## 已知

1. 表型（ph）之间的相似度（）,值越大，越相似
2. 表型已知的部分致病基因（dg），ph与dg之间是多对多的关系，大部分基因并未被分配到具体某一表型
3. 所有基因的PPI关系（），通过dijkstra方法基于PPI网络已求得所有基因间的局部最短距离
4. 组学数据

## 假设

1. 组学数据中胃癌的17个致病基因在所有人中的表达存在可挖掘的规律
2. 上一步挖掘的规律可直接与胃癌的致病基因对应
3. 相似表型对应的致病基因也相似
4. 极不相似的表型对应的致病基因也不相同

## 目标

1. 推测某些与胃癌相似的表型的致病基因是否可能与胃癌相关
2. 在无确定表型的基因中寻找可能与胃癌相关的致病基因

## 模型输入输出

1. 输入

* 通过组学数据，每个基因分别在胃癌和正常患者中的表达量的统计值（e.g.，平均值、标准差、最大值、p值）
* 基因所在pathway和GO
* 基因与胃癌17种已知基因的PPI距离向量
* 基因所对应表型与胃癌的 。特别地，如果是无已知表型的基因，还不知道怎么处理；如果是对应多种表型的基因，则求最短距离。

1. 输出

训练使用基因是否属于胃癌的二分类问题。但在测试阶段，输出基因属于胃癌的概率值。

## 训练/测试数据

1. 训练数据

15个胃癌致病基因与N个确定与胃癌不相关表型的对应基因

1. 测试数据

利用剩下2个胃癌致病基因做测试（10 fold）