应用与环境生物学报 2010, 16(3): 420~423
Chin J Appl Environ Biol=ISSN 1006-687X
2010-06-25
DOI: 10.3724/SP.J.1145.2010.00420

一种新的加权功能相似性算法在蛋白质相互作用研究中的应用*

王 东[1] 须 涛[1] 李令锦[1] 杜林方[1,2**]
([1]四川大学生物资源与生态环境教育部重点实验室 成都 610064)
([2]四川大学纳米生物医学技术与膜生物学研究所 成都 610041)

摘 要 基于GO术语衡量基因产物之间的相似性主要是通过语义相似性进行, 此类方法主要是先计算GO术语之间的相似性, 然后再通过与GO术语相关的基因产物的相似性两步来实现的. 第一步GO术语之间的相似性已经有了大量的研究, 但是关于第二步的功能相似性研究却很少, 针对平均化计算功能相似性的一些缺陷, 提出了一种新的加权的功能相似性计算方法——wfcSim方法. 应用DIP数据库中人类和线虫的蛋白质相互作用数据, 通过ROC曲线对基于平均化、基于最大化、Wang和wfcSim等4种功能相似性研究方法进行了评估, 结果表明, 与前人通常使用的基于平均的方法相比, wfcSim方法在功能相似性分析方面具有较高的优越性. 图3 参15

关键词 语义相似性; 功能相似性; 蛋白质相互作用; GO术语; wfcSim方法
CLC Q811.4

A Novel Weighted Functional Similarity Algorithm Applied to Analyze the Protein-protein Interaction*

WANG Dong[1], XU Tao[1], LI Lingjin[1] & DU Linfang[1, 2**]
([1]*Key Laboratory of Bio-resources and Eco-environment of Ministry of Education, College of Life Sciences, Sichuan University*, Chengdu 610064, China)
([2]*Institute for Nanobiomedical Technology and Membrane Biology, Sichuan University*, Chengdu 610041, China)

Abstract Semantic similarity calculation is one of the main methods measuring the similarity among gene products based on Gene Ontology. Most semantic-based applications following a two-step approach include semantic similarity calculations of paired GO terms and functional similarity calculations of all possible combinations of related GO terms. There are many studies focused on the semantic similarity calculation in the first step. However, the methods for elevating the similarity of GO terms to the similarity of proteins are still considered with some problems. In this study, a novel algorithm called weighted functional similarity calculation (wfcSim) was proposed to address some drawbacks of average-based methods. To analyze *Homo sapiens* and *Caenorhabditis elegans* protein-protein interactions (PPIs) from the Database of Interacting Proteins (DIP), the analysis by the receiver-operator characteristic (ROC) curves for four functional similarity algorithms, such as Ave and wfcSim, showed that the wfcSim method was proved the best among the other average methods. Fig 3, Ref 15

Keywords semantic similarity; functional similarity; protein-protein interaction; Gene Ontology; wfcSim algorithm
CLC Q811.4

在生物信息学中, Gene Ontology (GO)[1]应用的一个重要方面就是对GO术语的语义相似性进行度量. 基于GO的语义相似性具有很广泛的应用价值. 早在2003年, Lord等通过研究发现序列之间的相似性与语义相似性具有很大的相关性[2]. 一些基于表达谱聚类的方法也开始结合基因之间的功能相似性来更准确地研究基因之间的功能, 并取得了不错的效果. 同样在生物之间功能模块的研究中, 基因之间的功能相似性也具有一席之地, Wu等通过整合序列信息、基因间功能相似性、物种间进化信息, 用贝叶斯方法获得了微生物基因组中大量保守的功能模块[3]. Tao等通过解析GO的注释网络用基于GO之间的语义相似性预测了基因的新功能, 并用10倍交叉验证预测算法得到了97%的准确率[4]. 同样, 语义相似性在减少蛋白质相互作用的假阳性方面也获得了一定的成效[5].

一般来说, 如果两个基因产物的功能具有相似性, 那么它们在GO中注解的术语就比较相近, 所以可以通过求出GO术语对的相似度, 来近似估计两基因产物功能之间的相似程度[6]. 前人对于第一步求出GO术语对的相似度已经取得了不错的成果, 其中基于信息量的相似度计算更是由于其简单实用及效果优越得到了广泛的应用. 然而从语义相似性计算功能相似性却没有得到人们的重视, 从而使得如何更好地计算功能相似性没有一个比较好的结论.

Tao等关于功能相似性方法评估的一个研究认为, 蛋白质相互作用或者共表达的两个基因对之间的关系往往由两个基因之间功能最接近的两个GO术语所表明, 但是只考虑单个GO节点往往会被一些外在的因素如注释错误和不精确所影响[7]. 因此, 为了从全局考虑基因间GO注释的关系, 我们提出了一个系统的加权的功能相似性的计算方法——wfcSim算法. 将此算法应用到人的蛋白质相互作用数据集以

收稿日期: 2009-05-25 接受日期: 2009-06-08
* "十一五"国家科技支撑计划项目(No. 2006BAF07B01)资助
Supported by the National Science and Technology Pillar Program of "11th Five-year-plan" of China (No. 2006BAF07B01)
**通讯作者 Corresponding author (E-mail: dulinfang@yahoo.com)

及线虫的蛋白质相互作用数据集进行评估，结果表明此算法兼容了基于平均化和基于最大化功能相似性计算方法的优点，具有较好的实用价值.

1 材料与方法

1.1 数据来源及处理

蛋白质相互作用数据主要来自于蛋白质相互作用数据库（Database of Interacting Proteins，DIP；http: //dip.doe-mbi.ucla.edu）[8]所提供的蛋白质相互作用数据. 由于蛋白质相互作用数据库DIP内的数据有着来源广泛的可靠实验以及高质量的评估方法支持，所以经常被用来确定最可靠的互作子集.

本文所采用的DIP数据库是2009年1月26日版本. 该版本包括了2 172个人类的蛋白质相互作用数据，4 044个线虫的蛋白质相互作用数据，并且存在互作关系. ROC曲线被用来评估wfcSim算法，ROC曲线的分析需要两类数据集，一个是阳性数据集，在本研究中所用的是真实数据集，为蛋白质相互作用数据；另一个是阴性数据集，随机从某一物种的蛋白质库中抽样，并组成蛋白质对，再剔除真实的蛋白质互作关系，从而形成与阳性数据集同样大小的集合. 在实现过程中，ROC曲线的运行、ROC曲线下面积的计算及曲线的绘制是通过R语言中的ROC和ROCR [9]包进行的.

1.2 Resnik语义相似性的计算

Resnik [10]定义两个GO节点之间的功能相似性主要通过GO术语的信息量来进行. 在GO这种层级结构的词汇分类系统中，从父节点到子节点，其中的生物学含义是逐层深入的关系，越往节点上层，概念越笼统；换言之，越往下层，节点所包含的信息含量就越大，在这类情况下，根节点的信息量就是0.

因此Resnik等认为GO术语之间的语义相似性为两个GO节点c_1、c_2的公共最近祖先节点的信息量，即：

$$sim(c_1,c_2)=-\log[p_{ms}(c_1,c_2)]$$

这种方法是目前GO系统中研究语义相似性时最常用到的一种方法，它既简洁又有效，对于链接密度的可变性敏感性小，并得到了多种证据的验证，取得了很好的实用效果. 因此本文采用Resnik方法作为语义相似性的计算.

1.3 功能相似性的计算

相对于GO术语之间的相似性，我们更关注基因之间功能的相似性. 因为一个基因可能由多个GO术语注释，所以在得到两个基因g_1、g_2的GO术语的两个集合A_1和A_2后（A_1和A_2集合的大小为M和N），计算两个集合内的所有GO术语之间的语义相似性，进而研究这些相似性之间的关系，最终得到基因之间的功能相似性.

关于基因功能相似性计算方法的描述见图1.

1.3.1 **最大化的方法**（Max）[3, 11~12]　该方法把两个基因的功能相似性定义为集合A_1和A_2所有语义相似性的最大值：

$$sim(g_1,g_2)=\max_{c_k\in A_1;c_p\in A_2} sim(c_k,c_p)$$

1.3.2 **平均化的方法**（Ave）[2, 13]　提出平均化的方法的学者认为，既然基因有多个GO术语注释，那么就需要将这些术语综合考虑起来研究基因的功能相似性. 为此他们提出了一个相对简单的方法，即平均集合A_1和A_2所有术语之间的语义相似性：

$$sim(g_1,g_2)=\frac{1}{m\times n}\times\sum_{c_k\in A_1;c_p\in A_2} sim(c_k,c_p)$$

1.3.3 **Wang的方法**（Wang）[14]　为了更精确地衡量基因的功能相似性，Wang等认为既要考虑到两个基因之间被注释的所有GO术语之间语义相似性对于功能相似性的贡献，同时也不能通过简单的平均化来获得，为此提出了这个方法：

首先定义一个集合中任意一个GO术语与另外一个集合所有GO术语之间的局部相似性. $Sim(c, A_i)$即为单个GO术语c与另外一个集合A_i中任意一个GO术语的最大语义相似性.

$$sim(c,A_i)=\max_{i=1,2}(Sim(c,A_i))$$

假定两个基因被GO术语注释的两个集合A_1和A_2分别表示为$A_1=\{c_{11},c_{12},c_{13},\ldots,c_{1m}\}$，$A_2=\{c_{21},c_{22},c_{23},\ldots,c_{2n}\}$. 则基因$g_1$和$g_2$之间的功能相似性为：

$$sim(g_1,g_2)=\frac{\sum_{1\le i\le m} sim(c_{1i},A_2)+\sum_{1\le j\le n} sim(c_{2j},A_1)}{m+n}$$

1.3.4 **一种新的加权的功能相似性的计算方法**（wfcSim）　为了弥补基于平均化和基于最大化的功能相似性的一些缺点，从而让求出的功能相似性更加符合实际情况，结合这两种方法，我们提出了一种新的加权的功能相似性的计算方法——wfcSim.

wfcSim计算功能相似性的过程如下：

1) 首先计算相似性矩阵S，包括了A_1所有GO术语与A_2中所有GO术语之间的相似性.

$$S_{ij}=\underset{c_k\in A_1;c_p\in A_2}{sim(C_k,C_p)}$$

2) 比较集合A_1和集合A_2的大小，假设集合A_1小，则计算集合内第一个GO术语c_1与另外一个集合A_2内所有GO术语的语义相似性（图1中表示为→），求其最大值为：

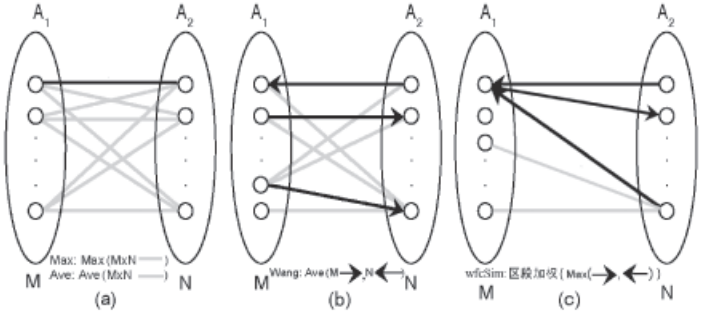

图1 Max, Ave, Wang和wfcSim功能相似性算法示意图

Fig. 1 Graphical illustration of Max, Ave, Wang and wfcSim functional similarity measures

$$s_1 = \max_{1 \le j \le N}(s_{1j})$$

然后寻找集合A_2内任意一个GO术语与集合A_1内具有最大语义相似性的GO术语（图1中表示为←），如果此GO术语为c_1，与c_1相关的A_1内的GO术语的序号则形成集合$\varphi(c_1)$. 至此，我们找到了与GO术语c_1所有相关的最大的语义相似性. 并寻找在c_1术语与集合A_2具有最大相似性的GO术语，则术语c_1的局部最大相似性：

$$Local(S_{1j}) = \max_{1 \le j \le N}(s_1, \ s_{1j} \ )_{j \in \phi(c_1)}$$

以此类推获得集合A_1内每一个术语的局部最大相似性. 如果集合A_2小，反之亦然.

3) 计算GO分支内GO术语的最大语义相似性S_{max}，在区间[0,S_{max}]内划分b个区段，设定参数α，从小到大则每个区段内的权值为α、α^2、α^3等等. 将上一步计算到的所有的局部最大相似性定位到各个区段内，计算每一个区段内的局部最大相似性的平均值Local(s_i). 最后g_1、g_2内的功能相似性为：

$$sim(g_1, g_2) = \frac{\alpha * Local(s_1) + \alpha^2 * Local(s_2) + \cdots + \alpha^b * Local(s_b)}{\alpha + \alpha^2 + \cdots + \alpha^b}$$

2 结果

2.1 DIP人类蛋白质相互作用数据集的ROC曲线分析

从DIP数据库中获得了人类的2 172个蛋白质相互作用对，其中含有1 575个蛋白质，将蛋白质注释到生物过程（BP）分支上，最后获得了可以用来进行ROC曲线分析的1 480个蛋白质互作关系（1 103个蛋白质）. 评估了4种算法Max、Ave、Wang和wfcSim算法在ROC曲线中的表现. 其中在用wfcSim算法计算功能相似性中，相似性区间参数b为3，参数α为2.

从图2可以看到，所有的曲线都靠近图形左上角，并且ROC曲线下面积（AUC值）都大于0.8，这表明人类的蛋白质相互作用数据适合进行功能相似性算法的分析. Max（黑色曲线）和wfcSim（青色曲线）更靠近图形左上角，表明了wfcSim算法在评估过程中取得了最好的效果. 而Ave（红色曲线）则靠近对角虚线，在ROC曲线分析过程中表明此方法的评估效果最差. 改变相似性区间参数b为2，发现在此参数下wfcSim算法的ROC曲线有微量提升（未发表实验数据）.

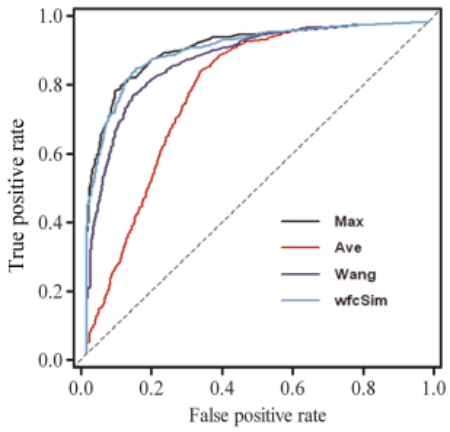

图2 DIP人类蛋白质相互作用在生物功能（BP）分支中功能相似性算法的ROC曲线评估

Fig. 2 ROC curves analysis of human PPIs derived from DIP in the BP ontology

2.2 DIP线虫蛋白质相互作用数据集的ROC曲线分析

在第2个数据集中用到了DIP数据库中线虫的4 044个蛋白质相互作用对，包括2 201个蛋白质. 同样，将蛋白质注释到生物过程分支上，最后获得了1 458个蛋白质互作关系（1 085个蛋白质）. 通过构建阳性数据集合阴性数据集，进一步进行ROC曲线的分析. 其中在用wfcSim算法计算功能相似性中相似性区间参数b为3，参数α为2.

从图3可以看到，4种功能相似性计算方法的评估效果与人类的蛋白质相互作用数据集类似. 但是从总体趋势可以看出，Max和wfcSim算法取得的效果最好，Wang的方法其次，Ave的方法则最差，其曲线下面积只有0.62. 同样，改变相似性区间参数b为2，发现在此参数下wfcSim算法的ROC曲线也有微量提升.

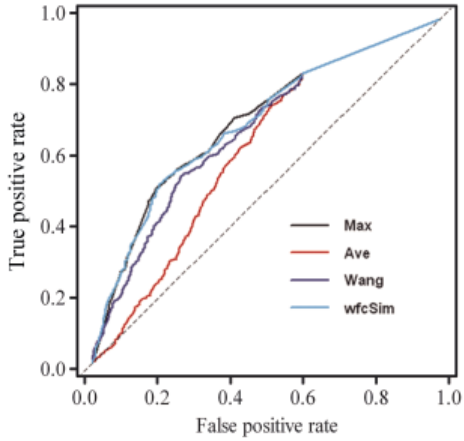

图3 DIP线虫蛋白质相互作用在生物功能（BP）分支中功能相似性算法的ROC曲线评估

Fig. 3 ROC curves analysis of *C. elegans* PPIs derived from DIP in the BP ontology

3 讨论

我们提出的wfcSim算法注重基因间共享的或者更相关的GO术语之间的关系，而不忽视GO术语之间的一些弱相关性，从而从语义相似性推算基因间的功能相似性更具有全面性. 实验结果表明，这种算法进行基因功能性计算更具有准确性.

wfcSim采用了加权的思想，结合基于最大化功能相似性和基于平均化功能相似性的方法，加入了针对不同数据集来人工设定相似性阈值区分区段的想法，对生物学问题更具有针对性，从而能够更好地计算基因间的功能相似性，对认清基因间的关系，进行生物学意义上的解释更具有帮助.

基于平均化的功能相似性的计算方法，认为两个基因所具有的所有可能的功能在某一时刻同时发挥作用，从而算术平均化基因间所有GO术语之间的语义相似性. 但是现实中当两个蛋白质进行相互作用时，往往是发生在某一时刻点，如两个具有互作关系的蛋白质SEC23和BOS1虽然具有很多不同的GO术语注释，而“水泡介导的内质网到高尔基体的转运”是它们共享的注释. 根据先前的报道[15]，它们就是在“水泡介导的内质网到高尔基体的转运”过程中发生相互作用. 这种情况下，基于平均化的方法相比于基于最大化的方法就

显示出了很大的劣势. 然而基于最大化的方法也具有一定的缺陷，如两个蛋白质可能在多种情况下发生相互作用，既有可能在细胞核外，也有可能在细胞核内发生相互作用，虽然发生在细胞核外的相互作用更强烈，然而计算在细胞核内发生相互作用的基因间的功能相似性，基于最大化的方法则会显得力不从心了. wfcSim通过设定语义相似性区段，这样就在加强高相似性区的权值的同时不抛弃低相似区段的数值，从而具有比较好的应用性. 在将来的研究中，我们将整合进入蛋白质相互作用的时间信息，以及蛋白质的定位信息，相信会进一步提升功能相似性的计算性能，从而能更好地从生物学意义上解释功能相似性.

Reference

1 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 2000, **25** (1): 25~29

2 Lord PW, Stevens RD, Brass A, Goble CA. Investigating semantic similarity measures across the Gene Ontology: The relationship between sequence and annotation. *Bioinformatics*, 2003, **19**: 1275~1283

3 Wu HW, Su ZC, Mao FL, Olman V, Xu Y. Prediction of functional modules based on comparative genome analysis and Gene Ontology application. *Nucleic Acids Res*, 2005, **33**: 2822~2837

4 Tao Y, Sam L, Li J, Friedman C, Lussier YA. Information theory applied to the sparse gene ontology annotation network to predict novel gene function. *Bioinformatics*, 2007, **23**: 529~538

5 Mahdavi MA, Lin YH. False positive reduction in protein-protein interaction predictions using gene ontology annotations. *BMC Bioinformatics*, 2007, **8**: 262

6 Popescu M, Keller JM, Mitchell JA. Fuzzy measures on gene ontology for gene product similarity. *IEEE/ACM Trans Comput Biol Bioinform,* 2005, **3:** 263~274

7 Xu T, Du LF, Zhou Y. Evaluation of GO-based functional similarity measures using *S. cerevisiae* protein interaction and expression profile data. *BMC Bioinformatics*, 2008, **9**: 472

8 Xenarios I, Salwínski L, Duan XJ, Higney P, Kim SM, Eisenberg D. DIP: The database of interacting proteins. *Nucleic Acids Res*, 2000, **28**: 289~291

9 Sing T, Sander O, Beerenwinkel N, Lengauer T, ROCR: Visualizing classifier performance in R. *Bioinformatics*, 2005. **21**: 3940~3941

10 Resnik P. Using information content to evaluate semantic similarity in a taxonomy. Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada, 1995. 448~453

11 Brameier M, Wiuf C. Co-clustering and visualization of gene expression data and gene ontology terms for *Saccharomyces cerevisiae* using self-organizing maps. *J Biomed Inform*, 2007, **40**: 160~173

12 Speer N, Spieth C, Zell A. A memetic clustering algorithm for the functional partition of genes based on the gene ontology. Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, La Jolla, CA, USA, 2004. 252~259

13 Wang H, Azuaje F, Bodenreider O, Dopazo J. Gene expression correlation and gene ontology-based similarity: An assessment of quantitative relationships. Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, La Jolla, CA, USA, 2004. 25~31

14 Wang JZ, Du ZD, Payattakool R, Yu PS, Chen CF. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 2007, **23**: 1274~1281

15 Newman AP, Shim J, Ferro-Novick S. BET1, BOS1, and SEC22 are members of a group of interacting yeast genes required for transport from the endoplasmic reticulum to the Golgi complex. *Mol Cell Biol*, 1990, **10**: 3405~3414