Poster K-2

Literature Mining at the Protein Information Resource (PIR)

Authors:
Hu, ZZ (*Protein Information Resource, Georgetown University Medical Center*)
Liu, H (*Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University Medical Cente*)
Han, B (*Center for Information Science and Technology, Temple University*)
Huang H (*Protein Information Resource, Georgetown University Medical Center*)
Mazumder R (*Protein Information Resource, Georgetown University Medical Center*)
Yuan, X (*Protein Information Resource, Georgetown University Medical Center*)
Obradovic, Z (*Center for Information Science and Technology, Temple University*)
Vijay-Shanker, K (*Department of Computer and Information Sciences, University of Delaware*)
Vucetic, S (*Center for Information Science and Technology, Temple University*)
Wu, CH (*Protein Information Resource, Georgetown University Medical Center*)

Short Abstract: The Protein Information Resource conducts research and provides resources on literature mining in areas, i) named entity recognition, providing BioThesaurus of gene/protein names and database associations; ii) information retrieval, developing algorithms for text categorization; iii) information extraction, developing online text mining tool for protein phosphorylation to assist protein annotation.

Long Abstract:

With the increasing volume of scientific literature now available electronically such as from PubMed, efficient text mining tools will greatly facilitate the extraction of information buried in free text and will assist in database annotation. Many methods, including natural language processing, machine learning, and rule-based approaches have been employed for biological literature mining, especially in the areas of entity recognition (ER, e.g., gene/protein name), information retrieval (IR, e.g., text categorization), and information extraction (IE, e.g., protein post-translational modification). The Protein Information Resource (PIR) group (http://pir.georgetown.edu/), actively collaborating with several other groups, conducts research and provides resources on literature mining in the above three areas. The literature mining resource publicly provided at PIR – iProLINK (integrated Protein Literature INformation and Knowledge) (http://pir.georgetown.edu/iprolink/) is aimed to provide annotated literature data sets for developing new literature mining algorithms, such as protein named entity recognition (ER), text categorization (IR), and protein annotation extraction (IE), and to provide literature mining tools for scientific users and curators.

I. Name Entity Recognition Using Gene/Protein Thesaurus

Recognizing gene and protein names and associating them to database entries is an essential step in text mining, which facilitates automatic mappings to specific database entries of papers describing the gene/protein (citation mapping), or of extracted information (IE) for protein annotations. Due to the lack of name standardization, gene/protein names in literature and annotated in bioinformatics databases often vary greatly. A comprehensive

collection of gene/protein names covering the wide range of names can be used to develop automatic name recognition methods. We developed a gene/protein name thesaurus, BioThesaurus, which currently collects 3.2 million gene/protein names from over 23 biological databases, including the major gene (e.g., Entrez Gene) and protein databases (e.g. UniProt), and the model organism databases (e.g. MGD, SGD). The thesaurus also associates all names with over 2.6 million protein entries in UniProt Knowledgebase (UniProtKB). A web-based BioThesaurus is developed to provide online query for finding gene/protein synonyms of any given UniProtKB entry, and for solving name ambiguities for the same name shared by more than one protein entry (http://pir.georgetown.edu/iprolink/biothesaurus/). The thesaurus of names and their database associations can be downloaded for automatic entity tagging. The coverage of protein names from literature in BioThesaurus is estimated to be >95%.

II. Text Categorization Using Machine Learning Method

Biological document classification based on specific information contained in the paper, such as protein-protein interactions, protein phosphorylation, or pathogenesis-related microbial proteins, is an important task in analyzing large-scale biomedical literature. One critical step in development of document classification systems is the attribute selection. As a standard practice, words are stemmed and the most informative ones are used as attributes in classification. Due to high complexity of biomedical terminology, general-purpose stemming algorithms are often conservative and could also remove informative stems. This can lead to accuracy reduction, especially when the number of annotated documents is small. A new algorithm omitting stemming and instead using the most discriminative substrings as attributes was tested on five annotated sets of abstracts from iProLINK, which report on the experimental evidence about five types of protein post-translational modifications, glycosylation, phosphorylation, acetylation, methylation and hydroxylation. The experiments showed that Naive Bayes and Support Vector Machine classifiers perform consistently better (with Area Under the ROC Curve (AUC) accuracy in range 0.92-0.97) when using the proposed attribute selection than when using attributes obtained by the Porter stemmer algorithm (AUC in 0.86-0.93 range). Therefore the proposed substring-based approach is highly effective in document classification even when relatively small annotated datasets are available for learning.

III. Online Text Mining Tool for Protein Phosphorylation

Protein phosphorylation is a fundamental molecular event essential to cellular processes. We previously developed and benchmarked a rule-based text mining system, RLIMS-P, for extracting protein phosphorylation information from literature regarding three objects—the protein kinase, the protein substrate, and the phosphorylation sites. We now developed a web-based version of the RLIMS-P for online mining of protein phosphorylation information from MEDLINE abstracts (http://pir.georgetown.edu/iprolink/rlimsp/). The online RLIMS-P text mining tool is designed for easy accessibility and with enhanced functionality. The online tool can process over 100 PubMed IDs in one query and presents three extracted phosphorylation objects in summary tables and in full reports with evidence tagged in abstracts. The tool further allows mapping of phosphorylated proteins to entries in the UniProtKB based on PubMed ID and/or on protein names using BioThesaurus. With a user-friendly interface for phosphorylation information mining, evidence tagging, and protein mapping to UniProtKB, the online RLIMS-P facilitates biological studies of phosphorylated proteins and the database annotation of phosphorylation features.