# Classifying Biomedical Literature Providing Protein Function Evidence

Joon-Ho Lim and Kyu-Chul Lee

**Because protein is a primary element responsible for biological or biochemical roles in living bodies, protein function is the core and basis information for biomedical studies. However, recent advances in bio technologies have created an explosive increase in the amount of published literature; therefore, biomedical researchers have a hard time finding needed protein function information. In this paper, a classification system for biomedical literature providing protein function evidence is proposed. Note that, despite our best efforts, we have been unable to find previous studies on the proposed issue. To classify papers based on protein function evidence, we should consider whether the main claim of a paper is to assert a protein function. We, therefore, propose two novel features — protein and assertion. Our experimental results show a classification performance with 71.89% precision, 90.0% recall, and a 79.94% F-measure. In addition, to verify the usefulness of the proposed classification system, two case study applications are investigated — information retrieval for protein function and automatic summarization for protein function text. It is shown that the proposed classification system can be successfully applied to these applications.**

**Keywords: Information retrieval, document classification, protein function evidence, automatic text summarization.**

Manuscript received Jan. 29, 2014; revised May 19, 2015; accepted June 9, 2015.

This work was supported by the Industrial Strategic Technology Development program (10035197) funded by the Ministry of Knowledge Economy (MKE, Rep. of Korea).

Joon-Ho Lim (joonho.lim@etri.re.kr) is with the SW & Contents Research Laboratory, ETRI, Daejeon, Rep. of Korea.

Kyu-Chul Lee (corresponding author, kclee@cnu.ac.kr) is with the Department of Computer Science & Engineering, Chungnam National University, Daejeon, Rep. of Korea.

## I. Introduction

According to the central dogma of biology, the genome in DNA is translated into protein through a messenger RNA. Protein is in charge of biological or biochemical roles, and is a primary element of the function of living bodies [1]. Therefore, many biomedical researchers have made their best efforts to find new protein functions using all types of available methods, such as homology-based analysis, sequence motif analysis, 3D protein structure analysis, and association network analysis [2]. In addition, based on known protein function information, researches including new drug development and other areas have been conducted. Based on this, protein function information is the core and basis information that many researchers need to find and manage.

However, because of recent advances in computational bioinformatics technologies, the number of published biomedical papers has increased explosively. According to MEDLINE fact sheets, the MEDLINE literature database contains over 19 million references, and 2,000 to 4,000 new papers are being added daily.[1)] This explosive increase in the amount of literature has made it difficult for biomedical researchers to find necessary protein function information.

To solve this difficulty, PubMed, which is the most widely used biomedical literature search engine, provides a separate literature list for protein functions. For example, if a query string of "MICU1" is entered into PubMed, it provides a separate list of papers previously input into the database. In addition, Swiss-Prot, which is the most widely used protein database in the world, provides protein function information and their evidences, as shown in Fig. 1, which are manually

1) http://www.nlm.nih.gov/pubs/factsheets/medline.html

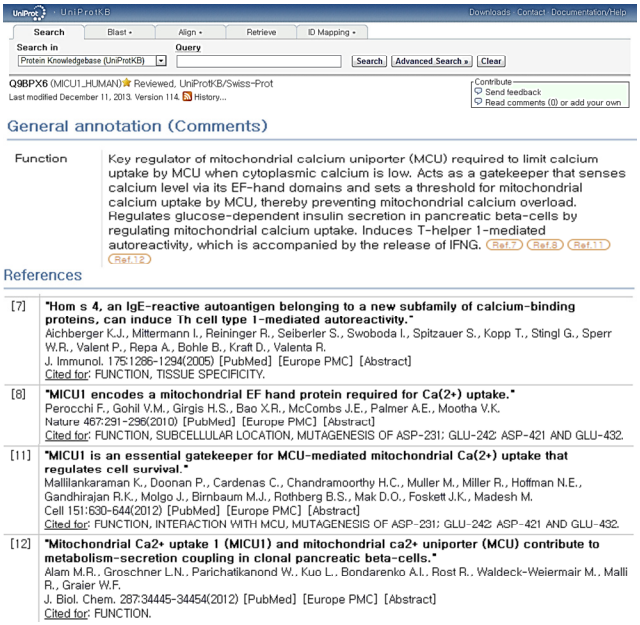


Fig. 1. Example of MICU1_HUMAN entry in Swiss-Prot, which provides function annotation and supporting references.

annotated by biomedical domain experts.

The literature providing protein function evidence is made up of papers such as reference papers for the protein function of Swiss-Prot and papers separately provided by PubMed. The literature does not simply mention previously identified protein functions, rather it provides evidence for such protein functions, such as experimental and sequence analysis results. In general, papers that first identify and assert a particular function of a target protein are papers that provide protein function evidence.

However, because PubMed and Swiss-Prot provide only the inputted literature previously registered in the database for protein functions, they inevitably show only limited information, which is much less than the actual information revealed through papers that can be found through PubMed. That is, because these services cannot provide papers that have not been reviewed when annotating the database or papers with new recently revealed research results, biomedical researchers are constantly searching the literature associated with their research, and spend a lot of time and effort to find new function information that they previously missed. Therefore, an automatic classification study on finding biomedical literature providing protein function evidence from a huge set of documents is valuable and can provide literature information that previously could not be retrieved using PubMed and Swiss-Prot.

However, classifying whether a biomedical document contains protein function evidence is a difficult problem. For this problem, it is not sufficient to apply a typical conventional classification method, which determines the class of a document by determining whether a protein function is mentioned in a document. To be classified as such, the main claim of the paper should be that it is providing a protein function. That is, for a correct classification, the main claim should be separated from comments on existing research and functional words ordinarily used, such as protein names. Thus, in addition to existing document classification methods, an additional feature study on classifying only literature providing protein function evidence is needed.

In this paper, we propose document classification research to classify biomedical papers based on protein function evidence. To solve this problem, we automatically construct a document classification corpus using protein function evidence information in Swiss-Prot; in addition, we propose two novel features — protein and assertion. We then show the effectiveness of the proposed system using two case study applications — information retrieval and automatic text summarization for protein functions.

## II. Related Work

In this section, we introduce numerous studies on biomedical literature classification. The difference between literature classification of the biomedical domain and non-biomedical domain is that the classification standard varies greatly depending on the purpose. For example, if the purpose is the construction of a protein–protein interaction (PPI) network, then the literature is classified according to the inclusion of PPI information. In addition, if the purpose is a specific disease, then the literature is classified according to the specific disease.

First, studies classifying literature including PPI information are as follows [3]–[7]. Kolchinsky and others [3] used features of word-pair and protein mentions, as well as proposing a linear classifier based on term-relevance to obtain a high performance. Chen and others [4]–[5] proposed a semi-supervised self-training method to use large unlabeled articles in [4] and a feature selection method based on context similarity in [5]. Garcia and others [6] extracted unigram, bigram, and trigram features of high information gain and classified the literature based on these extracted features. Matos and Oliveira [7] proposed a vector-space classification method, which retrieves similar articles to a target article and then classifies the retrieved articles using assigned scores.

For the TREC Genomics Track 2005 Categorization Task, [8] and [9] classify biomedical papers according to alleles of mutant phenotypes, embryologic gene expression, gene ontology, and tumor biology. Li and others [8] proposed a meta-classification method, which separately classifies full-text, abstracts, and MeSH ontology information using an SVM and then combines the classification results using logistic

regression. Cohen [9] performed repeated resampling of the training data to train multiple SVM classifiers; he then determined the test document category using these multiple classifiers.

The studies classifying biomedical literatures according to disease categories, such as arrhythmias, coronary heart, and cholangiocarcinoma diseases, are as follows [10]–[11]. Dollah and Aono [10] classified a document by performing ontology alignment between a hierarchy of extracted MeSH keywords and the OHSUMED disease hierarchy. Sibunruang and Polpinij [11] proposed a cholangiocarcinoma document classification method using the Cancer Technical Term Net Ontology.

Other kinds of biomedical literature classification studies are as follows. Polavarapu and others [12] studied biomedical document classification for human genome epidemiological research using the CDC HuGENet database. Krallinger and others [13] classified biomedical literature using an SVM for cell cycle information. In addition, as part of the BioCaster project, Conway and others [14] carried out a classification for biomedical news that is of interest to health professionals.

The previous works described above mainly classify the biomedical literature according to whether papers contain PPI information, disease categories, and so on. Despite our best efforts, a biomedical literature classification for providing protein function evidence has yet to be found.

Among previous studies, the most relevant work with the proposed study is classification according to PPI information. However, they are completely different in purpose and scope. The purpose of such PPI document classification is two-fold in that it is to extract PPI relationships using an existing relation extraction method [15]–[17] and to construct a PPI network to find candidate proteins that affect the target protein directly or indirectly. However, a protein function is how a protein works in an organism, and requires much broader information including the PPI. For example, in the case of the MICU1 protein in Fig. 1, document classification for PPI classifies documents only according to whether they contain *MICU1-regualte-MCU* information. However, document classification for protein function evidence considers much more general information, such as when MICU1 protein regulates MCU protein (that is, when cytoplasmic calcium is low); how MICU1 protein regulates MCU protein (that is, senses the calcium level through its EF-hand domains and sets a threshold); and what effects are allowed through this regulation (that is, preventing mitochondrial calcium overload, regulating glucose-dependent insulin secretion in pancreatic beta-cells, and inducing T-helper 1-mediated auto-reactivity).

If the proposed literature classification for protein function evidence is effectively solved, then it enables many applications for protein functions, such as information retrieval, automatic text summarization, and database curation [18]–[19].

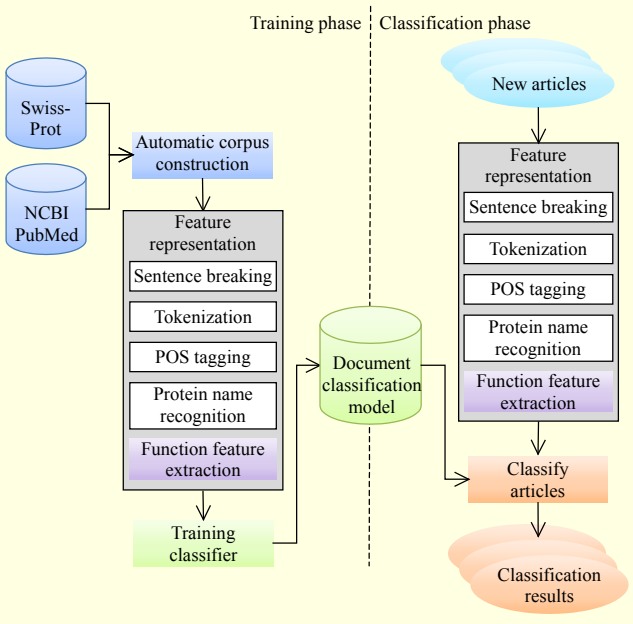


Fig. 2. Overall system architecture.

## III. Classifying Biomedical Literature Providing Protein Function Evidence

The overall system architecture is shown in Fig. 2. During the training phase, a corpus for document classification is first automatically constructed. After the collected documents are represented as feature vectors through natural language processing (NLP) pipelines, a document classification model is then trained using a machine learning algorithm.

During the classification phase, new test articles are classified by whether they provide protein function evidence. The input articles are represented as feature vectors using the same method as in the training phase and are classified using the trained document classification model.

### 1. Automatic Corpus Construction

The Swiss-Prot database is a high-quality, non-redundant, and manually annotated protein database [20]. It provides function annotation information and the supporting scientific literature for each protein, as shown in Fig. 1. To annotate the protein function information, reviewers and biomedical domain experts identify relevant papers by searching through literature databases such as PubMed and using literature mining tools. The full text of each paper is reviewed, and all experimental findings are compared with both the current knowledge on the related proteins and the sequence analysis results. The new findings captured from the scientific literature

Table 1. Simplified XML structure of Swiss-Prot.

```
<entry dataset="Swiss-Prot" created="2008-03-18Z" …>
  <name>MICU1_HUMAN</name> …
  <comment type="function" evidence="1 2 3 4">
    <text> Key regulator of mitochondrial calcium uniporter (MCU)
           required to limit calcium uptake by MCU when cytoplasmic
           calcium is low. …</text>
  </comment> …
  <evidence key="1" type="ECO:0000006">
    <source>
      <dbReference type="PubMed" id="16002733"/>
    </source>
  </evidence> …
</entry>
```

are then added to the protein entry.[2)]

To construct a document classification corpus, a protein function annotation and its evidence information are used. The XML structure of Swiss-Prot is shown in Table 1. First, all PubMed reference IDs (PMIDs) and their comment type are extracted for each protein entry. An extracted PMID is then classified according to whether it is referred to as a function type. Because all enrolled papers are carefully reviewed and annotated by biomedical domain experts, if a paper is referred to as a function type at least once, then it definitely includes protein function evidence. On the other hand, if a paper is not referred to as a function type after an expert review, then it is assumed that the literature will not include protein function evidence. As a result, the total number of articles in the Swiss-Prot database is 79,276, of which 44,021 articles are labeled as providing protein function evidence, and 35,255 articles are not. Lastly, all articles are fetched using the PubMed E-utilities API,[3)] and their title and abstract texts are then extracted.

## 2. Document Classification

As a classification method, the maximum entropy (ME) model is used [21], which has been successfully applied to many biomedical text mining tasks. It has a good performance with a fast speed [22]. In the ME model, the conditional probability of class $y$ given a feature vector $x$ is defined as follows:

$$p(y \mid x) = \frac{1}{Z(x)} \exp\left[\sum_{i=1}^{k} \lambda_i f_i(x, y)\right],$$

where $f_i(x, y)$ is the $i$th lexical feature function, $\lambda_i$ is the weighting parameter of $f_i(x, y)$, $k$ is the number of features, and $Z(x)$ is the normalization factor for $\Sigma y\ p(y|x) = 1$. The ME model has an advantage in that the relative importance of each feature can be analyzed, because it has weight $\lambda_i$ for each feature.

## 3. Feature Representation

In this section, an input document is converted into a feature set for classification. For the conversion, a basic NLP pipeline of sentence breaking, tokenization, and part-of-speech (POS) tagging is performed [23]. Using the NLP pipeline result, a feature set is extracted.

To correctly classify the literature, we need to know the lexicon used to represent the protein function. The lexicon for the protein function can be found using the ME model. Each document is represented as a bag-of-word feature, and the ME training algorithm then automatically learns the appropriate weights of each lexical feature. At this time, the lexical features representing a protein function have higher weights. The results of the feature importance will be shown in the experiment section.

It should then be distinguished whether the lexical features representing protein function are used for the main contribution of the research paper. To capture these cases, we propose two novel features — protein and assertion. A *protein* feature distinguishes an ordinarily used functional lexicon in the protein name, which is used regardless of the contribution of the paper. In addition, an *assertion* feature helps to distinguish whether the main argument of the paper contains a lexicon expressing protein function.

First, the protein feature separates lexicons used normally and as a protein name, because a lexicon used for a protein name is ordinarily used for an indication of a particular protein and has little relation to the main contribution of the paper. Table 2 shows title texts of existing research papers retrieved from PubMed, and it shows examples of a lexicon used normally and one used as a protein name. In the case of normal usage, the first example represents a *glutamate uptake* function, and the second example indicates a function of *binding between BST-2 and cellular MT1-MMP*. However, in the case of the protein name, mentions of *iron uptake protein* and *calcium-binding protein* are used for an indication of particular proteins and have little relationship with the main contents of the paper, which are marked as underlined.

To recognize protein names, we use the GENIA named entity tagger. Using the result of protein name recognition, we propose three separation methods: firstly, the *protein.lexical* feature separates the protein lexicon from the normal sentence lexicon by adding a "P_" prefix. For example, it affects whether the *uptak* feature, used as a single feature, is separated

2) http://www.uniprot.org/faq/45
3) http://www.ncbi.nlm.nih.gov/books/NBK25500/

Table 2. Examples of lexicons used for normal usage and protein name.

| Usage | Title text (retrieved from PubMed) |
|---|---|
| Normal usage | High glucose stimulates glutamate **uptakes** in pancreatic β-cells. (PMID: 22232641) |
| | BST-2 **binding** with cellular MT1-MMP blocks cell growth and migration via decreasing MMP2 activity. (PMID: 22065321) |
| Protein name | Bacillus cereus iron **uptake** protein fishes out an unstable ferric citrate trimer. (PMID: 23027976) |
| | Hom s 4, an IgE-reactive autoantigen belonging to a new subfamily of calcium-**binding** proteins, can induce Th cell type 1-mediated autoreactivity. (PMID: 16002733) |

Table 3. Detailed example of protein features.

| Input sentence | It requires [a calcium uptake protein 1]$_{PROTEIN}$, ... |
|---|---|
| *protein.lexical* | *requir P_calcium P_uptak P_protein P_1 ...* |
| *protein.tag* | *requir PROT* |
| *protein.tag lexical* | *requir PROT P_calcium P_uptak P_protein P_1 ...* |

as *uptak* or *P_uptak* based on the protein name. Second, the *protein.tag* feature replaces the protein lexicon with a protein tag. This has a generalization effect, which diminishes the high sparseness of the protein lexicon. Finally, the *protien.tag.lexical* feature is a combination of the *protein.tag* and *protein.lexical* features. A detailed example is shown in Table 3. Note that this approach to a protein feature is compared with a study by Li and others [8], whereby they regarded a protein name as a single token, such as *10-kD*skeletal*extracellular*matrix*protein*, and used lexical information in the protein name.

Second, we propose an assertion feature, which is a sentence related with the main argument of the paper. Generally, a research paper reveals previously unknown facts. An assertion sentence is a clearly described sentence of the main contribution of the paper. Examples of an assertion sentence are as follows:

- "We have identified a new member of the TGF-beta superfamily, CET-1, …"
- "In the present study, we demonstrate that upon engagement of the T cell receptor (TCR), …"
- "Our results show that cet-1 controls diverse biological processes …"
- "These results suggest that Xin may participate in a BMP-Nkx2.5-MEF2C pathway to control cardiac …"

Because whether the literature provides evidence of a protein function is determined by whether the main contribution of the paper is related with a protein function, the assertion sentences of the paper are crucial features for classifying literature providing protein function evidence. An algorithm to extract an assertion feature is shown in Table 4.

Table 4. Assertion feature extraction algorithm.

Document = $\{S_1, S_2, S_3, \ldots, S_n\}$
$S_i = \{w_1, w_2, w_3, \ldots, w_m\}$
AF[] = { }

For $i = 1, \ldots, n$ that
 For $j = 1, \ldots, m$ that
  If ($w_j$.lexical = "we") then AF = AF ∪ $S_i$
  If ($w_j$.lexical = "our") then AF = AF ∪ $S_i$
  If ($w_j$.lexical ∈ {"result," "results"}) then
   If ($w_{(j-1)}$.lexical ∈ {"the," "this," "these"}) then AF = AF ∪ $S_i$
  If ($w_j$.lexical = "paper") then
   If ($w_{(j-1)}$.lexical = "this") then AF = AF ∪ $S_i$
  If ($w_j$.lexical = "paper") then
   If ($\forall k \in \{1, \ldots, j-1\}$: $w_k$.pos != "verb") then AF = AF ∪ $S_i$

Return AF

The assertion sentence extraction algorithm proposed in this paper is relatively simple. Note that it is consistent with a tendency in which, if an author writes *we*, *this paper*, or *these results* in the abstract of the paper, as shown in the above examples, the author will often explain what they have studied, the meaning of the experimental results, and so on. In addition, the proposed algorithm is designed for an extraction of obvious assertion sentences with minimal error rather than extracting every possible sentence without exception. To verify the extraction algorithm, the proposed algorithm is applied to the training data, and we make sure that it effectively extracts assertion sentences in a consistent manner.

## IV. Experimental Results

For the experimental data, Swiss-Prot version 2012-07 is used. The dataset is randomly divided into a training set (90%) and a test set (10%). The numbers of the training and test instances are 71,434 and 7,842, respectively, and the accuracy of random guessing is 55.03%. To evaluate the performance of document classification, we use the precision, recall, and F-measure. For a text preprocessor and protein name recognizer, we use a GENIA named entity tagger.[4] In addition, the Maxent toolkit[5] and L-BFGS parameter estimation algorithm are used.

### 1. Experiments of Individual Features

In this section, we show detailed feature representation methods of three individual features and experimental results.

4) http://www.nactem.ac.uk/GENIA/tagger/
5) http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

First, a basic feature represents basic text processing results such as tokenization, normalization, and POS tagging. More specifically, a *basic.bow* feature is a bag-of-words of the tokenization results of a GENIA tagger with stop-words removed. The *basic.normalize* feature is the lexical normalization, which uses Porter Stemmer and separates symbol-connected tokens. In addition, the classification capability of each POS tag is analyzed, and it is shown that only three POS tags, *noun*, *verb*, and *adjective*, have significant effects among nineteen POS tags. A *pos.adj_noun_verb* filters out only three POS tags from the *basic.normalize* feature.

The experimental results of the basic features are shown in Table 5, and *basic.bow* provides a performance of the baseline system. The *basic.normalize* improves the recall performance much more than *basic.bow*. The performance of the *basic.normalize* feature indicates that researchers can find 80.17% of the correct literature with 76.68% precision. In addition, the *pos.adj_noun_verb* experiment shows that using only three POS tags can achieve the same level of performance as using all nineteen POS tags.

As the protein features, three protein features — *protein.lexical*, *protein.tag*, and *protein.tag.lexical* — were tested. The experimental results of the protein features are shown in Table 5. Because the *protein.tag* feature generalizes the sparseness of the protein lexicon, the recall is increased to 81.24% from 80.17% of the *basic.normalize* feature. In addition, the usage of both the *protein.tag* and *protein.lexical* features increases the precision and recall, and this combination shows the best performance.

In the assertion feature experiment, a classification using only assertion sentences was tested to determine a representation method of an assertion sentence for the combination phase. First, the *assert.normalize* feature is the assertion sentence feature applied with *basic.normalize* of the basic feature. Second, because assertion sentences of each paper have variations in their length, it is assumed that all documents have the same weight, and each feature of *assert.normalize.weight* is assigned weights based on the document length. Third, the *assert.protein.tag.lexical* feature applies the *protein.tag.lexical* feature to protein names in the assertion sentences.

In Table 5, because assertion features use only assertion sentences, each document has a short length, which causes a low precision. From the *assert.normalize.weight* feature experiment, document-length weighting is effective when documents have differences in their length. In addition, the *assert.protein.tag.lexical* feature also has a good performance in terms of precision and recall, and is slightly better than the *assert.normalize.weight* feature. Lastly, a combination of the weighting and *assert.protein.tag.lexical* feature can be considered. However, this combination method assigns more weights to the recognized protein names because a weight is assigned to the *PROT* tag and *P_*protein name lexicon, separately. Actually, if a weighting method is applied to the *assert.protien.tag.lexical* feature, the precision and recall are slightly decreased, and consequentially, the F-measure is decreased by about 0.5%.

Table 5. Experimental results of individual features.

| | Precision | Recall | F-measure |
|---|---|---|---|
| *basic.bow* | 77.21% | 76.87% | 77.04% |
| *basic.normalize* | 76.68% | 80.17% | 78.39% |
| *pos.adj_noun_verb* | 77.30% | 79.51% | 78.39% |
| *protein.lexical* | 76.47% | 80.71% | 78.54% |
| *protein.tag* | 76.00% | 81.24% | 78.53% |
| *protein.tag.lexical* | 76.99% | 81.01% | 78.95% |
| *assert.normalize* | 65.65% | 79.76% | 72.02% |
| *assert.normalize.weight* | 65.42% | 82.94% | 73.14% |
| *assert.protein.tag.lexical* | 65.74% | 82.89% | 73.32% |

## 2. Experiments of Combinational Features

In this section, various combination methods for previous individual features are tested. First, a document is classified into three types — the title sentence, assertion sentence, and non-assertion sentence — and the combination methods of these sentence types are then tested. Second, to improve the recall performance, a cut-off experiment is performed based on the output conditional probability of the ME model. Improving the recall performance is important for the following reason: a document classified as a false-positive may be re-examined by the user or system in the next phase. However, a document classified as a false-negative is not provided to the next phase and has no chance of re-examination. The risk of a false-negative case is therefore larger.

Based on previous experimental results, all of the combinational features apply the *protein.tag.lexical* of the protein features.

First, the *combination.T-N-N* feature appends *T* to the title and *N* to the assertion sentences and non-assertion abstract. This has an effect in that the lexicons in the title and abstract are separated as being different, and thus they have a different $f_i(x, y)$ and weight $\lambda_i$. Similarly, the *combination.T-A-N* feature appends *T* to the title; *A* to the assertion sentence; and *N* to the non-assertion abstract. In addition, the *combination.T-T-N* feature combines only the lexicon of the title and assertion sentences, and separates the non-assertion abstract; in addition, the *combination.N-A-N* feature separates only assertion


http://dx.doi.org/10.4218/etrij.15.0114.0041

Table 6. Experimental results of combinational features.

| | Precision | Recall | F-measure |
|---|---|---|---|
| *combination.T-N-N* | 77.14% | 81.10% | 79.07% |
| *combination.T-A-N* | 76.90% | 81.58% | 79.17% |
| *combination.T-T-N* | 75.89% | 83.17% | 79.36% |
| *combination.N-A-N* | 76.12% | 81.60% | 78.76% |

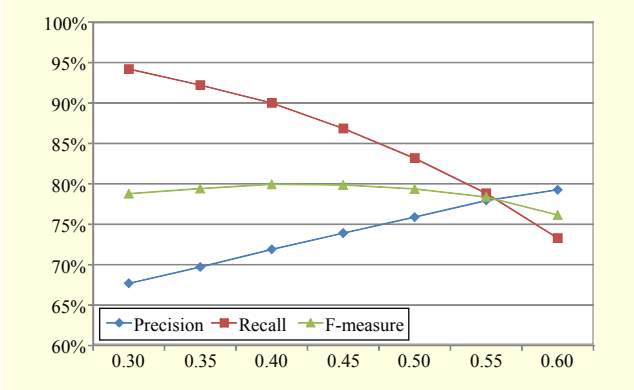


Fig. 3. Performance results of probability cut-off experiment.

sentences.

In Table 6, the performances of *combination.T-A-N* and *combination.T-T-N* are higher than *combination.T-N-N*, which means that the separation of an assertion sentence from the abstract is effective. In addition, the title sentence and assertion sentences have a similar nature. This leads to a result in which the combination of these two sentences, *combination.T-T-N*, achieves the best performance.

Second, to improve the recall performance, a probability cut-off experiment is performed. In this experiment, if $p(y = 1 \mid x) \geq$ cut-off, then the paper is then classified as including a protein function. Figure 3 shows the experimental results of the cut-off using the *combination.T-T-N* feature. By increasing the cut-off, the recall decreases faster than the precision increase. When the cut-off is 0.4, it achieves the best performance with a 71.89% precision, 90.0% recall, and 79.94% F-measure.

### 3. Feature Analysis

In this section, we analyze the model trained using the ME learning algorithm and then analyze the important features of the higher weights for each experiment [24]. Note that the conditional probability in the ME model contains $Z(x)$ in the denominator; thus, the relative weights in a single model are only compared, and the absolute values of the weights are not considered.

First, it is analyzed whether the major features for the document classification are lexicons representing a protein function. The trained model of the *basic.normalize* experiment is used to analyze the features. The "Basic feature" column of Table 7 shows the ten features among the top features and their weights. Note that the lexicons in Table 7 are stemmed and normalized. The analysis results show that the lexicons for a protein function are used as major features to classify a document providing protein function evidence. For example, major features used for protein function include *knockdown*, *uptak*, *act*, *transfect*, *autophagi*, *movement*, *synthet*, and *inject*.

Second, through a protein feature analysis, the weight difference between whether the same lexicon is used normally or as a protein name is analyzed. For the feature analysis, the trained model of the *protein.tag.lexical* experiment is used. The analysis result is shown in the "Protein feature" column of Table 7, which shows the ten features among the top features and their weights for normal usage and as a protein name, respectively. The analysis results show that lexicons such as *act* and *knockdown* have a significant weight difference between normal usage and protein name usage. Moreover, in the case of lexicons such as *overexpress* and *reduc*(*e*), if they are used normally, then they increase the probability for a positive class, but if they are used as a protein name, then they decrease the probability, reversely. In this paper, we propose the hypothesis that if a lexicon is used for a protein name, then it has different importance with normal usage, and this analysis result is evidence for its correctness.

Finally, to show the effectiveness of the separation of assertion sentences from other non-assertion sentences, the *combination.T-A-N* model is analyzed, and features that become important by separating an assertion sentence are investigated. The "Assertion feature" column of table 7 shows the feature analysis results. According to the results, a lexicon such as *act* is equally important regardless of whether it is used as an assertion feature or a non-assertion feature. However, lexicons such as *requir*(*e*), *function*, *role*, *promot*(*e*), and *control* have higher weights and are more important when used as an assertion feature than a non-assertion feature. Actually, although the two different training models are difficult to compare, we confirm that the assertion features in Table 7 have relatively low weights from the model of the *combination.T-A-A* experiment. This weight difference effect is caused by the proposed assertion feature, and is evidence for the effectiveness of the assertion feature.

## V. Two Case-Study Applications

### 1. Information Retrieval for Protein Function

Generally, an information retrieval system including

Table 7. Feature analysis results.

| Basic feature | | Protein feature | | | | Assertion feature | | | |
|---|---|---|---|---|---|---|---|---|---|
| knockdown | 0.6269036 | act | 0.4223537 | P_act | 0.0201349 | A_requir | 0.2948434 | N_requir | 0.0261408 |
| uptak | 0.3721180 | knockdown | 0.3218143 | P_knockdown | 0.0028082 | A_act | 0.2921403 | N_act | 0.3374955 |
| belong | 0.3638166 | overexpress | 0.2791324 | P_overexpress | −0.0050674 | A_function | 0.2393178 | N_function | 0.0512650 |
| act | 0.3605847 | uptak | 0.2425000 | P_uptak | 0.0080121 | A_role | 0.2264551 | N_role | 0.1079808 |
| transfect | 0.3501983 | reduc | 0.2336003 | P_reduc | −0.0049258 | A_promot | 0.1983949 | N_promot | −0.0393265 |
| autophagi | 0.3409845 | exhibit | 0.2309426 | P_exhibit | 0.0051353 | A_control | 0.1686988 | N_control | 0.0949963 |
| repressor | 0.3376251 | prolifer | 0.2303643 | P_prolifer | −0.0129291 | A_mediat | 0.1562042 | N_mediat | −0.0107241 |
| movement | 0.3194586 | function | 0.2240967 | P_function | −0.0122153 | A_induc | 0.1459353 | N_induc | 0.0314095 |
| synthet | 0.3008330 | antimicrobi | 0.2186139 | P_antimicrobi | 0.0250351 | A_encod | 0.1404075 | N_encod | 0.0105218 |
| inject | 0.2983965 | suppress | 0.2174351 | P_suppress | −0.0043959 | A_essenti | 0.1385595 | N_essenti | −0.0019322 |

PubMed provides search results when query words inputted by the user and the words in the literature are matched. Therefore, problems often occur in that the contents of the retrieved document are not related with the intention of the user.

As one solution to this problem, the proposed literature classification system can be used. That is, if the user's intention is to find the protein function information, the proposed classification system is applied to the literature search results. Then, only papers classified as providing protein function evidence are provided to the user.

To measure the reducing ratio of the proposed system for a protein function as compared with the existing PubMed system, a new metric of *reducing-ratio* (of search result) is used.

$$\text{Reducing-ratio} = \frac{\text{false-negative} + \text{true-negative}}{\text{total} - \text{instance}}.$$

For the experiments, three proteins — *CD177 antigen*, *Sequestosome-1 Protein*, and *Calcium uptake protein 1, mitochondrial* (*MICU1*) — are used.

First, the *CD177 antigen* uses a query of (CD177 antigen), and a total of 152 papers are searched. After classification, a 69.74% reducing ratio is shown. Because not a lot of studies have been carried out to identify the function of this protein, it is likely that a small amount of the literature is classified as providing protein function evidence.

Second, *Sequestosome-1 Protein* uses a query of (Sequestosome-1[Title]) OR SQSTM1[Title], and a total of 160 papers are searched. After classification, it achieves a reducing ratio of 40.0%. Among the top papers retrieved from PubMed, Table 8 shows examples of papers classified as providing protein function evidence or not.

Finally, *Calcium uptake protein 1, mitochondrial* (*MICU1*) uses a query of (*Calcium uptake protein 1, mitochondrial*) or

Table 8. Examples of retrieved papers for *Sequestosome-1 Protein*.

| PMID | Title text (retrieved from PubMed) | $p(y=1\|x)$ | Result |
|---|---|---|---|
| 24291777 | Heat Shock Factor 1 Confers Resistance to Hsp90 Inhibitors through p62/SQSTM1 Expression and Promotion of Autophagic Flux. | 0.9670 | +1 |
| 24270048 | TRAF6-mediated ubiquitination of NEMO requires p62/Sequestosome-1. | 0.7063 | |
| 24240628 | Overexpression of p62/SQSTM1 promotes the degradations of abnormally accumulated PrP mutants in cytoplasm and relieves the associated cytotoxicities via autophagy-lysosome-dependent way. | 0.6768 | |
| 24121507 | p62/Sequestosome-1 Up-regulation Promotes ABT-263-induced Caspase-8 Aggregation/Activation on the Autophagosome. | 0.9743 | |
| 24065390 | Knockdown of p62/Sequestosome 1 attenuates autophagy and inhibits colorectal cancer cell growth. | 0.5732 | |
| 24138988 | SQSTM1 mutations in Han Chinese populations with sporadic amyotrophic lateral sclerosis. | 0.0017 | −1 |
| 24086340 | Association of p62/SQSTM1 Excess and Oral Carcinogenesis. | 0.3162 | |
| 24042580 | SQSTM1 Mutations in French Patients With Frontotemporal Dementia or Frontotemporal Dementia With Amyotrophic Lateral Sclerosis. | 0.0012 | |
| 23812289 | Sporadic ALS with compound heterozygous mutations in the SQSTM1 gene. | 0.1042 | |
| 23658060 | Common susceptibility alleles and SQSTM1 mutations predict disease extent and severity in a multinational study of patients with Paget's disease. | 0.0049 | |

(*Atopy-related autoantigen CALC*) or *MICU1* and filters the publication date to between 1900/01/01 and 2012/12/31. The total search results are 805 papers. Because MICU1 protein is an important protein associated with cell apoptosis and atopic disease, numerous studies on identifying the function of MICU1 protein have been conducted. After classification, only a 9.81% reducing ratio is achieved. As shown in the above experiment results, the reducing ratio has a high variation according to how many studies have been performed related to the functions of the target protein

Finally, we investigate the average reducing ratio of the proposed retrieval system. To measure the average ratio, we use an assumption generally used in the information retrieval domain; that is, that all documents have the same prior probability of being retrieved. For the experiment, the *combination.T-T-N* system with a 0.4 probability cut-off is applied to the test data in the experiments section. The resulting reducing ratio on the test data is 30.49%, which means that a researcher will receive 30.49% fewer papers on average.

## 2. Automatic Text Summarization for Protein Function

Like Swiss-Prot, which provides a function annotation text for each protein, if a summary of protein functions is automatically generated based on the literature classified as providing protein function evidence, biomedical researchers can quickly learn the necessary protein function information from a large amount of literature. The text summarization technique has been widely studied in the NLP domain, and is mainly studied to automatically generate news summaries containing the key phrases of the contents from extensive numbers of news articles. In this paper, the mead automatic summarization library, which is most widely used in the summarization domain, is used [25].

First, to verify the effectiveness of the automatic summarization for protein function, the generated summary is compared with the gold-standard summary. MICU1 protein in Swiss-Prot provides four documents as evidence of protein function, as shown in Fig. 1. The MEAD summarization results using these four documents and annotation text manually inputted by human experts are compared. As a result, there is a significant amount of overlapping information in the automatically generated summary with the gold-standard summary, such as *MICU1 is an essential gatekeeper for MCU-mediated mitochondrial Ca 2 uptake*, *it regulates insulin secretion in pancreatic beta-cells*, *it induces T-helper 1- mediated autoreactivity*, and so on.[6] This result shows that a summary text for protein function can be successfully generated using automatic summarization techniques.

Table 9. Automatic summarization results for *Sequestosome-1 Protein* function using abstracts of positively classified papers in Table 8.

| |
|---|
| [1] Knockdown of p62/Sequestosome 1 attenuates autophagy and inhibits colorectal cancer cell growth. p62/Sequestosome-1 is a multifunctional adapter protein implicated in selective autophagy, cell signaling pathways, and tumorigenesis, and plays an important role at the crossroad between autophagy and cancer. |
| [2] Human colorectal cancer tissues from patients were analyzed for expression of p62 and Microtubule-associated protein light chain 3 LC3, an autophagosome marker using immunostaining, western blotting, real-time PCR, and confocal microscopy. |
| [3] To study the effects of p62 on autophagy and cell growth, shRNA for p62 was applied and cell growth curve was monitored in human colorectal cancer cell. |
| [4] Protein aggregates can form in the cytoplasm of the cell and are accumulated at aggresomes localized to the microtubule organizing center MTOC where they are subsequently degraded by autophagy. |
| [5] p62/Sequestosome-1 Up-regulation Promotes ABT-263-induced Caspase-8 Aggregation/Activation on the Autophagosome. |
| [6] Autophagy and apoptosis regulate cancer cell viability in response to cytotoxic stress; however, their functional relationship remains unclear. p62/Sequestosome 1 is a multifunctional protein and a signaling hub that shuttles ubiquitinated proteins to the lysosome during autophagy. |
| [7] Up-regulation of p62 was shown to enhance ABT-263-induced caspase-8 activation that was Bax-dependent and resulted from mitochondrial amplification. |
| [8] Ectopic wild-type p62, but not p62 mutants with loss of ability to promote apoptosis, was shown to co-localize with caspase-8 and to promote its self-aggregation in ABT-263-treated cells, shown using a bimolecular fluorescence complementation assay. |
| [9] A direct activator of caspase-8, i.e., TRAIL, alone or combined with ABT-263, induced caspase-8 aggregation and co-localization with p62 that was associated with a synergistic drug interaction. |
| [10] p62/SQSTM1 is required for cell survival of apoptosis-resistant bone metastatic prostate cancer cell lines. |

Second, to show the possibility of summarization using the proposed classification system, we compare two summaries generated using the proposed classification system. For the experiment, the abstracts of the retrieved papers in Table 8 are used. Table 9 shows the automatic summarization results for *Sequestosome-1 protein* using the abstracts of five positively classified papers, and the function information is underlined. The results show that the summarization library properly extracts the protein function information related to SQSMT1 protein.[7] However, the summary of the results using negatively classified papers is not related with protein function and shows

6) The summarization result is not provided because there is no space.

7) In sentence 1 and 6 of Table 9, two sentences are extracted because of preprocessing error of mead library.

information such as the relationship between SQSMT1 mutation and a disease.

The MEAD library extracts key sentences, which are predicted to have a central role among the inputted documents. Thus, to successfully summarize protein function text, only papers providing protein function evidence should be used.

## VI. Conclusion

In this paper, we proposed a biomedical literature classification problem for providing protein function evidence and its solution. Despite our best efforts, we were unable to find previous research on the proposed problem, despite its significance.

To solve the document classification problem, we automatically construct a document classification corpus using the function evidence information of the Swiss-Prot database. To correctly classify papers providing protein function evidence, a classification should not be determined using only the function words used in a document, rather it should consider whether the contributions of a paper are related to the protein function. To solve this difficulty, we propose two novel features — protein and assertion. In addition, through experiments and feature analysis, we show that these two features are helpful in effectively classifying papers providing protein function evidence. The performance of the proposed classification system shows 71.89% precision, 90.0% recall, and a 79.94% F-measure.

As case study applications for the proposed classification system, we show the possibilities of two application systems. The first application is an information retrieval system, which retrieves only papers providing protein function evidence among PubMed search results. The second application is an automatic summarization system for protein functions, such as a function annotation of Swiss-Prot. It was confirmed that the proposed classification system can be successfully applied to these applications. Using the proposed classification system, biomedical researchers can obtain protein function information easily and efficiently.

## References


[1] H. Lodish et al., "Molecular Cell Biology," 5th ed., New York, USA: W.H. Freeman, 2004.

[2] B. Rost et al., "Automatic Prediction of Protein Function," *Cell Molecular Life Sci*., vol. 60, no. 12, Dec. 2003, pp. 2637–2650.

[3] A. Kolchinsky et al., "Classification of Protein-Protein Interaction Full-Text Documents Using Text and Citation Network Features," *IEEE/ACM Trans. Comput. Biology Bioinformat*., vol. 7, no. 3, July–Sept. 2010, pp. 400–411.

[4] Y. Chen, P. Hou, and B. Manderick, "An Ensemble Self-Training Protein Interaction Article Classifier," *BioMed. Mater. Eng*., vol. 24, no. 1, 2014, pp. 1323–1332.

[5] Y. Chen, Y. Sun, and B.-Q. Han, "Improving Classification of Protein Interaction Articles Using Context Similarity-Based Feature Selection," *BioMed Res. Int*., Article ID 751646.

[6] F.C. Garcia et al., "Attribute Analysis in Biomedical Text Classification," *Proc. BioCreative Challenge Evaluation Workshop*, Madrid, Spain, Apr. 23–25, 2007, pp. 113–118.

[7] S. Matos and J.L. Oliveira, "Classification Methods for Finding Articles Describing Protein-Protein Interactions in PubMed," *J. Integr. Bioinformat*., vol. 8, no. 3, Sept. 2011, pp. 178–190.

[8] Y. Li, H. Lin, and Z. Yang, "Two Approaches for Biomedical Text Classification," *Int. Conf. Bioinformat. Biomed. Eng*., Wuhan, China, July 6–8, 2007, pp. 310–313.

[9] A.M. Cohen, "An Effective General Purpose Approach for Automated Biomedical Document Classification," *AMIA Annual Symp. Proc*., 2006, pp. 161–165.

[10] R.B. Dollah and M. Aono, "Ontology Based Approach for Classifying Biomedical Text Abstracts," *Int. J. Data Eng*., vol. 2, no. 1, 2011, pp. 1–15.

[11] C. Sibunruang and J. Polpinij, "Ontology-Based Text Classification for Filtering Cholangiocarcinoma Documents from PubMed," *Int. Conf. Brain Informat. Health*, Warsaw, Poland, Aug. 11–14, 2014, pp. 266–277.

[12] N. Polavarapu et al., "Investigation into Biomedical Literature Classification Using Support Vector Machines," *IEEE Comput. Syst. Bioinformat. Conf*., Stanford, CA, USA, Aug. 8–11, 2005, pp. 366–374.

[13] M. Krallinger, F. Leitner, and A. Valencia, "Retrieval and Discovery of Cell Cycle Literature and Proteins by Means of Machine Learning, Text Mining and Network Analysis," *Int. Conf. Practical Appl. Comput. Biology Bioinformat*., Salamanca, Spain, June 4–6, 2014, pp. 285–292.

[14] M. Conway et al., "Classifying Disease Outbreak Reports Using N-Grams and Semantic Features," *Int. J. Med. Informat*., vol. 78, no. 12, Dec. 2009, pp. e47–e58.

[15] H.C. Jang et al., "Finding the Evidence for Protein-Protein Interactions from PubMed Abstracts," *Bioinfomat*., vol. 22, no. 14, 2006, pp. e220–e226.

[16] L. Li et al., "An Approach to Improve Kernel-Based Protein-Protein Interaction Extraction by Learning from Large-Scale Network Data," *Methods*, Apr. 2015.

[17] N. Papanikolaou et al., "Protein-Protein Interaction Predictions Using Text Mining Methods," *Methods*, vol. 74, no. 1, Mar. 2015, pp. 47–53.

[18] D. Kwon et al., "Assisting Manual Literature Curation for Protein-Protein Interactions Using BioQRator," *Database*, vol. 2014, July 2014. pp. 1–7.

[19] H. Almeida et al., "Machine Learning for Biomedical Literature

Triage," *PLoS ONE*, vol. 9, no. 12, Dec. 2014, pp. 1–21.
[20] UniProt Consortium, "Ongoing and Future Developments at the Universal Protein Resource," *Nucleic Acids Res*., vol. 39, Jan. 2011, pp. D214–D219.
[21] A.L. Berger, V.D. Pietra, and S.D. Pietra, "A Maximum Entropy Approach to Natural Language Processing," *Comput. Linguistics*, vol. 22, no. 1, Mar. 1996, pp. 39–71.
[22] A. Zaeri and M. Nematbakhsh, "A Framework for Semantic Interpretation of Noun Compounds Using Tratz Model and Binary Features," *ETRI J*., vol. 34, no. 5, Oct. 2012, pp. 743–752.
[23] S. Lim et al., "Domain-Adaptation Technique for Semantic Role Labeling with Structural Learning," *ETRI J.*, vol. 36, no. 3, June 2014, pp. 429–438.
[24] Y. Bae, P. Ryu, and H. Kim, "Predicting the Lifespan and Retweet Times of Tweets Based on Multiple Feature Analysis," *ETRI J.*, vol. 36, no. 3, June 2014, pp. 418–428.
[25] D. Radev et al., "MEAD - a Platform for Multidocument Multilingual Text Summarization," *Proc. Int. Conf. Language Resources Evaluation*, Lisbon, Portugal, May 26–28, 2004, pp. 699–702.

**Joon-Ho Lim** received his BS and MS degrees in computer science from Korea University, Seoul, Rep. of Korea, in 2002 and 2005, respectively. Currently, he is a senior researcher at the Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea. His research interests include bio-medical text mining; natural language processing; and machine learning and question answering.

**Kyu-Chul Lee** received his BE, ME, and PhD degrees in computer engineering from Seoul National University, Rep. of Korea, in 1984, 1986, and 1996, respectively. In 1994, he worked as a visiting researcher at the IBM Almaden Research Center, San Jose, CA, USA. From 1995 to 1996, he worked as a visiting professor at the CASE Center at Syracuse University, Syracuse, NY, USA. He is currently a professor with the Department of Computer Engineering, Chungnam National University, Daejeon, Rep. of Korea. His current areas of research interest include multimedia database systems, hypermedia systems, object-oriented systems, and digital libraries. He has authored over 100 technical articles published in various journals and conferences. He is a member of ACM, the IEEE Computer Society, and the Korea Information Science Society.