ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

A Framework of Protein-Drug Association for Malaria by Text Data Mining of Biomedical Literature.

Bhanumathi Selvaraj[1]* and Sakthivel Periyasamy[2].

[1]Department of Computer Science and Engineering, Sathyabama University, Chennai-600119, India.
[2]Department of Electronics and Communication Engineering, Anna University, Chennai-600025, India.

ABSTRACT

The protein-drug association, the finding of the important relationship between diseases related proteins and drugs, provides useful information for the discovery of new drugs. In this paper, we first time propose a new framework to find an association of protein-drug for human malaria parasite *Plasmodium falciparum* using text data mining approaches. The framework begins with three phases of text data mining: (1) data collection from MEDLINE, UniprotKB and MeSH databases; (2) data pre-processing such as tokenization, stop word removal and stemming; (3) data analysis to retrieve actual information and extract useful information. Finally, regularized log-odds function is used to create an association matrix between *Plasmodium falciparum* proteins and their drug terms. The proposed framework could be useful for a new drug candidate discovery for malaria.

Keywords- Text data mining, biomedical literature, protein-drug association, malaria.

**Corresponding author*

INTRODUCTION

Malaria, a life-threatening infectious disease, causes death of over one million people every year, mostly less than 5 years old children [1]. In 2015, 90% of malaria deaths and 88% of malaria cases were found in Sub-Saharan Africans. According to the recent report of World Health Organization, 3.2 billion people were at risk of malaria [2]. In addition, the Center for Disease Control and Prevention reported 1500 malaria cases per year in the United States [3].

Malaria is caused in human by five Plasmodium parasites: *Plasmodium falciparum*, *Plasmodium malariae*, *Plasmodium vivax*, *Plasmodium knowlesi* and *Plasmodium ovale*. Among them, the most dangerous form of malaria is caused and death occurred by *Plasmodium falciparum* (*P. falciparum*) [4]. This parasite lives in tropical and subtropical climates and is transmitted by the bite of infected female Anopheles mosquito. The available vaccines for malaria are not effective to date. Therefore, chemotherapy and antimalarial control programmes are used for treatment and prevention of this disease, respectively [5]. However, chemotherapy-based treatments compromise their efficiency, drug resistance. Due to parasite resistance, some of the antimalarial drugs, for example, atovaquone [6], chloroquine [7], sulfadoxine [8] and pyrimethamine [9], have been withdrawn in several regions [10]. This highlights the urgent need for discovery of new antimalarial drugs and thus various efforts have been made including computer-aided drug discovery methods [11, 12].

The human genome project, structured chemical and biological data have accumulated numerous biomedical/biotechnical literature, which play important roles in the generation of new knowledge, such as finding specific relationships between entities, using text data mining approaches [13]. For example, Li *et al.* developed a novel framework to find new drug candidates for breast cancer using text and structured data mining of biomedical literature [14]. Singh-Blom *et al.* used two approaches, the Katz measure [15] and CATAPULT, to predict the gene-disease and gene-phenotype association, and showed that CATAPULT is better for identification of gene-disease association correctly [16]. In addition, Quan *et al.* reported a text mining approach to study gene-disease association using various approaches including maximum entropy classifier, probabilistic context-free grammars and network analysis [17]. Like Quan *et al* approach*,* Zhang et *al*. designed a new network-based computational method to extract pair-wise associations among genes, diseases and drugs from Semantic MEDLINE [18]. Recently, Yu *et al.* suggested a new method to find an association between drug-disease using protein complexes [19], and Kissa *et al.* integrated an unsupervised approach and text data mining concepts to derive drug-gene associations by investigating co-occurrence of drug and gene in MEDLINE articles [20]. Apart from these, several text mining tools, for example GenDisFinder [21], PolySearch2 [22], BeFree [23] and DISEASES [24], have been developed to extract protein/gene-disease and drug association from biomedical literature.

In this study, we propose first time a new framework to discover the protein-drug association for *P. falciparum* parasite using text data mining approaches. In this framework, the protein list and abstracts that are related to *P. falciparum* are first collected from the UniProtKB and MEDLINE databases, respectively. Subsequently, the relevant abstracts are retrieved based on the collected protein list. The drug terms are gathered from the MeSH database, and then they are used to obtain list of drug terms from the relevant abstracts. Finally, the association matrix is created between protein list and drug terms. The proposed framework may be useful for new drug candidate discovery for malaria.

MATERIALS AND METHODS

System architecture for drug-protein association

Figure 1 describes the system architecture that explains how the association matrix is formed for drug terms and protein list. The last five year's abstracts and titles that are relevant to *P. falciparum* were locally downloaded from the MEDLINE database [25]. The protein lists for *P. falciparum* were collected from the UniProtKB database [26]. Two types of protein lists are available in the UniProtKB database: UniProtKB/Swiss-Prot contains the reviewed proteins while UniProtKB/TrEMBL has the un-reviewed proteins. Only reviewed proteins were collected as they are frequently updated.

The text data mining applied on the collected abstracts to quickly identify the relevant abstract. Each protein name with its synonyms was used to retrieve relevant abstracts from the obtained abstracts of

MEDLINE. The drug terms were collected from chemicals and drug category and its subcategory in MeSH database [27]. The collected drug terms were matched with the relevant abstracts and final list of drug terms were found. Finally, association matrix is created between drug terms and protein list.

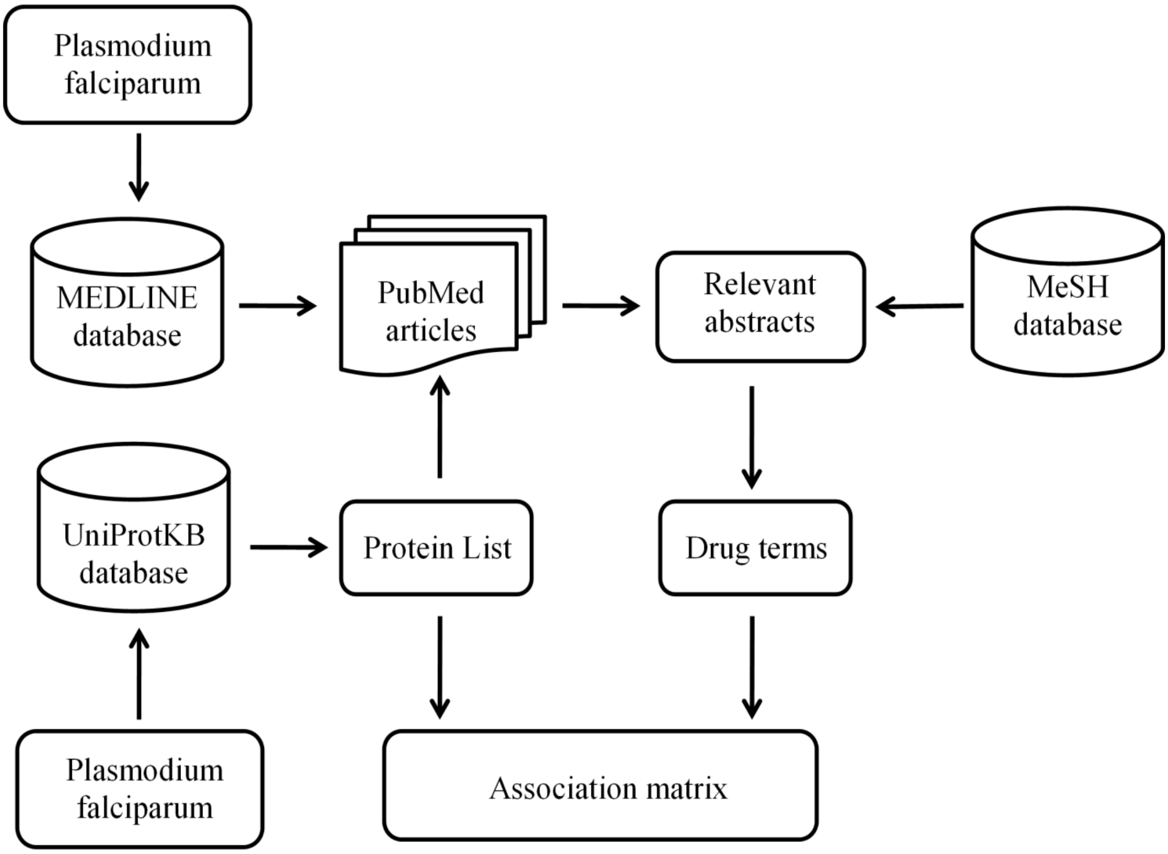

Figure 1. System architecture

Text data mining

Text data mining technique is often used to process unstructured textual data and extract the previously unknown useful knowledge from the biomedical literature [28, 29]. It comprises the following phases: text collection, text pre-processing and data analysis.

Text collection: The unstructured text data were gathered from biomedical literature for example MEDLINE [25] and databases for instance UniProtKB [26] and MeSH [27]. All these databases contain huge amounts of textual data in the form of natural language. The relevant documents were downloaded and stored as a corpus.

Text pre-processing: Text pre-processing is an important step in the text data mining, which is used to remove unwanted information of the corpus. Only title and abstracts of the documents were kept and all other information were removed. It allows use of most relevant information in the data analysis phase. The upper case alphabets of each document were converted into lower case alphabets. The following steps were performed in the pre-processing:

(1) *Tokenization*: The content of each document in the corpus was split into separate tokens.

(2) *Stop word removal*: The special symbols, numeric values and punctuation marks were removed from tokenized documents. Also, the stop words such as the, in, we, at, etc., were collected and stored into a list. The list was used to match the stop words with the tokens one by one and then matched tokens were deleted from the document.

(3) *Stemming*: It is used to identify the root words, for example, stemmer, stemmed, stems, stemming are based on the root word stem. For this, the popular Porter's stemmer algorithm [30, 31] was used.

Data analysis: In general, this phase involves the actual information retrieval and information extraction process to get useful information from the vast amounts of documents. In this study, the protein list, abstracts and drug terms are retrieved from various databases. Then, they are used to extract the significant drug terms which are helpful to form an association matrix with protein list.

Creation of association matrix between drug terms and protein lists

The association matrix is a two-dimensional table which allows one row and one column for each protein and each drug term, respectively. It is used to inspect the relationship among all the protein-drug pairs. For this, the regularized log-odds function was used [32, 14]. The association matrix is calculated based on the equation (1).

$$Score_{dtp} = \ln(tdf_{pdt} \times n + \varepsilon - \ln(tdf_p \times tdf_{dt} + \varepsilon) \qquad (1)$$

Where,
tdf_{dt}: represents the total number of documents that has drug terms
tdf_p: represents the total number of documents that has protein names
tdf_{dtp}: represents the total number of documents that has both drug terms and protein names
n: represents the total size of the MEDLINE documents
ε: small constant (ε=1).

For the proposed framework implementation, we used the Python 2.7.10 [33] and Natural Language Tool Kit (NLTK) [34].

RESULTS AND DISCUSSION

Construction of protein list for *P. falciparum*

Steps for constructing protein list are shown in Figure 2. The proteins for *P. falciparum* are retrieved from the UniProt Knowledgebase. There were 353 reviewed and manually annotated proteins in UniProtKB/Swiss-Prot. Subsequently, duplicates were removed, which resulted in 268 proteins. The synonyms and gene names were included with this protein list to retrieve relevant abstracts from the MEDLINE database. Some of identified *P. falciparum* proteins has been listed in Table 1 (Full list of proteins are available on request).

Table 1. Some of the proteins of *P. falciparum* retrieved from UniProtKB/Swiss-Prot

S. No	Entry	Entry name	Protein name(synonyms)	Gene names
1	P23746	ABRA_PLAFF	101 kDa malaria antigen (Acidic basic repeat antigen)	ABRA
2	Q9U8D3	PURA_PLAFA	Adenylosuccinate synthetase (AMPSase) (AdSS	Adss
3	P50492	AMA1_PLAF8	Apical membrane antigen 1 (Merozoite surface antigen)	AMA-1 PF83
4	Q94650	ARF1_PLAFA	ADP-ribosylation factor 1 (plARF)	ARF1 ARF
5	Q08853	ATC_PLAFK	Calcium-transporting ATPase (EC 3.6.3.8) (Calcium pump)	ATP6
6	Q8IBZ9	CRT_PLAF7	Putative chloroquine resistance transporter (PfCRT)	CG10 MAL7P1.27
7	Q8IHZ9	KC1_PLAF7	Casein kinase I (EC 2.7.11.1)	CK1 PF11_0377
8	P62344	CDPK1_PLAF7	Calcium-dependent protein kinase 1 (EC 2.7.11.1)	CPK1 CDPK1 PFB0815w
9	Q07785	CDC2H_PLAFK	Cell division control protein 2 homolog (EC 2.7.11.22)	CRK2 PK5
10	Q9N623	CRT_PLAFA	Chloroquine resistance transporter (*Pf*CRT)	CRT

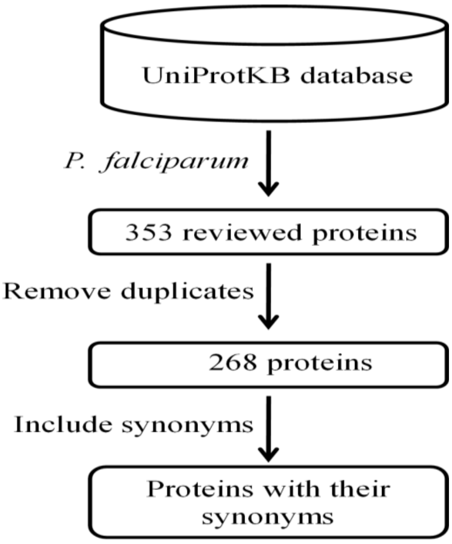

Figure 2. Construction of protein list

Retrieve MEDLINE abstracts and identify drug terms relevant to protein list

Plasmodium falciparum was used as a query to download the title and abstracts from the MEDLINE database through PubMed. The last five year's abstracts (8036) were locally downloaded. Text data mining process including tokenization, stop word removal stemming were carried out. The abstracts relevant to each protein were retrieved, which resulted in 1554 abstracts. To identify drug terms relevant to protein list, chemicals and drug category terms were downloaded from MeSH database (Figure 3), which is grouped into 16 categories [27]. All terms of MeSH database were searched with the 1554 abstracts and 80 drug/MeSH terms out of 292 were identified. The occurrence of the identified drug terms for *P. falciparum* parasite is shown in Table 2. The inorganic chemicals occurred 373 times in 1554 abstracts, indicating that most of the reported drugs for *P. falciparum* are inorganic chemicals. The association between the protein list and identified drug terms can be constructed by two-dimensional matrix, in which protein names and drug terms will be listed in rows and columns, respectively. The matrix is filled with the score obtained using the equation (1). The protein-drug pair is over- and under-represented based on the positive and negative value of the score$_{dtp}$, respectively. So, the higher score is more significant in the matrix [14].

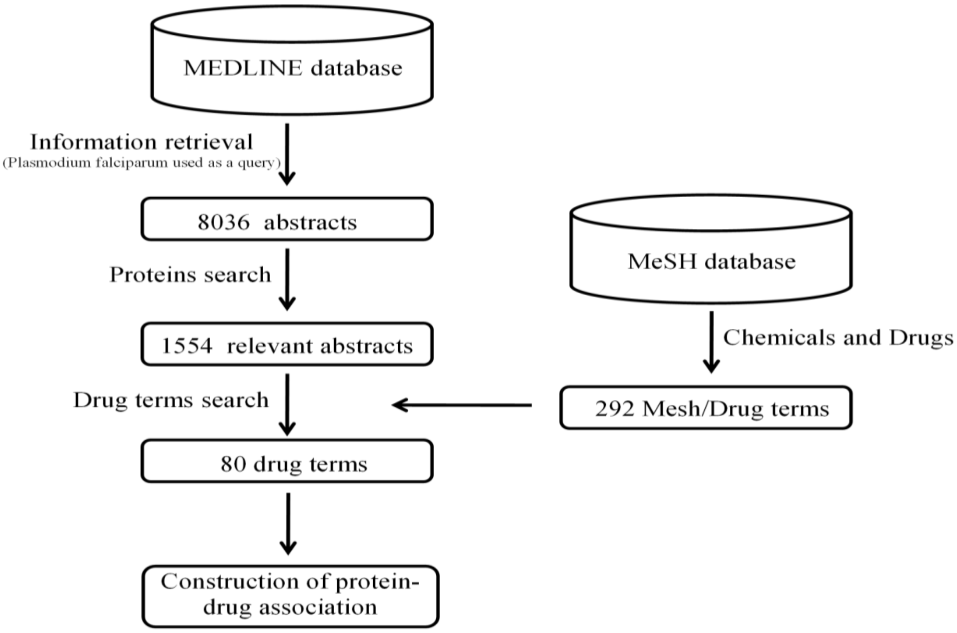

Figure 3. Workflow for drug term identification

Table 2. Frequencies of drug terms identified from abstracts

S. No.	Category	Total
1	Inorganic chemicals	373
2	Organic chemicals	196
3	Heterocyclic compounds	66
4	Polycyclic compounds	4
5	Macromolecular substances	2
6	Hormones, hormone substitutes, and hormone antagonists	0
7	Enzymes and coenzymes	95
8	Carbohydrates	3
9	Lipids	9
10	Amino acids, peptides, and proteins	50
11	Nucleic acids, nucleotides, and nucleosides	9
12	Complex mixtures	14
13	Biological factors	139
14	Biomedical and dental materials	223
15	Pharmaceutical preparations	255
16	Chemical actions and uses	0

CONCLUSION

In the past few decades, several advances have been made for the treatment and control of malaria. However, there is still extreme need to discover and improvement of effective antimalarial drugs due to the serious drug-resistance of malaria parasites. To the best of our knowledge, there are no researches on the protein-drug association for *P. falciparum*. In this study, we first time reported a framework to understand protein-drug association for *P. falciparum*. The proteins and drug terms related to *P. falciparum* were extracted from the biomedical literature using text data mining approaches. The protein-drug association matrix could be helpful to gain knowledge for novel drug candidate discovery for malaria.

ACKNOWLEDGEMENTS

BS acknowledges the Sathyabama University for the support.

REFERENCES

[1] http://www.netsforlifeafrica.org/malaria/malaria-statistics

[2] http://www.who.int/mediacentre/factsheets/fs094/en/

[3] http://www.cdc.gov/malaria/

[4] Perlmann P, Troye-Blomberg M. Folia Biol (Praha). 2000; 46(6): 210–8.

[5] Moorthy VS, Ballou WR. Malar J 2009; 8: 312.

[6] Srivastava IK, Morrisey JM, Darrouzet E, Daldal F, Vaidya AB. Mol Microbiol. 1999; 33: 704–711.

[7] Wellems TE and Plowe CV. J Infect Dis. 2001; 184: 770–776.

[8] Triglia T, Wang P, Sims PF, Hyde JE, Cowman AF. EMBO J. 1998; 17:3807–3815.

[9] Wu Y, Kirkman LA, Wellems TE. Proc Natl Acad Sci USA. 1996; 93: 1130–1134.

[10] Dondorp AM, Yeung S, White L, Nguon C, Day NP, Socheat D, von Seidlein L. Nat Rev Microbiol. 2010; 8:272–280.

[11] Elumalai P, Farah EM, Sergio W, Carmen de K, Margaret AP, Kelly C. J Chem Inf Model. 2016.

[12] Peter MN, Eric MG, Elumalai P, Kelly C. *ACS Infect Dis.* 2016; 2 (1): 8–31.

[13] Robert M, Michael H. DDT. 2002; (7) 11 (Suppl.): S89-S98.

[14] Li J, Zhu X, Chen JY. Int J Data Mind Bioin. 2010; 4(3): 241-255.

[15] Katz L. Psychometrika. 1953; 18: 39-43.

[16] Singh-Blom UM, Natarajan N, Tewari A, Woods JO, Dhillon IS, Marcotte EM. PloS one 2013; 8(5): e58977.

[17] Quan C, Ren F. Proc of the 5th International Workshop on Health Text Mining and Information Analysis. Louhi, EACL, 2014; 54-63.

[18] Zhang Y, Tao C, Jiang G, Nair AA, Su J, Chute CG, Liu H. J Biomed Semantics. 2014; 5: 33.

[19] Yu L, Huang J, Ma Z, Zhang J, Zou Y, Gao L. BMC Med Genomics. 2015; 8: S2.

[20] Kissa M, Tsatsaronis G, Schroeder M. Methods 2015; 74:71-82.

[21] Subramani S, Natarajan J. An Integrated Text Mining System based on Network Analysis for Knowledge Discovery of Human Gene-disease Associations (GenDisFinder). Proc of the fifth BioCreativeChallenge Workshop 2015; pp. 426-434.

[22] Liu Y, Liang Y, Wishart D. Nucleic Acids res; 2015; gkv383.

[23] Bravo À, Piñero J, Queralt-Rosinach N, Rautschka M, Furlong LI. BMC Bioinformatics. 2015; 16(1): 1.

[24] Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ. Methods 2015;74: 83-89.

[25] http://www.ncbi.nlm.nih.gov/pubmed/

[26] http://www.uniprot.org/uniprot/

[27] http://www.ncbi.nlm.nih.gov/mesh

[28] Hearst MA. Untangling Text Data Mining. Proc of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics 1999; pp. 3-10.

[29] Bakthavatsalam K, Bhuvaneswari V. A Study and Analysis of Gene Drug Association for Diabetic Gene-A Text Mining Approach. Int. Conf. on Intelligent Computing Applications (ICICA), 2014; pp. 416-420.

[30] Porter MF. Snowball: A Language for Stemming Algorithms. 2001.

[31] Porter MF. An Algorithm for Suffix Stripping. Program 1980; 14(3): 130-137.

[32] Korbel JO, Doerks T, Jensen LJ, Perez-Iratxeta C, Kaczanowski S, Hooper SD, Andrade MA, Bork P. PLoS Biol. 2005; 3(5): e134.

[33] https://www.python.org/downloads/release/python-2710/

[34] http://www.nltk.org/