基于异质网络分析方法的关键长非编码RNA的计算与识别

汇报人：张颖颖

指导老师：中南民族大学 朱剑林老师
天津大学 杜朴风老师

2020.5.29

中南民族大学
South-Central University For Nationalities

中南民族大学
South-Central University For Nationalities

目录 CONTENTS

一、研究背景

随着人类基因组计划的完成，人们发现人类基因组中仅有1.5%负责编码蛋白质基因，而剩下占98.5%比例的DNA序列也会在某个时刻进行转录，将会产生大量非编码RNA（ncRNA），这表明ncRNA在复杂生物体中可能起着重要的调控作用。

长非编码RNA(Long Noncoding RNA，lncRNA) 是一类长度大于200个核苷酸且几乎没有蛋白质编码能力的非编码RNA。

关键长非编码RNA是指在众多长非编码RNA中功能非常重要的长非编码RNA。

二、研究方法与思路

构建lncRNA-protein异质信息网络

- 从NPIter v4数据库中下载了lncRNA-protein关联数据，对数据进行筛选，预处理等操作
- 基于已知的关联关系，构建lncRNA-protein异质信息网络

HeteSim算法计算lncRNA之间的关联度

- 选择lncRNA-protein-lncRNA（LPL）作为元路径
- 将关联关系转化lncRNA-protein的0/1邻接矩阵
- 将0/1矩阵分别按照行向量和列向量进行标准化得到2个转移概率矩阵
- 2个矩阵相乘，对结果进行标准化处理，将最后的值化为【0,1】之间
- 得到lncRNA-lncRNA关联得分矩阵

二、研究方法与思路

构建lncRNA-lncRNA网络，用网络节点中心性识别关键lncRNA

- 用lncRNA-lncRNA关联得分矩阵构建lncRNA-lncRNA网络
- 用网络节点中心性识别网络中的关键lncRNA

实验结果评估

- 采用GIC分数作为参照，验证网络节点中心性识别关键lncRNA的有效性
- 各个中心性计算得出的值按照从大到小的顺序排列，分别取Top100，Top200，Top300,Top400为关键lncRNA
- 分别取GIC分数0.45，0.55,0.65,0.75,0.85作为阈值来区别关键lncRNA和非关键lncRNA（如值≥0.45，为关键lncRNA；值<0.45，为非关键lncRNA）
- 采用灵敏度、特异性、阳性预测值、阴性预测值、F-measure、准确性以及ROC曲线来评估实验结果

三、主要难点与结论 难点

A

无直接的lncRNA-lncRNA关联信息

使用lncRNA-protein数据集，构建lncRNA-protein异质信息网络，采用HeteSim算法，以lncRNA-protein-lncRNA为元路径，计算lncRNA与其他lncRNA之间的关联度，得到关联度矩阵

B

如何识别关键lncRNA

构建lncRNA-lncRNA虚拟网络，在网络中认为该节点越重要则越有可能是关键lncRNA，则用网络节点中心性来识别关键lncRNA。

C

各个网络节点中心性计算出来的值如何设定阈值?

由于使用网络节点中心性计算出来的是一个具体的值，并没有区分关键lncRNA。所以将这些值按照从大到小的顺序排列，分别取排名Top100，Top200，Top300，Top400作为识别出来的关键lncRNA

D

无关键lncRNA数据库

因为没有准确的关键lncRNA库来对结果进行验证，故采用GIC分数进行代替。GIC分数可以根据lncRNA的序列信息计算出lncRNA的重要性。分数越高，重要性越高，则越有可能是关键lncRNA。

三、主要难点与结论 结论1

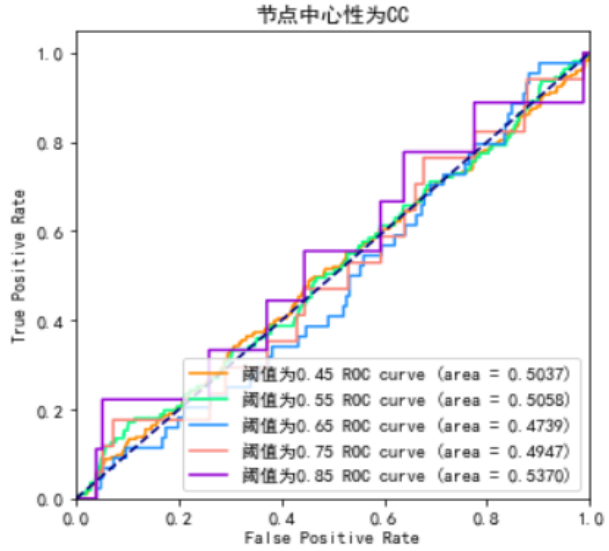
2、GIC分数分别取0.45, 0.55, 0.65, 0.75, 0.85作为阈值，在各个中心性的阈值取Top100，分别计算各个中心性的灵敏性、特异性、阳性预测值、阴性预测值、F-measure和准确值，并进行比较

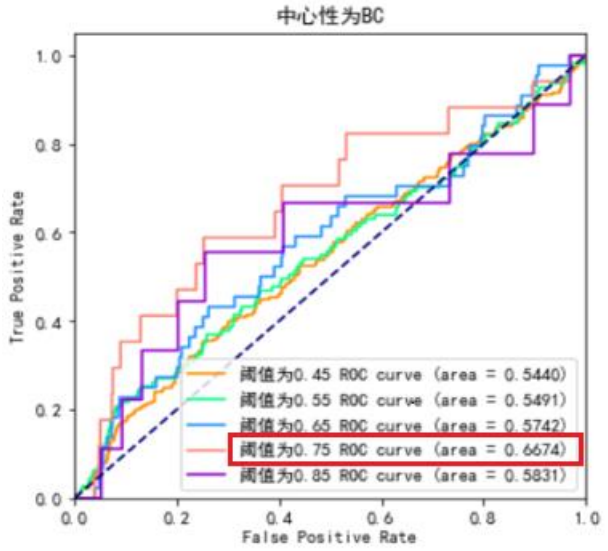
结果：从表5-1可以看出，除了阈值为0.85时，CC的各项指标的值要高于其他节点中心性。其他情况下，BC的各项指标均比其他中心性要高。由此可见，BC的识别效果要更好一些。

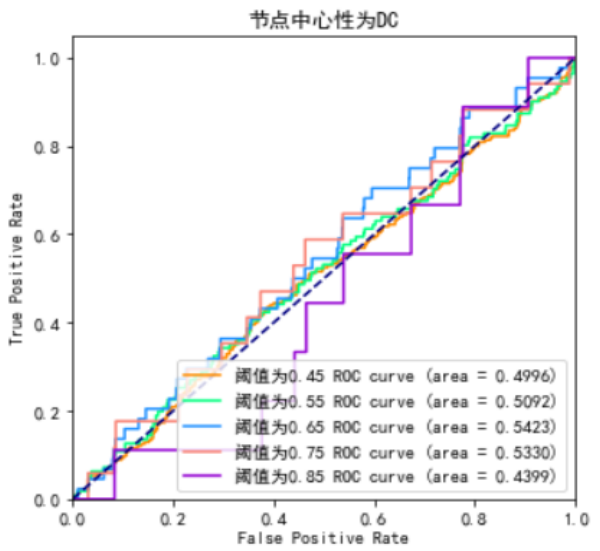
表 5-1：各个节点中心性的灵敏性、特异性、阳性预测值、阴性预测值、F-measure 和准确性值情况

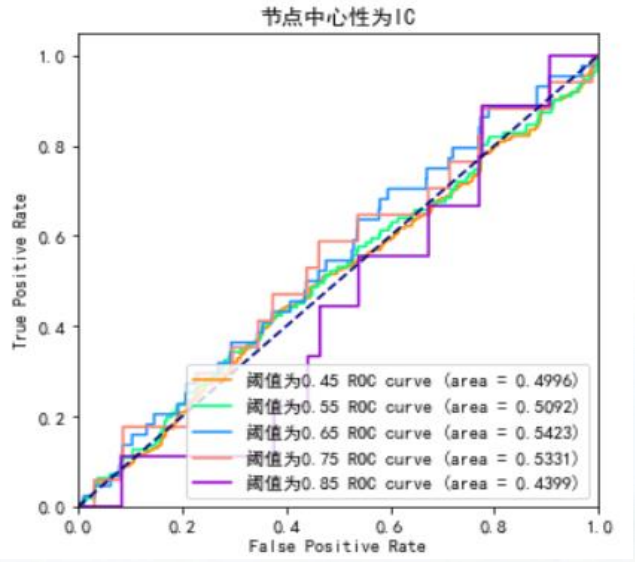
阈值	中心性	Sensitivity (SN)	Specificity (SP)	Positive Predictive Value(PPV)	Negative Predictive Value(NPV)	F-measure (F)	Accuracy (ACC)
0.45	CC	0.102	0.911	0.23	0.795	0.142	0.743
	BC	**0.142**	**0.921**	**0.32**	**0.803**	**0.196**	**0.759**
	DC	0.089	0.907	0.2	0.792	0.123	0.738
	IC	0.089	0.907	0.2	0.792	0.123	0.738
	EC	0.097	0.909	0.22	0.793	0.135	0.741
0.55	CC	0.135	0.913	0.15	0.903	0.142	0.833
	BC	**0.196**	**0.92**	**0.22**	**0.909**	**0.208**	**0.845**
	DC	0.108	0.91	0.12	0.9	0.114	0.828
	IC	0.108	0.91	0.12	0.9	0.114	0.828
	EC	0.108	0.91	0.12	0.9	0.114	0.828
0.65	CC	0.091	0.908	0.04	0.959	0.056	0.875
	BC	**0.2**	**0.913**	**0.09**	**0.964**	**0.124**	**0.883**
	DC	0.136	0.91	0.06	0.961	0.083	0.879
	IC	0.136	0.91	0.06	0.961	0.083	0.879
	EC	0.136	0.91	0.06	0.961	0.083	0.879
0.75	CC	0.176	0.909	0.03	0.986	0.051	0.898
	BC	**0.278**	**0.911**	**0.05**	**0.987**	**0.085**	**0.901**
	DC	0.176	0.909	0.03	0.986	0.051	0.898
	IC	0.176	0.909	0.03	0.986	0.051	0.898
	EC	0.176	0.909	0.03	0.986	0.051	0.898
0.85	CC	**0.222**	**0.909**	**0.02**	**0.993**	**0.037**	**0.903**
	BC	0.10	0.908	0.01	0.991	0.018	0.901
	DC	0.111	0.908	0.01	0.992	0.018	0.902
	IC	0.111	0.908	0.01	0.992	0.018	0.902
	EC	0.111	0.908	0.01	0.992	0.018	0.902

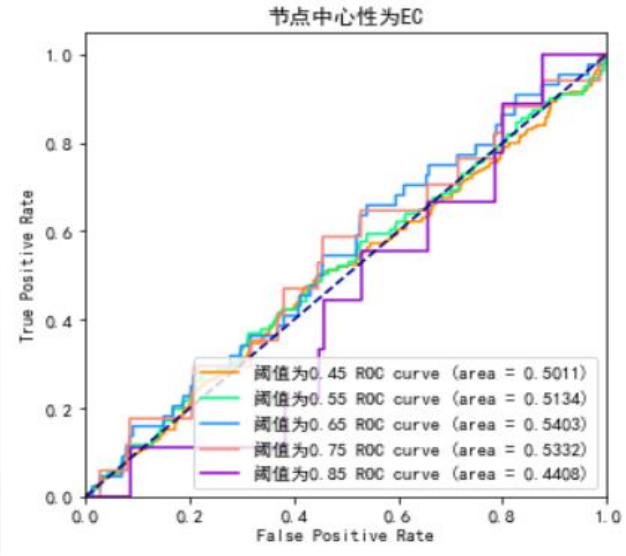
三、主要难点与结论 结论2





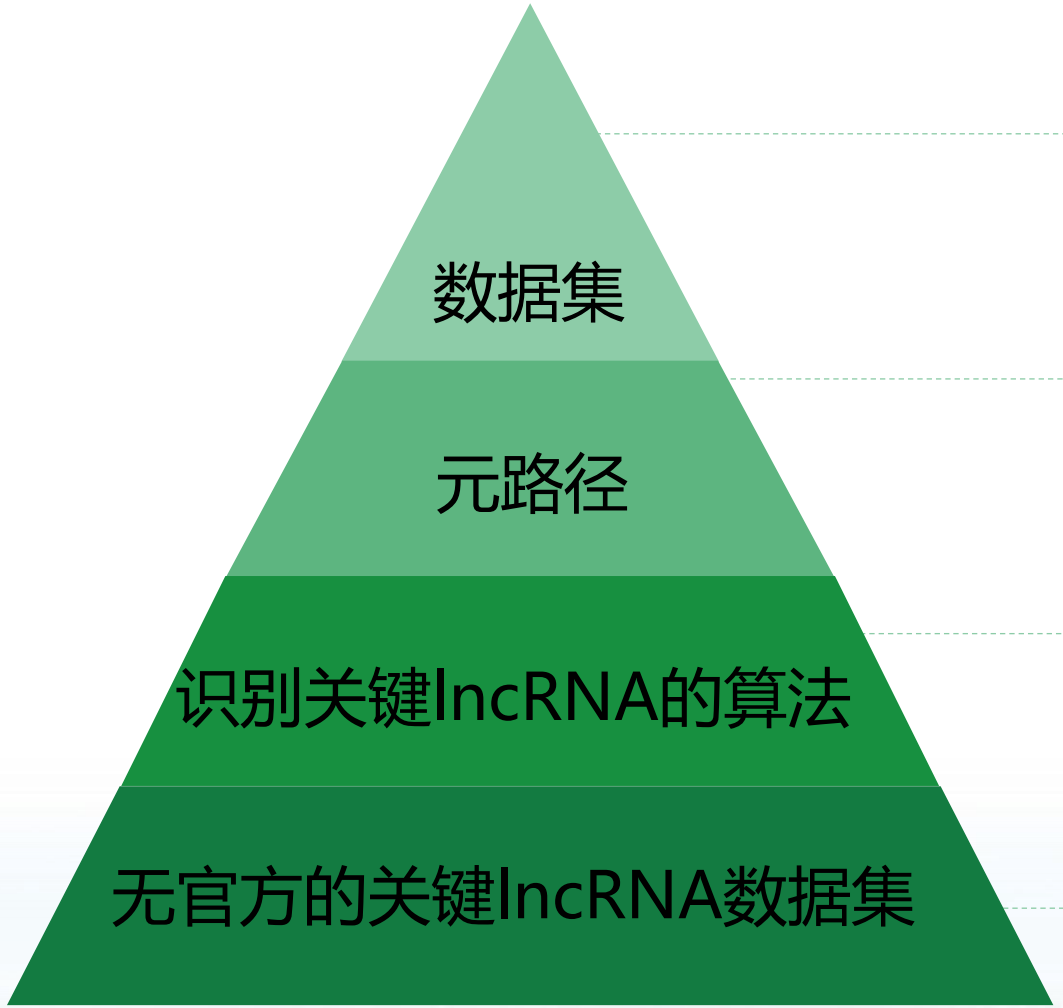






3、由各个节点中心性的AUC值对比可以看出，在阈值为0.45、0.55、0.65、0.75和0.85时，BC的AUC值均最大，分别为0.5440、0.5491、0.5742、0.6674和0.5831，所以还是BC的识别效果相对较好，其中GIC分数阈值取0.75时，BC的识别效果最佳。

四、总结 不足之处



本文数据集只采用了lncRNA-protein关联数据，过于单一。可加入protein-protein数据集，共同构建网络。

本文采用lncRNA-protein-lncRNA作为元路径，也可加入protein-protein关联数据，使用lncRNA-protein-protein-lncRNA作为元路径

本文采用网络节点中心性的方式识别关键lncRNA，也可采用别的方式来识别计算。

本实验无真实准确的关键lncRNA数据库来判断识别结果的准确性，只能采用GIC分数作为代替，这也是导致实验效果不好的原因之一。

四、总结 创新点

1. 采用HeteSim算法在lncRNA-protein异质信息网络中计算得出lncRNA-lncRNA的间接关联关系

2. 基于复杂网络的拓扑结构，采用网络节点中心性的方法在加权lncRNA-lncRNA网络中识别关键lncRNA

3. 采用GIC分数作为是否为关键lncRNA的参照，弥补无关键lncRNA数据库的缺憾

基于异质网络分析方法的关键长非编码RNA的计算与识别

谢谢观赏

THANK YOU FOR WATCHING

汇报人:张颖颖

指导老师：中南民族大学 朱剑林老师
天津大学 杜朴风老师

2020.5.29