文章编号：2095-2163(2019)01-0108-07 中图分类号：TP391.4 文献标志码：A

基于基因本体的相似度计算方法

荣河江，王亚东
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：对基因或基因产物功能上的相似性的研究是生物信息学关注的热点。基因本体数据描述了基因和基因产物的属性，基于基因本体术语相似度的研究对基因功能的分析、比较和预测具有重要意义。本文研究了基于基因本体的语义相似度计算方法，基于信息量的计算方法，结合最低和最近公共祖先节点，提出了一种改进的相似度计算方法。通过相似度值与专家打分结果相比较的方法进行评价，实验结果表明本方法比传统方法更有效。

关键词：基因本体；语义相似性；信息量

Computation method for semantic similarity based on gene ontology

RONG Hejiang, Wang Yadong
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 Research on the functional similarity of genes or gene products is a focus of attention in Bioinformatics. Gene Ontology describes the attributes for gene and gene products. Go based term similarity calculation is of great benefit to analysis, comparison and prediction of gene function. This paper studies the semantic similarity calculation method based on gene ontology. Based on information content calculation method, combining the least common ancestor with most recent common ancestor nodes, the paper proposes an improved similarity calculation method. By comparing similarity values with expert score results to conduct the evaluation, experimental results show that this method is more effective than traditional methods.

【Key words】 gene ontology; semantic similarity; information content

0 引 言

本体作为知识工程、知识表示等计算机学科的研究热点和重要领域知识描述工具，也成为了生物数据的重要描述方式。本体通过明确的语言，以形式化的方式描述给定领域的概念及概念之间的关联。每一个事物都有一些性质和关系，一个事物的性质和关系统称为事物的属性。比较事物的相似度就是定量评估事物间的属性。本体是研究不同术语关系的一种有效方法。基因本体在过去十年间，成为应用最为广泛的生物医学本体之一。

基因作为遗产信息的携带者，其相似性的研究在生物医学领域受到高度关注。研究基因及基因产物之间的相互作用关系，对于探索生命奥秘具有重要的意义。

基因本体(Gene Ontology)可以计算不同分子功能术语或生物过程术语之间的相似度，从而衡量不同分子功能或生物过程之间的关系；基于疾病本体(Disease Ontology)可以计算不同疾病术语之间的相似度，从而衡量不同疾病之间的关联关系。

随着高通量生物技术的进步，人类基因组测序和多物种基因测序工作的发展，推动了基因识别的研究进程，积累了海量的基因数据。为了更好地利用生物数据及其包含的信息，通过相似性分析获得已有的基因信息来推断基因的结构、功能和进化关系，可以为生物学家的研究提供帮助。相似性分析是生物信息学的基础之一，也是生物信息学的热点研究领域。目前，如何充分利用现有的基因数据和注释信息来计算基因的相似性，提高计算精度，是将基因技术转化为生物医学应用亟需解决的问题之一。

1 基因相似性计算现状

1.1 基因本体研究解析

基因本体项目(Gene Ontology project)从 1998 年创立至今，基因本体数据得到了急速积累。该项目的重点是将基因做系统性的注释，不仅仅用基因的序列来描述，而是用更加丰富多元化的方式来描述基因的性质，基因本体就是基因的注释分类。由于生物系统惊人的复杂性，并且要分析的数据集数量不断增加，生物医学研究越来越依赖于以计算机

作者简介：荣河江(1989-)，男，硕士研究生，主要研究方向：人工智能、知识工程；王亚东(1964-)，男，教授，博士生导师，主要研究方向：人工智能、知识工程、生物信息学等。

收稿日期：2018-06-10

形式存储的知识。基因本体(GO)项目提供了目前可用于计算基因和基因产物功能的最全面资源[1]。

基因本体可用于形式化、规范地注释基因和基因产物的特性。基因本体包含 3 个分支，可阐释如下。

(1)分子功能(Molecular function)：描述分子的化学功能。

(2)生物过程(Biological process)：描述基因产物可行使功能的生物学目的和过程。与分子功能相比，生物过程必须有多个清晰的步骤。

(3)细胞组件(Cellular component)：描述亚细胞结构、位置和大分子复合物。

基因本体信息可以构建一个有向无环图(Directed Acyclic Graph，DAG)。其设计结构可如图 1 所示。节点表示本体中的术语。边表示术语之间的关系，包括“is-a”、“part-of”、“regulate”三种类型。如果一个节点是另一个节点的子孙节点，那么前一个节点是后一个节点的子类型。

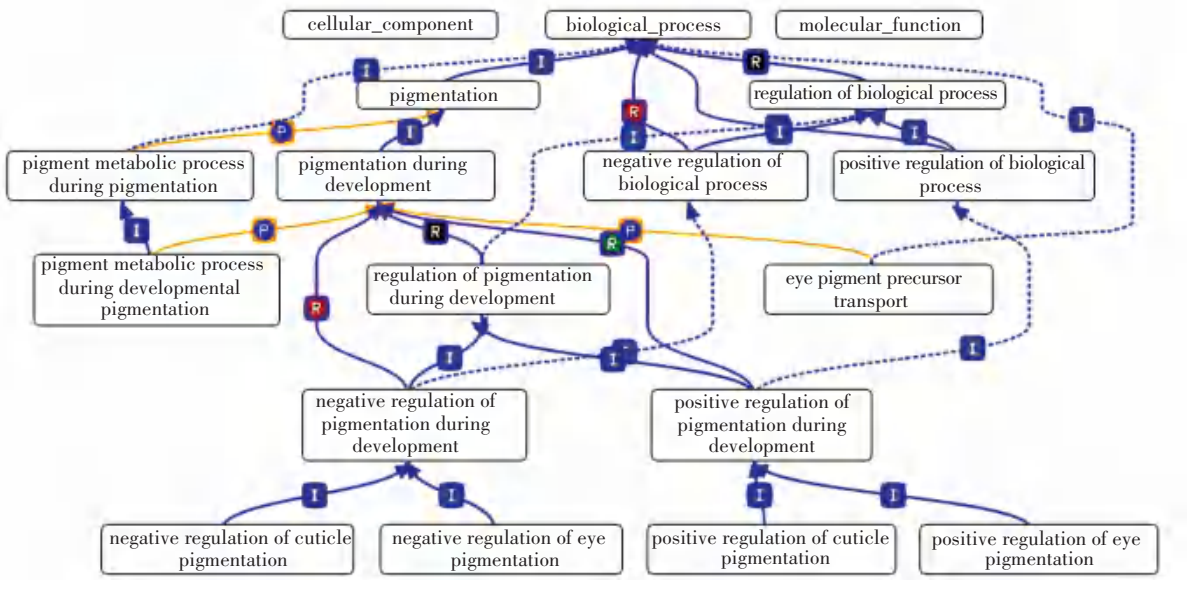

图 1　GO 结构：有向无环图

Fig. 1　GO structure：Directed Acyclic Graph(DAG)

根据 Pesquita 等人[2]2009 年发表文章综述，现有医学本体节点语义相似性的计算分为 3 类，分别是：基于边的方法(edge-based)、基于点(node-based)的方法和混合方法(hybrid)。其中，基于边的方法、即基于距离的方法，主要计算图中 2 个术语之间的路径数目。该技术利用最短路径或所有路径的平均值来定义概念间的距离，如果数据间的距离越大，表明数据之间的相似度越小。基于点的方法，又称为基于信息量(Information Content)的计算方法。该方法认为 2 个术语共享的信息越多，相似度越高。混合方法则基于术语的有向无环图的计算方法，既考虑术语祖先对其信息量的影响，也考虑术语所在位置的影响。诸如，基于语义覆盖的 Combine 算法就是混合方法。

1.2　术语的相似性分析

术语的相似性主要基于 2 个因素：术语包含的信息量(information content，IC)和术语的层次结构。直观上看，如果术语经常被用来注释实体，那么其中所含有的信息量就会减少，即术语的信息量与其所注释实体的数量成反比。

在这里，Sheldon[3]对术语信息量(information content)做出了如下定义：

$$IC(t) = -\log(p(t))$$

$$p(t) = \frac{anno(t) + \sum_{d \in descendant(t)} anno(d)}{\sum_{d \in descendant(root)} anno(d)} \quad (1)$$

其中，$anno(t)$ 是术语 t 的直接注释实体数量，则分子表示 t 及其子孙对应的注释实体数量，分母表示根节点($root$) 子孙节点注释实体数量，即为集合内实体的总数。

$p(t)$ 公式也可以简化表示为：

$$p(t) = \frac{Freq(t)}{|I(DO)|} \tag{2}$$

其中，$Freq(t)$ 表示 t 及其后代的注释实体数量，$|I(DO)|$ 表示集合内实体的总数。

从数学角度解释，基因术语 C 的信息量定义为其在所有基因术语实体中出现的概率 $p(c)$ 的负对数函数。由式(1)、式(2)可以很简明地看出，$p(t)$ 越接近 0，则信息量 $IC(t)$ 越大，这也符合信息增益理论；$p(t)$ 越小，其在信息本体构成的有向无环图中的层次越低(远离根节点)，其注释到的基因也越少，表示的内容越具体，因此信息量越大。

1.3 基于节点的算法研究总述

按照前文信息量的定义，每个基因术语的信息量都可以通过计算自身及其后代节点的注释实体个数得到。研究人员提出如下一种假设：2 种术语，如果这 2 种术语的祖先节点信息量越大，则两者之间就会越相似。由此产生了一系列基于节点的相似度计算方法。

将 2 个节点的祖先(common ancestor)节点定义如下：对于基因术语 C_1 和 C_2，两者通常有一个到多个共同祖先，令 $CA(C_1, C_2)$ 表示这些祖先集合。Resnik[4] 认为 2 个节点共享的信息越多，则这 2 个节点间就越相似，由此提出采用一种语义相似度计算方法：C_1 和 C_2 的相似度用 $CA(C_1, C_2)$ 中信息量最大的那个祖先(most informative common ancestor, MICA)的信息量来表示。研究推得数学公式表述如下：

$$Sim_{Resnik}(C_1, C_2) = IC(MICA(C_1, C_2)) \tag{3}$$

其中，$MICA(C_1, C_2) = arg \max_{c \in CA(C_1, C_2)} IC(c)$

Resnik 提出最大信息量共同祖先的概念。这是较早的、也是经典的语义相似性计算方法，但是该方法只考虑了节点的共性，而忽略了本体架构方面的信息，当 2 个节点的最低公共祖先相同时，其下层的节点语义相似性将不能进行区分。在此基础之上，Jiang 等人[5] 就随即提出了最低共同祖先(lowest/least common ancestor, LCA)的概念。根据最低共同祖先的概念，疾病 C_1 和 C_2 的相似度为 $IC(LCA(C_1, C_2))$。Wu 等人[6] 也相继提出了最近共同祖先(most recent common ancestor, MRCA)的概念，由此得到 MRCA 的数学定义可见如下：

$$MRCA(C_1, C_2) = arg \min_{c \in CA(C_1, C_2)} [dist(C_1, C) + dist(C_2, C) \tag{4}$$

其中，$dist(C_1, C)$ 表示 C_1 与 C_1 和 C_2 的共同祖先 C 的最短距离。根据 Wu 等人的计算方法，疾病 C_1 和 C_2 的相似度为 $IC(MRCA(C_1, C_2))$。

Couto 等人[7] 引入公共不相容祖先结合的概念(common disjunctive ancestors, CDAs)。不相容祖先节点的数学公式描述如下：

$$DisjAnc(t) = \{(C_1, C_2) \mid (\exists P: (P \in Paths(C_1, t)) \wedge (C_2 \notin p)) \wedge (\exists P: (P \in Paths(C_2, t)) \wedge (C_1 \notin p))\} \tag{5}$$

从上述定义可以看出，节点 t 的不相容祖先节点为：如果存在一条路径，从 C_1 到 t 而不通过 C_2，同时有另外一条路径，从 t 到 C_2 而不通过 C_1，那么节点 C_1 和 C_2 就是节点 t 的不相容祖先节点。综合上述定义推得，疾病 C_1 和 C_2 的相似度计算公式为：

$$Sim_{couto}(C_1, C_2) = \frac{1}{|CDAs(C_1, C_2)|} \sum_{c \in CDAs(C_1, C_2)} IC(c) \tag{6}$$

其中，$|CDAs(C_1, C_2)|$ 表示共同不相容祖先的个数。

上述基于祖先节点信息量的疾病相似性方法，都是基于共同祖先的概念，从单个或部分祖先节点的角度考虑 2 个术语的相似性，而忽略了属于自身的信息和不同层次的节点信息差异。以图 2 的 DAG 图为例，利用 *MICA* 的计算方式，2 个节点的相似度为这 2 个节点在 DAG 结构中拥有最大信息量的共同祖先节点的信息量。而在图 2 中，节点 B、C 和节点 B、D 拥有相同的 *MICA*，得到的相似度相同。但是分析后却会发现，C 和 D 属于不同的层次关系，B、C 的相似度应该高于 B、D 的相似度。因此不同层次的节点信息差异也是研究参考因素之一。

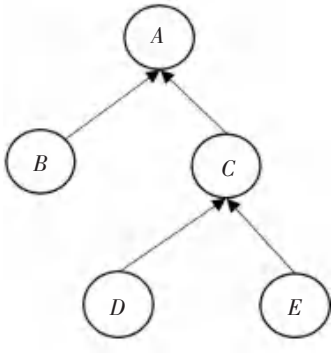

图 2 基因本体 DAG 图示

Fig. 2 Gene ontology DAG illustration

Lin[8] 认为 Resnik 的方法忽略了 2 个术语自身的信息量，当 2 个术语信息量较大时，对相似度准确性的影响也较大。基于此，Lin 则提出了一种新的相似度计算方法，定义疾病 C_1 和 C_2 的相似度为：

$$Sim_{lin}(C_1, C_2) = \frac{2\, Sim_{Resnik}(C_1, C_2)}{IC(C_1) + IC(C_2)} \tag{7}$$

Lin 的方法融入了术语自身的信息量，但是却并未考虑不同层次节点信息差异。直观上，2 个疾病的 *MICA* 越靠近根节点，其所共享的信息量就越小。为了解决这一问题，Schlicker 等人[9]研发提出了修正后的 Sim_{lin} 公式，采用系数 $1-p(c)$ 进行修正。研究中遵循的相似度计算公式如下：

$$Sim_{Schlicker}(C_1,C_2) = Sim_{lin}(C_1,C_2) \times (1-p(c)) \quad (8)$$

完整的计算公式即为：

$$Sim_{Schlicker}(C_1,C_2) = \frac{2IC(t_{MICA})}{IC(C_1)+IC(C_2)} \times (1-p(t_{MICA})) \quad (9)$$

其中，t_{MICA} 表示 t 为疾病 C_1 和 C_2 的 *MICA*，当 t 为根节点时，$p(t_{MICA})=0$，这与实际不符。在此基础上，Li 等人[10]又提出了新的修正系数，调整后的节点 C_1 和 C_2 相似度计算公式如下：

$$Sim_{Li}(C_1,C_2) = Sim_{lin}(C_1,C_2) \times \left(1-\frac{1}{1+IC(MICA(C_1,C_2))}\right) \quad (10)$$

接下来，在基于边和基于点的相似度计算方法之上，Wang 等人[11]设计提出了一种利用混合技术的相似度计算方法。对 DAG 图中的每一条边依据节点的关系（is-a 关系或 as-a 关系）赋予不同的权值。根据定义，一个节点的子有向无环图为该节点及其祖先节点。同时，研究中又定义了一个指标，用于描述位于节点 A 的子有向无环图中的任意一个节点 t，叫做节点 t 对节点 A 的语义贡献值。该值是节点 t 到节点 A 的最优路径中所有边权的乘积。公式描述如下：

$$\begin{cases} S_A(A) = 1 \\ S_A(t) = \max\{w_e \times S_A(t) \mid t \in childrenof(t) \text{ if } t \neq A\} \end{cases} \quad (11)$$

约定 w_e 为语义的贡献因子，根据节点关系的不同，赋予不同范围区间的值。

根据 Wang 等人的定义，2 个节点 C_1 和 C_2 的相似度为：

$$Sim_{Wang}(A_1,B_2) = \frac{\sum_{t \in T_A \cap T_B} (S_A(t)+S_B(t))}{SV(A)+SV(B)} \quad (12)$$

其中，$SV(C_1)$ 为节点 C_1 在 DNA 图中总的语义贡献值。即使是同一个祖先，由于相对于 C_1 和 C_2 的位置不同，该祖先对 C_1 和 C_2 的贡献度可能也不同。Wang 的方法利用了本体节点之间的拓扑结构，考虑到了本体上路径信息特征，虽然不是基于信息量的，但是具有较好的实验效果。

此外，李荣等人[12]提出了语义路径覆盖的概念。该算法根据路径的相交程度来计算彼此之间的相似性的值。采用基于语义链接（semantic links）的方法对每一概念的信息量进行计算。分别计算 2 个节点语义路径上交集节点的信息量之和与并集节点的信息量之和，将两者的比率作为其相似性的值。Combine 算法求解 v_1 和 v_2 的计算过程可分述如下。

（1）计算出 GO 中节点的总个数 N。

（2）对每个节点 φ，计算出该节点的子孙节点的个数 φ'。

（3）计算每个 φ 的子孙节点出现的概率 $p(\varphi)=\varphi'/N$。

（4）计算 φ 的信息量 $IC(\varphi)$。

（5）计算节点 v_1 和节点 v_2 语义路径交 α、语义路径并 β。

（6）分别计算 α 所包含节点的信息量 $IC(\alpha)$，β 所包含节点的信息量 $IC(\beta)$。

（7）定义节点 v_1 和 v_2 的相似性为 $IC(\alpha)$ 和 $IC(\beta)$ 的比率，即：

$$sim(v_1,v_2) = \frac{IC(\alpha)}{IC(\beta)} \quad (13)$$

2　改进的基因相似度计算方法

利用 2 种疾病的祖先节点信息计算 2 个疾病的相似性，方法简单且利于理解。但是只利用祖先节点信息，忽略自身的信息量，以及忽略注释实体之间关系，都会对相似度的准确性造成影响。研究可知，疾病本体 DO 包括 16 层节点，每层节点信息的分布箱线图如图 3 所示。通过对图 3 的考察分析可以发现，节点的位置与信息量成明显相关性[13]，必须充分考虑不同层次节点信息量的差异。

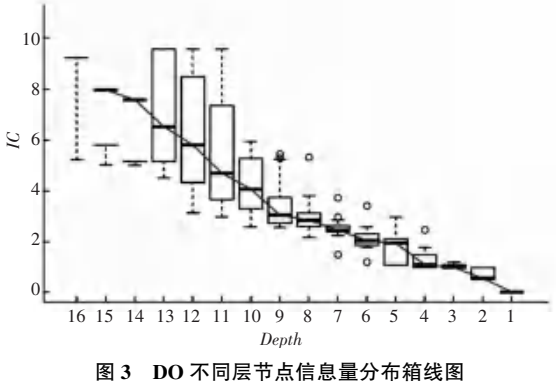

图 3　DO 不同层节点信息量分布箱线图

Fig. 3　Boxplot of information content distribution among nodes from different layer of DO

本文在两节点最近公共祖先点（most recent common ancestor）和最低公共祖先节点（least

common ancestor)的基础上,定义 Ri 公共节点(Ri common ancestor, RiCA)的公式具体如下:

$$RiCA = \alpha \times IC(LCA(C_1,C_2)) + \beta \times IC(MRCA(C_1,C_2)) \quad (14)$$

其中,α,β 为常系数因子;$LCA(C_1,C_2)$ 表示 C_1,C_2 节点的最低公共祖先节点,从 C_1,C_2 的祖先节点集合中,选择距离根节点(root)距离最远的节点,并用信息量 IC 公式计算其信息量。$MRCA(C_1,C_2)$ 表示最近祖先节点,将其定义为 C_1,C_2 的所有祖先节点中,与 C_1,C_2 的距离最短的点,即所跨越的边个数最短的点。换言之就是,LCA 考虑了标定的 2 个术语之间的相似性。而 MRCA 结合了拓扑结构中边的概念,使得注释实体在本体路径上得到了新式的更佳利用。通过乘以 α,β 两个常数因子,将 2 种方法得到的信息量整合起来,形成新的信息量评价值 $RiCA$。

同时,由于两节点的祖先节点相似度计算仅利用到了祖先节点信息,忽略了其它实体关系,而 Li 的公式包含了属于自身信息量和不同层次节点的信息差异,且有较好表现,基于 Li 的公式,引入 $RiCA$ 算子,得出两节点相似度的 Ri 计算即如式(15)所示:

$$Sim_{Ri}(C_1,C_2) = \frac{2RiCA}{IC(C_1)+IC(C_2)} \times \left(1 - \frac{1}{1+RiCA}\right) \quad (15)$$

3 实验对比分析

3.1 实验数据和评价指标

本体术语间相似度的评价是一个难题,一些本体术语间的实际功能相似性仍然无法准确知道。以疾病本体基因为例,一些疾病的关联性需要通过临床验证才能得出准确评价。没有一个公认的评价标准。在这里,研究采用了专家评分的方式进行评价。

李荣的研究团队从 GO 术语库筛选了 25 对术语,分别让 10 位与研究项目无关的生物学专家对术语之间的相似性进行打分[12],然后将专家打分的平均分作为评价基准。计算结果与专家结果的相似度高,则认为该方法准确性高。

本文采用 25 对术语进行筛选,去掉一些已经过时的术语,选取其中的 15 对进行测试。挑选的术语对列表可见表 1。

表 1 挑选的 GO 术语对

Tab. 1 The selected GO items

术语对序号	术语对	
1	oliGOsaccharide-transporting ATPase activity	maltose-transporting ATPase activity
2	cycloartenol synthase activity	pseudouridine synthase activity
3	glucitol transporter activity	propanediol transporter activity
4	Phospholipase activity	phospholipase D activity
5	P-element binding	ribonuclease III activity
6	hydroxypyruvate reductase activity	isocitrate dehydrogrnase(NADP+) activity
7	Interleukin-27 receptor activity	A3 adenosine receptor activity,G-protein coupled
8	shikimate kinase activity	phosphoenolpyruvate-protein phosphotransferase acitvity
9	Spermidine porter activity	KDEL sequence binding
10	phosphomevalonate kinase activity	alpha-I,3-mannosyltransferase activity
11	Lipid-linked peptidoglycan transporter activity	pantothenate transporter activity
12	3-dehydroshikimate dehydratase activity	ion transporter activity
13	Eclosion hormone activity	phytoene synthase activity
14	Oxysterol bingding	DNA ligase activity
15	Sorbose porter activity	tRNA dihydrouridine synthase activity

采用 ZZL 方法、Rensik 方法、Lin 方法、Combine 方法和本文提出方法对 15 对术语的相似度进行计算,计算结果见表 2。

表 2　人工打分各项对比

Tab. 2　The items comparison of manual score

术语对序号	ZZL	Rensik	Lin	Combine	本文方法	人工打分
1	0.999 9	7.901	0.961 2	0.839 6	0.494 7	8.30
2	0.984 4	6.109	0.732 4	0.545 1	0.307 9	3.45
3	0.968 8	5.761	0.721 9	0.511 9	0.292 3	3.45
4	0.998 0	5.886	0.867 2	0.550 1	0.336 9	8.45
5	0.891 6	3.306	0.447 3	0.125 9	0.074 4	2.00
6	0.980 5	3.903	0.494 9	0.386 5	0.229 8	4.35
7	0.941 9	2.832	0.379 6	0.118 6	0.067 0	3.00
8	0.953 1	2.804	0.376 6	0.247 4	0.143 1	4.00
9	0.757 9	1.762	0.264 4	0.032 7	0.018 1	0.90
10	0.886 7	1.835	0.272 2	0.127 5	0.070 8	2.90
11	0.812 5	1.813	0.280 3	0.146 4	0.081 8	3.35
12	0.570 3	0	0.107 1	0.073 0	0.040 1	1.45
13	0.570 3	0	0.074 7	0.053 8	0.023 9	0
14	0.546 9	0	0.079 3	0.059 6	0.034 3	1.45
15	0.816 4	0.526	0.131 3	0.058 0	0.031 9	0.90

3.2　实验结果和分析

经过反复测试数据，当 $\alpha = 0.6$，$\beta = 0.4$ 时，相关系数取得最大值。得到的本文方法与人工打分的相关系数为 0.872 1。研究得到的各种算法的运算结果与人工打分的相关系数可详见表 3。

表 3　各方法的相关系数

Tab. 3　The correlation of different methods

计算术语语义相似度的方法	相关系数
ZZL 方法	0.714 4
Rensik 方法	0.824 1
Lin 方法	0.849 6
Combine 方法	0.863 8
本文方法	0.872 1

从上述的比较结果中可以看出，本文的方法获得了较高的相关系数，表明在计算术语相似度方面取得较好的效果。主要是因为本文研究中不但基于两术语祖先节点的计算方法，而且结合了最近祖先节点和最低祖先节点的信息量，获得较优的信息量评价指标。同时考虑到了语义节点自身包含的信息量和不同层次语义结构的差异性影响，借鉴 Li 提出的语义相似度的方法，有效综合这些因素，因此获得了更好的准确性。

4　结束语

基因术语相似度计算，具有重要的意义。研究在 Li 方法的基础上，结合最近公共祖先节点和最低层次公共祖先节点，提出一种改进的相似度计算方法。该方法能够在评测术语节点间信息量的同时，考虑了 2 个术语自身的信息量，加入信息量均值的处理，使得算法具有较好的表现。实验结果表明，相比其它算法，取得了较高的相关系数。在多数情况下，本算法求出的结果具有更优的生物学意义。在后续的工作中，将进一步综合影响术语相似性度量的因素，以提高算法的正确性。

参考文献

[1] MUSEN M A, NOY N Y, SHAH N H, et al. The national center for biomedical ontology[J]. Journal of the American Medical Informatics Association, 2011, 19(2): 190-195.

[2] PESQUITA C, FARIA D, FALCÃO A O, et al. Semantic similarity in biomedical ontologies[J]. PLoS computational biology, 2009, 5(7): e1000443.

[3] SHELDON R. A first course in probability[M]. London: Pearson Prentice Hall, 2009.

[4] RESNIK P. Using information content to evaluate semantic similarity in a taxonomy[J]. arXiv preprint arXiv: cmp-lg/9511007, 1995.

[5] JIANG J J, CONRATH D W. Semantic similarity based on corpus statistics and lexical taxonomy[J]. arXiv preprint arXiv: cmp-lg/9709008, 1997.

[6] WU Xiaomei, ZHU Lei, GUO Jie, et al. Prediction of yeast protein-protein interaction network: Insights from the Gene Ontology and annotations[J]. Nucleic acids Research, 2006, 34(7): 2137-2150.

（下转第 118 页）