SNN-Cliq 算法实验简述

概述

SNN-Cliq由两个步骤组成。首先，使用输入数据集 (即gene expression matrix) 构建shared nearest neighbor (SNN) 图。 其次，在SNN图中通过递归地合并quasi-cliques来找到稠密子图。这两个步骤在相应的程序中执行：

1) SNN.m (or SNN.R) 2) Cliq.py

SNN.m

这是用matlab来实现的第一部分的程序。该matlab程序是用来将gene expression matrix转换成SNN图的，我认为是这个图的作用等同于distance matrix，即用来记录细胞与细胞之间的相似度的信息。

输入数据格式examples：

RPKM					
	gene 1	gene 2	gene 3	gene 4	gene 5
cell 1	5.90876	2.24234	4.74742	9.14232	6.92391
cell 2	6.92391	2.24234	9.14232	4.74742	5.90876
cell 3	4.74742	6.92391	5.90876	9.14232	2.24234
cell 4	6.92391	9.14232	2.24234	5.90876	4.74742
cell 5	9.14232	2.24234	5.90876	4.74742	6.92391

即行为cell，列为gene的格式。

输出数据格式examples：

2	**1**	**3.5**
3	1	3.5
3	2	3
4	1	3.5
4	2	3
4	3	3
6	5	3.5
7	5	3
7	6	3.5
8	5	3

Cliq.py

这是用python实现的第二部分的程序。该python程序实现聚类的过程，即搜寻最大最近子图合并的过程。

具体可调参数如下：

- r	**用于调整聚类结果中每一类的粒度，r越小每一类中包括的细胞越多**
- m	与参数r的效果类似，控制合并过程中什么时候进行剪枝
- n	SNN图中数据点的数量，可以不给

输入数据格式即为SNN.m程序的输出数据格式。

输出数据格式examples：

6
6
6
6
5
5
5
5

进行的实验

- 首先，使用之前的代码对Human_Cancer_Cell这个数据集进行整理，得到行是cell，列是gene的expression matrix (为了符合程序要求)，矩阵大小为86*3960。
- 接下来，在matlab中导入SNN.m文件，首先使用命令导入expression matrix

  data=importdata('Human_Cancer_Cell_Expression_Matrix.txt');

  然后使用命令得到SNN图, k为最邻近邻居的数量(使用默认值3)

  SNN(data.data, 'SNN_graph.txt',k);
- 最后，使用python程序Cliq.py对SNN_graph.txt进行聚类操作，参数全部使用原文推荐的默认参数，得出结果，具体结果在文件Cluster_Result.txt中

实验结果分析

- 目前得到的结果是86个cell被全部聚类为11个类簇，而原本的数据集内86个细胞总共被分成10个类别，目前还在想办法区分哪些细胞被正确分类