基于功能相似性预测疾病基因*

袁 芳[1], 李 靖[2]
(1. 深圳信息职业技术学院, 广东 深圳 518172; 2. 华为技术有限公司, 广东 深圳 518129)

摘 要: 如何从连锁定位区域中的众多基因中有效选取疾病候选基因是疾病诊断治疗和预防的基础。基于基因功能注释信息,设计和实现了一种新的基于基因功能相似性的疾病基因预测工具 DGP,分析候选基因和已知疾病基因的 GO 之间的相似性,对候选疾病基因进行打分排序。从 OMIM 数据库中提取一个包含 1 045 个已知疾病基因、涉及 305 种疾病的数据集来测试 DGP 的性能,其中 56.7% 的疾病基因在候选基因中排名前 5%, 68.5% 的疾病基因位于前 10%,结果显示 DGP 具有很高的准确率,能够从某个染色体区间中有效地识别出疾病基因。

关键词: 候选基因; 基因本体; 相似性; 预测
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2012)11-4213-03
doi:10.3969/j.issn.1001-3695.2012.11.053

Identifying disease genes based on functional similarity

YUAN Fang[1], LI Jing[2]
(1. *Shenzhen Institute of Information Technology*, *Shenzhen Guangdong* 518172, *China*; 2. *Huawei Technologies Corporation*, *Shenzhen Guangdong* 518129, *China*)

Abstract: Identifying disease genes is essential for elucidating pathogenesis and developing diagnosis and prevention measures. This paper developed a computational tool, named DGP, to assess candidate genes in interested chromosome regions for their possibility relating to a given disease. DGP prioritized the candidate genes by measuring the functional similarity to the known causative genes of the disease. It evaluated the performance of DGP with a dataset containing 1045 genes related to 305 diseases. The validation results show that 56.7% and 68.5% of disease-associated genes are at the top 5% and top 10% of the list prioritized by DGP. Therefore, DGP can effectively help the selection of candidate genes in interested chromosome regions for mutation analysis.

Key words: candidate gene; gene ontology; similarity; prediction

0 引言

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[1]。利用计算机分析染色体区间中众多的基因与疾病的关联程度是寻找疾病基因的一个很重要的方法。

目前,研究者已经提出了许多利用基因功能信息预测疾病基因的方法[2~8]。这类方法认为:表型相似的疾病具有相同的分子机制,即表型类似的疾病可能是由功能相同或者相似的基因导致。

基于这个假设,G2D[2,3]、POCUS[4] 和 FP[5] 挖掘疾病表型同基因功能之间的相似性来预测疾病基因,并取得一定的效果。这些方法都依据 GO 注释来评价基因与表型的相似性,但都只考虑共有 GO 注释的情况,没有充分考虑如离子通道活性和阳离子通道活性功能间的上下层次关系以及分子功能、生物过程等不同功能类别的发生特异性等。

随着 GO 数据完整性的提高,GO 注释的数目增多,两个基因拥有同一个 GO 注释的机率将会降低,只有充分挖掘 GO 注释的结构信息,才能有效地判断基因之间的功能相似程度[6]。基于这种思想,Perez-Lratxeta 等人于 2007 年在 G2D 中增添的 Known Genes[7] 方法采用 Resnik 提供的基于信息含量的语义相似性算法[9] 来评估 GO 注释之间的相似性,使预测效果得到显著的提高。然而 Franke 等人[10] 的研究表明,GO 注释是当前用于疾病基因预测的所有数据资源中最具预测效力的资源;同时他们还发现,在 GO 注释的基础上整合其他数据资源后,预测效果并没有明显的提高。所以如果能够提高基于 GO 注释的预测方法的预测效果,对提高多信息融合方法总体效果都有所帮助。

本文提出一个预测疾病基因的软件 DGP,该软件在充分挖掘不同 GO 之间相似性的基础上,进一步建立评价基因与表型的相似性的算法,从而根据待测基因与已知疾病基因的功能相似程度评估其是疾病基因的风险程度。与 G2D 使用的 Resnik 相似性算法不同,本文使用一个基于 GO 注释间最短路径距离的 GO 相似性算法,并考虑了不同 GO 对疾病基因的贡献程度不同。

收稿日期: 2012-03-06; **修回日期**: 2012-04-08 **基金项目**: 国家自然科学基金重大研究计划资助项目(90608020);国家自然科学基金基础科学基金资助项目(J1103514);广东省自然科学基金资助项目(S2012010010206)

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@163.com);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技术和生物信息技术.

1 方法

1.1 流程

用户给定待处理的疾病对应的 OMIM ID 及染色体上的定位区间信息(如 5q33 ~ 5q34 或 D1S2840 ~ D1S249 等),DGP 系统将计算定位区域中所有基因与特定疾病间的相关程度,并依此进行降序排列,以列表形式输出排序后的待测基因及其相关连接信息,便于领域专家结合具体情况作进一步分析判断。

本文的算法流程如图 1 所示。主要包括以下步骤:

a)数据准备。根据给定的疾病表型 OMIM ID,找出所有已知致病基因(或者根据用户提供的该表型的已知基因)及其 GO 注释;根据给定的染色体定位区间,找出该区间所有的待测基因,以及每个待测基因的 GO 注释。

b)功能相似性分析。首先对每一个已知疾病致病基因的 GO 注释与疾病的相关程度进行打分,然后根据待测基因的每一个 GO 和每一个疾病已知的 GO 在 DAG 中的最短路径对它们之间相似度进行打分,最后根据这两个打分来计算待测基因与给定疾病之间的相关程度。

c)根据与给定疾病之间的相关程度对待测区间内所有候选基因进行降序排列,排在前面(相关程度高)的基因更有可能是疾病的致病基因。

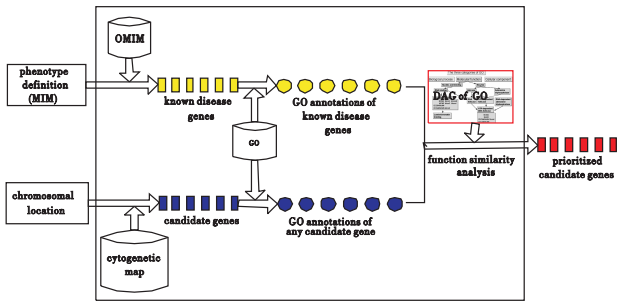

图1 候选疾病基因排序的主要流程

1.2 GO 之间功能相关度的计算

DGP 设计了一种新的利用 GO 之间的路径来评价候选基因与疾病相似程度的算法。比较两个 GO 的功能相似程度是评价两个基因或者两组基因之间功能相似程度的基础。现有的大部分基于 GO 注释的疾病基因注释工具只考虑了两个基因共有 GO 注释的情况。然而,由于基因功能本身具有比较复杂的层次关系,因此仅由两个基因间没有相同的 GO 注释并不能说明它们之间不存在功能上相似。当两个基因注释的 GO 注释不同时,可通过 DAG 提供的结构信息来建立 GO 注释的功能相似性。与 SUSPECTS 和 G2D 使用的信息含量的语义相似性算法 Resnik similarity 不同,本文使用的是基于 GO 在 DAG 上路径距离的算法。

在 DAG 中,任意一个 GO 节点都必须遵守这个路径规则:如果子注释可以描述一个基因产物的功能,则其所有祖先注释都可以描述其功能。这也就是说,两个不同的 GO 在 DAG 中向上回溯,如果可以找到共同的祖先节点(祖先 GO),即具有功能相关性。这里,任意两个 GO 注释 go_i 和 go_j 之间的相似程度定义为

$$go2go(go_i, go_j) = 1/(1 + sp(go_i, go_j)) \tag{1}$$

其中:$sp(go_i, go_j)$ 表示 go_i 与 go_j 间的最短路径的距离。这里假设路径上直接相邻的两 GO 间的距离为单位距离 1,相同 GO 之间的距离为 0。

1.3 基因与疾病之间相关度的计算

显然,如果待测基因的 GO 注释与疾病基因的 GO 注释越相似,则此 GO 注释与疾病的关联程度就越大。然而,对于一个特定的疾病来说,并不是疾病基因的每个 GO 都对此疾病有相同的贡献。所以在定义待测基因的 GO 注释与疾病的相关程度之前,需要评估每个疾病相关 GO 对此疾病的贡献程度。

对于给定一个疾病 d,它所有的已知致病基因集合为 G:$\{g_1, g_2, \cdots, g_k\}$,这些基因所有的 GO 注释组成的集合为 DS:$\{go_1, go_2, \cdots, go_m\}$,其中 go_i 在 G 中出现的次数为 n_i(即 G 中有 n_i 个基因的 GO 注释中含有 go_i)。DS 集合中某个 go_i 与疾病 d 的相关程度定义为

$$dgo2d(d, go_i) = n_i / \sum_{k=1}^{m} n_k \tag{2}$$

待测基因的 go_k 与疾病 d 的相关程度定义为

$$cgo2d(d, go_k) = dgo2d(d, go_{\text{nearest}}) \times go2go(go_k, go_{\text{nearest}}) \tag{3}$$

这里

$$go_{\text{nearest}} = \underset{go_h \in DS}{\text{argmax}}\ go2go(go_k, go_h)$$

基于以上定义,候选基因 g 与疾病 d 之间的相关度定义为

$$g2d(d, g) = \frac{1}{m} \sum_{i=1}^{n} cgo2d(d, go_i) \tag{4}$$

这里 m 是 DS 集合中 GO 的个数,n 是候选基因相关的 GO 的个数。

2 测试与比较

2.1 测试与结果

为了评估 DGP 的性能,获取 OMIM 数据库中含有两个以上的具有 GO 注释的疾病基因的 305 种疾病(涉及 1 045 个已知疾病基因)作为测试集。对于每一个疾病,本文选取其中一个致病基因作为目标基因,其余的作为已知基因。然后,获取目标基因上、下游各 15 Mb 区间内的基因作为待测基因。本文利用 DGP 对所有待测基因与疾病的相关度进行打分排序,如果目标基因排序越靠前,则说明系统预测准确度越高。为从整体上评测 DGP 的预测性能,采用了 Aerts 等人[11]定义的 *SP*, *SN* 和 ROC 作为评价标准。

本文基于这个测试集对 DGP 进行保留—交叉验证(其定位区间中的平均候选基因个数为 190)。实验结果显示,有 16.1%(168 个)的疾病的致病基因排名第 1,56.7% 的疾病基因在候选基因中排名前 5%,68.5% 的疾病基因排名前 10%,而 98.0% 疾病基因排在前 50%。

2.2 与其他预测方法的比较

在已报道的同类研究中,POCUS、FP 和 G2D 的 Known Genes 方法同样是基于功能相似性。本文比较了它们和 DGP 的性能。

对于 POCUS 和 FP 算法,由于均不提供服务,无法直接测试其预测精度。为此,根据其文献中给出的精度与本文方法进行间接比较。Turner 使用一个包含 29 种寡基因疾病、163 个疾病基因的测试集对 POCUS 进行了测试。对其文中提到的测试结果进行转换后,当 1 - *SP* 是 3.69% 时,*SN* 是 36.8%;而 DGP

在 1 – *SP* 为 3.69% 时,*SN* 达到 51.6%。从这个测试点上看,DGP 比 POCUS 的预测效果要好。Freudenberg 和 Propping 从 OMIM 数据库中提取一个包含 878 个疾病基因的测试集,在全基因组上对 FP 算法进行了测试。有三分之一的疾病基因位于所有候选基因排序的前 3%,有三分之二的疾病基因位于所有候选基因排序的前 15%。这个测试数据经过转换后,结果为:当 1 – *SP* 是 3% 时,*SN* 是 33.4%;当 1 – *SP* 是 15% 时,*SN* 是 66.7%;而 DGP 在 1 – *SP* 为 3% 和 15% 时,*SN* 分别是 41.2% 和 76.0%(图 2)。显然,在这两个测试点上,DGP 具有更好的预测效果。

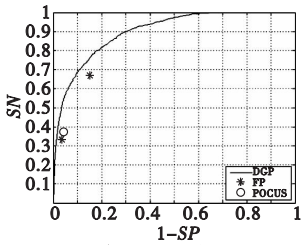

图2　DGP、FP与POCUS交叉验证的结果

G2D 的 Known Genes 是当前预测效果最好的基于功能注释信息的方法,并且提供网络服务,可以通过输入已知疾病基因对定位区间中的基因进行测试。然而,G2D 最多只能返回分值最高的 100 位的候选基因,而测试集中的平均待测基因数目接近 200,不便于系统地比较。为了更加完整地与 G2D 比较,挑选测试集与 G2D 网页上提供 Benchmark 共有的 213 种疾病作为比较对象,根据各自的预测性能画 ROC 曲线如图 3 所示,并计算曲线下面积。由图 3 可以看出,代表 Known Genes 的虚线完全位于 DGP 的曲线之下,说明 DGP 的预测效果更优。

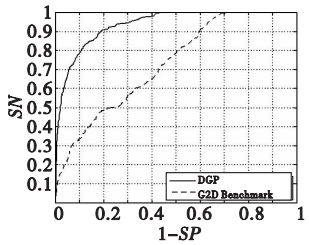

图3　DGP与G2D的Known Genes交叉验证的结果比较

SUSPECTS 除利用了 Resnik 相似性深层次挖掘了 GO 注释间关联外,还融合了序列信息、InterPro domain 以及表达信息,并提供了网络服务。为了比较其与 DGP 的预测性能,用同样的测试集对 SUSPECTS 进行了交叉验证。然而,测试集中的有些目标基因、定位区间无法被 SUSPECTS 识别。于是从测试集中提取能够被 SUSPECTS 用于交叉验证的含有 282 种疾病的 928 个基因的子集,分别对 SUSPECTS 和 DGP 进行测试,测试结果如图 4 所示。代表 DGP 的实线完全位于 SUSPECTS 的虚线之上,DGP 的曲线下面积(AUC)为 0.89 大于 SUSPECTS 的 0.84,说明 DGP 在整体的预测性能上更优。

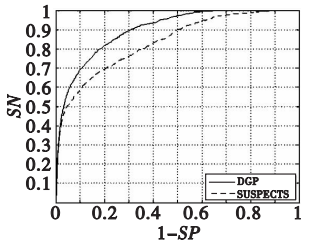

图4　DGP与SUSPECTS交叉验证的结果比较

3　结束语

本文提出了一个新的疾病基因预测工具 DGP,基于不同 GO 在 DAG 中的最短距离来分析待测基因的 GO 与已知致病基因的 GO 之间的相似性,能够从定位区间中有效地识别出疾病相关基因。从测试结果来看,DGP 的预测性能略优于国际同类方法。随着人类基因组注释的完整性和准确性(数量和质量)的不断提高,可以预见本文的算法准确率也会逐步得到提高。

虽然 GO 功能注释是目前除了序列注释之外最丰富的基因注释,但是现在依然有很多疾病没有任何已知致病基因,或者已知致病基因还没任何相关的功能注释,这意味着这些疾病没有任何直接相关的功能注释。为此,DGP 系统还将进行进一步的完善,首先,对某些可获取的相关信息量较少的疾病,通过提取具有相似表型的疾病的相关数据,扩增可用的信息量,提高系统的预测范围和性能;其次,融合蛋白质功能预测等方法,尽可能多地利用目前尚无有效功能注释的疾病基因信息;此外,针对代谢类疾病整合生化通路信息(如 KEGG 数据库)等进行更加深入的分析预测,提高预测复杂疾病致病基因的能力。

参考文献:

[1] YAN Shen. Positional candidate cloning of disease genes[J]. Life Sciences,1999,11(5): 205-508.

[2] PEREZ-LRATXETA C, BORK P, ANDRADE M A. Association of genes to genetically inherited diseases using data mining[J]. Nature Genetics,2002,31(3):316-319.

[3] PEREZ-LRATXETA C, WJST M, BORK P, *et al*. G2D: a tool for mining genes associated with disease[J]. BMC Genetics,2005,6(1): 45.

[4] TURNER F S, CLUTTERBUCK D R, SEMPLE C A. POCUS: mining genomic sequence annotation to predict disease genes[J]. Genome Biology,2003,4(11):R75.

[5] FREUDENBERG J, PROPPING P. A similarity-based method for genome-wide prediction of disease-relevant human genes[J]. Bioinformatics,2002,18(suppl2):110-115.

[6] LORD P W, STEVENS R D, BRASS A, *et al*. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation[J]. Bioinformatics,2003,19(10):1275-1283.

[7] PEREZ-LRATXETA C, BORK P, ANDRADE-NAVARRO M A, *et al*. Update of the G2D tool for prioritization of gene candidates to inherited diseases[J]. Nucleic Acids Res,2007,35:W212-W216.

[8] ANDREAS S, THOMAS L, MARIO A. Improving disease gene prioritization using the semantic similarity of gene ontology term[J]. Bioinformatics,2010,26(18): 561-567.

[9] RESNIK P. Semantic similarity in a taxonomy: an information based measure and its application to problems of ambiguity in natural language[J]. J Artif Intelligence Res,1999,11:95-130.

[10] FRANKE L, Van BAKEL H, FOKKENS L, *et al*. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes[J]. Am J Hum Genet, 2006,78(6): 1011-1025.

[11] AERTS S, LAMBRECHTS D, MAITY S, *et al*. Gene prioritization through genomic data fusion[J]. Nat Biotechnol,2006,24(5):537-544.