1. 文章简介

这篇文章是基于现有的双聚类方法识别的基因往往难以解释等问题，提出了BiCoN的算法，即将网络作为双聚类的约束，将双聚类限制在分子相互作用网络中连接的功能相关基因上，以基因表达数据作为输入，并将患者分为两个亚组，为每个亚组确定一个基因子网络。该算法可以最大化两个亚组患者之间基因表达的差异，使得能够同时精确定位负责患者分组的分子机制。

2. 数据获取

这篇文章的实验数据分为如下两个部分。

（1）基因表达数据

- TCGA泛癌数据集，包括有腔或基底部乳腺癌的患者。TCGA乳腺癌数据通过UCSC-Xena获得：https://xenabrowser.net/。
- GSE30219-来自GEO的非小细胞肺癌数据集，用于腺癌或鳞状细胞癌患者。 NSCLC数据集使用GEO2R获得：https://www.ncbi.nlm.nih.gov/geo/geo2r/，其登录号为GSE30219。

这两个数据集与包含注释癌症亚型的相应元数据一起检索。

（2）分子相互作用网络

使用了BioGRID (version 3.5.176) 中H. Sapiens 的PPI。这个网络由16830个基因之间的343563个独特的相互作用组成。

这些数据集可以在上述网站中下载，如：

也可以直接使用作者在github上留下的链接进行数据下载：

There are 2 examples of gene expression datasets that can be placed in the "data" folder

- GSE30219 - a Non-Small Cell Lung Cancer dataset from GEO for patients with either adenocarcinoma or squamous cell carcinoma.
- TCGA pan-cancer dataset with patients that have luminal or basal breast cancer. Both can be found here

BiCoN_data

文件

下面将截取三个数据集的一部分。

- TCGA泛癌数据集：

	A	B	C	D	E	F	G	H
		TCGA-E2-	TCGA-BH	TCGA-AN	TCGA-A7-	TCGA-E9-	TCGA-E9-	TCGA-AR
6883		-0.04219	0.6179346	1.0703416	0.7998117	0.4823714	-1.50814	-1.287304
90557		0.9341749	-1.353575	-0.038278	0.1631239	-1.030677	-0.677466	-1.174363
26038		-1.388329	0.0237522	-1.388329	2.0096627	0.3283737	0.1886332	0.4496736
6843		-0.908202	-0.137955	-1.000439	-0.414562	0.4680238	-1.255688	-1.372737
9064		0.2623244	0.1200251	-0.065785	-0.848431	-0.011063	1.2545344	-0.456942
201973		-0.262488	-1.142425	0.8182881	0.8114533	0.620955	1.5603028	-1.760362
388325		-1.771162	-1.633238	0.588366	0.5840212	-0.556598	-2.18357	-0.099123
643376		-0.023675	-0.023675	-0.023675	-0.580533	1.1611968	-1.505061	-1.505061
7752		-0.703116	-0.17209	0.0950799	0.8981654	-0.37917	-3.11048	0.8492737
144717		0.4865413	0.3371181	1.3191643	1.0519185	1.0234637	-1.60381	0.4379251
90527		-1.0427	0.2527091	-0.427092	1.613108	0.3607973	-1.838778	-1.0427

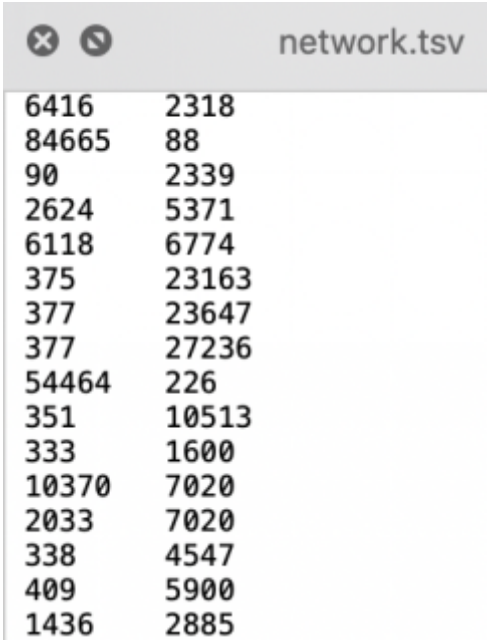
- 非小细胞肺癌数据集GSE30219：

	A	B	C	D	E	F	G	H	I	J
1		GSM74805	GSM74805	GSM74806	GSM74806	GSM74806	GSM74806	GSM74806	GSM74806	GSM74807
2	6311	7.3487663	6.946423	6.474783	6.7287264	7.314376	7.283062	7.636509	6.613424	7.2604547
3	133584	6.5662155	5.7979374	5.630046	6.9633284	6.5118327	6.642319	6.2873397	5.6970778	6.057786
4	283165	3.8878543	3.7710621	3.958817	3.739808	3.3642237	3.9328575	3.8695502	3.4576929	3.562493
5	255082	3.420246	3.76282	3.7249117	3.4283502	3.36081	3.2705097	3.289275	3.4421132	3.071813
6	26505	6.5806975	6.9130874	6.695494	6.289358	7.1318245	6.585806	6.565159	6.510072	6.9293737
7	60401	5.3771124	4.4122896	5.652992	5.382945	4.997994	5.1870623	5.2857995	5.2696733	5.489599
8	4170	6.7456437	7.3968853	7.8013723	7.337647	6.8374387	6.92385	7.1598335	7.7071865	6.8175298
9	390892	3.3782783	3.5057878	3.4751196	3.3607326	3.4166331	3.1645	3.5020025	3.2778106	3.4790726
10	57677	3.1521454	2.8597941	3.2013474	3.1883752	3.2491946	3.5539403	3.5143259	3.2170575	3.3571987
11	83900	2.9318435	2.9121494	2.9820054	3.2022967	2.9802587	2.7716067	2.8311498	3.1226323	2.9211993
12	64073	8.817757	9.743748	10.525072	8.646335	6.521598	7.3323746	11.123371	4.4957433	7.106431
13	10016	7.539694	6.60214	6.397893	7.1896505	7.806694	5.835632	7.1447673	6.9832544	6.4247656
14	92342	7.1315603	7.2222023	7.4213443	7.4834046	7.163394	7.268465	7.128778	7.2334795	7.111697
15	2121	5.538143	5.101856	5.522111	5.4132133	5.6213865	5.5385733	5.722379	5.4568667	5.206876
16	54212	2.8537483	2.911261	3.1684833	2.9829319	2.776379	3.1347811	2.7891147	2.8514624	2.8028822

以上两组数据集中，一行代表某一种基因，一列代表某一个病人。第一列是基因的ID号。

- PPI网络：

  表格中有两列代表两个相互作用的基因。注意输入数据中不应存在标题。

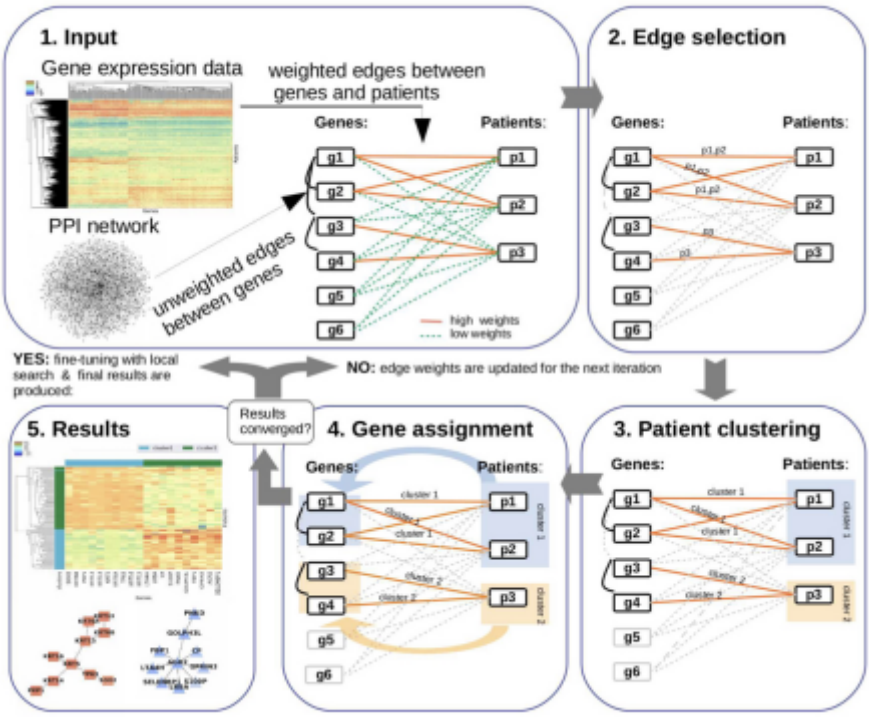


3. 算法

3.1 算法总览

BiCoN是一种启发式算法，它可以发现差异表达的子网络，从而从机制上解释患者分层。这种组合问题可以通过各种元启发式框架来解决，例如遗传算法或群体智能。这篇文章选择了蚁群算法（ACO）作为探索搜索空间和局部搜索的主要框架，以确保最终解的局部最优性。将蚁群算法与局部搜索相结合，可以有效地求解复杂组合优化问题，并导致与ACO或单独的本地搜索相比的显著改进。由于在ACO处理类似问题之前已有良好的经验），这个方法预计与本地搜索的结合将产生高质量的结果。

蚁群算法（ACO）是一种自然启发的概率方法，用于解决计算问题，可以归结为通过图寻找最优路径。这里使用ACO来确定每个患者的一组相关基因，然后将这些基因聚合到一个全局解决方案中。下图将辅助说明来描述算法的详细过程。简单来说，蚂蚁在关已关联的图上分三个阶段移动，重复进行直到收敛：



- 第一步是将基因表达数据与PPI网络输入，建立二部图。以基因表达数据为基础，将数据分为基因和病人两个部分，基因表达的情况数据作为有权边连接在基因和病人之间。PPI网络中的关系作为无权边连接在基因与基因之间。
- 第二步将为每个患者确定最相关的特征，其中边缘用患者 ID 进行注释。
- 第三步运用第二步中的结果对患者进行聚类。
- 第四步将基因通过蚁群算法分配到相应的簇，若结果收束则进入第五步，否则回到第二步进行迭代。
- 第五步，输出最终结果。

3.2 详细描述

下面对算法进行详细的描述：

1. 蚂蚁在与患者节点高度相连的节点内执行随机游走，并通过选择与患者最相关的基因（第二步的橙色边，即高权值的边）根据目标函数进行贪婪选择。其中，为某个病人选择一个基因的概率取决于基因表达值（边的权值）和蚂蚁的“记忆”的组合信息。蚂蚁的“记忆”存储在信息素矩阵中。蚂蚁在行走过程中将记录其信息素，用来标识自己的行走路径。由于之前的每一次迭代都会产生当前的最优分配策略，即局部最优解，从而在更优路线中，蚂蚁留下的信息素浓度也就越高。增加迭代次数，局部最优解就能接近全局最优解。
2. 然后用k-means算法将选定的基因用于聚类，其中k=c=2。在步骤4提取每个患者群的相关基因。通过目标函数得分来评价候选解。
3. 最佳解用于更新信息素和概率矩阵，以便下一次迭代。
4. 当获得最佳解时，进行局部搜索以获得可能的局部改进，即迭代地将更改应用于子网（例如节点插入、删除或替换），并保持导致目标函数最大化的更改。这使得能够检索到鲁棒和稳定的解决方案，以及确保局部最优。

3.3 代码实现

具体的伪代码实现如下，主要分为三个部分：

- 数据准备（数据预处理以及初始化信息素矩阵）
- 迭代搜索
- 结果输出

Algorithm 1: BiCoN

input : $X^{n\times m}$ - expression matrix , $G^{n\times n}$ - molecular interaction network, a - pheromone importance, b - heuristic information importance, cl - cost limit, L_{min}- minimum solution subnetwork size, L_{max}- maximum solution subnetwork size, ρ - evaporation coefficient, K - number of ants, $IterMax$ - maximum number of iterations allowed, ε - convergence criteria

output: solutionFinal that consists of genes clusters U, patients clusters V

```
Initialization;
H ← HeuristicMat(X);
C ← CostCalc(H);
T ← InitialPher();
p ← ProbUpd(t,H,a,b);
IterCount ← 0;
scores ← [];
while (sMax-sMean> ε) AND (IterCount <IterMax) do
    foreach i = 1 to K do
        paths ← [];
        foreach j = 1 to m do
            N ← SearchRad(j,H);
            path ← RandomWalk(j,G,cl,N,P);
            paths.add(path);
        end
        patientsClusters ← ClusterPatients(paths,X);
        genesClusters ← ClusterGenes(paths,
         patientsClusters);
        genesClusters ← NetReduce(genesClusters);
        s ← Score(genesClusters,
         patientsClusters,L_min,L_max);
        scores.add(s);
    end
    sMax ← Max(scores);
    sMean ← Mean(scores);
    Update parameters with respect to the best solution found;
    T ← PherUpd(T,sMax,genesClusters,patientsClusters,ρ);
    p ← ProbUpd(T,H,a,b);
end
solutionFinal ←LocalSearch (patientsClusters,genesClusters)
```

实现时，首先利用setup.py文件将环境安装好，在文件目录下执行：

```
python3 setup.py install
```

需要设置输入路径，这里选用提供的非小细胞肺癌数据集GSE30219与PPI网络：

```
path_expr,path_net ='/data/gse30219_lung.csv', '/data/biogrid.human.entrez.tsv'
```

然后对数据进行预处理：

```
GE,G,labels, _= data_preprocessing(path_expr, path_net)
```

运行该模型（参数设置为`L_g_min = 10`，`L_g_max = 15`）：

```
model = BiCoN(GE,G,L_g_min,L_g_max)

solution,scores= model.run_search()
```

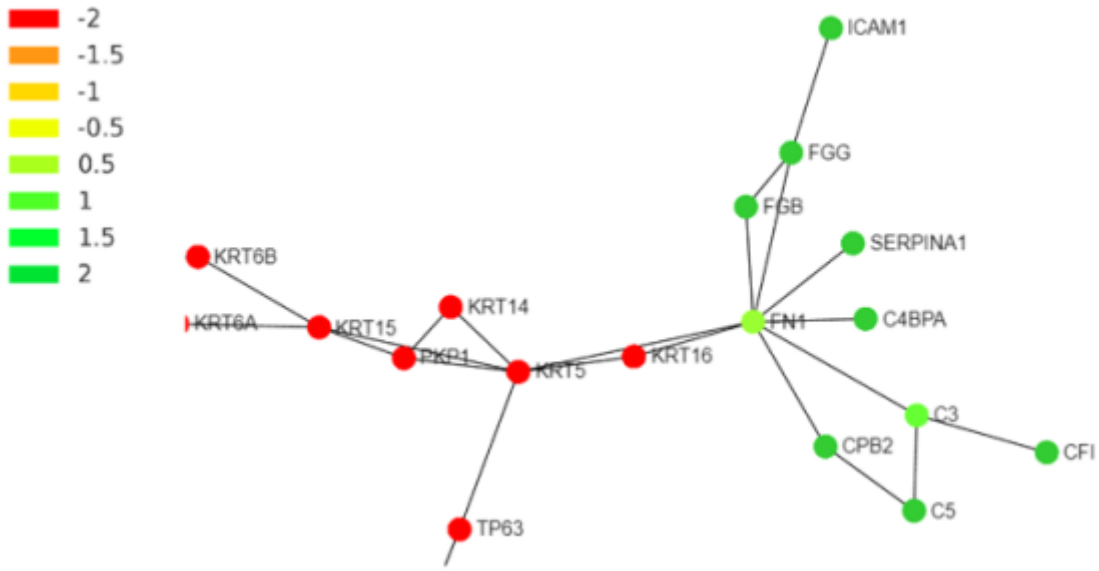
4. 结果

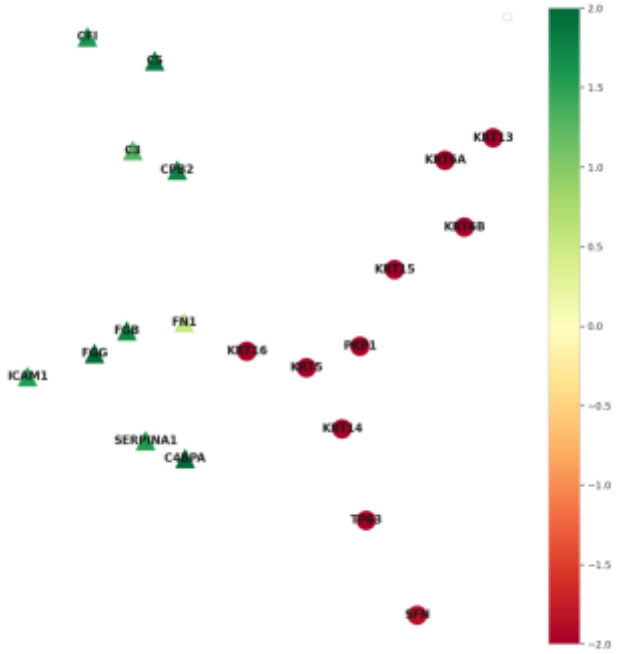
结果输出的文件包含`gene1`，`gene2`，`patient1`，`patient2`四列数据，均存在一行中。显示最终双聚类的结果。在`result.csv`文件中。可以根据github上的使用手册具体访问病人与基因的信息。

在这里，我们调用`/results`中的图表进行结果展示。

（1）提取的子网

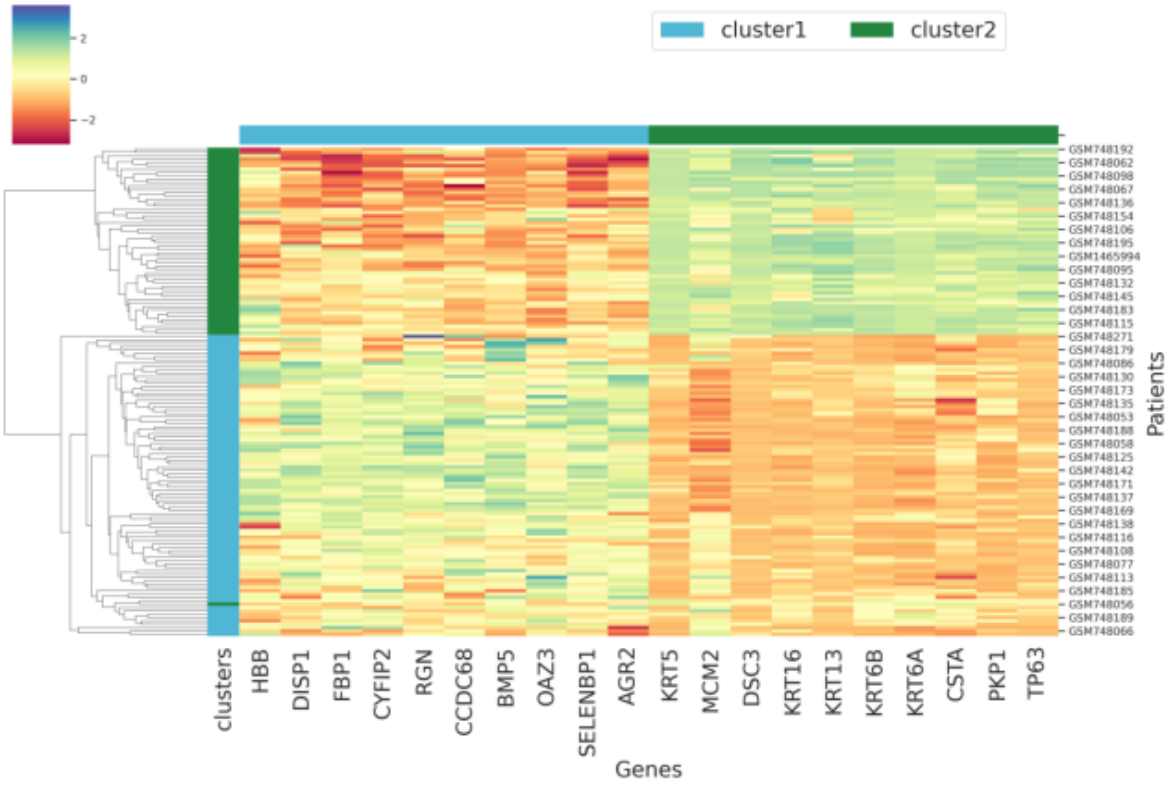
节点的颜色对应于两组患者表达水平的平均差异。图中显示了分析基因之间的相互作用。





（2）热图

热图显示了生成的双聚类，即部分基因组与特定患者的直接关系。



（3）收敛行为

展示了训练过程中，评分随迭代次数的变化，可以看到在迭代15次之后，结果大致收敛。