

以细胞系为中心的基因 - 突变 - 疾病语义网络构建研究 *

吴 萌 李 姣 侯 丽

(中国医学科学院医学信息研究所 北京 100020)

〔摘要〕 阐述语义模型在生物医学领域的研究和应用现状, 提出构建一个以细胞系为中心, 包含基因、突变与疾病间语义关系的语义网络, 详细介绍语义网络的设计流程、数据处理方法与步骤, 比较、总结、验证所完成的语义网络与其他相关数据库的全面性、完整性与正确性。

〔关键词〕 知识图谱; 语义网; 链接数据

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2019.07.012

Study of the Building of Semantic Network for Connecting Genes, Mutations and Diseases Centered by Cell Line WU Meng, LI Jiao, HOU Li, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China

〔Abstract〕 The paper expounds the study and application status quo of semantic model in biomedicine, proffers the building of a semantic network containing the semantic relation of genes, mutations and diseases with cell line as its core, expounds the semantic network's design, process, data processing methods and procedures. It also compares, summarizes and verifies the comprehensiveness, integrity and correctness of this semantic network and other related databases.

〔Keywords〕 knowledge graph; semantic web; linked data

1 引言

语义网广义上来说是对未来网络的一个设想,

〔修回日期〕 2019-03-07

〔作者简介〕 吴萌, 硕士, 实习研究员; 通讯作者: 侯丽。

〔基金项目〕 国家重点研发计划“精准医学本体和语义网络构建”(项目编号: 2016YFC0901901); 互联网医疗系统与应用国家工程实验室项目“面向跨院的电子病历数据融合关键技术与标准构建研究”(项目编号: NELIMSA2018 P02); 国家人口与健康科学数据共享服务平台国际合作项目(项目编号: NCMI-MCE01-201706)。

狭义上来说是一种智能网络, 不但能够理解词语和概念, 而且还能够理解它们之间的逻辑关系^[1]。随着语义网概念的提出, 互联网逐步从仅包含网页之间超链接的文档万维网转变为描述各种实体之间关系的数据万维网。基于此, 知识图谱这一概念于 2012 年 5 月由谷歌率先提出, 其目标在于描述真实世界中存在的各种实体和概念, 及实体与概念之间的关联关系, 从而改善搜索结果。同时, 资源描述框架(Resource Description Framework, RDF)及其模式(Resource Description Framework Schema, RDFS)在语义 Web 中处于核心地位, 是实现 Web 信息共享和数据交换的基础。

伴随生物医学领域测序技术的飞速发展和精准医学概念的提出, 越来越多的科学研究开始关注于

疾病发生的内在复杂机制,以及各个生物学实体之间的网络调控通路和关联关系,以提供个性化的治疗方案。知识图谱等语义网络技术为多源异构的生物学数据的整合和复杂关系网络的建模提供新的解决方案,通过利用统一的数据表示标准,为生物学数据的检索、分析、挖掘提供基础。在癌症生物学的研究过程中,人类癌症细胞系作为一种易于获取、方便使用的生物模型,广泛应用于探索癌症的分子特征以及相应的治疗反应。由于临床试验复杂且昂贵,而借助细胞系进行临床前实验有助于极大地提高临床实验的成功率。目前,许多项目都致力于细胞系及其遗传学和基因组学数据提供系统的整合方案,例如癌症体细胞突变目录(Catalogue of Somatic Mutation In Cancer, COSMIC),使用户在进行生物实验和药物测试时可以选择更合适的癌症细胞系,也为生物研究提供临床依据^[2]。本研究从 NCBI gene、ClinVar、COSMIC、Cellosaurus、OMIM 与 NCI 6 个数据库中分别获取基因、突变、细胞系与疾病及其间的语义关系数据,拟以细胞系数据为核心,构建一个包含基因、突变与疾病数据及其语义关系的 RDF 语义网络。旨在对疾病基因组学等相关领域中的生物学数据进行建模与整合,以期为进一步发现新的医学实体语义关系,理解与分析疾病的致病机制提供数据支撑。

2 研究现状

2.1 语义模型在医学数据领域的应用

2.1.1 生物学语义模型 随着 2015 年美国总统奥巴马提出精准医学计划,世界多个国家陆续开始部署精准医学项目,更多的科学研究开始关注于疾病发生过程中内在的分子机制,而所催生的大量多来源异构的生物学数据,迫切需要统一的数据整合方案。语义模型技术为多来源异构的生物学数据整合提供方案,并致力于提供一套统一的生物学实体表示标准,使机器和人都可以理解,其灵活性、可扩展性以及可对语义关系进行模型等特点非常适用于表示复杂的生物学网路数据。如上海曙光医院构建的中医药知识图谱、医学系统命名法 -

临床术语(Systematized Nomenclature of Medicine - Clinical Terms, SNOMED - CT)和 IBM Watson Health 等系统^[3]。基于链接数据,也可以识别出新的语义关系。如 Dalleau 等对药物、疾病和基因相关的 6 个数据库进行整合与链接,构建药物基因组学相关的 RDF 格式语义网络,共包含 2 640 793 个 3 元组。基于构建的链接数据,分别用两种基于图的机器学习的方法——随机森林和图核,对药物与基因是否相关进行预测,从而发现新的药物 - 疾病关系^[4]。

2.1.2 细胞系相关语义模型 细胞系目前已在许多生物学实验和研究中被广泛使用。复杂疾病,如癌症的发生通常开始于一系列体细胞 DNA 变化所导致的失控的细胞增殖,这些大部分变化指的是突变等特定的 DNA 序列变化。研究认为细胞从正常状态转变到完全的恶性形态的过程,必须积累 5 ~ 10 个体细胞突变,每一种突变都会引发不一样的细胞功能改变^[5]。对癌症细胞系进行基因组测序,可以发现引发细胞机制发生变化的重要突变,整合并分析肿瘤发生过程中细胞系发生的突变信息,有利于更好地理解肿瘤发生的内在分子机制,从而发现新的治疗方案。而现有医学数据的语义模型中,整合细胞系、突变和疾病等相关实体及其语义关系的语义模型还非常少见。大多数模型只联系突变和疾病,或细胞系和疾病等实体。如细胞系本体(The Cell Line Ontology, CLO),主要描述细胞系、癌症、细胞和有机体之间的关联^[6]。COSMIC 数据库整合细胞系、突变和疾病等实体之间的关联,但是对于疾病的描述没有采用通用的表示方法,不利于与其他疾病数据库进行映射,其表示方式没有采用语义模型。

2.2 关系数据向语义数据的转化

随着下一代测序技术的发展,越来越多丰富的组学数据被生产、注释出来。其大多数存储形式为关系型数据库或类似的表格文件形式。而如何将关系型数据的数据资源和语义关系信息转化为 RDF 语义网络格式,成为一种普遍需求。许多项目如 Bio2RDF^[7], the EBI platform^[8], PDBj^[9] 以及

Linked Open Drug Data (LODD)^[10]等都致力于推动健康科学数据转化为统一的链接数据形式。其中, Bio2RDF 是一个开源的项目, 采用语义网技术构建并提供生命科学数据的链接数据网络。Bio2RDF 定义一套简单的规则, 为多来源异构的数据集合创建 RDF (S) 兼容的链接数据形式。目前已为 clinical-trials. gov, dbSNP, GenAge 等 35 个数据库提供 RDF 链接数据结构^[7]。R2RML^[11]是 W3C RDB2RDF 工作组于 2012 年 9 月发布的一种映射语言, 可以定义关系型数据库与 RDF 格式数据之间的映射规则, 从而将关系型数据转化为 RDF 3 元组形式。基于 R2RML 的常用工具如 D2RQ、db2triples、OpenLink Virtuoso 等都可以实现关系型数据库向 RDF 的转化。

由此可见, 在生物医学领域, 构建不同实体间的语义模型对于数据的整合、复杂网络的表示及发现新的语义关系等都具有突出作用。而目前的研究多集中在药物、疾病、症状等传统医学的关系层面, 围绕基因、突变、细胞系等精准医学领域实体的语义网络尚不多见。上述关系数据转化技术中, 对比发现 D2RQ 有明显的优势, 它支持任何关系数据库的数据转换、通用性强, 支持灵活的映射配置文件, 提供一种标准的转换方式生成一个虚拟的 RDF (S), 确保数据库的内容更新便捷^[12]。鉴于此, 本文采用 D2RQ 工具将数据资源转化为 RDF 格式。

3 研究思路与框架

3.1 语义模型

基因、突变、细胞系与疾病之间存在多种语义关系。对语义关系进行有效以及规范的定义, 是语义模型构建的基础, 也为后续语义网络在文本挖掘等领域中的应用提供潜能。Verspoor 等人提出一种人类变异组信息注释模式, 对 11 种实体类型和关系进行规范。将这种模式应用于一个小型的以肠癌主题的全文预料库中, 通过使用这个模式进行全文

注释, 注释结果的一致性得到显著提高^[13]。本研究参考这种模式以及多种数据库对语义关系的命名方式, 最终确定 4 种语义类型 Gene、Mutation、Cell-line、Disease 之间的 6 种语义关系的定义模式。根据收集的数据库中的数据形式与数据内容, 选择部分基本信息进行提取, 例如基因的类型、突变的位点和疾病的别名等。这样, 每种语义类型都有多种相关属性进行更为全面的描述。具体语义关系模型设计, 见图 1。

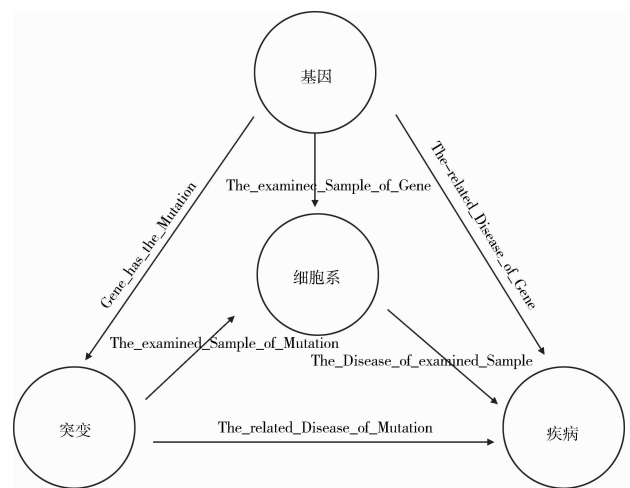


图 1 语义关系模型及数据来源

3.2 实验流程

根据所设计的语义模型, 选取 NCBI gene、ClinVar、COSMIC、Cellosaurus、OMIM 与 NCI 等 6 个数据库作为基因、突变、细胞系与疾病等实体数据的来源。通过数据筛选、格式转化及融合等数据预处理流程, 形成待处理数据集, 将数据集存入 MySQL 数据库中, 依据数据特性共存储为 4 个实体表, 以及 6 个关系表。其后, 利用 D2RQ 映射工具, 根据本研究设计的语义模型, 定义关系型数据与 RDF 数据的映射规则。最后, 将关系型数据库转化为 RDF 语义格式并利用 D2RQ 工具部署本地 Web 应用, 实现对语义网络的生成、检索与分析。实验流程, 见图 2。

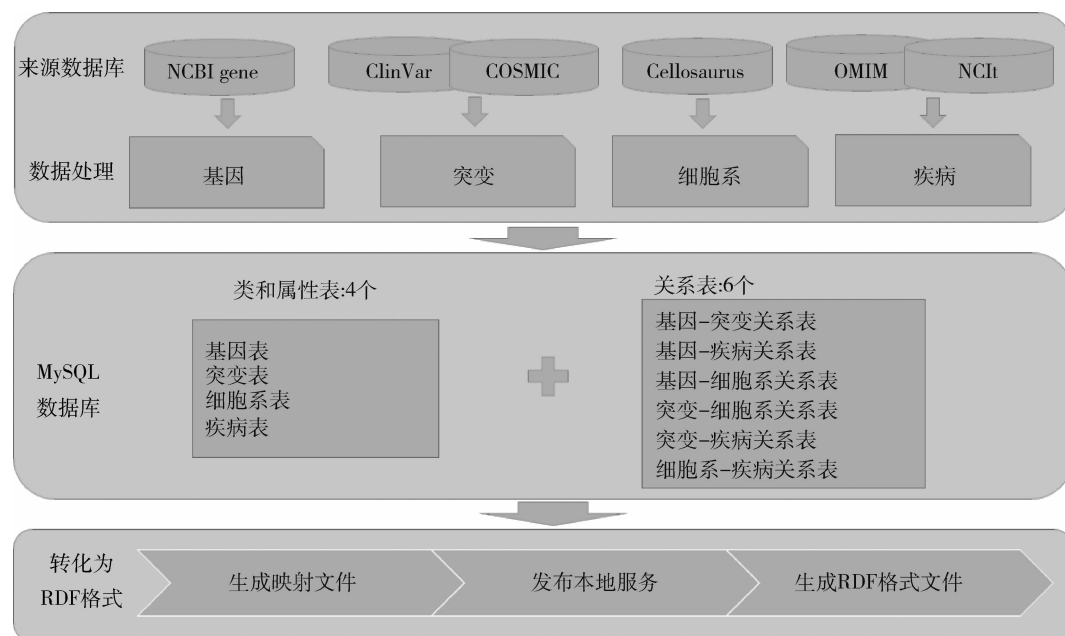


图 2 语义关系模型的构建

4 实验过程与结果

4.1 数据处理

4.1.1 数据提取 (1) 基因数据选自 NCBI gene 数据库。NCBI gene 数据库提供了多物种的基因序列信息, 包括序列、表达、结构、功能以及引用等信息, 基因的唯一标识——Entrez_ ID, 在美国国家生物技术信息中心 (National Center for Biotechnology Information, NCBI) 研发的所有数据库中都是通用的^[14]。本研究选择 NCBI gene 人类物种的基因数据作为实验数据的基因部分, 共 60 195 条, 包括基因 ID、基因名称、在染色体上的位置和基因类型等信息。以 NCBI gene 数据库的 Entrez_ ID 为基因数据的标准标识。(2) 突变数据与突变-基因、突变-疾病之间的关系数据选自 ClinVar 数据库。ClinVar 是一个公开的数据库, 其中收集与疾病相关的人类遗传变异^[15]。本研究选择突变概要文件 variant_summary.txt 中基因组参考序列版本为 GRCh38 的突变数据, 共 254 030 条, 筛选所在基因、突变名称、突变类型、所在染色体、相关疾病等信息。ClinVar 中突变所在的基因使用 NCBI gene 的 Entrez_ ID 进行标识, 可与 NCBI gene 进行链接。突变相关

的疾病整合了 OMIM 的疾病 ID 标识, 所以也可与 OMIM 中的疾病链接。ClinVar 的突变标识以 RCV000000000.0 形式表示, 在突变对应多个疾病的情况下, 一个突变会对应多个突变标识所以本研究自定义突变的唯一标识, 并保留 RCVaccession 的信息。(3) 疾病数据与疾病-基因之间的关系数据选自 OMIM 数据库。在线人类孟德尔遗传数据库 (Online Mendelian Inheritance in Man, OMIM) 是一个关于人类基因和表型的权威数据库, 目前每日更新并支持免费获取^[16]。本研究选择 morbidmap.txt 文件中 7 326 条疾病数据, 保留疾病名称、基因名称、基因的 OMIM ID 以及基因的位置信息。利用 mim2gene.txt 中 OMIM 基因与 NCBI 基因的映射, 获取疾病-基因之间的关系。由于有些疾病的 OMIM ID 缺失, 所以自定义疾病的唯一标识。(4) 细胞系的数据选自 Cellosaurus 数据库。Cellosaurus 数据库由瑞士生物信息研究所的团队在 Biocuration 2016 大会上提出^[17]。是目前整合细胞系信息较为全面的数据库。下载并处理 68 406 条细胞系数据为标准格式, 筛选细胞系名称、编码、别名、相关疾病等信息, 自定义细胞系的唯一标识。

4.1.2 数据关联 基于以上数据处理的步骤, 可以得到大部分实验数据。但疾病-细胞系、突变-

细胞系、基因-细胞系的关系是无法直接从已整合的数据库获取的,需要引入中间数据库,才能将这些实体进行关联。Cellosaurus 中存在疾病-细胞系的关系,但是 Cellosaurus 中的疾病数据使用 NCIt^[18]的疾病术语进行表示,提供其在 NCIt 中 ID 编码。所以利用 NCIt 将 Cellosaurus 细胞系与 OMIM 疾病进行关联。参考一体化医学语言系统(Unified Medical Language System, UMLS)中已对 NCIt 的疾病术语和 OMIM 的疾病术语进行整合的信息。利用两个来源的术语在 UMLS 数据中是否在同一个概念下进行同义判断,将 NCIt 与 OMIM 的疾病术语进行映射。对于突变-细胞系与基因-细胞系的关系,利用 COSMIC 数据库作为中间数据库来获取这些信息。COSMIC 是世界上最大最全的研究人类体细胞突变对癌症影响的数据资源^[2],其中包含细胞系和原代细胞的基因测序信息和识别的突变信息。本研究重点关注对细胞系与突变和基因的关联信息的获取。获取突变数据与细胞系的关系,先将 COSMIC 的细胞系与 Cellosaurus 的细胞系的名称以及别名进行匹配,再根据突变在染色体中的位置,以及突变的类型,将 COSMIC 中的突变信息与 ClinVar 中的突变信息进行映射,以获得 ClinVar 突变与 Cellosaurus 细胞系之间的关联。而 COSMIC 中存在 NCBI 基因与细胞系关系,利用之前 COSMIC 的细胞系与 Cellosaurus 的细胞系的映射,可获得 NCBI 基因与 Cellosaurus 细胞系之间的关联。

4.2 数据映射

4.2.1 数据存储 本研究采用 D2RQ 工具将融合的数据资源转化为 RDF 格式。D2RQ 是一个开源的平台,提供以虚拟只读的 RDF 数据形式访问关系数据库的功能。通过 D2RQ,可以使用 SPARQL 语言对关系型数据进行检索,通过其自带 Web 应用浏览数据,也可以使用工具获取生成 RDF 格式的文件。根据 D2RQ 工具所需要的输入文件格式,将基因、突变、细胞系、疾病 4 种实体类型以及相关属性存为 MySQL 数据库中相应的 4 张表,6 种实体间的关系存为数据库中相应的 6 张表。

4.2.2 映射规则 D2RQ 提供映射语言来描述关系型数据库模式向 RDFS 转换的映射规则。一个 D2RQ 映射文件本身就是用 Turtle 语法编写的 RDF 文档。D2RQ 提供一些便利的工具,例如 generate-mapping 工具可以自动生成一个映射文件 mapping.ttl。但是自动生成的映射文件只包含一些基础的规则设置,更多复杂的映射规则可以参考 D2RQ 的映射语言^[19]。

本研究首先通过声明一个数据库为 d2rq: Database 类来设置数据库的相关属性,包括 JDBC 数据库的 URL, JDBC 驱动程序类名,数据库用户名密码等。声明一个相应的实体为 d2rq: ClassMap 类来设置类的属性,包括 URI 的模式,例如设置 gene 的 URI 模式为“gene/@ @ gene. Gene_ ID | urlify @ @”,其中 gene. Gene_ ID 为数据库中 gene 表的 Gene_ ID 列。利用 d2rq: PropertyBridge 来设置属性,例如属性名称、属性值。实体之间的语义关系,利用其中的 d2rq: refersToClassMap 来定义,例如基因-突变之间的语义关系“Gene_ has_ the_ Mutation”,将两个实体进行链接。

4.3 实验结果

4.3.1 实验结果浏览与获取 本研究根据基因、突变、细胞系、疾病 4 种实体类型以及之间的 6 种语义关系对应的表格,编写 D2RQ 映射语言文件 mapping.ttl。利用生成的文件,运行 d2r-server 工具,启动 D2RQ 部署的本地服务 http://localhost:2020/。通过这个 Web 应用,可以对数据进行浏览,也可以利用 SPARQL 执行搜索并设定展示的条目数量。以 AKT3 基因为例,数据浏览结果,见图 3。基因 AKT3 的信息页面包含多种信息。其中,该基因相关的突变,用字段“Gene_ has_ the_ mutation”表示;检测该基因的细胞系样本,用字段“The_ examined_ Sample_ of_ Gene”表示,其值对应的链接,可以链接到相应的突变和细胞系的信息页面。其他属性,例如,基因所在的染色体、基因全名和基因 ID 等信息也在该页面中详细列出。

Property	Value
vocab:Gene_has_the_mutation	<http://localhost:2020/resource/mutation/201739>
vocab:Gene_has_the_mutation	<http://localhost:2020/resource/mutation/201745>
vocab:Gene_has_the_mutation	<http://localhost:2020/resource/mutation/201829>
vocab:Gene_has_the_mutation	<http://localhost:2020/resource/mutation/201836>
vocab:Gene_has_the_mutation	<http://localhost:2020/resource/mutation/201935>
vocab:Gene_has_the_mutation	<http://localhost:2020/resource/mutation/201936>
vocab:Gene_has_the_mutation	<http://localhost:2020/resource/mutation/201941>
vocab:Gene_has_the_mutation	<http://localhost:2020/resource/mutation/232133>
vocab:Gene_has_the_mutation	<http://localhost:2020/resource/mutation/239746>
vocab:Gene_has_the_mutation	<http://localhost:2020/resource/mutation/242595>
vocab:Gene_has_the_mutation	<http://localhost:2020/resource/mutation/52602>
vocab:The_examined_Sample_of_Gene	<http://localhost:2020/resource/cell/17836>
vocab:gene_Chromosome	1
vocab:gene_Full_name_from_nomenclature_authority	AKT serine/threonine kinase 3
vocab:gene_Gene_ID	10000
vocab:gene_Symbol	AKT3
vocab:gene_Symbol_from_nomenclature_authority	AKT3
vocab:gene_dbXrefs	MIM:611223 HGNC:HGNC:393 Ensembl:ENSG00000117020 Vega:OTTHUMG00000039994
vocab:gene_map_location	1q43-q44
vocab:gene_type_of_gene	protein-coding
rdfs:label	gene #10000
rdf:type	vocab:gene

图3 AKT3 基因信息浏览页面

此外，D2RQ 也提供通过命令行进行 SPARQL 语言搜索的功能。通过 dump - rdf 命令可以将生成的 RDF 文件导出生成“TURTLE”、“RDF/XML”、“RDF/XML - ABBREV”、“N3”以及“N - TRIPLE”等多种 RDF 语法格式。实验最终构建的语义网络，共包含基因60 195个，突变 254 030 个，细胞系 68 406 个，疾病 7 326 个，共构建 3 元组 726 236 个。其中，基因与突变之间的 3 元组 254 030 个，基因与疾病之间的 3 元组 15 477 个，基因与细胞系之间的 3 元组 287 342 个，细胞系与突变之间的 3 元组 195 个，细胞系与疾病之间的 3 元组 36 377 个，突变与疾病之间的 3 元组 132 815 个。

4.3.2 实验结果对比分析 对 NCBI Gene，ClinVAR，Cellosaurus，COSMIC 等几个数据库进行分析，对数据类型和数据格式进行调研，与本研究模型进行对比，发现本研究构建的语义关系模型，其

细胞系相关的数据类型覆盖程度更为全面，方便用户对细胞系及相关信息进行浏览和查询，而不用跨越多个数据库，缘于已有效地将细胞系相关的数据进行整合。每种数据类型都采用业内通用的数据库的名称和标识，提供良好的互操作性，为细胞系的研究提供帮助。相比于 NCBI Gene，本研究构建的语义网络还整合了突变和细胞系的信息，较 ClinVar 多细胞系信息，比 Cellosaurus 数据库的信息主要多基因和突变的数据，而在数据类型覆盖层面，COSMIC 数据库同样覆盖了基因、突变、细胞系和疾病的信息，但同时本语义网络在疾病数据的整合方面提供疾病术语命名和编码的标准，见表 1。当然该语义网络中的许多数据是从其中几个数据库中提取并整合的，因此整合更多细胞系相关的新数据、发现更多的语义关系也是本语义网络应该继续努力的方向。

表 1 本研究模型数据类型覆盖范围与其他几个相关数据库对比

数据类型	NCBI Gene	ClinVar	Cellosaurus	COSMIC	本研究模型
Gene	√	√	-	√	√
Mutation	-	√	-	√	√
Cell - line	-	-	√	√	√
Disease	√	√	√	√*	√

注：* 即 COSMIC 没有提供表型标准名称以及与其他数据库的映射

本研究通过 D2RQ 部署的语义网络的 Web 应用,目前仅限于内部实验使用,后期将持续完善细胞系语义网络浏览平台,发布为外网可访问的形式并提供多种语义网络获取方式。

5 结语

本研究利用基于 RDB 到 RDF 映射语言 (R2RML) 的 D2RQ 映射工具,对以细胞系数据为中心的包含基因、突变与疾病数据以及语义关系的数据进行建模与整合,分析不同来源生物医学数据的特性,发现数据融合的有效方法,所构建的语义网络数据类型覆盖更为全面,可为用户提供更加便捷的服务。但是仍有许多不足之处有待改进,今后的工作将从以下内容展开:进一步优化多来源异构数据的融合方法,多方面考虑数据的不同特性,利用实体相似计算方法提高数据的映射成功率。研究突变对于癌症等复杂疾病带来的内部细胞机制的变化,丰富实体之间的语义关系。最终将数据来源扩展至文献、电子病历等形式,从中挖掘更多的生物医学实体之间的语义关系,从而对语义网络进行扩展,提高语义模型的实用性。

参考文献

- 1 百度百科. 语义网 [EB/OL]. [2018 - 10 - 31]. <https://baike.baidu.com/item/语义网/118508?fr=aladdin>.
- 2 Forbes S A, Beare D, Boutselakis H, et al. COSMIC: Somatic Cancer Genetics at High - resolution [J]. *Nucleic acids research*, 2016, 45 (D1): D777 - D783.
- 3 Williams AJ, Harland L, Groth P, et al. Open PHACTS: semantic interoperability for drug discovery [J]. *Drug discovery today*, 2012, 17 (21): 1188 - 1198.
- 4 Dalleau K, Marzougui Y, Da Silva S, et al. Learning from Biomedical Linked Data to Suggest Valid Pharmacogenes [J]. *Journal of Biomedical Semantics*, 2017, 8 (1): 16.
- 5 Hanahan D, Weinberg RA. Hallmarks of Cancer: the next generation [J]. *Cell*, 2011, 144 (5): 646 - 674.
- 6 Sarnitviji S, Lin Y, Xiang Z, et al. CLO: the cell line ontology [J]. *Journal of Biomedical Semantics*, 2014, 5 (1): 37.
- 7 Belleau F, Nolin MA, Tourigny N, et al. Bio2RDF: towards a Mashup to Build Bioinformatics Knowledge Systems [J]. *Journal of Biomedical Informatics*, 2008, 41 (5): 706 - 716.
- 8 EBI. RDF platform [EB/OL]. [2018 - 10 - 31]. <https://www.ebi.ac.uk/rdf/>.
- 9 Kinjo AR, Bekker GJ, Suzuki H, et al. Protein Data Bank Japan (PDBj): Updated User Interfaces, Resource Description Framework, Analysis Tools for Large Structures [J]. *Nucleic Acids Research*, 2017, 45 (Database issue): D282 - D288.
- 10 W3C. HCLSIG/LODD [EB/OL]. [2018 - 10 - 31]. <https://www.w3.org/wiki/HCLSIG/LODD>.
- 11 Souripriya Das. R2RML: RDB to RDF Mapping Language [EB/OL]. [2018 - 09 - 17]. <https://www.w3.org/TR/r2rml/>.
- 12 杨王黎, 郑雪芸, 袁满. 关系模型向 RDF (S) 模型转换研究 [J]. *微型电脑应用*, 2017, 33 (9): 3 - 8.
- 13 Verspoor K, Jimeno Y A, Cavedon L, et al. Annotating the Biomedical Literature for the Human Variome [EB/OL]. [2018 - 10 - 31]. <https://academic.oup.com/database/article/doi/10.1093/database/bat019/331953#>.
- 14 Murphy M, Brown G, Wallin C, et al. Gene Help: Integrated Access to Genes of Genomes in the Reference Sequence Collection [EB/OL]. [2018 - 10 - 31]. <https://www.ncbi.nlm.nih.gov/books/NBK3841/>.
- 15 Landrum MJ, Lee JM, Benson M, et al. ClinVar: public archive of interpretations of clinically relevant variants [J]. *Nucleic Acids Research*, 2015, 44 (D1): D862 - D868.
- 16 Amberger JS, Bocchini CA, Schiettecatte F, et al. OMIM. Org: Online Mendelian Inheritance in Man (OMIM®), an Online Catalog of Human Genes and Genetic Disorders [J]. *Nucleic Acids Research*, 2014, 43 (D1): D789 - D798.
- 17 Bairoch A. The Cellosaurus, a Cell - Line Knowledge Resource [J]. *Journal of Biomolecular Techniques*, 2018, 29 (2): 25 - 38.
- 18 Golbeck J, Frago G, Hartel F, et al. The National Cancer Institute's Thesaurus and Ontology [J]. *Journal of Web Semantics*, 2003 (1): 75 - 80.
- 19 Cyganiak R, Bizer C, Garbers J. The D2RQ Mapping Language [EB/OL]. [2018 - 10 - 31]. <http://d2rq.org/d2rq-language>.