Chapter 7.16
Mining BioLiterature:
Toward Automatic Annotation of Genes and Proteins

Francisco M. Couto
Universidade de Lisboa, Portugal

Mário J. Silva
Universidade de Lisboa, Portugal

ABSTRACT

This chapter introduces the use of Text Mining in scientific literature for biological research, with a special focus on automatic gene and protein annotation. This field became recently a major topic in Bioinformatics, motivated by the opportunity brought by tapping the BioLiterature with automatic text processing software. The chapter describes the main approaches adopted and analyzes systems that have been developed for automatically annotating genes or proteins. To illustrate how text-mining tools fit in biological databases curation processes, the chapter presents a tool that assists protein annotation. Besides the promising advances of Text Mining of BioLiterature, many problems need to be addressed. This chapter presents the main open problems in using text-mining tools for automatic annotation of genes and proteins, and discusses how a more efficient integration of existing domain knowledge can improve the performance of these tools.

INTRODUCTION

Bioinformatics aims at understanding living systems using biological information. The facts discovered in biological research have been mainly published in the scientific literature (BioLiterature) since the 19th century. Extracting knowledge from such a large amount of unstructured information is a painful and hard task, even to an expert. A solution could be the creation of a database where authors would deposit all the facts published in BioLiterature in a structured form. Some generic databases, such as UniProt, collect and distribute biological information (Apweiler et al., 2004). However, different communities have different needs and views on specific topics, which

change over time. As a result, researchers do not look only for the facts, but also for their evidence. Before a researcher considers a fact as relevant to his work, he checks the evidence presented by the author, because facts are normally valid only in a specific context. This explains why Molecular Biology knowledge continues to be mainly published in BioLiterature. Another solution is Text Mining, which aims at automatically extracting knowledge from natural language texts (Hearst, 1999). Text-mining systems can be used to identify the following types of information: entities, such as genes, proteins and cellular components; relationships, such as protein localization or protein interactions; and events, such as experimental methods used to discover protein interactions. Bioinformatics tools to collect more information about the concepts they analyze also use Text Mining. For example, information automatically extracted from the BioLiterature can improve gene expression clustering (Blaschke, Oliveros, & Valencia, 2004).

Text Mining of BioLiterature has been studied since the last decade (Andrade & Valencia, 1998). The interest in the topic has been steadily increasing, motivated by the vast amount of publications that curators have to read in order to update biological databases, or simply to help researchers keep up with progress in a specific area. Text Mining can minimize these problems mainly because BioLiterature articles are quite often publicly available. The most widely used BioLiterature repository is MEDLINE, which provides a vast collection of abstracts and bibliographic information. For example, in 2003, about 560,000 citations have been added to MEDLINE. Reading 10 of these documents per day, it would take around 150 years to read all the documents added in 2003. Moreover, the number of new documents added per year increased by more than 20,000 from 2000 to 2003. Hence, text-mining systems could have a great impact in minimizing this effort by automatically extracting information that can be used for multiple purposes and could not possibly be organized by other means.

This chapter starts by providing broad definitions used in Text Mining and describes the main approaches. Then, it summarizes the state-of-the-art of this field and shows how text-mining systems can be used to automatically annotate genes or proteins. Next, the chapter describes a tool designed for assisting protein annotation. Finally, the chapter discusses future and emerging trends and presents concluding remarks.

TEXT MINING

Text Mining aims at automatically extracting knowledge from unstructured text. Usually the text is organized as a collection of documents, or corpus.

TextMining = NLP + DataMining

Data Mining aims at automatically extracting knowledge from structured data. (Hand, Mannila, & Smyth, 2000). Thus, Text Mining is a special case of Data Mining, where input data is text instead of structured data. Normally, text-mining systems generate structured representations of the text, which are then analyzed by Data Mining tools. The simplest representation of a text is a vector with the number of occurrences of each word in the text (called a bag-of-words). This representation can be easily created and manipulated, but ignores all the text structure. Text-mining systems may also use Natural Language Processing (NLP) techniques to represent and process text more effectively. NLP is a broad research area that aims at analyzing spoken, handwritten, printed, and electronic text for different purposes, such as speech recognition or translation (Manning & Schütze, 1999). The most popular NLP techniques used by text-mining systems include: tokenization, morphology analysis, part-of-speech tagging, sense disambiguation, parsing, and anaphora resolution.

Tokenization aims at identifying boundaries in the text to fragment it into basic units called