



# Improvement of PubMed Literature Searching using Biomedical Ontology

Zuoshuang Xiang, Yongqun He - University of Michigan, Ann Arbor, MI USA

Unit for Laboratory Animal Medicine  
Department of Microbiology and Immunology  
Center for Computational Medicine and Bioinformatics  
University of Michigan Medical School  
018 ARF, 1150 W. Medical Center Dr.  
Ann Arbor, MI 48109-5614, USA  
URL: <http://www.violinet.org/vaccineontology>  
URL: <http://www.violinet.org/pubvo>

## Abstract

PubMed articles are annotated using the Medical Subject Headings (MeSH) to increase search efficiency. However, MeSH contains limited information on many biomedical domains (e.g., vaccine). Biomedical ontologies may be used to improve PubMed searching capability. This study demonstrates that Vaccine Ontology (VO) can be used to significantly improve PubMed searching efficacy in the vaccine domain. The recall and precision of the ontology-based literature mining approach are analyzed and discussed.

## Introduction

MeSH is the controlled vocabulary of medical and scientific terms that are used by biomedical scientists to manually index articles in the PubMed literature database. The MeSH terminology has been used in PubMed to improve literature searching. However, MeSH does not cover many biomedical domains (e.g., vaccine) well. An ontology represents the consensus-based controlled vocabularies of terms and relations which are logically formulated to promote intelligent information retrieval and modeling. We hypothesize that ontology-based PubMed search will significantly improve literature search efficacy. To test this hypothesis, we apply the Vaccine Ontology (VO; <http://www.violinet.org/vaccineontology>) to search for PubMed literature associated with *Brucella* vaccines. *Brucella* is an intracellular bacterium that causes brucellosis, the most common zoonotic disease worldwide (1).

VO development does not allow multiple inheritances. Each subclass in VO has an 'is\_a' relationship with its parent class. This ensures that all subclasses (e.g., *Brucella* RB51) can be included in a literature searching when a parent class (e.g., "*Brucella* vaccine") is searched. This is a basic principle in VO-based vaccine literature search.

In this presentation, we demonstrate that VO-based literature search dramatically increases PubMed searching sensitivity compared to the default searching engine in PubMed.

## Methods and Results

**Methods:** A user case study is to search "live attenuated *Brucella* vaccine" from the PubMed literature database. As of June 10, 2009, a direct PubMed search of this string of keywords returns 58 papers (or PubMed hits). VO includes 16 *Brucella* vaccines (including 4 licensed vaccines) that have phenotypes of 'live' and 'attenuated'. We developed an algorithm that recursively searches for all ontology labels and synonyms of the class '*Brucella* vaccine' and all its subclasses in VO. The phenotypes 'viable' (synonym: live) and 'attenuated' are used for filtering out unqualified *Brucella* vaccines. All names are assembled into a searching string for a PubMed keywords search. The searching results were manually analyzed in order to calculate and compare the precisions and recalls of the VO-based method and the default PubMed method.

**Results:** VO-based method dramatically improves PubMed searching sensitivity (Table 1).

**Table 1. Enhanced literature search using VO.**

PubMed Search Keywords	Hits	True	Precision
live attenuated <i>Brucella</i> vaccine	58	55	95%
Consider "live attenuated <i>Brucella</i> vaccine" in VO:			
<i>Brucella</i> (RB51 OR SRB51)	182	182	100%
<i>Brucella</i> (strain 19 OR S19)	537	510	95%
<i>Brucella</i> Rev. 1	145	144	99%
<i>B. suis</i> (strain 2 OR S2)	56	12	21%
<i>Brucella bacA</i> mutant vaccine	1	1	100%
Other 12 live attenuated <i>Brucella</i> vaccines in VO	62	59	95%
Total (unique ones)	763	695	91%

Our search using this VO-based method increased the recall of searching "live attenuated *Brucella* vaccine" by 13 fold (698/55) compared to the searching without VO (Table 1). The precision of the searching remains high in VO-based literature searching method (in total 91%) compared to PubMed searching (95%).

## Approaches to Improve VO-based PubMed Search

Our study suggests that the following approaches are necessary to improve VO-based PubMed searching:

(1) **Inclusion of synonyms in ontology terms significantly increases searching recall.** For example, inclusion of S19 as a synonym to strain 19 results in 100 more hits in PubMed search.

(2) **Increased recall and precision can only be achieved with well-assigned names for labels and synonyms.** For example, the search for the term '*Brucella abortus bacA* mutant vaccine' returns higher searching precision than '*Brucella abortus bacA* mutant' without lose of recall. On the other hand, "*Brucella* RB51" is better than "*Brucella abortus* RB51" for obtaining higher recall without lose of precision.

(3) **Ontology term labels containing numbers may be quoted to increase PubMed searching precision.** For example, using "strain 2" instead of (strain 2) dramatically improves precision to >95% from current ~20% (Table 1).

These approaches are generically applied for other ontology-based literature search. They also provide guidance in better ontology development.

## PubVO: VO-based Vaccine PubMed Search

We have developed a web server PubVO (<http://www.violinet.org/pubvo>) to implement our VO-based PubMed searching algorithm. SPARQL is used in PubVO to extract VO vaccine hierarchical information and apply it for increased power of PubMed literature search. The use case study shown in Table 1 has also been implemented in PubVO (Fig. 1). PubVO provides the final keywords submitted for PubMed search. It also lists all searched hits in PubVO. For each searched paper, the citation information is provided and the article title can be clicked to link to the paper abstract in PubMed. Alternatively, a link is provided for the user to directly visualize all the results in the PubMed web server (Fig. 1).

PubVO has been successfully applied to search other types of *Brucella* vaccines and vaccines against other pathogens. More advanced features are being added into PubVO.



**Final keywords to be searched:** (live attenuated *Brucella* vaccine) OR (B. abortus RB51) OR (Brucella RB-51) OR (Brucella RB51) OR (B. abortus strain 19) OR (Brucella S19) OR (Brucella strain-19) OR (B. abortus bacA mutant) OR (attenuated B. abortus with deletion of znuA) OR (B. abortus RB51WboA) OR (B. abortus strain RB51SOD) OR (B. abortus strain 45/20) OR (B. abortus P39 deletion vaccine) OR (B. melitensis P39 deletion vaccine) OR (B. melitensis strain VTRM1) OR (B. melitensis bp26 deletion vaccine) OR (Brucella CGV26) OR (B. melitensis WR201) OR (microencapsulated B. melitensis mutant vaccine) OR (Brucella melitensis strain Rev. 1) OR (B. suis strain VTRS1) OR (B. suis strain 2) OR (Brucella suis S2)

**Found 763 records. First 100 returned. Show results in PubMed.**

1. Kang SI, Her M, Heo EJ, Nam HM, Jung SC, Cho D. 2009. Molecular typing for epidemiological evaluation of *Brucella abortus* and *Brucella canis* isolated in Korea. *Journal of microbiological methods* [PMID: 19463862]
2. Martins H, Garin-Bastuji B, Lima F, Flor L, Pina Fonseca A, Boinas F. 2009. Eradication of bovine brucellosis in the Azores, Portugal-Outcome of a 5-year programme (2002-2007) based on test-and-slaughter and RB51 vaccination. *Preventive veterinary medicine* 90:80-89 [PMID: 19439382]
3. Tsolis RM, Seshadri R, Santos RL, Sangari FJ, Lobo JM, de Jong MF, Ren Q, Myers G, Brinkac LM, Nelson WC, Deboy RT, Angiuoli S, Khouri H, Dimitrov G, Robinson JR, Mulligan S, Walker RL, Elzer PE, Hassan KA, Paulsen IT. 2009. Genome degradation in *Brucella ovis* corresponds with narrowing of its host range and tissue tropism. *PLoS one* 4:e5519 [PMID: 19436743]
4. Galindo RC, Muñoz PM, de Miguel MJ, Marin CM, Blasco JM, Gortazar C, Kocan KM, de la Fuente J. 2009. Characterization of possible correlates of protective response against *Brucella ovis* infection in rams immunized with the B. melitensis Rev 1 vaccine. *Vaccine* 27:3039-3044 [PMID: 19428917]

Fig. 1. PubMed literature searching using PubVO. This example illustrates how the use case study is implemented.

## Conclusion and Discussion

Our study shows that biomedical ontologies (e.g., VO) can be used to significantly improve literature searching. While MeSH provides reasonable searching specificity, sensitivity may be very low for many biomedical research areas (e.g., vaccine). Therefore, it is possible to dramatically increase PubMed searching sensitivity by incorporating ontology information. Proper naming of ontology terms and their synonyms is critical to improve ontology-based literature search. To evaluate whether an ontology term is named accurately, one possible approach is to test if this name would increase true positive results and decrease false negative results in PubMed search. The conclusion of our VO-based literature search study is most likely applicable for literature searching using other ontologies.

## Acknowledgements

This research is supported by a Rackham Pilot Research grant to YH at the University of Michigan.

## References

1. Xiang Z, Zheng W, He Y. BBP: *Brucella* genome annotation with literature mining and curation. *BMC Bioinformatics*. 2006 Jul 16;7(1):347.