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

一、参考文章以及发现的不足：

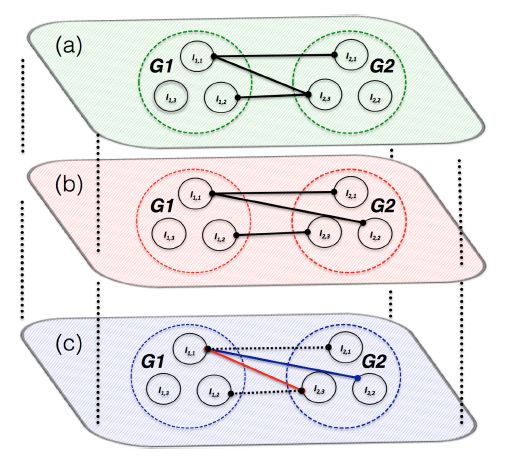
参考的文章：

*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：
  + 一个所能拥有的全部有个，而每个基因又有个样本.
  + 因此可以用一个的矩阵来描述一个表达
    1. 矩阵中的元素自然是的表达量.
* 根据不同的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、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：
    2. 假设，其中协方差矩阵
    3. 问题最终转化为假设检验：
       - 接受不加边，接受则加边.
    4. 为了降低复杂度，对求LDT.
       - 由于研究目标从转化成了LDT，因此也需要将等价转化成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. 在样本数量较少的情况下Correlation极不可靠！
3. 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.
4. 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.
5. 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：
   1. 求的LDT值极为复杂：
   2. 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.

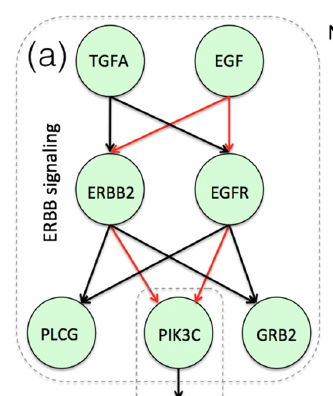
三、研究的新问题以及新方法：

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）.



1. 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** | | | |
|  |  | … |  |  |  | … |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

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：
2. 绿色部分是疾病样本的gene1-gene2 co-expression matrix，数学符号是：

* 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：

* 就是比较和是否存在差异，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.*

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