Prediction by Partial Matching for Identification of Biological Entities

Thirumalaiswamy Sekhar, Arvind Kumar

Permanent Link: http://hdl.handle.net/1805/2266

Keywords: Biological Entities ; Identification ; Partial Matching ; Prediction

Date: 2010-09-29

Sponsorship:

Malika Mahoui, Ph.D., Chair; Narayanan Perumal,Ph.D.; Pedro Romero,Ph.D.

Abstract:

As biomedical research and advances in biotechnology generate expansive datasets, the need to process this data into information has grown simultaneously. Specifically, recognizing and extracting these “key” phrases comprising the named entities from this information databank promises a plethora of applications for scientists. The ability to construct interaction maps,identify proteins as drug targets are two important applications. Since we have the choice of defining what is “useful”, we can potentially utilize text mining for our purpose. In a novel attempt to beat the challenge, we have put information theory and text compression through this task. Prediction by partial matching is an adaptive text encoding scheme that blends together a set of finite context Markov models to predict the probability of the next token in a given symbol stream. We observe, named entities such as gene names, protein names, gene functions, protein-protein interactions – all follow symbol statistics uniquely different from normal scientific text. By using well defined training sets that allow us to selectively differentiate between named entities and the rest of the symbols; we were able to extract them with a good accuracy. We have implemented our tests, using the Text Mining Toolkit, on identification of gene functions and protein-protein interactions with f-scores (based on precision & recall) of 0.9737 and 0.6865 respectively. With our results, we foresee the application of such an approach in automated information retrieval in the realm of biology.

Files in this item

Name: SekharThesis.pdf View/Open
Size: 709.7Kb
Format: PDF

This item appears in the following Collection(s)

- Informatics Graduate Theses and PhD Dissertations