·医学信息研究·

生物医学领域本体网络整合工具研究进展*

李 猛 吴正荆

（吉林大学公共卫生学院 长春 130021）

〔摘要〕 介绍 3 类国内外生物医学领域本体网络整合工具的研究成果，包括生物医学本体网络整合平台、疾病－药物本体知识发现工具、基因－蛋白质本体集成分析工具，分析其特点及不足，总结本体整合工具开发过程中应该注意的问题，希望能为相关研究者提供借鉴。

〔关键词〕 生物医学；本体；整合工具；关系抽取

Research Progress on Ontology Web Integrating Tools in Biomedicine Field *LI Meng, WU Zheng－jing, School of Public Health, Jilin University, Changchun 130021, China*

〔**Abstract**〕 The paper introduces three kinds of ontology web integrating tools research achievements in biomedicine field both domestic and abroad, including biomedicine ontology web integrating platform, disease－medicine ontology knowledge discovering tool, gene－protein ontology integrating analysis tool. It mainly analyzes the features of every tools as well as their disadvantages, summarizes the problems in the development process of the ontology integrating tools in order to provide reference to researchers.

〔**Keywords**〕 Biomedical science; Ontology; Integration tools; Relation extractesion

20 世纪 90 年代以来，随着计算机科学、信息学、生物医学的基因组学、蛋白组学、代谢组学等学科的不断迅速发展，相关医学文献资源数量正在呈现几何级数增长，如何有效利用这些庞杂的资源成为许多学者关注的焦点，目前逐渐发展成熟的本体研究正在致力于解决这些问题。早在公元前 4 世纪哲学家亚里士多德所确立的重要哲学分支“metaphysics”就是“关于存在的科学”，在很长时间里，本体论也一直被看做是 metaphysics 的同义词[1]，他将本体定义为整个现实世界（本体）的基本特征。目前，基于本体的研究与应用呈现快速的发展趋势，从而不断完善了相关的基础理论，乃至开发了许多应用于不同领域的技术产品，这些都为医学研究提供了帮助。基于已有的数据库或本体，来整合、分析、获取有价值的数据相互关系，受到越来越多的重视。国内外专家学者正在研究创建基于典型本体的网络整合工具，并取得了很大的进展和成果。

1 生物医学本体网络整合平台——PubOnto

PubOnto 是一种基于本体的开放性生物医学 Medline 检索工具，由美国密歇根大学的 Weijian Xuan 等人开发完成[2]，可在网上免费使用[3]。这个系统能从开放式生物医学本体（Open Biomedical

〔修回日期〕 2010－08－29

〔作者简介〕 李猛，硕士研究生，发表论文 2 篇；通讯作者：吴正荆，博士，教授，发表论文 40 余篇。

〔基金项目〕 教育部社会科学基金项目（项目编号：07JA870001）。

Ontology，OBO）中抽取多种的本体，通过多种途径影响和过滤检索的结果，从而方便用户使用。数据来自：基因本体（Gene Ontology，GO），解剖基础模型（Foundational Model of Anatomy，FMA），哺乳动物表型本体（Mammalian Phenotype Ontology，MPO），环境本体（Environment Ontology，EO），集成了 MeSH，PubMed，Gene 数据库的数据管理功能。基于最新的 Adobe Flex 3.0 开发平台，能让用户进行互动，交互浏览数据资源。

主要特点：（1）有效的本体遍历和聚类。（2）基于本体的研究成果检索。（3）Medline 引文显示和分析。（4）数据的可视化。（5）查询服务的开放式结构。（6）其他脚本工具：包括自动化实现检索过程、自定义用户菜单、用户可以保管曾经的查询历史记录。PubOnto 能进行不同数据库之间的跨本体查询，这种功能对于那些对不熟悉领域内研究结果感兴趣的研究人员非常有用。把 Medline 结果映射成大量的可视化数据，然后进行统计分析，为研究人员提供了多种视野角度，对于了解研究进展非常有帮助。不足在于用户查询之前需要了解学科背景，选择正确的数据库，如果 PubOnto 从开始的查询点起，就能自动选择本体，将是非常完美的。这个系统有望合并到更加综合的 PubViz 系统里去。

2 疾病－药物本体知识发现工具

2.1 Disease－Drug Correlation Ontology（DDCO）

DDCO 是疾病－药物关联本体，由美国辛辛那提大学的 Xiaoyan A Qu（AXQ）和 Ranga C Gudivada（RCG）等人开发完成[4]，现无免费版本使用。DDCO 的主要开发目的是能找到某种药物与疾病治疗的隐性关系，DDCO 用网络本体语言（Web Ontology Language，OWL）和资源描述框架（Resource Description Framework，RDF）编辑而成，集成了大量的本体、可控词表、数据模式表，并且把从药理学和生物学领域抽取出来的各种数据表相互链接。数据来源包括：MeSH 表，美国国立癌症研究所叙词表（the National Cancer Institute Thesaurus，NCI Thesaurus），解剖治疗学药品分类系统（The Anatomical Therapeutic Chemical Classification System，ATC System），京都基因与基因组百科全书药物目录（Kyoto Encyclopedia of Genes and Genomes，KEGG Drug Category），常用有害事件术语标准（Common Terminology Criteria for Adverse Event，CTCAE），Gene Ontology 和医学系统命名法——临床术语（Systematized Nomenclature of Medicine ——Clinical Terms，SNOMED CT）。最终设计了 3 个关键子域：Pharmacological 子域（药物及相关化合物子域），Phenomical 子域（疾病及相关临床症状子域），Biological 子域（生物医学子域），Biological 子域在 Pharmacological 子域和 Phenomical 子域之间起着桥梁作用，通过 Biological 子域内的 pathway，gene，molecular phenotype 和 function 等组件来实现其他两子域的连接。目前的 DDCO 有 2 046 种分类（不包括从 GO 直接导入的），平均每个分类有 17 个子类（最多 35 个，最少 1 个）；DDCO 总共有 221 种特性（properties）；有 67 种规则；用语法学方法分析、集成了超过 1 400 种经国家食品药品管理局认证的药物（Food and Drug Administration－approved Drug，FDA－approved Drug）；含 15 068 种人类基因（用 7 124 种唯一的 GO 词条标识）；14 899 种基因－路径（gene－pathway）关联。

主要特点：（1）创建多维集成的 RDF biological－centric 网络结构，收集了 FDA 认证的药物、人类疾病谱的关联、与药物相关的临床特性，最终设计成药理学领域的标识系统。（2）用图表论的分析方法去验证药物－疾病的关联特性和拓扑特性，用 RDF 查询系统产生的子图表相关联的药物候选分级特性，以语义网络结构为基础能推导出一些关键实体和关键过程。（3）首次把网络化集中程度应用到语义学的集成的 pharmacome－diseasome 知识库中，在 RDF 网络中预测的关联程度。不足在于了解疾病－药物机理不够充分，以此去定义最佳特征值，会产生偏倚，当数据、图表多时系统运行比较慢。

2.2 Comparative Toxicogenomics Database（CTD）

CTD 是一种化学药物－基因－疾病（chemical－gene－disease）关联的知识发现工具，由美国

Mount Desert Island 生物学实验室的 Allan Peter Davis 等人开发完成[5]，可在网上免费使用[6]。CTD 开发的主要是目的是通过基因之间的双向关系发现化学药物与疾病之间的关系，CTD 在 270 个类中演示了 3 900 种化学药物和 13 300 种基因的 116 000 种相互作用关系；获得 5 900 种基因 - 疾病和 2 500 种化学药物 - 疾病直接关系，通过集成这些数据，推理出了 350 000 种基因 - 疾病和 77 000 种化学药物 - 疾病关系。查询功能主要由 8 个词表集成：化学药物词表、化学药物修饰语词表、基因词表、基因修饰语词表、作用词表、疾病词表、有机体词表、参考资料词表。

CTD 利用的外部数据来自 3 个类别：（1）CTD 化学药物页面：化合物别名 ID（ChemIDPlus）、化学性致癌研究信息系统的报道、基因 - 毒素（GENE - TOX）、危险物质（Hazardous Substances）、数据库（Data Bank）、药物库（DrugBank）、MeSH 和毒理学资料联机（TOXLINE）。（2）CTD 基因页面：Gene Ontology（GO）标识，KEGG 路径，整合蛋白数据库（UniProt）的核苷酸氨基酸序列，日本 DNA 数据库（DNA Data Bank of Japan，DDBJ），欧洲核酸序列数据库（The European Molecular Biology Laboratory，EMBL），基因数据库（GenBank），美国生物信息技术中心基因数据库（NCBI Entrez - Gene）页面的相关链接，EDGE 数据库的微阵列报道，蛋白质序列页面与基因页面相联系，并且依次链接到蛋白质序列数据库包括 GenPept，InterPro，PRINTS，PROSITE，ProDom，SMART，Pfam 序列的相关记录，如果匹配的话也可以通过果蝇基因数据库（FlyBase）或斑马鱼信息网络库（The Zebrafish Information Network，ZFIN）链接到特种（species - specific）基因页面。（3）CTD 疾病页面：KEGG 的路径，MeSH 和联机孟德尔人类遗传数据库（On - line Mendelian Inheritance in Man，OMIM）的定义和同义词表。CTD 的特点是相关数据库不断更新，集成了化学药物、基因和疾病的核心数据资源，然后通过推理能产生假定存在的新的知识。

2.3 dbNEI 2.0（db neuro - endocrine - immune 2.0）

dbNEI 2.0 是一种能展现药物 - 神经内分泌免疫系统（NEI）- 疾病（drug - NEI - disease）关联的多维网络结构工具[7]，是由清华大学信息科学与技术国家实验室（筹）的李梢教授带领其学生在 2006 年开发的，可在网上免费使用[8]。dbNEI 2.0 开发的主要目的是探索基于 NEI 系统的药物和疾病之间的关联，也能促进建立未来的系统性医学知识体系的一体化视图。dbNEI 系统主要收集了 NEI 分子及其相互作用的数据，能把概念性的 NEI 数据转换成系统的 NEI 网络结构，主要通过 3 个步骤实现其功能：（1）通过基于本体的 GO 数据扩展策略把 NEI 的分子数据扩展到 2 242 种基因和 7 657 种化学物。（2）通过 KEGG 的指令转换和新陈代谢相关路径、次黄嘌呤鸟嘌呤磷酸核糖转移酶（HPRD）的蛋白质 - 蛋白质相互作用关系（protein - protein interactions，PPI），转录因子和 microRNA 序列数据来构建 NEI 分子的多样化相互作用。（3）通过 NEI 分子的多样化相互作用把 611 种药物和 823 种疾病关联起来。

dbNEI 2.0 提供多种检索途径：基因 ID、gene 名称、化合物 ID、KEGG ID、MeSH 词、组织名称、转录因子（TF）和 microRNA，键入检索词可以看到其相关数据，同时根据其检索结果画出不同类型的网络结构图。dbNEI 2.0 的主要特点是提供 drug - NEI - disease 相互作用的多维网络图，并且可以为探索药物和疾病潜在的新关联提供帮助。

3 基因 - 蛋白质本体集成分析工具

3.1 GS2PATH

GS2PATH 是一种能够查询基因本体和生物化学的路径数据之间关联的网络集成分析工具[9]，由韩国的生命工学研究院（KRIBB）生物信息中心（KOBIC）的 Jin Ok Yang，Charny Park，Byungwook Lee 等人开发完成。GS2PATH 能提供 GO 本体和生物学路径数据库的 GO 词条的聚集程度，并且能通

过计算超几何（hyper -geometric）概率 GO 词条的 P - 值分析评估基因集的聚集程度，把相关的生物化学路径（KEGG 和 BioCarta 路径数据库）转化为图像，在以下领域为用户提供帮助：新陈代谢、指令转换、遗传的信息进程、环境的信息进程、细胞学的进程、疾病和药物的发展。GS2PATH 含有 4 个关键模块：询问处理程序模块（Query Processor）、GO 数据存取模块（Query Processor）、KEGG 数据存取模块（KEGG Accessor）、BioCarta 数据存取模块（BioCarta Accessor）。

主要特点：（1）提供 GO 词条和路径之间的功能关联。（2）基因集的聚集程度的超几何测验。（3）向上规则（up -regulation），向下规则（down -regulation）基因集的多重检索。（4）GO 词条的多样化过滤选项。（5）用户自定义基因路径的颜色。

3.2 Word Add in For Ontology Recognition

Word Add in For Ontology Recognition 能进行本体识别，能增强科学文献的语义学功能，是由美国加里福尼亚大学的 J. Lynn Fink 等人开发完成[10]，可在网上免费下载并使用[11]。这个系统主要是应用 Microsoft Word 2007，. NET 平台，可以在 Windows 界面上安装，主要默认 XML 格式和扩展的 Word 格式的文件。系统具有自定义、自动化功能，有 3 个主要医学本体数据库：蛋白质数据库（Protein Data Bank，PDB），整合蛋白质知识库（UniProt Knowledgebase，UniProtKB），NCBI 数据库基因库（GenBank）和参考序列（RefSeq）。

主要特点：系统能为作者在写文稿的时候添加语义学短语，增强书面语言的语义学意义，为相关短语提供上下文语境，从而得到这个短语的更多评价信息。如果系统中没有可供选择的本体库，可以再添加本体库，有用的词可以定位到信息面板（InfoPane）内，然后应用需要的词或短语。

3.3 Lists2Networks

Lists2Networks 是一种根据基因/蛋白质列表的集成分析工具，是由美国纽约系统生物学中心的 Alexander Lachmann，Avi Ma′ayan 等人开发完成[12]，在网上经注册后可以免费使用[13]。Lists2Networks 是基于网络的系统工具，用户可以上传人类 genes/proteins 列表到客户端进行分析。这个系统一次性应用许多已有的列表，用不同的操作对列表进行分析，包括对蛋白质 - 蛋白质（protein - protein）的关联分析，共现（co -expression）关联分析，背景知识共标识（co -annotation correlation）分析，也可以对基因列表的路径、基因本体术语、激酶底物、microRNA - mRAN、蛋白质 - 蛋白质相互作用、代谢产物、蛋白质领域分析。L2N 系统用 PHP，JSP/Java 和 JavaScript 编辑而成，数据储存在 MySQL database 数据库中，用 Asynchronous JavaScript 和 XML（AJAX）更新其部分数据库。系统主要包括 6 大模块：上传列表模块、后台网络扩展列表模块、操作列表模块、分析重复列表和丰富功能性术语模块、蛋白质 - 蛋白质相互关系浏览模块、同其他用户分享列表进行交流模块。

主要特点：（1）简化了基因列表的分析和扩展过程，可以进行许多不同实验数据的集成分析，也为假说的提出提供依据。（2）用蛋白质 - 蛋白质相互作用、共现、共标识方法来扩展分析列表的功能。（3）交流、分享列表的合作方式。（4）合并、补充、剔除网络结构的功能。（5）集成已有数据的功能。

4 结论

以上介绍的 3 类本体整合工具应用到了不同的领域，这些工具有不同的数据资源、语言和系统开发，经过修订有的已形成了概念的统一性，避免了概念的等级关系混乱性，但有的工具没有。目前在不同的数据库或本体中集成分析整合工具存在着一些问题，主要表现在以下几个方面：（1）基础本体构建方法不统一。本体的建构从本质上说是一种组织或团体意义上的决策行为，专家的知识是语境相关且独立构建的，功能强大但难免比较片面，因此很难构建一个可以满足所有成员使用需求的本体[14]。所构建的本体只在专家熟知的领域内应用，无法扩大资源的共享，甚至在熟知的领域内也未达

到良好的共识，本体构建还缺乏成熟的方法论。这样使基于不同本体的集成分析工具过于庞杂，无法建立有影响力的整合工具。（2）缺乏不同本体之间的资源转换工具。在不同的表达形式之间的本体转换是迫切需要解决的问题，而缺乏自动转换工具，必然会带来很多问题[15]。本体一般都是由不同组织独立开发完成，同是生物医学领域的本体也存在着不同程度的重叠，这影响了本体在信息交互中的作用，也影响生物医学领域本体网络整合工具的开发进程。所以进一步研究本体转换工具，使用一种高效准确的转换工具来剔除、添加、修订等形式来规范不同数据库的资源，以达到领域本体的整合，将是未来的研究方向之一。（3）缺乏本体整合工具评价标准。目前缺乏对整合工具的评价标准，大多数工具都是放在网上供专家学者免费使用，然后进行反馈，这样的反馈缺乏理论性和操作性，因此需要进一步研究对网络本体整合工具的评价标准。

参考文献

1 Alexander Maeche. Ontology Learning for the Semantic Web〔M〕. Norwell: Kluwer Academic Publishers, 2002: 15 - 17.

2 Weijian Xuan, Manhong Dai, Barbara Mirel, et al. Open Biomedical Ontology - based Medline Exploration〔J〕. BMC Bioinformatics, 2009, 10 (Suppl 5): S6.

3 PubOnto: open biomedical ontology - based Medline exploration〔EB/OL〕. 〔2010 - 4 - 12〕. http://brainarray.mbni.med.umich.edu/brainarray/prototype/PubOnto/.

4 Xiaoyan A Qu, Ranga C Gudivada, Anil G Jegga, et al. Inferring Novel Disease Indications for Known Drugs by Semantically Linking Drug Action and Disease Mechanism Relationships〔J〕. BMC Bioinformatics, 2009, 10 (Suppl 5): S4.

5 Allan Peter Davis, Cynthia G. Murphy, Cynthia A. Saraceni - Richards, et al. Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical - gene - disease networks〔J〕. Nucleic Acids Research, 2009, (37): 786 - 792.

6 Comparative Toxicogenomics Database〔EB/OL〕. 〔2010 - 4 - 12〕. http://ctd.mdibl.org./.

7 Jing Zhang, Tao Ma, Yanda Li, et al. dNEI2.0: buiding multilayer network for drug - NEI disease〔J〕. Bioinformatics Applications Note, 2008, 24 (20): 2409 - 2411.

8 dbNEI Introduction: what's the dbNEI?〔EB/OL〕. 〔2010 - 4 - 12〕 http://bioinfo.au.tsinghua.edu.cn/dbNEIweb/index.php.

9 Jin Ok Yang, Charny Park, Byungwook Lee, et al. GS2PATH: a web - based integrated analysis tool for finding functional relationships using gene ontology and biochemical pathway data〔J〕. Bioinformation, 2007, 2 (5): 194 - 196.

10 J. Lynn Fink, Pablo Fernicola, Rahul Chandran, et al. Word Add - in for Ontology Recognition: semantic enrichment of scientific literature〔J〕. BMC Bioinformatics, 2010, (11): 103.

11 Word Add - in For Ontology Recognition〔EB/OL〕. 〔2010 - 4 - 12〕. http://ucsdbiolit.codeplex.com/.

12 Alexander Lachmann, Avi Ma'ayan. Lists2Networks: integrated analysis of gene/protein lists〔J〕. BMC Bioinformatics, 2010, (11): 87.

13 Lists2Networks〔EB/OL〕. 〔2010 - 4 - 12〕. http://amp.pharm.mssm.edu/lachmann/upload/register.php.

14 E Daniel , O Leary1. Impediments in the Use of Explicit Ontologies for KBS Development〔J〕. International Journal of Human - Computer Studies , 1997, 46 (223) : 327.

15 M Uschold, P Clark, M Healy, et al. An Experiment in Ontology Reuse〔C〕. Banff , Canada: The 11th Knowledge Acquisition Workshop, 1998.