绝密 ★ 启用前

南开大学数学科学学院

生物信息学2024-2025期末测试卷

注意事项:

1. 命题人：李晗、王子叶
2. 考试限时：100 分钟
3. 考试时间：2025 年 5 月 20 日 (数院 520 必有考试)
4. 数据难以记起，部分数字系人为捏造.

一、解答题

1. 给定两个序列：ATCGATAA 与 ACGACC，利用 Smith-Waterman 算法求解下列问题，其中 open = extension = -2，打分函数为

$$s(i,j)=\begin{cases}1, & x_i=y_j\\ -1, & \text{Otherwise.}\end{cases}$$

（1）试写出打分矩阵；

（2）根据打分矩阵的结果，写出序列比对结果.

2. 给定如下的距离矩阵：

	1	2	3	4	5
1	-	0.3	0.6	0.6	0.7
2	0.3	-	0.6	0.6	0.8
3	0.6	0.6	-	0.3	0.4
4	0.6	0.6	0.3	-	0.2
5	0.7	0.8	0.4	0.2	-

（1）使用 UPGMA 方法，构建进化树.

（2）计算进化树的总枝长.

3.（12 分）

给定如下的预测概率和真实标签：

真实标签	预测概率
1	0.3
1	0.1
0	0.65
0	0.5

（1）以 0.2, 0.4, 0.6 为阈值，计算混淆矩阵，并说明预测准确率；

（2）对上一问每个阈值得到的混淆矩阵计算召回率和精确率.

4. 给定如下的 HMM 模型：

$$A = \begin{pmatrix} 0.5 & 0.2 & 0.3 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \end{pmatrix}, B = \begin{pmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \\ 0.7 & 0.3 \end{pmatrix}, \pi = \begin{pmatrix} 0.2 \\ 0.4 \\ 0.4 \end{pmatrix}$$

利用前向算法求解观测序列 $O = \{1, 0, 1\}$ 的概率 $P(O \mid \lambda)$.

5. 现在我们有一个 DNA 序列，碱基有四种：A,T,C,G （分别为腺嘌呤、胸腺嘧啶、胞嘧啶、鸟嘌呤），我们想将其输入到模型当中；

（1）请说明至少两种将其编码为数值变量的方式；

（2）请说明你写出的编码方式的优缺点.

6. 生物信息学的生物基础：

（1）请说明生物学中心法则的内容；

（2）请说明蛋白质结构的四级结构是什么？并说明各级结构的特点；

（3）请从如下的多个主题选择一个：基因组学、宏基因组学、蛋白质结构预测、单细胞组学，阐述这个主题当中的一个经典问题，并说明解决它的经典方法和模型解释.