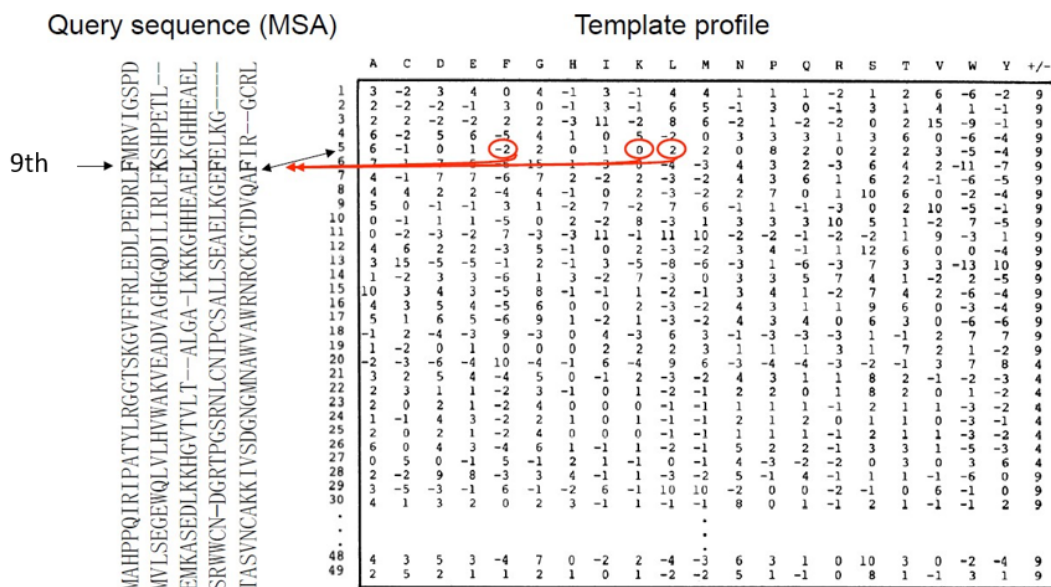


2024 春 生物信息学 结课考试 本科生试题 回忆版

命题：高建召

回忆：栗山未来

1. 用 N-W 算法对两条序列进行比对，空位罚分起始=延伸=-7，打分矩阵为两字符相等+5，两字符不相等-3，序列长度大概是 1 条 5 字符一条 7 字符
2. 推导不带权重的 profile-profile 打分公式 $S(i,j)=\langle f_i,p_j \rangle$ ，并用公式计算：



S (9, 5) (跟原题不一定完全一样)

3. 给定距离矩阵，用 UPGMA 方法构建进化树，类似下题：

D_{ij}	Bear	Raccoon	Weasel	Seal
Bear 熊	-	0.26	0.34	0.29
Raccoon 浣熊		-	0.42	0.44
Weasel 黄鼠狼			-	0.44
Seal 海豹				-

4. HMM 的前向算法：

例 10.2 考虑盒子和球模型 $\lambda = (A, B, \pi)$, 状态集合 $Q = \{1, 2, 3\}$, 观测集合 $V = \{\text{红}, \text{白}\}$,

$$A = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}, \quad B = \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{bmatrix}, \quad \pi = (0.2, 0.4, 0.4)^T$$

设 $T = 3$, $O = (\text{红}, \text{白}, \text{红})$, 试用前向算法计算 $P(O | \lambda)$.

- 给定 label 和 prediction 的概率, 在阈值 0.25, 0.5, 0.75 下分别计算混淆矩阵和 TPR, TNR
- 给定公式和两对序列 (包括观测序列和状态序列), 用极大似然法计算 HMM 参数 π, A, B :

□ 使用 $\gamma_i(t), \xi_t(i, j)$ 表示结果

$$\pi_i = \frac{P(O, q_1 = S_i | \lambda')}{P(O | \lambda')} = \gamma_i(\mathbf{1})$$

$$a_{ij} = \frac{\sum_{t=1}^{T-1} P(O, q_t = S_i, q_{t+1} = S_j | \lambda')}{\sum_{t=1}^{T-1} P(O, q_t = S_i | \lambda')} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

$$b_j(k) = \frac{\sum_{t=1}^T P(O, q_t = S_j | \lambda') I(o_t = v_k)}{\sum_{t=1}^T P(O, q_t = S_j | \lambda')} = \frac{\sum_{t=1, o_t=v_k}^T \gamma_t(j)}{\sum_{t=1}^T \gamma_t(j)}$$