Using Bigram Language Model for Protein Name Recognition

Serhan TATAR, Ilyas Cicekli

Department of Computer Engineering, Bilkent University 06800 Bilkent, Ankara, Turkey
{statar, ilyas}@cs.bilkent.edu.tr

1 Motivation

As one of the basic tasks in automatic discovery and extraction of information from biological texts, protein name extraction is still a challenge. Extracting protein names from unstructured texts is a prerequisite for the increasing demand in automatic discovery and extraction of information from biological texts. Locating the information on different levels can be seen as a layered structure and this layered structure makes different extraction tasks interdependent. Because the output of a task at a layer is input to the next layer, the success of a former task affects the performance of the others. For instance, how well we locate the protein names in a text has an impact on how well we find the interactions between the proteins.

2 Method

In order to identify protein names, we study using bigram language model, a special case of N-gram which is used in various areas of statistical natural language processing, along with the hierarchically categorized syntactic word types. We determine 21 syntactic token types categorized under five main classes to generalize protein names: *single*, *abbreviation*, *delimiter*, *regular*, and *other*.

After learning the necessary model parameters, a probability estimate is produced for every possible fragment in the test data. We use sliding-window technique to determine the fragments. Fragments with the highest likelihood, exceeding a certain threshold value, are extracted as protein names.

3 Results

Table 1 compares performance values of our method (Bigram) with the values published for several methods. Our method has a comparable performance to the others with respect to F-score. The comparison also shows that our method is effective and competitive.

Table 1. Comparison of methods for protein name extraction.

	Recall	**Precision**	**F-score**
Bigram	67.5%	60.2%	63.6 %
YAPEX [1]	59.9%	62.0%	61.0 %
SemiCRF [2]	76.1%	58.9%	66.1 %
DictHMM [2]	45.1%	69.7%	54.8 %
Prob [3]	60.1%	66.9%	63.3 %

References

1. Proteinhalt i text, http://www.sics.se/humle/projekt/prothalt/
2. Kou, Z., Cohen, W. W., and Murphy, R. F. 2005. High-recall protein entity recognition using a dictionary. Bioinformatics 21, 1 (Jan. 2005), 266-273.
3. Seki, K. and Mostafa, J. 2003. A Probabilistic Model for Identifying Protein Names and their Name Boundaries. In Proceedings of the IEEE Computer Society Conference on Bioinformatics (August 11 - 14, 2003). CSB. IEEE Computer Society, Washington, DC, 251.