CHAPTER 4

Extracting Information for Meaningful Function Inference through Text-Mining

Hong Pan, Li Zuo, Rajaraman Kanagasabai, Zhuo Zhang, Vidhu Choudhary, Bijayalaxmi Mohanty, Sin Lam Tan, S.P.T. Krishnan, Pardha Sarathi Veladandi, Archana Meka, Weng Keong Choy, Sanjay Swarup and Vladimir B. Bajic*

Abstract

One of the emerging technologies in computational biology is text-mining which includes natural language processing. This technology enables extraction of parts of relevant biological knowledge from a large volume of scientific documents in an automated fashion. We present several systems which cover different facets of text-mining biological information with applications in transcription control, metabolic pathways, and bacterial cross-species comparison. We demonstrate how this technology can efficiently support biologists and medical scientists to infer function of biological entities and save them a lot of time, paving way for more focused and detailed follow-up research.

Introduction

Text-mining of biomedical literature has received an increased attention in the past several years.[1-6] This is caused by several reasons:

a. a huge volume of the scientific documents available over internet to an average user;
b. inability of an average user to keep track of all relevant documents in a specific domain of interest;
c. inability of humans to keep track of associations usually contained in, or implied by, scientific texts; these associations could be either explicitly stated, such as 'interaction of A and B', or they need not necessarily be explicitly spelled out in a single sentence;
d. inability of humans to simultaneously deal with a large volume of terms and their cross-referencing;
f. necessity to search a number of different documents (or sometimes resources) to extract a set of relevant information;
e. inability of a single user to acquire the required information in a relatively short (acceptable) time.

As an illustration, currently PubMed repository (http://www.ncbi.nlm.nih.gov/entrez) contains over 14 million indexed documents.[1] It is common that searches of PubMed frequently provide several hundreds or more returned documents. Studying these large document sets is not an easy task for a single user. If the analysis has to be repeated several times with different selection of documents, then such a task is usually not feasible.

*Corresponding Author: Vladimir B. Bajic—Knowledge Extraction Lab, Institute for Infocomm Research, 21 Heng Mui Keng Terrace, 119613, Singapore. Email: bajicv@i2r.a-star.edu.sg

Discovering Biomolecular Mechanisms with Computational Biology,
edited by Frank Eisenhaber.

Text-mining is seen as an interesting and powerful supporting technology to complement research in biology and medicine. A number of text-mining systems which tackle different problems aimed at supporting biological and medical research, and which focus on different aspects of genomics, proteomics, or relations to diseases, have been reported.[7-19]

Computational biology produces answers, which form the bases and lead to better designs for further experimentation. Among the various computational biology approaches, text-mining systems provide a unique front where large quantum of knowledge put out by experimental biologists can be efficiently screened using "vocabularies" or standard terms adopted and used widely by biologists. Hence, such systems analyze the existing knowledge and uncover potential associations among biological entities or phenomena that can lead to further experimentation. In effect, text-mining-based approaches allow biologists to focus on certain unique aspects of information that would have been reported independently thus not lending them for establishment of readily recognizable associations or correlations. Many such associations in biology go unnoticed till more directed studies are done to address the specific associations. Text mining approaches, therefore, have the inherent capacity to help speed-up the rate of biological discovery.

In this chapter we present several text-mining systems developed in our Knowledge Extraction Lab at the Institute for Infocomm Research, Singapore, two of which are the result of an on-going collaboration with Department of Biological Sciences, National University of Singapore. We show how these systems can assist an average (nonexpert) user to better understand specific problems in biology and bring them closer to the answers about functions of biological entities inferred on the basis of an in silico method. Before we present these systems, we also define the problem we intend to deal with and describe some of the general features that text-mining system should provide to the end-users.

Scope and Nature of Text-Mining in Biomedical Domain

By automated knowledge extraction, we understand an automated extraction of names of entities, such as genes, gene products, metabolites, pathways, etc., which appear in biomedical and molecular biology literature, as well as the relationships between these entities. The basic relation between two entities is characterized by the cooccurrence of their names in the same document, or in a specific segment of the document. However, the actual relation between these entities is not easy to characterize by the computer program. It is customary to leave it to the user to assess the actual nature of such relations based on the associated documents. To the best of our knowledge, very few text-mining systems exist which can accurately extract such relations.

Characteristics of Text-mining Systems

There are several basic features that text-mining systems should provide. These systems should:

a. be easy to use;
b. be interactive;
c. allow several ways of submitting data;
d. allow user to select categories of terms to be used in the analysis;
e. provide suitable interactive summary reports;
f. show association maps in suitable graphical format;
g. preferably have built-in intelligence to filter out irrelevant documents;
h. preferably be able to extract large-volume of useful information in reasonable timeframes.

While, in principle, any free text document can be analyzed, for the purposes of discussion here, we will assume that documents are abstracts of scientific articles, such as those contained in PubMed[1] repository. Then, generally speaking, there could be three levels at which text analysis can be conducted: the 'Abstract level', the 'Sentence level', the 'Relation level'. At the 'Abstract level', the system analyzes the whole abstract to determine if it contains relations between the utilized biological entities or not. At the 'Sentence level', the system assesses whether the abstract analyzed contains sentences that explicitly claim relations between the entities or