

基于语义路径覆盖的 Gene Ontology 术语 间语义相似性度量方法*

李 荣^{1,2} 曹顺良² 李园园² 谭 灏² 朱扬勇^{1,2} 钟 扬³ 李亦学^{2,4**}

1. 复旦大学计算机与信息技术系, 上海 200433; 2. 上海生物信息技术研究中心, 上海 200235;

3. 复旦大学生命科学学院, 上海 200433; 4. 中国科学院上海生命科学院生物信息中心, 上海 200031

摘要 进行 GO(gene ontology)语义相似性度量是解决生物学数据集成中语义异构问题的重要方法. 传统方法主要是基于距离的度量和基于信息量的度量. 文中提出了一种基于语义路径覆盖的度量方法, 并实现了其中 Combine 算法. 该算法首先计算出每个节点的信息量, 然后分别计算两个节点的语义路径的交的节点信息量之和以及这两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相似性度量值. 实验结果表明该算法具有较高的相关系数.

关键词 Gene Ontology 语义相似性度量 信息量

从特定蛋白质的功能信息出发, 查找与其功能相似或者相关的蛋白质, 并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. 由于功能相似或相关的蛋白质并不一定具有较强的序列上的相似性, 所以基于序列比对的查询方法不能完全满足需要. 由于各种生物学数据库的广泛分布以及高度的异构等特性, 使得在不同的数据源中查找相关蛋白质的信息极为不便^[1]. 对不同数据源的数据进行集成是解决这一问题的方法之一.

除了数据源间语法上的异构以外, 生物数据集成的主要障碍是语义上的异构. 这种异构既存在于不同数据源之间, 也存在于同一个数据源内. 进行生物学的数据集成必须解决语义上的异构问题, 即需进行语义数据集成^[2], 消除由于系统建立者的应用背景知识不同而带来的理解差异, 由此产生了本体(ontology)这一概念, 以减少或消除概念及术语的混乱. 本体的建立成为数据共享、互操作以及数据集成的基础. 生物本体中使用最广、影响最大的

是由 Gene Ontology 协会开发和维护的 GO^[3]. GO 的结构见图 1, 是一个有向无环图, 图中的节点为 GO 的条目(entry)所对应的术语(term).

GO 应用的一个重要方面就是通过对 GO 术语的语义相似性进行度量, 在被 GO 注释的异构数据源中找出在功能、生物过程或组成成分等方面具有相似性的条目^[4], 如在一个数据库中找出与特定蛋白质具有功能相似性的蛋白质.

计算有向无环图中各个节点语义的相似度的方法可以分为两类: 基于边或距离的方法和基于信息量的方法. 给定一对术语 c_1 和 c_2 , 基于边或距离的度量方法^[5]通过计算分类体系(taxonomy)或本体中与这两个术语相关的节点之间的距离来获得术语间的相似性. 这个距离称为路径距离. 距离越短, 相似性越高.

基于距离的方法的主要缺陷是假设本体中的节点和链接都是均匀分布的^[6]. 然而在本体或分类体系中链接密度可变的情况下, 这一假设并不成立. 比如用 “photoreceptor” (光受体, GO: 0009881)

2005-10-20 收稿, 2005-12-05 收修改稿

* 国家“八六三”计划项目(2002AA231011)和上海市科委重大项目(02DJ14013)资助

** 通讯作者, E-mail: yxli@scbit.org

与“transmembrane receptor”（跨膜受体，GO: 0004888）之间的关系和“chaperone”（分子伴侣，GO: 0003754）与“signal transducer”（信号传感

器, GO: 0004871)之间的关系做比较, 如图 1 所示, 虽然它们有直接的共同祖先, 而且它们之间的距离都相等, 但前两者明显比后两者要相似得多。

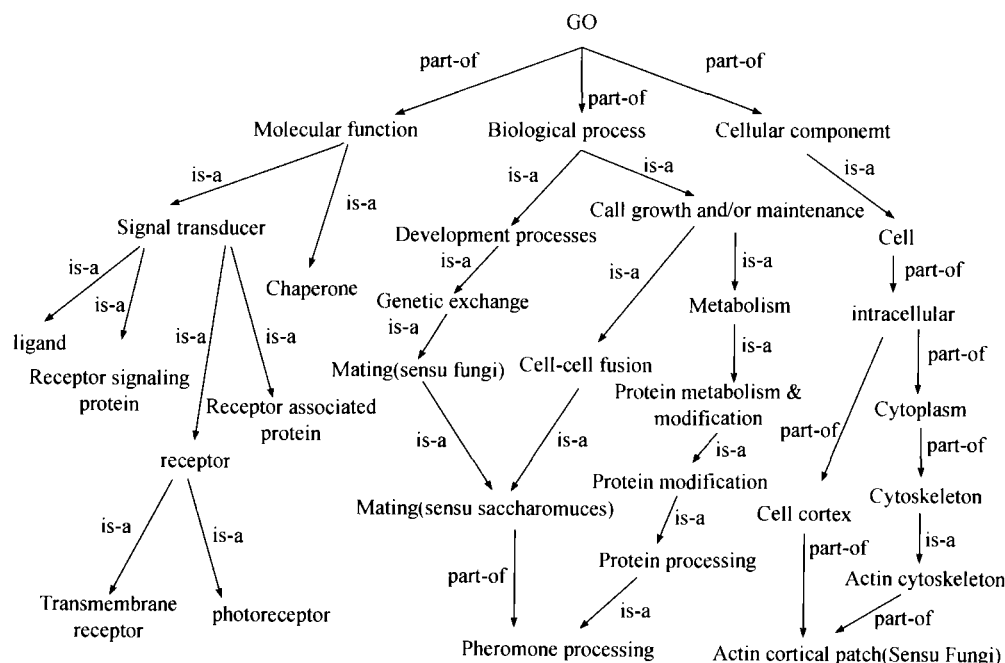


图 1 GO 结构示意图

Zhong 等^[7]提出的基于距离的语义相似性度量算法源于概念层次中上层概念之间的距离比下层概念之间的距离大的想法,也支持兄弟之间的距离比父子之间的距离大的含义。这种算法根据两个概念各自所处的概念层次之间的距离来计算语义相似性。

度量语义相似性的另一种方法是采用信息论^[6]中信息量的概念. 基于信息量的方法对链接密度的可变性问题具有较小的敏感性, 有时甚至是不敏感的^[8].

基于信息量的方法基于如下的假设：两个术语共享的信息越多，则它们越相似。在这种情况下两个术语共享的信息可以使用在本体中包含它们的节点的信息量来进行计算。Resnik^[8]定义的语义相似性度量的取值为两个节点的共同祖先的信息量的最大值。

Lin^[9]提出的方法不仅考虑了两个术语的共同祖先,还考虑了术语本身的信息量.

基于距离的度量方法和基于信息量的度量方法

从完全不同的角度来考虑语义相似性。基于距离的方法比较直观，而基于信息量的方法更具有理论上的合理性。

基于信息量的方法忽略了结构方面的信息，它具有如下一些缺点：所产生的术语比较的结果通常是比较粗糙的，特别是如果术语对的祖先是相同的，那么任何子层的术语对的相似性的值就没有区别，不能加以比较。例如术语对(receptor signaling protein, receptor associated protein)与术语对(receptor signaling protein, transmembrane)相似性的值是相同的。同样，其他类型的链接关系信息也被忽略了。

1 基于语义路径覆盖的 GO 术语间相似性度量方法

为了克服上述两种方法的缺陷,使求出的 GO 术语间的相似性更加符合实际的情况,结合这两种方法我们提出了一种新的基于语义路径覆盖的 GO 术语间相似性度量方法,并实现了一个具体算

法——Combine 算法。由于 GO 的 3 个本体：分子功能 (molecular function)、生物过程 (biological process) 和细胞结构 (cellular component) 是互相独立的，所以分别考虑每个本体。

首先给出在算法中用到的几个基本定义。

定义 1 关系图 关系图是一个有向无环图 $G = \langle V, E, W, r \rangle$ ，其中 $V \neq \emptyset$ 是节点的集合， r 是图的根。 $E \subset V \times V$ 是节点间有向边的集合。 W 是 V 到正实数集合的函数，称为权重函数。

定义 2 路径 $G = \langle V, E, W, r \rangle$ 是一个关系图，设有节点 $v_0, v_1, \dots, v_n \in V, n \geq 0$ ，如果有向边序列 $(e_1, e_2, \dots, e_n)$ 满足任意 $e_i \in E (1 \leq i \leq n)$ 是 v_{i-1} 到 v_i 的边，则称节点序列 $P = (v_0, v_1, \dots, v_n)$ 是从 v_0 到 v_n 的路径。如果两个节点之间没有路径，则定义为空路径。

定义 3 语义路径 设有关系图 $G = \langle V, E, W, r \rangle$ ，称从 r 到任意节点 $v \in V$ 的有向路径是一条 v 的语义路径。 v 的语义路径至少有一条，它们的集合记为 $\phi(v)$ 。

定义 4 语义路径的交 设关系图 G 上有语义路径 $P = (v_0, v_1, \dots, v_n)$ 和 $Q = (v'_0, v'_1, \dots, v'_m)$ ，不妨设 $n \geq m > 0$ ，则 P 和 Q 的交是一语义路径，记为 $P \cap Q$ ，具体定义是：如果存在一个最大的 $k \geq 0$ 满足 $v_i = v'_i, 0 \leq i \leq k, k \leq m$ ，且 $v_{k+1} \neq v'_{k+1}$ ，则 $P \cap Q = (v_0, v_1, \dots, v_k)$ 。

定义 5 语义路径的并 设关系图 G 上有语义路径 $P = (v_0, v_1, \dots, v_n)$ 和 $Q = (v'_0, v'_1, \dots, v'_m)$ ， $n > 0, m > 0$ ，语义路径 P 和 Q 中的所有节点的集合构成 P 和 Q 语义路径的并，记为 $P \cup Q$ 。

根据上述相关论述我们给出一种基于语义路径覆盖的 GO 术语间相似性度量算法——Combine 算法。该算法根据路径的相交程度来计算出它们之间相似性的值，即利用 GO 本身的结构，采用基于语义链接 (semantic links) 的方法来对每一个概念的信息量进行计算。

设 v_1 和 v_2 为 GO 中的两个节点，采用 Combine 算法求解 v_1 和 v_2 相似性值的过程如下：

(1) 计算出 GO 分类体系结构中节点的总个数 N ；

(2) 对每个节点 φ ，计算出它的子孙节点的个数 φ' ；

(3) 计算出 φ 的子孙节点出现的概率 $p(\varphi) = \frac{\varphi'}{N}$ ；

(4) 根据信息论计算出 φ 的信息量 $IC(\varphi) = -\log(p(\varphi))$ ；

(5) 求出节点 v_1 与节点 v_2 的语义路径的交 α ：

$$\alpha = \left(\left(\bigcup_{NP_i \in \phi(v_1)} NP_i \right) \cap \left(\bigcup_{NP_j \in \phi(v_2)} NP_j \right) \right),$$

其中 $\phi(v_1)$ 为节点 v_1 的语义路径的集合， $\phi(v_2)$ 为节点 v_2 的语义路径的集合；

(6) 求出节点 v_1 与节点 v_2 的语义路径的并 β ：

$$\beta = \left(\left(\bigcup_{NP_i \in \phi(v_1)} NP_i \right) \cup \left(\bigcup_{NP_j \in \phi(v_2)} NP_j \right) \right);$$

(7) 求出 α 所包含节点的信息量的和 $IC(\alpha)$ ：

$$IC(\alpha) = \sum_{\omega_i \in \alpha} IC(\omega_i);$$

(8) 相应地，求出 β 所包含节点的信息量的和 $IC(\beta)$ ：

$$IC(\beta) = \sum_{\omega_i \in \beta} IC(\omega_i);$$

(9) 定义节点 v_1 和 v_2 之间的语义相似性为 $IC(\alpha)$ 与 $IC(\beta)$ 的比率，即：

$$\text{sim}(v_1, v_2) = \frac{IC(\alpha)}{IC(\beta)}.$$

2 结果与讨论

如图 2 所示，采用 Combine 算法求 GO: 0042711 与 GO: 0042712 的语义相似性。为简便起见，我们只选取了 GO 中相关的部分。图中标注了每个节点的信息量。这两个节点的语义路径的交为 {GO: 0008150, GO: 0007610, GO: 0000003, GO: 0019098}，其节点信息量的和为 12.7575，语义路径的并为 {GO: 0008150, GO: 0007610, GO: 0000003, GO: 0019098, GO: 0042712, GO: 0042711}，其节点信息量的和为 31.1389。GO: 0042712 与 GO: 0042711 的语义相似性的值为

0.4097.

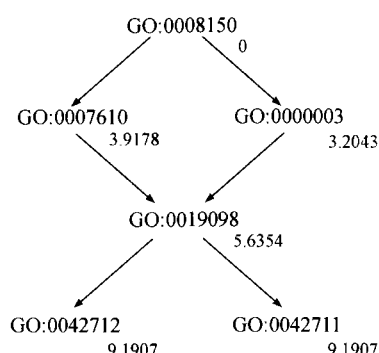


图 2 Combine 算法示例

目前还没有一个标准的方法来对 GO 的语义相似性进行度量,在此,我们采用 Resnik 的做法^[8]来比较各种语义相似性的算法,将相似性算法的计算结果与人工打分的结果相比较,计算它们之间的相

关性。

首先采用 Resnik 提供的算法计算出 GO 中每一术语对的相似性的值,挑选出 25 对覆盖高、中、低分值的术语对,挑选出的术语对见表 1。我们邀请了课题组外 10 位长期专注于研究工作的生物学家(这些生物学家并不了解我们的工作),发放打分表(表中的术语对是无序的),请他们分别对这些挑选出的术语对进行打分。完全没有语义相似性的术语对打分为 0 分,完全相同的(即互为同义词)为 10 分,其他依次类推。将各术语对在打分表中的分值求平均,作为该术语对的相似性度量标准。

其次,分别采用 ZZL 算法^[7]、Lin 算法^[9]以及 Combine 算法等分别求出这 25 对术语对的相似性值,结果见表 2。

表 1 挑选的 GO 术语对

序号	术语对
1 oligosaccharide-transporting ATPase activity	maltose-transporting ATPase activity
2 store-operated calcium channel activity	voltage-gated calcium channel activity
3 cycloartenol synthase activity	pseudouridine synthase activity
4 glucitol transporter activity	propanediol transporter activity
5 phospholipase activity	phospholipase D activity
6 interleukin-20 receptor binding	interleukin-22 receptor binding
7 ATP dependent RNA helicase activity	ribose porter activity
8 P-element binding	ribonuclease III activity
9 adenine deaminase activity	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in other compounds
10 hydroxypyruvate reductase activity	isocitrate dehydrogenase (NADP+) activity
11 interleukin-27 receptor activity	A3 adenosine receptor activity, G-protein coupled
12 wishful thinking binding	death receptor adaptor protein activity
13 shikimate kinase activity	phosphoenolpyruvate-protein phosphotransferase activity
14 spermidine porter activity	KDEL sequence binding
15 phosphatidylinositol-3-phosphatase activity	FAT10 hydrolase activity
16 phosphomevalonate kinase activity	alpha-1, 3-mannosyltransferase activity
17 lipid-linked peptidoglycan transporter activity	pantothenate transporter activity
18 thioredoxin reductase (NADPH) activity	theta DNA polymerase activity
19 3'-5' exodeoxyribonuclease activity	cdysone O-acyltransferase activity
20 chaperonin ATPase activity	dTDP-4-dehydrorhamnose 3, 5-epimerase activity
21 trans-2-methyl-5-isopropylhexa-2, 5-dienal dehydrogenase activity	chlorophyll synthetase activity
22 3-dehydroshikimate dehydratase activity	ion transporter activity
23 eclosion hormone activity	phytoene synthase activity
24 oxysterol binding	DNA ligase activity
25 sorbose porter activity	tRNA dihydrouidine synthase activity

表 2 GO 术语对相似性值

术语对 序号	Resnik	Lin	ZZL	Combine	人工打分
1	7.901	0.9612	0.9999	0.8396	8.30
2	6.109	0.8019	0.9961	0.5309	8.00
3	6.109	0.7324	0.9844	0.5451	3.45
4	5.761	0.7219	0.9688	0.5119	3.45
5	5.886	0.8672	0.9980	0.5501	8.45
6	4.304	0.5380	0.9688	0.5029	7.00
7	4.040	0.5097	0.9966	0.2495	2.90
8	3.306	0.4473	0.8916	0.1259	2.00
9	3.979	0.5347	0.9570	0.2862	3.90
10	3.903	0.4949	0.9805	0.3865	4.35
11	2.832	0.3796	0.9419	0.1186	3.00
12	2.223	0.3139	0.8125	0.1514	3.50
13	2.804	0.3766	0.9531	0.2474	4.00
14	1.762	0.2644	0.7579	0.0327	0.90
15	1.814	0.2700	0.8809	0.0997	3.45
16	1.835	0.2722	0.8867	0.1275	2.90
17	1.813	0.2803	0.8125	0.1464	3.35
18	0.526	0.1313	0.7617	0.0577	1.35
19	0.526	0.1395	0.7529	0.0422	1.35
20	0.526	0.1313	0.7656	0.0698	1.80
21	0.526	0.1313	0.7656	0.0829	2.35
22	0	0.1071	0.5703	0.0730	1.45
23	0	0.0747	0.5703	0.0538	0
24	0	0.0793	0.5469	0.0596	1.45
25	0.526	0.1313	0.8164	0.0580	0.90

最后，将各算法计算出的结果与人工打分的结果求相关系数，结果见表 3。

表 3 各算法的相关系数

相似性算法	相关系数
ZZL 的算法	0.7144
Resnik 的算法	0.8241
Lin 的算法	0.8496
Combine 算法	0.8638

从表 3 中可以看出，Combine 算法的相关系数最高，达到 0.8638，基于距离的 ZZL 的算法的相关系数最低，为 0.7144，而基于信息量的 Resnik 和 Lin 的算法的相关系数介于两者之间。这是因为 Combine 算法综合考虑了相比较的节点(术语对)在本体中所处的相对位置，同时也考虑了每一个术语信息量的大小，因而具有较好的准确性。

3 结论及未来工作

GO 术语语义相似性度量是解决生物学数据集

成中语义异构的重要方法。综合基于路径的度量方法和基于信息量的度量方法，我们提出了一种基于语义路径覆盖的度量方法，并实现了其中的 Combine 算法，相比其他算法具有较高的相关系数。

我们将 Combine 算法应用于解决不同数据库之间或同一数据库之间术语的语义异构问题以及不同基因之间相似性的查找。为了加快语义相似性查询的速度，将计算结果存入 Semantic Similarity 数据库表中，方便地得到 GO 中任意两个术语之间的相似度，也就可以方便地得到被不同的 GO 术语注释的基因或基因产物之间的相似性，从而有效地解决由于语义异构引起的生物信息学数据整合的困难。

参 考 文 献

1 Davidson S B, Overton G C, Buneman P. Challenges in integrating biological data sources. *Journal of Computational Biology*, 1995, 2(4): 557—572

2 Zoe L. *Bioinformatics: Managing Scientific Data*. San Fransisco: Morgan Kaufmann Publishers, 2003, 11—34

3 Ashburner M, Ball C, Blake J, et al. Gene Ontology: Tool for the unification of biology. *Nature Genetics*, 2000, 25(1): 25—29

4 Cao S L, He W Z, Zhong Y, et al. Semantic search among heterogeneous biological databases based on gene ontology. *Acta Biochim Biophys Sin*, 2004, 36(5): 365—370

5 Rada R, Mili H, Bicknell E, et al. Development an application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 1989, 19(1): 17—30

6 Budanisky A, Hirst G. Semantic distance in Word Net: An experimental, application-oriented evaluation of five measures. *Word Net and Other Lexical Resources Workshop Proceedings of NAACL 2001*. Pittsburgh, 2001. 2001, 29—34

7 Zhong J W, Zhu H P, Li J M, et al. Conceptual graph matching for semantic search. *Proceedings of the 10th International Conference on Conceptual Structures: Integration and Interfaces*. Borovets, 2002. Berlin: Springer, 2002, 92—196

8 Resnik P. Using information content to evaluate semantic similarity in a taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Montreal, 1995. San Fransisco: Morgan Kaufmann Publishers, 1995, 448—453

9 Lin D. An information-theoretic definition of similarity. *Proceedings of 15th International Conference on Machine Learning*. Madison, 1998. San Fransisco: Morgan Kaufmann, 1998, 296—304