

Extracting Human Protein Information from MEDLINE Using a Full-Sentence Parser

Róbert Busa-Fekete* and András Kocsor*†

Abstract

Today, a fair number of systems are available for the task of processing biological data. The development of effective systems is of great importance since they can support both the research and the everyday work of biologists. It is well known that biological databases are large both in size and number, hence data processing technologies are required for the fast and effective management of the contents stored in databases like MEDLINE. A possible solution for content management is the application of natural language processing methods to help make this task easier.

With our approach we would like to learn more about the interactions of human genes using full-sentence parsing. Given a sentence, the syntactic parser assigns to it a syntactic structure, which consists of a set of labelled links connecting pairs of words. The parser also produces a constituent representation of a sentence (showing noun phrases, verb phrases, and so on). Here we show experimentally that using the syntactic information of each abstract, the biological interactions of genes can be predicted. Hence, it is worth developing the kind of information extraction (IE) system that can retrieve information about gene interactions just by using syntactic information contained in these text. Our IE system can handle certain types of gene interactions with the help of machine learning (ML) methodologies (Hidden Markov Models, Artificial Neural Networks, Decision Trees, Support Vector Machines). The experiments and practical usage show clearly that our system can provide a useful intuitive guide for biological researchers in their investigations and in the design of their experiments.

Keywords: feature extraction, human gene interaction, data mining, machine learning, MEDLINE

*Research Group on Artificial Intelligence of the Hungarian Academy of Sciences and University of Szeged H-6720 Szeged, Aradi vértanúk tere 1., Hungary, E-mail: {busarobi,kocsor}@inf.u-szeged.hu

†The author was supported by the János Bolyai fellowship of the Hungarian Academy of Sciences.