第36卷 第20期 计 算 机 工 程 2010年10月
Vol.36 No.20 Computer Engineering October 2010
·人工智能及识别技术· 文章编号：1000—3428(2010)20—0209—02 文献标识码：A 中图分类号：TP391

基于基因本体的语义相似度研究

魏 韡[1,2]，向 阳[2]，陈 千[2]
(1. 井冈山大学信息科学与传媒学院，江西 吉安 343009；2. 同济大学电子与信息工程学院，上海 201804)

摘 要：针对基因本体的有向无环图结构，提出一种新的计算基因本体中术语间语义相似度的方法。该方法通过计算2个术语的公共祖先及符合条件的不相交祖先，得到不相交祖先的信息量平均值和2个术语的信息量平均值，并将2个平均值的比值作为2个术语的语义相似度。实验结果证明该方法准确度较高。

关键词：语义相似度；基因本体；不相交祖先

Research on Semantic Similarity Based on Gene Ontology

WEI Wei[1,2], XIANG Yang[2], CHEN Qian[2]
(1. College of Information Science and Media, Jinggangshan University, Ji'an 343009, China;
2. School of Electronics and Information, Tongji University, Shanghai 201804, China)

【Abstract】Basedon the directed acyclic graph structure of gene ontology, this paper proposes a novel method to measure the semantic similarity of gene ontology terms. By calculating the common ancestors of two terms and the disjunctive ancestors according with the condition, the average information content of disjunctive ancestors and the average information content of the two terms are obtained. The semantic similarity of two terms is the ratio of the average information content of disjunctive ancestors to the average information content of two terms. Experimental results show that the method has high accuracy.

【Key words】 semantic similarity; gene ontology; disjunctive ancestor

1 概述

大量的生物数据经过各种数据库集成后提供给生物学家使用，但是由于不同的数据库中存在语法尤其是语义的不同，因此具有相同名字的术语会具有不同意义，而意义相同的术语却又具有不同名字。例如，“基因”这个概念在生物学中就有不同的含义。这使得生物学家们浪费很多时间和精力在搜寻生物知识上。本体是“对领域内共享概念模型的明确的规范化说明”，本体论在生物信息学中有着重要的应用。在生物学中使用最广泛的本体是基因本体。基因本体[1]是为了实现对各种数据库中基因产物功能描述相一致而产生的本体。基因本体最初是由1988年对果蝇数据库、酵母基因组数据库和小鼠基因组数据库3个模式生物数据库的整合开始的。之后基因本体不断发展扩大，现在已整合包含数十个动物、植物、微生物的数据库中的基因产物。基因本体旨在建立一个适用于各个物种、对基因和蛋白质进行限定和描述并且能够随着研究的不断深入而更新的语言词汇标准。基因本体包括3个子本体，即分子功能本体、生物过程本体和细胞组件本体。分子功能本体描述基因产物个体的功能；生物过程本体描述分子功能的有序组合；细胞组件本体描述亚细胞机构、位置和大分子复合物。此外，基因本体定义的每一个子本体构成一个有向无环图(Directed Acyclic Graph, DAG)。在有向无环图中的一条路径上，基因本体词汇遵循“真路径”的原则：如果子术语可以描述该基因产物，那么其所有父亲术语也可以描述该基因产物，即可以认为功能上父子术语具有传递性。现有基因本体中，术语之间的关系包括以下2种：父子关系is_a，部分关系 part_of。利用基因本体，每个基因及其产物就可以从其生物功能、参与的生物过程及出现在细胞中的位置进行描述。于是，大量描述性的基因及其产物的注释表现为简洁的标准化词汇，从而为更好地利用生物数据、发掘生物知识提供帮助。

基因本体在生物信息学中的一个重要应用是比较被基因本体注释的术语或概念的语义相似度。例如，通过比较注释相关蛋白质的术语的语义相似度而非仅仅对相关蛋白质的序列比对能够更好地发现功能相似的蛋白质。几种代表性的基于信息量理论的语义相似度计算方法如下：文献[2]给出的计算2个术语的语义相似度方法是基于2个术语的公共祖先的最大信息量。一个术语的信息量与它在本体中出现的频率呈反比，即出现频率越高的术语其信息量越少。就像在语言中经常出现的虚词(例如：这)，其代表的信息量就比较少。文献[3]提出的方法是基于2个术语的公共祖先的最大信息量与2个术语本身所含信息量的比值。基于语义路径覆盖的Combine[4]方法首先计算出每个术语对应节点的信息量，然后分别计算2个节点的语义路径的交的节点信息量之和以及这2个节点语义路径的并的节点信息量之和，将这两者之间的比率作为2个术语的语义相似性度量值。文献[5]实现了一个软件包GOSim。该软件包综合各种基于信息理论的语义相似度算法来计算基因本体中术语的语义相似度。通过计算基因产物相关术语的相似度，间接度量成簇基因的功能相似度。

基金项目：国家自然科学基金资助项目(70771077)；国家“863”计划基金资助项目(2008AA04Z106)

作者简介：魏 韡(1983－)，男，讲师、博士研究生，主研方向：语义网，生物信息学；向 阳，教授、博士生导师，陈 千，博士研究生

收稿日期：2010-03-31 **E-mail：**weiweihzkd@163.com

文献[6]提出了基于信息量理论的计算基因集合之间语义相似度的算法。该算法综合考虑不同的术语对基因集合的语义贡献。

上述方法都把基因本体看作树形结构，只考虑 2 个术语的公共祖先中最大的信息量。但是，基因本体的结构并不是严格的树形结构，它是有向无环图。在树形结构中，一个节点最多只有一个直接祖先节点，但是在有向无环图中，一个节点可以有多个直接祖先节点。

2 基于不相交祖先的语义相似度计算方法

基于基因本体的有向无环图结构，本文考虑用术语的所有不相交的公共祖先来计算语义相似度。首先给出本文方法中的 4 个基本定义。

定义 1 路径：$G=<V, E>$是一个有向无环图，设节点 v_a 和 v_b 之间的路径为 $P=(v_0, v_1,\cdots,v_n)$，其中，$v_0=v_a$；$v_n=v_b$；V_i 是 $V_{i+1}(0\leq i\leq n\text{-}1)$的直接祖先，即 v_i 和 v_{i+1} 存在有向边连接。

定义 2 节点祖先：$G=<V, E>$是一个有向无环图，设节点 v 的所有祖先 $Anc(v)$是所有根节点到节点 v 的路径的并集。

定义 3 公共祖先：$G=<V, E>$是一个有向无环图，设节点 v_a 和 v_b 的公共祖先 $ComAnc(v_a, v_b)$是 2 个节点祖先的并集，即 $ComAnc(v_a, v_b)=Anc(v_a)\cap Anc(v_b)$。

定义 4 不相交的节点：$G=<V, E>$是一个有向无环图，设节点 v_a 和 v_b 分别和某个节点 v 存在路径，但 v_a 和 v 之间的路径不包括 v_b 且 v_b 和 v 之间的路径不包括 v_a，则节点 v_a 和 v_b 为节点 v 的不相交节点。

根据上述定义，本文提出基于基因本体的有向无环图，通过基因本体中 2 个术语的所有不相交的公共祖先来计算其语义相似度。设 v_a 和 v_b 是 2 个术语在基因本体的有向无环图中对应的节点，其语义相似度的计算方法如下：

(1)根据定义 1~定义 3，求出 2 个节点的公共祖先集合 $ComAnc(v_a, v_b)$。

(2)计算公共祖先集合 $ComAnc(v_a, v_b)$中每个节点 v 在有向无环图中出现的频率 $Pro(v)=\dfrac{Fre(v)}{MaxNum}$。其中，$Fre(v)$是节点在有向无环图中出现的次数，即节点的子孙节点个数；$MaxNum$ 是有向无环图中所有节点的个数。

(3)根据信息论的定义，计算 $ComAnc(v_a, v_b)$中每个节点 v 的信息量 $IC(v)=-1\mathrm{b}(Pro(v))$。

(4)按照信息量的大小，将 $ComAnc(v_a, v_b)$中的节点排序。

(5)初始化 2 个节点不相交的公共祖先集合 $DisComAnc(v_a, v_b)$，即集合中只有信息量最大的节点。

(6)根据定义 4 将排好序的 $ComAnc(v_a,v_b)$中的节点依次与 $DisComAnc(v_a,v_b)$集合中的节点比对，如满足不相交节点的定义，则将 $ComAnc(v_a,v_b)$中的节点加入 $DisComAnc(v_a,v_b)$集合。

(7)对 $DisComAnc(v_a, v_b)$集合中的节点的信息量求平均值 $IC(average)$。

(8)计算 v_a 和 v_b 的语义相似度 $sim(v_a, v_b)$：

$$sim(v_a,v_b)=\frac{2\times IC(average)}{IC(v_a)+IC(v_b)}$$

如图 1 所示，采用基于不相交公共祖先的方法计算 GO:0048513 和 GO:0048608 的语义相似度。GO:0048513 的所有祖先集合是{GO:0008150, GO:0032502, GO:0048856, GO:0048731}。GO:0048608 的所有祖先集合是{GO:0008150, GO:0032502, GO:0048856, GO:0003006}。GO:0048513 和 GO:0048608 的公共祖先集合是{GO:0008150, GO:0032502, GO:0048856}。分别计算出公共祖先集合中节点的信息量，并按照信息量的大小排序。GO:0008150 的信息量最小为 0，GO:0032502 的信息量为 2.131 1，GO:0048856 的信息量为 2.511 0。GO:0048513 和 GO:0048608 不相交的祖先集合是{GO:0032502, GO:0048856}。计算出不相交祖先的信息量平均值是 4.642 1。最后计算出 GO:0048513 和 GO:0048608 的语义相似度为 0.607 4。

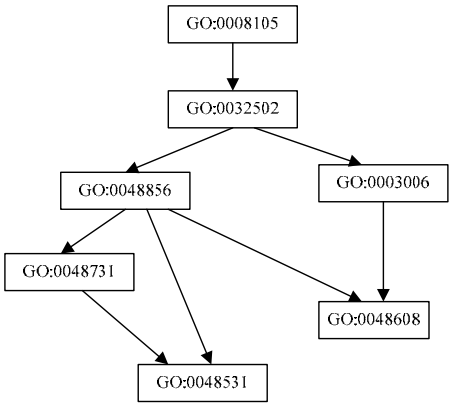

图 1 GO 的部分有向无环图

3 实验对比与分析

本文选取了文献[4]中使用的 25 组基因本体中的术语对来比较各种语义相似度方法的优劣。这 25 组术语对分别用文献[2]方法、文献[3]方法、Combine 方法和改进的基于不相交公共祖先的方法计算其语义相似度，再与通过专家判断的语义相似度进行对比得到这些方法的相关系数，其结果如表 1 所示。从中可以看出，本文方法获得的结果优于其他方法。

表 1 4 种方法的相关系数

语义相似度计算方法	相关系数
文献[2]方法	0.824 1
文献[3]方法	0.849 6
Combine 方法	0.863 8
本文方法	0.875 2

4 结束语

基于信息量计算术语间相似度的方法是基于层次树形来考虑的。Combine 方法虽然在这类方法的基础上有所改进，但仍没有完全考虑基因本体的有向无环图结构。本文提出的基于不相交公共祖先的方法继承了该方法的优点，同时由于针对基因本体的有向无环图结构，因此在理论上比 Combine 方法更精确和全面，实验结果也证明，用本文方法能进一步提高准确率。

参考文献

[1] Ashburner M, Ball C A, Blake J A, et al. Gene Ontology: Tool for the Unification of Biology[J]. Nature Genet, 2000, 25(1): 25-29.

[2] Resnik P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy[C]//Proceedings of the 14th International Joint Conference on Artificial Intelligence. [S. l.]: Morgan Kaufmann Publishers, 1995: 448-453.

[3] Lin Dekang. An Information-theoretic Definition of Similarity[C]// Proceedings of the 15th International Conference on Machine Learning. [S. l.]: Morgan Kaufmann Publishers, 1998: 296-304.

(下转第 219 页)