Identification of cancer related gene regulatory towards alternative splicing and gene pathways

一、参考文章以及发现的不足：

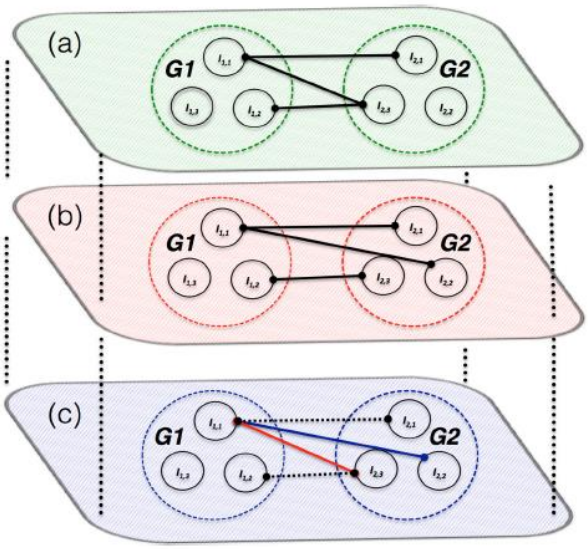
参考的文章：

SpliceNet: recovering splicing isoform-specific differential gene networks from RNA-Seq data of normal and diseased samples, Hari Krishna Yalamanchili, et al.

1.1 原文研究的问题：

比较正常和疾病情况下两个基因之间的 **Isoform 交互网络**的**差异**. （数据来源是 TCGA）

- 这种差异是通过边的**有无的变化**来描述的.
- 核心是建立交互网络，具体落实是**计算两个 Isoform 之间的关联度**，关联度大则连边，否则无边.



意义：在 Isoform 层面上认识正常和疾病之间的差异，发现疾病背后的 Isoform 表现出的行为.

1.2 具体操作（概括性描述）：

✧ 数学表示：

- 一个$gene$所能拥有的全部$exon$有p个，而每个基因又有n个样本.
- 因此可以用一个$p \times n$的矩阵X来描述一个$Isoform$表达
  1. 矩阵中的元素自然是$exon$的表达量.

✧ 根据不同$gene$的$Isoform$两两之间的关联度决定是否在网络中加边：

- 实际上是计算两个矩阵X^i、X^j的 Correlation，然后根据 Correlation 的大小决定是否加边.
- 具体步骤：
  1. 建立 co-expression matrix：$X = \left[X^i, X^j\right]^T$
  2. 假设$X \sim N(\mu, \Sigma)$，其中协方差矩阵$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$
  3. 问题最终转化为假设检验：$H_0: \Sigma_{12} = 0 \quad vs \quad H_1: \Sigma_{12} \neq 0$
     - 接受H_0不加边，接受H_1则加边.
  4. 为了降低复杂度，对Σ_{12}求 LDT.
     - $L_n = tr(A_{21}A_{11}^{-1}A_{12}^{-1}A_{22}), \quad A_{ij} = (n-1)\Sigma_{ij}$
     - 由于研究目标从X转化成了 LDT，因此也需要将$X \sim N(\mu, \Sigma)$等价转化成 LDT 的渐进分布.
  5. 最后在 LDT 的分布中求 P-value 来决定接受哪个假设.

✧ 利用以上方法分别建立正常样本和疾病样本的$Isoform$交互网络，最后对网络进行比较得出差异.

1.3 存在的问题：

1. 通过 Correlation 计算的关联度并不能完全描述直接交互，间接交互也会包含在其中.
2. 由于目前 Isoform 层面还没有 Gold Standard，还没有检测 Isoform 网络好坏的数据支持.
3. 参考文献最终是用 Isoform 的交互差异来推测基因交互的差异，因此 Isoform 层面有几率将噪音和误差进一步传递上去.
4. 计算步骤多且复杂（特别是在 Isoform 层面上），过程中放大噪音的机会大大增加：
   a) 求Σ_{12}的 LDT 值极为复杂；
   b) 求 LDT 渐进分布的参数也非常复杂.

二、思考和改进（动机和目的）：

1. 由于现实中研究基因调控比直接研究 Isoform 交互更直观和普遍，因此可以直接研究 gene-gene 之间的交互关系.
   - 但还是利用 level3 的 exon 表达量来计算.
2. 基于此，可以研究 pathway 中已经确定存在交互关系的 gene 对，然后研究这些 gene-gene 交互在正常和疾病状态下的差异.
   - 最后在 pathway 中找出那些 normal-disease 下存在显著差异的通路.
3. 放弃使用大量的矩阵运算，改用其它数学模型来描述 gene-gene 之间的交互关系.

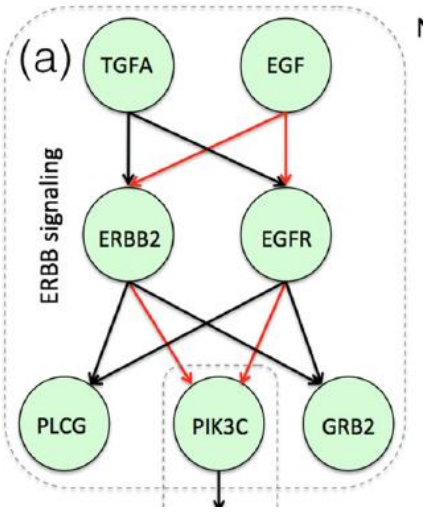
三、研究的新问题以及新方法：

3.1 问题描述：

基于选择性剪切以及基因 pathways 鉴别疾病（癌症）相关的基因调控.

核心要素：

1. 基于基因 pathway：研究的目标基因都来自 pathways.
   - 在 pathways 中找到一条**已知交互**的 gene-gene 对.
   - 然后判断该对基因的交互在正常和疾病状况下是否存在差异.
   - 如果存在差异则在**原 pathway 中标记**这条通路（高亮标记）.



2. 基于选择性剪切：识别差异的过程中利用 gene 中 isoform 的各个 exon 的表达量.

3.2 建立初步的数学模型：

			Normal Sample				Disease Sample			
			S_1	S_2	…	S_t	S_1	S_2	…	S_t
$Gene_1$	Iso_{11}	$exon_{11}$								
		$exon_{11}$								
		⋮								
		$exon_{1p}$								
	Iso_{12}	⋮								
	⋮	⋮								
	Iso_{1m}	$exon_{11}$								
		$exon_{11}$								
		⋮								
		$exon_{1p}$								
$Gene_2$	Iso_{21}	$exon_{21}$								
		$exon_{21}$								
		⋮								
		$exon_{2q}$								
	Iso_{22}	⋮								
	⋮	⋮								
	Iso_{2n}	$exon_{21}$								
		$exon_{21}$								
		⋮								
		$exon_{2q}$								

这样便获得了两个“基因 co-expression”矩阵：

1. 红色部分是正常样本的 gene1-gene2 co-expression matrix，数学符号是：NM_mat
2. 绿色部分是疾病样本的 gene1-gene2 co-expression matrix，数学符号是：DE_mat

◆ 目标就是比较两个矩阵是否有“差异”，如果有，则在 pathway 中标记 gene1-gene2 通路

3.3 差异比较：

- 就是比较NM_mat和DE_mat是否存在差异，这里有待选的若干思路：

1. K-S Test，Kolmogorov–Smirnov Test：
   - 将 sample 投影到同一维度上，然后对两个一维向量进行差异判断.
   - 可以有效降低计算复杂度，不会用到大量的、复杂的矩阵运算.
2. MRDM，Multi-Dimensional Regulatory Module，是一种探测 MRDM 的算法，但其中计算一对基因交互关系的方法值得借鉴.
   - 参考文献：

   Identifying multi-layer gene regulatory modules from multi-dimensional genomic data, Wenyuan Li, et al.
3. ...其它方案有待发掘.

三、研究计划：

3.1 深入研究原命题：

1. 继续深入阅读原参考文献，发现新的弱点以及可以改进的思路.
2. 参考其它相关文献，在问题框架和具体算法方面寻找更优的解答.

3.2 数据源：

1. 学习 TCGA 数据库，弄懂常用的数据格式以及 API 工具.
2. 在原文的基础上进一步处理数据，得到可以适用于本文算法的数据.

3.3 尝试：

1. 先选择一个实现度强的方法进行初步实验.
2. 得到初步结果后进行反思和总结.
3. 进一步探索其它实用方法以及新潮的方法.

3.4 成果：

1. 对所有的方法进行总结和对比.
2. 尝试对方法之间的优劣进行解释.
3. 整理素材，撰写论文.