

## An Ontology Improves Text Information Access: A Case Study Using Human Disease Genes in Yeast

Crangle, Colleen<sup>\*1,2</sup>, Sopchak, Lynne<sup>1</sup>

<sup>1</sup>ConverSpeech LLC, Palo Alto, CA, USA; <sup>2</sup>Faculty of Informatics, University of Ulster, Ulster, Northern Ireland

**BACKGROUND:** Gene-related discoveries are increasingly being reported in scientific publications and annotated databases. In these reports, critical information is expressed as natural language in free-text form. The sheer volume of such text data has created a need for improved methods of information access. A formidable challenge lies in the fact that genes and proteins typically have multiple names, and they are referred to using terms that are variations of these names [1]. In addition, the same name can apply to many different biomedical entities, including distinct genes, substances, characteristics, and procedures. No simple method of synonym management gives reliable access to information on a specific gene of interest. **OBJECTIVE:** To understand how a biomedical ontology can improve access to free-text information about gene-related discoveries. **METHOD:** Case Study. We chose a set of genes identified as new candidate genes for the mitochondrial-related disorder of spastic paraplegia 5A [2]. The MEDLINE citation database was then searched for all articles relevant to this set of genes, using the query “CGI-11 OR LOC85479 OR PDE7A” submitted to PubMed. We submitted this same query to the ConverSpeech Distiller, a front-end to the PubMed database that uses the Entrez Programming Utilities from NCBI. The Distiller has integrated into it a biomedical ontology, BioMedPlus, for term expansion and results filtering. (Term expansion adds aliases and other synonyms to a search; filtering examines returned citations for words related to a variety of hierarchical and other terminology information, including definitions.) BioMedPlus is automatically constructed from publicly available resources such as NCBI’s LocusLink and the SGD<sup>TM</sup> database. It organizes gene and protein names, symbols, and aliases into synonym sets with hypernym (more general), hyponym (more specific), and holonym (part-whole) links. Results from (a) the PubMed search, (b) the Distiller search with term expansion only, and (c) the Distiller search with term expansion and results filtering were compared, using the measures of recall and precision. A failure analysis was performed on the results returned by the Distiller judged irrelevant and those omitted. **RESULTS AND CONCLUSIONS:** The PubMed search returned 19 citations. Of the 252 citations returned by the Distiller search, 52 were judged relevant by a human domain expert (2<sup>nd</sup> author). The PubMed search had precision of 100% (19/19) but recall of at most 37% (19/52). (There are at least 52 MEDLINE citations that are relevant.) The Distiller search with term expansion had low precision of 21% (52/252) but recall possibly as high as 100%. After results filtering, the Distiller search achieved precision of 100% (46/46) and recall of at most 88% (46/52). Failure analysis showed that term expansion introduced references to over two dozen different biomedical entities, and that results filtering eliminated all citations containing them but also excluded six relevant citations. We conclude that an ontology can improve access to gene-related text information if it is used not only for term expansion but also results filtering.

### References

1. Yu H, Hatzivassiloglou V, Friedman C, Rzhetsky A, Wilbur WJ. Proc AMIA Symp. 2002;:919-23. Automatic extraction of gene and protein synonyms from MEDLINE and journal articles.
2. Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, Jones T, Chu AM, Giaever G, Prokisch H, Oefner PJ, Davis RW. Nat Genet. 2002 Aug;31(4):400-4. Systematic screen for human disease genes in yeast.