**大作业**

**背景介绍：**细胞是动植物的结构和功能的基本单位。人体大约由几十万亿个细胞组成，但即使是同一组织的细胞，在形态、基因表达等方面仍存在着差异。为了探究细胞与细胞之间的差异及其原因，单细胞测序技术应运而生。

**问题：**

1. 在附件1中，我们提供了12类细胞的单细胞ATAC-seq原始数据（数据说明见下）。请对该数据进行处理，设计分类器，并给出测试集标签（10分）。
2. 在附件2中，我们提供了已经处理好的12类细胞的单细胞ATAC-seq相对于870个调控因子的数据。请对该数据进行处理，设计分类器，给出测试集标签（20分），并说明哪些调控对于分类是最有效的（10分）。
3. 针对于附件1，2中的数据，请尝试使用多种方法进行降维、聚类和可视化（25分）。

**注意：**

1. 最终作业形式请以报告的形式给出，并附上代码。请详细写出每一问的处理方法和原因，并给出结论。
2. 附件1的数据维度极高，请考虑对数据进行特征筛选和降维。
3. 附件1的数据保存为稀疏矩阵，python下导入数据请使用scipy.io.mmread。
4. 附件1的训练数据读入python后是一个239255x1000 sparse matrix，即1000个细胞，每个细胞239255维特征；测试数据读入后是则是239255x131 sparse matrix。附件2的训练数据读入后是870x1000 matrix；测试数据则是870x131 matrix。如果发现数据有任何问题，请及时联系助教。
5. 每一问的分数已经标出，剩余35分是报告的分数，所以不管结果如何一定要提交报告，失败也是一次尝试。
6. 问题（1）（2）的测试集的预测标签请以文本的形式给出，答案直接给出一行标签即可，分隔符请务必使用“\t”。标签必须和问题提供的标签相同（即以细胞名称形式给出，切记不要给成数字），顺序也必须和测试集样本顺序相同，否则最后检查测试集正确率时会出现0分的情况。

**评分：**

大作业的评分将以**排名的高低**给出。正确率最高的对应为满分，最低的对应为0分。中间的同学将按照梯度进行插空。具体的方法以下例为说明。

假如共6名同学，对于第（2）问的正确率（20分）分别为20%，50%，50%，75%，90%，100%，那么得分就有5个梯度（正确率相同的合并为一个梯度）：0分，5分，10分，15分，20分。然后依次将得分匹配到各个分数。

如果出现了6名同学正确率为20%，90%，90%，90%，90%，100%，那么将只有三个梯度：0分，10分，20分。

**数据说明：**

[ATAC-seq](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4374986/)是一种2013年提出的实验方法，单细胞ATAC-seq则是对单个细胞进行该实验。下面是对实验所得数据的简单介绍。

人类的染色质根据其accessibility（译为开放性或可接近性）可以分为Open和Close状态。处于Open状态的染色质比较疏散，可以结合一些调控元件（转录因子等），进而调控基因表达。而处于Close状态的染色质固缩，核小体排布紧密，转录不活跃，关闭基因的表达。

ATAC-seq实验通过Tn5转座酶随机插入到Open状态的染色质，然后通过测序的技术来检测插入位置，进而推断出Open状态染色质的位置（如图1），对于分析基因调控网络等有着重要的意义。

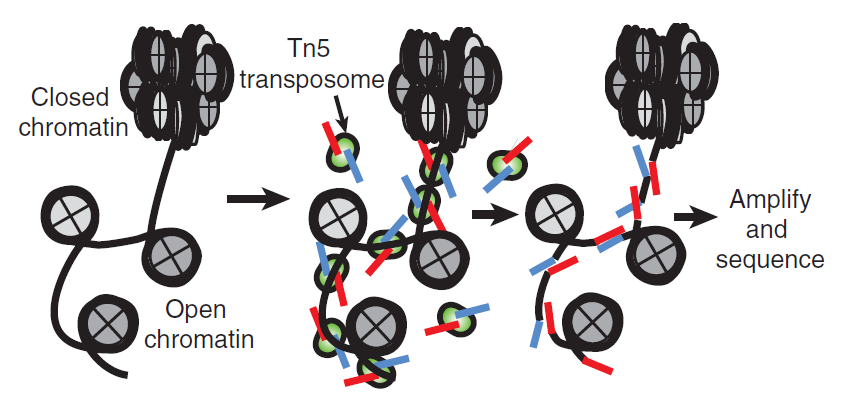


图1 ATAC-seq实验示意图。红色和蓝色代表接头序列，在测序时有重要作用。Tn5可以插入到Open区域。

当拿到数据后，我们可以统计出基因组上每个位置被插入的情况。第（1）问中的peak则是将同一类细胞混合后，依据统计模型找出的插入位点的位置较多的区域（这种区域是可靠的Open区域，如图2所示），最后将所有不同类细胞找到的peaks合并在一起进行筛选，最后得到的就是我们附件1中239255的peak数。

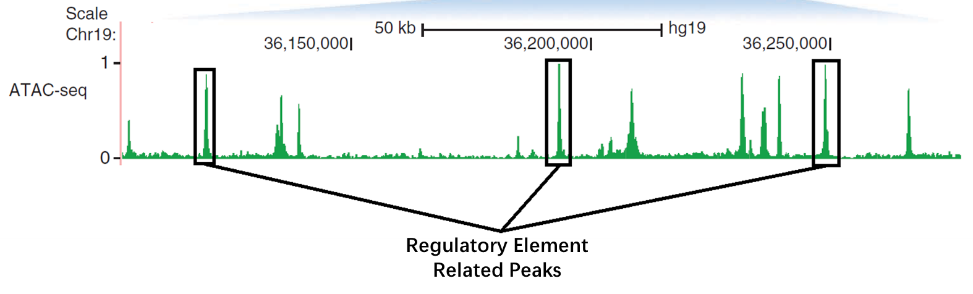


图2 peak示意图，此处的数据时同一类细胞混合之后的结果。横轴表示基因组坐标，纵轴表示归一化后的插入次数，注意附件1中给出的数据是未归一化的数据。黑框是圈出的与某个调控因子相关的三个peaks，附件2中的数据则是希望大家基于不同的调控因子对细胞进行分类和聚类。

附件2中相对于转录因子的数据则是对原始数据处理后的结果。结合已有的数据库（JASPAR等），找出了每个调控因子最相关的peaks，然后将这些peaks的数据进行合并、归一化等处理，即可得到附件2的数据，这一部分数据已经处理好了，大家可以直接拿来进行分类和聚类。

该问题目的是让大家应用模式识别知识到真实数据中。如果对该生物过程不理解，**没有必要**专门去查阅文献，简单的将这个问题看成是小样本高维数据的分/聚类问题进行处理即可。编程上不对大家做要求，可以使用各种包，可以引用别人的方法（但一定要说明出处），鼓励大家多去尝试各种方法（使用近年来比较新颖的方法会考虑酌情加分），失败的结果也可以写在报告中，报告分数还是比较高的。

**参考文献：**

Buenrostro, Jason D., et al. "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position." Nature methods 10.12 (2013): 1213.

Buenrostro, Jason D., et al. "Single-cell chromatin accessibility reveals principles of regulatory variation." Nature 523.7561 (2015): 486.

Ji, Zhicheng, Weiqiang Zhou, and Hongkai Ji. "Single-cell regulome data analysis by SCRAT." Bioinformatics 33.18 (2017): 2930-2932.

Zamanighomi, Mahdi, et al. "Unsupervised clustering and epigenetic classification of single cells." Nature communications 9.1 (2018): 2410.