PSE: A Tool for Browsing a Large Amount of PubMed Abstracts with Gene Names and Common Words as the Keywords

Takashi Yoneya
t-yoneya@kirin.co.jp
Pharmaceutical Research Laboratories, Pharmaceutical Division, Kirin Brewery Co. Ltd, 3 Miyahara, Takasaki, Gunma 370-1295, Japan

Keywords: text mining, gene name, PubMed, cooccurrence sentence

1 Introduction
In research target discovery process, PubMed [3] abstracts are one of the most important data resources. Most of the abstracts contain the important sentences leading the estimation. In many cases, the importance is decided by the relationship between a gene name and common words. For example:
   a) c-Kit is constitutively activated in various tumors.
   b) c-Kit expression increases in various tumors.
   c) c-Kit expression is observed in various tumors.
Although all these sentences indicate the relation between c-Kit and tumor, relative importance of the meanings is different and decided by the underlined common words. In this presentation, I introduce a web-based software, PubMed Sentence Extractor (PSE), which parses huge PubMed abstracts, extracts and displays the cooccurrence sentences of gene names and other keywords. I believe that PSE is helpful to collect necessary literatures efficiently in order to find research targets.

2 Method and Results
2.1 Keyword Extraction
24,126 PubMed abstracts were retrieved using "result", which was postulated not to be biased in particular fields of articles, as a keyword. The score for each term was defined with the maximum value of the term frequency – inverse document frequency (tf-idf) score of all abstracts. The words included more than 0.1% of abstracts and of which the tf-idf scores were less than 0.1 were saved. Next, the words included more than 1% of abstracts and of which the tf-idf scores were more than 0.1 were listed. The words suited for keywords in the latter list were removed, manually, and the remaining list was saved. The two lists were merged and saved as the stopword list (SL1). SL1 contains 8,814 words at last. SL1 is used for both creation of a gene name dictionary and common word extraction. For the actual process, each word in abstracts is extracted and compared with SL1. If the word is not included in SL1, the occurrence is counted, recorded and displayed as a pull-down list. Twenty to twenty-five percents of total words are selected as keywords with SL1.

2.2 Compilation of Gene Name Dictionary
Gene name information was extracted from EntrezGene records [1]. The records having an official symbol and originating from human, mouse and rat were used. Aliases and symbols were extracted from each record and a series of aliases were created based on several rules, for example: (a) add/remove hyphen/space, convert Arabic number into Roman number if the name is ended by number (e.g. Akt-1 => Akt1, Akt 1, AktI, Akt-I, Akt 1), (b) trim the name if it contains some specific words (e.g. one of the words is "isoform", then "fibulin 1 isoform A" => "fibulin 1"). Each word or phrase was compared with SL1 and an after-mentioned stopword list (SL2), which was compiled only for gene name extraction, and the word or phrase, which did not match with one or both of them, was added as an alias after converted into uppercase. The homonymous gene names of common words, e.g. yes, generate many false positives. Therefore, elimination of homonyms for common words is important. Three sets of abstracts containing around 1,500 records were retrieved using
“cancer AND kinase”, "autoimmune" and "hypertension" as keywords, respectively, and the candidate words of gene names, which contained false positives, were marked. Browsing the results, I created two stopword lists, an ordinary stopword list (SL2) and an adjacent pattern list, (SL3).

### 2.3 Gene Name Extraction

Each abstract is split into sentences. If a word is in parenthesis or located between two commas, the adjacent word is compared with SL3. This comparison is carried out by a backward match manner. For example, "OSUS" is in SL3 and let a phrase, "systematic lupus erythematosus (SLE) ", is found in a sentence. Because the terminus of "erythematosus" matches with "OSUS", the word, "SLE", is ruled out of candidates of gene names. This process is effective to remove many acronyms of chemical compounds, disease names and so on. The exclusion using SL3 is applied only to each abstract, but those using SL1 and SL2 are applied to all records. After the extraction of stopwords, gene name extraction is carried out. Starting with the top of the sentence, 6 continuous words are compared with the dictionary and are phased out if the phrase does not match. All letters are converted into uppercase prior to the comparison in case of more than two words. To improve the precision, following different rules are applied to single word: (a) a word containing non-alphabetical letters or more than two uppercase letters at the beginning of the sentence, or (b) a word containing other than lowercase alphabets and not locating at the top of the sentence. If the word matches with one of the rules, it is converted into uppercase and compared with the dictionary. Therefore, PSE cannot detect single words containing only lowercase as gene names. The sensitivity of this system was evaluated using randomly selected two sets of forty abstracts from whole PubMed records at first. The precision, recall and F-measure were 77-78%, 45-66% and 57-71%, respectively. In order to collect some extent of enriched abstracts of mammalian gene names, I retrieved two sets of forty abstracts using following keywords, "gene AND disease AND activation". The precision, recall and F-measure were 86-93%, 58-65% and 69-76%, respectively. Homonymy was not considered in this evaluation.

### 2.4 System Implementation

PSE is written with Perl and PHP, and runs on UNIX family OS. User retrieves PubMed abstracts with xml format and registers them into the system. In registration process, PSE splits the whole abstract into sentences, assigns the sentence IDs, counts keywords, searches gene names and records the sentence IDs for each gene symbol. Users select a gene symbol and a common word from the pull-down menus, and PSE displays the extracted sentences and related information. Because these sentences and information are linked to the whole abstracts and the external web sites, EntrezGene, Refseq [4] and OMIM [2], users can get more information on the interested genes and sentences. Although PSE has no rule of assigning homonyms, it can display the list of them in case of incorrect assignments, and using the list, users select the correct symbol and get information on the correct gene. This program also contains a simple text search function. Therefore, users can search cooccurrence sentences with free keywords in the registered abstracts.

### 3 Discussions

I developed PSE in order to extract useful records focusing on gene information efficiently from PubMed. This system has following four advantages; the reduction of the amount of records, the showing of keyword lists, the consideration of the synonyms and the variations of gene names, and the links to external databases. In previous version of PSE [5], I used only the dictionary to extract gene names, but the dictionary and additional rules, SL1, SL2 and SL3, were used for the process in new version PSE. Because the precision was significantly improved from 48% to 77% by this modification, the addition of other rules was quite effective. Assignment of homonyms is another important issue and is not regarded in this algorithm. I would like to adopt some algorithm to this problem as the future plan.

### References