小组作业：DNA 分析

学习目标：

- 获得使用循环、条件（if 语句）和字符串操作编写 Python 代码的经验
- 熟悉从命令行运行 Python 程序并使用命令行参数定位数据文件
- 编写 Python 代码来分析 DNA 数据集

背景

你们将使用、修改和扩展程序来计算 DNA 数据的 GC 内容。DNA 的气相色谱含量是 G 或 C 核苷酸的百分比。

DNA 可以看作是核苷酸序列。每个核苷酸是腺嘌呤、胞嘧啶、鸟嘌呤或胸腺嘧啶。这些碱基缩写为 A、C、G 和 T。核苷酸也被称为核苷酸碱基、含氮碱基、核碱基或仅仅是碱基。

生物学家对 GC 含量感兴趣的原因有很多：

- GC 含量可以识别 DNA 中的基因，也可以识别基因的类型。基因的 GC 含量往往高于 DNA 的其他部分。编码区较长的基因具有更高的 GC 含量。
- 具有较高 GC 含量的 DNA 区域需要较高的温度进行某些化学反应，例如复制 DNA 时。
- GC 含量可用于物种分类。

百科有更多关于 GC 内容的信息。这些阅读是可选的，这项作业不需要完成课外阅读。

你的程序将读取由高通量测序器产生的数据文件——这是一台机器，它接收一些 DNA 作为输入，并产生一个包含核苷酸序列的文件作为输出。

以下是特定序列器的前 8 行输出：

@SOLEXA-1GA-2_2_FC30DNN:1:2:574:1722

CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC

+SOLEXA-1GA-2_2_FC30DNN:1:2:574:1722

hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh

@SOLEXA-1GA-2_2_FC30DNN:1:2:478:1745

GTGGGGGTGATGTCCACGATTACGCCGACCGGCTGG

+SOLEXA-1GA-2_2_FC30DNN:1:2:478:1745

hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh

核苷酸数据位于第二行、第六行、第十行等。程序将不使用文件的其余部分，那些是有关测序器和创建核苷酸数据的测序过程的信息。

问题 1：获取文件，添加你们的姓名

通过下载 homework2.zip 文件获得所需的文件。（这是一个很大的下载-请耐心等待。）

解压缩 homework2.zip 文件以创建 homework2 目录/文件夹。你将在这里工作。homework2 目录/文件夹包含：

- dna_analysis.py，你们将完成的部分 Python 程序
- answers.txt，一个用于回答文本问题的文件
- 数据，一个目录。其中包含要处理的数据：
    - *.fastq 文件，从 DNA 测序器输出；这是程序分析的数据
- expected_output,一个包含 dna 分析程序最终结果的示例运行的目录。

你们将通过修改两个文件（dna_analysis.py 和 answers.txt）来完成工作，然后提交修改后的版本。将你们的姓名用注释添加到这些文件的顶部。

每个问题都会要求你们对程序 dna_analysis.py 进行一些更改（或者在 answers.txt 文件中写入文本，或者两者兼有）。当你们这样做时，通常会添加代码到程序中。在处理以后的问题时，不要将之前解决问题的代码修改或删除；最终的程序应该应用到所有问题的代码。

在这两个文件中，请将代码行上的字符数保持在 80 以下，否则你们的文件将变得难以读取。在 Python 中实现这一点的一种技术是通过变量存储子表达式，将大型方程分解为较小的方程。

在作业结束时，我希望 dna_analysis.py 能够产生准确形式的输出：

```
GC-content: ___
AT-content: ___
G count: ___
C count: ___
A count: ___
T count: ___
Sum count: ___
Total count: ___
seq length: ___
AT/GC Ratio: ___
GC Classification: ___
```

其中__用要计算的值替换。当然，每个类别中的确切值将根据你们使用的输入数据而有所不同。我希望程序输出的格式与此完全匹配。

你们可以将输出与 homework2 文件的 expected_output 文件夹中给定的文件进行比较。

你们将以文本文件的形式提交 answers.txt。纯文本是程序员之间交流信息的标准，因为它可以在任何计算机上阅读，而无需安装专有软件。可以使用空闲或其他文本编辑器编辑文本文件。如果使用文字处理器，请确保将文件保存为文本。Windows 用户请避免使用记事本，因为记事本会损坏文件中的行结尾；写字板 或 Notepad++是更好的选择。

问题 2：运行程序

当编写分析数据的程序（或任何其他类型的程序）时，检查程序的正确性是很重要的。一种方法是将程序的输出与以其他方式（如手动或

其他程序）完成的计算进行比较。为此，我提供了 test-small.fastq 文件。这个文件足够小，你们可以在文本编辑器中轻松地打开它并手动计算 GC 内容。然后，运行程序以验证它是否为此文件提供了正确的答案。

对于这个赋值，你们将通过打开一个 shell 或命令提示符来运行程序（*NOT*IDLE 的 Python 解释器）。按照本页上的说明进行操作，本页将教你们操作系统的命令行导航的基本知识。你们应该通过 anaconda prompt 或者 shell 中里通过 cd 目录名 来导航到 teamwork1 目录，然后键入以下命令：

在 Mac/Linux 上：

python dna_analysis.py data/test-small.fastq

在 Windows 上：

python dna_analysis.py data\test-small.fastq

如果出现“无法打开文件'dna_analysis.py'”错误或“没有此类文件或目录”错误，则可能是你们的不在你们的 teamwork1 中，或者你们键入的文件名不正确。

确认程序在 test-small.fastq 上正确运行后，通过执行 6 个命令如：

python dna_analysis.py data / sample_1.fastq

或者如果你是 Windows 用户，

python dna_analysis.py data\sample_1.fastq

通过在上面的命令中将 sample_1.fastq 更改为不同的文件名，在不同的数据文件上运行程序。耐心点，你正在处理大量数据，可能需要一分钟左右的时间才能运行。

（如果你们感兴趣，sample_3.fastq 和 sample_4.fastq 来自肺炎链球菌，sample_5.fastq 来自疱疹病毒。）

如果你们已经使用了输出比较工具（在页面底部引用），你们可能会注意到一些结果与示例结果不同。别担心，这个问题将在**问题 6** 中解决。

在 sample_1.fastq 上运行时，将程序生成的有关 GC 内容的输出行剪切并粘贴到 answers.txt 文件中。例如，你们的答案可能是：

GC-content：0.42900139393

（请注意，这不是你们应该得到的答案，这只是你们的答案应该采用的格式的一个示例。）

问题 3：删除一些行

1. 在你们的程序中，注释掉以下几行：

```
seq = ""
linenum = 0
```

在它们前面加上#字符。重新运行程序，就像在**问题 2** 所做的那样。在 answers.txt 中，解释发生了什么以及为什么发生。

2. 现在，删除 1 中添加的#，将这些行还原到其原始状态。尝试把这一行注释掉会怎么样：

```
gc_count =0
```

在 answers.txt 中解释。还原修改（why？）。

问题 4：按内容计算

扩充你们的程序，以便除了计算和打印 GC 比率（ratio）外，它还计算和打印 AT 含量（content）。A T 含量是 A 或 T 核苷酸的百分比。

计算 AT 含量的两种方法是：

1. 复制检查每个基对的现有循环并对其进行修改。现在有两个循环，其中一个计算 GC 计数（count），另一个计算 AT 计数。或者
2. 在现有循环中添加更多语句，以便一个循环同时计算 GC 计数和 AC 计数。

你可以选择你喜欢的方法。

通过手动计算 test-small.fastq 文件的 AT 含量来检查你们的工作，然后将其与在 test-small.fastq 上运行程序的输出进行比较。

在 sample_1.fastq 上运行程序。将相关的输出行剪切并粘贴到 answers.txt 中。

问题 5：计算核苷酸

扩充你的程序，这样它也可以计算并打印 A 核苷酸的数量，T 核苷酸的数量，G 核苷酸的数量，和 C 核苷酸的数量。

执行此操作时，最多向程序中添加一个额外循环。通过重用现有循环，你们应可以在不添加任何新循环的情况下解决此部分。

检查你们的工作，手动计算 test-small.fastq 文件的结果，然后将它们与在 test-small.fastq 上运行程序的输出进行比较。

在 sample_1.fastq 上运行程序。将输出的相关行剪切并粘贴到 answers.txt（表示 G count、C count、A count 和 T count 的行）。

问题 6：检查数据

对于所给的 11 个.fastq 文件中的每一个，比较以下三个数量：

- Sum count 为综合：A 计数、C 计数、G 计数和 T 计数
- total count 为核苷酸碱基总数
- seq length 为 seq 的长度。你可以用 len（seq）来计算。

换句话说，计算 test-small.fastq 的三个数值，并确定它们是相等的还是不同的。然后对 test-high-gc-1.fastq 等执行同样的操作。

对于至少一个文件，这些值中至少有一处不同。在 answers.txt 文件中，说明不同的文件和数值。（如果每个文件的所有数值都相等，则代码中包含至少一个错误。）在 answers.txt 文件中，编写一个简短的段落来解释这些不同的原因。

解释为什么（或者如果所有指标都相同，则调试代码纠错）可能需要你们执行一些检测工作。例如，要理解这个问题，可能需要将数据文件加载到文本编辑器中并进行检查。我强烈建议你们从数字不完全相同的最小数据文件开始。如果失败，可以手动计算每个计数，然后将手动结果与程序计算的结果进行比较，以确定错误所在。最后一种方法是修改程序或创建一个新程序，分别计算数据文件每行的三个数值（不同于你们现在运行的这个文件）：如果整个文件的数值不同，则它们必须在某些特定行中不同。检查该行将帮助你们理解问题在哪。

如果在**问题 6** 中测量的三个数值都相同，那么在计算 GC 含量时，分母中使用哪一个并不重要。然而事实上，你看到的数字是不一样的。在 answers.txt 文件中，说明这些数值中哪些可以用在分母中，哪些不能用，以及为什么。

如果你们的程序错误地计算了 G C 数值（应该等于（G+C）/（A+C+G+T）），那么在 answers.txt 文件中声明这个事实。然后，返回并更正程序，同时更新 answers.txt 文件中其他地方的所有错误答案。

**如果你们不确定是否计算正确，你们可以将输出与 homework2 文件的 expected_output 目录中给定的文件进行比较。你们尚未完成所有数

值分配，因此你们的输出将不完全相同。但像 GC 含量（GC-content），AT 含量（AT-content）和单独核苷酸计数应该是相同的。在下面的**问题 7** 和**问题 8** 中，你们将在预期的输出文件中生成最后两行输出。

问题 7：计算 AT/GC 比率

有时生物学家使用 A T/G C 比值，定义为（A+T）/（G+C），而不是 GC 含量，定义为（G+C）/（A+C+G+T）。

修改你们的程序，以便它也计算 AT/GC 比率。

通过手动计算 test-small.fastq 文件的结果来检查你们的工作。将它们与在 test-small.fastq 上运行程序的输出进行比较。

在 sample_1.fastq 上运行程序。将输出的相关行剪切并粘贴到 answers.txt（表示 AT/GC 比率的行）。

问题 8：生物分类

气相色谱含量可用于微生物的分类。

修改程序以打印出使用这些分类给出的数据文件中描述的生物体分类：

如果气相色谱含量高于 60%，则认为该生物体“气相色谱含量高（high）”。

如果气相色谱含量低于 40%，则认为该生物体“气相色谱含量低（low）”。

否则，该生物体被视为“中等 GC 含量（medium）”。

生物学家可以使用 GC 含量对物种进行分类，测定 DNA 的熔化温度（对生态学和实验都有用，例如，对 GC 含量高的生物体来说，PCR 更难），以及用于其他目的。下面是一些例子：

天蓝色链霉菌 Streptomyces coelicolor A3(2) 的气相色谱含量为 72%。

酵母(Saccharomyces cerevisiae)的气相色谱含量为 38%。

拟南芥(Arabidopsis thaliana)的气相色谱含量为 36%。

恶性疟原虫（Plasmodium falciparum）的气相色谱含量为 20%。

再次，测试你们的程序是否能在一些具有已知输出的数据上工作。test-small.fastq 文件的 GC 含量较低。我提供了另外四个测试文件，它们的名称解释了它们的 GC 内容：test-medium-GC-1.fastq、test-medium-GC-2.fastq、test-high-GC-1.fastq、test-high-GC-2.fastq。

在你们的程序对所有测试文件都有效之后，在 sample_1.fastq 上运行它。只将程序的相关输出行剪切并粘贴到 answers.txt 中。

提交你的作品

你们快完成了！

在 answers.txt 文件的底部，在“协作”部分，说明哪些学生或其他人（除了课程工作人员）帮助你们完成作业，或者没有人帮助你们完成作业。

通过作业提交页面提交以下文件。

- dna_analysis.py
- answers.txt

在提交作业之前，请确保将输出与 homework2 文件的预期输出目录中给定的文件进行比较。

这时，在提交框里写下你们组花了多少时间思考和完成这份作业。

点下提交键。

现在你们完成了！