数据挖掘研究方案

--基于蛋白质互作网络中的极大团预测疾病蛋白质

研究组成员 34 组

组长 孙一鸣 14051628 email: bearsugar@foxmail.com

组员 吕宏愿 14051624

组员 毛煊 14051625

数据

蛋白质互作网，已知疾病标识的数据集

目标

使用数据挖掘的原理和方法，通对蛋白质互作网络进行分析、处理，从而预测人体可能的疾病蛋白质。

思路

根据蛋白质互作网络的分布属性，den(S) >= 0.5 时，子网 S 是一个密集子网。当 den(S) =1 时，密集子网 S 是一个极大团。由于极大团中每个节点都存在互作关系，因此，使用极大团预测疾病蛋白质是一个值得研究的方向。

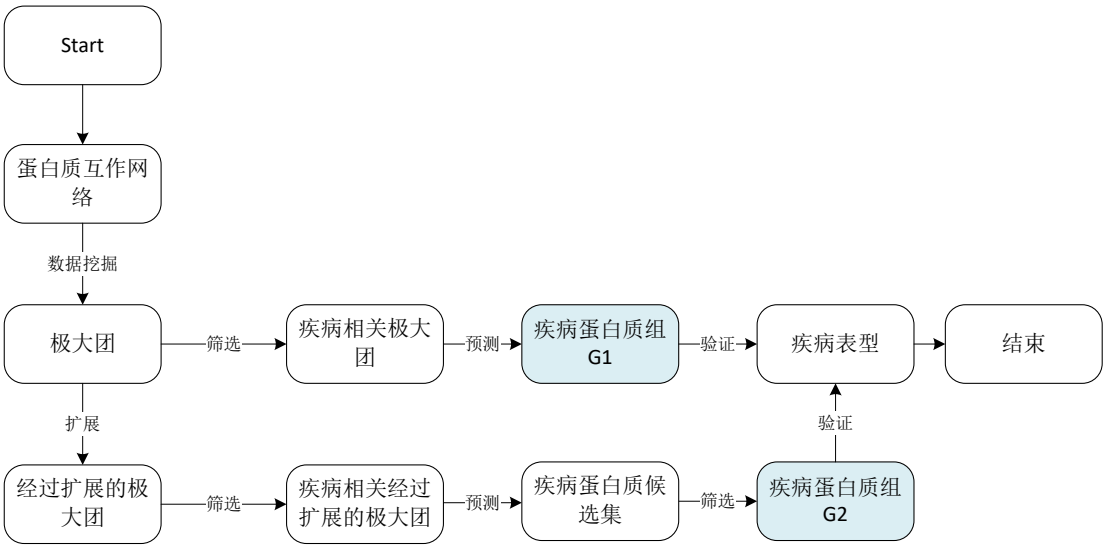
目前，利用网络的拓扑属性来预测致病蛋白质的方法主要有：

1. 利用网络的 Hub 结点
2. 利用瓶颈结点
3. 拓扑模块

极大团是一种特殊的拓扑模块，拥有目前拓扑模块的优点。它能关联大量结点。并且处于蛋白质互作网络中重要的位置。基于这样的事实，我们提出了**基于极大团预测疾病蛋白质**的方案。

方案基本步骤：

1. 在蛋白质互作网络中，挖掘出极大团；
2. 分析，处理极大团，经过筛选、扩展、预测等步骤得到疾病蛋白质组 G1，疾病蛋白质组 G2；
3. 验证结果的准确性；



关键步骤设计思路:

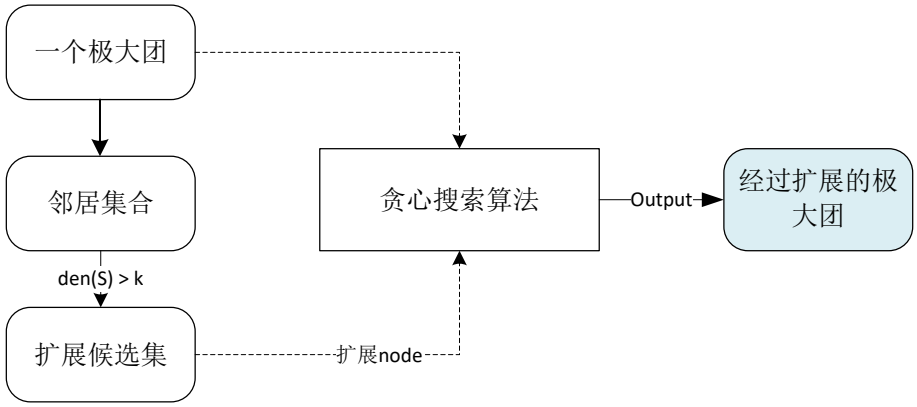
Ⅰ. 挖掘极大团

使用 Bron–Kerbosch 极大团算法

Ⅱ. 扩展极大团

对于已知极大团，和邻居节点组成候选集 S，den(S) > k 的候选集 S 组成扩展候选集。我们采用贪心搜索算法，从扩展候选集选出最终的扩展极大团。

蛋白质互作网络节点图



III．从疾病相关的极大团和疾病相关经过扩展的极大团，预测出疾病蛋白组

统计疾病相关极大团中致病蛋白质的个数，采用显著性检验的方式预测　疾病蛋白质组。

IV．检验预测结果的隹确性

将预测得出的疾病蛋白质组，与已知疾病标识的数据集进行比对，计算得出验证结果的隹确性。

研究方案设计心得

杨昆老师的《数据挖掘》课实践性很强。

通过一学期的学习，我先后接触了 K-means 算法，K 最近邻分类器，svm 支持向量机分类器，交叉验证方法等。值得一提的是，我掌握了 python 语言。

最后一次课的作业是，设计一个数据挖掘研究方案。对于生物，分子方面的专业知识要求较高。设计期间，查阅了大量论文，期刊，文献，收获极大。

感谢老师一学期的教诲。