PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2015, 42(2): 115~122
www.pibb.ac.cn

基于疾病本体的疾病相似性计算方法 *

李 杰[1)**,***] 初砚硕[1)**] 程 亮[1)] 王亚东[1)***] 孔蕾蕾[2)]
([1)]哈尔滨工业大学计算机科学与技术学院，哈尔滨 150001；[2)]黑龙江工程学院计算机科学与技术学院，哈尔滨 150050)

摘要 疾病相似性研究对于复杂疾病发病机制的理解、诊断、预测和药物研发具有重要意义．最近，研究人员通过集成多种疾病术语库，构建了描述疾病关系的疾病本体(disease ontology，DO)，这为从 DO 角度研究疾病相似性打下了基础．本文综述了基于 DO 及其注释信息的疾病相似性计算方法，探讨了疾病相似性计算存在的问题和挑战，为疾病相似性进一步的研究提供有益参考．

关键词 疾病本体，疾病相似性
学科分类号 Q5，Q6，R36
DOI: 10.16476/j.pibb.2014.0140

1 背 景

疾病相似关系的研究对于复杂疾病发病机制理解、重大疾病早期预防、诊断、新药物研发及药物安全评估具有重要作用．比如一些牙齿疾病与一些重要系统性疾病，如消化系统疾病、糖尿病、甚至神经系统疾病(如老年痴呆)等有着密切关系．牙齿疾病和这些疾病相似关系的研究对于疾病的预防具有重要作用．另外，相似疾病常常采用类似药物来治疗，通过疾病相似关系的研究可以发现已在临床上应用的“新”药物，降低药物研发成本和临床试验成本．

疾病种类繁多，国际疾病分类标准 ICD (international classification of diseases)根据疾病的病因、部位、病理及临床表现 4 个主要特征对疾病进行划分．传统上，根据这些特征来描述疾病之间的相似性，在一定程度上是有效的，但在缺乏明显症状的情况下，该方法往往难以奏效．比如，一些一直以来被认为不相关的疾病，却有着相同的生物过程[1]．疾病之间关系复杂，描述疾病关系的信息除了分布在国际疾病分类标准库，还分布在其他医学知识库中，如医学主题词(medical subject headings，MeSH)知识库、孟德尔人类遗传学数据库(online mendelian inheritance in man，OMIM)、临床医学术语标准(systematized nomenclature of medicine——clinical terms，SNOMED CT)、国家癌症研究词库(NCI thesaurus, NCIt) 等．另一方面，随着高通量生物技术的发展，疾病相关分子数据得到快速积累，基于分子数据的疾病注释系统也得到快速发展．目前，如何充分利用疾病知识库及疾病的注释信息来计算疾病相似性，提高计算精度，将一种疾病的治疗知识用于相似疾病的治疗或药物研发，成为转化医学急需解决的问题之一．

本体是有效研究不同术语关系的有效方法，为了更好地描述疾病之间的关系，Schriml 等[2]整合了 MeSH、ICD、NCIt、SNOMED CT、OMIM 等医学知识库中的疾病术语，定义了疾病术语间的复杂逻辑关系，构建了疾病本体库(disease ontology，DO)，从而为基于 DO 的疾病相似性研究打下基础．DO 相关的疾病知识分布在多个医学知识库

* 国家高技术研究发展计划(863)(2012AA02A602,2012AA020404),黑龙江省博士后科研发展基金(LBH-Q13084),哈尔滨市科技创新基金(2014RFQXJ090)和国家自然基金(61471147)资助项目．
** 共同第一作者．
*** 通讯联系人．Tel: 0451-86413316
王亚东．E-mail: ydwang@hit.edu.cn
李 杰．E-mail: jieli@hit.edu.cn
收稿日期：2014-05-19，接受日期：2014-08-26

中，这些知识库含有丰富的疾病注释信息．我们在DO的基础上，集成多个高质量疾病知识库，构建了基于DO的疾病注释系统[3]，通过该系统可以详细查询8000多种疾病的分子成分，浏览疾病之间的关系．这为联合DO结构及其注释信息研究疾病相似性打下了基础.

目前基于DO的疾病相似性计算方法通常利用DO的结构或/和其注释信息，本文将综述基于这两类信息的疾病相似性计算方法，分析不同方法的特点.

2 疾病相似性计算方法

疾病相关的数据多种多样，比如临床数据(临床记录、组织切片、生化检验结果等)、疾病知识库、疾病网络、疾病本体、疾病注释信息等，这些都可用于疾病相似性计算，但是大规模搜集几千种疾病详细的临床数据比较困难，目前疾病相似性计算主要利用DO及其注释信息．这些信息主要包括DO节点信息、DO注释信息、DO节点路径信息．基于这些信息开发了多种疾病相似性计算方法，这些方法可分为两大类(图1)：基于DO注释实体信息和基于DO拓扑结构信息的方法，前者又包括基于两疾病祖先节点注释实体信息、基于两疾病注释实体互作信息两类.

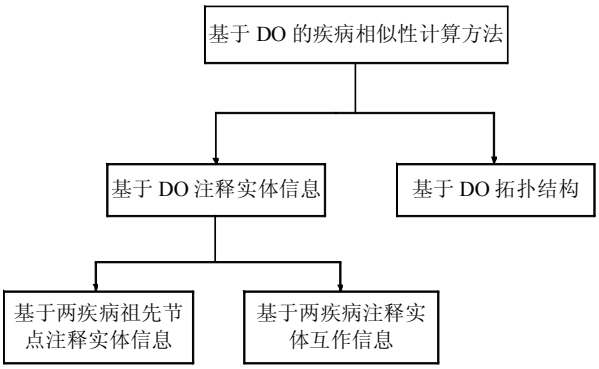

Fig. 1 DO-based methods for disease similarity measurement

图1 基于DO的疾病相似性计算方法

2.1 基于DO注释实体信息量的方法

DO中的每一个疾病术语都和一些分子成分(如基因、蛋白质、小分子、药物等)相关，通常称这些分子成分为疾病的注释实体．通过疾病注释实体信息计算两个疾病相似性的方法，这里称之为基于DO注释实体信息量的方法．两个疾病的相似性也与它们共同的祖先有关，来自同一个祖先节点的两个疾病的相似性通常要大于不属于同一祖先节点的两个疾病的相似性．因此，可以通过计算两个疾病祖先节点的信息量来计算两个疾病的相似度．同样，DO中，每个疾病与其注释实体相关，两个疾病的相似性也可以通过计算它们注释实体之间的关系来计算．前者称之为基于两疾病祖先节点注释实体信息的计算方法，后者称之为基于两疾病注释实体互作信息的方法.

2.1.1 基于两疾病祖先节点注释实体信息量的方法

a．疾病本体节点信息量.

要计算两个疾病祖先节点注释实体的信息量，首先需要计算两个疾病节点的信息量，注释是指用概念描述实体的过程，例如“帕金森综合症”这个具体疾病可以用“疾病”、“神经疾病”等抽象概念进行描述，而这个描述过程成为注释，“帕金森综合症”则为实体．DO中每个节点(疾病)的信息量有多种计算方法．根据信息增益理论，对于疾病 c，Ross[4]认为可将它的信息量(information content，IC)表示为：

$$IC(c)=-\lg p(c)，\text{其中 } p(c)=\frac{freq(c)}{|\mathrm{I(DO)}|}$$

该式将疾病 c 的信息量 $IC(c)$定义为其在所有疾病中出现的概率 $p(c)$的负对数函数．根据信息增益理论，$p(c)$越接近于0，疾病的相关信息越具体，信息量 $IC(c)$就越大．例如一个具体的疾病名称，“Parkinson's disease”(帕金森综合症)，要比一个最普通的疾病名称，“disease”(疾病)，出现的概率低，这样“Parkinson's disease”所含的信息量就要比“disease”大．基于这样的认识，不同研究人员对上式的解释不同，Resnik[5]认为I(DO)为包含DO所有注释实体的集合，|I(DO)|为集合内实体的个数；$freq(c)$为疾病 c 及其后代的注释实体总数，即 $freq(c)_{\mathrm{Resnik}}=|\mathrm{I}(c\cup \mathrm{D}(c))|$，$\mathrm{D}(c)$表示疾病 c 的后代节点．而Zhang等[6]则认为I(DO)为DO所有节点的集合，$freq(c)$为节点 c 及其后代节点的个数，即 $freq(c)_{\mathrm{Zhang}}=|c\cup \mathrm{D}(c)|$．对于DO节点的信息量，不论哪种定义，后代节点的信息量都大于其祖先节点的信息量，即如果 c_1“is_a” c_2，则有 $IC(c_1)\geqslant IC(c_2)$.

DO是有规律组建起来的术语集合，与具有多个后代节点的节点比较，叶子节点通常具有更高的信息量．当一个节点的信息量足够多且足够具体

时，该节点不会再派生出更多的子节点．基于这样的假设，另外一些学者提出了基于子代节点数量和本体节点总量关系的信息量计算方法[7-9]，该类方法与基于信息论的方法不同．Seco 等[7]认为本体中节点的信息量与其后代的多少有关，根据 Seco 对节点信息量的定义， 疾病 c 的信息量为：

$$IC_{\text{Seco}}(c)=1-\frac{\lg[|\text{D}(c)|+1]}{\lg[|\text{I}(\text{DO})|]}$$

其中 |I(DO)| 表示本体中所有节点的数量，|D(c)| 表示集合 D(c)中的元素数．Zhou 等[9]在 Seco 工作的基础上，又考虑了节点深度，将本体节点信息量设为节点后代数量和节点深度的函数，并用参数 k 来调整两部分的贡献比例，应用到 DO 上，疾病 c 的信息量可表示为：

$$IC_{\text{Zhou}}(c)=k\left\{1-\frac{\lg[|\text{D}(c)|+1]}{\lg[|\text{I}(\text{DO})|]}\right\}+(1-k)\left\{\frac{\lg[dp(c)]}{\lg[dp(\text{DO})]}\right\}$$

其中 $dp(c)$ 表示节点 c 的深度，$dp(\text{DO})$ 表示 DO 的最大深度．Vafaee 等[8]则将本体节点的信息量设计为节点深度和局部密度函数的乘积，应用到 DO 上，疾病 c 的信息量可用下式表示：

$$IC_{\text{Vafaee}}(c)=\frac{dp(c)}{\max\limits_{c_k\in \text{D}(c)} dp(c_k)}\times\frac{|Node_{hc}|}{|\text{D}(c)|}$$

其中 $|Node_{hc}|$ 表示节点 c 所在层的节点数．

b．基于两疾病祖先节点注释实体信息量．

DO 是一个有向无环图(directed acyclic graph，DAG)，DO 的每个节点代表一个疾病，都连接到一定数量的实体上，每个疾病的信息量可以根据它们自身及后代节点的注释实体数量来计算．对于两个疾病，如果它们的祖先节点信息量越大，它们越相似，基于这种假设，研究人员提出了一系列可用于计算疾病相似性的方法．

对于疾病 c_1 和 c_2，在 DO 中它们通常有一个或多个共同祖先(common ancestor)，这些祖先用集合 $CA(c_1, c_2)$表示，根据 Resnik 提出术语相似度方法[5]，c_1 和 c_2 的相似度可以用 $CA(c_1, c_2)$中信息量最大的那个祖先(most informative common ancestor，MICA，图 2 中菱形节点)的信息量来表示．

$$\text{Sim}_{\text{Resnik}}(c_1, c_2)=IC(\text{MICA}(c_1, c_2))$$

其中，$\text{MICA}(c_1, c_2)=\arg\max\limits_{c\in \text{CA}(c_1, c_2)} IC(c)$，$IC(c)=-\lg p(c)$，$p(c)=\dfrac{freq(c)}{|\text{I}(\text{DO})|}$

Resnik[5]提出了最大信息量共同祖先的概念(表 1)，并用其标定两个术语的相似性，在 Resnik[5]方法的基础上，Jiang[10]和 Lin 等[11]又先后提出改进的方法．Jiang 等[10]提出了最低共同祖先(lowest/least common ancestor，LCA，图 2 中五角星形节点)的概念，疾病本体中 c_1 和 c_2 的 LCA 可表示为：

$$\text{LCA}(c_1, c_2)=\arg\max_{c\in \text{CA}(c_1, c_2)} dp(c)$$

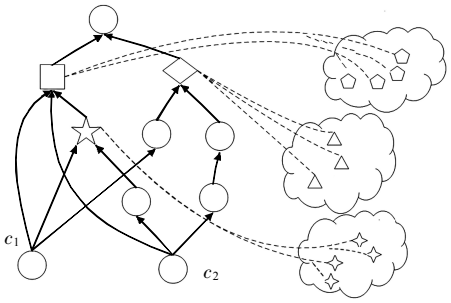

Fig. 2 Types of common ancestor

图 2 共同祖先节点类型

Table 1 Information content of ancestor nodes-based methods for disease similarity measurement

表 1 基于祖先节点信息量的疾病相似性计算方法

计算方法	所用疾病本体信息
Resnik *et al*.[5]	最大信息量共同祖先节点
Jiang *et al*.[10]	最低共同祖先节点
Wu *et al*.[12]	最近共同祖先节点
Couto *et al*.[13]	不相容祖先节点集合

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[12]提出了最近共同祖先(most recent common ancestor，MRCA，图 2 中正方形节点)的概念，c_1 和 c_2 的 MRCA 为：

$$\text{MRCA}(c_1, c_2)=\arg\min_{c\in \text{CA}(c_1, c_2)} [dst(c_1, c)+dst(c_2, c)]$$

其中，$dst(c_1, c)=\min(\#\text{ of edges separating } c_1 \text{ and } c)$，表示 c_1 到 c_1 和 c_2 的共同祖先 c 的最短距离，即最短路径所跨的边的个数．根据该概念，疾病 c_1 和 c_2 的语义相似度为 $IC(MRCA(c_1, c_2))$．Couto 等[13]则提出共同不相容祖先集合的概念(common disjunctive ancestors，CDAs，图 3)，并用其计算术语的相似性．节点 c 的不相容祖先节点集合为所有的、且不经过对方路径的祖先节点对的集合．比如图 3a 中，节点 c 的不相容祖先集合为：c_1 和 c_2、c_1 和 c_3、c_2 和 c_3、c_1 和 c_4、c_2 和 c_4．求 c_1 和 c_2 共同不相容祖先集合的计算方法为：首先将 c_1 和 c_2 的不相容祖先集合取交集，其次对交集中每一对不相容

祖先的信息量进行计算，取信息量最小的祖先节点构建集合．图 3b 展示了 c_1 和 c_2 共同不相容祖先的概念，c_1 的不相容祖先对集合包括 c_3 和 c_4、c_3 和 c_6、c_3 和 c_7、c_3 和 c_8、c_4 和 c_6、c_4 和 c_7、c_4 和 c_8；c_2 的不相容祖先对集合包括 c_4 和 c_5、c_4 和 c_6、c_4 和 c_7、c_4 和 c_8、c_5 和 c_6、c_5 和 c_7、c_5 和 c_8．这两个集合的并集包括的祖先对有 c_4 和 c_6、c_4 和 c_7、c_4 和 c_8，在这 3 对节点中，信息量最小的祖先分别为 c_6、c_7、c_8，这 3 个节点构成的集合称为 c_1 和 c_2 的共同不相容祖先集合．根据以上定义，疾病 c_1 和 c_2 的相似性公式为：

$$\mathrm{Sim}_{\mathrm{Couto}}(c_1, c_2)=\frac{1}{|\mathrm{CDAs}(c_1, c_2)|}\sum_{c\in \mathrm{CDAs}(c_1, c_2)} IC(c)$$

式中 $|\mathrm{CDAs}(c_1, c_2)|$ 表示共同不相容祖先的个数.

研究人员提出了 MICA、LCA 和 MRCA 等祖先的概念，对于 DO 中的两个疾病 c_1 和 c_2，这些概念在某些情形下并不重合(图 2)，这样基于这些概念的计算方法和结果会不相同．以节点 c_1 和 c_2 为例，图 2 中正方形节点到这两个节点的最短路径为 2，是两个节点的共同祖先中距离 c_1 和 c_2 最近的节点；图中五角星节点是正方形节点的子节点，且在祖先节点集合中的“辈分”最小，为 LCA；图 2 中，五边形星形表示被本体注释的实体，虚线表示相应的注释关系，菱形节点对应的注释实体为 c_1 和 c_2 的共同祖先中注释实体数最少的祖先节点，为 MICA.

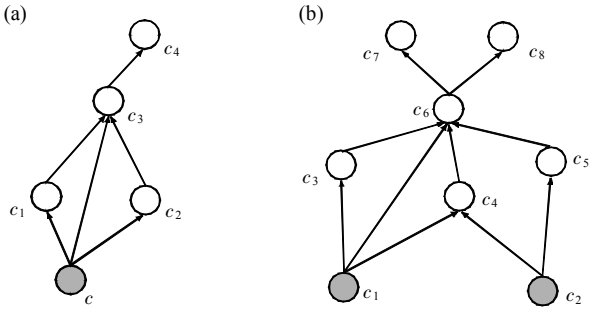

Fig. 3　Disjunctive ancestors (a) and common disjunctive ancestors (b)

图 3　不相容祖先 (a) 和共同不相容祖先 (b)

以上方法提出各种共同祖先的概念(表 1)，根据这些祖先的信息量，可以设计不同的疾病相似性计算方法．这些方法概念易于理解，公式简洁，但这些方法从单个或部分祖先节点的角度考虑两个术语的相似性，忽略了节点自身的信息和不同层次节点信息量的差异．比如疾病本体 DO 包括 16 层节点，每层节点的信息量分布如图 4 所示(根据 Resnik 计算 *IC* 方法计算)，图 4 表明节点的位置越深，信息量越大．因此，需要设计有效的计算方法来消除不同层节点信息量的差异．Lin[11]认为 Resnik[5]的方法仅考虑了两个术语的信息量，没考虑两个术语自身的信息量，当两个术语比较大，且两个术语自身信息量也比较大时，两个术语的语义相似度仍比较小．按照 Lin 提出的方法，疾病 c_1 和 c_2 的语义相似度应为：

$$\mathrm{Sim}_{\mathrm{lin}}(c_1, c_2)=\frac{2\mathrm{Sim}_{\mathrm{Resnik}}(c_1, c_2)}{IC(c_1)+IC(c_2)}$$

Schlicker 等[14]认为，Resnik[5]和 Lin[11]的方法没有考虑本体不同层节点信息量的差异．直观上，两个疾病的 MICA 越靠近根节点，它们共享的信息量就越小；相反，MICA 越远离根节点，MICA 所描述的概念越具体，相应的信息量就越越大．为了克服这一缺陷，Schlicker 等用系数$[1-p(c)]$修正了 Lin 的公式．根据 Schlicker 等的方法，疾病 c_1 和 c_2 的语义相似度如下：

$$\mathrm{Sim}_{\mathrm{Schlicker}}(c_1, c_2)=\mathrm{Sim}_{\mathrm{Lin}}(c_1, c_2)\times[1-p(c)]$$

其中 $p(c)=\dfrac{freq(c)}{|\mathrm{I(DO)}|}$，$c$ 为疾病 c_1 和 c_2 的 MICA. 但当 c 为根节点时，$p(c)$为 0 时，这显然不符合实际．为了解决这一问题，Li 等[15]提出了自己的修正系数，根据 Li 的方法，疾病 c_1 和 c_2 的语义相似度为：

$$\mathrm{Sim}_{\mathrm{Li}}(c_1, c_2)=\mathrm{Sim}_{\mathrm{Lin}}(c_1, c_2)\times\left\{1-\frac{1}{1+IC[\mathrm{MICA}(c_1, c_2)]}\right\}$$

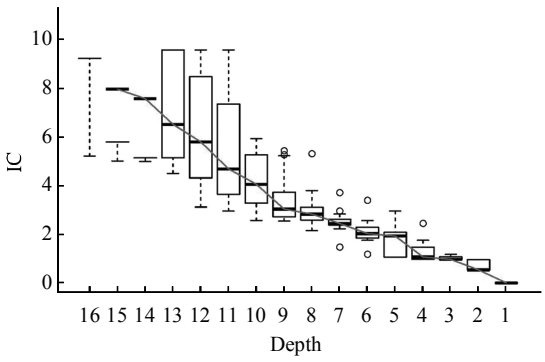

Fig. 4　Boxplot of information content distribution among nodes from different layers of DO

图 4　DO 不同层节点信息量分布箱线图

2.1.2 基于疾病本体注释实体互作信息的方法

基于两疾病祖先节点注释实体信息量的方法计算疾病间的相似性，算法简单、易于理解，但这些方法只利用了祖先节点注释实体的数量信息，没有考虑注释实体之间的关系信息．本体中两个术语的相似性与它们注释实体之间关系密切的程度相关，如两个疾病的注释实体之间有着更紧密的联系，它们应更相似．最近基因组学快速发展，导致更多疾病分子成分被发现和实验证实，疾病注释实体(如基因，蛋白质等)之间的关系得到更准确、完整地描述，在此基础上，研究人员构建了各种分子网络(如蛋白质互作、基因共表达、基因调控、代谢等网络)和本体(基因本体、生物序列本体、蛋白质本体等)．这为从注释实体互作信息的角度精确描述疾病间的关系打下了基础.

为了充分利用基因间的功能关系信息，Mathur 等[16]在基因本体(gene ontology，GO)的基础上提出了如下的疾病相似性计算方法.

首先，获得疾病 c_1 和疾病 c_2 相关的基因集合 $G_1=\{g_{11}, g_{12}, \cdots, g_{1m}\}$ 和 $G_2=\{g_{21}, g_{22}, \cdots, g_{2n}\}$；然后，用 GO 分别对 G_1 和 G_2 进行生物过程注释，获得显著的生物过程术语集合 $T_{c_1}=\{T_{c_1,1}, T_{c_1,2}, \cdots, T_{c_1,l}\}$ 和 $T_{c_2}=\{T_{c_2,1}, T_{c_2,2}, \cdots, T_{c_2,q}\}$．然后将 $T_{c_1,i}$ 和 $T_{c_2,j}$ 的相似性定义为：

$$sim_{\text{go}}(T_{c_1,i}, T_{c_2,j})=\frac{|\text{I}(T_{c_1,i})\cap \text{I}(T_{c_2,j})|}{|\text{I}(T_{c_1,i})\cup \text{I}(T_{c_2,j})|}\times\frac{[IC(T_{c_1,i})+IC(T_{c_2,j})]}{2}$$

$\text{I}(T_{c_1,i})$是与术语 $T_{c_1,i}$ 相关的基因集合，$IC(T_{c_1,i})$ 是 GO 术语 $T_{c_1,i}$ 的信息量．最后，Mathur 等用下面的公式计算疾病 c_1 和疾病 c_2 的相似度：

$$\text{Sim}_{\text{Mathur}}(c_1, c_2)=\frac{1}{2}\left[\frac{\sum_{1\leqslant i\leqslant l}F_{c_2}(T_{c_1,i})}{l}+\frac{\sum_{1\leqslant j\leqslant q}F_{c_1}(T_{c_2,j})}{q}\right]$$

其中，$F_{c_2}(T_{c_1,i})=\max\limits_{1\leqslant j\leqslant q}(sim_{\text{go}}(T_{c_1,i}, T_{c_2,j}))$，$F_{c_1}(T_{c_2,j})=\max\limits_{1\leqslant i\leqslant l}(sim_{\text{go}}(T_{c_2,j}, T_{c_1,i}))$，在两个已知数据上的实验结果表明，该方法具有比较好的计算精度.

两个基因共享生物过程，但并不一定在分子层面存在紧密的关系，与 GO 相比，基因网络更精确地描述了基因之间的关系，基于这种考虑我们提出了基于基因功能网络的相似性计算方法[17]．计算过程为：对于两个基因 g_1 和 g_2，我们用下面的公式描述它们在基因功能网络[18]上的关系：

$$ew(g_i, g_j)=\begin{cases}1 & \text{如果基因 } g_i=g_j\text{，且同属于疾病 } c_1 \text{ 和 } c_2\\ cc(g_i, g_j) & \text{如果基因 } g_i\neq g_j\text{，且在网络上直接相连，它们的关系用相关系数表示}\\ 0 & \text{如果基因 } g_i\neq g_j\text{，且在网络上无直接连线}\end{cases}$$

对于疾病 c_1 和疾病 c_2，与它们相关的基因集合分别为 $G_1=\{g_{11}, g_{12}, \cdots, g_{1m}\}$ 和 $G_2=\{g_{21}, g_{22}, \cdots, g_{2n}\}$．我们用基因 g 和 G_1 内所有基因功能关系的最大值来表示基因 g 和 c_1 的关系，即：

$$F_{c_1}(g)=\max_{1\leqslant i\leqslant m}(ew(g, g_{1i})), g_{1i}\in G_1$$

同样，基因 g 和 c_2 疾病的关系，可用 $F_{c_2}(g)=\max\limits_{1\leqslant j\leqslant n}(ew(g, g_{2j})), g_{2j}\in G_2$ 来表示．在此基础上，将 c_1 和 c_2 的相似性定义为：

$$\text{Sim}(c_1, c_2)=\frac{\sum_{1\leqslant j\leqslant n}F_{c_1}(g_{2j})+\sum_{1\leqslant i\leqslant m}F_{c_2}(g_{1i})}{m+n}, g_{1i}\in G_1, g_{2j}\in G_2$$

实验结果表明，该方法与其他方法相比，具有更好的性能.

2.2 基于 DO 拓扑结构的方法

利用疾病本体注释实体信息计算两个疾病的相似性，方法简单，易于理解．但这些方法从单个或部分祖先节点的角度考虑两个术语的相似性，没有充分利用两个术语在本体上的路径信息，即本体节点之间的拓扑结构信息．因此，另外一些学者从本体拓扑结构的角度来计算术语的相似性.

DO 是一个有向无环图，对于两个疾病 c_1 和 c_2 及它们的上代 n 个节点集合 $CA=\{Ac_1, Ac_2, \cdots, Ac_n\}$，$c_1$ 到集合 CA 的各节点的最短距离分别为 $dc_{11}, dc_{12}, \cdots, dc_{1n}$，$c_2$ 到集合 CA 各节点的最短距离分别为 $dc_{21}, dc_{22}, \cdots, dc_{2n}$，疾病 c_1 和 c_2 的距离定义为 $dst(c_1, c_2)=\min(dc_{11}+dc_{21}, dc_{21}+dc_{22}, \cdots, dc_{1n}+dc_{2n})$，比如，图 2 中 c_1 和 c_2 的距离为 2．两个疾病距离越短，即两个节点之间跨过的边数越少，其相似性越高，基于这种认识，Rada 等[19]较早提出了基于节点距离的术语相似性计算方法，根据 Rada 的方法，疾病 c_1 和 c_2 的相似性为：$\text{Sim}_{\text{Rada}}(c_1, c_2)=dst(c_1, c_2)$．根据上式计算出的，术语距离可能为 0 或无穷，这与实际情况不符．针对这一问题，Lee 等[20]提出了改进的方法，将原来的相似度值域[0, ∞]用归一化参数 λ 进行了处理，使得相似度的范围变为[0, 1]，计算公式为：

$$\text{Sim}_{\text{Lee}}(c_1, c_2)=\frac{\lambda}{\lambda+dst(c_1, c_2)}, 0\leqslant\lambda\leqslant+\infty$$

本体中不同边蕴含的信息量不同，上层节点之间的边所含信息量少，底层节点之间的边所含信息量多，因此在计算两个术语的相似度时，对下面的边应赋予更高的权重．另外，边所含信息量还与边的类型、强度、本体网络复杂性有关．基于这种考虑，Jiang 等[10]将 4 个因素：局部网络复杂度、节

点层级、边的类型、边的强度综合起来，对边进行加权．根据 Jiang 等[10]提出的本体术语相似性方法，DO 中疾病 c 到其父节点 c_p 的边的权重为：

$$ew(c, c_p)=\left[\beta+(1-\beta)\frac{\overline{\mathrm{E}}(\mathrm{DO})}{\mathrm{E}(c_p)}\right]\times\left[\frac{dp(c_p)+1}{dp(c_p)}\right]^{\alpha}\times [IC(c)-IC(c_p)]\times et(c, c_p)$$

这里 $\mathrm{E}(c_p)$表示 c_p 的子节点总边数(本地密度)，$\overline{\mathrm{E}}(\mathrm{DO})$表示 DO 层级结构的平均密度，$dp(c_p)$表示疾病 c_p 在 DO 中的深度，$et(c, c_p)$表示边的类型因素．参数 $\alpha\geqslant 0$ 和 $0\leqslant\beta\leqslant 1$ 用于控制节点深度和密度因素对边权重的贡献程度，$IC(c)$和 $IC(c_p)$分别表示疾病c 和其父节 c_p 的信息量，两者之差表示边的强度．根据 Jiang[10]提出的本体术语相似性方法，疾病 c_1 和 c_2 的相似性为它们之间最短路径所跨过的边的权重之和，即：

$$\mathrm{Sim}(c_1, c_2)=\sum_{i=1}^{N} ew_i$$

这里 ew_i是疾病 c_1 和 c_2 在疾病本体上最短路径所跨过的第 i 个边的权重，N 表示最短路径的边数．两个术语在本体中往往具有多个祖先节点，它们到这些祖先节点的路径有多条，融合这些路径的信息能够进一步提高计算精度．Wang 等[21]提出了一种计算方法来解决这一问题．根据 Wang[21]的方法，疾病 c_1 和 c_2 的相似性为：

$$\mathrm{Sim}_{\mathrm{Wang}}(c_1, c_2)=\frac{\sum_{c_p\in \mathrm{CA}(c_1, c_2)}(k_{c_1}(c_p)+k_{c_2}(c_p))}{\sum_{c_{p_1}\in \mathrm{A}(c_1)}k_{c_1}(c_{p_1})+\sum_{c_{p_2}\in \mathrm{A}(c_2)}k_{c_2}(c_{p_2})}$$

其中，c_p 为 DO 中 c_1 和 c_2 的共同祖先节点，c_{p_1}和 c_{p_2}分别为疾病 c_1 和 c_2 的祖先节点．k_{c_1}的定义如下：

$$\begin{cases}k_{c_1}(c_1)=1\\ k_{c_1}(c_p)=\max\{ew\times k_{c_1}(c')|c'\in \mathrm{childrenof}(c_p)\}\ \mathrm{if}\ c_p\neq c_1\end{cases}$$

这里，ew 为连接 c_1 和 c'的边的权重，c'是 c_p 的子节点．与其他基于本体结构的方法相比，Wang 的方法将 c_1 和 c_2 的共享信息由 MICA 扩展至全部 $\mathrm{CA}(c_1, c_2)$集合，克服依赖单条路径所面临的信息不全问题．Wu 和 Palmer[22]则提出基于本体层级结构的术语相似性计算方法，根据他们的方法，疾病 c_1 和 c_2 的相似性为：

$$\mathrm{Sim}_{\mathrm{Wu\ and\ Palmer}}(c_1, c_2)=\frac{2N_3}{N_1+N_2+2N_3}$$

这里，N_1 为节点 c_1 到 c_1 和 c_2 的最小共同祖先所经历的节点数，N_2 表示从节点 c_2 到 c_1 和 c_2 的最小共同祖先所经历的节点数．N_3 表示从最小共同祖先到根节点所经历的节点数．该方法利用了 3 条路径的信息．Wu 等[12]则提出了另外一种方法，该方法用与本体结构相关的三个参数：α，β，γ 来刻画术语的相似性．

$$\mathrm{Sim}_{\mathrm{Wu}}(c_1, c_2)=\frac{dp(\mathrm{DO})}{dp(\mathrm{DO})+\gamma}\times\frac{\alpha}{\alpha+\beta}$$

α 是从根节点到 c_1 和 c_2 的最近邻共同祖先(MRCA，most recent common ancestor)的最大路径长度．β 是 c_1 与 c_2 到达其叶子节点的最小路径长度中的最大值．γ 是 c_1 和 c_2 经过的 MRCA 的路径之和．该方法不仅利用了祖先节点的路径信息，还利用了叶子节点之间的路径信息，使之不同于别的基于本体结构的计算方法．

3 讨论及展望

通过疾病的相似性可以加快疾病机制研究，相关药物的研发，相似疾病可能具有相似的发病机制，治疗方法和药物．另外，复杂疾病(比如癌症)有效药物的研发成本持高不下，通过寻找相似疾病的药物靶点是研发复杂疾病有效药物的一条有效途径．比如，通过我们的疾病相似度计算方法[17]，发现系统性硬皮病(systemic scleroderma)与风湿性多肌痛(polymyalgia rheumatica)相似度极高，且在化合物与疾病关联的数据库中只有 11 个化合物与系统性硬皮病相关，而没有风湿性多肌痛相关的化合物．我们推测这 11 个化合物中的若干化合物可能与风湿性多肌痛相关．通过查找最近的文献，我们发现 11 个化合物中有 4 种化合物已被证实与风湿性多肌痛相关．

基于疾病相似性计算的重要性，国内外学者提出了多种计算方法．尽管如此，但精确计算疾病的相似度依然存在如下问题：

a．DO 结构仍需进一步完善．疾病关系复杂，DO 的层次结构及其术语之间属性关系的描述都需要进一步完善．临床数据和海量基因组数据的急剧增长，为改进和完善 DO 提供了很好的条件，将来的 DO 需要融合这些数据中的信息．另外，DO 每个边的语义不一样，距离的含义不一样，所含信息量也不同，这样完全基于本体拓扑结构的疾病相似性计算方法就有一定的局限性，将来需要进一步考虑边的生物含义．

b．疾病本体注释信息仍需进一步完善．DO 的注释信息来源于多种医学知识库，这些知识库质量差异较大，来源各异，既有分子数据，也有临床

表型数据，它们从不同的层面对疾病进行注释，如何评价这些数据对疾病的注释准确度，仍是一个挑战性问题．比如，一个实体(比如基因)注释到很多疾病上，那么该实体的信息量会被“冲淡”，利用这样的注释实体计算疾病的相似度会出现偏差．近年来，疾病相关的基因组数据(基因、microRNA、lincRNA 等)快速增长，急需发展有效的方法将这些数据用于注释疾病，增加对疾病的发展机制的理解和治疗，并用于疾病相似度的计算．

c．基于 DO 的疾病相似性计算方法仍需进一步发展．当前大多数方法利用了疾病本体的结构信息和 / 或注释信息，但没有充分利用疾病注释实体之间的网络信息，比如疾病网络、蛋白质网络、基因共表达网络信息等信息，相应的计算方法也比较少．因此，将来需要发展能够融合分子网络、注释实体和本体结构多层面信息的疾病相似性计算方法，进一步提高计算疾病相似性的精度．我们的初步研究成果[17]表明，融合以上信息能够提高计算精度，在已有数据上的比较结果表明，我们的方法计算精度要高于 PSB、Wang 和 Resnik 的方法，具体比较结果见文献[17].

参 考 文 献

[1] Butte A J, Kohane I S. Creation and implications of a phenome-genome network. Nat Biotechnol, 2006, **24**(1): 55−62

[2] Schriml L M, Arze C, Nadendla S, *et al*. Disease ontology: a backbone for disease semantic integration. Nucleic Acids Research, 2012, **40**(Database issue): D940−946

[3] Cheng L, Wang G, Li J, *et al*. SIDD: a semantically integrated database towards a global view of human disease. PloS One, 2013, **8**(10): e75504

[4] Ross S. A First Course in Probability. [S.l.]: Pearson. 2012

[5] Resnik P. Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007 1995

[6] Zhang P, Zhang J, Sheng H, *et al*. Gene functional similarity search tool (GFSST). BMC Bioinformatics, 2006, **7**(3): 135−143

[7] Seco N, Veale T, Hayes J. An intrinsic information content metric for semantic similarity in WordNet. In: ECAI: 2004: Citeseer; 2004: 1089

[8] Vafaee F, Rosu D, Broackes-Carter F, *et al*. Novel semantic similarity measure improves an integrative approach to predicting gene functional associations. BMC Systems Biology, 2013, **7**: 22

[9] Zhou Z, Wang Y, Gu J. A new model of information content for semantic similarity in WordNet. In: Future Generation Communication and Networking Symposia, 2008 FGCNS'08 Second International Conference on: 2008: IEEE; 2008: 85−89

[10] Jiang J J, Conrath D W. Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008 1997

[11] Lin D. An information-theoretic definition of similarity// Preceedings: ICML, 1998: 296−304

[12] Wu X, Zhu L, Guo J, *et al*. Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. Nucleic Acids Research 2006, **34**(7): 2137−2150

[13] Couto F M, Silva M J, Coutinho P M. Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors. In: Proceedings of the 14th ACM international conference on Information and knowledge management: 2005: ACM; 2005: 343−344

[14] Schlicker A, Domingues F S, Rahnenfuhrer J, *et al*. A new measure for functional similarity of gene products based on Gene Ontology. BMC bioinformatics 2006, **7**: 302

[15] Li B, Wang J Z, Feltus F A, *et al*. Effectively integrating information content and structural relationship to improve the go-based similarity measure between proteins. arXiv preprint arXiv: 10010958 2010

[16] Mathur S, Dinakarpandian D. Finding disease similarity based on implicit semantic similarity. J Biomed Informat, 2012, **45**(2): 363−371

[17] Cheng L, Li J, Peng J, *et al*. SemFunSim: a new method for measuring disease similarity by fusing semantic association and gene co-function. PloS one, 2014, **9**(6): e99415

[18] Lee I, Blom U M, Wang P I, *et al*. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res, 2011, **21**(7): 1109−1121

[19] Rada R, Mili H, Bicknell E, *et al*. Development and application of a metric on semantic nets. Ieee T Syst Man Cyb, 1989, **19**(1): 17−30

[20] Lee J H, Kim M H, Lee Y J. Information retrieval based on conceptual distance in IS-A hierarchies. Journal of Documentation, 1993, **49**(2): 188−207

[21] Wang J Z, Du Z, Payattakool R, *et al*. A new method to measure the semantic similarity of GO terms. Bioinformatics, 2007, **23** (10): 1274−1281

[22] Wu Z, Palmer M. Verbs semantics and lexical selection. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics, 1994: 133−138

Disease Ontology-based Methods for Disease Similarity Measurement*

LI Jie[1)**,***], CHU Yan-Shuo[1)**], CHENG Liang[1)], WANG Ya-Dong[1)***], KONG Lei-Lei[2)]

([1)] *School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001, China;

[2)] *School of Computer Science and Technology, Heilongjiang Institute of Technology, Harbin* 150050, China)

Abstract Disease similarity studies for understanding the pathogenesis of complex diseases, diagnosis, prognosis and drug development are important. Recently, researchers constructed Disease Ontology disease (DO) by integrating a variety of disease terms, which builds up the foundation of disease similarity measurement based on DO. We summary disease similarity calculation methods based on DO-related information in this paper and discuss the issues and challenges in disease similarity measurement, offer helpful reference for further research.

Key words disease ontology, disease similarity

DOI: 10.16476/j.pibb.2014.0140

* This work was supported by grants from The Hi-Tech Research and Development Program (2012AA02A602, 2012AA020404), Postdoctoral Science Research Developmental Foundation of Heilongjiang Province (LBH-Q13084), Harbin Science and Technology Innovation Fund(2014RFQXJ090) and National Natural Science Foundation of China(61471147).

**These authors contributed equally to this work.

***Corresponding author. Tel: 86-451-86413316

WANG Ya-Dong. E-mail: ydwang@hit.edu.cn

LI Jie. E-mail: jieli@hit.edu.cn

Received: May 19, 2014 Accepted: August 26, 2014