Tech Review

Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery

The implementation of precision medicine would transform how some diseases are treated and prevented. Understanding the underlying mechanism of disease sits as a barrier in implementing precision medicine. However, utilizing the techniques of text and data mining researchers are slowly making good headway and tackling the challenges as they arise. Data collection is the first step in the discovery process, translating the data using the techniques of analytical and computational methods is the second step.

Data mining involves the collection of large amounts of data mainly referred to as big data, the data collected is then applied through mining algorithms to generate meaningful information. Data mining methods used include classification algorithms, frequent pattern algorithms, and clustering algorithms. With the amount of data exponentially growing over the last decade or so, these novel algorithms will provide a breakthrough in the implementation of precision medicine.

Text mining on the other hand is applied to unstructed sources. Using the techniques of text mining, researchers can extract insightful information from a document within a collection using methods including PLSA (Probabilistic latent semantic analysis) and LDA (Latent Dirichlet Allocation). NER (named entity recognition) is one of the ways to extract and identify data from the text and tag it based on the category. This can be applied to genes, drugs, and diseases making It easier to identify in a document.

Some Applications:

Gene Prioritization:

Diseases with diverse symptoms are caused by multiple genes. Many gene prioritization methods are used to determine the causative genes, some of these include (GeneWanderer, GeneSeeker, GeneProspector). We use similarity scores to rank the genes to look for causative genes.

Precision Medicine:

Building treatment and prevention strategies based on person's genes is the idea behind precision medicine. There has been a lot of growth in this area over the last few years, because of the ease of data access. Using the techniques of data/text mining, there have been a bunch of advancements made in this field. One of the examples would be to cure and treat cancer, genetic data is used.

Pharmacogenomics:

The study of how genes effect on how we react to medicinal drugs. There has been a significant use of data/text mining in understanding how drugs are related to genes and diseases. There is a PharmGKB database where all the pertinent Pharmacogenomics information is stored. The study of Pharmacogenomics has made significant progress, and we are getting close to seeing the vision of personalized medicine turn into reality.

Toxicology:

To study the impact of chemical/biological substances on humans, the techniques of data and text mining plays an important role in determining the potential toxic substances. To understand this, in addition to large amount of data, methods such as HTS (High Throughput

Screening) are used. Several studies have been done to understand the chemical induced toxicity using context specific algorithms and unsupervised methods.

In conclusion, there have been few recent advancements in the field of biomedical science utilizing the techniques of data and text mining. However, in today's world we are lacking the tools needed to properly analyze data and support timely decisions. Even though we have massive amount of data available to us, the data mining tools are not efficient to integrate all the data sources and provide intelligent information. There remains a major challenge of integrating the structured and unstructured data sources to provide even further meaningful insights.

References:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4719073/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6230431/