# Course Project 2020.2【占总成绩的 70%】

1. 对于给定的如下三条蛋白质序列，试采用SP评分函数 $S(m_i)=\sum_{k<l} s(m_i^k, m_i^l)$（对第*i*列），给出相应的最优比对路径，其中精确匹配（exact match）定义为3分，错配（mismatch）为-1，且空位（gap）罚分为-2。

>d1g08b_ a.1.1.2 (B:) Hemoglobin, beta-chain {Cow (Bos taurus)}
mltaeekaavtafwgkvkvdevggealgrllvvypwtqrffesfgdlstadavmnnpkvk
ahgkkvldsfsngmkhlddlkgtfaalselhcdklhvdpenfkllgnvlvvvlarnfgke
ftpvlqadfqkvvagvanalahryh
>d1itha_ a.1.1.2 (A:) Hemoglobin {Innkeeper worm (Urechis caupo)}
gltaaqikaiqdhwflnikgclqaaadsiffkyltaypgdlaffhkfssvplyglrsnpa
ykaqtltvinyldkvvdalggnagalmkakvpshdamgitpkhfgqllklvggvfqeefs
adpttvaawgdaagvlvaamk
>d1cqxa1 a.1.1.2 (A:1-150) Flavohemoglobin, N-terminal domain {Alcaligenes eutrophus}
mltqktkdivkatapvlaehgydiikcfyqrmfeahpelknvfnmahqeqgqqqqalara
vyayaeniedpnslmavlkniankhaslgvkpeqypivgehllaaikevlgnaatddiis
awaqaygnladvlmgmeselyersaeqpgg

2. 采用Feng-Doolittle累进式多重序列比对算法，对附件P1_fd.fasta中给定的44条序列进行比对，试给出以下结果：

(1) 利用如下定义的距离：

$$D=-\log S_{eff}=-\log\frac{S_{obs}-S_{rand}}{S_{\max}-S_{rand}}\quad ,$$

计算出该44条序列的距离矩阵；

(2) 利用（1）中的距离矩阵构建引导树（guide tree）；

(3) 给出多重序列比对结果。

请每位同学在规定的截止日期前独立地提交自己的完整报告。报告中必须包含相关的基本概念、基本原理以及对计算结果的分析和讨论，这些将是考核的重点。格式请严格按照《计算机学报》或《清华大学学报（自然科学版）》进行排版。另请附程序说明文件和程序代码文件。