行政院國家科學委員會專題研究計畫　期中進度報告

自然語言處理技術在生物資訊自動粹取上的研究(1/2)

計畫類別：個別型計畫
計畫編號：NSC91-2213-E-002-088-
執行期間：91 年 08 月 01 日至 92 年 07 月 31 日
執行單位：國立臺灣大學資訊工程學系暨研究所

計畫主持人：陳信希

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 92 年 5 月 19 日

1 Introduction

Mining biological relationships from scientific text becomes more important. Most researchers [1, 2, 3, 4, 5, 6, 7] in discovering molecular relationships are based on some special verbs and their related noun forms which are selected by intuition. In this way, we cannot assure if the keyword set is complete for mining biological relationships. The purpose of this paper is to identify significant biological keywords/key phrases automatically, which often accompany with molecular entities in the biological documents.

2 Method and Results

2.1 Method

Step 1 (Tagging the unannotated corpus): The corpus that has size of 1,514 MEDLINE abstracts about protein structures for the experiment was downloaded from the PASTA website in Sheffield University [http://www.dcs.shef.ac.uk/nlp/pasta]. By looking up the protein lexicon, the protein names in the biological corpus are identified.

Step 2 (Preprocessing): First we exclude the stopwords, and then stemming procedure is applied.

Step 3 (Computing collocation statistics): The collocation words with proteins specify that they often co-occur with protein names. In this study, we calculate three collocation statistics as follows.

1. Frequency: In general, more occurrences in the collocation windows are preferred, but the standard criteria for frequencies are not acknowledged. Hence, another model is considered to assist this task.
2. Mean and variance: The mean value of collocations can indicate how far collocates are typically located from protein names. Furthermore, variance shows the deviation from the mean. If the standard deviation is equal to zero, it says that the collocates and the protein names always occur at exactly the same distance equal to the mean value. If the standard deviation is low, two words usually occur at about the same distance. If the standard deviation is high, then they co-occur at random.
3. *t*-test model: To get the statistical confidence that two words have a collocation relationship, *t*-test hypothesis testing is adopted. In the *t*-test model, if the *t*-value is larger than 2.576, the word is regarded as a good collocate with 99.5% confidence.

Step 4 (Extraction of collocation keywords): From the previous discussion about the extraction of the biological information, verbs are the main words because many of the subject and object terms related to these verbs are names of genes or proteins. Thus, verbs are targets in this phase.

2.2 Results of Extracting Keywords

Of the 4,782 different stemmed words appearing in the collocation windows, there are 154 collocation keywords with verbal part-of-speech in Step 4. Partial results and the corresponding evaluation results are illustrated in Tables 1 and 2, respectively. In Table 1, the “Word” column lists the collocates found in Step 4. The “Freq” column shows the results about the frequency. The “Avg-Dist” column represents the average distance in the collocation windows. The “STD-Dev” column denotes the standard deviation. The “*t*-value” column describes the results in the *t*-test model. In Table 2, “A”(Author) is the surname of the first author in the references [1, 2, 3, 4, 5, 6, 7]. The number of verbs listed in the literatures to support the interactions or pathways of proteins or genes are listed in the "SV" (suggested_verbs) colum. The "F" (finding) column is the output in Step 4. The "U" (uncontained) column counts the suggested verbs that are not contained in the biological corpus. The "NF" (not_found) column denotes the number they are not considered as good keywords by our method. The "P" column that denotes the performance is calculated as follows. |*finding*| / (|*suggested_verbs*| - |*uncontained*|)

Table 1. The Collocates with the Highest 10 *t*-value in Step 4

Word	Freq	Avg-Dist	STD-Dev	*t*-value
structure	819	-1.591	2.999	26.713
Bind	365	-0.164	3.269	17.830
complex	348	0.198	2.904	17.410
determine	193	1.233	3.079	12.965
activate	188	0.138	3.254	12.796

Table 2. Evaluation Results in Step 4

A	SV	F	U	NF	P
Blaschke	14	10	2	2	83.33%
Ng	8	4	3	1	80%
Ono	4	4	0	0	100%
Park	12	5	2	5	50%
Sekimizu	7	7	0	0	100%
Thomas	10	10	0	0	100%
Yakushiji	5	3	2	0	100%

After computing the overall performance, the average value is 73.33%, where 30 keywords have appeared in the reference corpus and we found 22 collocates. The result shows there remains some improving space.

2.3 Results of Extracting Key Phrases

Because compound words may be interesting key phrases, we find them in the second experiment. The idea here is to apply the proposed method with window size one word on each side of the original keywords. Consequently, the performance in the first row is improved to 91.67%. If the influence of pathways, i.e. Ng [2] and Park [4] used, is removed, the average value is 95.45% (21/22). Otherwise, it is 76.67% (23/30). As a result, we get a more objective suggestion about compound collocation keywords rather than the researchers' subjective opinions. That is the key idea of this paper.

3 Concluding Remarks

We have shown an automatic way of mining biological keywords from scientific text in the protein domain. The methods are based on collocation statistics, and the performance reaches to 95.45% if we focus on the interaction between proteins. The same approach can be extended to gene domain when the gene dictionary replaces the protein dictionary. Another considerable domains include DNA, RNA, or drugs. In summary, the work presented herein represents significant evidence toward mining biological relationships among textual sources, and lots of applications, such as name and interaction extraction, will get better feedback.

Acknowledgements

Part of research results was supported by National Science Council under the contract NSC-91-2213-E-002-088. We also thank Dr. George Demetriou in the Department of the Computer Science of the University of Sheffield, who kindly supported the resources in this work.

References

[1] Blaschke, C., Andrade, M.A., Ouzounis, C., and Valencia, A. Automatic Extraction of Biological Information from Scientific Text: Protein-Protein Interactions. *Proceedings of the seventh International Conference on Intelligent Systems for Molecular Biology* (*ISMB99*), Heidelberg, Germany - AAAI Press, 60-67, 1999.

[2] Ng, S.K. and Wong, M. Toward Routine Automatic Pathway Discovery from On-line Scientific Text Abstracts. *The Tenth Workshop on Genome Informatics* (*GIW99*), 10:104-112, 1999.

[3] Ono, T., Hishigaki, H., Tanigami, A., and Takagi, T. Automated Extraction of Information on Protein-Protein Interactions from the Biological Literature. *Bioinformatics*, 17(2):155-161, 2001.

[4] Park, J.C., Kim, H.S., and Kim, J.J. Bidirectional Incremental Parsing for Automatic Pathway Identification with Combinatory Categorial Grammar. *Proceedings of the Pacific Symposium on Biocomputing* (*PSB 2001*), 6:396-407, 2001.

[5] Sekimizu, T., Park, H.S., and Tsujii, J. Identifying the Interaction Between Genes and Genes Products Based on Frequently Seen Verbs in Medline Abstract. *The Ninth Workshop on Genome Informatics* (*GIW98*), 62-71, 1998.

[6] Thomas, J., Milward, D., Ouzounis, C., Pulman, S., and Carroll, M. Automated Extraction of Protein Interactions from Scientific Abstracts. *Proceedings of the Pacific Symposium on Biocomputing* (*PSB 2000*), 4-9 January 2000, Honololu, Hawaii, 5:538-549, 2000.

[7] Yakushiji, A., Tateisi, Y., and Miyao, Y. Event Extraction from Biomedical Papers Using a Full Parser. *Proceedings of the Pacific Symposium on Biocomputing* (*PSB 2001*), 6:408-419, 2001.