Open Journal of Nature Science 自然科学, 2017, 5(2), 185-189
Published Online May 2017 in Hans. http://www.hanspub.org/journal/ojns
https://doi.org/10.12677/ojns.2017.52025

Hans汉斯

Advances in HIV Protein Ontology

Liwen Zhang, Heng Chen*

Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai
Email: *chenheng@sibs.ac.cn

Received: May 4th, 2017; accepted: May 18th, 2017; published: May 25th, 2017

Abstract

As a way of knowledge expression and organization, ontology plays an important role in information integration and data mining. At present, the biological ontology has widely received more and more attention and focus in the research on knowledge organization and management. In this paper, the relevant concepts of ontology are combed, and the classification and application of the biological ontology are summarized, and the research status, problems and prospects of the HIV protein ontology are introduced so as to provide guidance and enlightenment for the subsequent studies.

Keywords

Ontology, Biological Ontology, HIV Protein Ontology, Advance

HIV蛋白本体研究进展

张丽雯，陈　恒*

中国科学院上海生命科学研究院，上海
Email: *chenheng@sibs.ac.cn

收稿日期：2017年5月4日；录用日期：2017年5月18日；发布日期：2017年5月25日

摘　要

本体作为一种知识的表达与组织方式引入到数据挖掘与信息整合中，发挥着重要的作用。目前，生物学领域的本体在知识组织和管理研究中越来越受到广泛关注和重视。本文梳理了本体的相关概念，综述了生物学本体的分类及应用，并详细介绍了HIV蛋白本体的研究现状、存在的问题和未来展望，以便为后续研究工作提供指导和启发。

*通讯作者。

文章引用: 张丽雯, 陈恒. HIV 蛋白本体研究进展[J]. 自然科学, 2017, 5(2): 185-189.
https://doi.org/10.12677/ojns.2017.52025

关键词

本体，生物学本体，HIV蛋白本体，研究进展

1. 本体

本体概念从哲学引入，20 世纪 90 年代被应用到计算机领域中，随着本体理论与技术研究的不断深入，本体的应用逐渐扩展到知识工程、信息检索及自然语言处理等领域。德国学者 Studer 在对本体进行深入研究后，指出本体的新定义[1]：“共享概念模型的明确的形式化规范说明”。该定义主要包含了共享、概念模型、明确和形式化 4 重含义，共享反映了本体描述的是领域内共同认知的知识；概念模型是指抽象于具体客观世界的概念的模型，体现了概念独立于具体事物的状态；明确则表明概念以及概念之间的关系具有明确的定义；形式化表示本体能够被计算机所处理，反映的是本体与计算机及计算机之间的交互[2]。

通过对本体特定领域的知识进行描述，可得到相应领域内大家所共同理解与认可的概念，并以一定的形式化模式对这些术语及术语间的相关关系进行准确的定义与限定，实现本体所描述的领域知识的统一认知与理解，能够更好地实现知识的组织、管理、发现与抽取，提高知识的挖掘质量与效率[3]。如今本体已被广泛应用于图书情报、数字图书馆、信息检索、Web 异构信息处理和软件复用等领域。典型应用有：1) 基于语义的信息检索，特别是网络搜索引擎和数字化图书馆；2) 基于本体的数据集成、机器学习等；3) 领域本体的应用。比如，在生物信息学中建成的基因本体；4) 语义 Web 服务；5) 在线元数据管理和自动信息发布[4]。

2. 生物学本体

生物学本体是生物领域的科学知识和本体方法有机结合的产物，能够用来建立生物科研领域的知识结构和概念模型。在研究过程中，生物学家收集事实现象与经验信息，并用自己的语言加以记录，然后使用这些知识解释未知现象，但这些语言描述的知识对于计算机而言可读性极差，难以进行直接识别和应用[5]。基于本体的知识表示能够促进信息抽取与检索，并且支持数据的互操作性，所以生物本体可以规范地表示已定义的生物术语之间的关系，有效地组织生物数据并能充分体现其语义信息，使这些专业术语可以同时被人类和计算机识别，帮助生物领域的研究人员对相关知识达成一致理解，从而顺畅地进行数据的交换和探索[6]。

一般来说，生物学相关本体研究和应用主要集中在 7 个方向：基因表达研究；基因表达的信号传导研究；蛋白质相关研究；基因及染色体相关研究；本体相关的系统及软件开发；与本体相关的词表的研究；计算生物学和基因组学的方法学研究等[7] [8]。目前最具有权威性的生物学本体研究组织为开放生物医学本体组织(Open Biomedical Ontology, OBO)，该组织下的本体项目有基因本体(Gene Ontology)、蛋白本体(Protein Ontology)、序列本体(Sequence Ontology)、植物本体(Plant Ontology)、疾病本体(Disease Ontology)等[9]。与蛋白质相关的本体主要有基因本体(GO)和蛋白本体(PRO。GO 是一个结构化的术语系统，旨在统一各种基因产物数据库的信息表达方式。从结构看，GO 主要包含结构组件、分子功能和生物过程 3 个

子本体，通过“is a”和“part of”等语义关系将生物学概念互相关联起来构成一个大型的语义网络。PRO是一个由EBI开发的关于蛋白的本体库，作为OBO项目的一个子项目，主要针对UnitProtKB/SWISS-Prot和MGI中的人和鼠蛋白，并且是以疾病相关的蛋白为主。可从两方面进行分类，第一个是针对蛋白质domain的进化关系，另一个针对蛋白质的各种存在形式。PRO不仅自身有自己的词汇结构和结构关系，而且它还和OBO中其他的本体相关联。

3. HIV蛋白本体

3.1. HIV概述

艾滋病毒(HIV)是一种逆转录病毒，它感染人类的免疫系统细胞(主要是T淋巴细胞)，摧毁或损害其功能。感染初期没有症状，但是随着感染的发展，免疫系统开始变弱，患者更加容易遭受机会性感染[10]。HIV可分为HIV-1与HIV-2两型。最初发现的是HIV-1病毒，其感染性更强。多数国家的HIV感染是由HIV-1造成的，并且感染HIV-1后超过90%的患者会在10~12年内发病成为艾滋病；HIV-2主要分布在非洲西部，其感染往往没有相关的病症[11]。

HIV病毒颗粒呈球型，直径约为100~120 nm (如图1所示)，双链RNA位于核衣壳内，外膜上镶嵌着gp41蛋白以及与其非共价结合的gp120蛋白，这两种包膜蛋白共同组成HIV刺突结构，在HIV进入宿主细胞的过程中起到重要的作用。HIV-1基因组全长为9.7 kb，其基因组有9个开放阅读框，包含3个结构基因(*gag*, *pol*, *env*)和6个调控蛋白基因(*tat*, *rev*, *nef*, *vif*, *vpu*, *vpr*))。3个结构基因编码结构蛋白和酶，其中*gag*基因编码基质蛋白(MA, p17)、衣壳蛋白(CA, p24)及核衣壳蛋白(p6, p7)；*pol*基因编码逆转录酶(RT)包含核糖核酸酶H(RNase H)活性、蛋白酶(PR)和整合酶(IN)；*env*基因编码包膜糖蛋白Gp160，Gp160可以裂解为囊膜蛋白Gp120和穿膜蛋白Gp41；6个调控蛋白基因编码两个调节蛋白Tat及Rev与4个附属蛋白Nef、Vif、Vpr、Vpu [12]。

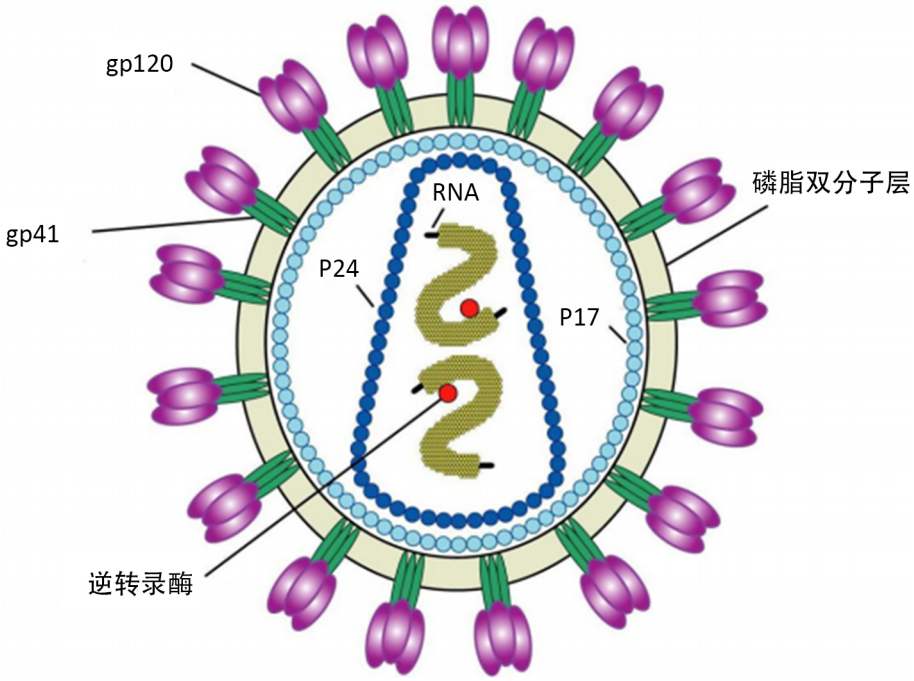

Figure 1. Structure of the HIV Virus [13]
图1. HIV病毒结构[13]

3.2. HIV 相关本体研究现状

中国医学科学院医学信息研究所建立的重大传染病知识服务平台[14]，以症状、治疗、传播途径、易感人群、世界艾滋病日、全球首个艾滋病病例及专家等为分类依据，构建了艾滋病本体[15]，对艾滋病相关的本体框架做了一个较好地分类梳理。

美国西北大学遗传医学中心同马里兰大学医学院合作，构建了人类疾病相关的疾病本体(Disease Ontology, DO) [16] [17]。DO 将各种疾病以传染病、细胞增殖性疾病、心理健康疾病、代谢疾病、基因疾病及综合征等进行了分类。以 HIV 为主题词进行检索，可以清楚判断出 HIV 在该本体中的上下等级关系(属于传染病中的病毒性传染病)，也可以看到有艾滋病的部分信息，并同维基百科中 HIV 的注释说明也进行了整合关联。

OBO 组织构建的蛋白本体 PRO，规范地整合了各类型蛋白质的表述，包括了人类、小鼠、大肠杆菌蛋白质及其关系：蛋白质家族关系、进化学关系及由基因变异、选择性剪接、溶蛋白裂解、翻译后修饰产生的不同蛋白质的关系。PRO 主要包括 ProEvo (基于进化亲缘关系的蛋白质)、ProForm (基于给定基因位点获得的各种蛋白质形式)、ProComp (各蛋白复合物)三个子本体，其基本框架如图 2 所示。PRO 涵盖了 HIV 蛋白的相关信息，以 HIV 的 GP160 蛋白为例进行检索(如图 3 所示)，在 PRO 中蛋白是按照基因水平、序列水平、修饰水平及家族亲缘性水平进行分类排列的。

4. 存在问题与展望

目前国内外还没有专门针对 HIV 蛋白的本体。疾病本体 DO 是从宏观上整体把握所有疾病的上下位

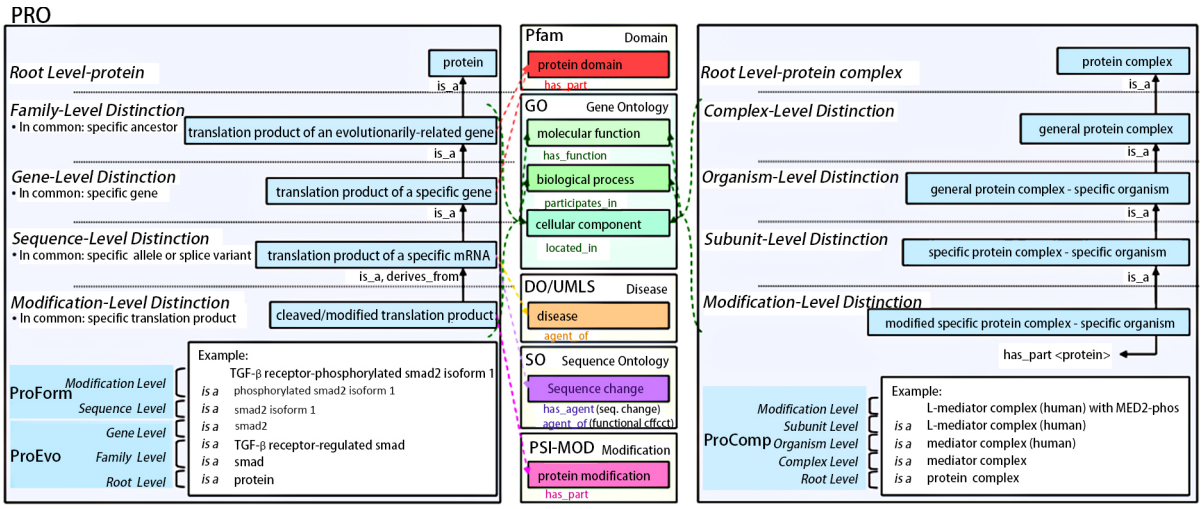

Figure 2. Framework of PRO
图 2. 蛋白本体(PRO)基本框架图

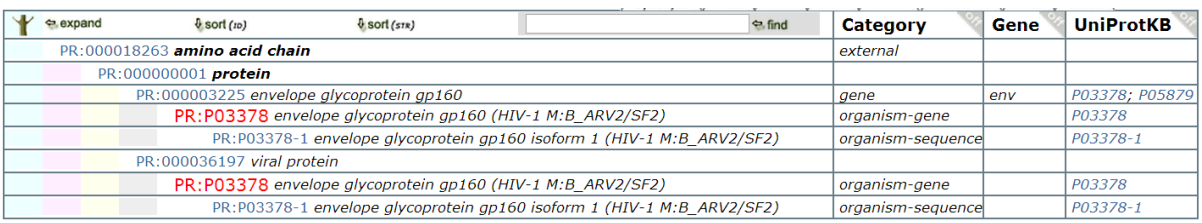

Figure 3. Example of GP160 protein retrieval
图 3. GP160 蛋白检索例

关系，没有深入到特定疾病及特定蛋白。蛋白本体 PRO 涵盖了 HIV 蛋白的相关信息，但只是对其基本信息进行简单罗列，从 HIV 相关的某个特定蛋白质属性出发形成一定的结构，却并没有从 HIV 自身出发形成更加专业清晰的知识组织结构，不能对 HIV 所有蛋白间的关系及结构有一个更好的清晰认识。构建一个 HIV 蛋白本体并应用于知识组织与管理的专题数据库中，可以对 HIV 所有蛋白及蛋白之间的关系与相互作用有一个更好的认识与理解，同时又可以减少 HIV 病毒信息中一词多义、同物异名现象的发生，提供一个更加准确及直观的检索结果。过去肝炎病毒领域蛋白本体[18] [19]的成功构建及其在基于语义检索中的成功应用，在一定程度上克服了传统检索方式所带来的信息冗余或信息丢失等问题，也为 HIV 蛋白本体的进一步研究和构建提供了一个较好的模板和参照。

基金项目

国家科技支撑计划项目课题(2013BAH21B06)和地方配套基金资助。

参考文献 (References)

[1] Studer, R., Benjamins, V.R. and Fensel, D. (1998) Knowledge Engineering: Principles and Methods. *Data & Knowledge Engineering*, **25**, 161-197.

[2] Du, X.-Y. (2006) A Survey on Ontology Learning Research. *Journal of Software*, **17**, 1837.

[3] 宋朋. 本体构建最新研究进展综述[J]. 中国科技资源导刊, 2015, 47(3): 73-83.

[4] 王向前, 张宝隆, 李慧宗. 本体研究综述[J]. 情报杂志, 2016, 35(6): 163-170.

[5] Bodenreider, O. and Stevens, R. (2006) Bio-Ontologies: Current Trends and Future Directions. *Briefings in Bioinformatics*, **7**, 256-274. https://doi.org/10.1093/bib/bbl027

[6] Malone, J., Stevens, R., Jupp S., Hancocks, T., Parkinson, H. and Brooksbank, C. (2016) Ten Simple Rules for Selecting a Bio-Ontology. *PLoS Computational Biology*, **12**, e1004743. https://doi.org/10.1371/journal.pcbi.1004743

[7] Zulkarnain, N.Z., Meziane, F. and Crofts, G. (2016) A Methodology for Biomedical Ontology Reuse. In: Metais, E., Meziane, F., Saraee, M., *et al.*, Eds., *Natural Language Processing and Information Systems*, *NLDB* 2016, *Lecture Notes in Computer Science*, Vol. 9612, Springer, Cham, 3-14. https://doi.org/10.1007/978-3-319-41754-7_1

[8] 吴明智, 崔雷. 生物医学相关的本体研究现状[J]. 医学信息学杂志, 2009, 30(7): 41-44.

[9] http://www.obofoundry.org/

[10] Poorolajal, J., Hooshmand, E., Mahjub, H., Esmailnasab, N. and Jenabi, E. (2016) Survival Rate of AIDS Disease and Mortality in HIV-Infected Patients: A Meta-Analysis. *Public Health*, **139**, 3-12. https://doi.org/10.1016/j.puhe.2016.05.004

[11] Gilbert, P.B., McKeague, I.W., Eisen, G., Mullins, C., Gueye-Ndiaye, A., Mboup, S. and Kanki, P.J. (2003) Comparison of HIV-1 and HIV-2 Infectivity from a Prospective Cohort Study in Senegal. *Statistics in Medicine*, **22**, 573-593. https://doi.org/10.1002/sim.1342

[12] Woodman, Z. and Williamson, C. (2009) HIV Molecular Epidemiology: Transmission and Adaptation to Human Populations. *Current Opinion in HIV and AIDS*, **4**, 247-252. https://doi.org/10.1097/COH.0b013e32832c0672

[13] 孙德福. 艾滋病毒蛋白酶异位抑制剂体系的分子动力学研究[D]: [硕士学位论文]. 济南: 山东师范大学, 2012.

[14] 高东平, 方安, 李杨, 孙晓北, 刘浩, 池慧. 知识服务平台的设计与应用——以重大传染病信息知识服务平台为例[J]. 情报理论与实践, 2011, 34(7): 111-115.

[15] 方安, 洪娜, 高东平, 李亚子, 池慧. 传染病本体构建及其在知识服务平台中的应用[J]. 现代图书情报技术, 2012, 28(1): 7-12.

[16] Schriml, L.M., Arze, C., Nadendla, S., Chang, Y.W.W., Mazaitis, M., Felix, V., Feng, G. and Kibbe, W.A. (2012) Disease Ontology: A Backbone for Disease Semantic Integration. *Nucleic Acids Research*, **40**, D940-D946. https://doi.org/10.1093/nar/gkr972

[17] http://disease-ontology.org/

[18] 魏晓萍. 肝炎病毒蛋白领域本体的构建及应用研究[D]: [硕士学位论文]. 上海: 上海交通大学, 2013.

[19] 张珅, 张永娟, 金毅, 陈恒. 生命科学领域本体研究及应用初探[J]. 情报杂志, 2013(32): 32-38.