Automatic extraction from scientific abstracts of synonyms for proteins and genes

Hong Yu
Dept. Medical Informatics, Columbia University

Introduction: Protein and gene names change frequently as research reveals details about these entities.[1] Because authors often use synonyms, information retrieval requires identification of these alternate names. Many biological databases — such as GenBank[2] and SWISSPROT[3]— have synonym databases; however, the databases may not be complete. Furthermore, to our knowledge, the extraction of synonyms is mainly done by laborious manual curating and review. It is desirable to automate the process due to the enormous volume of publication. We observed that many scientific abstracts have summaries of synonyms. The synonyms are often specifically proposed or mentioned and may be classified into a set of patterns to be recognized by automation.

Methods: We manually classified and evaluated patterns used by authors to represent synonyms of proteins or genes in scientific abstracts. We implemented a program, SRE (for synonym recognition and extraction), to recognize and extract the terms associated with the patterns. SRE is written in Perl. The output of SRE is sets of two or more synonyms. We applied SRE to 2,312 scientific abstracts, a subset of abstracts we downloaded from PubMed by the keyword "human." We then evaluated the precision of SRE's results, using our own judgment as the standard. *Precision* is the number of correct sets of synonyms of proteins or genes divided by the total sets of terms retrieved.

Results: We classified several patterns that express synonyms of proteins and genes in scientific abstracts. The simplest patterns are "synonym" or "a synonym of, " such as in "*Thermoactinomyces candidus should be considered a synonym of Thermoactinomyces vulgaris...,*"[5] where synonyms *Thermoactinomyces candidus* and *Thermoactinomyces vulgaris* can be extracted as noun phrases before and after the string "a synonym of." To evaluate whether the patterns of "synonym" and "a synonym of" would help us to find synonyms of proteins or gene names, we retrieved all the PubMed abstracts that contained the keyword *synonym* and manually analyzed whether the associated terms are proteins or genes. A search on the keyword *synonym* for abstracts from 1966 to present retrieved a total of 540 abstracts. A subset of 30 randomly selected abstracts contained no protein or gene names; in most cases, terms were names of species. We therefore discarded this approach. "Called" and "known as" are frequently used to introduce synonyms ("*...Apo3 (also known as DR3, WSL-1, TRAMP or LARD)*"[6]), as are various separation symbols, such as the solidus and comma, (e.g., *Apo3/DR3/Ws1-1/lymphocyte-associated receptor of death*[7]). We therefore implemented those patterns into SRE. Next, we had SRE search on the patterns in all the 2,312 abstracts; it extracted 453 sets of terms. Of them, we judged 15 (3.3%) sets to be genuine synonyms of protein and gene names. For example, one set is "Apo3|DR3|WSL-1|TRAMP|LARD". SRE erroneously paired *fibrosis|pulmonary*, *ig-beta|gamma*, and *CD94|NKG2A*. It also listed *Apo3|DR3|Wsl-1|lymphocyte*; the first three are synonyms, but the fourth, *lymphocye,* is not.

Discussion: SRE has a precision of 3.3%. Since many false positives are not protein or gene names, we shall increase precision greatly by sorting the retrieved entries and discarding those that are not protein or gene names. For this task, we may need an exhaustive list of protein and gene names. Many public databases include the names and descriptions of proteins and genes.[1-3] We may also screen out English words. However, research indicated that 5.6% gene names belonged to general English terms[8]. Even assuming we could check that the retrieved entries are indeed names of proteins or genes, however, we would not eliminate all false positives. For example, *CD94* and *NKG2A* are both binding-related proteins, but they are not synonyms. A further strategy is to link the retrieved protein and gene names to their primary sequences: if their sequences are identical, then the terms are synonyms. It may be possible to compare all the primary sequences in GenBank and SWISSPROT and to extract synonyms. This approach, however, would be challenging due to the enormous volume of primary data in GenBank and SWISSPROT. A good approach may combine the natural language processing (NLP) for validation. For example, if we could identify in an article that *CD94* binds to *NKG2A,* we would detect that the two were not synonyms.

Acknowledgements: I want to thank my advisor Andrey Rzhetsky for his contribution to this project. Hong Yu is supported by LM07079 "Research Training Grant".

References:

1. Genomics 1997, 45: 464-8.
2. Nuc Acids Res 2000, 28(1): 126-8.
3. Nuc Acids Res 2000, 28(1): 45-8.
4. Nuc Acids Res 2000, 28(1): 56-9.
5. Int J Syst Evol Microbiol 2000, 50 Pt 5: 1905-8.
6. Curr Biol 1998, 8(9): 525-8.
7. Arch Immunol Ther Exp (Warsz) 1999,47(4): 217-21.
8. Genome Inform Ser Workshop Genome Inform 1998. 9: 72-80.