j}\exp(F_{j-1}^{\mathrm{I}}(i)) \\
&\quad + a_{\mathrm{D}_{j-1}\mathrm{D}_j}\exp(F_{j-1}^{\mathrm{D}}(i))]
\end{aligned}
$$

其中，a 为状态转移概率，e 为发射概率，j 为序列位置，i 为 j 位置上的氨基酸残基；

Viterbi 算法计算最优路径：

$$
V_j^{\mathrm{M}}(i) = \log\frac{e_{\mathrm{M}_j}(x_i)}{q_{x_i}} + \max\begin{cases} V_{j-1}^{\mathrm{M}}(i-1) + \log a_{\mathrm{M}_{j-1}\mathrm{M}_j} \\ V_{j-1}^{\mathrm{I}}(i-1) + \log a_{\mathrm{I}_{j-1}\mathrm{M}_j} \\ V_{j-1}^{\mathrm{D}}(i-1) + \log a_{\mathrm{D}_{j-1}\mathrm{M}_j} \end{cases}
$$

$$
V_j^{\mathrm{I}}(i) = \log\frac{e_{\mathrm{I}_j}(x_i)}{q_{x_i}} + \max\begin{cases} V_j^{\mathrm{M}}(i-1) + \log a_{\mathrm{M}_j\mathrm{I}_j} \\ V_j^{\mathrm{I}}(i-1) + \log a_{\mathrm{I}_j\mathrm{I}_j} \\ V_j^{\mathrm{D}}(i-1) + \log a_{\mathrm{D}_j\mathrm{I}_j} \end{cases}
$$

$$
V_j^{\mathrm{D}}(i) = \max\begin{cases} V_{j-1}^{\mathrm{M}}(i) + \log a_{\mathrm{M}_{j-1}\mathrm{D}_j} \\ V_{j-1}^{\mathrm{I}}(i) + \log a_{\mathrm{I}_{j-1}\mathrm{D}_j} \\ V_{j-1}^{\mathrm{D}}(i) + \log a_{\mathrm{D}_{j-1}\mathrm{D}_j} \end{cases}
$$

其中，a 为状态转移概率，e 为发射概率，j 为序列位置，i 为 j 位置上的氨基酸残基；

(2) 实验结果：运行 python main.py, 函数 viterbi 计算最优路径，函数 forward 求 log-odds

得到两个序列的最优路径和 log-odds:

>d1fsla_ a.1.1.2 (A:) Leghemoglobin {Soybean (Glycine max), isoform A}

IIIIIDDDDMMMMMMMMMMMMMMMMMMMIIIIDDDDMMMMMMMMMMMMMMMMM
MMMMMMMMMMMMMMMMMMDIDMMMMMMMMMMMMMMMMMMMMMMMMMMM
MMMMMMMMMMDMMMMMDMMMMMMMMMMDDDDMMMMMMMMMMMMMMM
MMMMMMMMMMMMMMMMMMMMMMMMMMDDDDI

log-odds: -311.52355435612134

>d1ea0a3 d.153.1.1 (A:1-422) Alpha subunit of glutamate synthase, N-terminal

IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIDDDDMMMMMMMMMMMM
MMMMMIDDDDMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMDI
DMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMDMMMMMDM
MMMMMMMMDDDDMMMMMMMMMMMMMMMMDDMMMMMMMMMMMMMMM
MMMMDDDDDDI

log-odds: 338.726315994433