链接预测方法概览

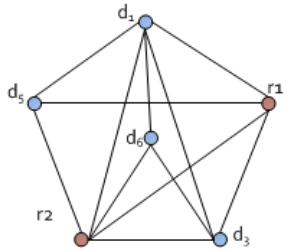
miRNA-Disease 关系预测方法总结(文中方法也可用于解决其他链接预测的问题)

本帖的目的：介绍一些常用的链接预测方法，这些方法都已应用在生物信息以及社会网络的链接预测工作中，本文主要关注在 miRNA 和疾病关系预测中前人使用过的方法。

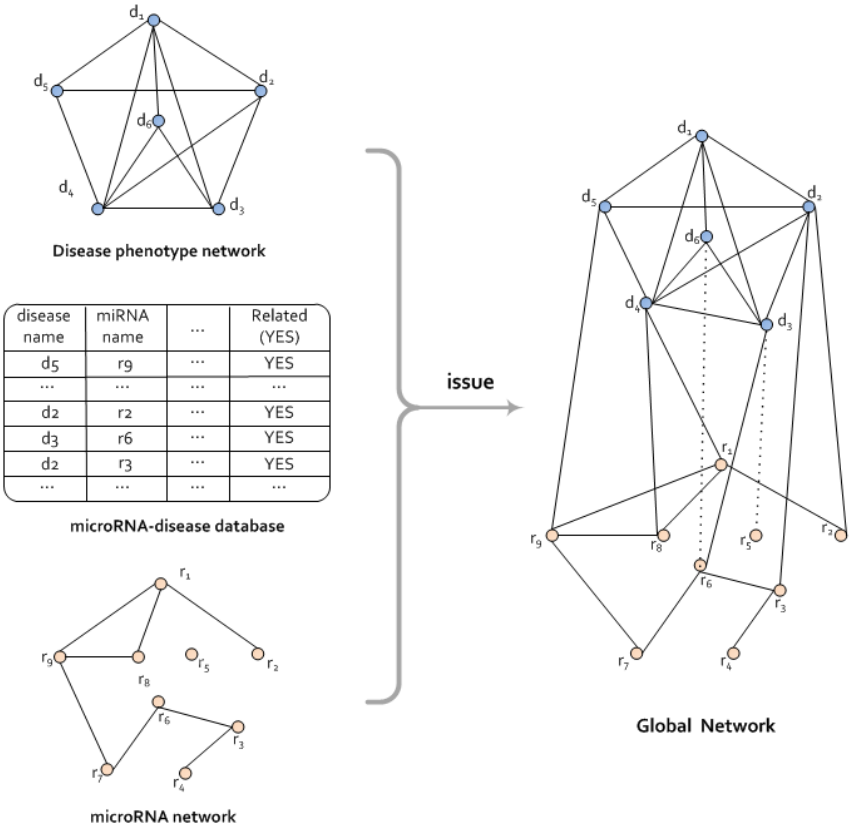
介绍思路：将常用的链接方法大致分为两类，一类是基于相似性测量的；另一类是基于机器学习方法的；针对每一类中的方法，注意进行介绍（包括方法思想和代码的介绍）；涉及的方法有（1）**Katz 方法** （2）**Random Walk 方法** （3）**Hetesim 方法** （4）**分步的 Random Walk** 等方法，之后会继续补充。

基础知识：

（1）由于大部分方法是基于数据网络的（network based methods），因此需要先介绍下基于网络方法的一些基本定义。简单的说，就是从网络的角度看待数据，使用数据。将数据点想像成网络中的点，数据与数据之间的关联想象成网络中的边。举个例子，我们下载到的 RNA-Disease 关联数据，其中包含两种数据点若干（RNA 和 disease），把他们看做网络中的点；他们之间的关联看做网络中存在一条边连接两点。如下图，通常用矩阵存储网络数据。



对于我们需要解决的 RNA-Disease 关联预测，通常研究者会构建如下的数据网络，如图



（2）上图左侧由三个数据网络构成，其中 disease 相似性网络和 RNA 相似性网络可由多种方法计算获得。disease 相似性网络和 RNA 相似性网络与可靠，预测结果也会更好。

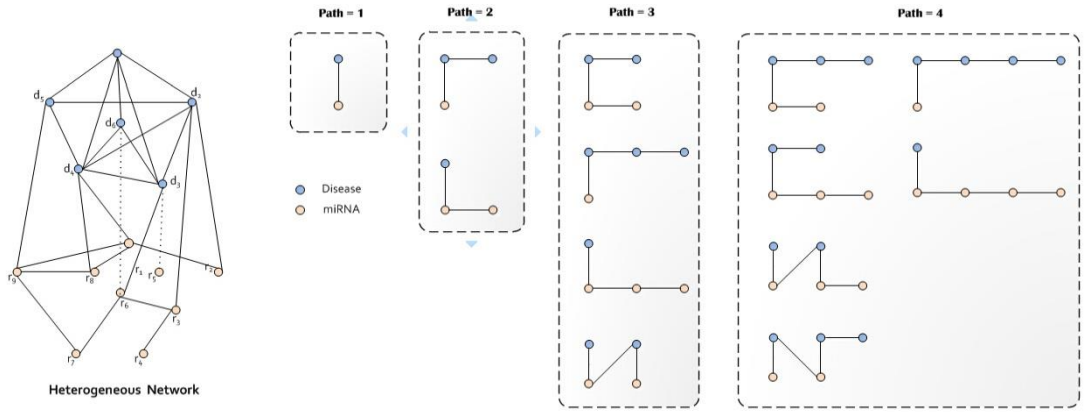
方法介绍：以下方法均应用在上文构建好的数据网络上

类一（基于相似性测量）：

（1）Katz 方法

思想：数路径条数，在网络中，从 A 点到 B 点产生的可达路径条数越多，两点产生关联的可能性就更大；同时，认为路径长度不同对链接产生的贡献就不同，越长的路径贡献越小(若 A->B ，AB 关联强；若 A->C->B，AB 关联次强)，从而赋予不同长度路径不同的权值。

图解：在完整路径中可能产生的不同长度的路径种类



MATLAB 代码：

```
load('weights.mat');
W = weights;
% % 逐个计算路径并归一化
% 计算路径总数,总路径种类有1+2+3+4 +1(lable)
%路径长度为2
F1 = RSmat*RDIrain;
F2 = RDIrain*DSmat;
%路径长度为3
A1 = F1*DSmat;
A2 = RDIrain * RDIrain' * RDIrain;
A3 = RSmat*F2;
%路径长度为4
B1 = A1 * DSmat;
B2 = RSmat * A1;
B3 = RSmat * A2;
B4 = A2 * DSmat;
% 按列归一化每一个特征矩阵
Q = normalize(RDIrain);
F1 = normalize(F1);
F2 = normalize(F2);
A1 = normalize(A1);
A2 = normalize(A2);
A3 = normalize(A3);
B1 = normalize(B1);
B2 = normalize(B2);
B3 = normalize(B3);
B4 = normalize(B4);
ScoreMatrix = W(1)*Q + W(2)*F1+ W(3)*F2 +W(4)*A1+ W(5)*A2+ W(6)*A3 + W(7)*B1+ W(8)*B2 +W(9)*B3+ W(10)*B4;
```

（2）Random Walk 方法

思想：选择一些种子节点（赋初值，且从该点出发），考虑从这些点随机的到达到其他可达点的概率，不断迭代计算，使概率值最终达到稳定状态，概率越大认为从种子节点到该点产生链接的可能性越大。

MATLAB 代码：

```
% 初始化 S(0)
S0 = RDmat(:,d);
S0 = S0/length(find(S0)); % 赋初始概率值
X = S0;% 设置重启概率参数
% 迭代过程
j=0;
S1 = S0;
S2 = 0.4*Mq*S0 + 0.6*X + 0.1*Mu*S0 + 0.9*X;
while norm(S2-S1)>10^-10
    S1 = S2;
    S2 = 0.4*Mq*S1 + 0.6*X + 0.1*Mu*S1 + 0.9*X;
    j = j+1;
end
score = S2;
```

（3）分步的 Random Walk

思想：有点将 katz 方法和 Random walk 相结合的意思，在 Random walk 的基础上，不考虑达到稳态时的结果，仅考虑前几次迭代的结果（路径长度短贡献大）作为最终结果。

MATLAB 代码：修改 Random walk 方法的迭代次数即可

（4）Hetesim 方法

-------未完待续

论文参考：

（1）Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks, *Briefings in bioinformatics*, 17(2)(2016), 193 - 203

（2）Prediction and validation of disease genes using HeteSim Scores, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*