Automatic Extraction of Gene and Protein Synonyms from MEDLINE and Journal Articles

Hong Yu,[1] Vasileios Hatzivassiloglou,[2] Carol Friedman,[1,3] Andrey Rzhetsky,[1,4] W. John Wilbur [5]

[1] Dept. Medical Informatics, Columbia University, New York, NY 10032, USA
[2] Dept. Computer Science, Columbia University, New York, NY 10027, USA
[3] Dept. Computer Science, Queens College, City University of New York, New York, NY 11367, USA
[4]Columbia Genome Center, Columbia University, New York, NY 10032, USA
[5] National Center for Biotechnology Information, Bethesda, MD 20894, USA

ABSTRACT

Genes and proteins are often associated with multiple names, and more names are added as new functional or structural information is discovered. Because authors often alternate between these synonyms, information retrieval and extraction benefits from identifying these synonymous names. We have developed a method to extract automatically synonymous gene and protein names from MEDLINE and journal articles. We first identified patterns authors use to list synonymous gene and protein names. We developed SGPE (for synonym extraction of gene and protein names), a software program that recognizes the patterns and extracts from MEDLINE abstracts and full-text journal articles candidate synonymous terms. SGPE then applies a sequence of filters that automatically screen out those terms that are not gene and protein names. We evaluated our method to have an overall precision of 71% on both MEDLINE and journal articles, and 90% precision on the more suitable full-text articles alone.

1. INTRODUCTION

Genes and proteins often have multiple names; as research reveals more details about these entities, additional names are often given for the same substance. Because authors often use different synonyms to refer to the same gene or protein across articles or sub-domains, information retrieval and information extraction benefits from identifying these alternate names. Many biological databases — such as GenBank and SWISSPROT— include synonyms; however, the databases may not be always up to date. Furthermore, synonym relationships between gene and protein names are mainly extracted by laborious manual curating and review. It is desirable to automate the process due to the enormous volume of published information about genes and proteins.

We distinguish between two types of synonymy for gene and protein names. Type I consists of the correspondence between the short and long forms of gene and protein names (e.g., *LARD* and *lymphocyte associated receptor of death*). Type II consists of the correspondence between all short forms (e.g., *Apo3*, *DR3*, *TRAMP*, *LARD*, and *wsl*). Previously, we developed a method [1] for linking the short and long forms of gene and protein names when both appear in published abstracts. In this study, we focus on the harder problem of automatically identifying Type II synonymy between single-word gene and protein names. In future work, we will explore automatic methods for identifying the remaining type of synonymy (between multi-word names) by linking those names to their short forms and using the equivalence between short forms.

We first identified the patterns that authors use to define synonymous short gene and protein names by analyzing a sample of MEDLINE abstracts and full-text journal articles. We then coded these patterns in SGPE (for synonym extraction of gene and protein names), a software program that automatically extracts synonymous gene and protein names associated with the patterns. We added two additional filters for screening out terms that are not genes and proteins, thus reducing SGPE's output to the cases of synonymous gene and protein names. We evaluated SGPE on 11 million MEDLINE abstracts (1966-2001) and more than 51,000 full-text articles from several leading biology journals.

2. BACKGROUND

Synonyms are different lexemes with the same meaning. Synonymous gene and protein names represent the same biological substances. This might be recognized by identical biological function or because they have the same gene or amino acid sequence.

Work in computational linguistics related to synonym detection has mainly focused on detecting semantically related words rather than exact synonyms, often using

surrounding words to cluster words according to their similarity. For example, approaches [2] and [3] identified "beer" and "wine" as similar words because both had similar surrounding words such as "drink", "people", "bottle" and "make". Hatzivassiloglou et al (1993) [4] used similar techniques to link scalar adjectives such as "hot" and "cold." In the biomedical domain, Hole (2000) [5] selected biomedical multi-word terms as candidate synonyms if those terms shared any word(s). For example, the string "*cerebrospinal fluid"* leads to "*cerebrospinal fluid protein assay," "CSF protein assay,"* and "*CSF protein."* Wilbur and Kim (2001) [6] developed a trigram matching algorithm to identify multi-word synonymous phrases.

A number of systems have identified Type I synonyms by mapping abbreviations to their full forms. This research includes the approaches of Hisamitsu and Niwa (1998) [7] and Bowden et al (1998) [8] that mapped common abbreviations to full forms, the approach of Yu et al (2002) [9] that mapped common biomedical abbreviations to full forms, and the approaches of Yoshida et al (2000) [10] and Yu et al [1] that mapped gene and protein short names to their long forms. Little work has been reported on the identification of Type II synonymous gene and protein names. This paper presents a method that uses patterns to detect Type II synonymous gene and protein names from MEDLINE abstracts and articles.

3. METHODS & RESULTS

Our study has two components: 1) Identification of Synonymous Patterns; 2) Applying SGPE to MEDLINE and Journal Articles.

3.1 Identification of Synonymous Patterns

3.1.1 Methods

We randomly selected a few sets of synonymous gene and protein names (e.g., *APO3, DR3, TRAMP, LARD* and *wsl*) from the SWISSPROT databank and extracted from MEDLINE and journal articles sentences in which at least two synonymous gene and protein names occur. We manually identified the commonalities among those sentences, which usually incorporate similar strings of symbols or words that are in our interest. The common patterns were chosen and implemented into SGPE to exclude those sentences that had the patterns that SPGE implemented. We then iteratively repeated the process until all sentences were excluded (i.e., covered by SPGE). In the future, we will examine methods used for hyponym identification [11] for applicability to our task.

We also manually explored several alternative patterns. These patterns included "synonym" or "a synonym of, " such as in "*Thermoactinomyces candidus should be considered a synonym of Thermoactinomyces vulgaris…,*" where the two synonyms *Thermoactinomyces candidus* and *Thermoactinomyces vulgaris* can be extracted as noun phrases before and after the string "a synonym of."

3.1.2 Patterns found

We found that various separation symbols such as slash and comma are the most frequent patterns that authors use to list synonymous gene and protein names in their MEDLINE abstracts. For example, sentences that appear in MEDLINE included "*We found neither evidence for expression of the recently identified TWEAK receptor Apo3/TRAMP/WSL/DR3/LARD, nor indications for direct interactions of tweak with TNFR"* and *"…the mouse homologue for the Wsl-1 (DR3, Apo3, TRAMP, LARD, TR3, TnfRSF12) gene."*

We found that the patterns of slash and comma also apply to journal articles for listing synonymous gene and protein names. In addition, many of the patterns of slash and comma are also associated with one of the following phrases "also called," "known as" and "also known as" for listing synonyms. For example, "*…a subset of the family, comprising cd95, tnf-r1, p75 ngf-r, dr3 (also called wsl-1, tramp, apo-3 or lard), car1 and the two trail receptors (dr4/trail-r1 and trail-r2) also share a…."* and "*…which include WAF1 (also called p21/Cip1/Sdi1), p16, p27, and p15."*

To evaluate whether the patterns of "synonym" and "a synonym of" would help us to find synonyms of protein 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 2001 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 those patterns.

3.2 Applying SGPE to MEDLINE and Journal Articles

3.2.1 Methods

We applied SGPE to 11 million MEDLINE abstracts (1966-2001) and 51,814 journal articles in our database. The journal articles include mostly *Cell, Science, J Biol Chem, and Curr Opin Biotechnol*, up to

year 2002. Since MEDLINE covers not only the biological domain, but also other domains such as medicine, we applied keywords such as *protein, gene, peptide,* and *receptor* to select abstracts that are relevant to the biological domain.

We developed pattern-matching methods to extract candidate synonymous gene and protein names that were associated with the patterns we recognized. For the patterns of slash and comma, we directly selected the terms that lie on either side of these symbols.

For the patterns of "known as", "also known as" and "also called", we further separated into the patterns with and without parentheses. An example of a pattern with parentheses is "*The transcription factors: Egr-1 (also called NGF-I-A, Krox, Tis 8, Zif 268) and Egr-2 and Egr-3…*", where the candidate terms are within the parentheses and precede the left parenthesis. Examples that do not include parentheses include "*These include: galectin-1, also known as galaptin, L-14-I, L-14, and BHL, a homodimer with subunit molecular mass*" and "*it is coregulated with GRP78, also known as the immunoglobulin heavy chain binding protein, BiP…*"

We developed pattern-matching methods to recognize automatically the terms that are associated with parentheses. The methods were similar to the methods of slash and comma except that the boundaries of terms were the symbols of left and right parenthesis instead of white space. In addition, SGPE recognized the terms that precede the left parenthesis. In our study, we discarded the patterns without parentheses due to the contextual complexity that makes it difficult to develop a simple pattern-matching approach for automatic identification.

The output of the first stage of SGPE is sets of two or more terms with their PubMed ID or journal article ID. The terms are candidate synonymous gene and protein names.

We then implemented two filters that apply external and internal knowledge to filter out those terms that are not gene and protein synonymous names.

- **SGPE-Filter-1** applies a dictionary of units (e.g., min and sec) and a dictionary of common English words to filter out those terms that are not gene and protein names. One rule excludes a set of candidate synonymous terms if one or more of the terms is a unit. The second rule excludes a set of candidate synonymous terms if two-thirds or more of terms are common English words. Note that we do not exclude all the common English words because a small number of gene and protein short names are also English words [12]. For example, our method therefore kept the synonymous set "*APO3|DR3|TRAMP|LARD|WSL,*" which contain two English common words.

- **SGPE-Filter-2** applies knowledge within the abstracts and journal articles to filter out the sets that are not synonymous gene and protein names.

  We first implemented into SGPE-Filter-2 an approach we developed in a previous study [9] that maps the abbreviations to full forms (when the full forms are defined within parentheses), from which SGPE-Filter-2 filtered out those sets that contain any term that is not a gene or a protein name. For example, SGPE-Filter-2 filtered out *HIV-2/SIVSMM/MAC* since *HIV-2* was defined in the abstract as *human immunodeficiency virus type 2*, which SGPE-Filter-2 recognized as a term outside the gene/protein set. The methods for identifying whether abbreviations and full forms are gene or protein names were developed in our previous study [1].

  Note that not all the terms in our candidate synonymous terms have their definitions in the abstracts. After this step of filtering, the candidate sets still consist of a lot of other terms that are not gene and protein synonymous names.

  SGPE-Filter-2 then filters out those candidate synonymous terms that are listed two or more times in the same abstracts or journal articles. We implemented this approach into SGPE-Filter-2 because we recognize that most of the authors introduce a synonym list only once in their abstracts. This filter will exclude some terms that are units (e.g., "*nl|min|kg*") and chemical compounds (e.g., "*sulfate|urea*"). The filter also exclude protein terms that are not synonyms but functionally related. For example, in the case of "*CD94|NKG2A,*" both *CD94* and *NKG2A* are binding-related proteins, but they are not synonyms.

  The short gene and protein names usually consist of less than seven letters [13]. SGPE-Filter-2 therefore excludes sets where any single term has more than six letters. For example, this rule filtered out the set *hypocalcaemic|hypomagnesaemic|*

hyperphosphataemic. SGPE-Filter-2 also removes sets with any term that contains two or more dashes (e.g., *d-galactose|n-acetyl-d-galactosamine|sialic*), as such terms are usually not gene or protein names.

3.2.2 RESULTS

SGPE identified a total of 1,666,966 sentences (160,282 from MEDLINE; 1,506,684 from journal articles) that match the slash pattern, 3,696,466 sentences (73,101 from MEDLINE; 3,623,365 from journal articles) that match comma pattern, and 10,440 sentences from journal articles that match the phrasal patterns "known as", "also called" and "also known as."

From the slash and comma sentences from MEDLINE, SGPE output a total of 822 unique sets of candidate synonymous gene and protein names. From 51,814 journal articles, SGPE output a total of 1,784 sets of candidate synonymous gene and protein names that were associated with parentheses.

4. EVALUATION

4.1 Evaluation Methods

We first randomly selected a total of 60 SGPE output sets (30 from MEDLINE; 30 from journal articles) and manually evaluated the precision. We then mapped the synonymous gene and protein names from the 60 SGPE output sets to SWISSPROT (version 40).

Note that, automatic gene and protein term identification is still a work in progress. In our case, it is more difficult to recognize gene and protein synonymous terms when they are not associated with parentheses. For example, it is difficult to apply simple pattern-matching rules to identify the protein term "*the transcription factor ternary complex factor*" from the sentence "*MAP kinases phosphorylate the transcription factor ternary complex factor, known as p62…*" In this study, we did not implement methods to extract automatically candidate synonymous terms from sentences that are associated with the phrases "known as," "also called," and "also known as" and that do not include parentheses for listing. However, we evaluated whether those sentences have lists of synonymous gene and protein names and in the future, we will develop methods to identify automatically those synonyms.

We randomly selected 30 sentences extracted from 51,814 journals that are not associated with parenthesis and phrases "known as," "also known as," and "also called." We manually evaluated whether the sentences included synonymous gene and protein names.

4.2. Evaluation Results

We identified a total of 9 and 27 synonymous gene and protein names from the 30 SGPE MEDLINE sets and 30 SGPE journal articles sets, respectively. The precision of SGPE in extracting synonymous gene and protein names from MEDLINE and journal articles is 0.30 (>0.17 at 95% confidence) and 0.90 (>0.76 at 95% confidence), respectively. The overall precision can be estimated as 0.71 [(0.30*822+0.90*1784)/2606]. Twenty-three out of thirty sentences that have the patterns of "*known as,*" "*also called,*" and "*also known as*" and that are not associated with parentheses included lists of synonymous gene and protein names.

A total of 11 sets of synonymous gene and protein names out of 36 sets extracted from SGPE (nine from journal articles and two from MEDLINE) were partially matched to SWISSPROT; the remaining sets all matched to SWISSPROT. An example of a partially matching synonymous set is "*|Egr-1 |Zif |NGF1-A |Krox 24 |TIS |CEF5*", where *TIS* and *CEF5* were not included in SWISSPROT.

5. DISCUSSION

Applying external and internal knowledge to extract information is commonly used in information retrieval and extraction. For example, Klavans and Muresan (2001) [14] extracted definitions of biomedical concepts by recognizing patterns authors use to define those concepts. Applying a dictionary of common English words to gene and protein name recognition was first introduced by Fukuda et al. (1998) [15] for protein name identification and Proux et al. (1998) [12] for gene name identification.

SGPE automatically identified synonymous gene and protein names from the literature with an overall estimated precision of 71%. Our results indicate that SGPE identifies some valid synonymous gene and protein names that do not exist in SWISSPROT. Therefore, SGPE is a useful tool for identifying newly coined synonymous gene and protein names from the literature. We did not measure the recall because we do not have exhaustive lists of synonymous genes and proteins that can be used as a gold standard for measuring the recall.

Further, SGPE's performance was much higher (90%) on full-text articles, and such articles become increasingly available online. The reason that SPGE

performs better in full text than in abstracts is that synonymous gene and protein terms are more frequently listed in the introduction section of full articles than abstracts.

The limitations of SGPE include that it relies on authors to list synonymous gene and protein names in the literature. Not all the authors will list synonymous gene and protein names and therefore the extraction may not be complete. We will experiment with methods based on surrounding words [2] [3] [4] for identifying synonymous gene and protein names, and therefore overcome the limitation of relying on authors. We may also apply functional relationships for synonymous gene and protein name identification. The hypothesis is that synonymous gene and protein names will be described to have the same functional relationships with other gene and protein names. The functional relations may be identified by GENIES [16], a natural language semantic parser that was developed for the biological domain.

SGPE does not identify the patterns of "known as", "also called", and "also known as" that do not include parentheses; therefore SGPE missed some of the synonymous gene and protein names. In order to capture those patterns, we may apply morphological cues such as Fukuda's [15] upper case and numerical features for gene and protein name identification, from which we recognize the patterns that may be implemented into SGPE. We may also apply the part-of-speech tagger that was developed by Tanabe and Wilbur (2002) [17] and that was specifically trained on the biomedical domain to assist pattern recognition.

ACKNOWLEGEMENTS

This research was supported in part by National Science Foundation Innovative Technology research grant EIA-0121687 and National Institutes of Health grant RO1 GM61372-01A2. Hong Yu was also supported by Research Training Grant LM07079 from the National Library of Medicine. Any opinions, findings, or recommendations are those of the authors and do not necessarily reflect the views of the sponsoring agencies.

REFERENCES

1. Yu, H., et al. *A rule-based approach for automatically identifying gene/protein terms in MEDLINE abstracts*. Submitted
2. Lin, D.K. *Automatic retrieval and clustering of similar words*. in *Proceedings of ACL-98*. 1998.
3. Li, H. and N. Abe. *Word clustering and disambiguation based on co-occurrence data.* In *CMP-lg*. 1998.
4. Hatzivassiloglou, V. and K. McKeown. *Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning*. In *ACL 1993*. 1993.
5. Hole, W.T. and S. Srinivasan. *Discovering missed synonyms in a large concept-oriented metathesaurus.* in *Proc AMIA Symp*. 2000.
6. Wilbur, W.J. and W. Kim. *Flexible phrase-based query handling algorithms*. In *Aversa E, Manley C, eds. Proceedings of the ASIST 2001 Annual Meeting.* 2001. Washington, D.C.
7. Hisamitsu, T. and Y. Niwa. *Extraction of useful terms from parenthetical expression by using simple rules and statistical measures.* In *CompuTerm98.* 1998.
8. Bowden, P.R., L. Eventt, and P. Halsted. *Automatic arconym acquisition in a knowledge extraction program*. in *CompuTerm98.* 1998.
9. Yu, H., G. Hripcsak, and C. Friedman, *Mapping abbreviations to full forms in biomedical articles.* J Am Med Inform Assoc, 2002. 9(3): p. 262-72.
10. Yoshida, M., K. Fukuda, and T. Takagi, *PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary.* Bioinformatics, 2000. 16(2): p. 169-75.
11. Hearst, M. *Automatic acquistion of hyponyms from large text corpora*. in *Proceedings of the fourteenth international conference on computational linguistics.* 1992. Nantes, France.
12. Proux, D., et al., *Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction.* Genome Inform Ser Workshop Genome Inform, 1998. 9: p. 72-80.
13. Maltais, L.J., et al., *Rules and guidelines for mouse gene nomenclature: a condensed version. International Committee on Standardized Genetic Nomenclature for Mice*. Genomics, 1997. 45(2): p. 471-6.
14. Klavans, J. and S. Muresan. *Evaluation of the DEFINDER System for Fully Automatic Glossary Construction.* In *Proc AMIA Symp.* 2001.
15. Fukuda, K., et al., *Toward information extraction: identifying protein names from biological papers.* Pac Symp Biocomput, 1998: p. 707-18.
16. Friedman, C., et al., *GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles.* Bioinformatics, 2001. 17(Suppl 1): p. S74-82.
17. Tanabe, L. and W.J. Wilbur, *Tagging gene and protein names in biomedical text.* Bioinformatics, 2002. In press.