TEXTMINING IN PREDICTING GENE-CANCER RELATIONSHIPS

AUTHORS

1.Aischarya Brahma
3rd year B.tech biotechnology
(tulsibrahma@yahoo.com)

2. Garima Chaudhry
3rd year B.tech biotechnology
(garima_2910@yahoo.com)

SCHOOL OF BIOTECHNOLOGY, CHEMICAL AND BIOMEDICAL ENGINEERING,
VELLORE INSTITUTE OF TECHNOLOGY UNIVERSITY,
TAMIL NADU, INDIA.

ABSTRACT

The NCBI describes text mining as a modular process involving document categorization, named entity tagging, fact and information extraction, and collection-wide analysis. In document categorization, a subset of potentially relevant documents is retrieved to increase the efficiency of subsequent steps. Named entity tagging identifies the important entities or objects mentioned in the article, often using a list of synonyms. Fact and information extraction identifies the relationships between entities. Finally, in collection-wide analysis, information extracted from different documents is integrated. Thus text mining involves the application of techniques from areas such as information retrieval, natural language processing, information extraction and data mining.

Cancers are complex diseases with multiple genetic and environmental factors contributing to their development. Studies require logical hypotheses regarding the genes to be tested and clear criteria for case definition. Cases may be defined as people who have any of several types of cancer, if those types are related. For example, epidemiologic studies of BRCA1 mutation carriers might benefit from information collected about both breast and ovarian cancer cases. In several ways, groups of cancers that have shared genetic factors are anticipated to lead to further etiologic hypotheses and advances regarding environmental agents. Using text mining, firstly grouping cancers will be especially useful if a group combines several cancers that are rare and difficult to study individually. Secondly, knowledge of genetic pathways might suggest

an environmental factor associated with all of the cancers. For example, a grouping defined by a vitamin receptor gene would suggest vitamin intake as a possible environmental agent in the etiology of all of the cancers. Thirdly, text mining will allow us to design studies that might extend gene-cancer associations to include cancers at other sites.

There is a serious need for enhanced technology in this area, including tools which facilitate the extraction of semantic information from databases and texts, the establishment of standardized vocabularies and terminologies, the codification of these semantic relations between terms in ontologies of concepts connected by semantic relations, and the merger and linkage of such ontologies once created. This paper motivates the need for a concerted effort in the area of integration of biological knowledge, both by outlining the genes related to a particular type of cancer and also various types of cancer related to a particular gene.