Bridging Vaccine Ontology and NCIt vaccine domain for cancer vaccine data integration and analysis

Yongqun He[1], Guoqian Jiang[2]
[1] University of Michigan Medical School, Ann Arbor, MI 48109, USA;
[2] Mayo Clinic, Rochester, MN, 55906, USA

ABSTRACT

The Vaccine Ontology (VO) is a community-based ontology in the domain of vaccines and vaccination. VO is aligned with the Basic Formal Ontology (BFO) and developed by following OBO Foundry principles. National Cancer Institute (NCI) Thesaurus (NCIt) serves as a reference ontology to facilitate interoperability and data sharing for cancer translational and basic research. To facilitate better cancer vaccine research, we compared the VO and NCIt vaccine domain (NCIt-vaccine) and examined the possibility of bridging and integrating these two ontologies. Our results showed that only a small portion of vaccine terms overlap between the two ontologies, and VO and NCIt-vaccine are complementary in different aspects. It is possible to integrate, map, and merge them. This study can be used as a use case for achieving the broader goal of merging and integrating NCIt and OBO library ontologies.

1 INTRODUCTION

Cancer clinical and biology research studies have generated large volumes of data. Barriers to data normalization, standardization, and quality assurance make it difficult to annotate and integrate cancer data in meaningful ways and hence delay widespread research data reuse within the broader scientific community. In cancer vaccine study domain, for example, there is an urgent need to develop an integrated data and knowledge repository that can facilitate translational research studies in developing treatment vaccines against many types of cancer. To this end, ontology-based data integration approaches have been increasingly used to address this challenge (Mate et al., 2015).

Notably, NCI has developed NCIt that serves as a reference ontology to facilitate interoperability and data sharing for cancer translational and basic research (de Coronado et al., 2004). NCI has been exploring new approaches to broaden external participation in the ontology development and quality assurance process, including introducing a solid upper-level ontology. NCIt includes a vaccine branch (NCIt-vaccine) that covers many different cancer-related vaccines. Concurrently, the Open Biological and Biomedical Ontologies (OBO) Foundry, as a collaborative initiative, has aimed at establishing a set of ontology development principles and incorporating ontologies following these principles in an evolving non-redundant and interoperable suite (Smith et al., 2007). The OBO library currently includes >160 ontologies covering >3 million terms. The OBO ontologies related to clinical and biological vaccine studies include the Vaccine Ontology (VO) (He et al., 2009;Ozgur et al., 2011).

The importance of merging and integrating NCIt and OBO library ontologies has been well recognized (de Coronado et al., 2007). Here we compared VO and NCIt-vaccine with the aim to possibly align and merge these two ontologies together. Our results show both promise and challenges.

* To whom correspondence should be addressed: yongqunh@umich.edu and Jiang.Guoqian@mayo.edu

2 METHODS

2.1. Vaccine module extraction and ontology metrics comparison

The current versions (as of April 24, 2015) of VO and NCIt (version 15.03e) were obtained from their download websites. We used an OWL-based ontology module extraction tool (https://sites.google.com/site/ontologymodularity/) and extracted the vaccine module from each respectively, anchored by the VO code "vaccine (VO_0000001)" and the NCIt code "Vaccine (C923)" and their subclasses. We compared the ontology metrics of the two vaccine modules using the Protégé 5 Ontology Metrics plugin.

2.2. Ontology alignment and coverage analysis

We first manually aligned direct subclasses of the vaccine codes in two modules, and then used a UMLS-based lexical mapping tool called the Sub-Term Mapping Tools (STMT) (Lu and Browne, 2012) to retrieve the UMLS CUIs for all subclasses of the vaccine code VO_0000001 in VO. As each NCIt code has already had a corresponding UMLS CUI asserted, we produced the mappings of vaccine terms between these two ontologies. The content coverage for the vaccine terms (matched and unmatched) between the two ontologies was analyzed.

3 RESULTS

3.1 Ontology module extraction and metrics comparison

VO currently covers 4,751 terms including ~800 terms imported from other existing ontologies (http://www.ontobee.org/ontostat.php?ontology=VO). If we only count the classes under VO:vaccine (VO_0000001),

the VO vaccine branch has 28 direct subclasses and 2,140 descendants. In comparison, the NCIt-vaccine section has 11 direct subclasses and 703 descendants (Table 1).

Ontology Metrics	VO Vaccine	NCIt Vaccine
Axiom	26704	13954
Logical axiom count	8151	1570
Class count	3047	874
Class axioms		
SubClassOf axioms count	7776	1513
EquivalentClasses axioms	144	5
DisjointClasses axioms count	8	15
Object property count	82	18
DL expressivity	SROIQ	S

Table 1. Comparison of VO and NCIt-vaccine ontology metrics

3.2. VO-NCIt vaccine domain ontology alignment

Our analysis found that 10 of 11 NCIt high-level vaccine codes had exact matches with VO vaccine codes. VO has an additional 17 subclasses (e.g., 'allergy vaccine' and 'prime-boost vaccine') that do not have any NCIt match.

For the lexical mappings, in total, 280 matches were identified for VO vaccine terms with UMLS CUIs. These may serve as bridging points between VO and NCIT. NCIt is more focused on cancer vaccines. VO is more focused on infectious disease vaccines. In addition to various vaccines, VO also represents various vaccine components such as vaccine antigens, adjuvants, DNA vaccine plasmids, etc. These can be used to logically represent specific vaccines.

3.3. Bridging VO and NCIt-vaccine

NCIt-vaccine includes many cancer vaccines not included in VO. Unlike VO vaccines, these cancer vaccines are not fully represented. Therefore, it is possible to apply VO representation methods to logically represent NCIt-vaccine. For example, we modeled the NCIt-vaccine 'Alpha Fetoprotein Plasmid DNA Vaccine' (NCIt: C48373; synonym: phAFP) (Fig. 1). This vaccine consists of a plasmid DNA encoding alpha fetoprotein. After vaccination, expressed alpha fetoprotein may stimulate a cytotoxic T lymphocyte (CTL) response against tumor cells that express alpha fetoprotein, resulting in tumor cell lysis (Hanke et al., 2002).

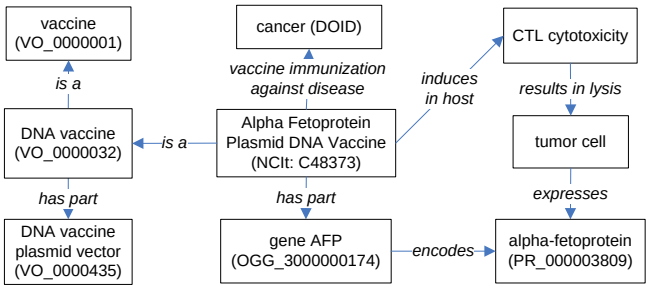

Fig. 1. Modeling of a NCIt cancer vaccine using VO approach

4 DISCUSSION

Although the topics of these ontologies are also covered by NCIt and its associated ontologies, NCIt, in general, lacks of granular terms that could be complemented by OBO ontologies that cover more terms in the biological domain. Our pilot study demonstrated that differences of ontology metrics of the vaccine modules extracted from the two ontologies, in terms of axiom richness and DL expressivity. While only a small portion of vaccine terms overlap between the two ontologies, both ontologies are complementary to each other in different ways. For better cancer vaccine study data integration, it is possible to align, map, and possibly merge these two vaccine modules. The merged ontology could be used to annotate the data and metadata available in various cancer vaccine resources, such as the CanVaxKB knowledgebase (http://www.violinet.org/canvaxkb/).

ACKNOWLEDGEMENTS

This research was supported in part by a bridge fund to Y.H. in the University of Michigan and a NCI U01 caCDE-QA grant (1U01CA180940-01A1).

REFERENCES

De Coronado, S., Haber, M.W., Sioutos, N., Tuttle, M.S., and Wright, L.W. (2004). NCI Thesaurus: using science-based terminology to integrate cancer research results. *Stud Health Technol Inform* 107**,** 33-37.

De Coronado, S., Tuttle, M.S., and Solbrig, H.R. (2007). Using the UMLS Semantic Network to validate NCI Thesaurus structure and analyze its alignment with the OBO relations ontology. *AMIA Annu Symp Proc***,** 165-170.

Hanke, P., Serwe, M., Dombrowski, F., Sauerbruch, T., and Caselmann, W.H. (2002). DNA vaccination with AFP-encoding plasmid DNA prevents growth of subcutaneous AFP-expressing tumors and does not interfere with liver regeneration in mice. *Cancer Gene Ther* 9**,** 346-355.

He, Y., Cowell, L., Diehl, A.D., Mobley, H.L., Peters, B., Ruttenberg, A., Scheuermann, R.H., Brinkman, R.R., Courtot, M., Mungall, C., Xiang, Z., Chen, F., Todd, T., Colby, L.A., Rush, H., Whetzel, T., Musen, M.A., Athey, B.D., Omenn, G.S., and Smith, B. (Year). "VO: Vaccine Ontology", in: *The 1st International Conference on Biomedical Ontology (ICBO-2009)*: Nature Precedings), http://precedings.nature.com/documents/3552/version/3551.

Lu, C.J., and Browne, A.C. (Year). "Development of Sub-Term Mapping Tools (STMT)," in: *AMIA 2012 Annual Symposium, November 3-7, 2012*), Page 1845.

Mate, S., Kopcke, F., Toddenroth, D., Martin, M., Prokosch, H.U., Burkle, T., and Ganslandt, T. (2015). Ontology-based data integration between clinical and research systems. *PLoS One* 10**,** e0116656.

Ozgur, A., Xiang, Z., Radev, D.R., and He, Y. (2011). Mining of vaccine-associated IFN-gamma gene interaction networks using the Vaccine Ontology. *J Biomed Semantics* 2 Suppl 2**,** S8.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., and Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 25**,** 1251-1255.