ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(33), DOI: 10.17485/ijst/2016/v9i33/99620, September 2016* ISSN (Online) : 0974-5645

Generating Drug-Gene Association for Vibrio Cholerae using Ontological Profile Similarity

C. Geethanjali*and S. Bhanumathi

Department of Computer Science and Engineering, Sathyabama University, Chennai - 600119, Tamil Nadu, India; Geethanjali.jc@gmail.com, banujun8@gmail.com

Abstract

Objective: The measure of biomedical literature has been expanding quickly in the most recent decade. Text mining systems can analysis this vast scale information, shed light onto complex medication instruments and concentrate connection data that can bolster computational polypharmacology. The key idea of the paper is to find the drug candidate for the disease cholera, which is caused due to the organism Vibrio cholerae. **Method:** The technique estimates the Pointwise Mutual Information (PMI) among protein name obtained from the UniprotKB and the Medical Subject Headings that contain drug terms. Based on the PMI scores, gene/protein profiles and drug are produced and candidate drug-gene/protein associations are constructed when evaluating the relevancy of their profiles. **Findings:** The association between protein and drug is found and the drug candidates are proposed. **Applications:** The similar technique can be applied to find the drug candidate for various diseases which reduces the drug release cost in the market.

Keywords: Association, Corpus, MEDLINEDatabase, MESH, Ontology, Text Mining

1. Introduction

The recent research on biomedical literature widely includes the procedure of text mining. Text mining is used to get highly useful information by analyzing the biomedical data. Biomedical content in databases like MEDLINE comprises of trial results identified with different organic elements like protein, drugs, gene and diseases. Mining of biomedical data to comprehend the connection available on these organic elements from the corpus gets to be fundamental. Text mining has turned into a significant device for breaking down biomedical literature and extracting useful information[1]. The extraction of valuable data on organic substances with protein name and its related disease from biomedical text data.

The center of this paper is to consequently separate connection among protein, drug and disease for cholera. Text mining is characterized as the programmed revelation of new, already obscure data from unstructured literary information[2]. The existing system implies the drug-gene association in general since there is no such research carried out of finding the drug candidate for the disease cholera, the proposed work is predicted drug candidates through the protein drug similarity.

The paper is organized as follows. Section 1 discusses literature related to gene-drug association and also analyzed the tools utilized for finding biomedical terms of the biomedical data. Section 2 explains the materials and methods for the gene/protein-drug association and drug candidates. Section 3 discusses the algorithm that is used to match protein names with indexed abstracts. Section 4 showed the results and Section 5 presents the conclusion.

This section discusses the literature for discovering the association among protein, drugs and diseases. The tools utilized for obtaining biomedical terms are investigated.

In[3] discussed the r-scaled scoring mechanism to calculate the hidden relationship among biomedical concepts. In[4] presented a new applications and extensions of supervised machine learning algorithms to digout relationships among drugs and genes automatically.

With this study new focuses for as of now marketed medications were tentatively affirmed. Lamb et al. fabricated a Connectivity Map and utilized comparable

* *Author for correspondence*

quality expression marks to interface medications, qualities and sicknesses[5]. Advanced a medication unfavorable occasion target system of 73 new off-targets and 656 promoted drugs with substance highlights and successive data[6].This technique prompted the trial affirmation of 125 novel drug-target collaborations. Different methodologies additionally utilized similitude in light of pharmacological impacts for producing in silico forecasts of drug-target collaborations[7].

There are a couple of unsupervised techniques towards the expectation of quality medication affiliations. Knox made a system of supposed semantically connected elements to sedate in light of openly accessible stores which contain drug-related data[8].

In[9] deduce quality drug affiliations through an administered system based approach and demonstrate that it performs better contrasted with drug based and target-based medication quality affiliation expectation,also use the thought of an incorporated system and learn worldwide and nearby components towards target distinguishing proof and medication repurposing[10].

In give a review on a few drug-target system applications and point to data fulfillment as the primary impediment towards the proficient recognizable proof of drug-target collaborations[11]. In this study, they concentrate on the forecast of medication quality affiliations. The proposed system uses the book reference to quantify corpus-based semantic relatedness between ontological terms. To the best of our insight, it is the initially unsupervised technique that predicts new medication quality affiliations exclusively by breaking down deliberately the co-event of biomedical terms in all the logical distributions filed by MEDLINE. Middle of the road ontological ideas are utilized to shape the connections in the middle of medication and qualities. Previously, the issue of building up backhanded connections between two ideas A and C through an arrangement of intermediate ideas of concept B has been tended by[12].

In[13], recommended the MedMiner tool for text mining. It is a popular online tool for mining biomedical literatures. The main purpose of this tool is disambiguate gene names from the biomedical text.

The implied method finds the co-occurrences of GO and Medical Subject Headings (MeSH). The titles and abstracts that related to Vibrio cholera are downloaded from the MEDLINE database using PubMed. The co-event data is utilized to rank the most related GO and MeSH Disease biomedical ideas to the medication and the quality individually. These ideas frame an individual profile for every medication and quality, which is used to survey the relationship between them by evaluating the level of the relatedness between their profiles. Also, the produced profiles can give an understanding into biomedical properties for medications and qualities and construe relationship between them that have not been incorporated into a database nor reported. To this end, we tentatively assess the methodology in organizing drug quality affiliations. The outcomes demonstrate that the co-occurrence based profiles execute on a similarity with the physically curates profiles.

2. Materials and Methods

Proteins are found in all living *organisms*. Problems occur when *proteins* become denatured. The drug is an important substance that used to protect people who suffered from different kinds of disease. The disease is some issue brought into the human life. The dataset gives the related information between gene and protein and their protein names and drugs. This helps to map a drug candidate or disease with its proteins and drugs.

2.1 Cholera an Overview

Cholera is a substance which has a therapeutic, intoxicating or other equivalent effect when given to a living structure. The great indication is a lot of watery loose bowels that endures a couple of days. Retching and muscle issues might likewise happen. Cholera is caused by various sorts of Vibrio cholerae, with a few sorts delivering more extreme illness than others. It is spread for the most part by water and food that has been polluted with human excrement containing the microscopic organisms. Cholera antibodies that are given by mouth give sensible protection to around six months. They have the included benefit of ensuring against another kind of looseness of the bowels brought about by E. coli. The essential treatment is oral rehydration treatment - the supplanting of liquids with marginally sweet and salty arrangements.

2.2 Generating Drug Candidates

With the hierarchical progress to the bio medical writing the proposal has been made in a way that the usage of the

proper open NLP in the abstract of all the related datasets towards the disease and along with the protein mixtures had made betterment in the drug prediction. Thus the Indexing process tends to allocate the use of all the related total set of items in the dataset and pre-processing in to be done so that the finding of all the recent and related information towards the package. Information gathering techniques are regularly approximately controlled, bringing about out-of-extent qualities, analyzing information that has not been precisely screened for such issues can deliver deceiving results. Accordingly, the representation and nature of information is above all else before running an investigation. These procedures expels clamor from information; standardization, which sorts out information for more proficient get to; and highlight extraction, which hauls out determined information that is noteworthy in some specific setting. The PMI score is calculated by the method used in. The system architecture is shown in Figure 1 gives details about how the drug candidate and drug-protein association is predicted.

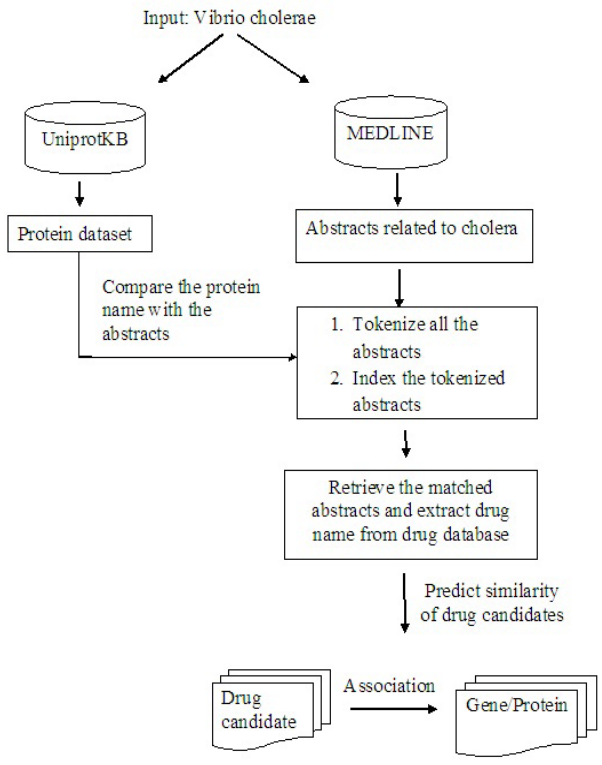

Figure 1. System architecture.

2.3 Modules Description

- Data collection.
- Preprocessing the data.
- Tokenization.
- Indexing.
- Retrieving the drug names from matched abstracts.
- Drug candidate prediction.

2.3.1 Data Collection

- Input Protein Set: The protein dataset is obtained from UniprotKB database**.** The input file for the manipulation of the data analyzing is the protein sets which does have all the protein set and all the components needed and this set are present in the form of rows and columns of an Excel sheet. Thus the affective need for the combination of the protein set along with the needed molecules been stored for the prediction purpose.
- Abstract Dataset: All the abstracts related with the disease cholera is extracted from PubMed database. Abstract data set is the place where the information and the disease is been kept thus it have so much in number of information which are in need and are not. Thus, the data set is present in the XML file. Thus, XML parser is needed to retrieve all the information from the xml file where all the data to be pre-processing and other process can be done**.** A XML Parser is a parser that is intended to peruse XML and make a path for projects to utilize XML.

2.3.2 Preprocessing

Preprocessing is an information mining system that includes changing crude information into a justifiable arrangement. True information is regularly inadequate, conflicting, and/or ailing in specific practices or drifts, and is liable to contain numerous blunders. Preprocessing is a demonstrated strategy for determining such issues. Preprocessing plans crude information for further preparing.

- Data Cleaning: Information is scrubbed through procedures, for example, filling in missing qualities, smoothing the boisterous information or determining the irregularities in the information.
- Data Integration: Information with various

representations is assembled and clashes inside of the information are determined.

- Data Transformation: Information is standardized, collected and summed up.
- Data Reduction: This stride plans to show a lessened representation of the information in an information distribution center.
- Data Discretization: Includes the diminishment of various estimations of a ceaseless property by isolating the scope of quality interims.

The OpenNLP library is a machine learning based toolbox for the handling of characteristic dialect content. It bolsters the most well-known NLP assignments, for example, tokenization, sentence division, grammatical feature labelling, named element extraction, lumping, parsing and co reference determination. These undertakings are generally required to construct more propelled content preparing administrations. OpenNLP additionally included greatest entropy and perception based machine learning.

The objective of the OpenNLP venture will make an experienced toolbox for the aforementioned undertakings. An extra objective is to give an extensive number of pre-fabricated models for an assortment of dialects and additionally the commented on content assets that those models are derived from.

- Indexing:At the point when indexing an archive, the indexer likewise needs to pick the level of indexing exhaustively, the level of subtle element in which the report is depicted. All in all the higher the indexing exhaustively, the more terms filed for every report. The class of filed dialects has reasonable significance in characteristic dialect handling as a computationally moderate speculation of setting free dialects, since filed sentence structures can depict a significant number of the nonlocal imperatives happening in normal languages. To help both indexers and searchers in utilizing and controlling the system. Helps clients separate in the middle of terms and decreases equivocalness in the dialect.
- Tokenization:Tokens are consistently inaccurately suggested as terms or words, notwithstanding it is on occasion fundamental to make a sort/token capability. A token is an event of a course of action of characters in some particular document that are assembled as an accommodating semantic unit for planning. A class of all tokens containing same characters is known as Type. A term is a sort that is incorporated into the IR framework's lexicon. The arrangement of list terms could be totally unmistakable from the tokens, for occasion, they could be semantic identifiers in a scientific classification, yet practically speaking in current IR frameworks they are emphatically identified with the tokens in the archive. In any case, instead of being precisely the tokens that show up in the report, they are typically gotten from them by different standardization forms.

2.3.3 Retrieving Drug Name from Matched Abstracts

The lists of all the drugs are obtained from the drug database. After the abstracts are preprocessed and tokenized, the abstracts are placed in a separate database. Thus the tokenized abstract is checked for the match of proteins which is obtained from the UniprotKB.

2.3.4 Drug Prediction

Along these lines with all the further above philosophies and handling this procedure helps in foreseeing the medications for the infection which expands the precision of all the diverse and esteemed dataset.

2.3.5 Algorithm Steps

Step 1: Retrieve the protein set from UniprotKB database and abstracts from MEDLINE database.
Step 2: Preprocess the abstracts by tokenizing and indexing with the help of OpenNLP.
Step 3: Retrieve the matched abstracts and extract drug name from drug database.
Step 4: Genetic Algorithm is implied to match the protein name with the indexed abstracts.
Step 5: Based on the similarity, drug candidate is predicted.

3. Genetic Algorithm

Genetic Algorithm (GA) is a stochastic inquiry technique which has been generally utilized by the information digging group for finding characterization rules. The exactness of the standards that GA finds is practically identical and a few times considerably more precise than the tenets acquired by the other grouping calculations.

3.1 Uses of Genetic Algorithm

- It considers each rule set for a class as a tiny component. It mines and adds rules independently to these components whenever the data distribution changes.
- This avoids unnecessary mining activity there by reducing learning cost and this makes the algorithm scalable with respect to size.
- The efficiency of the algorithm is proved by using standard data sets.
- In GA adaptability is for the most part tended to by parallel making so as to handle or the answer for focalize rapidly and thereby decreasing the quantity of era which thusly lessens the learning time.
- They create similarly precise results and they likewise impressively decrease the computational time.

4. Results

The displayed strategy uses the co-events of GO and network disease ideas with medications in MEDLINE titles and abstracts. In light of that co-event data, profiles of ontological terms are made for both medications and qualities. At that point, with the assistance of a corpus-based factual measure, PMI, we relate medications to qualities by surveying the semantic relatedness of their profiles.

Thus, Figure 2 implies the drug-gene association of the drug cholix and the protein acetaldehyde dehydrogenase which signifies that the drug candidate cholix can be predicted for the cure of cholera.

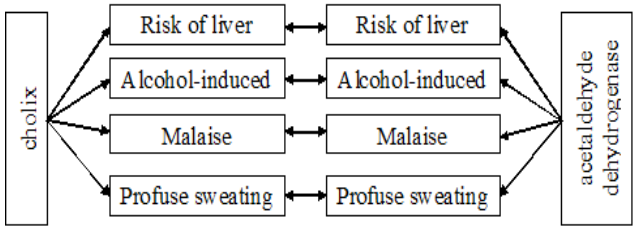

Figure 2. Gene/protein-drug association.

5. Conclusions

Accuracy of the system can be expected to a full extend for the usage of OpenNLP. Thus, the pre-processing improves and gets the perfect specifies term. Indexing helps in easier retrieval of matched abstracts. Thus, the overall system predicts the drug candidate with the ontological similarity between the drug and the protein of the organism Vibrio cholera.

6. References

1. Kissa M, Tsatsaronis G, Schroeder M. Prediction of drug gene associations via ontological profile similarity with application to drug repositioning. Methods. 2015 Mar; 74:71–82.
2. Weeber M. Advances in literature-based discovery. Journal of the American Society for Information Science and Technology. 2003; 54(10):913–25.
3. Takarabe M, KoteraM, Nishimura Y, Goto S, Yamanishi Y. Drug target prediction using adverse event report systems: Apharmacogenomic approach. Bioinformatics. 2012; 28(18):611–8.
4. Chang J. Using machine learning to extract drug and gene relationships from text. [Doctoral Dissertation]. 2004.
5. Yamanishi Y, Kotera M, Kanehisa M, Goto S. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatics. 2010; 26(12):246–54.
6. Chen B, Ying D, David W. Assessing drug target association using semantic linked data. PLoS Computational Biology. 2012 Jul; 8(7):1–10.
7. Wu Y, Liu M, Zheng WJ, Zhao Z, Xu H. Ranking gene-drug relationships in biomedical literature using latent dirichlet allocation. Pacific Symposium on Biocomputing; 2012. p.422–33.
8. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, DjoumbouY. DrugBank 3.0: A comprehensive resource for omics research on drugs. Nucleic Acids Research. 2011; 39:1035–41.
9. Cheng F, Liu C, Jiang J, Lu W, Li W, Liu G, Zhou W, Huang J, Tang Y. Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Computational Biology.2012 May; 8(5):1–12.
10. Emig D, Ivliev A, Pustovalova O, Lancashire L, Bureeva S, Nikolsky Y, Bessarabova M. Drug target prediction and repositioning using an integrated network-based approach. PLoS One. 2013; 8(4):606–18.
11. Vogt I, Jordi M. Drug-target networks. Molecular Informatics, 2010; 29(2):10–4.
12. Srinivasan P. Text mining: Generating hypotheses from MEDLINE. Journal of the American Society for Information Science and Technology. 2004 Mar; 55(5):396–413.
13. Vazquez M, Krallinger M, Leitner F, Valencia A. Text mining for drugs and chemical compounds: Methods, tools and applications. Molecular Informatics. 2011; 30(6-7):506–19.