**B题**

RNA干扰 (RNA interference, RNAi) 技术能够研究一个单独基因的作用, 其效应因子为小干扰RNA(small interfering RNAs, siRNA)分子。由于使用RNAi技术可以特异性剔除或关闭特定基因的表达，所以该技术已被广泛用于探索基因功能和传染性疾病及恶性肿瘤的基因治疗领域。在过去的几十年里，使用RNAi治疗癌症、代谢性疾病、呼吸系统疾病和传染性疾病已经进入临床实践。最近，一些研究人员报道了一组内源性siRNAs通过抑制DNA翻译发挥了作用。 但是，RNAi基因沉默（是指生物体中特定基因由于种种原因不表达或者是表达减少的现象）的有效性依赖于siRNA的靶向特定基因，因此选择有效的siRNA是十分重要的。

一般来说，研究人员通过各种算法设计有效的siRNA，这些算法所使用的特征包含经验规则，核苷酸频率，二元编码模式，热稳定性，以及许多混合特征。事实上，对于这些常用的特征，并没有直接的实验证据表明它们能够影响siRNA的有效性，也不能给出有效和无效siRNA的明确界限。

Hencken数据集包含针对不同病毒的siRNA序列数据库 <http://crdd.osdd.net/servers/virsirnapred/dataset.php?dataset=1725_unique>

4.Unique (1725)数据集提供了1725个长度为19bp的 siRNA的实验指标（见附件），通常定义效率超过70 % 靶向基因敲除被认为是有效siRNA（效率为负表示促进DNA翻译）。根据该数据集，从中提取特征，构造分类方法，对有效和无效的siRNA进行分类，完成下面三个问题：

1. 对你们的分类方法进行评估。
2. 预测siRNA的效率。
3. 对你们的预测效率进行评估。