Evaluating gene/protein name tagging and mapping for article retrieval

Chong Min Lee, Manabu Torii, Jinesh Shah, Yi-Ting Tsai, Zhang-Zhi Hu, Hongfang Liu*

Lab of Text Intelligence in Biomedicine
Georgetown University Medical Center
Washington, DC 20007
{cml54, mt352, jcs222,yt235,zh9, hl224}@georgetown.edu

Abstract

Background: Tagging gene/protein names in text and mapping them to database entries are critical tasks in biological literature mining. Most of the existing tagging and normalization approaches, however, have not been evaluated for practical use in article retrieval towards efficient biocuration.

Results: By utilizing literature cross-reference information provided by NCBI Entrez Gene database, we found that the coverage of gene/protein databases with respect to gene/protein names found in text is around 94%. The upper bound of the recall in retrieving MEDLINE citations by gene/protein names is around 70-80% when citations cross-referred by many genes are overlooked and flexible matching of names are used. Of genes/proteins failed to be retrieved by names, over 30% are caused by citations not discussing cross-referred genes/proteins in the abstracts and around 60% are caused by the gene/protein name tagging system trained on the BioCreAtIvE II gene mention corpus.

Conclusions: The study demonstrates that existing gene/protein databases have a decent coverage of gene/protein names used in MEDLINE abstracts. Approaches and data resources for gene/protein tagging and mapping need to be selected appropriately for individual practical tasks.

1 Background

Literature mining has become important part of modern biomedical research and involved in tasks ranging from helping biologists retrieve research articles to automatically extracting designated types of information from articles (Krallinger and Valencia 2005; Krallinger, Valencia et al. 2008). One of the practical applications of biomedical literature mining is to detect articles describing a specific gene or protein. Many existing molecular databases provide literature cross-reference information. For example, the National Library of Medicine (NLM) began an initiative to link scientific publications to Entrez Gene entries via Gene Reference Into Function (GeneRIF). Similar to the sequence submission mechanism in GeneBank, GeneRIF records can be provided by individual researchers. For protein annotations, the UniProt consortium has devoted to providing annotation evidences, including those from literature, during the curation of protein records.

Given the current level of maturity of biomedical literature mining applications, it may be difficult to fully automate the knowledge acquisition at the level that is comparable with expert curators. However, there have been evidences that literature mining applications can significantly boost the efficiency and the quality of the curators' work. For example, Textpresso (Muller, Kenny et al. 2004) is an information retrieval system that can retrieve sentences from full-length articles. The system is equipped with a semantic classification system consisting of 33 term categories. Target documents retrieved and stored in the system are pre-processed, and phrases identified in documents are automatically labeled with semantic categories. Category annotation allows users to formulate sentence retrieval queries that consist of term categories as well as key phrases. Another retrieval system is PubSearch (http://pubsearch.org/) (Harris, Clark et al. 2004), a web-based curation tool for genes, that allows curators to search for documents containing designated genes and also to annotate documents. PreBind (Donaldson, Martin et al.

2003) was designed to support human curation of BIND, an online database of protein-protein interaction. In the PreBind system, protein names and their synonyms were first extracted from sequence databases, RefSeq and SGD, and MEDLINE records containing protein names and their synonyms were retrieved. Then, MEDLINE citations potentially containing protein interaction information were identified using a text categorization system. It reportedly reduced the duration of the task of extracting protein interaction information by 70%.

Recently, automated gene/protein tagging and mapping systems have achieved reasonable performance when species information is provided, as evidenced in BioCreAtIvE workshops (Morgan, Hirschman et al. 2004; Hirschman, Colosimo et al. 2005; Hirschman, Yeh et al. 2005; Krallinger, Leitner et al. 2007; Altman, Bergman et al. 2008; Krallinger, Morgan et al. 2008; Krallinger, Valencia et al. 2008; Morgan, Lu et al. 2008). However, it is not clear how these systems perform in retrieving articles relevant to a specific gene or protein. Additionally, it is not clear how important it is to have a comprehensive list of gene/protein names and to be able to handle variant forms of a gene/protein name in text.

Utilizing literature cross-reference information provided by Entrez Gene (i.e., GeneRIF), we designed an experiment to answer the following several questions related to gene/protein tagging, and mapping for gene/protein curation:

- what is the coverage of an existing gene/protein dictionary assembled from existing databases, BioThesaurus, regarding to gene/protein names mentioned in abstracts;
- what is the performance of an existing gene/protein tagging system, BioTagger-GM, when evaluated in a practical curation setting;
- how flexible matching criteria needs to be when dictionary lookup is employed for gene/protein name mapping; and
- what is the upper bound of the recall when using such automated systems to link MEDLINE citations to gene/protein records in databases based on gene/protein names mentioned in abstracts.

In the following, we first describe the resources and systems used in the study. The study design and assessment method are presented next. We then present and discuss the results and conclude the paper.

2 Methods

The study was designed to utilize existing resources and systems

2.1 Resources

Resources used in the data include text and gene resources available from National Center of Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov) and gene/protein terminology resources and tagging systems available at Lab of Text Intelligence at Georgetown University (http://biomine.dbb.georgetown.edu). The following provides a brief summarization for each of them.

MEDLINE is an NLM's premier bibliographic database. We used the 2010 distribution of MEDLINE that contains citations information over 18 million articles published in the life science domain.

Entrez Gene (Wheeler, Church et al. 2004) is NCBI's database for gene-specific information. Each gene is given a unique identifier (GENEID) in the database. Among the information in the database, literature cross-reference information of genes is provided in GeneRIF.

BioThesaurus (Liu, Hu et al. 2006) is a web-based thesaurus designed to map protein and gene names to protein entries in the UniProt Knowledgebase (UniProtKB) or gene entries in Entrez Gene. The latest gene-centric version (July 1, 2010) contains over 11 million names extracted from 32 molecular biological databases according to the cross-references provided by UniProtKB or Entrez Gene, as well as cross-reference information provided in each individual database.

BioTagger-GM (Torii, Hu et al. 2009) is a gene/protein name tagger utilizing BioThesaurus and Conditional Random Field (CRF). The tagger was trained on the training data of the BioCreAtIvE II gene mention task. The trained CRF model together with a post-processing module yielded an F-score over 86% on the test data of BioCreAtIvE II gene mention.

2.2 Data preparation

We obtain a collection of paired IDs (PMID, GENEID) from GeneRIF, where a gene record GENEID has a literature reference PMID. For each pair, gene/protein names associated with

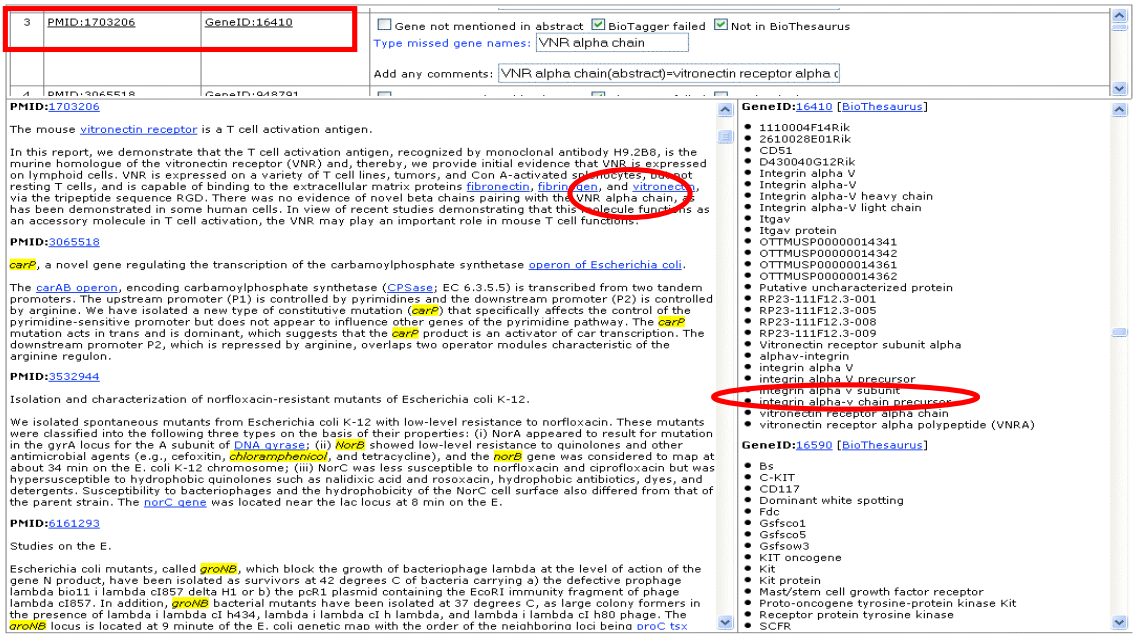


Figure 1. The manual assessment interface for (PMID, GENEID) pairs failed to be mapped.

gene record GENEID are retrieved from BioThesaurus, and additionally the abstract of the cited reference PMID is processed by BioTagger-GM for detection of gene/protein names. For example, we retrieved two pairs (19570885, 20393) and (19570885, 20497) from GeneRIF, where two genes with GENEIDs 20393 and 20497 are cross-referred with one literature citation with PMID 19570885. BioTagger-GM identifies several gene/protein names including one name (i.e., "SGK1") for the gene with GENEID 20393 and one name (i.e., "NCC") for the gene with GENEID 20497.

2.3 Name mapping

For each pair (PMID, GENEID), we used two approaches to find mappings between names identified by BioTagger-GM in the abstract of the cited reference PMID and names of the gene record GENEID in BioThesaurus. When a mapping with the best score between a pair of names is found, the pair is considered to be detectable through automated approaches. The first approach is a relaxing method, where exact matching was tested first and then the name without the first and/or the last words were tried for dictionary lookup. Possible number of removed words is limited to two words. The number of removed words is recorded as the penalty score of the mapping ranging from 0 to 2. The second approach is to use a similarity measure, Jaccard Index (JI) defined as $\frac{|C_1 \cap C_2|}{|C_1 \cup C_2|}$. In the formula, C_i is the set of words. The similarity measured by JI ranges from 0 to 1. When two names do not have a shared word, the score is 0, and 1 if identical. Additionally, a normalization procedure is used to accommodate variations caused by lexical variants of words and punctuation marks. Specifically, punctuation marks are ignored, all letters are lower-cased, and lexical variants are normalized based on Specialist Lexicon provided in the Unified Medical Language System (UMLS).

2.4 Assessment

The statistics and assessment measures are calculated in each of the five groups of abstracts, grouped according to the number of distinctive genes referred in an abstract (1, 2:4, 5:16, 17:256, and > 256). We report the coverage of (PMID, GENEID) pairs on how many candidate pairs of detected names and GENEIDs are generated by varying thresholds of scores in each range.

There are several causes for those failed to be mapped:

- **Gene not mentioned in the abstract** - Any name for a listed gene does not appear in an abstract, e.g., genes are mentioned in the full-length article but not in the abstract.

Table 1. The overall statistics of the data and assessment results using the relaxing method.

Group	#PMIDs	#Names/art.	#(PMID, GID)	Before Normalization (%)			After Normalization (%)		
				BN0	BN1	BN2	AN0	AN1	AN2
1	261,593	4.20	261,593	63.63	72.57	76.37	74.74	83.02	86.23
2:4	164,882	5.69	414,269	53.9	60.08	63.65	66.73	72.33	74.13
5:16	37,667	6.10	282,996	29.64	33.68	37.15	38.07	43.26	51.47
17:256	4,634	4.80	199,945	4.99	6.41	8.82	6.44	9.08	19.44
>256	497	2.88	1,251,434	0.04	0.06	0.55	0.06	0.35	1.82

- **BioTagger-GM failed** – BioTagger-GM failed to identify a name for a listed gene mentioned in the abstract.
- **Names not in BioThesaurus** – BioThesaurus does not contain a name for a listed gene mentioned in the abstract.

To estimate the numbers of (PMID, GENEID) pairs failed to be identified in the full-scale, we sampled 100 pairs from each group that are failed to map before removing any word. Figure 1 shows the evaluation interface we built to analyze the results. Note that BioTagger-GM and BioThesaurus could fail at the same time (the second and the third causes listed above). For example, as shown in Figure 1, given a GeneRIF pair (1703206, 16410), the name to be detected in the abstract is "VNR alpha chain", while BioTagger-GM failed to tag it and BioThesaurus did not cover the name, even though a similar name "vitronectin receptor alpha chain" can be found in BioThesaurus.

3 Results

There are 2,410,237 (PMID, GENEID) pairs extracted from GeneRIF associated with 469,273 articles with an average of 5.14 genes per article. Over 90% of the articles have less than five genes cross-referred. In average, there are 4.88 gene/protein names per abstract identified by BioTagger-GM. Table 1 and Figure 2 show the statistics and assessment results for abstracts in the five groups (i.e., 1, 2:4, 5:16, 17:256, and > 256). In articles with one gene cross-referred, 63.3% of them are identifiable using exact string matching (i.e., BN0). Generally, the measurement increases around 10% (e.g., 63.3% BN0 to 74.4% AN0) after string normalization and around additional 9% (e.g., 83.02% AN1) if a leading or a trailing word in a name was removed. For some articles with many genes cross-referred, the chance of finding their names in the abstract decreases to almost 0%. The results obtained using JI are similar to the ones obtained using the relaxing approach. Figure 2 also shows the percentage of mapped pairs decreases when the number of genes cross-referred by the abstract increases.

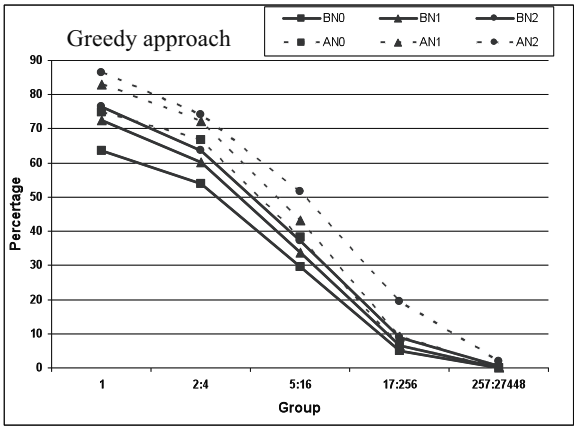
Table 2 shows the distribution of the causes for failed mapping of pairs. Two analysts agreed most of the times with the causes. Note that some pairs can have two causes of failed mapping: "BioTagger-GM failed" and "Not in BioThesaurus" (11 pairs in Group 1 and 5 in Group 2:4). When the number of genes cross-referred is less than 5, around 34% of them could not be identified because the genes are not mentioned in the abstract and around 60% of them are failed because of BioTagger-GM failures. For abstracts with many genes cross-referred, the dominant cause of pairs failed to be mapped is genes not mentioned in abstracts.

4 Discussion

We assessed gene/protein entity tagging and mapping in article retrieval to assist gene/protein curations by utilizing literature cross-reference information provided in GeneRIF and publicly available resources.

The results suggest that papers linked to many genes very rarely contain the corresponding names in their abstracts. In fact, most of the papers associated with more than 256 genes (the fifth group) are either about genome sequencing projects or about databases, which would be irrelevant to the curation of particular genes or proteins.

The manual evaluation indicates that gene/protein names listed in databases are comprehensive in including gene/protein names mentioned in text. Among the pairs failed to be mapped in Group 1 and Group 2:4 (about 30% of the total), less than 20% of them are caused by names not in BioThesaurus, which indicates over



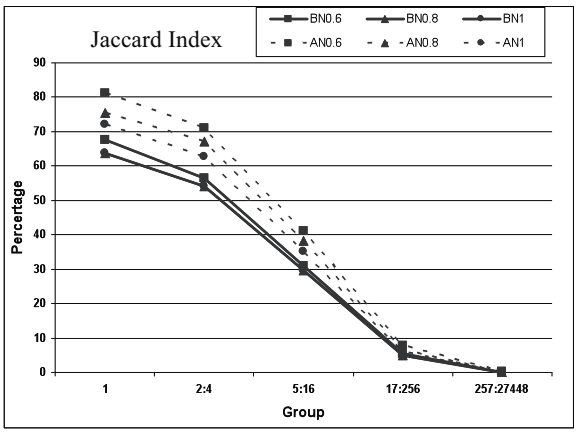


Figure 2. Percentages of (PMID, GENEID) pairs detectable with different dictionary lookup flexibility for different groups.

94% coverage of names by BioThesaurus (i.e., 1-20%×30%=0.94).

Relaxing dictionary lookup is important for article retrieval. As observed in Figure 2, there is an increase up to 10% in the coverage of (PMID, GENEID) pairs after normalization. Also, an increase of 20% in coverage is observed when allowing at most two-word difference. One main factor of such a big increase is due to additional modifiers in names detected by BioTagger-GM. For example, species names frequently occur as modifiers of gene/protein names in text, while in BioThesaurus species names seldom occur in the names. Also, words indicating semantic categories such as "gene" or "protein" may be present in the names used in text (e.g., "SMAD2 gene" vs. "SMAD2").

The study demonstrates that among pairs failed to be mapped in Group 1 and Group 2:4, over 30% are caused by names not mentioned in the abstract. It indicates that an upper bound of a recall is around 91% (1-30%*30%=0.91) for article retrieval when using abstracts only.

Table 2. Manual assessment results of 500 (PMID, GENEID) failed pairs with 100 pairs per frequency group.

Group	1	2:4	5:16	17:256	>256
Not in BioThesaurus	17	12	4	3	0
BioTagger-GM failed	60	59	29	8	0
Gene not mentioned in abstract	34	34	67	92	100
# Agreed by two analysts	87	83	90	94	100

Among the cases where a name could not be properly mapped, 60% of them were names not detected by BioTagger-GM. The tagger failed even when names were included in BioThesaurus, although BioThesaurus lookup results are used as features in the tagger. This might be attributed to the fact that mere lookup of BioThesaurus yields a low precision by itself, even though the recall can be high. Another important consideration is the definition of "genes/proteins". The annotation guidelines of genes/proteins for the BioCreAtIvE corpus may not conform to the notion of genes/proteins for the purpose of GeneRIF annotation.

5 Conclusions

We have conducted an assessment of the coverage of gene/protein names in databases with respect to gene/protein names in text, and automated gene/protein tagging and mapping for retrieving articles relevant to specific genes or proteins. The study demonstrates that existing gene/protein databases have a decent coverage of gene/protein names mentioned in the text. The study provides an upper bound of recall when using automated methods to retrieve articles. The study suggests that the most appropriate approach and data resources to facilitate gene/protein tagging and mapping needs to be selected for the specific task in hand.

Acknowledgments

This work was supported by NIH 1-R01-LM009959-01A1 and NSF CAREER 0845523.

References

1. Altman, R. B., C. M. Bergman, et al. (2008). "Text mining for biology--the way forward: opin-ions from leading scientists." Genome Biol 9 Suppl 2: S7.
2. Donaldson, I., J. Martin, et al. (2003). "Pre-BIND and Textomy-mining the biomedical literature for proteinprotein interactions using a support vector machine." BMC Bioinformatics 4(11): 1471-2105.
3. Harris, M. A., J. Clark, et al. (2004). "The Gene On-tology (GO) database and informatics re-source." Nucleic Acids Res 32(Database is-sue): D258-61.
4. Hirschman, L., M. Colosimo, et al. (2005). "Overview of BioCreAtIvE task 1B: normalized gene lists." BMC Bioinformatics 6 Suppl 1: S11.
5. Hirschman, L., A. Yeh, et al. (2005). "Overview of BioCreAtIvE: critical assessment of informa-tion extraction for biology." BMC Bioinfor-matics 6 Suppl 1: S1.
6. Krallinger, M., F. Leitner, et al. (2007). Assessment of the second biocreative PPI task: automatic extraction of protein-protein interactions. Proceedings of the Second Biocreative Chal-lenge Evaluation Workshop.
7. Krallinger, M., A. Morgan, et al. (2008). "Evaluation of text-mining systems for biology: overview of the Second BioCreative community chal-lenge." Genome Biol 9 Suppl 2: S1.
8. Krallinger, M. and A. Valencia (2005). "Text-mining and information-retrieval services for mo-lecular biology." Genome Biol 6(7): 224.
9. Krallinger, M., A. Valencia, et al. (2008). "Linking genes to literature: text mining, information extraction, and retrieval applications for bi-ology." Genome Biol 9 Suppl 2: S8.
10. Liu H, Hu ZZ, Zhang J, Wu C. BioThesaurus: a web-based thesaurus of protein and gene names. Bioinformatics. Jan 1 2006;22(1):103-105.
11. Morgan, A. A., L. Hirschman, et al. (2004). "Gene name identification and normalization using a model organism database." J Biomed In-form 37(6): 396-410.
12. Morgan, A. A., Z. Lu, et al. (2008). "Overview of BioCreative II gene normalization." Genome Biol 9 Suppl 2: S3.
13. Muller, H. M., E. E. Kenny, et al. (2004). "Textpres-so: an ontology-based information retrieval and extraction system for biological litera-ture." PLoS Biol 2(11): e309.
14. Torii M, Hu Z, Wu CH, Liu H. BioTagger-GM: a gene/protein name recognition system. J Am Med Inform Assoc. Mar-Apr 2009;16(2):247-255.
15. Wheeler, D. L., D. M. Church, et al. (2004). "Data-base resources of the National Center for Biotechnology Information: update." Nucleic Acids Res 32(Database issue): D35-40.