

基于蛋白质相互作用网络预测癌症致病基因*

袁芳¹, 李靖², 周艳红¹

(1. 华中科技大学 湖北生物信息与分子成像重点实验室, 武汉 430074; 2. 华为技术有限公司, 广东 深圳 518129)

摘要: 基于现有的蛋白质相互作用数据, 提出利用邻居曲线方法来分析癌症基因产物在蛋白质相互作用网络中的中心度和聚集度, 据此获取与癌症高度相关的候选致病基因。癌症基因大规模测试显示, 有26%的目标基因在候选基因中排名前5%, 90%的目标基因在候选基因中排名前50%, 该方法能有效地识别癌症致病基因。

关键词: 疾病基因; 癌症; 蛋白质相互关系; 预测

中图分类号: TP391 **文献标志码:** A **文章编号:** 1001-3695(2012)09-3221-03

doi: 10.3969/j.issn.1001-3695.2012.09.005

Predicting cancer-causative genes based on protein-protein interaction network

YUAN Fang¹, LI Jing², ZHOU Yan-hong¹

(1. Key Laboratory of Hubei Bioinformatics & Molecular Imaging, Huazhong University of Science & Technology, Wuhan 430074, China;

2. Huawei Technologies Corporation, Shenzhen Guangdong 518129, China)

Abstract: By analyzing the topological features in PPIs (protein-protein interactions) network based on nearest neighbor search, this paper evaluated the possibilities of candidate genes to be cancer-causative genes. A large scale test shows that 26% and 90% target genes are at the top 5% and top 50% of the list prioritized, the approach can identify cancer genes with high performance.

Key words: disease gene; cancer; protein-protein interaction; prediction

通过定位候选策略和全基因组关联研究等方法, 很多致病基因定位在染色体上的一个区间或多个染色体区间。基于生物信息学的疾病基因预测可以快速获得疾病候选基因, 从而加快疾病研究的进程。现在已经出现很多基于已知疾病基因信息的疾病基因预测软件^[1]。随着蛋白质相互作用 (protein-protein interaction, PPI) 数据的丰富, 基于疾病基因的产物在 PPI 网络中独特的拓扑特征成为疾病基因预测的新途径。

研究发现, 癌症基因的蛋白质产物在 PPI 网络中有着两个独特的拓扑特征: a) 中心性, 即癌症致病蛋白质倾向于位于网络的中央^[2]; b) 聚集性, 即癌症致病蛋白质之间更容易发生相互作用^[3]。根据这些癌症基因在蛋白质相互作用网络中的拓扑特征, 可以从 PPI 网络中挖掘潜在的癌症基因。

基于 PPI 的疾病基因预测方法面临两个难题。一是 PPI 数据的缺陷。当前的 PPI 数据很不完整, 如 OPHID^[4] 的人类 PPI 数据只有约 10 万条, 而人类 PPI 数据总共估计有 65 万条^[5]; 而且这 10 万条相互作用数据中大约有三分之一是基于计算机预测的, 还有一部分是基于高通量实验的, 这些相互作用的可靠性都不高; 剩下的数据中还有大量不同来源的冗余数据。已知的可靠的相互作用仍很少。二是拓扑特征难以量化。例如, 通常以 degree, 即邻居数来衡量蛋白质的中心性, 但一个 degree 高的蛋白质, 既可能是一个 hub 蛋白, 也可能只是一个研究较多的普通蛋白。

当前基于 PPI 的疾病基因预测方法都难以解决这两个问

题。George 等人^[6]将已知疾病基因的相互作用对象预测为疑似疾病基因, 他们研究了三级相互作用, 发现基于直接相互作用的灵敏度很低, 而基于非直接相互作用预测的假阳性很多。Xu 等人^[7]使用了五个拓扑特征, 提供了预测方法的性能, 但无法说明这些拓扑特征的生物学意义。

本文基于蛋白质相互作用数据, 提出利用邻居曲线方法来分析待测基因在 PPI 网络中的拓扑特征, 将蛋白质的中心度和聚集度进行量化, 并根据这两个特征的量化值, 用朴素贝叶斯模型打分, 从而筛选出与癌症高度相关的候选致病基因。邻居曲线方法侧重于特征量化的物理意义, 比已有的方法更好地描述了中心性和聚集性这两个拓扑特征, 通过癌症基因大规模测试显示, 该方法能高精度地识别癌症致病基因, 加快疾病研究的进程。

1 材料和方法

1.1 数据

1) 蛋白质相互作用 从 OPHID 下载了人类蛋白质相互作用数据。OPHID 的数据集包含了基于文献注释的、高通量蛋白质相互作用实验识别的、通过模式生物的相互作用数据预测得来的三类不同来源的蛋白质相互作用。这里没有采用模式生物的相互作用数据, 并去掉了不同来源的重复冗余数据。经过处理后, 蛋白质相互作用网络包含了 100 716 条互作, 涉及

收稿日期: 2012-02-14; **修回日期:** 2012-03-29 **基金项目:** 国家自然科学基金资助项目(90608020); 国家自然科学基金基础科学基金资助项目(J1103514)

作者简介: 袁芳(1977-), 女, 湖北武汉人, 博士, 主要研究方向为生物信息技术和数据库技术(yuancopper@yahoo.com.cn); 李靖(1977-), 男, 湖北武穴人, 工程师, 硕士, 主要研究方向为通信网络和生物信息技术; 周艳红(1966-), 男, 湖南湘潭人, 博导, 主要研究方向为生物信息技术。

到 11 456 个蛋白。

2) 癌症基因 癌症基因数据来自于 Cancer Gene Project (CCP)^[8] 的癌症基因普查数据 (2010 年 5 月 30 日的版本)。从 Cancer Gene Census 数据库中获取 527 个经实验验证与癌症相关的基因, 除去其中重复的、没有蛋白质相互作用注释的基因, 共得到 37 种含有蛋白质相互作用注释的癌基因的癌症类型, 共 218 个基因。

1.2 流程

分析待测区间中的某个基因与某个癌症是否高度相关, 通过对该基因进行打分来评估, 主要步骤如下:

a) 通过基因图谱获取待测区间中的所有待测基因; 基于 PPI 数据, 利用邻居曲线方法分析待测基因在 PPI 网络中的拓扑特征, 对中心度和聚集度进行量化。

b) 基于中心度和聚集度的量化值, 利用朴素贝叶斯模型对待测基因进行打分。

c) 根据打分对待测区间中的所有基因进行降序排序, 排名越靠前的基因被认为更有可能是癌症致病基因。

1.3 利用邻居曲线方法进行中心度和聚集度的量化

根据与已知疾病基因发生相互作用的情况预测潜在疾病基因。但是已知疾病基因中直接相互作用的情况较少, 而随着相互作用距离的增加, 假阳性指数增加。本文研究了随着相互作用距离 x 增加, 连接蛋白质数目 $L(p)$ 和癌症蛋白质数目 $L(c)$ 的变化 (图 1), 将这一变化曲线称为邻居曲线。从图 1 可以看出癌症蛋白质比普通蛋白质更快地扩散到网络中, 更快地连接到其他癌症蛋白质。由于邻居曲线是典型的 S 型曲线, 可以用 Logistics 方程来拟合, 即

$$L(x) = \frac{T}{1 + e^{ax+b}} \quad (1)$$

其中: T 表示函数的上限, 对于 $L(p)$, 其等于网络中蛋白质的总数减 1, 对于 $L(c)$, 其等于网络中癌症蛋白质的总数; a 和 b 对于每个蛋白质是不同的, 可以用加权最小二乘法求得。

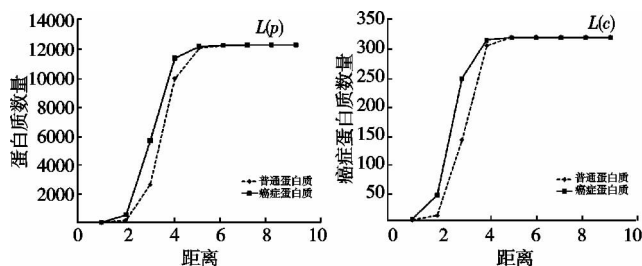


图1 $L(p)$ 和 $L(c)$ 随 x 增加的变化

根据 Logistics 方程的性质, 其拐点横坐标为

$$x = -\frac{b}{a} \quad (2)$$

纵坐标为

$$L = \frac{T}{2} \quad (3)$$

$L(p)$ 的拐点横坐标 x_p 表示扩散到网络的一半所需的作用距离, 可以用来衡量蛋白质的中心度, x_p 越小则蛋白质越容易在网络中扩散; $L(c)$ 的拐点横坐标 x_c 表示连接一半已知癌症基因所需的作用距离, 可以用来衡量蛋白质的聚集度, 或者称为绝对聚集度, x_c 越小则蛋白质越靠近已知癌症基因。由于处于网络中央的蛋白质更快连接到已知癌症基因, x_p 和 x_c 正相关, 如图 2 所示, 因而取它们的差值 m 来衡量蛋白质的相对

聚集度:

$$m = x_p - x_c \quad (4)$$

m 越大, 表明同等中心度情况下, 蛋白质越靠近已知癌症蛋白质。

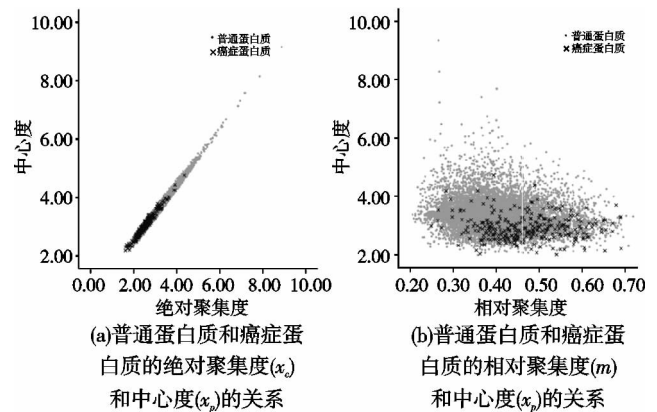


图2 x_p 与 x_c 的关系

1.4 利用邻居曲线方法进行中心度和聚集度的量化

分析可发现癌症基因拥有更小 x_p 和更大的 m , 如图 3 所示, 因而可以根据这两个特征预测癌症蛋白质。假设正常蛋白质和癌症蛋白质的 x_p 和 m 都服从正态分布并相互独立, 其均值和方差可以由先验知识求得, 可以通过朴素贝叶斯模型求得某一蛋白质是癌症蛋白质的概率:

$$P_{NC} = \frac{f_{x_p}^+ \cdot f_m^+}{f_{x_p}^+ \cdot f_m^+ + f_{x_p}^- \cdot f_m^-} \quad (5)$$

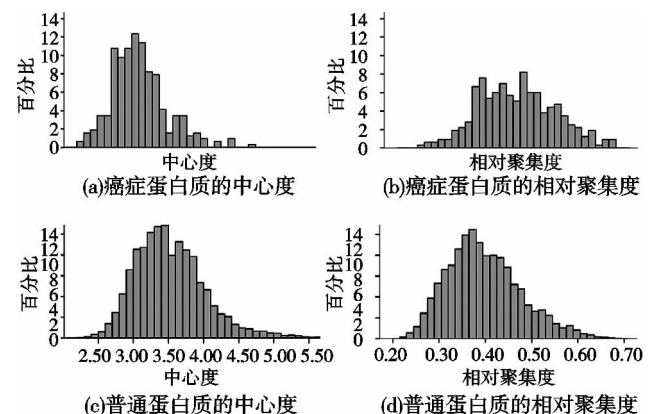


图3 蛋白质中心度和聚集度

2 测试和结果

把从 CCP 中提取的涉及 37 种癌症的 218 个癌症基因作为测试集。对于每一个癌症, 选取其中一个致病癌症基因作为目标基因, 其余的作为已知基因。然后, 获取目标基因上、下游各 15Mb 区间内的基因作为待测基因, 对邻居曲线方法、Xu 等人 and George 等人的方法进行了保留交叉验证。

为从整体上评测预测工具的性能, 采用了 Aerts 等人^[1] 定义的特异性 (SP)、敏感度 (SN) 和受试者工作特征曲线 (ROC) 作为评价标准。

SP 表示样本中低于阈值排位 (threshold position, ThP) 的基因所占的比例, 取输出的基因列表的前 ThP 个基因作为候选疾病基因, GN 表示待测基因的总数, 则 SP 定义为

$$SP = \frac{ThP}{GN} \quad (6)$$

SN 反映选择的候选基因中包含真正的致病基因的可靠程

度,在给定阈值排位的情况下,对于含有 DN(本测试中, DN = 218)个样本的测试集,有 TN 个样本的候选基因中含有致病基因,则 SN 定义为

$$SN = \frac{TN}{DN} \quad (7)$$

若 SN 和 SP 分别为 0.7 和 0.9,则表示 70% 目标基因被准确地预测到待测基因的前 10% (即 $1 - SP$)。显然,在 $1 - SP$ 相同时,SN 越大则预测效果越好。

为了能够从整体上更直观地比较不同预测工具的性能,以 SN 为纵轴、 $1 - SP$ 为横轴作受试者工作特征曲线 (receiver operating characteristic, ROC),并计算曲线下面积 (area under the ROC curve, AUC)。AUC 是预测工具总体性能评价标准,在 $AUC > 0.5$ 的情况下, AUC 越接近于 1,说明预测效果越好; $AUC = 0.5$ 时,说明预测方法完全不起作用。

图 4 显示了这几种基于 PPI 数据预测癌症基因的方法的性能。根据测试结果可以看出,邻居曲线方法、Xu 等人方法和 George 等人方法分别有 26% (SN = 0.26)、23% (SN = 0.23) 和 12% (SN = 0.12) 的目标基因在候选基因中排名前 5% ($1 - SP = 0.05$); 分别有 90% (SN = 0.90)、89% (SN = 0.89) 和 81% (SN = 0.81) 的目标基因在候选基因中排名前 50% ($1 - SP = 0.5$)。显然,在这两个评价点上,邻居曲线方法的预测效果比其他两个方法要高一点。计算它们的 ROC 曲线下面积 (AUC),分别为 0.764、0.761 和 0.705,邻居曲线方法在总体性能上也略好一些。

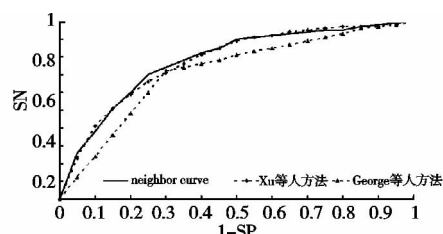


图4 邻居曲线方法、Xu等人方法和George等人方法的测试结果比较

3 比较和讨论

人们使用“枢纽蛋白”这一术语来描述那些在 PPI 网络中起中枢作用的蛋白质,认为那些邻居众多、处于网络中央的节点往往是更为重要的。但是使用节点邻居数 degree 来衡量节点的中心度有三个问题: degree 易受人为误差影响,一个研究较多的蛋白质更容易获得高 degree; degree 只能反映网络的一小部分的拓扑结构 (而且还是不准确的); degree 的取值很有限,对于 degree 相等的蛋白质无法比较其重要性。类似地,用疾病基因邻居数描述节点的聚集度也会遇到这样的问题。Xu 等人采用到已知疾病基因的平均距离来描述节点的聚集度,考虑到了全局网络的结构,可以在一定程度上减少误差,同时扩大了特征的取值范围,但缺点是与中心度正相关,位于网络中央的节点到已知疾病基因的平均距离显然比位于网络边缘的要小。而 George 等人将已知疾病基因的相互作用对象预测为疑似疾病基因,他们研究了三级相互作用,发现基于直接相互作用的灵敏度很低,而基于非直接相互作用预测的假阳性很多。

使用邻居曲线参数 x_n 和 m 来衡量节点的中心性和聚集性则可以解决这些问题,其不仅考虑了全局网络的结构,将人为误差分散在网络中,获得连续变量,还可以使两个特征相互独立。这一方法的缺点是要求网络是闭合的,比较两个不同网络

中节点的 x_n 和 m 没有意义。不过由于大部分蛋白质是在一个闭合网络中,因此并没有太大影响。

比较测试的结果显示,邻居曲线方法比 Xu 等人方法和 George 等人的方法性能要略好一些,其中 Xu 等人的方法考虑了 5 个拓扑特征,而邻居曲线方法只用到 2 个特征,更表明了这 2 个特征的选择优势。

4 结束语

基于生物信息学的疾病基因预测可以快速地获得疾病候选基因,从而加快疾病研究的进程。随着 PPI 数据的完善和相互作用组学的发展,基于疾病基因的产物在 PPI 网络中独特的拓扑特征发现疾病基因成为新的途径。但现有的基于 PPI 的疾病基因预测方法只考虑到网络的局部结构,由于目前 PPI 数据的不完整、不可靠、不平衡引起的误差较大,且量化拓扑特征的方法不合理,制约了基于 PPI 网络局部结构的疾病基因预测方法的预测效果。

本文提出的 protein rank 算法和邻居曲线方法都能很好地解决这个问题,取得更好的预测效果。邻居曲线方法侧重于特征量化的物理意义,比传统方法更好地描述了中心性和聚集性这两个拓扑特征。测试结果表明这两种方法的预测效果都优于国际同类的方法。尽管如此,基于 PPI 疾病基因预测的效果还有很大的提升空间,其中最关键的是 PPI 数据的补充和确认。由于当前相互作用组是一个热门的领域,PPI 数据增长速度很快,可以预见当数据更为完整时还可以挖掘更多、更可靠的拓扑特征,进一步提升预测效果。

此外,利用 PPI 数据不仅可以鉴别新的疾病基因或者挖掘其网络属性,利用 PPI 网络还可以鉴别疾病关联的子网络和基于网络的疾病分类。后两个方向正处于起步阶段,其不只是单纯地推测某基因是潜在疾病基因的概率,还能根据在网络中的功能推测其对疾病的贡献。利用 PPI 推测基因在各种疾病中所起的作用是进一步研究的主要方向。

参考文献:

- [1] AERTS S, LAMBRECHTS D, MAITY S, *et al.* Gene prioritization through genomic data fusion[J]. *Nature Biotechnology*, 2006, 24 (5): 537-544.
- [2] JONSSON P F, BATES P A. Global topological features of cancer proteins in the human interactome [J]. *Bioinformatics*, 2006, 22 (18): 2291-2297.
- [3] GANDHI T K B, ZHONG Jun, MATHIVANAN S, *et al.* Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets[J]. *Nature Genetics*, 2006, 38(3): 285-293.
- [4] BROWN K R, JURISICA I. Online predicted human interaction database[J]. *Bioinformatics*, 2005, 21(9): 2076-2082.
- [5] STUMPF M P H, THORNE T, De SILVA E, *et al.* Estimating the size of the human interactome[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2008, 105 (19): 6959-6964.
- [6] GEORGE R A, LIU J Y, FENG L L, *et al.* Analysis of protein sequence and interaction data for candidate disease gene prediction[J]. *Nucleic Acids Research*, 2006, 34(19): e130.
- [7] XU Jian-zhen, LI Yong-jin. Discovering disease-genes by topological features in human protein-protein interaction network[J]. *Bioinformatics*, 2006, 22(22): 2800-2805.
- [8] FUTREAL P A, COIN L, MARSHALL M, *et al.* A census of human cancer genes[J]. *Nature Review Cancer*, 2004, 4(3): 177-183.