The 2nd Annual Conference of the Korean Society for Bioinformatics (KSBI2003)

Natural language processing techniques for bioinformatics

Jun-ichi Tsujii

Department of Computer Science
Graduate School of Information Science and Technology
University of Tokyo, Japan

Abstract

With biomedical literature expanding so rapidly, there is an urgent need to discover and organize knowledge extracted from texts. Although factual databases contain crucial information the overwhelming amount of new knowledge remains in textual form (e.g. MEDLINE). In addition, new terms are constantly coined as the relationships linking new genes, drugs, proteins etc. As the size of biomedical literature is expanding, more systems are applying a variety of methods to automate the process of knowledge acquisition and management.

In my talk, I focus on the project, GENIA, of our group at the University of Tokyo, the objective of which is to construct an information extraction system of protein - protein interaction from abstracts of MEDLINE. The talk includes

(1) Techniques we use for named entity recognition
  (1-a) SOHMM (Self-organized HMM)
  (1-b) Maximum Entropy Model
  (1-c) Lexicon-based Recognizer
(2) Treatment of term variants and acronym finders
(3) Event extraction using a full parser
(4) Linguistic resources for text mining (GENIA corpus)
  (4-a) Semantic Tags
  (4-b) Structural Annotations
  (4-c) Co-reference tags
  (4-d) GENIA ontology

I will also talk about possible extension of our work that links the findings of molecular biology with clinical findings, and claim that textual based or conceptual based biology would be a viable alternative to system biology that tends to emphasize the role of simulation models in bioinformatics.