Biological Text Mining for Extraction of Proteins and Their Interactions*

Kiho Hong[1], Junhyung Park[2], Jihoon Yang[2], and Sungyong Park[2]

[1] IT Agent Research Lab, LSIS R&D Center,
Hogae-dong, Dongsan-Gu, Anyang-Shi, Kyungki-Do 431-080, Korea
khhong1@lsis.biz

[2] Department of Computer Science and Interdisciplinary Program
of Integrated Biotechnology, Sogang University,
1 Shinsoo-Dong, Mapo-Ku, Seoul 121-742, Korea
jhpark@mllab.sogang.ac.kr, {yangjh, parksy}@sogang.ac.kr

Abstract. Text mining techniques have been proposed for extracting protein names and their interactions. First, we have made improvements on existing methods for handling single word protein names consisting of characters, special symbols, and numbers. Second, compound word protein names are extracted using conditional probabilities of the occurrences of neighboring words. Third, interactions are extracted based on Bayes theorem over discriminating verbs that represent the interactions of proteins. Experimental results demonstrate the feasibility of our approach with improved performance.

1 Introduction

In biologically significant applications such as developing a new drug and curing an inveterate disease, understanding the mutual effects of proteins (or genes which will be used interchangeably in the paper) are essential [1]. In order to achieve the goal, extracting gene names must be proceeded. Motivated by this background, we propose a new approach to extracting gene names and their relations.

2 Extraction of Protein Names

A protein is named either as a single word (i.e. singular protein name (SPN)) or multiple words (i.e. multiple protein name (MPN)). We describe extracting methods for each case.

2.1 SPN Extraction

A SPN is extracted by two steps:

1. Word Class Tagging: We used the Brill's tagger for tagging the text [2]. We added a word class GENE and prepared a list of words in the class.

* This work was supported by grant No. R01-2004-000-10689-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

S.K. Pal et al. (Eds.): PReMI 2005, LNCS 3776, pp. 645–648, 2005.

GenBank[1] database was adopted for making the list. To define lexicon rules and context rules during the tagger's learning stage, we used GENIA CORPUS [3,4].

2. SPN Extraction: Generally, protein names are usually irregular and ambiguous. Even though there exist some rules for protein naming (some can be found at Nature Genetics site [5]), it is hard to apply the rules to existing protein names. Also as the rules are not generalized, some of the special characters are used frequently. For this reason, processing them plays a great role for the whole efficiency. The HMM(Hidden Markov Model) with the Viterbi algorithm is applied for SPN extraction [6]. Also, in order to handle special characters, a substitution method was considered (e.g. & for digits and ? for roman letters).

2.2 MPN Extraction

Usually an SPN makes up an MPN with near (or neighboring) words. However, an MPN not including any SPN should be considered as well (e.g. tumor necrosis factor). Based on the technique used in TagGeN [5], we developed an enhanced probability model. First, if GENE tag is included, the range of an MPN is determined by expanding words in bidirection (i.e. right and left). If an MPN does not include any GENE word, we use SEED word (e.g. the words appearing in MPNs frequently). To determine the range of an MPN, it is needed to expand the search from a GENE word or a SEED word, considering the following probability:

$$P(W_{next}|W_{current}, M_{current} = 1) \tag{1}$$

where W_i represents a word occurring at position i, and M_i is a binary value which represents whether the word at position i belongs to GENE word class or not.

3 Extraction of Protein Interactions

There could be a pattern like *'Protein(A)-Type(interaction)-Protein(B)'* [2]. We define the verbs for the interactions and extract events from these predefined patterns. Then we are able to know that entity A has a relation with B. We first extract the discriminating verbs and then extract associated protein interactions.

3.1 Discriminating Verb Extraction

A discriminating verb is extracted as follows:

1. Pre-processing: The set of types (i.e. interactions) we are interested in would be the discriminating verb set. To define the set, words tagged as verbs by Brill's tagger are extracted.

[1] http://www.ncbi.nlm.nih.gov/Genbank/index.html

2. P-Score: We use the Bayesian probability model for estimating the P-Score of each verb in the document. Then, we determine the set of discriminating verbs based on the P-Score. Therefore, the P-Score exhibits how well a verb describes the interaction between proteins. This was proposed for extracting a word set to classify documents by Marcotte [7]. We applied the method for extraction discriminating verbs and calculate the following probability:

   $$P(n|N,f) \approx e^{-Nf}\frac{(Nf)^n}{n!} \qquad (2)$$

   where n means how many times a verb is used as a protein interaction, N is the total number of words in a document, and f is the total occurrences of each verb. The Poisson distribution can be an alternative for $P(n|N,f)$ while N is big enough and f is fairy small.
3. Discriminating Verb Selection: Calculate P-Score for every word, and then choose a set of arbitrary number of words with the highest P-Scores.

3.2 Protein Interaction Extraction

The steps of extracting protein interactions are as follows:

1. Complex Sentence Processing: To handle the ambiguity in a sentence, we used Toshihide Ono's method [1] used for processing complex and negative sentences.
2. Interaction Extraction: If there is a pattern like *'Protein(A)-Type(Verb)-Protein(B)'* and a discriminating verb in a sentence, we calculate $Confidence$ of the sentence and then add the sentence into the *event* (protein interaction) set. The $Confidence$ is calculated by sum of b (binary value which represents whether the pattern is included in the sentence or not) and a reciprocal of sd (sum of distances from proteins to a verb in the sentence). A sentence with no discriminating verb is added into candidate event set. We re-calculate $Confidence$ with $Frequency$ (how many times protein(A) and (B) are found in documents).

4 Experiments

We obtained the following extraction results of proteins and their interactions. Data used for the experiments are 600 papers from the GENIA Corpus. Our results are compared with those by ABGene and TagGeN [3,5] in Table 1.

- SPN: Since we used a substring matching method, our system produced high recall value at the cost of precision. Our system also showed outstanding extraction time than other methods.
- MPN: 'Exact' and 'partial' respectively mean perfect and subset match of MPN. Our approach outperformed TagGeN in MPN extraction.

Table 1. Performance of SPN and MPN Extraction

	SPN			MPN (exact/partial)	
	Precision(%)	Recall(%)	Time(sec)	Precision(%)	Recall(%)
Our system	85.00	96.27	6.23	86.65/91.35	84.25/91.56
ABGene	87.01	55.22	113.00	-	-
TagGeN	83.24	76.27	36324	87.81/91.15	80.23/86.51

– Protein Interaction Extraction: We used 80 discriminating verbs. Arbitrarily selected 100 sentences including 14 negative, 8 compound sentence structures, and 121 protein interactions were used. From the sentences, we extracted 139 protein interactions and obtained 76.58% precision, 92.70% recall and 83.87% F-measure value.

5 Conclusion

We developed an extraction system for proteins and their interactions. Our method with character substitution and abundant lexicon improved overall performance. We also defined discriminating verbs and extracted them using a probabilistic model. We extracted 80 discriminating verbs by Poisson distribution. Finally, we defined events, and extracted the interactions considering the confidence values of the events. We observed improved performance in all experiments.

References

1. Ono, T.: Automated extraction of information on protein-protein interactions from the biological literature. Bioinformatics **17** (2001) 155–161
2. Brill, E.: Some advances in transformation-based part of speech tagging. In: AAAI. (1994)
3. J. D. Kim, T. Ohta, Y.T., Tsujii, J.: Genia corpus - a semantically annotated for bio-textmining. Bioinformatics **19** (2002) 180–192
4. Rinaldi, F.: Mining relations in the genia corpus. In: Proceedings of the Second European Workshop and Text mining for Bioinformatics. (2004)
5. Tanabe, L., Wilbur, W.J.: Tagging gene and protein names in full text article. In: Proceedings of Association for Computational Linguistics. (2004) 9–13
6. Duda., R.O.: Pattern Classification. second edition edn. Wiley-interscience. Inc. (2000)
7. M.Marcotte, E.: Mining literature for protein-protein interactions. Bioinformatics **17** (2002) 359–363