INFORMATION SCIENCE PUBLISHING
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITB13377

This chapter appears in the book, *Advances in Electronic Business, Volume II*
edited by Eldon Li © 2007, Idea Group Inc.

Chapter XII

Web Mining for Protein-to-Protein Interaction Information

Hsi-Chieh Lee
Yuan Ze University, Taiwan &
National Kinmen Institute of Technology, Taiwan

Szu-Wei Huang
Yuan Ze University, Taiwan

Eldon Y. Li
National Chengchi University, Taiwan, &
California Polytechnic State University, USA

Abstract

This study proposes a mining system for finding protein-to-protein interaction literatures from the databases on the Internet. In this system, we search for discriminating words for protein-to-protein interaction by way of statistics and the results from literatures. A threshold is also evaluated to check if a given literature is related to protein-to-protein interactions. In addition, a keypage-based search mechanism is used to find related papers for protein-to-protein interactions from a given document. To expand the search space and ensure better performance of the system, mechanisms for protein name identification and databases for protein names are also developed. The system is designed with a web-based user interface and a job-dispatching kernel. Experiments are conducted and the results have been checked by a biomedical expert. The experimental results indicate that by using the proposed mining system, it is helpful for researchers to find protein-to-protein literatures from the overwhelming pieces of information available on the biomedical databases over the Internet.

Introduction

In recent years, due to advances in information technology, more and more biomedical-related information is available electronically on the Internet. For example, the MEDLINE database contains over 12 million citations dating back to the mid-1960's. Therefore, it has become an important issue for mining valuable biomedical information from the literature (Valencia-García, Ruiz-Sánchez, Vicente, Fernández-Breis, & Martínez-Béjar, 2004; Wang, Kuo, Chen, Hsiao, & Tsai, 2005), especially information on the Internet (Hong & Han, 2002). Expert systems and data mining techniques have been used for years in medical diagnosis domain (Chou, Lee, Shao, & Chen, 2004; Alonso, Caraça-Valente, González, & Montes, 2002).

In the post-genomic era, some scientists focus on finding meaningful information of DNA or try to use the information of gene sequence in solving problems. However, the spirit of post-genomic era can be view broadly in three ways. The first one is the sequences from DNA level, and the second one is the expressed sequence tag (EST) from RNA level. The last one is proteome from the protein level. People can use the analyzed information to understand the interaction between each other and discover the meaning behind it. In other words, after decoding the sequence, scientist can analyze the interaction between gene and protein, and understand the role the gene is playing on an organism. It has been shown that the protein and genomics would become the main issue in the post-genomic era (Eisenberg, Marcott, Xenarios, & Yeates, 2000).

Moreover, scientists try to understand the interaction and relation between proteins from biochemistry and gene-related angles. For example, the database of interaction proteins (DIP) developed in UCLA (Xenarios et al., 2000) has data about over 5900 proteins and 10500 protein-to-protein interactions. Besides DIP, there exist many other databases with the collection of the data regarding protein function and pathway. However, if people want to know the relationship between proteins, they have to search different literatures and try to find some relationships. It is considered mission impossible to check on MEDLINE manually where there exist more than 15 million biomedical citations. It is time-consuming and ineffective. It would be helpful if the job can be processed automatically and the database can be updated as soon as new literatures are available.

Generally, mining the literatures of protein-to-protein interactions requires natural language processing. The literature discussing protein-to-protein interactions does not contain a language that a computer can understand. As a result, there are two typical approaches in solving the problem. The first approach is transferring the format into a way that computer can understand by natural language processing. For example, in Ono, Hishigaki, Tanigami, and Takagi (2001), they brought up an idea to extract biomedical-related information with two steps. The first step is to scan the full document with a protein name dictionary. The second step is to extract content related to protein-to-protein interactions by predefined rules. The second approach is to extract biomedical-related information using statistics. The most typically way of statistics is calculating the frequency of words. In Marcotte, Xenarios, and Eisenberg (2001), they used statistics to find 83 words as discriminative words to check whether a paper is discussing protein-to-protein interaction.