实验一 序列数据分析

1. 背景知识

FASTQ 格式是一种常用的序列文件格式，是当前高通量测序数据的标准格式。illumina 测序得到的原始图像数据经过 Base Calling 转化为序列数据，结果以 FASTQ 文件格式来存储，包含测序 read 的序列信息以及测序质量信息。

FASTQ 文件格式如下所示：

@K00169:186:HM5C2CCXX:6:1101:8136:2962 1:N:0:CTGGCATA
CCACTCATAATCCAGCAAATACTAAATCTGCTGCAGGAAAAGAAATGCGGTTGAGCTT
+
AFFKKFKKFFKKKKFKAFKKAAKFAFFKKFKKFFKKKKFKAFKKAAKFAFFKKFKKFFKF

第一行：区分不同 reads 的 ID 号。以'@'开始，后面跟着序列的描述信息。
第二行：序列信息，由碱基以及 N 组成。
第三行：'+'，或者与第一行相同，无特殊意义。
第四行：第二行序列中每个碱基对应的测序质量值，以 ASCII 码表示。

近年来，测序质量多采用 Phred33 编码方式，碱基质量得分 Q 与 ASCII 值的关系是：**ASCII** 值 = **Q** + **33**。一般，碱基质量为 0 ~ 40，即 ASCII 值范围为 33 ~ 73，对应字符为！~ I 。

根据碱基质量得分可以评估测序出错率，碱基质量得分 Q 与测序错误率 P 的换算关系为：$\mathbf{Q = -10log_{10}P}$ 。

对于测序得到的 FASTQ 文件，通常需要进行常规的序列分析，来评估测序质量。比如，测序得到的 reads 数量、碱基含量分布（包括错误碱基 N 的含量分布）、GC 含量分布、碱基质量分布等统计。

2. 实验目的

熟练使用 python 语言，对序列数据进行分析和可视化。

3. 实验任务

给定 FASTQ 文件“data1.fq”，进行如下分析。

1）GC 含量统计并作图显示。计算每条 read 中的 GC 含量（即 G+C 的总含量），并用直方图显示。

2）统计所有 reads 在各位置上 ACGT 碱基以及 N 的含量分布，并作图显示。

3）将 FASTQ 文件中测序质量序列转换为碱基质量，统计所有 reads 在各位置上碱基质量分布，并作图显示。

4）产生低质量 FASTQ 文件“data1_low.fq”。对于给定的文件“data1.fq”，随机选择给定比例 p（比如 p=0.05）的 reads，并对选择的 reads 随机选择 k（k < len(read), 比如 k = 15）个位置，将这 k 个位置上的碱基替换为字符“N”。用参数“-p 0.05 –k 15”运行脚本，得到低质量 FASTQ 文件“data1_low.fq”；然后，用题 2）中的脚本重新统计各位置上的 ACGT 碱基以及 N 的含量分布，看是否有变化。

5）去除低质量 FASTQ 文件“data1_low.fq”中质量较低的 read 条目，生成高质量 FASTQ 文件“data1_high.fq”。考虑 reads 中 N 的数量以及 reads 中碱基的质量，当 read 中 N 的数量大于 n 或者 reads 中低质量碱基比例超过 r（将质量低于 q 的碱基视为低质量碱基），则去除该 read 条目。用参数“-n 10 –q 20 –r 0.1”运行脚本，得到处理后的 FASTQ 文件“data1_high.fq”；然后，用题 2）中的脚本重新统计各位置上的 ACGT 碱基以及 N 的含量分布，看是否有变化。

4. 实验报告要求

将实验任务的题目、以及对应的代码及图表结果等信息编辑在一个 word 文件中（注意代码缩进，代码用五号字号，其他文字用小四字号）。