Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 165–168

ISBN 978-83-60810-22-4

Discovering similarities for the treatments of liver specific parasites

Pınar Yıldırım
Okan University, Faculty of Engineering and Architecture, Department of Computer Engineering, Tuzla Campus, 34959,Akfırat, Tuzla, Istanbul, Turkey
Email: pinar.yildirim@okan.edu.tr

Kagan Ceken
Akdeniz University, Department of Radiology, Dumlupınar Bulvarı, Antalya, Turkey
Email:kceken@akdeniz.edu.tr

Osman Saka
Akdeniz University, Department of Biostatistics and Medical Informatics, Dumlupınar Bulvarı, Antalya, Turkey
E-mail:saka@akdeniz.edu.tr

***Abstract*—Medline articles are rich resources for discovering hidden knowledge for the treatments of liver specific parasites. Knowledge acquisition from these articles requires complex processes depending on biomedical text mining techniques. In this study, name entity recognition and hierarchical clustering techniques were used for advanced drug analyses. Drugs were extracted from the articles belonging to specific time periods and hierarchical clustering was applied on parasite and drug datasets. Hierarchical clustering results revealed that some parasites have similar in terms of treatment and the others are different. Our results also showed that, there have not been major changes in the treatment of liver specific parasites for the past four decades and there are problems associated with the development of new drugs. Both pharmaceutical initiatives and healthcare providers should investigate major drawbacks and develop some strategies to overcome these problems.**

***Index Terms*—Biomedical Text Mining, Clustering Analysis, Liver, Parasite.**

I. Introduction

MEDLINE articles are rich resources for discovering and tracking medical knowledge. Biomedical text mining techniques play an important role to acquire knowledge from these articles and they are applied for numerous studies in biomedical domain so far.

Parasitic diseases affect hundreds of millions of people worldwide and result in significant mortality and devastating social and economic consequences [1]. Parasitic diseases are especially harmfully on the liver which supports almost every organ and liver specific parasites also affect many people [2].

In this study, we developed a knowledge discovery on the treatment of parasites affecting liver and we hope that our study makes substantial contributions to all scientists and medical experts working on parasites affecting the efficiency of the liver's mechanism.

II. Methods

We used the Medline distribution available through the PubMed Web portal at the National Library of Medicine (NLM)[1] as well as on the in house distribution at the EMBL-EBI[2]. In the first phase of our article analysis procedure, two clinicians proposed a list of species which induce liver-specific diseases. They also proposed classes of drugs that could be used in the treatment of these diseases.

Medline is a collection of biomedical documents and administered by the National Center for Biotechnology Information (NCBI) of the United States National Library of Medicine (NLM). PubMed web site provides a service of the National Library of Medicine that include over 20 millions bibliographic citations from Medline and other life science journals for biomedical articles back to 1950s. The full text of articles are not stored; rather, links to the provider's site to obtain the full-text articles are given, is available [3-4].

TABLE I
NUMBER OF ARTICLES FOR SELECTED PARASITES

Parasite	Number of Articles
Clonorchis Sinensis	178
Echinococcus Multilocularis	229
Echinococcus granulosus	400
Entamoeba histolytica	1075
Fasciola Hepatica	917
Schistosoma Japonicum	446
Schistosoma Mansoni	1731
Opisthorchis Viverrini	213
Total	5189

Medline abstracts are in XML format and they contain logical markup to organize meta information such as the journal, author list, affiliations, publication dates and related MeSH headings [5].

We used a complex query for the retrieval of the Medline abstracts that are relevant for the liver specific parasites. The query resulted in a document set of 17,377 articles and all articles were processed with the text minig solution available at

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

[2] http://www.ebi.ac.uk/citexplore/

the EBI (European Bioinformatics Institute) called "Filter Server". In EBI architecture, a filter server specializes in recognizing the vocabulary of a particular terminology and receives a stream of text and annotates it with XML tags [6].

In our study, the filter servers identified the species mentions and its variants and the mention of drugs in the text. Species in the articles were annotated and parasites' names affecting liver were selected by two medical doctors. The frequencies of parasites were calculated. The frequency of the parasite provides the number of times a considered parasite appeared in the selected articles. Most ranked eight of them were used for analysis. Relevant articles for each parasite belonging to specific time periods (e.g., 1970-1980, 1980-1990, 1990-2000 and 2000-2009) were collected from PubMed. Table 1 shows the number of articles for selected parasites.

After retrieving the articles in specific time periods, drugs names were found by using drug filter server which tags drugs'names from DrugBank. The DrugBank database is a bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug information[3] [7].

Drugs have some variations such as synonyms and brand names. DrugBank was searched for each drug, synonyms and brand names of drugs were found. After finding variations, these names were mapped to one specific name.

A. Clustering Analysis

Clustering analysis is one area of machine learning of particular interest to data mining. It provides the means for the organization of a collection of patterns into clusters based on the similarity between these patterns, where each pattern is represented as a vector in multidimensional space [8].

Hierarchical clustering methods produce a hierarchy of clusters from small clusters of very similar items to large clusters that include more dissimilar items. Hierarchical methods usually produce a graphical output known as a dendogram or tree that shows this hierarchical cluster structure. Some hierarchical methods are divisive; those progressively divide the one large cluster compromising all of the data into smaller clusters and repeat this process until all clusters have been divided. Other hierarchical methods are agglomerative and work in the opposite direction by first finding the clusters of the most similar items and progressively adding less similar items until all items have been included into a single large cluster [8].

The Euclidean distance is one of the common similarity measures and it is defined as the square root of the squared discrepancies between two entities summed over all variables measured [9].

The pseudo code of agglomerative hierarchical clustering algorithm is as follows:

1. Compute the proximity matrix
2. Merge the closest two clusters
3. Update the proximity matrix to reflect the proximity between the new cluster and the original clusters
4. Repeat step 2 and 3 until only one cluster remains

Figure 1 shows an example of dendogram for A,B,C,D and E objects [10].

[3] http://www.drugbank.ca/

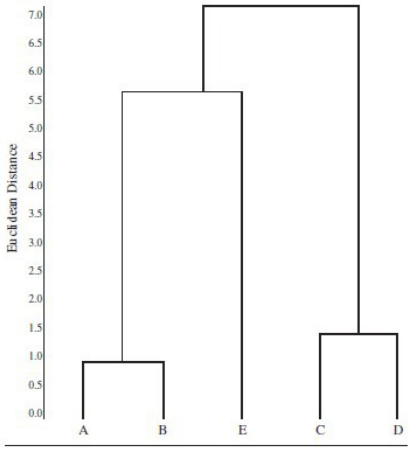

Fig. 1 An Example of Dendogram for A,B,C,D and E objects

III. RESULTS

Hierarchical clustering analysis was used to post-process the results from the co-occurrence analysis for the treatment of parasites. R statistical software is used for clustering and heatmap analysis. A heatmap is a graphical way of displaying a table of numbers by using colors to represent the numeric values.

The main categories of drugs that are used for the treatment of parasites comprise anthelmintic, anti-inflammatory and antiprotozoal drugs. Figures 2, 3, 4 and 5 show the heatmaps of the cluster analysis.

According to the observed results, one cluster consists of Echinococcus Multilocularis, Echinococcus Granulosus and Fasciola Hepatica. They share the following commonality:

- Albendazole, mebendazole and praziquantel from the standard treatment for all three parasites. This raises the notion that drugs developed for the treatment of one specie could in principle be exploited for the other two species.

Clonorchis Sinensis, Schistosoma Mansoni and Opisthorchis Viverrini form the second cluster of the analysis. In this cluster, praziquantel is seen as the common treatment. Apart from this drug, these parasites show little treatment with the other antihelmintic drugs. It is possible that these species would still profit from treatment with any of the other drugs.

Schistosoma Mansoni forms its own cluster. In the cluster, patients undergo treatment with antiinflammatory drugs similar to patients suffering from Fasciola Hepatica, but the type of antiinflammatory treatment differs significantly from the treatment of Fasciola Hepatica. Furthermore, patients suffering from Schistosoma Mansoni receive additional novel anthelmintic drugs such as levamisole and oxamniquine from 1980 to 2000.

Althogether, the treatment of parasites seems to be fairly stable over the past four decades with regards to the reporting of treatments in the scientific literature.

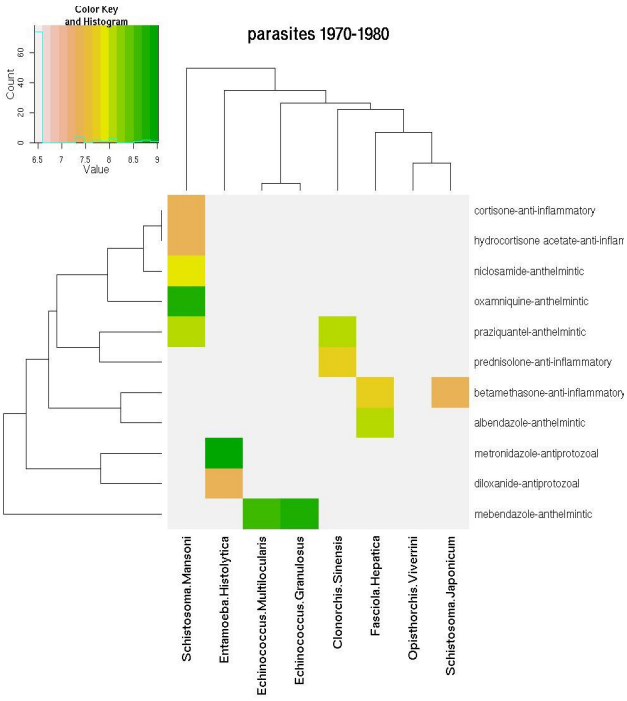

Fig. 2 Drug heatmap of parasites for 1970-1980 time period

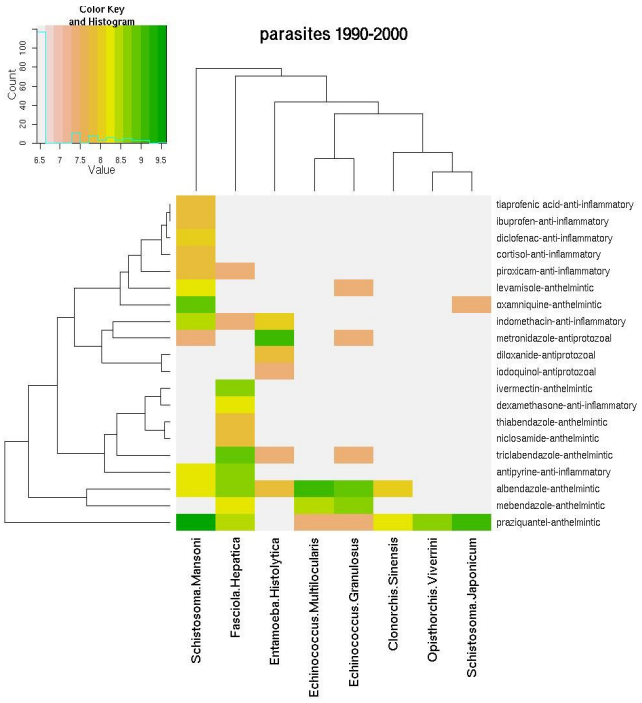

Fig. 4 Drug heatmap of parasites for 1990-2000 time period

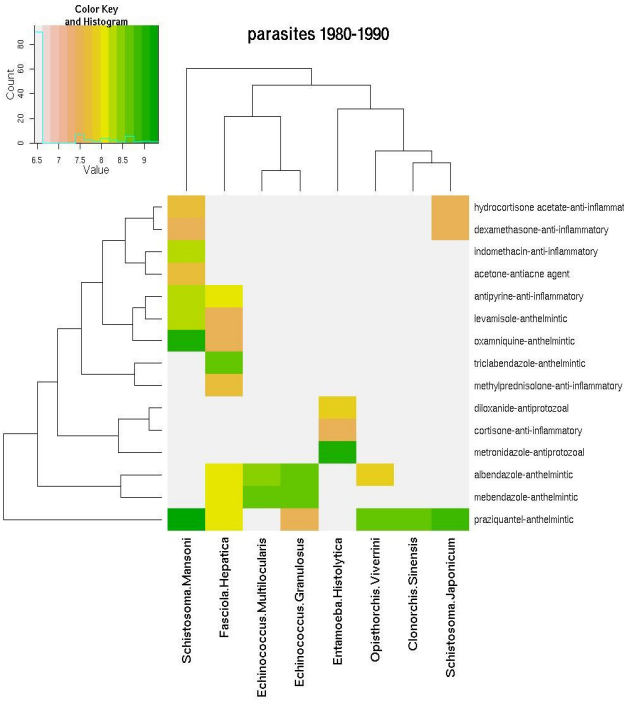

Fig. 3 Drug heatmap of parasites for 1980-1990 time period

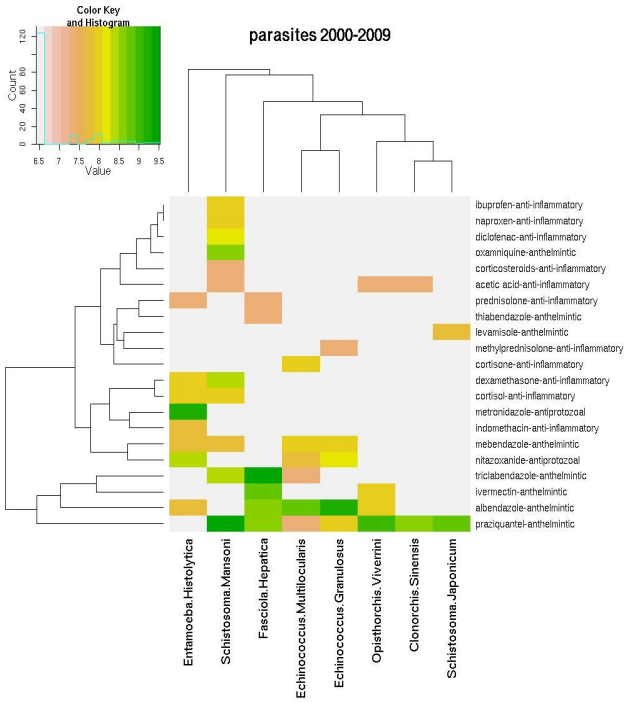

Fig. 5 Drug heatmap of parasites for 2000-2009 time period

IV. DISCUSSION

Infections with parasites are important causes of morbidity and mortality [11]. The control of parasitic disease requires a complex interplay of activities in the fields of public health, education, political will and medicine science. There is a need for treatment, and the search for better drugs is a perpetual process. Advances in science, especially in the field of parasite genomics and its attended technology, have opened up possibilities for new drugs.

We analyzed Medline abstracts to extract hidden knowledge and according to our results, there are no big changes among time periods in the treatment of liver specific parasites. Despite the large global burden of parasitic diseases, there has been very little recent effort by the pharmaceutical industry to develop agents to treat human parasitic infections [12].

Parasitic diseases, though globally massive in their impact, mainly affect poor people in poor regions of the world. As such, they would never be viewed as viable target markets for the pharmaceutical industry, particularly in today's post merger climate. In parallel, funding for basic research on these organisms and the pathogenesis of the diseases they produce has been woefully inadequate compared with funding for diseases of much lower prevalence but more direct impact in the developed countries of Europe and North America [1].

The protection of proprietary rights and the return of investments are also important issues for drug makers. With the long payback period associated with these indications, costs often are not recovered when a compound runs off patent and generic products may be introduced [13].

Regulatory requirements are another major concern that has a considerable impact on the length and costs of the drug development process and, hence, on the ultimate market price of the drug product. Paradoxically, increasingly demanding standards favour the larger wealthy companies which are those least interested in tropical diseases. Nevertheless, dossiers do not always undergo the same level of review worldwide, sometimes because of limited health budgets, and sometimes owing to a misconception about the regulatory process [13].

Medline abstracts are generally publicly available and therefore easy to share and distribute, while full text papers are not always available. However, there are some works to make them easily available and they can be public and sharable soon. As future work, it would be of interest to develop an efficient way to analyze full text papers to compare the results of abstracts [14].

V. CONCLUSIONS

Biomedical literature provides valuable knowledge for clinical studies and research. Medical experts can not read all the articles in a specific medical problem and discover hidden connections between entities. In this study, we worked with medical doctors and considered their needs and the point of view of them. Liver specific parasites were selected for the research. The combination of data mining and text mining techniques were used to get some facts hidden in Medline articles. Drugs which are based on specific time periods were extracted from the articles by using named entity recognition techniques and hierarchical clustering techniques were applied on parasite-drug datasets. Hierarchical clustering results revealed that some treatments of parasites are similar and the others are different. Our results also show that there have not been major changes in the treatment of liver specific parasites for the past four decades. We investigated the reasons for the challenge of drug discovery and development for parasitic diseases. We believe that our results will make an important contribution to medical research, clinical studies and pharmaceutical research.

ACKNOWLEDGMENT

We would like to thank Antonio Jose Jimeno Yepes, Dietrich Rebholz Schuhmann and Rabin Saba for their help and contributions.

REFERENCES

[1] A. R. Renslo, J. H. Mckerrow, "Drug discovery and development for neglected parasitic diseases, Nature Chemical Biology, vol. 2, no.12, 2006.

[2] L. A. Marcos, A. Terashima, E. Gotuzzo, "Update on hepatobiliary flukes: fascioliasis, opisthorchiasis and clonorchiasis, Current Opinion in Infectious Diseases, 2008, pp 523-530.

[3] N. Uramoto, H. Matsuzawa, T. Nagano, A. Murakami, H. Takeuchi, K. Takedo, "A Text Mining System for Knowledge Discovery from Biomedical Documents", IBM Systems Journal, vol. 43, Issue.3, pp 516-533.

[4] W. Zhou, N. D. Smalheiser., C. Yu, "A tutorial on information retrieval: basic terms and concepts, Journal of Biomedical Discovery and Collaboration", 2006, pp 1-8.

[5] D. Rebholz-Schuhmann, M. Arregui, A.J.J. Yepes, H. Kirsch, G. Nenadic, "Automatic Text Analysis Based on Web Services", Handout for the ISMB 2007 Tutorial, ISMB, Vienna, 20.07.2007.

[6] D. Rebholz-Schuhmann, H. Kirsch, S. Gaudan, M. Arregui, G. Nenadic, "Annotation and Disambiguation of Semantic Types in Biomedical Text: a Cascaded Approach to Named Entity Recognition", NLPXML '06 Proceedings of the 5th Workshop on NLP and XML, 2005.

[7] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang & J. Woolsey , "DrugBank: a comprehensive resource for in silico drug discovery and exploration", *Nucleic Acids Research,* 34**,** D668-D672, 2006.

[8] S. M. Holland, "Cluster Analysis", Department of Geology, University of Georgia, Athens, GA 30602-2501, 2006.

[9] J. W. Beckstead, "Using Hierarchical Cluster Analysis in Nursing Research", *Western Journal of Nursing Research*, Vol. 24, No. 307, pp-307-319, 2002.

[10] K. Vipin, "Introduction to Data Mining", Addison-Wesley, 2006.

[11] A. J. J. Wood, "Drug Therapy", The England Journal of Medicine, 1996, pp 1178-1184.

[12] A. C. White, "Nitazoxanide: a new broad spectrum antiparasitic agent", Expert Rev. Anti-infect. 2004, pp 43-49.

[13] P. Trouiller, P. L. Olliaro,"Drug development output from 1975 to 1996: what proportion for tropical diseases. International Journal of Infectious Diseases", 1998; 3: 61-63.

[14] A. Vlachos, "Evaluating and combining biomedical named entity recognition systems", In Poster Proceedings of BioNLP at ACL, Prague, 2007.