一种改进的基因功能网络术语相似度计算方法*

唐 智 刘志明 罗凌云 欧阳纯萍 万亚平

(南华大学计算机学院 衡阳 421001)

〔摘要〕 阐述现有基因术语间语义相似度计算方法，提出基于融合高斯核函数的重启随机游走的基因本体术语相似度算法（Random Walk with Restart – based Similarity Measure，RWRSM），测试算法性能并进行分析，结果表明该算法优于其他算法，可以提高准确性及稳定性。

〔关键词〕 基因功能网络；术语相似度；NETSIM2；随机游走；基因本体

〔中图分类号〕R – 056 〔文献标识码〕A 〔DOI〕10. 3969/j. issn. 1673 – 6036. 2019. 02. 010

An Improved Similarity Calculation Method of Terminology of Gene Function Network *TANG Zhi, LIU Zhiming, LUO Lingyun, OUYANG Chunping, WANG Yaping, School of Computer University of South China, Hengyang 421001, China*

〔**Abstract**〕 The paper elaborates on the existing semantic similarity calculation methods between Gene Ontology (GO) terminologies, puts forward the Random Walk with Restart – based Similarity Measure (RWRSM) based on the Gauss Kernel Function, tests the performance of the calculation method and carries out analysis. The result shows that the calculation method is superior to any other calculation method since it increases the accuracy and stability.

〔**Keywords**〕 gene function network; term similarity; NETSIM2; random walk; Gene Ontology (GO)

1 引言

基因本体（Gene Ontology，GO）是生物医学领域最成功的本体之一，为描述基因（基因产物）的分子功能、生物过程等相关信息提供一个规范、准确的术语集，目前被广泛应用于生物医学相关研究领域[1]。主要体现在基因功能比较与分析、蛋白质相互作用预测、基因集合富集分析等诸多领域，由此成为不可或缺的生物医学本体。基因本体最初由基因本体组织（Gene Ontology Consortium）于 1998 年建立[2]。随着基因本体的发展，越来越多的模式生物数据库加入基因本体组织中，包括大多数主要的植物、动物和微生物数据库，截至目前基因本体

〔收稿日期〕 2018 – 09 – 30

〔作者简介〕 唐智，硕士研究生；通讯作者：刘志明，博士，硕士生导师，发表论文 40 余篇。

〔基金项目〕 国家自然科学基金“大型复杂医学领域本体质量评估理论研究”(项目编号：61502221)；南华大学研究生科学基金项目“基于基因本体相似度计算算法研究”(项目编号：2018KYY083)；国家自然科学基金“面向资源型社交网站的知识图谱构建方法研究”(项目编号：61402220)；湖南省哲学社会科学基金“面向新闻评论的网络舆情演化分析研究”(项目编号：16YBA323)。

已经能够为100 多个物种提供注释信息[3]。

在生物学中，从特定的功能信息出发，查找与其功能相似或者相关的蛋白质，对这些蛋白质间关联程度进行比较量化是生物信息学研究中经常遇到的问题[4]。然而功能相似或者相关的蛋白质并不一定具有较强的序列上的相似性，基因本体的出现为解决这一问题提供新的途径。研究表明如果两个基因产物的功能相似，那么它们在GO 中注释的术语就越相近[5]。所以GO 应用的一个重要方面就是对术语语义的相似性进行度量，从而计算基因产物的相似度[6]。为更好地计算术语的相似度，本文提出一种基于融合高斯核函数的重启随机游走的基因本体术语相似度算法RWRSM。

2 现有基因术语间语义相似度计算方法

2.1 基于基因本体的方法

近年来高通量生物学技术的显著改进导致生物学数据的指数增加。基因本体是用于解释生物实验结果的最流行的生物信息学资源之一。GO 提供结构化、受控制的术语词汇表，通过分子功能、生物过程和细胞成分3 种属性来描述基因[7]。在每个类别中术语被构造为有向无环图（Directed Acyclic Graph，DAG）。基于GO 的语义相似性已成功应用于许多研究领域，如基因功能预测[8]、基因网络分析[9-10]、同源性分析[11]、基因关联可视化[12]和缺失价值估算[13-14]。基于基因本体来计算基因功能相似性的方法[15-20]可分为4 种类型：基于路径长度的方法、基于节点信息的方法、综合方法和基于基因功能网络的方法。基于路径长度的方法通过考虑GO 拓扑结构信息[21]来计算相似度。最近提出的方法称为相对特异性相似性（Relative Specificity Similarity，RSS），考虑两种类型的长度信息：从给定术语对到其最接近的叶子术语的边长和到最低共同祖先（Lowest Common Ancestor，LCA）节点的边长关系。但是基于边缘的方法完全依赖于基因本体的拓扑结构。这种方法不能在相同的拓扑水平上区分术语。对于基于节点的方法，这些方法依赖于特定的分类法。该方法利用信息量最大的共同祖先（Most Informative Common Ancestor，MICA）的信息内容（Information Content，IC）来测量两个GO 术语之间的相似性[22]。评估测试显示结果与蛋白质序列相似性一致，但是基于节点的方法仅考虑注释，忽略GO 的拓扑信息。在综合组中，提出在GO 中使用更多信息的方法。混合相对特异性相似度（Hybrid Relative Specificity Similarity，HRSS）使用4 种类型的信息（信息内容、结构拓扑、注释和MICA）来计算语义相似性[21]。InteGO 方法提出一种基于秩的方法来整合多种现有相似性方法，称为种子方法，以考虑GO 的更多方面[16]。InteGO2 方法通过投票方法从一组方法中选择最合适的方法，并基于元启发式搜索方法[12]整合这些选定的方法。评估测试表明综合方法比种子方法表现更好。但是所有这些方法都只基于GO 结构信息，忽略GO 的不准确表示和缺失的信息等问题。如只有37% 的拟南芥基因具有GO 的所有3 个结构域的实验注释[23]。因此GO 中的不完整信息可能导致低质量的相似性。

2.2 基于网络的方法

最近提出一种基于网络的计算方法（Network-based Similarity Measure，NETSIM），通过整合基因关联和GO 拓扑结构和注释来解决这些问题[18]。基于代谢反应图的实验表明通过结合基因关联可以增强语义相似性，但NETSIM 只考虑网络中的直接链路，使计算过程中只使用基因协同功能网络中的部分信息，实际上除直接连接的基因对外还应考虑基因功能网络中包含的间接基因相互作用。因此一种新的基于网络的NETSIM2 被提出，通过随机游走的方法考虑基因协同网络中的直接和间接相互作用，代谢反应图的实验表明结果远高于之前所有算法，但是NETSIM2 方法测量出的LFC 值不够稳定，大部分LFC 值很小。为解决这个问题本文通过使用高斯核函数来计算边权重，生成归一化加权矩阵 P^T，通过随机游走算法得到每对基因之间的相关性得分。最后基于NETSIM2 原有算法计算术语对相似度。与之前方法相比本文算法融合高斯核函数与随

机游走计算基因相关性得分，以考虑直接和间接的交互，通过有效地整合基因功能网络大大提高 LFC 值的稳定性。

3 改进的基因术语间语义相似度计算方法

3.1 融合高斯核函数与随机游走计算基因相关性得分

3.1.1 生成随机游走图 重启随机游走图构建基本思路是：将训练集合 D 中的每个训练数据 d ∈ D 映射为图中的一个点，根据两点之间的相关程度决定是否用边将两点相连。为方便求解，一般构建成完全图，将两点之间的相似度作为边的权重。当两点之间的相似度很小或者为 0 时，节点之间边的权重做零值处理。假设加权图用 G = （V，E）表示，V 是图 G 中的节点集合，E 是图 G 中边的集合，其中节点个数为 n。边权重使用高斯核函数计算，公式如下：

$$W(i,j) = exp\left(-\frac{d_{i,j}^2}{\sigma^2}\right) \tag{1}$$

其中 σ 是个参数变量，经验设定 $\sigma = \sum_{i \neq i} \| x_i - y_i \| / [n(n-1)]$。$d_{ij}$为两个图像节点之间的欧式距离，表示为：

$$d_{ij}^2 = \sum_{t=1}^{k} (x_i^t - x_j^t) \tag{2}$$

3.1.2 随机游走过程 在使用重启随机游走算法之前先根据权重矩阵计算图 G 的概率转移矩阵 P：

$$P(i,j) = \begin{cases} W(i,j) / \sum_{t=1}^{n} W(i,t) & (V_i, V_j \in G) \\ 0 & (V_i, V_j \in G) \end{cases} \tag{3}$$

通过以下步骤计算基因之间的相关性得分。首先根据计算出的概率转移矩阵 P，然后生成归一化加权矩阵 P^T。最后基于 RWR 的方法可以描述如下：

$$D_i(t+1) = a \cdot P^T D_i(t) + (1-a) e_i \tag{4}$$

其中 e_i 是表示初始状态。公式（4）的稳态解为：

$$D_i = (1-a)(1-aP^T)^{-1} e_i \tag{5}$$

其中 D_i 是 | V | ×1 向量，e_i 是 | V | ×1 起始向量。（1 - a）被定义为重启概率，其值在 0 和 1 之间。之后可以得到矩阵 S，保存 N（V，E）中每对基因之间的相关性得分。

3.2 计算两个 GO 术语之间的相似性

根据 Peng 等在以往的工作[24]中所表示的方法，计算两个术语之间的相似性，这两个基因本体术语对结合来自基因功能网络和 GO 的信息。设 t_1 和 t_2 为两个术语。将 $D(t_1, t_2)$ 定义为基因集距离，以计算由 t_1 和 t_2 注释的基因组之间的相似性。$D(t_1, t_2)$ 定义为：

$$D(t_1,t_2) = \frac{\sum_{g_i \in G_1} \prod_{g_i \in G_2} d_{ij} + \sum_{g_i \in G_2} \prod_{g_i \in G_1} d_{ij}}{2 | G_1 \cup G_2 | - \sum_{g_i \in G_1} \prod_{g_i \in G_2} d_{ij} - \sum_{g_i \in G_2} \prod_{g_i \in G_1} d_{ij}} \tag{6}$$

其中 G_1 和 G_2 分别是由 t_1 和 t_2 注释的基因集合。d_{ij}是两个基因之间的距离得分，$d_{ij} = 1 - S_{ij}$。S_{ij} 是基于 RWR 的方法计算基因 i 和 j 之间的相关性得分。所有术语对的基因集距离在 0 ~ 1 之间归一化。然后基于标记为 U 的路径约束注释来计算两个术语之间的相似性。在传统的基于 LCA 的方法中，考虑 LCA 的所有后代。路径约束注释方法仅使用与比较术语最相关的术语。相关术语集包括 3 个部分：由术语 t_1 和 t_2 注释的基因集、由共同父节点注释的基因集 p 及其后代项在 t_1 或 t_2 到 p 的路径上。设 t_1 和 t_2 为两个 GO 的两个术语对，p 为它们的共同祖先。t_1 和 t_2 之间的相似性根据 Peng 等在以往工作[24]中提出的方程定义。

$$S(t_1, t_2) = \frac{2log | G | - 2logf(t_1, t_2, p)}{2log | G | - (log | G_1 | + log | G_2 |)} \times \left(1 - \frac{h(t_1, t_2)}{| G |} \times \frac{G_p}{G}\right) \tag{7}$$

其中 G_p（或 G）是由共同祖先术语 p（或根词）及其后代注释的基因集。在等式中 $f(t_1, t_2, p)$ 基于路径约束注释计算相似性，定义如下：

$$f(t_1, t_2, p) = D(t_1, t_2)^2 \times | U(t_1, t_2, p) | + (1 - D(t_1, t_2)^2) \times \sqrt{| G_1 | \times | G_2 |} \tag{8}$$

$h(t_1, t_2)$ 是测量共同父母的特异性，定义如下：

$$h(t_1, t_2) = D(t_1, t_2)^2 \times | G | + (1 - D(t_i, t_2)^2) \times max(| G_1 |, | G_2 |) \tag{9}$$

在公式（9）中，左侧部分测量从项 t_1 和 t_2 到

p 的距离，右侧部分计算从 p 到根的距离。如果存在多于一个最低共同祖先则选择最高得分作为 t_1 和 t_2 之间的相似性。

4 术语相似度算法性能测试与分析

4.1 基于基因相似度的模型测试方法

虽然有很多方法可以评价基因功能相似度，但是没有直接的方法来衡量两个基因本体术语的相似度。因此首先基于术语相似度计算得到基因相似度，然后借助基因之间关系来衡量术语相似度计算模型性能。这一方法在相关研究中被广泛使用。为保证公平性，对于所有被比较的模型，通过使用相同的基因本体术语相似度来计算基因相似度并用此结果衡量术语相似度计算模型性能。如果计算某一物种对应的所有基因本体术语之间的相似度，就可以计算该物种中任意两个至少被一个基因本体术语注释的基因之间的相似度。给定两个基因 g_i 和 g_j 以及注释它们的术语集合 T_i 和 T_j，为计算基因相似度，利用 Wang 等提出的方法将多对术语相似度累积成基因相似度，如公式（10）所示：

$$GS(g_i \cdot g_j) = \frac{\sum_{t \in T_i} sim(t, T_j) + \sum_{t \in T_j} sim(t, T_i)}{|T_i| + |T_j|} \tag{10}$$

公式中对于属于集合 T_i 的每个术语，sim（t, T_i）表示术语 t 和术语集合 T_i 中术语之间相似度的最大值，如公式（11）所示：

$$sim(t, T_j) = , max_{t_j \in T_j} S(t, T_j) \tag{11}$$

公式中 S（t, t_1）表示 t 和 t_j 的所有最小公共祖先中 S（t, t_i, p）的最大值。

基于 EC 编号（酶学委员会）组信息进行评估，由相同 EC 编号标记的基因具有相似功能。基因根据其 EC 编号（完整的 4 位数）分组到不同类别。然后测试同一类别的基因是否具有更高的相似性。在数学上使用记录的倍数变化（Logged Fold Change，LFC）度量[24]进行定量评估。对 EC 编号的 e_i 类别 LFC 分数计算如下：

$$LFC(e_i) = \frac{1}{|EC|} \times \sum_{e_j \in EC; G(e_j) \cap G(e_i) = \emptyset} \frac{\sum_{g \in G(e_i)} diff_g(e_i, e_j)}{|G(e_i)|} \tag{12}$$

其中 G（e_i）是基因组，是 e_i 由标记类别的基因组成；其中满足 $G(e_i) \cap G(e_i) = \emptyset$，$diff_a$（$e_i$, e_j）定义如下：

$$diff_g(e_i, e_j) = In \frac{|G(e_i)| \times \sum_{g' \in G(e_j)} (1 - GS(g, g') + c)}{|G(e_j)| \times \sum_{g^* \in G(e_i)} (1 - GS(g, g^*) + c)} \tag{13}$$

其中 G（e_i）是不包括 g 的 e_i 基因集；c 是拉普拉斯平滑参数；GS（g, g'）、GS（g, g^*）由公式（10）定义。等式（13）测量 EC 间距离和 EC 内部距离之间的差异。基于公式（12）中对对数差异倍数（LFC）得分的定义，如果一个模型的对数差异倍数得分越高，那么该模型的性能越好。

4.2 实验数据与结果

4.2.1 实验数据 2018 年6 月从 GO 网站下载 GO 结构和注释（www. geneontology. org/）工作中只使用“is－a”、“part－of”两种术语之间的语义关系。考虑到数据量太大、时间复杂度过高，所以使用基因数目相对较少的 YeastNet 中包含的基因关联对酵母菌进行评估测试。酵母菌的 EC 分组可以从 www. yeastgenome. org/下载。

4.2.2 实验结果 通过比较不同 EC 类别和相同类别的基因之间的关系，根据 GO 的相似性来评估本文提出的改进算法性能。使用 YeastNet 中包含的基因关联对酵母进行评估测试。LFC 评分作为标准与目前该领域 LFC 评分最高的两种算法 NETSIM 和 NETSIM2 进行比较。NETSIM、NETSIM2 与本文中改进的算法 RWRSM 测量比较每个 EC 组的 LFC 得分，结果显示本文算法在所有 104 组 EC 编码中有 88 组具有最高 LFC 得分，占所有分组的 84.6%，而其他算法只有 15 个 EC 中具有最高 LFC 得分，基于 NETSIM、NETSIM2 及算法 RWRSM 对酵母菌数据表现最佳 LFC 评分的 EC 数量，见图 1。RWRSM 算法的 LFC 评分针对与酵母菌的 EC 分组结果显示在所有评估测量中 RWRSM 的平均值最高，见图 2。3 种计算方法计算出的 LFC 平均值分别是 RWRSM

0.575、NETSIM 20.547、NETSIM 0.534。3 种算法的折线统计，见图 3。其中 RWRSM 算法的折线趋于稳定在 0.6 左右，并且在所有 EC 类别中最高 LFC 分值占比达到 85% 以上，明显优于其他算法。

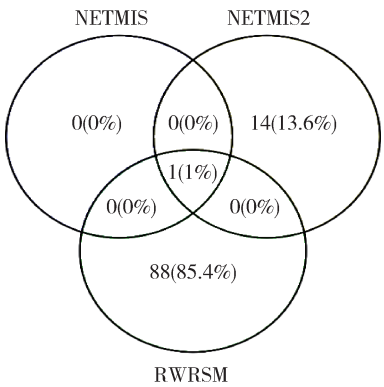

图 1 基于 NETSIM、NETSIM2 及 RWRSM 算法在酵母菌中表现最佳 LFC 评分的 EC 数量

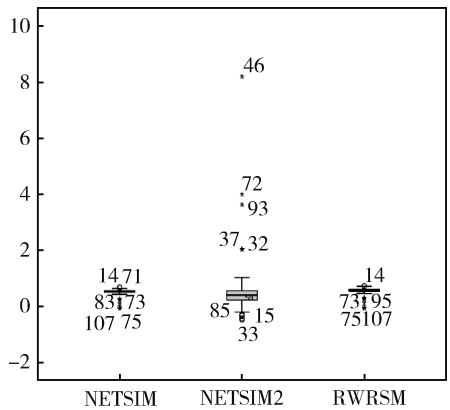

图 2 GO 酵母 LFC 相似性测量得分性能比较

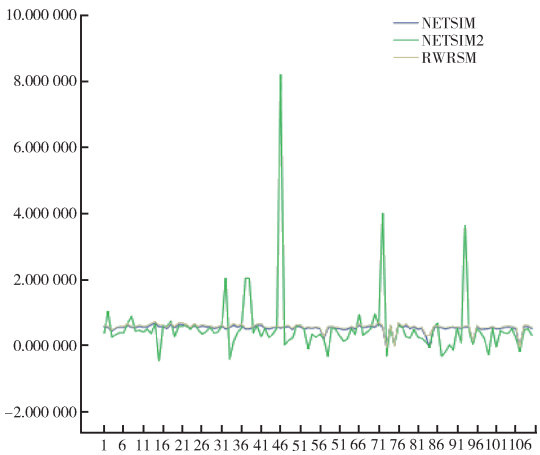

图 3 3 种算法 LFC 相似性得分性能折线

5 结语

基因本体论是用于描述基因和基因产物特性的最流行的生物信息学资源之一。计算基于 GO 的基因功能相似性已被广泛用于多个研究领域。然而低质量的相似性可能源于 GO 的不完整信息和有限的注释量。NETSIM 通过考虑基因关联、GO 结构和注释来解决这些问题。但其仅使用基因共功能网络中的本地关联信息，因为 NETSIM 仅考虑网络中的直接链路。后来提出的 NETSIM2 虽然考率到网络结构的全局性，但是实验结果表明 LFC 评分不够稳定。本文提出一种新的基于 NETSIM2 网络的改进算法，对基于 RWR 的方法考虑共功能网络全局结构，融合高斯核函数得到权重矩阵，包括 3 个步骤：首先，给定基因共功能网络融合高斯核函数得到权重矩阵，基于随机游走和重启方法计算两个基因之间的相关性得分矩阵；其次，通过组合来自共功能网络和 GO 的信息计算两个 GO 项之间的相似性；最后，选择 GO 术语对使用基于标准分数的方法测量两个基因的相似性。EC 实验结果表明本文算法在酵母数据集的所有测量中表现最佳，LFC 结果更加稳定。在所有 104 个 EC 中有 88 个具有最高 LFC 得分，占所有分组的 84.6%，而其他算法只有 15 个 EC 中具有最高 LFC 得分，平均值也明显高于其他算法，测量出的 LFC 评分结果更加稳定。

参考文献

1 Consortium G O. The Gene Ontology Project in 2008 [J]. Nucleic Acids Research, 2008, 36 (Database Issue): D440 – D444.

2 Ashburner M, Ball C A, Blake J A, et al. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium [J]. Nature Genetics, 2000, 25 (1): 25 – 29.

3 Consortium T G O. Gene Ontology Consortium: going forward [J]. Nucleic Acids Research, 2015, 43 (Database Issue): 1049 – 1056.

4 王晓峻，毛莺池，钱国锋. 基于语义的 QoS 感知 Web 服务发现机制 [J]. 计算机科学，2010，37 (8)：133 – 138.

5 蒋哲远，韩江洪，王钊. 动态的 QOS 感知 Web 服务选择

和组合优化模型 [J]. 计算机学报, 2009, 32 (5): 1014 - 1025.

6 Couto F M, Silva M J, Coutinho P M. Measuring Semantic Similarity between Gene Ontology Terms [J]. Data & Knowledge Engineering, 2007, 61 (1): 137 - 152.

7 Consortium T G O. Expansion of the Gene Ontology Knowledgebase and Resources [J]. Nucleic Acids Research, 2017, 45 (D1): 331 - 338.

8 Peng J, Wang T, Wang J, et al. Extending Gene Ontology with Gene Association Networks [J]. Bioinformatics, 2016, 32 (8): 1185 - 1194.

9 Díaz - Montaña JJ, Díaz - Díaz N, Gómez - Vela F. Gfd - net: a novel semantic similarity methodology for the analysis of gene networks [J]. J Biomed Inform, 2017 (68): 71 - 82.

10 Yu G, Fu G, Wang J, et al. Predicting Protein Function via Semantic Integration of Multiple Networks [J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2016, 13 (2): 220 - 232.

11 Nehrt N L, Clark W T, Radivojac P, et al. Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals [J]. Plos Computational Biology, 2011, 7 (6): e1002073.

12 Peng J, Li H, Liu Y, et al. Erratum to: InteGO2: a web tool for measuring and visualizing gene semantic similarities using Gene Ontology [J]. BMC Genomics, 2017, 18 (Suppl 5): 553 - 560.

13 Yang Y, Xu Z, Song D. Missing Value Imputation for MicroRNA Expression Data by Using a GO - based Similarity Measure [J]. BMC Bioinformatics, 2016, 17 (Suppl 1): 10.

14 Peng J, Lu J, Shang X, et al. Identifying Consistent Sisease Subnetworks Using DNet [J]. Methods, 2017 (131): 104 - 110.

15 Pesquita C, Faria D, Falcão A O, et al. Semantic Similarity in Biomedical Ontologies [J]. Plos Computational Biology, 2009, 5 (7): e1000443.

16 Peng J, Wang Y, Chen J. Towards Integrative Gene Functional Similarity Measurement [J]. BMC Bioinformatics, 2014, 15 (2): 1 - 10.

17 Peng J, Li H, Jiang Q, et al. An Integrative Approach for Measuring Semantic Similarities Using Gene Ontology [J]. BMC Systems Biology, 2014, 8 (S5): S8.

18 Peng J, Uygun S, Kim T, et al. Measuring Semantic Similarities by Combining Gene Ontology Annotations and Gene Co - function Networks [J]. BMC Bioinformatics, 2015, 16 (1): 1 - 14.

19 Peng J, Xue H, Shao Y, et al. A Novel Method to Measure the Semantic Similarity of HPO Terms [J]. International Journal of Data Mining & Bioinformatics, 2017, 17 (2): 173.

20 Peng J, Wang H, Lu J, et al. Identifying Term Relations Cross Different Gene Ontology Categories [J]. BMC Bioinformatics, 2017, 18 (Suppl 16): 573.

21 Wu X, Pang E, Lin K, et al. Improving the Measurement of Semantic Similarity Between Gene Ontology Terms and Gene Products: insights from an edge - and IC - based hybrid method [J]. Plos One, 2013, 8 (5): e66745.

22 Sevilla J L, Segura V, Podhorski A, et al. Correlation between Gene Expression and GO Semantic Similarity [J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2005, 2 (4): 330 - 338.

23 Lamesch P, Berardini T Z, Li D, et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools [J]. Nucleic Acids Research, 2012, 40 (Database Issue): 1202 - 1210.

24 Peng J, Zhang X, Hui W, et al. Improving the Measurement of Semantic Similarity by Combining Gene Ontology and Co - functional Network: a random walk based approach [J]. BMC Syst Biol, 2018 (12): 18.