Prediction of Protein Sub-cellular Localization using Information from Texts and Sequences

Hong-Woo Chun[1,2,3] **Chisato Yamasaki**[2,3] **Naomi Saichi**[2,3] **Masayuki Tanaka**[2,3]
chun@dbcls.rois.ac.jp, {chisato-yamasaki, nao-saichi, masa-tanaka}@aist.go.jp
Teruyoshi Hishiki[3] **Tadashi Imanishi**[3,5] **Takashi Gojobori**[3,6]
{t-hishiki, t.imanishi, t-gojobori}@aist.go.jp
Jin-Dong Kim[4] **Jun'ichi Tsujii**[4,7,8] **Toshihisa Takagi**[1,9]
{jdkim, tsujii}@is.s.u-tokyo.ac.jp, takagi@dbcls.rois.ac.jp
[1] Database Center for Life Science, Research Organization of Information and System, Engineering 12th Bldg., University of Tokyo, 2-11-16 Yayoi, Bunkyo-ku, Tokyo, 113-0032, Japan
[2] Japan Biological Information Research Center, Japan Biological Informatics Consortium
[3] Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, Japan
[4] Department of Computer Science, University of Tokyo, Japan
[5] Graduate School of Information Science and Technology, Hokkaido University, Japan
[6] Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics
[7] School of Informatics, University of Manchester, UK
[8] National Centre for Text Mining, UK
[9] Department of Computational Biology, University of Tokyo, Japan

Abstract

This paper presents a novel prediction approach for protein sub-cellular localization. We have incorporated text and sequence-based approaches.

1 Introduction

Natural Language Processing (NLP) has tackled and solved a lot of prediction problems in Biology. One practical research issue is *Protein Sub-Cellular Localization (PSL) Prediction*. Many previous approaches have combined information from both texts and sequences by a machine learning (ML) technique (Shatkay et al., 2007). All of them have not used traditional NLP techniques such as parsing. Our aim is to develop a novel PSL prediction system using information from texts and sequences. At the same time, we demonstrated the effectiveness of the traditional NLP and the sequence-based features in the viewpoint of the text-based approach.

2 Methodology

A Maximum Entropy-based ML technique has been used to combine information from both texts and sequences. To develop a supervised ML-based prediction system, an annotated corpus is needed to train the system. However, there is no publicly available corpus that contains the PSL. Therefore, we have constructed a corpus using GENIA corpus as an initial data, because the annotation of *Protein* and *Cellular component* in GENIA corpus is already done by human experts. The new types of annotation contain two tasks. The first annotation is to classify 1,117 cellular components in GENIA corpus into 11 locations, and the second annotation is to categorize a relation between a protein and a location into positive, negative, and neutral. Biologists selected 11 locations based on *Gene Ontology*: *Cytoplasm, Cytoskeleton, Endoplasmic reticulum, Extracellular, Golgi apparatus, Granule, Lysosome, Mitochondria, Nucleus, Peroxisome,* and *Plasma membrane*. The number of co-occurrences in GENIA corpus is 864. [1] Three human experts annotated with 79.49% of inter-annotator agreement. For calculating the inter-annotator agreement, all annotators annotated 117

[1] The co-occurrence in the proposed approach is a sentence that contains at least one pair of protein and cellular component names.

BioNLP 2008: Current Trends in Biomedical Natural Language Processing, pages 90–91,
Columbus, Ohio, USA, June 2008.

Location	# Relevant relations	Performance : F-score (Precision, Recall)			
		Baseline	Text	Sequence	Text + Sequence
Nucleus	173	0.282 (0.164, 1.0)	0.764 (0.736, 0.794)	0.725 (0.569, 1.000)	**0.778 (0.758, 0.798)**
Cytoplasm	94	0.163 (0.089, 1.0)	**0.828 (0.804, 0.852)**	0.788 (0.657, 0.984)	**0.828 (0.804, 0.852)**
Plasma membrane	23	0.043 (0.022, 1.0)	0.875 (0.814, 0.946)	0.857 (0.766, 0.973)	**0.885 (0.841, 0.932)**

Table 1: Performance of protein sub-cellular localization prediction for each location.

co-occurrences. From the texts, we used eight features: (1) protein and cellular component names annotated by human experts, (2) adjacent one and two words of names, (3) bag of words, (4) order of names, (5) distance between names, (6) syntactic category of names, (7) predicates of names, and (8) part-of-speech of predicates. To analyze the syntactic structure, we used the ENJU full parser whose output is predicate-argument structures of a sentence.

To combine the information from sequences, we attempted to predict PSL for all proteins in GENIA corpus by two existing sequence-based methods: WoLF PSORT (Horton et al., 2006) and SOSUI (Hirokawa et al., 1998). Approximately 14% of protein names in GENIA corpus obtained results. From the sequences, we used two features: (1) existence of the sequence-based results, and (2) the number of sequence-based results.

3 Experimental results and Conclusion

The proposed approach has integrated text and sequence-based approaches. To evaluate the system, we performed 10-fold cross validation using 864 co-occurrences including positive, negative, and neutral relations. We measured the precision, recall, and F-score of the system for all experiments. Among 864 co-occurrences in GENIA corpus, 301 positive or negative co-occurrences have been considered as relevant relations, and the remaining 563 neutral relations have been considered as irrelevant relations.

Four approaches have been compared based on three locations in Table 1. The four approaches are *baseline, text-based approach, sequence-based approach, and integration of the text and sequence-based approaches*. Baseline experiment used an assumption: *there is a relevant relation if a protein and a cellular component names occur together in a co-occurrence*. The three locations selected when there are the sequence-based results and the number of relevant relations is more than *one*. All experiments showed that the integration of text and sequence-based approaches is the best, even though the experiments for *Cytoplasm* showed the best performance at both the text-based approach and the integration approach.

A new prediction method has been developed for protein sub-cellular localization, and it has integrated text and sequence-based approach using an ML technique. The traditional NLP techniques contributed to improve performance of the text-based approach, and the text and sequence-based approaches reciprocally contributed to obtain a improved PSL prediction method. The newly constructed corpus will be included in the next version of GENIA corpus. There are weak points in the proposed approach. The current evaluation method has been focusing on evaluating the text-based approach, and the results of the sequence-based approach were obtained for only 14% of proteins in GENIA corpus, so these situations might be the reason that the sequence-based approach did contribute a little. Thus, we need to evaluate the proposed approach with a more reasonable method.

Acknowledgments

We acknowledge Fusano Todokoro for her technical assistance.

References

Paul Horton, Keun-Joon Park, Takeshi Obayashi and Kenta Nakai. 2006. *Protein Subcellular Localization Prediction with WoLF PSORT. Asia Pacific Bioinformatics Conference (APBC)*, pp. 39–48.

Takatsugu Hirokawa, Seah Boon-Chieng and Shigeki Mitaku. 1998. *SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics*, 14(4): pp. 378–379.

Hagit Shatkay, Annette Höglund, Scott Brady, Torsten Blum, Pierre Dönnes and Oliver Kohlbacher. 2007. *SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. Bioinformatics.*, 23(11): pp. 1410–1417