

**ACQUISITION OF LIVER SPECIFIC PARASITES-BACTERIA-DRUGS-
DISEASES-GENES KNOWLEDGE FROM MEDLINE**

**A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
THE MIDDLE EAST TECHNICAL UNIVERSITY**

BY

PINAR YILDIRIM

**IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF HEALTH INFORMATICS**

FEBRUARY 2011

Approval of the Graduate School of Informatics

Prof. Dr. Nazife Baykal
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Doctor of Philosophy.

Assist. Prof. Dr. Didem Gökçay
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Doctor of Philosophy.

Prof. Dr. Osman Saka
Supervisor

Examining Committee Members

Prof. Dr. Nazife Baykal (METU, II) _____

Prof. Dr. Osman Saka (METU, II, Akdeniz U. Med. Fac) _____

Dr. Ali Arifoğlu (METU, II) _____

Prof. Dr. Ergun Karaağaoğlu (Hacettepe U. Med. Fac.) _____

Prof. Dr. Mehmet R. Tolun (Çankaya U. Eng. Fac) _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Pınar Yıldırım

Signature : _____

ABSTRACT

ACQUISITION OF LIVER SPECIFIC PARASITES-BACTERIA-DRUGS- DISEASES-GENES KNOWLEDGE FROM MEDLINE

Yıldırım, Pınar

Ph.D., Department of Health Informatics

Supervisor: Prof. Dr. Osman Saka

February 2011, 138 pages

Biomedical literature such as MEDLINE articles are rich resources for discovering and tracking disease and drug knowledge. For example, information regarding the drugs that are used with a particular disease or the changes in drug usage over time is valuable. However, this information is buried in thousands of MEDLINE articles. Acquiring knowledge from these articles requires complex processes depending on the biomedical text mining techniques. Today, parasitic and bacterial diseases affect hundreds of millions of people worldwide. They result in significant mortality and devastating social and economic consequences. There are many control and eradication programs conducted in the world. Also, many drugs are developed for diseases caused from parasites and bacteria. In this study, research was conducted of parasites (bacteria

affecting the liver) and treatment drugs were tested. Also, relationships between these diseases and genes, along with parasites and bacteria were searched through data and biomedical text mining techniques. This study reveals that the treatment of parasites and bacteria seems to be stable over the last four decades. The methodology introduced in this study also presents a reference model to acquire medical knowledge from the literature.

Keywords: Biomedical text mining, Information extraction, Liver, Parasites, Bacteria

ÖZ

MEDLINE MAKALELERİNDEN KARACİĞERE ÖZGÜ PARAZİT-BAKTERİ- İLAÇ-HASTALIK-GEN BİLGİLERİNİN ELDE EDİLMESİ

Yıldırım, Pınar

Doktora, Sağlık Bilişimi Bölümü

Tez Yöneticisi: Prof. Dr. Osman Saka

Şubat 2011, 138 sayfa

MEDLINE makaleleri gibi biyomedikal literatürler, hastalık ve ilaç bilgilerini araştırmak ve izlemek için zengin kaynaklardır. Örneğin hangi ilacın belirli bir hastalık için kullanıldığı ve zaman içindeki ilaç kullanımındaki değişiklikler önemlidir ama binlerce MEDLINE makalesinin içine gömülmüştür. Bu makalelerden bilgi elde etme biyomedikal metin madenciliği tekniklerine dayalı karmaşık işlemler gerektirir. Günümüzde, parazit ve bakteri hastalıkları dünya çapında yüzlerce milyon insanı etkilemektedir ve önemli ölçüde ölüm oranına ve sosyal ve ekonomik yıkıcı sonuçlara neden olmaktadır. Dünyada birçok kontrol ve yokediciler programları vardır ve parazit ve

bakterilerilere dayalı hastalıkların tedavisi için birçok ilaç geliştirilmektedir. Bu çalışmada, veri ve biyomedikal metin madenciliği tekniklerinden yararlanılarak karaciğeri etkileyen parazit ve bakteriler ve ilaçlara ait bilgi edinme tanıtılmış ve ayrıca bu parazit ve bakterilerle hastalık ve gen ilişkileri araştırılmıştır. Bu çalışma, parazit ve bakterilerin tedavisinin son dört onar yıllık dönemde aynı görüldüğünü ortaya çıkarmıştır. Ayrıca bu çalışmada tanıtılan yöntemler dizisi, literatürden tıbbi bilgi elde etmek için referans model oluşturmaktadırlar.

Anahtar Kelimeler: Biyomedikal metin madenciliği, Bilgi çıkarma, Karaciğer, Parazitler, Bakteriler

ACKNOWLEDGMENTS

I would like to express my gratitude to all those who helped me to complete this thesis. At first I want to thank Prof. Dr. Osman Saka, Assoc. Prof. Dr. Ü. Erkan Mumcuođlu, and Prof. Dr. Mehmet R. Tolun for their support and academic guidance. Also, I want to thank Dr. Dietrich Rebholz-Schuhmann who gave me the opportunity to work in his research group at EBI (European Bioinformatics Institute) and supervised all aspects of my work there. In addition, I am thankful to Antonio Jose Jimeno Yepes and Christoph Grapmüller. Special thanks to Aydın T. Bakır for his technical assistance and Asst. Prof. Dr. Kađan Çeken for his help and inspiration. To all my family, I offer my sincere thanks for their unshakable faith in me and their willingness to endure the vicissitudes of my endeavors.

TABLE OF CONTENTS

ABSTRACT.....	iv
ÖZ.....	vi
ACKNOWLEDGMENTS	viii
TABLE OF CONTENTS.....	ix
LIST OF TABLES.....	xiii
LIST OF FIGURES	xviii
CHAPTER	
1 INTRODUCTION	1
1.1 Aim of Study.....	1
1.2 Related Work	5
2 INTRODUCTION TO BIOMEDICAL TEXT MINING	6
2.1 Biomedical Text Mining.....	6
2.2 How Does Text Mining Work?.....	7
2.3 Biomedical Text Mining Tasks.....	9
2.3.1 Named Entity Recognition.....	9
2.3.2 Synonym and Abbreviation Extraction	10
2.3.3 Relation Extraction	11
2.4 Natural Language Processing (NLP).....	11
2.5 Information Retrieval (IR).....	13
2.5.1 Public Document Repositories (MEDLINE)	14
2.6 Domain Knowledge.....	16
2.6.1 Ontologies	17

2.6.2	Taxonomies	17
2.6.3	The UMLS	18
3	METHODS.....	22
3.1	Steps of study.....	22
3.2	Normalization of Drugs.....	33
3.3	Statistical Co-occurrence Based Relation Extraction.....	36
3.4	Clustering Analysis	39
4	ANALYSIS OF PARASITES	41
4.1	Historical Review of Antiparasitic Drugs	41
4.2	Time Based Analysis of Parasites.....	46
4.3	Statistical Co-occurrence Based Relation Extraction of Parasites.....	46
4.4	Fasciola Hepatica	46
4.4.1	Drug Time Analysis of Fasciola Hepatica.....	46
4.4.2	Relation Extraction of Fasciola Hepatica	47
4.5	Schistosoma Mansoni.....	48
4.5.1	Drug Time Analysis of Schistosoma Mansoni.....	48
4.5.2	Relation Extraction of Schistosoma Mansoni.....	50
4.6	Schistosoma Japonicum.....	50
4.6.1	Drug Time Analysis of Schistosoma Japonicum.....	50
4.6.2	Relation Extraction of Schistosoma Japonicum.....	51
4.7	Entamoeba Histolytica	52
4.7.1	Drug Time Analysis of Entamoeba Histolytica	52
4.7.2	Relation Extraction of Entamoeba Histolytica.....	53
4.8	Echinococcus Granulosus.....	53
4.8.1	Drug Time Analysis of Echinococcus Granulosus.....	53
4.8.2	Relation Extraction of Echinococcus Granulosus	54
4.9	Echinococcus Multilocularis.....	54
4.9.1	Drug Time Analysis of Echinococcus Multilocularis	54
4.9.2	Relation Extraction of Echinococcus Multilocularis.....	56

4.10	Clonorchis Sinensis	56
4.10.1	Drug Time Analysis of Clonorchis Sinensis	56
4.10.2	Relation Extraction of Clonorchis Sinensis	56
4.11	Opisthorchis Viverrini	57
4.11.1	Drug Time Analysis of Opisthorchis Viverrini	57
4.11.2	Relation Extraction of Opisthorchis Viverrini	58
4.12	General Distribution of Drugs in Main Classes	58
4.13	Clustering Analysis of Parasites	61
5	ANALYSIS OF BACTERIA	67
5.1	Introduction	67
5.2	Time Based Analysis of Bacteria	67
5.3	Statistical Co-occurrence Based Relation Extraction	68
5.4	Salmonella Typhimurium	68
5.4.1	Drug Time Analysis of Salmonella Typhimurium	68
5.4.2	Relation Extraction of Salmonella Typhimurium	70
5.5	Staphylococcus Aureus	70
5.5.1	Drug time analysis of Staphylococcus Aureus	70
5.5.2	Relation Extraction of Staphylococcus Aureus	71
5.6	Helicobacter Pylori	71
5.6.1	Drug Time Analysis of Helicobacter Pylori	71
5.6.2	Relation Extraction of Helicobacter Pylori	73
5.7	Mycobacterium Tuberculosis	73
5.7.1	Drug Time Analysis of Mycobacterium Tuberculosis	73
5.7.2	Relation Extraction of Mycobacterium Tuberculosis	74
5.8	Listeria Monocytogenes	75
5.8.1	Drug Time Analysis of Listeria Monocytogenes	75
5.8.2	Relation Extraction of Listeria Monocytogenes	75
5.9	Klebsiella Pneumonia	76
5.9.1	Drug Time Analysis of Klebsiella Pneumonia	76

5.9.2	Relation Extraction of Klebsiella Pneumoniae	77
5.10	Pseudomonas Aeruginosa.....	77
5.10.1	Drug Time Analysis of Pseudomonas Aeruginosa.....	77
5.10.2	Relation Extraction of Pseudomonas Aeruginosa	78
5.11	Streptococcus Pneumoniae	79
5.11.1	Drug Time Analysis of Streptococcus Pneumoniae.....	79
5.11.2	Relation Extraction of Streptococcus Pneumoniae	80
5.12	General Distribution of Drugs in Main Classes.....	80
5.13	Clustering Analysis of Bacteria	83
6	DISCUSSION AND CONCLUSIONS	90
6.1	Discussion.....	90
6.2	Future Work.....	93
6.3	Conclusions.....	94
	REFERENCES.....	96
	APPENDICES.....	104
	A. RELATION EXTRACTION FOR PARASITES.....	104
	B. RELATION EXTRACTION FOR BACTERIA	120
	CURRICULUM VITAE.....	136

LIST OF TABLES

TABLE

Table 1: Geographic distribution, standard treatment of some liver parasites.....	3
Table 2: Steps of study.....	24
Table 3: List of filter servers.....	27
Table 4: Frequencies of selected parasites.....	29
Table 5: MeSH and NCBI Taxonomy Ids for Selected Parasites	29
Table 6: MeSH Ids for selected bacteria.....	30
Table 7: Number of articles for selected bacteria.....	30
Table 8: Number of articles according to time periods for Fasciola Hepatica.....	31
Table 9: Total number of articles for parasites	32
Table 10: Normalization of drugs for parasites.....	33
Table 11: Normalization of drugs for bacteria	34
Table 12: Summary of liver specific Fasciola Hepatica and albendazole co-occurrence data.....	37
Table 13: Timeline of some antiparasitic drugs.....	42
Table 14: Side effects of oxamniquine compared to praziquantel.....	48
Table 15: Development of antibiotics over time (Norrby et al. 2005)	68
Table 16: Fasciola hepatica/drugs co-occurrences based on the Drugbank filter server	104
Table 17: Fasciola Hepatica/diseases co-occurrences based on the UMLS disease filter server.....	105

Table 18: Fasciola Hepatica/genes co-occurrences based on the Swissprot filter server	105
Table 19: Schistosoma Mansoni/drugs co-occurrences based on Drugbank filter server	106
Table 20: Schistosoma Mansoni/diseases co-occurrences based on the UMLS disease filter server	107
Table 21: Schistosoma Mansoni/genes co-occurrences based on the Swissprot filter server	107
Table 22: Schistosoma Japonicum/drugs co-occurrences based on the Drugbank filter server	108
Table 23: Schistosoma Japonicum/diseases co-occurrences based on the UMLS disease filter server	109
Table 24: Schistosoma Japonicum/genes co-occurrences based on the Swissprot server	109
Table 25: Entamoeba Histolytica/drugs co-occurrences based on the Drugbank filter server	110
Table 26: Entamoeba Histolytica/diseases co-occurrences based on the UMLS disease server	111
Table 27: Entamoeba Histolytica/genes co-occurrences based on the Swissprot filter server	111
Table 28: Echinococcus Granulosus/drugs co-occurrences based on the Drugbank filter server	112
Table 29: Echinococcus Granulosus/diseases co-occurrences based on the UMLS disease filter server	113
Table 30: Echinococcus Granulosus/genes co-occurrences based on the Swissprot filter server	113
Table 31: Echinococcus Multilocularis/drugs co-occurrences based on the Drugbank filter server	114

Table 32: Echinococcus Multilocularis/diseases co-occurrences based on the UMLS disease filter server	115
Table 33: Echinococcus Multilocularis/genes co-occurrences based on the Swissprot filter server	115
Table 34: Clonorchis Sinensis /drugs co-occurrences based on the Drugbank Filter Server	116
Table 35: Clonorchis Sinensis/diseases co-occurrences based on the UMLS disease filter server.....	117
Table 36: Clonorchis Sinensis/genes co-occurrences based on the Swissprot filter server	117
Table 37: Opisthorchis Viverrini/drugs co-occurrences based on the Drugbank filter server.....	118
Table 38: Opisthorchis Viverrini/diseases co-occurrences based on the UMLS disease filter server	119
Table 39: Opisthorchis Viverrini/genes co-occurrences based on Swissprot filter server	119
Table 40: Salmonella Typhimurium/drugs co-occurrences based on the Drugbank filter server.....	120
Table 41: Salmonella Typhimurium/diseases co-occurrences based on the UMLS disease filter.....	121
Table 42: Salmonella Typhimurium/genes co-occurrences based on the Swissprot filter server.....	121
Table 43: Staphylococcus Aureus/drugs co-occurrences based on the Drugbank filter server.....	122
Table 44: Staphylococcus Aureus/diseases co-occurrences based on the UMLS disease filter server	123
Table 45: Staphylococcus Aureus/genes co-occurrences based on the Swissprot filter server.....	123

Table 46: Helicobacter Pylori /drugs co-occurrences based on the Drugbank filter server	124
Table 47: Helicobacter Pylori /diseases co-occurrences based on the UMLS disease filter server.....	125
Table 48: Helicobacter Pylori /genes co-occurrences based on the Swissprot filter server	125
Table 49: Mycobacterium Tuberculosis/drugs co-occurrences based on the Drugbank filter server	126
Table 50: Mycobacterium Tuberculosis/diseases co-occurrences occurrences based on the UMLS disease filter server	127
Table 51: Mycobacterium tuberculosis/genes co-occurrences based on the Swissprot filter server	127
Table 52: Listeria Monocytogenes/drugs co-occurrences based on the Drugbank filter server.....	128
Table 53: Listeria Monocytogenes/diseases co-occurrences based on the UMLS disease filter server	129
Table 54: Listeria Monocytogenes/genes co-occurrences based on the Swissprot filter server.....	129
Table 55: Klebsiella Pneumoniae/drugs co-occurrences based on the Drugbank filter server.....	130
Table 56: Klebsiella Pneumoniae/diseases co-occurrences based on the UMLS disease filter server	131
Table 57: Klebsiella Pneumoniae/genes co-occurrences based on the Swissprot filter server.....	131
Table 58: Pseudomonas Aeruginosa/drugs co-occurrences based on the Drugbank filter server.....	132
Table 59: Pseudomonas Aeruginosa/diseases co-occurrences based on the UMLS disease filter server	133

Table 60: Pseudomonas Aeruginosa/genes co-occurrences based on the Swissprot filter server.....	133
Table 61: Streptococcus Pneumoniae/drugs co-occurrences based on the Drugbank filter server.....	134
Table 62: Streptococcus Pneumoniae/diseases co-occurrences based on the UMLS disease filter server	135
Table 63: Streptococcus pneumoniae/genes co-occurrences based on the Swissprot filter server.....	135

LIST OF FIGURES

FIGURE

Figure 1: A Typical example of Named Entity Recognition	9
Figure 2: View of an entry for HIV Pneumonia in the UMLS metathesaurus	19
Figure 3: View of “ Sarcoma” in The SPECIALIST lexicon	20
Figure 4: Summary of the research procedure followed in this study.....	23
Figure 5: XML annotation of species	28
Figure 6: Single link clustering	40
Figure 7: Single link dendogram.....	40
Figure 8: Time series plot of anthelmintic drugs for Fasciola Hepatica.....	47
Figure 9: Time series plot of anthelmintic drugs for Schistosoma Mansoni	49
Figure 10: Time series of anthelmintic drugs for Schistosoma Japonicum.....	51
Figure 11: Time series plot of antiprotozoal drugs for Entamoeba Histolytica.....	52
Figure 12: Time series plot of anthelmintic drugs for Echinococcus Granulosus	54
Figure 13: Time series plot of anthelmintic drugs for Echinococcus Multilocularis	55
Figure 14: Time series plot of most ranked drugs for Clonorchis Sinensis.....	57
Figure 15: Time series plot of anthelmintic drugs for Opisthorchis Viverrini	58
Figure 16: Distribution of analgesic Drugs.....	59

Figure 17: Distribution of anthelmintic drugs.....	59
Figure 18: Distribution of antiinflammatory drugs	60
Figure 19: Distribution of antiprotozoal Drugs.....	60
Figure 20: Drug heatmap of parasites for 1970-1980 time period	61
Figure 21: Drug heatmap of parasites for 1980-1990 time period	63
Figure 22: Drug heatmap of parasites for 1990-2000 time period	64
Figure 23: Drug heatmap of parasites for 2000-2009 time period	65
Figure 24: Total drug heatmap of parasites	66
Figure 25: Time series plot of antibacterial drugs for Salmonella Typhimurium.....	69
Figure 26: Time series plot of antibacterial drugs for Staphylococcus Aureus	71
Figure 27: Time series plot of antibacterial drugs for Helicobacter Pylori	72
Figure 28: Time series plot of antibacterial drugs for Mycobacterium Tuberculosis	74
Figure 29: Time series plot of antibacterial drugs for Listeria Monocytogenes	75
Figure 30: Time series plot of antibacterial drugs for Klebsiella Pneumoniae.....	76
Figure 31: Time series plot of antibacterial drugs for Pseudomonas Aeruginosa	78
Figure 32: Time series plot of antibacterial drugs for Streptococcus Pneumoniae.....	79
Figure 33: Distribution of analgesic drugs for bacteria	81
Figure 34: Distribution of antibacterial drugs for bacteria	82
Figure 35: Distribution of anti-infective drugs for bacteria.....	80
Figure 36: Distribution of anti-inflammatory drugs.....	83
Figure 37: Drug heatmap of bacteria for 1970-1980 time period	85
Figure 38: Drug heatmap of bacteria for 1980-1990 time period	86
Figure 39: Drug heatmap of bacteria for 1990-2000 time period	87
Figure 40: Drug heatmap of bacteria for 2000-2009 time period	88
Figure 41: Total drug heatmap of bacteria.....	89
Figure 42: Time series plot of most ranked drugs for Fasciola Hepatica.....	104
Figure 43: Time series plot of most ranked drugs for Schistosoma Mansoni.....	106
Figure 44: Time series plot of most ranked drugs for Schistosoma Japonicum	108
Figure 45: Time series plot of most ranked drugs for Entamoeba Histolytica	110

Figure 46: Time series plot of most ranked drugs for Echinococcus Granulosus.....	112
Figure 47: Time series plot of most ranked drugs for Echinococcus Multilocularis	114
Figure 48: Time series plot of most ranked drugs for Clonorchis Sinensis.....	116
Figure 49: Time series plot of most ranked drugs for Opisthorchis Viverrini.....	118
Figure 50: Time series plot of most ranked drugs for Salmonella Typhimurium.....	120
Figure 51: Time series plot of most ranked drugs for Staphylococcus Aureus	122
Figure 52: Time series plot of most ranked drugs for Helicobacter Pylori	124
Figure 53: Time series plot of most ranked drugs for Mycobacterium Tuberculosis ...	126
Figure 54: Time series plot of most ranked drugs for Listeria Monocytogenes	128
Figure 55: Time series plot of most ranked drugs for Klebsiella Pneumoniae.....	130
Figure 56: Time series plot of most ranked drugs for Pseudomonas Aeruginosa.....	132
Figure 57: Time series plot of most ranked drugs for Streptococcus Pneumoniae	134

CHAPTER I

I INTRODUCTION

Biomedical literature such as MEDLINE articles are rich resources for discovering and tracking medical knowledge. For example, information the drugs that are used with a specific disease or changes in drug usage over time is valuable. Unfortunately, it is buried in thousands of articles. Acquiring knowledge from these articles requires complex processes depending on the data and biomedical text mining techniques.

In this study, interviews were held with medical doctors and knowledge was acquired about parasites and bacteria that affect liver and relevant treatment drugs. Diseases and genes from MEDLINE articles were considered as a subject for research.

1.1 Aim of Study

The treatment of parasites, in particular those caused by worms, is the focus of ongoing research due to the fact that hundreds of millions of people worldwide are affected, inducing devastating social and economic consequences (Renlso, et al. 2006). The antihelminthic treatment of parasites that comprise the health of the liver form a subgroup that requires special attention due to the significant damage caused by these parasites to the infected individuals.

There are several drugs available that are effective against different types of helminth and that can be used simultaneously to treat different types of helminth. This is advantageous if a patient is infected by many species. Medical staff has to be increasingly aware of drug resistencies and has to prepare alternative treatment methods. During the drug discovery process, it is required to identify drugs that have similar treatment profiles and compare it to a given drug. This drug must not have been tested for the parasite in this particular case. As part of this study, we filter all treatments of worms out of scientific literatures to give an overview for the ongoing research in this field and to enable researchers to quickly identify alternative treatments using constructive hypothetical criteries.

Although, scientific literature was screened for all manuscripts that evaluate a specific treatment of the parasite or bacterium induced disease, it is still a challenging task. In this study, we only consider parasites and bacteria that affect the structure and the function of the liver.

There are many control and eradication programs conducted in the world and many treatment drugs are developed for diseases caused from parasites. Parasites also cause harmful effects on main organs in the body. The liver and the lungs are most affected by them. Since nutrients that are absorbed in intestinal systems are first transmitted to liver by portal systems, liver is the most important host organ for parasites. Major health problems in East Asia, East Europe, Africa and Latin America are caused by liver infections from parasites. Recently, there have been heavy reports of millions of infected people world wide in specific geographic areas with risk factors. (Marcos, Terashima and Gotuzzo 2008). Table 1 shows geographic distribution and standard treatment of liver parasites (Marcos et al. 2008), (Craig 2003), (Shaikenov 2006) and (Mortele, Segatto and Ros 2004).

Table 1: Geographic distribution, standard treatment of some liver parasites

Parasite	Geographic Distribution	Standard Treatment
Fasciola Hepatica	America (mainly Peru and Bolivia), Europe, Asia, Western Pacific, North America	Triclabendazole
Opisthorchis Viverrini	Thailand, Laos, Cambodia, Vietnam	Praziquantel
Clonorchis Sinensis	North-east China, southern Korea, Japan, Taiwan, northern Vietnam and the far eastern part of Russia	Praziquantel
Echinococcus Granulosus	Europe, Northern Asia, North America	Albendazole
Echinococcus Multilocularis	Alaska, Canada, North Central USA, Northern Europe, Euroasia, Japan	Albendazole
Schistosoma Mansoni	Africa, Brazil, Suriname, Venezuela, Caribbean	Praziquantel
Schistosoma Japonicum	Southeast Asia, Western Pasific Countries (including China, Philipines, Indonesia)	Praziquantel
Entamoeba Histolytica	India, Africa, Far East, Central and South America	Metronidazole

In medical science, both the preclinical and clinical disciplines of medicine, such as gastroenterology, infectious diseases, microbiology, biology, surgery and radiology are interested in the treatment of liver specific parasites. Many scientists and clinicians are working on these flukes. There are many treatment options besides medical drug therapy, such as surgery or, in some of the parasitic infections like hydatid cyst, interventional percutaneous treatment. In these methods, surgery may have some harmful effects and generally is not considered as the first option. Drug therapy has an important role, not only in the treatment of individual patients, but also in conjunction with public

health and vector control measures to reduce the transmission of parasitic infections (Liu 1996). In addition, these parasitic diseases are mostly seen in undeveloped countries but can also be seen in some developed countries in Europe or North America. In the case of the latter, it is mostly affecting migrants and travelers. Since clinicians in these countries are not very familiar with these diseases, and doctors may use incorrect approaches to treat the patients. We hope that our study provides valuable information to all scientists and experts working on liver specific parasitic diseases.

Bacteria infections can also lead to serious life threatening complications and death. Bacteria are especially harmful on the liver which supports almost every organ and is therefore vital for survival. Because of its strategic location and multidimensional functions, the liver is also prone to many diseases. In this study, liver specific parasites, bacteria, drugs, disease and gene knowledge was extracted in MEDLINE articles and it provides new knowledge for clinical studies. Time analysis of drugs for each parasite and bacterium highlight that some drugs disappeared over time or that others are newly emerging. Furthermore, frequencies of drugs show which drugs are preferred more than others for the treatment of a particular parasite and bacterium. Therefore, both clinicians and researchers can make time-based comparisons with these information. In addition, diseases and genes which can be related to parasites or bacteria were searched and some hidden knowledge was discovered.

This study will make substantial contributions to the aims and tasks of medical information. Medical informatics is the discipline concerned with the systematic processing of data, information and knowledge in medicine and healthcare (Haux 1997).

The main aim is to improve the quality of healthcare and research in medicine. Medical informatics use appropriate models and methods for solving problems concerned with processing data, information, and knowledge. The method used in this study can be a reference model which allow the medical informants to understand the procedures of processing knowledge in biomedical literature for specific medical problem. When considering the point of view of physicians, this study helps them to get hidden facts

buried in medical articles, and also to interpret them in order to build up new medical knowledge.

Focusing on pharmaceutical researchers and initiatives who design antiparasitic and antibacterial drugs to fight infections, they can evaluate the history of the drug usage for treatment and develop new effective strategies.

1.2 Related Work

Knowledge acquisition techniques utilizing text mining have been used in many biomedical studies. For example Chen et al., applied a text mining approach for the automated acquisition of disease-drug associations in MEDLINE articles and summaries in the electronic medical records provided by the NewYork–Presbyterian Hospital. The above authors described the annotation of text sources provided by MeSH (Medical Subject Headings) and two NLP (Natural Language Processing) systems (BioMedLEE and MedLEE) that can be used to extract disease and drug entities. They also proposed various statistical techniques that can be used to identify strong disease-drug associations for a subset of eight diseases (Chen et al. 2008).

NLP techniques are also used to extract information in biomedical documents that can be used to improve patients' management, clinical decision support, quality assurance, and clinical research. Researchers have focused on the development of text mining solutions for specific tasks, particularly to seek concept pairs such as disease-drug. Many of these studies involved the use of knowledge resources such as the UMLS Metathesaurus. Some researchers focused on the hypotheses generation to provide useful information to health care providers and medical researchers. Identifying relationships between extracted entities (e.g., gene-drug, disease-gene, and disease-drug) is another research area for biomedical text mining studies. In addition, NLP and text mining have been applied to clinical documents for a range of applications, including detecting clinical conditions and medical errors, coding and billing, tracking physician performance, utilizing resources, improving communications with health care providers, and monitoring alternate methods for treatment (Chen et al. 2008).

CHAPTER 2

2 INTRODUCTION TO BIOMEDICAL TEXT MINING

2.1 Biomedical Text Mining

Text mining is the application of data mining techniques to the automated discovery of useful or interesting knowledge from text documents (Mooney R. 2005). Although traditional data mining techniques deal with the processing of structure databases or flat files, text mining techniques are dedicated to knowledge extraction from unstructured or semi-structured texts. These systems usually do not run their knowledge discovery algorithms on unprepared document collections. Preprocessing operations include a variety of different types of techniques used and adapted from information retrieval, information extraction and computational linguistics research that transform raw, unstructured, original-format content (like that which can be downloaded from PubMed) into a carefully structured, intermediate data format (Feldman and Sanger 2007).

The knowledge encoded in textual documents is organized around sets of domain-specific terms (e.g. names of proteins, genes, diseases, etc.) which are used as a basis for sophisticated knowledge acquisition. The basic problem is to recognize domain-specific concepts and to extract instances of specific relationships among them (Nenadic, Spasic and Ananiadou 2003). Therefore, within the many techniques, the integration of several resources such as terminologies and ontologies are required to extract efficient knowledge discovery.

The basic element in text mining is the document. From a linguistic perspective, a document demonstrates a rich amount of semantic and syntactical structure (Feldman and Sanger 2007). Text mining focuses on the document collection. At its simplest, a document collection can be any grouping of text-based documents. Most text mining solutions are aimed at discovering patterns across very large collections of documents. The number of documents in such collections can range from the many thousands to the tens of millions (Feldman and Sanger 2007).

In biomedical domain, with an overwhelming amount of textual information, there is a need for effective text mining that can help researchers gather and make use of the knowledge encoded in text documents. The amount of published papers makes it difficult for a person to efficiently localize the information of interest in a large collection of documents. For example, the MEDLINE database is a big resource for molecular biology, biomedicine and medicine. It is doubtful that any researcher could process such a huge amount of information, especially if the knowledge spans across domains (Nenadic et al. 2003). The goal of biomedical research is to discover knowledge and put it to practical use in the forms of diagnosis, prevention and treatment. Therefore, biomedical text mining allows researchers to identify needed information more efficiently and uncover relationships obscured by the sheer volume of available information (Cohen and Hersh 2005).

2.2 How Does Text Mining Work?

Text mining involves different techniques from areas such as information retrieval, natural language processing, information extraction and data mining. These various stages of a text mining process can be combined into a single workflow (Redfearn 2008).

A number of well-defined software components are used for processing natural language text. It is the goal to separate the natural language text into components that can be annotated either with syntactical information or with semantic information. The

syntactic information reports on the role of a word (or token) in the text and the semantic information gives the meaning of the word (Rebholz-Schuhmann D. 2007).

Tokenization is one of the most basic steps in text processing. In this step, the text is separated into single tokens, i.e. the sentence is separated into its words. From a naive perspective it is obvious that any token is separated by white space (a blank, a carriage return, or a tabular) and by punctuation signs from other tokens. In reality, it can be quite difficult to use such a solution. In particular, gene and protein names, as well as names for chemical entities that consist of combinations of characters, numbers and punctuation signs (e.g. HZF-1) (Rebholz-Schuhmann D. 2007).

Morphological analysis is used to extract additional features from the token. Part of the morphological analysis is, for example, the distinction between uppercase and lowercase representations. Morphological variation can raise ambiguities in comparison to derivational modifications of a word (e.g. an 's' at the end of a term can determine a plural variant of a term) (Rebholz-Schuhmann D. 2007).

Part-of-speech tagging refers to the processing step in which a token is denoted with its syntactical role in the sentence (e.g. distinction between a noun and a verb). These techniques are generally based on statistical IT components that have been trained on a selected corpus (Rebholz-Schuhmann D. 2007).

Chunking is used to separate the text into pieces (called "chunks") that consist of several tokens. In other words a selection of tokens are kept together if it is obvious that this selection of words or features form a unit. This is for example, the case, if the words from a noun phrase (e.g. "the white house" and "the green door") or any type of idiom ("for the time being") (Rebholz-Schuhmann D. 2007).

Parsing is a processing step that delivers the sentence structure. The parser delivers a data representation which gives insight into which components of the sentence are grouped together and how they act on other components of the sentence. There are different types of parsing techniques.

2.3 Biomedical Text Mining Tasks

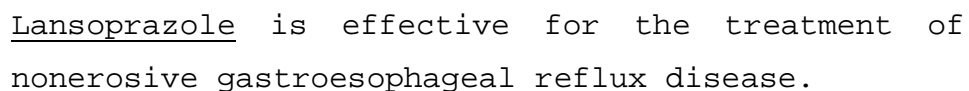
Tasks that are used for biomedical text mining include the following items:

- Named entity recognition
- Synonym and abbreviation extraction
- Relation extraction

2.3.1 Named Entity Recognition

Biomedical literature contains a special category of entities that refer to gene and protein names, chemical compounds, diseases, tissues, cellular components or other predefined biological concepts (Scherf, Epple and Werner 2005). The main goal of named entity recognition is to identify entities in text.

The approaches of this task generally fall into three categories: lexicon based, rules based and statistically based. Combined approaches also have been used. The output may be a set of tags that assign a predicted type to each word or phrase of interest, as in part-of-speech (POS) tagging (Cohen and Hersh 2005). Furthermore, identification of named entities (NEs) in a document can be viewed as a three-step procedure (Krauthammer and Nenadic 2004). In the first step, single or multiple adjacent words that indicate the presence of domain concepts are recognized (term recognition). In the second step, called term categorisation, the recognised terms are classified into broader domain classes (e.g., genes, proteins and species). The final step is the mapping of terms into referential databases. The first two steps are commonly referred to as named entity recognition (NER) (Rebholz-Schuhmann et al. 2006).



Lansoprazole is effective for the treatment of nonerosive gastroesophageal reflux disease.

Figure 1: A Typical example of Named Entity Recognition

In Figure 1, Lansoprazole is a drug name and nonerosive gastroesophageal reflux disease is a disease name.

On the other hand, NER in the biological domain (particularly the recognition part) profits from large, freely available terminological resources, which are either provided as ontologies (e.g., Gene Ontology, ChEBI, UMLS) or result from biomedical databases containing named entities (e.g., UniProt/Swiss-Prot) (Rebholz-Schuhmann et al. 2006).

2.3.2 Synonym and Abbreviation Extraction

Many biomedical entities have multiple names and abbreviations; therefore, it would be beneficial to have an automated means to collect these synonyms and abbreviations to help users that are performing searches for literature. Some words or phrases can refer to different things depending upon context (e.g., Ferritin can be a biological substance or a laboratory test). Conversely, many biological entities have several names (e.g., PTEN and MMAC1 refer the same gene). Biological entities may also have multi-word names (e.g., carotid artery), so the problem is additionally complicated by the need to determine name boundaries and resolve overlap of candidate names. Most of the work in this type of extraction has focused on uncovering gene name synonyms and biomedical term abbreviations (Cohen and Hersh 2005).

Due to the synonym/homonym problem, the task of entity recognition requires an identification and a disambiguation step. Identification strategies range from methods that use ad hoc rules about typical syntactic structures of entity identifiers to algorithms that search identifiers of a given dictionary with exact and inexact pattern matching methods. Typically, dictionaries are used in combination with pattern recognition approaches. The dictionaries are based on publicly available sources of standardized, structured data annotated by human experts. The disambiguation step implies a classification method deciding whether the text where entity has been identified refers to expected topic (Scherf et al. 2005).

There are a variety of techniques which include combination of named entity recognition, with statistical, support vector machine classifier based, and automatically

or manual pattern based matching rules algorithms. For example, many biological terms have some specific patterns such as upper case letters, numerical figures and non-alphabetical characters (e.g., T-Lymphocytes, PTEN). Because of this, some rules can be generated to recognize these names in the text documents.

2.3.3 Relation Extraction

Relation extraction techniques recognize occurrences of a pre-specified type of relationship between a pair of specific type of entities (e.g., relationships between diseases and drugs). These techniques usually consist of neighbor divergence analysis, vector space approach and k-medoids that cluster algorithm, fuzzy set theory on co-occurring dataset records, and type and part-of-speech tagging (Petric 2007).

2.4 Natural Language Processing (NLP)

The role of NLP in text mining is to provide the systems in the information extraction phase with linguistic data that they need to perform their task. Often this is done by annotating documents with information such as sentence boundaries, part-of-speech tags and parsing results. These can then be read by the information extraction tools.

In biomedical domain, there are two important sub domains for NLP methodologies. These are clinical medicine and molecular biology. In the clinical domain, the emphasis is on disease, anatomy, etiology and intervention along with the interaction among these phenomena. The other research area is molecular biology. A major challenge is recognizing entities such as genes (and other aspects of the genome) and proteins. Another way to investigate NLP systems is to consider the genre of the text being processed. Two relevant genres in biomedicine are clinical records (such as discharge summaries and imaging reports) (Rindfleisch 2006).

Various linguistic approaches have been used to process biomedical text. These can be broadly categorized as either statistical or symbolic rule-based systems. In medicine, the latter is predominating. Due to the complexity of language, systems often focus on one aspect of linguistic structure: words, phrases, semantic concepts or semantic relations. Words can be identified with little (or no) linguistic processing. Phrases are normally

identified on the basis of at least some syntactic analysis, using part-of-speech categories and rules for defining phrase patterns in English (Leroy, Chen and Martinez 2003). The identification of concepts and relations constitutes semantic processing and requires that text be mapped to a knowledge structure. In the biomedical domain, the Unified Medical Language System (UMLS) provides one such resource. Semantic processing is a method of automatic language analysis that identifies concepts and relationships to represent document content (Rindflesch 2006).

Textual information management systems based solely on words have enjoyed considerable popularity, largely because the underlying processing is relatively easy to implement. After grammatical function words such as determiners “the” and “this” and prepositions “of” and “with” are eliminated, the remaining words are taken as a surrogate representation of semantic content (Rindflesch 2006).

In recent years, research has continued to focus on text indexing and document coding to allow powerful, meaningful retrieval of documents. Document indexing uses terms from glossary or ontology (MeSH, UMLS, and SNOMED) or text features such as words or phrases. Various methods have been applied to medical scientific literature and to clinical narrative (de Bruijn and Martin 2002).

One major contrast between most NLP research in clinical medicine and the more recent ones in molecular biology is the type of language material: patient records versus scientific articles. Most NLP systems in clinical medicine work with text from patient records such as discharge summaries and diagnosis reports. NLP systems in bioinformatics use mostly articles or abstracts from the scientific medical literature. Differences between these two types of text affect the choice of techniques for NLP. Biomedical literature is carefully constructed and meticulously proof-read, so spelling errors and incomplete parses are less of a problem. On the other hand, new concepts may be introduced, such as newly unraveled molecule. The bulk of literature is in English. Clinical narrative, on the other hand, might be more colloquial with the use of ungrammatical constructs and unstandardized abbreviations. It is more likely to contain segments of “canned text” longer phrases or possibly entire paragraphs that are

repeatedly encountered between records. Unknown words are, may be spelling errors, but are often proper names such as patient names, doctor names, addresses or institution names. There is work on clinical narrative includes methods that handle languages other than English, or do cross language operations such as retrieval from multilanguage collections or interlingual translation (de Bruijn and Martin 2002).

2.5 Information Retrieval (IR)

IR systems identify the documents in a collection which match a user's query. The most well known IR systems are search engines such as Google, which identify those documents on the World Wide Web that are relevant to a set of given words. IR systems allow us to narrow down the set of documents that are relevant to a particular problem. As text mining involves applying very computationally intensive algorithms to large document collections, IR can speed up the analysis considerably by reducing the number of documents for analysis (Redfearn 2008).

The purpose of information retrieval is the selection and delivery of documents from a larger set of documents. As a result the retrieval engine has to process the documents first to identify the features (e.g., words) contained in the documents. The features are kept in a separate data structure that is called the "index." When the retrieval engine is queried with a single term, it identifies all documents that contain the term and then returns the set of documents. If the query contains several terms then the IR engine calculates a score on the basis of the information contained in the index to deliver the most relevant documents (Rebholz-Schuhmann D. 2007).

Information retrieval in general and IR for Web search engines has developed into its own reserach area. In general terms, features from the query have to be mapped to features from the available documents. This has lead into research which features are meaningful, how they can be identified and prioritized and how the retrieval engine has to make efficient use of assumptions on the queries and the documents to deliver the correct sample of documents. Briefly the retrieval engine has to classify (or prioritize)

the documents into the set of documents that meets the expectations of the user (Rebholz-Schuhmann D. 2007).

An information retrieval engine directs the query of a user to the index where the contained tokens point to relevant documents. Such an engine can be implemented based on a relational database (e.g., combination of Oracle and Apache Lucene for CiteXplore, EBI) and on special IT solutions (e.g., Apache Lucene). For example, Apache Lucene is an open source project with the goal to deliver an IT solution that allows indexing of a corpus of documents and querying the corpus for document retrieval. Lucene is implemented in Java, comes with its own tokenizer, which can be replaced if necessary, and provides a query language that allows optimization of the queried results. In the case of the complete distribution of MEDLINE abstracts each file of the MEDLINE repository contains a set of citations (references, documents, abstracts). Each file is processed, (i.e. the title, the abstract and the fields containing the MeSH terms associated to the abstract) are analyzed with tokenizers that identifier numbers and terms. If a field is analyzed, then the title of the field (e.g., ArticleTitle and AbstractText) are kept (Rebholz-Schuhmann D. 2007).

2.5.1 Public Document Repositories (MEDLINE)

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) (Uramoto et al. 2004). The documents are available on PubMed web site. PubMed is 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; instead, links to the provider's site to obtain the full-text of articles are given, is available (Zhou, Smalheiser and Yu 2006).

Each article in MEDLINE is indexed according to multiple fields, including title, abstract, author names, journal name, language of publication, year of publication and Medical subject Headings (MeSH) thesaurus (Table 1). The MeSH thesaurus is controlled vocabulary produced by the National Library of Medicine and used for

indexing, cataloging and searching for biomedical and health related information and documents. A list of entry terms (synonymous or closely related terms) is given for each descriptor. In the MeSH thesaurus, descriptors are related by parent/child relations: each descriptor has at least one parent and may have several. The arrangement of MeSH descriptors from other hierarchies is intended to serve the purpose of indexing and information retrieval and does not always follow strict classificatory principles (Bodenreider, Ananiadou and Mc Naught 2006). The set of MeSH terms is manually assigned by biomedical experts who scan each article (Zhou et al. 2006).

MeSH descriptors are organized in 16 categories. For example, category A for anatomic terms, category B for organisms, C for diseases, D for drugs and chemicals, etc. Each category is further divided into subcategories. Within each subcategory, descriptors are arrayed hierarchically from most general to most specific in up to eleven hierarchical levels. Because of the branching structure of the hierarchies, these lists are called as “trees”. Each MeSH descriptor appears in at least one place in the trees and may appear in as many additional places as may be appropriate. Those who index articles or catalog books are instructed to find and use the most specific MeSH descriptor that is available to represent each indexable concept¹.

PubMed employs the Boolean operators that are used to retrieve a set in which each record contains all the search terms. This operator places no condition on where the terms are found in relation to one another; the terms simply have to appear somewhere in the same record. For example, if one desired documents on the use of the drug propranolol in the disease hypertension, a typical search statement might be (propranolol AND hypertension). The OR operator retrieves documents that contain at least one of the specific search terms. The NOT operator excludes the specified from the search. Certain very common words (e.g., “this”) are placed on a stoplist and are automatically excluded from queries (Zhou et al. 2006).

Before PubMed begins to retrieve articles, it performs query preprocessing, to identify which fields of the MEDLINE record are relevant, and to alter or expand the query

¹ http://www.nlm.nih.gov/mesh/intro_trees2006.html

terms via automatic term mapping. For example, the query [high blood pressure] will be automatically mapped to the MeSH term “hypertension” (each MeSH term may have a set of synonyms as alternative entry terms. In this example, “high blood pressure” is one of the synonyms or entry terms of “hypertension”). PubMed will search using the mapped MeSH term within the MeSH field, as well as the term originally entered. MeSH comprises a hierarchy of terms, and the more specific terms corresponding to that MeSH term will also automatically be searched. In this example, three more specific MeSH terms “Hypertension Malignant”, “Hypertension, Renal”, and “Hypertension, Pregnancy-Induced” are also searched (Zhou et al. 2006).

The Mesh Browser provides an online vocabulary look-up aid available for use with MeSH. It is designed to help quickly locate descriptors of possible interest and to show the hierarchy in which descriptors of interest appear. Virtually complete MeSH records are available, including the scope notes, annotations, entry vocabulary, history notes, allowable qualifiers, etc. The browser does not link directly to any MEDLINE or other database retrieval system and thus is not a substitute for the PubMed².

2.6 Domain Knowledge

In text mining systems, concepts belong not only to the descriptive attributes of a particular document but generally also to domains. With respect to text mining, a domain has come to be loosely defined as a specialized area of interest for which dedicated ontologies, lexicons and taxonomies of information may be developed (Feldman and Sanger 2007).

Domain knowledge can be used in text mining preprocessing operations to enhance concept extraction and validation activities. Although not strictly necessary for the creation of concept hierarchies within the context of a single document or document collection, access to background knowledge can play an important role in the development of more meaningful, consistent and normalized concept hierarchies. Domain knowledge can be used to inform many different elements of a text mining

² <http://www.nlm.nih.gov/mesh/mbinfo.html>

system. In preprocessing operations, domain knowledge is an important adjunct to classification and concept extraction methodologies (Feldman and Sanger 2007).

2.6.1 Ontologies

An ontology is a formal, explicit specification of a shared conceptualization for a domain of interest and is organized by concepts and relationships. The use of ontologies is indispensable for any text mining applications, in particular, for the tasks of named entity recognition, information extraction and retrieval. Ontologies are also needed in the area of data integration. The incompatibilities among data formats, structure and models (flat files, relational databases, etc.) have become a major obstacle in biological research. To overcome this problem of integration of data, e.g., data warehouses or distributed databases were created. It is necessary to know and describe exactly which data entries in one data source relate to the data entries in another source and to know how they are related (Saric, Engelken and Reyle 2008). There are many biomedical ontologies; refer Open Biomedical Ontologies (OBO) for a comprehensive list of biomedical ontologies. Some of the most widely-used biomedical ontologies are UMLS, GO(Gene Ontology) and SNOMED (Hu 2006).

2.6.2 Taxonomies

Taxonomy is the practice and science of classification. Taxonomies or taxonomic schemes are composed of taxonomic units known as taxa (singular taxon) or kinds of things that are arranged frequently in a hierarchical structure. Typically they are related by subtype-supertype relationships, also called parent-child relationships. Originally the term taxonomy referred to the classifying of living organisms³. The NCBI taxonomy database contains the names of all organisms that are represented in the genetic databases with at least one nucleotide or protein sequence⁴.

³ <http://www.wiki.org/wiki/Taxonomy>

⁴ <http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy>

2.6.3 The UMLS

The Unified Medical Language System (UMLS) facilitates the development of computer systems that behave as if they “understand” the language of biomedicine and health. National Library of Medicine (NLM) produces and distributes the UMLS Knowledge Sources (databases) and associated software tools. Developers use the Knowledge Sources and tools to build or enhance systems that create, process, retrieve and integrate biomedical and health data and information. The Knowledge Sources are multi-purpose and are used in systems that perform diverse functions involving information types such as patient records, scientific literature, guidelines and public health data.

There are three UMLS Knowledge Sources:

- (a) The Metathesaurus, which contains over one million biomedical concepts from over 100 source vocabularies
- (b) The SPECIALIST Lexicon & Lexical Tools, which provide lexical information and programs for language processing
- (c) The Semantic Network, which defines 135 broad categories and fifty-four relationships between categories for labeling the biomedical domain⁵.

Metathesaurus

The UMLS Metathesaurus is a large, multi-purpose and multi-lingual vocabulary database that contains information about biomedical and health related concepts, their various names, and the relationships among them. It is built from the electronic versions of numerous thesauri, classifications, code sets and lists of controlled terms used in patient care, health services billing, public health statistics, indexing biomedical literature, and/or basic, clinical and health services research. The scope of the Metathesaurus is determined by the combined scope of its source vocabularies, including-in addition to Gene Ontology and MeSH-disease vocabularies (e.g., International Classification of Diseases), clinical vocabularies (e.g., SNOMED CT), nomenclatures of drugs and medical devices, as well as the vocabularies of many

⁵ <http://www.nlm.nih.gov/research/umls/>

subdomains of biomedicine (e.g., nursing, psychiatry, gastrointestinal endoscopy) (Bodenreider et al. 2006).

The Metathesaurus is organized by concept or meaning. In essence, it links alternative names and views of the same concept and then identifies useful relationships between different concepts. All concepts in the Metathesaurus are assigned at least one Semantic Type from the Semantic Network to provide consistent categorization at the relatively general level represented in the Semantic Network. Many of the words and multi-word terms that appear in concept names or strings in the Metathesaurus also appear in the SPECIALIST Lexicon. In Metathesaurus, each term is assigned to a unique string identifier, which is then mapped to a unique concept identifier (CUI). An entry for HIV pneumonia in the Metathesaurus main termbank (MRCON) look like this (Figure 2).

C0744975 | ENG | P | L1392183 | PF | S1657928 | HIV pneumonia | 3 |

C0744975 : The concept Identifier

ENG : The language of term

P : The term status

L1392183 : The term identifier

PF : The string type

S1657928 : The string identifier

HIV pneumonia : The string itself

3 : A restriction level

Figure 2: View of an entry for HIV Pneumonia in the UMLS metathesaurus

Terms from the constituent vocabularies are organized into more than a million concepts that reflect synonymous meaning. For example, the concept “Chronic childhood

Arthritis” contains synonymous terms “Arthritis, Juvenile Rheumatoid” (from MeSH and SNOMED) and “ Rheumatoid Arthritis in Children” (Library of Congress Subject Headings), among others (Rindflesch 2006).

SPECIALIST Lexicon

The SPECIALIST Lexicon is one of three knowledge sources developed by the National Library of Medicine (NLM) as part of the Unified Medical Language System(UMLS) project. It provides the lexical information needed for processing natural language in the biomedical domain. The lexicon entry for each word or multi-word term records syntactic (part of speech, allowable complementation patterns), morphological (base form, inflectional variants) and orthographic (spelling variants) information. It is, in fact, a general English lexicon that includes many biomedical terms. Lexical items are selected from a variety of sources, including lexical items from MEDLINE/PubMed citation records, the UMLS Metathesaurus, and a large set of lexical items from medical and general English dictionaries. Contrary to Wordnet, the SPECIALIST lexicon does not include any information about synonym or semantic relations among its entries.

The SPECIALIST lexicon is distributed as part of the UMLS and can be queried through application programming interfaces for java and XML. It is also available as an open source resource as part of the SPECIALIST NLP tools⁶.

```
sarcoma
    cat =noun
    variants = uncount
    variants = reg
    variants = greg
```

Figure 3: View of “ Sarcoma” in The SPECIALIST lexicon

⁶ [http:// SPECIALIST.nlm.nih.gov](http://SPECIALIST.nlm.nih.gov)

The SPECIALIST lexicon describes syntactic characteristics of biomedical and general English terms. This comprehensive resource provides the basis for NLP in the biomedical domain. In addition to part-of-speech labels for each entry, spelling variation when it occurs (particularly British forms) and inflection for nouns, verbs and adjectives are included. Inflection is encoded by referring to rules for regular variants (-s for nouns and -s, -ed, ing for verbs, for example) as well as Greco-Latin plurals. Irregular forms are listed where they apply. The variant annotation for sarcoma, for example, indicates that this form may either appear invariant (sarcoma), with a regular plural (sarcomas), or with Greco-Latin morphology (sarcomata) (Rindflesch 2006) (Figure 3).

Semantic Network

The UMLS Semantic Network constitutes an upper-level ontology of medicine. It provides a consistent categorization of all concepts represented in the Metathesaurus and provides a set of useful relationships between these concepts. All information about specific concepts is found in the Metathesaurus. The network provides information about the set of basic semantic types or categories that may be assigned to these concepts. It also defines the set of relationships that may hold between the semantic types. There are major groupings of semantic types for organisms, anatomical structures, biological function, chemicals, events, physical objects and concepts or ideas. Some examples are given in (Rindflesch 2006):

Therapeutic or Preventive Procedure' TREATS 'Injury or Poisoning

Organism Attribute' PROPERTY_OF 'Mammal

Bacterium' CAUSES 'Pathologic Function

The current release of the Semantic Network contains 135 Semantic Types and 54 relationships. The Semantic Network serves as an authority for the Semantic Types that are assigned to concepts in the Metathesaurus. The network defines these types, both with textual descriptions and by means of the information inherent in its hierarchies.

CHAPTER 3

3 METHODS

Our literature analysis procedure uses the MEDLINE distribution available through the PubMed Web portal at the National Library of Medicine (NLM) as well as on the in house distribution at the EMBL-EBI (European Molecular Biology Laboratory-European Bioinformatics Institute). In total, the MEDLINE distribution contains almost 20 million abstracts. The former contains the full set of available documents, but is only accessible through special retrieval engines that does not provide full access to all documents due to restrictions implemented by the setup of the retrieval engine. The latter tool enables processing without using restricted retrieval services, but is lagging behind in term of the number of stored documents (Rebholz-Schuhmann et al. 2007).

3.1 Steps of study

In the first phase of our document analysis procedure, our clinicians proposed a list of species which induce liver-specific diseases. Furthermore, they proposed categories of drugs that could be used in the treatment of such diseases. Our proposed list of species was compared against the entries in the NCBI taxonomy to make sure that all synonyms of the species have been included, and that no relevant species is excluded from the study.

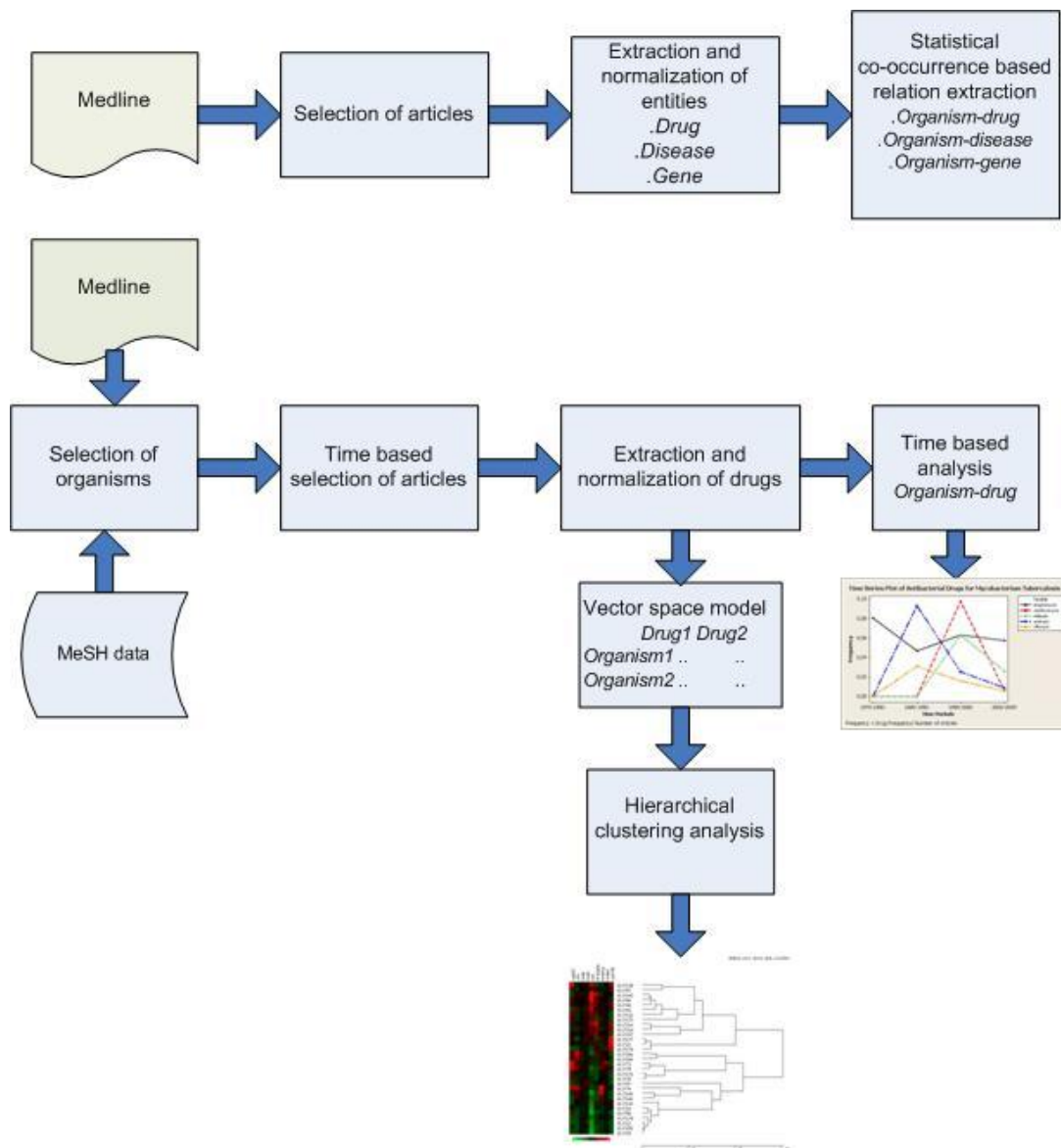


Figure 4: Flowchart of the research procedure followed in this study

MEDLINE delivers articles' abstracts in XML format. The full-text articles are available elsewhere (Zhou et al. 2006). The full MEDLINE abstracts include meta-data such as the Journal title, the author list, affiliations, publication dates as well as annotations inserted by the NLM such as creation date of the MEDLINE entry, list of chemicals associated with the document, as well as related MeSH headings (Rebholz-Schuhmann et al. 2006).

Table 2: Steps of study

Function of Step	Input	Output
Selection of organisms		
Parasites		
Selection of parasites	PubMed query <ul style="list-style-type: none"> Parasites list(organisms) and liver(organ) 	Article collection
Extraction of parasites	Article collection	Tagged organisms in the articles
Calculation and ranking of co-occurrence frequencies of parasites with liver	Tagged organisms in the articles	List of most ranked parasites with liver
Bacteria		
Selection of bacteria	PubMed query for bacteria with liver by using bacteria list in MesH Data	List of number of articles
Finding most frequent bacteria	List of number of articles	List of most ranked bacteria with liver(organ)
Time based analysis		
Time based selection of articles for each organism (parasite/bacterium)	PubMed query for the articles of liver related organism belonging to specific time periods <ul style="list-style-type: none"> 1970-1980 1980-1990 1990-2000 2000-2009 	Time based article collection
Extraction of drugs	Time based article collection	Drug tagged articles
Calculation of co-occurrence frequencies of drugs	Drug tagged articles	Co-occurrence frequencies of drugs
Normalisation of drugs	Drug name variations	Drug names mapped to specific names
Creation of time series plots <ul style="list-style-type: none"> Selection of main drug class for the treatment 	Time specific data for organism and main drug class	Time series plots
Statistical co-occurrence based relation extraction		
Collection of all articles from Pubmed for liver specific organism	PubMed query <ul style="list-style-type: none"> Liver and organism 	Article collection
Extraction of entities: <ul style="list-style-type: none"> Drug Disease Gene 	Article collection	Entity tagged articles

“Table 2 (cont.)”

Calculation of co-occurrence frequencies of entities	Entity tagged articles	Co-occurrence frequencies of entities
Normalisation of entities	Searching entity name variations	Entity names mapped to specific names
Applying statistical co-occurrence based relation extraction	Co-occurrence data for organism and entities	Co-occurrence frequencies and PMI
Hierarchical Clustering Analysis		
Creation of vector space model for organism and drug data (time specific)	Vector space model <ul style="list-style-type: none"> • Organism/drug 	Heatmap outputs of clustering analysis

Generally, our approach consists of selection of organisms (parasite/bacteria), time based analysis, and statistical co-occurrence based relation extraction and hierarchical clustering analysis. Figure 4 shows flowchart of the research procedure followed in this study and Table 4 illustrates each step of study by explaining the function, input and output of steps. We presented a reference model to discover hidden knowledge from MEDLINE articles, but at the selection of organism phase, we applied two methods to select liver specific parasites and bacteria. These two methods can be proposed to use for similar studies.

We used the PubMed interface and a complex query for the retrieval of the MEDLINE abstracts that are relevant for the liver-specific parasites. Following query resulted in a document set of 17,377 articles:

```
(("echinococcosis"[MeSH Terms] OR "echinococcosis"[All Fields]) AND
("liver"[MeSH Terms] OR "liver"[All Fields])) OR (("fasciola"[MeSH Terms] OR
"fasciola"[All Fields]) AND ("liver"[MeSH Terms] OR "liver"[All Fields])) OR
(("amoeba"[MeSH Terms] OR "amoeba"[All Fields] OR "amebic"[All Fields]) AND
("liver"[MeSH Terms] OR "liver"[All Fields])) OR (("schistosoma"[MeSH Terms]
OR "schistosoma"[All Fields]) AND ("liver"[MeSH Terms] OR "liver"[All
Fields])) OR (Clonorchis[All Fields] AND ("liver"[MeSH Terms] OR "liver"[All
Fields])) OR (("opisthorchis"[MeSH Terms] OR "opisthorchis"[All Fields]) AND
```

("liver"[MeSH Terms] OR "liver"[All Fields])) OR (("plasmodium"[MeSH Terms] OR "plasmodium"[All Fields]) AND ("liver"[MeSH Terms] OR "liver"[All Fields])) OR ("Liver Diseases, Parasitic"[Mesh])

All documents were processed with the text mining solution available at the EBI called “Filter Server” (Rebholz-Schuhmann et al. 2008). The filter server identifies the entities mentioned such as species, drugs and diseases and their variants in the text (Rebholz-Schuhmann D. 2007).

The architecture of biomedical text mining system in EBI provides a framework for the extraction of facts from biomedical literature. The architecture consists of some specific tasks:

- (a) Procure fast and reliable access to MEDLINE articles
- (b) Provide tools and means for extracting the terminology contained in various data sources and named entity recognition (NER)
- (c) Provide a framework for the integration of the different NER modules and the infrastructure necessary to develop and improve methods for the extraction of facts the biomedical literature
- (d) Develop and maintain applications

The architecture is based on regular expressions. The Java library binds regular expressions to actions that are automatically executed whenever a match occurs in the text stream being processed. At a match, the associated action can modify the stream or leave it unchanged. Commonly, XML tags are used to mark named entity and other regions of interest in the text. Several thousand regular expression/action pairs can be combined into one machine, called a Deterministic Finite Automation (DFA) and can be used in parallel from a computer application (Rebholz-Schuhmann et al. 2007). Available modules for linguistic tasks in EBI architecture:

Tokenizer: The tokenizer separates the input text into its components, called the tokens.

POS tagging: Two types of POS tagging is available, The TreeFilter Server and the POS Filter Server from the Centrum for information and speech processing.

Sentenciser: It splits the stream of text into its sentences

Parsing: The integration of parsing solutions is being developed.

In EBI architecture, there are some filter server solutions and each filter server specializes in recognizing the vocabulary of a particular terminology and performing specialized actions depending on the input it receives. A filter server receives send a stream of text and annotates it. The server runs its embedded software on the incoming text to recognize and tag the terminology with XML tags (Rebholz-Schuhmann et al. 2007). We used organism filter server which tags the names of species taken from Entrez NCBI Taxonomy. All species in the articles were annotated in XML. Following the article in XML shows annotation of species in a sample MEDLINE article (Figure 4) and Table 3 shows the list of filter servers used to extract species, drugs, diseases and genes names from collected MEDLINE articles.

Table 3: List of filter servers

Name	Prerequisite	Function	Output
Species	<plain>	Tags species names taken from Entrez Taxonomy	<z:species ids=""%1"">%0</z:species>
Drugbank	<plain>	Tags drug names taken from Drugbank	<z:drug ids=""%1"">%0</z:drug>
UMLS Individual Disease Server	<plain>	Tags disease names taken from UMLS diseases	<z:disease ids=""%1"">%0</z:disease>
Swissprot	<plain>	Tags protein or gene names taken from UniProt	<z:uniprot fb=""%1"" ids=""%2"">%0</z:uniprot>

After annotating the species in the articles, parasites' names affecting liver were selected by two clinicians. The frequencies of the amount of times a parasite was considered

appears in the selected articles were calculated. Most ranked of them were used for analysis (Table 4).

```
<PubmedArticle>
  <MEDLINECitation Owner="NLM" Status="In-Data-Review">
    <PMID>18953152</PMID>
    <Title>Journal of postgraduate medicine</Title>
    <Abstract>
      <AbstractText><plain>Echinococcal cysts usually involve the liver;
      extrahepatic localization is reported in 11% of all cases of abdominal hydatid disease.
      We report a case of a prevesical hydatid cyst. A 53-year-old man was admitted with a
      large suprapubic mass. Ultrasonography and computed tomography revealed a cystic
      mass situated in front of the urinary bladder. There were no cysts in any other location.
      Serological tests were positive for <z:species ids="6209">Echinococcus</z:species>.
      The patient was operated on and the cyst was completely excised. The pathologic
      examination confirmed the diagnosis of <z:species
      ids="6209">Echinococcus</z:species>. Isolated hydatid cyst situated in front of the
      urinary bladder has never been described in the literature. Hydatid cyst should always
      be considered in the differential diagnosis of abdominopelvic masses in endemic
      regions, before any procedure like puncture, biopsy or cystectomy, in order to avoid
      dissemination of the cystic contents or an anaphylactic shock.</plain></AbstractText>
    </Abstract>
```

Figure 5: XML annotation of species

Table 4 shows eight of the highest ranked parasites' names in the articles and their frequencies. Table 5 shows MeSH and NCBI Taxonomy Ids for these parasites. At the second part of our study, we analyzed bacteria. Bacteria names were selected from MeSH data in National Library of Medicine (NLM) and MEDLINE was searched to find

most frequent bacteria with liver (Table 6). Table 7 shows number of articles for selected bacteria.

Table 4: Frequencies of selected parasites

Liver Specific Parasites	Frequencies
Fasciola Hepatica	3378
Schistosoma Mansoni	3114
Schistosoma Japonicum	973
Entamoeba Histolytica	872
Echinococcus Granulosus	708
Echinococcus Multilocularis	401
Clonorchis Sinensis	372
Opisthorchis Viverrini	175

Table 5: MeSH and NCBI Taxonomy Ids for Selected Parasites

MeSH ID and Descriptor	NCBI Taxonomy ID
D005210 Fasciola Hepatica	6192
D012550 Schistosoma Mansoni	6183
D012549 Schistosoma Japonicum	6182
D004748 Entamoeba Histolytica	5759
D048209 Echinococcus Granulosus	6210
D048210 Echinococcus Multilocularis	6211
D003004 Clonorchis Sinensis	79923
D009891 Opisthorchis Viverrini	6198

Table 6: MeSH Ids for selected bacteria

Bacteria	MeSH ID
Salmonella Typhimurium	D012486
Staphylococcus Aureus	D013211
Helicobacter Pylori	D016480
Mycobacterium Tuberculosis	D009169
Listeria Monocytogenes	D008089
Klebsiella Pneumoniae	D007711
Pseudomonas Aeruginosa	D011550
Streptococcus Pneumoniae	D013296

Table 7: Number of articles for selected bacteria

Bacteria	Number of Articles
Salmonella Typhimurium	4025
Staphylococcus Aureus	1019
Helicobacter Pylori	634
Mycobacterium Tuberculosis	971
Listeria Monocytogenes	913
Klebsiella Pneumoniae	554
Pseudomonas Aeruginosa	489
Streptococcus Pneumoniae	316
Total	8921

A drug time analysis for each parasite and bacterium was developed and relevant articles belonging to specific time periods (e.g., 1970-1980, 1980-1990, 1990-2000 and 2000-2009) was found. PubMed offers many search options for users. For example, in order to find articles relevant for “Fasciola hepatica” published in 1970-1980, a researcher can use limits option to set “1970-1980” specific date range. Table 8 shows number of articles according to time periods for Fasciola hepatica and Table 9 shows total number of articles for selected parasites. The following query was used for finding the articles between this time period and this query was modified for other time periods to collect time specific articles.

"Fasciola hepatica"[All Fields] AND "liver"[All Fields] AND ("1970"[EDAT] : "1980"[EDAT])

Table 8: Number of articles according to time periods for Fasciola Hepatica

Time Period	Number of Articles
1970-1980	211
1980-1990	203
1990-2000	218
2000-2009	285
Total	917

MEDLINE includes biomedical articles back to 1950s. However, most of articles belonging to 1950-1960 and 1960-1970 time periods do not contain abstracts. For example, some articles published in between 1960-1970 were retrieved from PubMed for analysis of Fasciola hepatica, but few drug names were extracted in these articles and the result does not provide enough information to make comparison with other time periods. Therefore, this time period was removed in time analysis.

Table 9: Total number of articles for 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

After retrieving the articles in specific time periods, drug names were found by using drug filter server which tags drugs names taken from Drugbank database. The Drugbank database is a bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e., sequence, structure and pathway) information(Wishart et al. 2006). After using drug filter server, all drugs in the articles were annotated in XML. At the next step, frequencies of drugs were calculated for each time periods. A software developed in Java was used to find the frequencies. In addition, drugs' therapy classes were searched in Drugbank to find their categories. After finding categories of all drugs, antehelminthic drugs and antibacterial drugs which are main classes of treatment for parasites and bacteria were selected for time based analyses. Since the number of articles varies in each time period, the frequencies of drugs were divided by the number of articles to get the frequencies. Minitab statistical software was used to create time series plots for each

parasite and bacterium and the differences between frequencies were shown in these plots.

Time series plots provide drug time analysis for clinicians to identify drug usage over time and to compare drugs according to their frequencies.

3.2 Normalization of Drugs

Drugs have some variations such as synonyms and brand names. For example, Retinol is a synonym of Vitamin A. On the other hand, Vermox is a brand name of Mebendazole. In this study, Drugbank was searched for each drug, synonyms and brand names of drugs are found. Drugbank is one of the biggest resource for drugs and currently contains > 4100 drug entries, corresponding to >12000 different trade names and synonyms (Wishart et al. 2006). In addition, it provides a fully searchable web-enabled resource with many built-in tools and features for viewing, sorting and extracting drug or drug target data.

Table 10: Normalization of drugs for parasites

Drug Name Variations	Normalized Names	Drugbank ID
Ciclosporin, Cyclosporine, Cyclosporine A	Cyclosporine	DB00091
Caffeine, Triad	Triad	DB00201
Interlekun 2, IL-2	IL-2	DB00041
Phenazone, Antipyrine	Antipyrine	DB01435
Retinol, Vitamin A	Vitamin A	DB00162
Vermox, Mebendazole	Mebendazole	DB00643
Biltricide, Praziquantel	Praziquantel	DB01058
Flagyl, Metronidazole	Metronidazole	DB00916

After finding variations, these names were manually normalized to one specific name. For example, Ciclosporin, Cyclosporine and Cyclosporine A were mapped to Cyclosporine. Table 10 and 11 shows drug name variations for parasites and bacteria, normalized names, and Drugbank ID which uniquely identifies each drug in Drugbank database.

Table 11: Normalization of drugs for bacteria

Drug Name Variations	Normalized Names	Drugbank ID
Phenobarbital, Luminal	Phenobarbital	DB01174
Methyldopa,Aldomet	Aldomet	DB00968
Alpha-Tocopherol, Vitamin E	Vitamin E	DB00163
Amsacrine, Mamsa	Amsacrine	DB00276
Aspirin, Salicylic acid	Aspirin	DB00945
Sulfasalazine, Asulfidine,Salazopyrin	Sulfasalazine	DB00795
IL-2, Interleukin-2	IL-2	DB00041
Phenylbutazone, Butazolidin	Phenylbutazone	DB00812
Cyclosporine,Ciclosporin, Cyclosporin, Cyclosporin A	Cyclosporine	DB00091
Diphenhydramine, Benadryl	Diphenhydramine	DB01075
Dapsone, DDS	Dapsone	DB00250

“Table 11(cont.)”

Dicumarol,Dicoumarol	Dicumarol	DB00266
Benzphetamine, Didrex	Benzphetamine	DB00865
Theophylline,Elixophyllin, Lanophyllin, Slo-Phyllin, Respbid, Quibron	Theophylline	DB00968
Erythromycin, Erythrocin Stearate, Pce	Erythromycin	DB00277
Metronidazole, Flagyl	Metronidazole	DB00199
Triamterene, Maxzide	Triamterene	DB00698
Testosterone, Methyltestosterone	Testosterone	DB00384
Primidone,Mysoline	Primidone	DB00624
Procarbazine, Natulan	Procarbazine	DB00794
Oxytetracycline, Terramycin	Oxytetracycline	DB01168
Ampicillin, Polycillin, Principen	Ampicillin	DB00595
Praziquantel, Pyrantel	Praziquantel	DB01032
Zidovudine, Retrovir	Zidovudine	DB01058
Oxazepam, Serax	Oxazepam	DB00495
Caffeine, Triad	Caffeine	DB00842
Vitamin a, Retinol	Vitamin a	DB00201

After time based analyses of each parasite and bacterium, statistical co-occurrence based relation extraction and clustering analyses were performed respectively. In time based analyses, articles related to considered parasite or bacterium were selected in specific time periods. However, in order to perform statistical co-occurrence based relation extraction, all articles were collected from MEDLINE without using specific time ranges. Also, all drugs, diseases and genes in these articles were found by using filter servers solutions in EBI. After annotating these entities, their frequencies were calculated to use for statistical co-occurrence based relation extraction. In clustering analyses, drug data in different time periods were integrated to find similarities for the treatment of parasites or bacteria.

3.3 Statistical Co-occurrence Based Relation Extraction

Information about pairwise association between biomedical concepts, such as genes, proteins, diseases and chemical compounds constitutes an important part of biomedical knowledge. It is common for a researcher to need answers to questions like “What diseases are relevant to a particular gene?” Or “What chemical compounds are relevant to a particular disease?” Text mining complements biomedical databases by providing researchers with a convenient way to find such information from the literature (Tsuruoka, Tsujii and Ananiadou 2008).

Co-occurrence statistics have proven very effective in discovering associations between concepts. One important example comes from information retrieval. As early as the late 1950s, statistical co-occurrence was explored as a means of enlarging and sharpening literature searches by several researchers. In another example computational linguistics were applied for word sense disambiguation, word clustering and lexicon construction. Since the 1990s, co-occurrence statistics have also been widely used in synonym mining and word-to-word translation (Cao et al. 2005).

Several measures were selected to rank the co-occurrences that exploit different properties. From the available measures for co-occurrence analysis, we selected the following ones representatives of different issues:

Frequency $P(w_i, w_j)$: The most frequent co-occurrences appear at the top ranks. In this measure we do not consider the individual distributions of w_i and w_j (Jimeno Yepes 2008).

Pointwise mutual information: Pointwise Mutual Information (PMI) (or specific mutual information) is a measure of association used in information theory and statistics. The PMI of a pair of outcomes x and y belonging to discrete random variables quantifies the discrepancy between the probability of their coincidence given their joint distribution versus the probability of their coincidence given only their individual distributions and assuming independence⁷.

Pointwise mutual information is defined as $\log\left(\frac{P(w_i, w_j)}{P(w_i)P(w_j)}\right)$, where $P(w_i)$ is the proportion of the documents that match the query; $P(w_j)$ is the proportion of the documents that contain the concept; and $P(w_i, w_j)$ is the proportion of the documents that match the query and contain the concept. Pointwise mutual information gives an indication of how much more the query and concept co-occur than we expect by chance (Tsuruoka et al. 2008).

Table 12: Summary of liver specific Fasciola Hepatica and albendazole co-occurrence data

	Count
Total Articles of MEDLINE	18000000
Fasciola Hepatica and albendazole Co-occurrence Frequency	47
Fasciola Hepatica individual frequency	917
Albendazole individual frequency	466

⁷ http://en.wikipedia.org/wiki/Pointwise_mutual_information

In this study, $P(w_i, w_j)$ shows the frequency which specific parasite and drug are appeared together in the articles. $P(w_i)$ is the individual frequency of the parasite or bacteria and $P(w_j)$ is the individual frequency of the drug. Table 12 illustrates an example of data for relationship extraction. It lists a summary of liver specific Fasciola Hepatica and albendazole co-occurrence data.

An example of PMI calculation for Fasciola Hepatica and albendazole co-occurrence:

w_i : Fasciola Hepatica

$$P(w_i) = \frac{917}{18000000}$$

w_j : Albendazole

$$P(w_j) = \frac{466}{18000000}$$

$$P(w_i, w_j) = \frac{47}{917}$$

$$PMI = \left(\log \frac{P(w_i, w_j)}{P(w_i)P(w_j)} \right) \text{ (Equation 1)}$$

$$PMI = 28.52$$

There are several Fasciola Hepatica-drug pairs. For each drug, PMI was calculated and ranked in a table. PMI was also used for diseases and genes which can be related to each parasite and bacteria. Due to the sensitiveness of some of these measures, less frequent

pairs were not considered and most frequent ten pairs were selected to extract relationships.

3.4 Clustering Analysis

Clustering analysis is one area of learning by machine that is 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 a multidimensional space. Let us assume that X is a pattern. X typically consists of m components, represented in multidimensional space as:

$$X=(X_1, X_2, \dots, X_m)$$

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 dendrogram or tree that shows this hierarchical cluster structure. Some hierarchical methods are divisive; those progressively divide the one large cluster comprising 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 (M. 2006).

In numeric clustering methods, 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 (i.e., features) measured (Beckstead 2002).

The basic agglomerative hierarchical clustering algorithm is as follows:

1. Compute the proximity matrix
2. **Repeat**
3. Merge the closest two clusters

4. Update the proximity matrix to reflect the proximity between the new cluster and the original clusters
5. **Do this until** only one cluster remains

Figure 6 shows the result of applying the single link technique to data set comprising of six points. Figure 6 also shows the nested clusters as a sequence of nested ellipses, where the numbers associated with the ellipses indicate the order of the clustering. Figure 7 shows the same information, but as a dendrogram. The height at which two clusters are merged in the dendrogram reflects the distance of the two clusters (Vipin 2006).

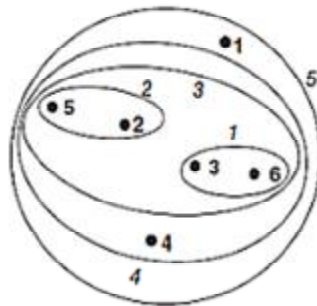


Figure 6: Single link clustering

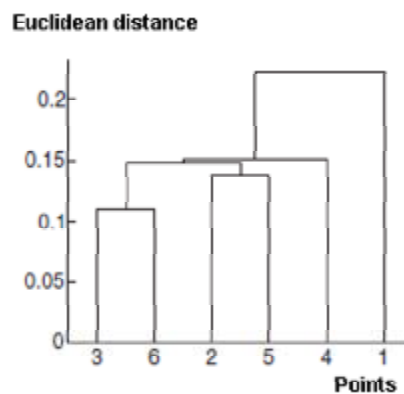


Figure 7: Single link dendrogram

CHAPTER 4

4 ANALYSIS OF PARASITES

4.1 Historical Review of Antiparasitic Drugs

Until the 1960's, the chemotherapy of human intestinal and systemic helminthiases was extremely unsatisfactory. The first of the benzimidazole compounds, thiabendazole was introduced into clinical medicine in the early 1960's. Others soon followed, and the most recently introduced is albendazole. Table 13 shows the timeline of some antiparasitic drugs (Harder 2002), (Cook 1991) and (White 2004)^{8,9,10,11}.

In the 1960s, the introduction of benzimidazole derivatives as potent, broad-spectrum anthelmintics opened up a new era in the control of parasitic diseases in veterinary medicine. Many benzimidazoles were screened, and some of them have been marketed and are now of widespread use in the animal-health industry. Mebendazole, and newer benzimidazole carbamate, albendazole, were further developed for the treatment of human intestinal helminthiases (Kern 2003). In 1979, albendazole, became available for the treatment of hydatid disease.

Albendazole or mebendazole resulted in greater clinical and parasitological efficacy, lower rates of morbidity and mortality, lower rates of disease recurrence, and shorter

⁸ <http://www.drugbank.ca>

⁹ <http://www.drugs.com>

¹⁰ <http://en.wikipedia.org/wiki>

¹¹ <http://www.chemindustry.com>

hospital stays when compared to subjects undergoing surgical intervention in humans with hepatic cystic echinococcosis. A similarly good therapeutic response using mebendazole and albendazole in human hydatid disease was reported by (Teggi, Lastilla and Derosa 1993) (Canete, et al. 2009).

Table 13:Timeline of some antiparasitic drugs

Year of introduction/discovery	Drug	Brand Names
Halogenated phenols and bisphenols		
1933	Bithionol	Actamer, Bidiphen, Bisoxyhen, Bithin, Bitin, Bitionol, Lorothidol, Lorothiodol, Neopellis, Nobacter, Prevenol, Vancide BL
Salicylanilides		
1960	Niclosamide	Niclocide
Benzimidazoles		
1961	Thiabendazole	Apl-Luster, Arbotect, Bioguard, Bovizole, Chemviron TK 100, Cropasal, Drawipas, Eprofil, Equivet TZ, Equizole, Hokustar hp, Lombristop, Mertec, Mertect, Mertect 160, Mertect 340f, Mertect Isp, Metasol TK 10, Metasol TK 100, Mintesol, Mintezol, Minzolum, Mycozol, Nemacin, Nemapan, Omnizole, Ormogal, Polival, RPH, Sanaizol 100, Sistesan, Storite, TBZ 6, TBZ 60W, Tebuzate, Tecto, Tecto 10P, Tecto 40F, Tecto 60, Tecto B, Tecto rph, Testo, Thiaben, Thibendole, Thibenzole, Thibenzole 200, Thibenzole att, Thiprazole, Tiabenda, Tibimix 20, Tobaz, Top form wormer, Triasox, Tubazole
1971	Fenbendazole	Panacur, Safe guard

“Table 13(cont.)”

1971	Flubendazole	Flubendazol
1972	Mebendazole	Bantenol, Besantin, Equivurm Plus, Lomper, MBDZ,MEBENDAZOLE 99%, Mebendazole (JAN/USP), Mebendazole(USAN), Mebendazole, Mebenoazole, Mebenvet, Mebex, Mebutar, Noverme, Ovitelmin, Pantelmin, Telmin, Vermicidin, Vermirax, Vermox, Vermox (TN), Verpanyl
1972	Oxfendazole	
1973	Oxamniquine	Mansil, Vansil
1979	Albendazole	Albenza, Eskazole, Valbazen, Zentel
1983	Triclabendazole	Fasinex
1983	Ivermectin	Ivermectin B1, Ivermectin-luminol, Mectizan, Stromectol
Isoquinoline derivatives		
1975	Praziquantel	Biltrice
Antibacterial, Antiprotozoal		
Early 1990s	Nitazoxanide	Alinia, Fental, Phavic-1

Niclosamide is a well established drug which has been used in several countries since 1960. The current review supports the continuous use of niclosamide as no major safety concerns have been raised (Ofori-Adjei et al. 2008).

In the early 1970s, praziquantel was jointly discovered by Bayer and Merck in Germany and the first studies on human volunteers were reported in 1978 (Cioli and Pica-Mattoccia 2003). The advent of praziquantel in the 1970s changed the landscape of research and development of drugs for treatment and morbidity control of schistosomiasis. Owing to its high efficacy against all five human schistosome species, good tolerability and ease of administration as a single oral dose, praziquantel has become the drug of choice for the treatment and morbidity control of schistosomiasis throughout the world (Shu-Hua 2005).

The first alarming reports of possible praziquantel resistance came from an intensive focus in Northern Senegal. Additional evidence for resistance to praziquantel was collected in Egypt. The schistosomes with decreased susceptibility to praziquantel exist and urgent efforts should be made towards the development of new antischistosomal drugs (Cioli and Pica-Mattoccia 2003).

Should the situation arise in which the use of praziquantel becomes no longer advisable (for lack of efficacy or for fear of increasing resistance), the obvious alternative would be oxamniquine. It has been introduced into clinical medicine for *Schistosoma Mansoni* and *Schistosoma haematobium* infections. It has an advantage because in most countries it is cheaper than praziquantel (Cook 1991). However its efficacy is restricted to *Schistosoma Mansoni* infections and its availability may be precarious in the future. It would be highly desirable to design and develop new antischistosome drugs (Cioli and Pica-Mattoccia 2003).

Ivermectin, which had also been widely used in veterinary medicine, first became available in clinical medicine in the early 1980s. The value of ivermectin was slowly recognized in clinical medicine. Extensive use in veterinary medicine over many years represents an excellent example of the low level of cooperation between these two major

disciplines, and a lack of extrapolation in application of chemotherapeutic regimens from animals to man (Cook 1991).

In 1985, resistance to ivermectin in parasites was an unrealized threat. Since then, the inevitable appearance of ivermectin resistant parasites has occurred. Although much remains to be learned about how the drug works and how resistance to it will develop, it has earned the title of “wonder drug” (Geary 2005).

Nitazoxanide is a new broad-spectrum antiparasitic agent. In contrast with other agents, it is being primarily developed to treat human infections. Initial studies were reported in 1984 but its clinical development only progressed after antiprotozoal activity was demonstrated in the early 1990s (White 2004).

The role of nitazoxanide as a broad spectrum anthelmintic is intriguing. Nitazoxanide may be able to replace benzimidazoles. If benzimidazole resistance spreads in human as it has in veterinary practice, it is certainly an advantage to have another agent for the treatment of intestinal tapeworms. If clinical studies confirm the *in vitro* activity, nitazoxanide may play a key role in chemotherapy of hydatid disease as well. The studies in fascioliasis suggest that nitazoxanide may prove an effective agent versus trematodes. If nitazoxanide can clear schistosomiasis, it may become the anthelmintic of choice for deworming programs in areas with both intestinal nematodes and schistosomiasis (White 2004).

Quinfamide and mebendazole are very effective drugs in the eradication of amebiasis, giardiasis and ascariasis. Similarly, nitazoxanide has demonstrated its effectiveness as a broad-spectrum antiparasitic drug in open-level and *in vitro* studies (Davila-Gutierrez et al. 2002).

A generic drug (generic drugs, short: generics) is a drug which is produced and distributed without patent protection. The generic drug may still have a patent on the formulation but not on the active ingredient. A generic must contain the same active ingredients as the original formulation. According to the U.S. Food and Drug Administration (FDA), generic drugs are identical or within an acceptable bioequivalent

range to the brand name counterpart with respect to pharmacokinetic and pharmacodynamic properties. Albendazole is one of the anthelmintics for the treatment of parasites and mebendazole, pyrantel and thiabendazole are the generic of albendazole¹².

4.2 Time Based Analysis of Parasites

Time based analysis was performed for each parasite and main drug class in the treatment of liver specific parasites was selected to create time series plots (methods were explained in chapter 3).

4.3 Statistical Co-occurrence Based Relation Extraction of Parasites

Statistical Co-occurrence Based Relation Extraction was applied on parasite/drugs, parasite/diseases and parasite/genes co-occurrence data (methods were explained in chapter 3).

4.4 Fasciola Hepatica

4.4.1 Drug Time Analysis of Fasciola Hepatica

For safety and efficacy, the drug of choice for the treatment of human fascioliasis caused by *Fasciola hepatica* is triclabendazole. In contrast to other trematodes, treatment with praziquantel is frequently unsuccessful ((White 2004) and (Fairweather 1999)). Triclabendazole has been used successfully to treat human cases of fascioliasis. Nevertheless, resistance to triclabendazole first appeared in farm animals in Australia in the mid-1990s and since then has been reported in a number of European countries (Ireland, the UK, the Netherlands and Spain). The heavy reliance on a single drug puts treatment strategies for fascioliasis at risk (Brennan et al. 2007). In recent studies, nitazoxanide is suggested for treatment (White 2004).

¹² <http://www.emedexpert.com>

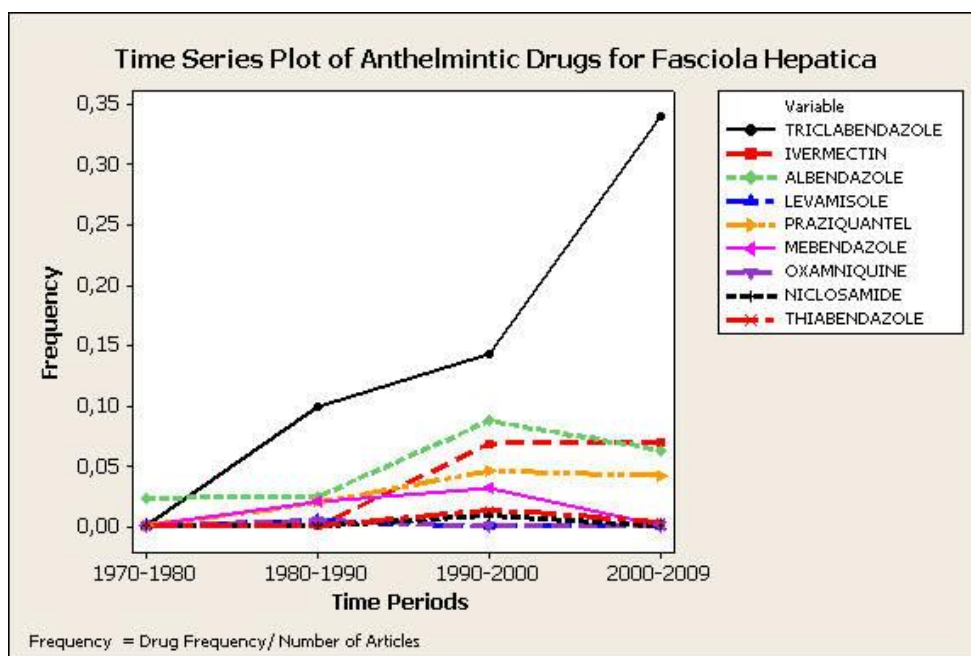


Figure 8: Time series plot of anthelmintic drugs for Fasciola Hepatica

Figure 8 shows drug time series plot of Anthelmintic drugs for Fasciola hepatica. According to the figure, triclabendazole is seen as the most preferred drug for the treatment of Fasciola hepatica. Other drugs such as albendazole and praziquantel have some frequencies but comparing to triclabendazole, their frequencies are very low.

4.4.2 Relation Extraction of Fasciola Hepatica

Appendix A Table 16 shows Fasciola hepatica/drugs co-occurrences based on the Drugbank filter server and Appendix A Figure 42 shows time series plot for most ranked drugs in both co-occurrence and PMI list. Appendix A Table 17 shows Fasciola Hepatica/disease co-occurrences based on the UMLS disease filter server and Appendix A Table 18 shows Fasciola Hepatica/genes co-occurrences based on the Swissprot filter server.

4.5 Schistosoma Mansoni

4.5.1 Drug Time Analysis of Schistosoma Mansoni

Schistosoma Mansoni is a significant parasite of humans and one of the major agents of schistosomiasis. Praziquantel that is active against all schistosome species is now the most widely used. No important long-term safety risks have been documented in people so far. It eliminates the already matured worms and has little or no effect on eggs and immature worms (Gryseels et al. 2006).

Oxamniquine acts only on Schistosoma Mansoni and it is as effective as praziquantel but can provoke more pronounced side-effects, most notably drowsiness, sleep induction and epileptic seizures (Gryseels et al. 2006).

Recent studies indicate no deaths associated with oxamniquine use in the treatment of schistosomiasis caused by Schistosoma Mansoni. The review compared oxamniquine with praziquantel and indicated important differences in some of the effects as shown in Table 14 (Ofori-Adjei et al. 2008).

Table 14: Side effects of oxamniquine compared to praziquantel

Side Effects	Oxamniquine (F/N(%))	Praziquantel (F/N(%))
Diarrhoea	38/544(7%)	74/536(14%)
Abdominal pain	115/571(20%)	240/563(42%)
Myalgia	7/352(2%)	0/327(0%)
Seizure	2/372(0.5%)	

In the absence of a vaccine, efficient vector control, and water sanitation, the treatment and control of schistosomiasis relies heavily on a single drug, praziquantel. Another alternative, oxamniquine, is also available, but its bioactivity is restricted to Schistosoma Mansoni and the drug has largely been replaced in favor of the more cost-effective.

Today, praziquantel is the recommended drug for disease treatment at either the community or individual level (Caffrey 2007).

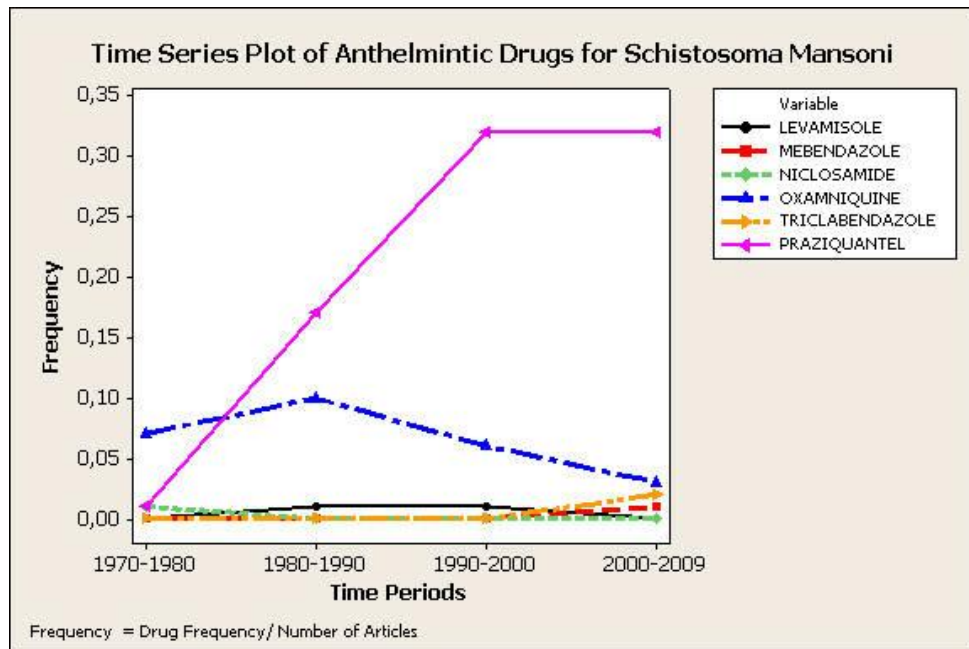


Figure 9: Time series plot of anthelmintic drugs for Schistosoma Mansoni

Since the first clinical trials, praziquantel has been a tremendous success because it is a single dose therapy that is effective, non-toxic and relatively cheap. Nonetheless, praziquantel has several short-comings, and there is an increased awareness that now is the time to identify new drugs, either to complement praziquantel or replace it should it fail (Caffrey 2007).

In the late 1980s, a massive outbreak of Schistosoma Mansoni occurred in the Senegal river basin (SRB) after the construction of a dam and subsequent water resource development (Gryseels et al. 2000).

Figure 9 shows time series plot of anthelmintic drugs for Schistosoma Mansoni. According to figure, some anthelmintic drugs are associated with this parasite. Praziquantel is seen as the most preferred drug in the treatment of Schistosoma Mansoni. In addition, oxamniquine had a high frequency between 1970-1990 but its frequency

dropped after 1990. The other anthelmintic drugs such as triclabendazole and niclosamide have very low frequencies. This shows that they are not effective in the treatment of *Schistosoma Mansoni*.

4.5.2 Relation Extraction of *Schistosoma Mansoni*

Appendix A Table 19 shows *Schistosoma Mansoni*/drugs co-occurrences based on Drugbank filter server and Appendix A Figure 43 shows time series plot for most ranked drugs in both co-occurrence and PMI list. Appendix A Table 20 shows *Schistosoma Mansoni*/diseases co-occurrences based on the UMLS disease filter server and Appendix A Table 21 shows *Schistosoma Mansoni*/genes co-occurrences based on the Swissprot filter server.

4.6 *Schistosoma Japonicum*

4.6.1 Drug Time Analysis of *Schistosoma Japonicum*

Schistosomiasis caused by *Schistosoma Haematobium*, *Schistosomiasis Mansoni* and *Schistosomiasis Japonicum* is a chronic and debilitating disease that exacerbates poverty. The treatment and control of Schistosomiasis virtually relies on a single drug, praziquantel. This drug is increasingly used, hence there is mounting concern about the development of resistance to praziquantel. The drug targets the adult worm, but has only minor activity against the young developing stages; Therefore, retreatment is necessary to kill those parasites that have since matured. There is no dedicated drug discovery and development program pursued for Schistosomiasis, either by the pharmaceutical industry or through public-private partnerships. (Keiser et al. 2009) Reports show that mefloquine, a marketed drug prophylaxis and treatment of malaria, shows antischistosomal properties in laboratory studies with mice. It might be possible that use of mefloquine against malaria reduces the burden of Schistosomiasis and the potential ancillary benefit of the antimalarial drug mefloquine should be investigated against schistosomiasis (Keiser et al. 2009).

Some 20 years after the introduction of praziquantel, several cases of resistance have arose. The schistosomes developing resistance to praziquantel was predicted in 1993 by

Coli and colleagues (Fallon et al. 1996). Figure 10 shows time series of anthelmintic drugs for *Schistosoma Japonicum* and some breakpoints in this graphics may reveal the time of drug resistance. According to the figure, the frequency of praziquantel has decreased after 1990-2000 time period. This change may prove the resistance of praziquantel in 1990s. The other drug such as levamisole and oxamniquine have very low frequencies.

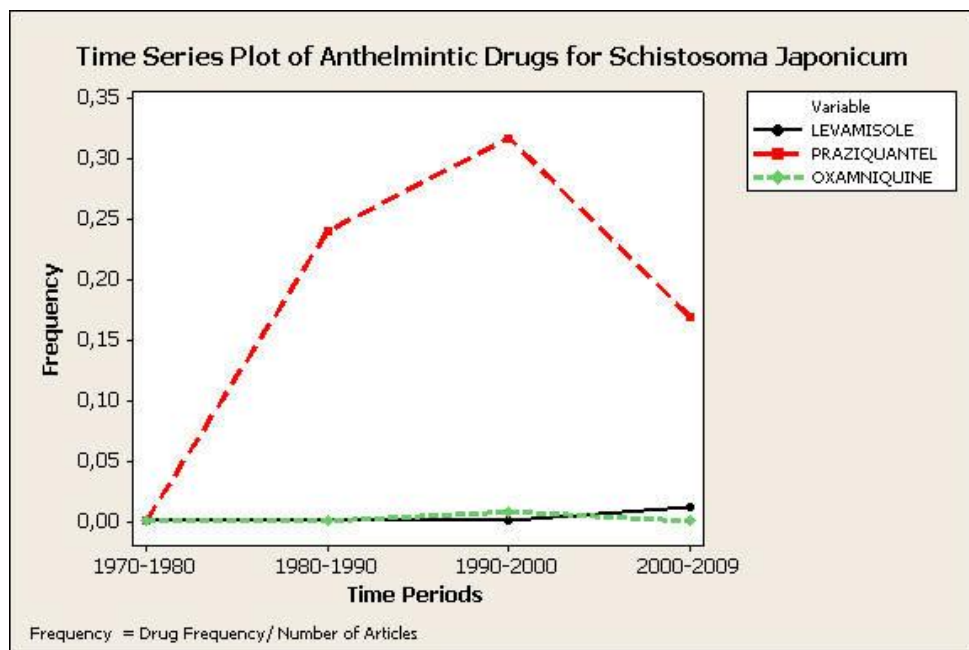


Figure 10: Time series of anthelmintic drugs for *Schistosoma Japonicum*

4.6.2 Relation Extraction of *Schistosoma Japonicum*

Appendix A Table 22 shows *Schistosoma Japonicum*/drugs co-occurrences based on the Drugbank filter server and Appendix A Figure 44 shows time series plot for most ranked drugs in both co-occurrence and PMI list. Appendix Table 23 shows *Schistosoma Japonicum*/diseases co-occurrences based on the UMLS disease filter server and Appendix A Table 24 shows *Schistosoma Japonicum*/genes co-occurrences based on the Swissprot server.

4.7 Entamoeba Histolytica

4.7.1 Drug Time Analysis of Entamoeba Histolytica

Entamoeba histolytica, associated with high morbidity and mortality continues to be a major public health problem throughout the world. It causes amoebiasis disease that is a major public health problem in developing countries (Bansal, Malla and Mahajan 2006).

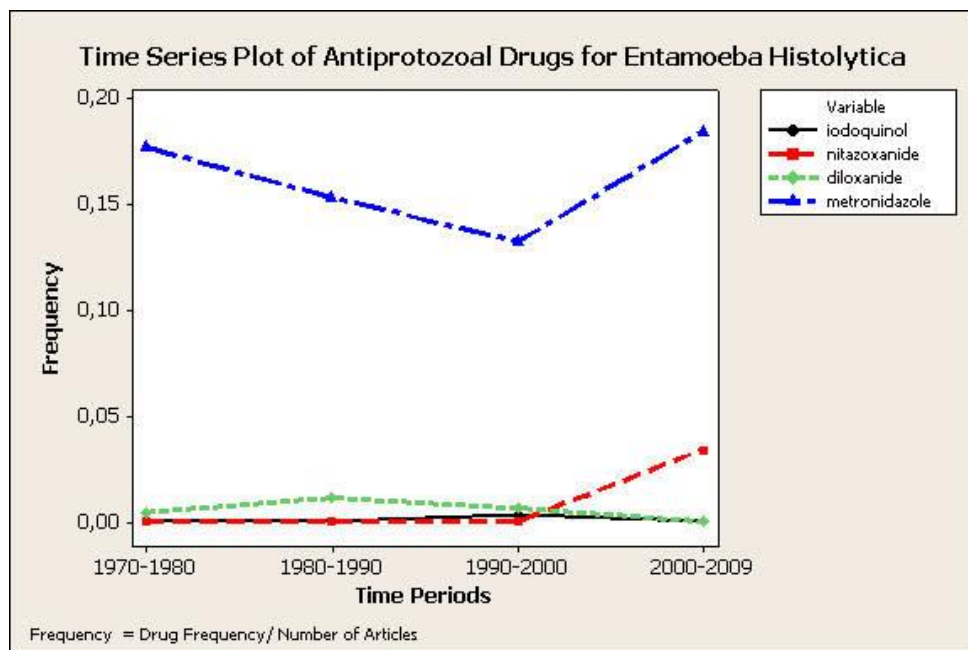


Figure 11: Time series plot of antiprotozoal drugs for Entamoeba Histolytica

Entamoeba histolytica is also a major cause of diarrhea and it infests millions of people worldwide each year, and approximately 40,000 to 100,000 people die annually from the disease. Metronidazole is most commonly available drug. Treatment with metronidazole is most commonly followed by the following agents: a luminal agent, paromomycin, diloxanide furoate and iodoquinol. Metronidazole remains the mainstay of treatment¹³. In most patients, the response is rapid and dramatic. Surgical drainage is considered only if there is no response to drug therapy, or when the diagnosis is uncertain (Ng et al. 2006).

¹³ http://utdol.com/patients/content/topic.do?topicKey=~KyryrqRcf_fR3tW/

Figure 11 shows time series plot of antiprotozoal drugs for *Entamoeba histolytica*. According to the figure, most frequent drug is metronidazole. Despite extensive worldwide use and some occasional reports of failure, acquired resistance to metronidazole is rare and does not appear to be a serious problem. From 2000-2009, nitazoxanide is seen and its frequency is higher than most of drugs in the table. Nitazoxanide is a new parasitic agent and may also prove to be an important luminal agent in amebiasis (White 2004).

4.7.2 Relation Extraction of *Entamoeba Histolytica*

Appendix A Table 25 shows *Entamoeba Histolytica*/drugs co-occurrences based on the Drugbank filter server and Appendix A Figure 45 shows a time series plot for most ranked drugs in both co-occurrence and PMI list. Appendix A Table 26 shows *Entamoeba Histolytica*/disease co-occurrences based on the UMLS disease server and Appendix A Table 27 shows *Entamoeba Histolytica*/genes co-occurrences based on the Swissprot filter server.

4.8 Echinococcus Granulosus

4.8.1 Drug Time Analysis of *Echinococcus Granulosus*

Hydatidosis is caused by the larval stage of *Echinococcus Granulosus* and its infection is found worldwide and has a higher prevalence in the world¹⁴.

Figure 12 shows time table of anthelmintic drugs for *Echinococcus Granulosus*. Albendazole has high frequency in all the time periods. Mebendazole also has a high frequency between 1970-1980. After this time period, its frequency has sharply decreased. Triclabendazole, praziquantel and levamisole are the other anthelmintic drugs associated with *Echinococcus Granulosus*, but they have very low frequencies.

¹⁴ <http://www.cdfound.to.it>

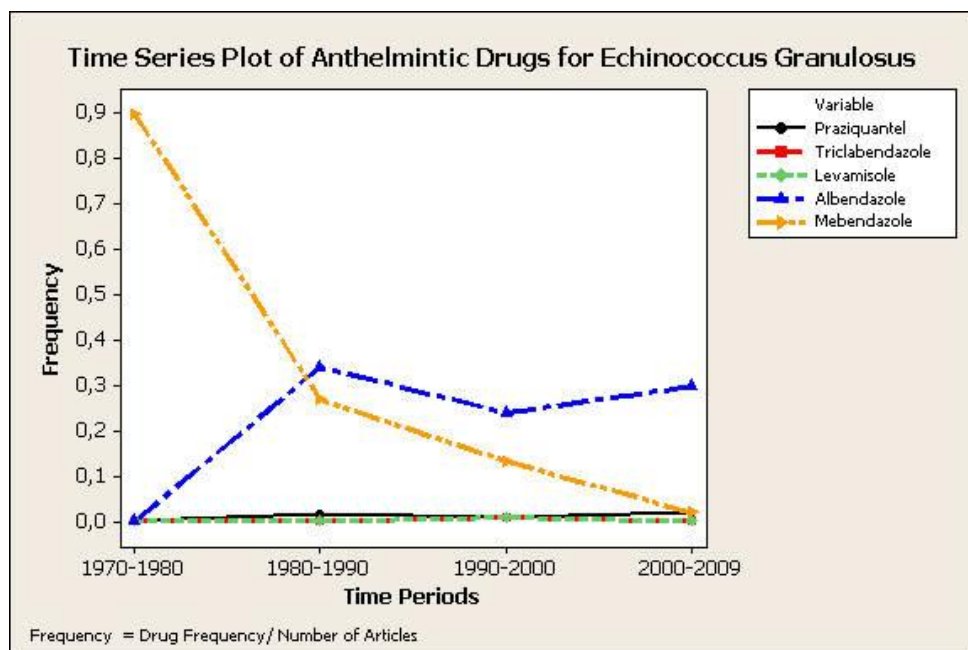


Figure 12: Time series plot of anthelmintic drugs for Echinococcus Granulosus

4.8.2 Relation Extraction of Echinococcus Granulosus

Appendix A Table 28 shows Echinococcus Granulosus/drugs co-occurrences based on the Drugbank filter server and Appendix A Figure 46 shows time series plot for most ranked drugs in both co-occurrence and PMI list. Appendix A Table 29 shows Echinococcus Granulosus/disease co-occurrences based on the UMLS disease filter server and Appendix A Table 30 shows Echinococcus Granulosus/genes co-occurrences based on the Swissprot filter server.

4.9 Echinococcus Multilocularis

4.9.1 Drug Time Analysis of Echinococcus Multilocularis

Echinococcus Multilocularis is one of four species of hydatid cyst-forming tapeworms that occur as adults in the small intestine of carnivores (Craig 2003). Mebendazole and albendazole are the only anthelmintic effective against cystic echinococcosis (Brunetti 2008).

Figure 13 shows a time series plot of anthelmintic drugs for *Echinococcus Multilocularis*. According to the figure, albendazole has high frequency between 1970 and 2000. Praziquantel and triclabendazole have been seen, but have very low frequencies. They should not be considered for time analysis. Mebendazole has high frequency in 1970-1980 time period, but since then its frequency has continuously decreased. Both mebendazole and albendazole are safe and side effects observed in studies are not severe and always reversible. However, many studies suggested that albendazole is significantly more effective than mebendazole for the treatment of liver cysts (Teggi et al. 1993) and (Brunetti 2008).

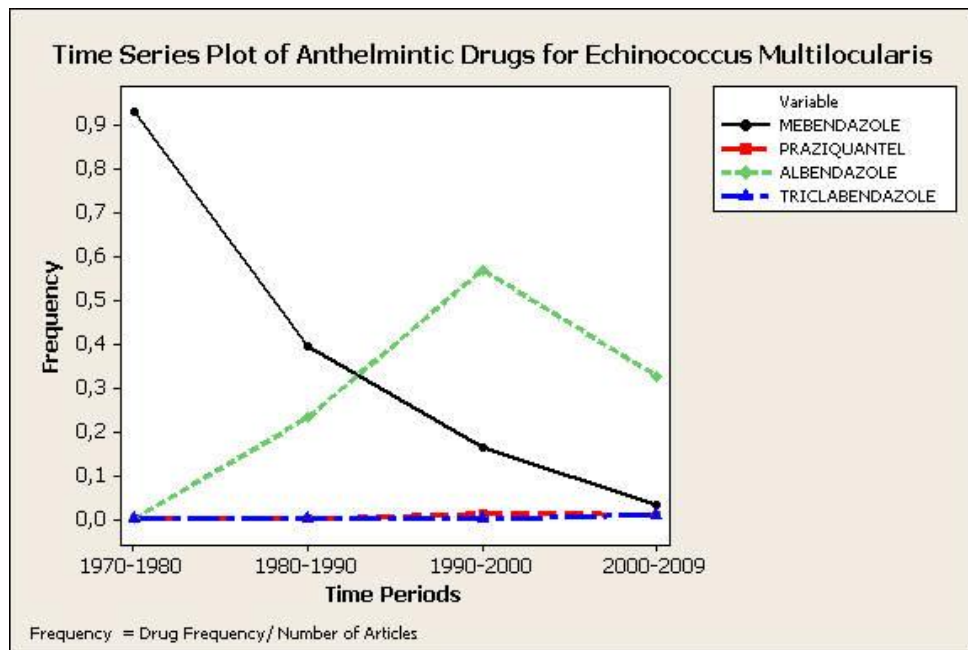


Figure 13: Time series plot of anthelmintic drugs for *Echinococcus Multilocularis*

The search for new drugs is ongoing. Oxfendazole has been tested for cystic echinococcosis in naturally infected animals. It seems at least as effective as albendazole and is easier to administer; determination of its relative efficacy warrants a comparison with albendazole. Due to *Echinococcus multilocularis*, Amphotericin B was proposed as salvage treatment in alveolar echinococcosis, but it does not seem to work in cystic echinococcus. The parasitocidal effect of nitazoxanide was recently proven in vitro, but has never been tested in human echinococcosis (Stamatakis et al. 2009).

4.9.2 Relation Extraction of Echinococcus Multilocularis

Appendix A Table 31 shows Echinococcus Multilocularis/drugs co-occurrences based on the Drugbank filter server and Appendix A Figure 47 shows a time series plot for most ranked drugs in both co-occurrence and PMI list. Appendix A Table 32 shows Echinococcus Multilocularis/diseases co-occurrences based on the UMLS disease filter server and Appendix A Table 33 shows Echinococcus Multilocularis/genes co-occurrences based on the Swissprot filter server.

4.10 Clonorchis Sinensis

4.10.1 Drug Time Analysis of Clonorchis Sinensis

Before introduction of praziquantel, treatment of clonorchiasis caused by Clonorchis Sinensis was with antimony preparations, gentian violet, emetine HCl, chloroquine diphosphate and dithiazanine iodide, but only temporary clinical improvement and negative or reduced egg counts could be achieved, but a complete cure was not obtained. Praziquantel proved to be safe and effective against trematodes in man (Rim 1984).

Figure 14 shows time series plot of anthelmintic drugs for Clonorchis Sinensis. According to the figure, praziquantel is the most effective drug for the treatment of Clonorchis sinensis. Albendazole has low frequency and only is seen in 1990-2000 time period.

4.10.2 Relation Extraction of Clonorchis Sinensis

Appendix A Table 34 shows Clonorchis Sinensis /Drugs Co-occurrences based on the Drugbank Filter Server and Appendix A Figure 48 shows time series plot for most ranked drugs in both co-occurrence and PMI list. Appendix A Table 35 shows Clonorchis Sinensis/diseases co-occurrences based on the UMLS disease filter server and Appendix A Table 36 shows Clonorchis Sinensis/genes co-occurrences based on the Swissprot filter server.

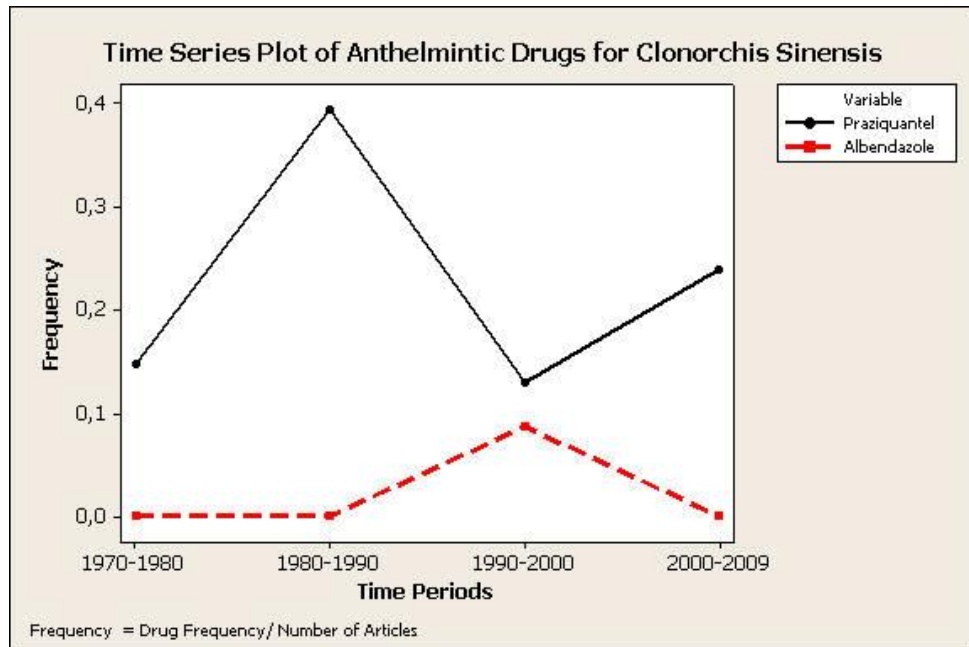


Figure 14: Time series plot of most ranked drugs for Clonorchis Sinensis

4.11 Opisthorchis Viverrini

4.11.1 Drug Time Analysis of Opisthorchis Viverrini

Opisthorchis Viverrini is an endemic disease and in Thailand, approximately 6 million people are infected (Marcos et al. 2008). The infection is associated with a number of hepatobiliary diseases, including cholangitis, obstructive jaundice, hepatomegaly, cholecystitis, cholelithiasis and cholangiocarcinoma (Kaewpitoon and Pengsaa 2008).

Figure 15 shows time series plot of anthelmintic drugs for Opisthorchis Viverrini. According to figure, praziquantel is the most effective drug for the treatment of Opisthorchis Viverrini. Ivermectin and albendazole have low frequencies.

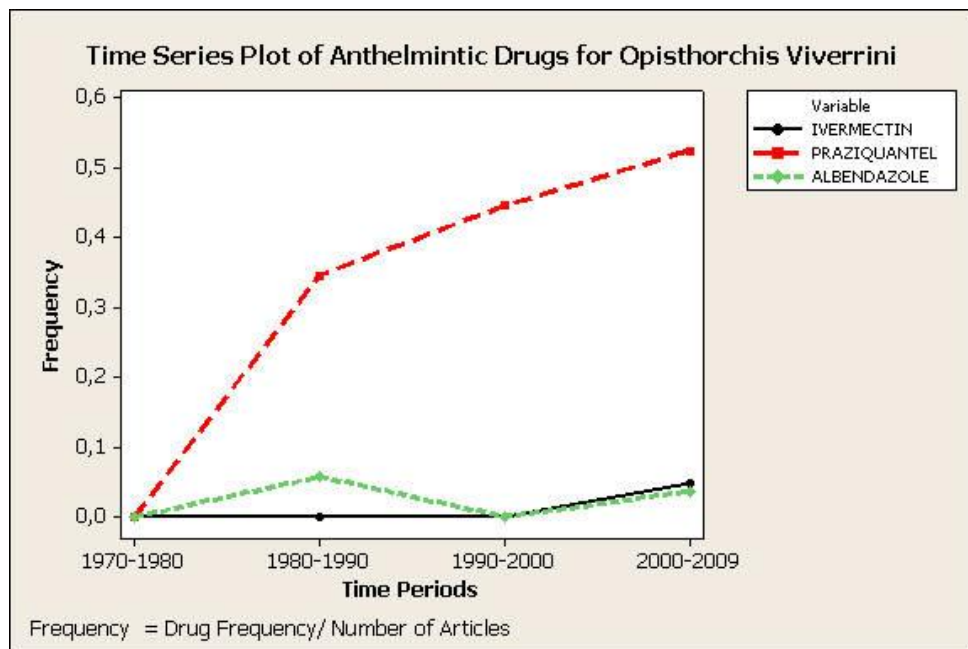


Figure 15: Time series plot of anthelmintic drugs for *Opisthorchis Viverrini*

4.11.2 Relation Extraction of *Opisthorchis Viverrini*

Appendix A Table 37 shows *Opisthorchis Viverrini*/drugs co-occurrences based on the Drugbank filter server and Appendix A Figure 49 shows time series plot for most ranked drugs in both co-occurrence and PMI list. Appendix A Table 38 shows *Opisthorchis Viverrini*/diseases co-occurrences based on the UMLS disease filter server and Appendix A Table 39 shows *Opisthorchis Viverrini*/genes co-occurrences based on Swissprot filter server.

4.12 General Distribution of Drugs in Main Classes

Figure 16, 17, 18 and 19 show distribution of analgesic, anthelmintic, antiinflammatory and antiprotozoal drugs in all of time periods respectively. These charts provide general view of main classes of drugs mentioned in the articles for liver specific parasites.

Distribution of Analgesic Drugs

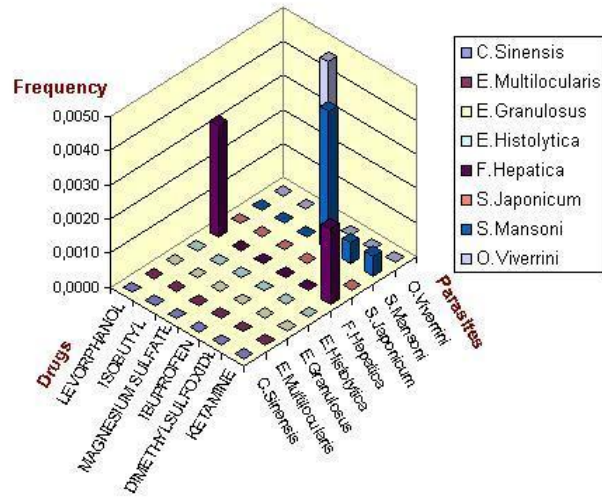


Figure 16: Distribution of analgesic Drugs

Distribution of Anthelmintic Drugs

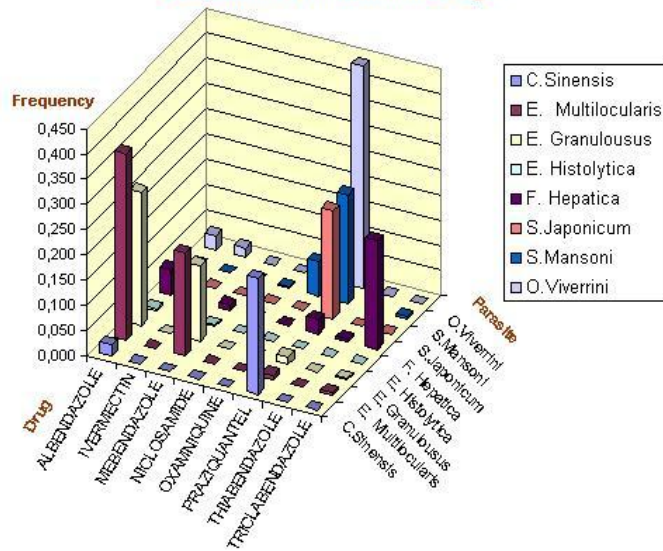


Figure 17: Distribution of anthelmintic drugs

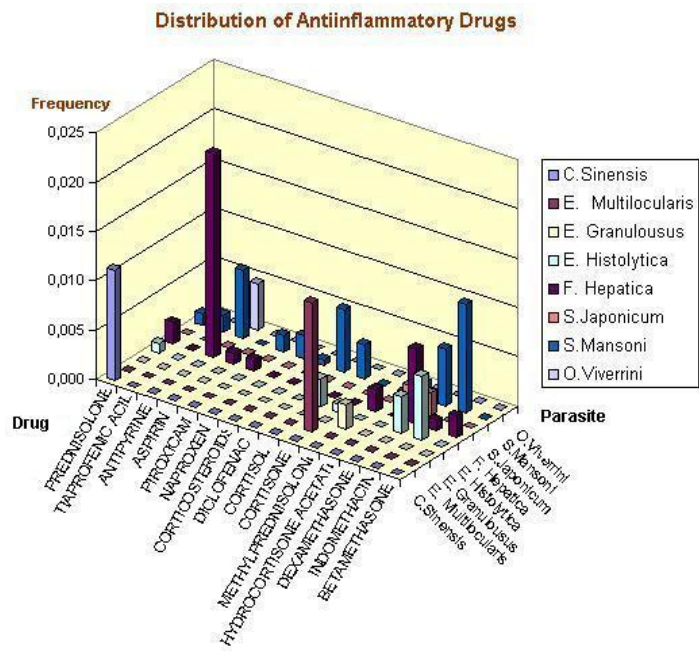


Figure 18: Distribution of antiinflammatory drugs

Distribution of Antiprotozoal Drugs

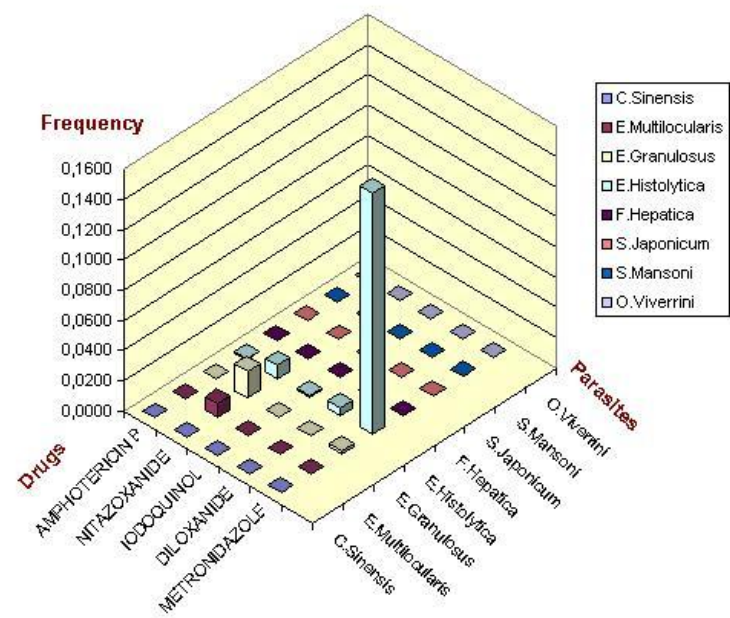


Figure 19: Distribution of antiprotozoal Drugs

4.13 Clustering Analysis of Parasites

Hierarchical clustering analysis was used to post-process the results from the co-occurrence analysis for the treatment of parasites with anthelmintic, antiprotozoal and antiinflammatory drugs. This analysis should enable the identification of drugs that can be applied across different species and that could be underexploited in a given species.

The main categories of drugs that are used for treatment of parasites comprise anthelmintic, antiinflammatory and antiprotozoal drugs. Figures 20, 21, 22, 23 and 24 show the heatmaps of the cluster analysis that were performed to examine the similarity of treatments.

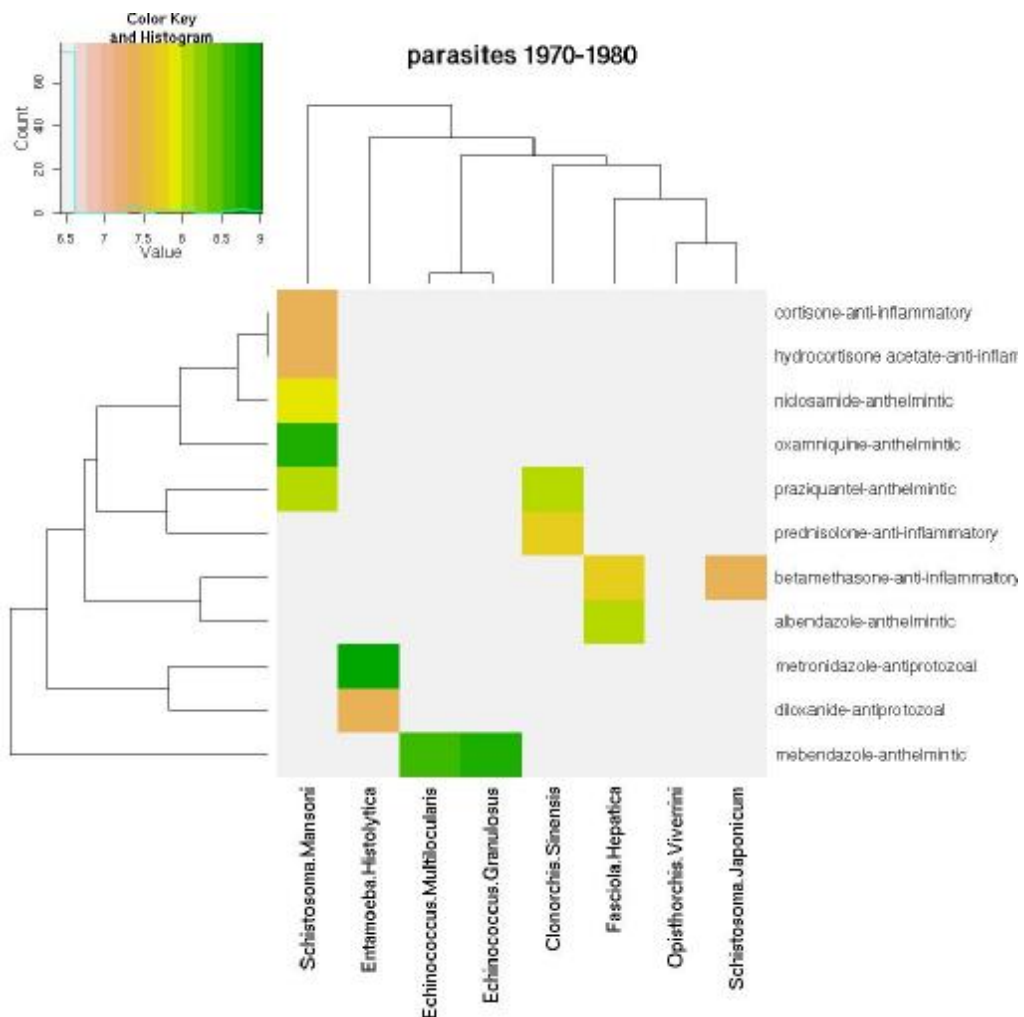


Figure 20: Drug heatmap of parasites for 1970-1980 time period

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 form the standart treatment for all three species. This raises the notion that drugs developed for the treatment of one species could, in principle, be exploited for the other two species.

The following characteristics could be seen for individual parasites:

During the period 1980-1990, the treatment of *Fasciola Hepatica* included a number of antiinflammatory drugs in addition to the anthelmintic treatment. Two conclusions explain this finding. The first, this parasite could induce significant inflammatory changes to the liver in contrast to the other two parasites, or second, *Fasciola Hepatica* could form the model parasite infection for the clinical trials of antiinflammatory drugs in inflammatory diseases in the liver (hepatitis model).

From 1990 to 2000, triclabendazole was added as treatment for *Echinococcus Granulosus*, and from 2000 to 2009, nitazoxanide has been included as a new antiparasitic agent in the treatment of *Echinococcus Multilocularis* and *Echinococcus Granulosus*.

The *Clonorchis Sinensis*, *Opisthorchis Viverrini* and *Schistosoma Mansoni* form the second cluster of 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:

Patients having an infectious disease with these species undergo treatment with anti-inflammatory 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.

Altogether, 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.

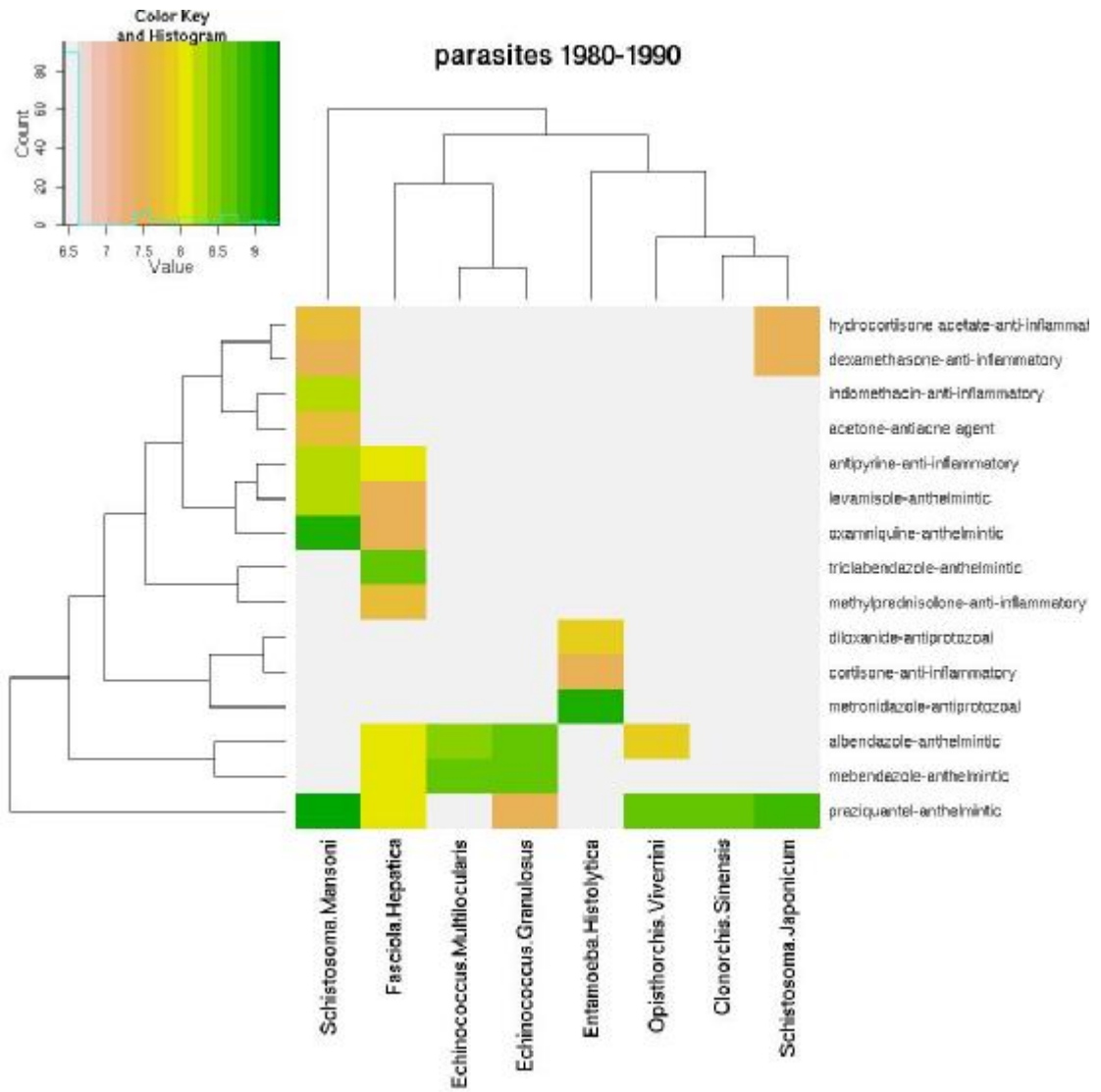


Figure 21: Drug heatmap of parasites for 1980-1990 time period

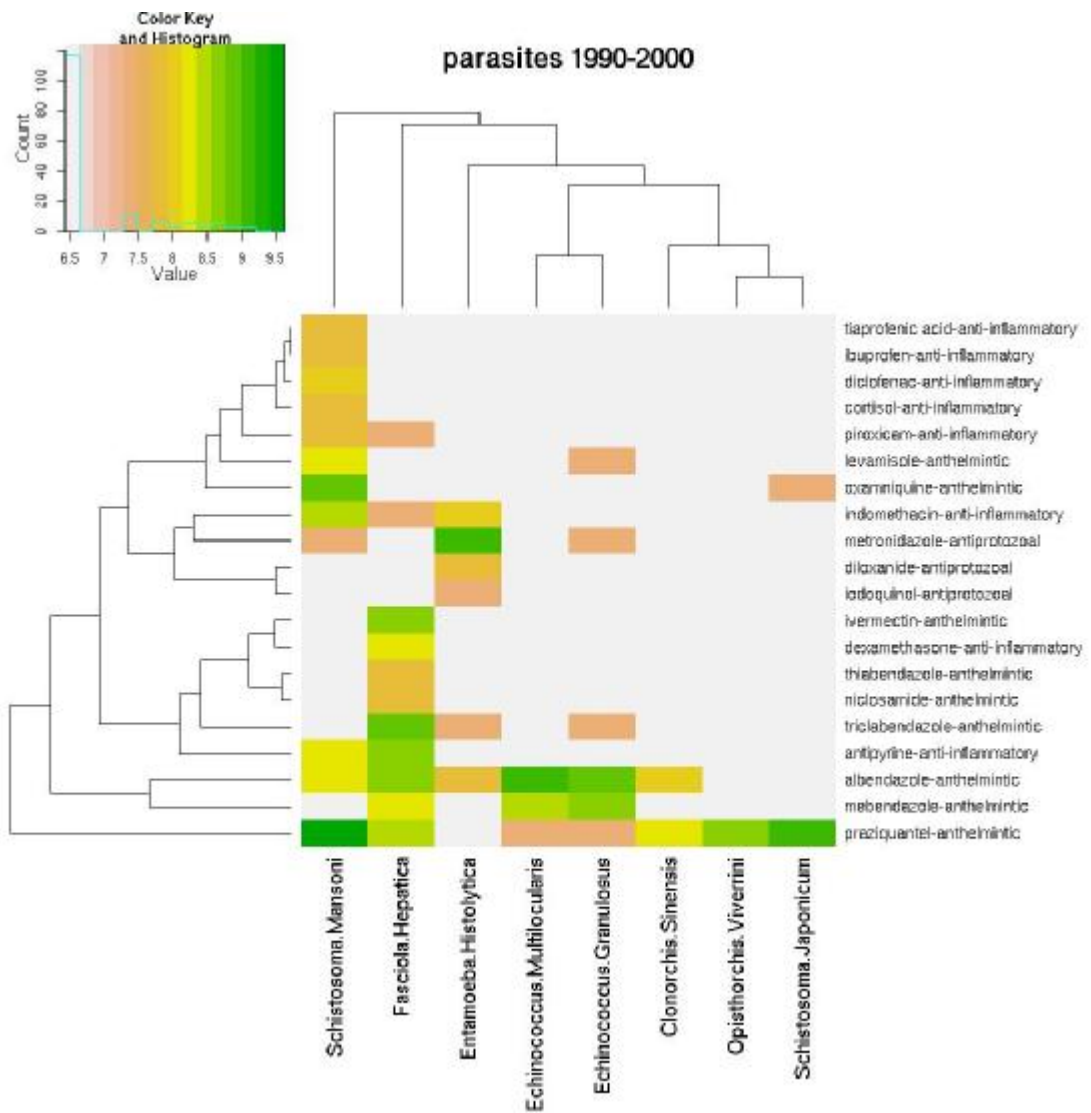


Figure 22: Drug heatmap of parasites for 1990-2000 time period

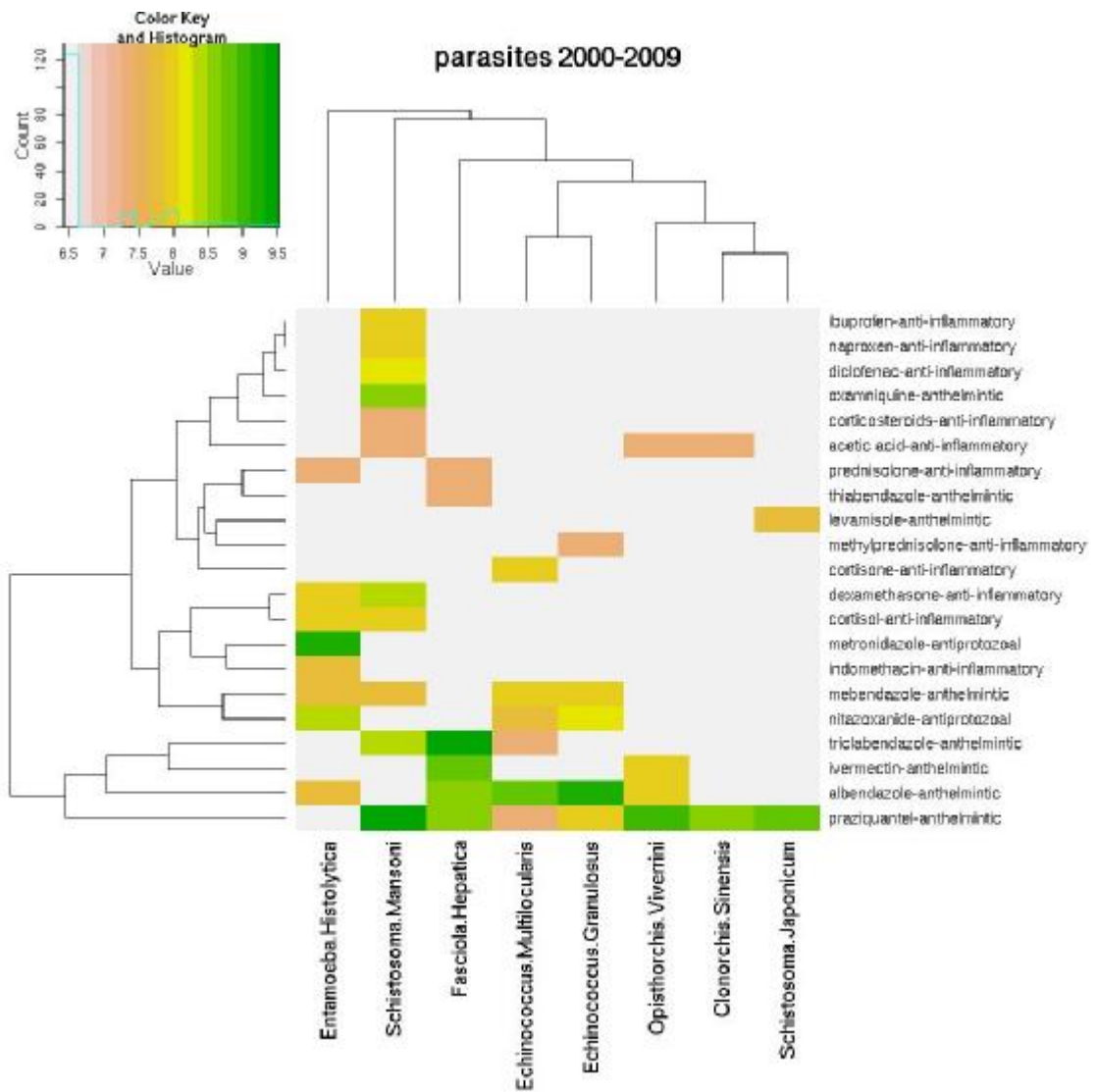


Figure 23: Drug heatmap of parasites for 2000-2009 time period

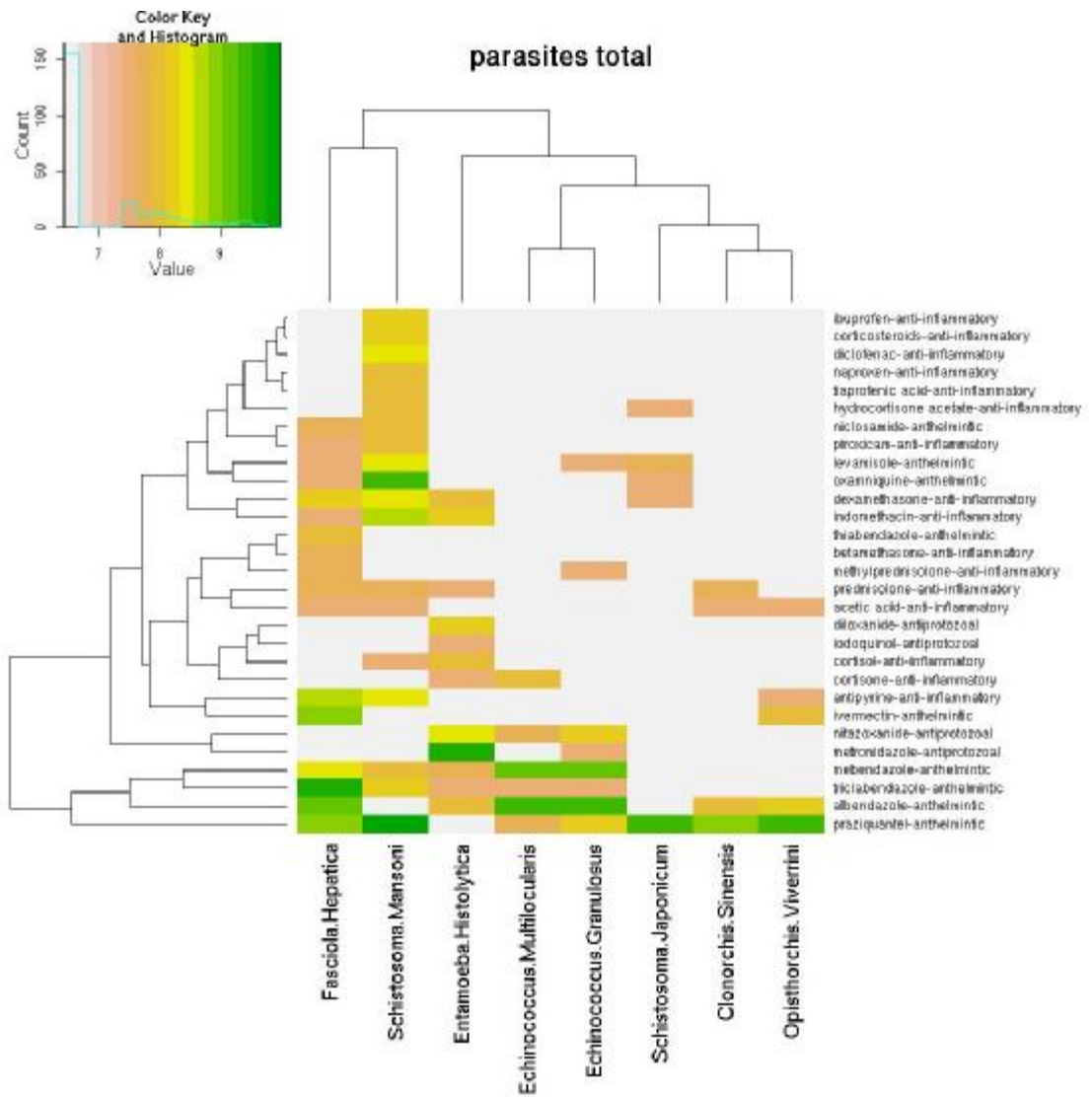


Figure 24: Total drug heatmap of parasites

CHAPTER 5

5 ANALYSIS OF BACTERIA

5.1 Introduction

More than one-third of the world's population is likely infected by bacterial pathogens. Two million fatalities occur each year from bacterial infections. Drug discovery research of infectious diseases, in particular dealing with antibacterial/antibiotic susceptibility and resistance, is in a process of continuing evolution (Monaghan and Barrett 2006). Table 15 shows development of antibiotics over time.

Several drugs are used the treatment of bacteria and drugs have some generic drugs. For example, amikacin, gentamicin, kanamycin, streptomycin and tobramycin are generic drugs. The other group of drugs such as azithromycin, erythromycin, clarithromycin, dirithromycin, roxithromycin and telithromycin are also generic drugs.

5.2 Time Based Analysis of Bacteria

Time based analysis was performed for each bacterium. Main drug class in the treatment of liver specific bacteria was selected to create time series plots (methods were explained in chapter 3).

Table 15: Development of antibiotics over time (Norrby et al. 2005)

Year of Introduction/Discovery	Drug Class
1936	Sulfonamides
1940	β -Lactams
1949	Tetracyclines
1949	Chloramphenicol
1950	Aminoglycosides
1952	Macrolides
1958	Glycopeptides
1962	Streptogramins
1962	Quinolones
1999	Oxazolidinones
2003	Lipopeptides

5.3 Statistical Co-occurrence Based Relation Extraction

Statistical Co-occurrence Based Relation Extraction was applied on bacterium/drugs, bacterium /diseases and bacterium /genes co-occurrence data (methods were explained in chapter 3).

5.4 Salmonella Typhimurium

5.4.1 Drug Time Analysis of Salmonella Typhimurium

Infections with nontyphoidal *Salmonella* have increased during the last 3-4 decades and although a decrease has been reported over the last decade, *Salmonella* infections continue to be a major public health concern in many countries and the resistance to antimicrobial drugs appear to pose a particular health risk. It was found to be resistant to the following five drugs in 1990s: ampicillin, chloramphenicol, streptomycin, sulfonamides and tetracycline ((Helms et al. 2005) and (Chen 2003)). Figure 25 shows a

time series plot of antibacterial drugs for Salmonella Typhimurium. According to the figure, the frequencies of ampicillin and tetracycline have decreased after 1980-1990 time period. It may reveal that the resistance of these drugs were seen in these time periods.

(Threlfall et al. 2006) investigated the changes of antimicrobial resistance in Salmonella Enteritidis and Typhimurium from human infection in England and Wales in 2000, 2002 and 2004 has shown that the incidence of resistance to nalidixic acid coupled with decreased susceptibility to ciprofloxacin has more than doubled between 2000 and 2004. Increase of resistance to nalidixic acid and ciprofloxacin is from 43% in 2000 to 76% in 2004. The occurrence of resistance to ampicillin increased from 5% in 2000 to 8% in 2004, but resistance to tetracyclines and trimethoprim changed very little over the 5-year period (Threlfall et al. 2006).

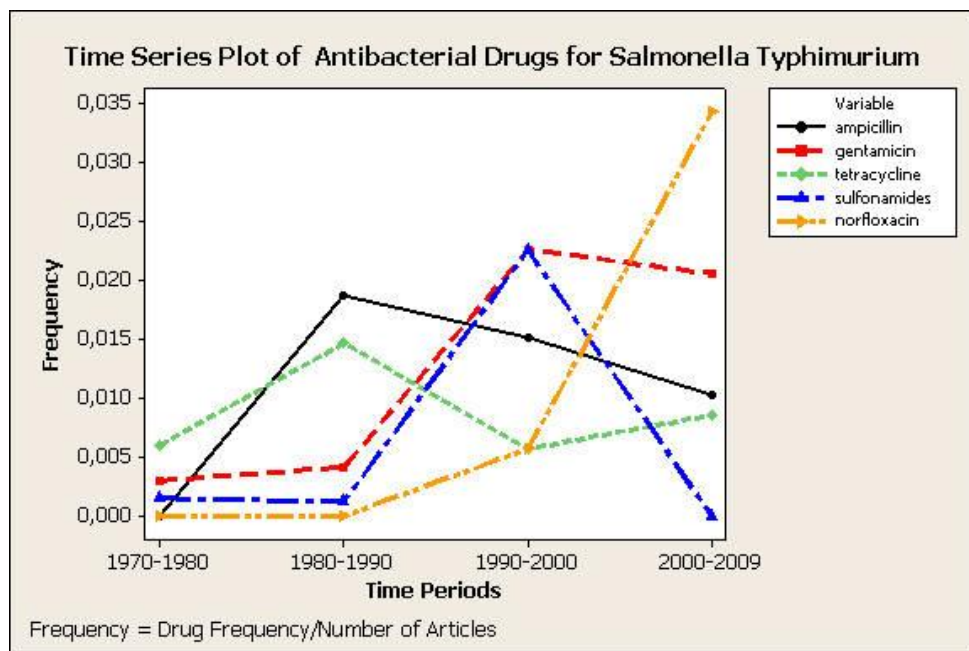


Figure 25: Time series plot of antibacterial drugs for Salmonella Typhimurium

(Chen 2003) analysed the trend of drug resistance of Salmonella typhimurium in Taiwan in 1991-2001. Their study showed that drug resistance ratio for single drug was the highest for streptomycin at 84.2%, followed by tetracycline at 82.5%, chloramphenicol

at 71.9%, ampicillin at 70.2%, and nalidixic acid at 18.4%. Changes in drug resistance of *Salmonella typhimurium* in the recent years were studied by using new generation antibiotics such as ciprofloxacin and ceftiraxone. According to results, ciprofloxacin of the floronquinolones group was 3.8% drug resistance.

Norfloxacin has proved to be a very broad spectrum anti-bacterial drug and it is one of the new 4-quinolone anti-bacterial agents. It was introduced in 1984 in the world market. Subsequent to norfloxacin, four more new quinolones compounds also have come into the market. Still norfloxacin and its successor's ciprofloxacin are able to hold their own market in their clinical use as popular agents for urinary tract infections¹⁵.

5.4.2 Relation Extraction of *Salmonella Typhimurium*

Appendix B Table 40 shows *Salmonella Typhimurium*/drugs co-occurrences based on the Drugbank filter server and Appendix B Figure 50 shows a time series plot for most ranked drugs in both co-occurrence and PMI list. Appendix B Table 41 shows *Salmonella Typhimurium*/diseases co-occurrences based on the UMLS disease filter and Appendix B Table 42 shows *Salmonella Typhimurium*/genes co-occurrences based on the Swissprot filter server.

5.5 *Staphylococcus Aureus*

5.5.1 Drug time analysis of *Staphylococcus Aureus*

The first cases of methicillin-resistant *Staphylococcus Aureus* (MRSA) infections were reported from the UK in 1961 and MRSA became a major problem in hospital settings worldwide in the 1980s. Although community-acquired MRSA are emerging worldwide, vancomycin-resistant *Staphylococcus Aureus* remain extremely rare. Up to 2007, three vancomycin-resistant *Staphylococcus Aureus* were reported from the US in 2002 and 2004. Linezolid is one of the new active agents and it is active against MRSA (Michel and Gutmann 1997) and (Nordmann et al. 2007). Figure 26 shows time series plot of antibacterial drugs for *Staphylococcus Aureus*. Despite their resistance, the frequencies

¹⁵ <http://www.dsir.gov.in/reports/techreps/tsr114.pdf>

of both methicillin and vancomycin have increased in all the time periods. Furthermore, linezolid has been seen after 1990-2000 time period and it has increasing frequency.

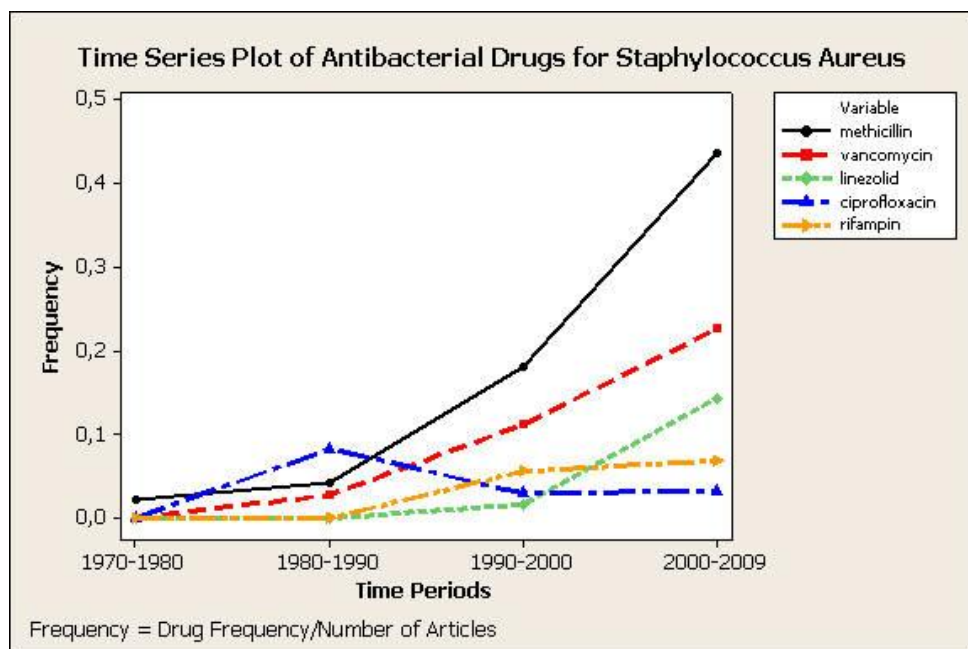


Figure 26: Time series plot of antibacterial drugs for Staphylococcus Aureus

5.5.2 Relation Extraction of Staphylococcus Aureus

Appendix B Table 43 shows Staphylococcus Aureus/drugs co-occurrences based on the Drugbank filter server and Appendix B Figure 51 shows a time series plot for most ranked drugs in both co-occurrence and PMI list. Appendix B Table 44 shows Staphylococcus Aureus/diseases co-occurrences based on the UMLS disease filter server and Appendix B Table 45 shows Staphylococcus Aureus/genes co-occurrences based on the Swissprot filter server.

5.6 Helicobacter Pylori

5.6.1 Drug Time Analysis of Helicobacter Pylori

The discovery that Helicobacter Pylori infection is the main cause of most gastroduodenal diseases and has been a major breakthrough in gastroenterology. It has

dramatically changed the management of these diseases which are now considered as infectious diseases and are treated with antibiotics (Megraud 2004).

Triple therapy, including two antibiotics, amoxicillin and clarithromycin, and a proton pump inhibitor given for a week has been recommended as the treatment of choice at several consensus conferences. However, this treatment may fail for several reasons, as reported elsewhere. In fact, the main reason for failure was found to be *Helicobacter Pylori* resistance to one of the antibiotics used (that is, clarithromycin). Other treatments have also been proposed, including metronidazole, a fluoroquinolones, and rifamycins for which resistance has become an emerging issue (Megraud 2004).

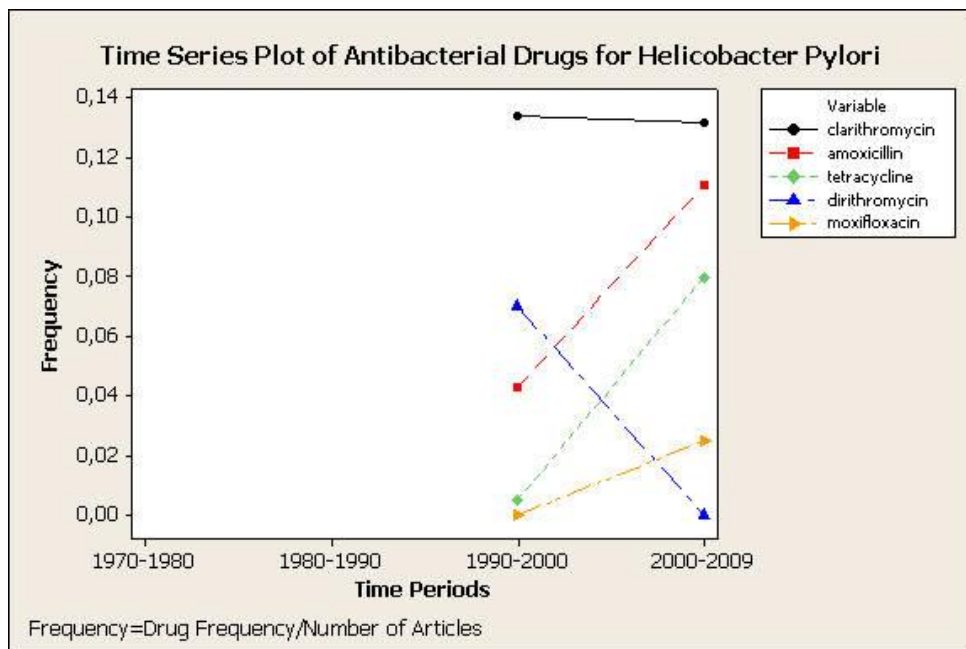


Figure 27: Time series plot of antibacterial drugs for Helicobacter Pylori

Figure 27 shows time series plot of antibacterial drugs for helicobacter pylori. Except clarithromycin and dirithromycin, the other drugs have increasing frequencies after 1990-2000 time periods. In the figure, dirithromycin has been seen only 1990s. It is a macrolide like the standard macrolide erythromycin, as well as clarithromycin and azithromycin. Some studies showed that it may not offer any unique clinical advantage over clarithromycin or azithromycin (Wintermeyer, AbdelRahman and Nahata 1996).

5.6.2 Relation Extraction of Helicobacter Pylori

Appendix B Table 46 shows Helicobacter Pylori /drugs co-occurrences based on the Drugbank filter server and Appendix B Figure 52 shows time series plot for most ranked drugs in both co-occurrence and PMI list. Appendix B Table 47 shows Helicobacter Pylori /diseases co-occurrences based on the UMLS disease filter server and Appendix B Table 48 shows Helicobacter Pylori /genes co-occurrences based on the Swissprot filter server.

5.7 Mycobacterium Tuberculosis

5.7.1 Drug Time Analysis of Mycobacterium Tuberculosis

Tuberculosis (TB) is caused by Mycobacterium Tuberculosis and kills approximately 2 million people each year. Due to the intrinsic resistance of M. tuberculosis to many antibiotics, chemotherapy of TB is restricted to a very limited number of drugs, which have to be used in combination for at least 6 months (Danilchanka, Mailaender and Niederweis 2008).

Multidrug-resistant (MDR) tuberculosis (TB) is a form of TB that is resistant to some of the first-line drugs used for the treatment of the disease. It is associated both with a higher incidence of treatment failures and of disease recurrence, as well as with higher mortality than forms of TB sensitive to first-line drugs. Levofloxacin (LFX) represents one of the few second-line drugs recently introduced in the therapeutic regimens for MDR TB (Richeldi et al. 2002).

Rifampicin (RFP) was developed as one of the anti-tuberculosis drugs in 1966 and has been used for almost 30 years. Establishment of combination therapy using RFP has been contributing to the treatment/eradication of tuberculosis. A number of rifamycin derivatives, as post RFPs, have been synthesized/developed over the the years. Chemical modification of rifamycins has largely been concentrated on the moiety of naphthalene ring because modification of the ansa chain moiety reduces the activity. In 1992, rifabutin was approved as a preventive drug for MAC infection in AIDS patients in the United States and in European countries (Hidaka 1999).

(Hsueh et al. 2006) summarized data from 1990-2002 in Taiwan and results showed that primary resistance ranged from 4.7 to 12% for isoniazid, 0.7 to 5.9% for rifampin, 1 to 6% for ethambutol, and 4 to 11% for streptomycin.

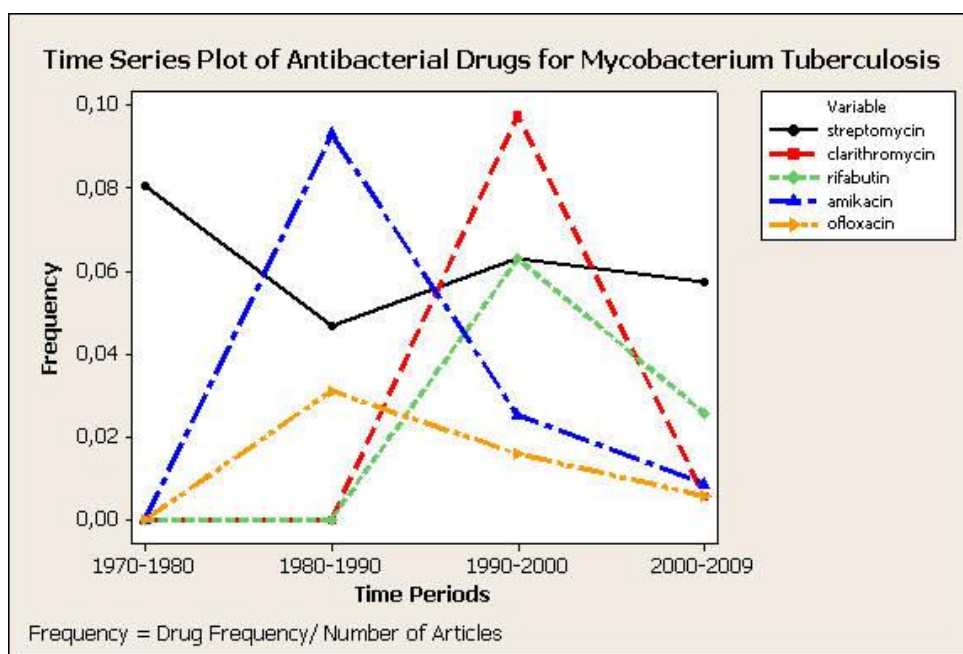


Figure 28: Time series plot of antibacterial drugs for Mycobacterium Tuberculosis

Figure 28 shows time series plot of antibacterial drugs for mycobacterium tuberculosis. According to the figure, in both 1980-1990 and 1990-2000 time periods, significant changes have been seen for some drugs such as clarithromycin, amikacin, rifabutin. In addition, streptomycin and ofloxacin have also had both increases and decreases in these time periods but these changes are not as sharp as the others are.

5.7.2 Relation Extraction of Mycobacterium Tuberculosis

Appendix B Table 49 shows Mycobacterium Tuberculosis/drugs co-occurrences based on the Drugbank filter server and Appendix B Figure 53 shows time series plot for most ranked drugs in both co-occurrence and PMI list. Appendix B Table 50 shows Mycobacterium Tuberculosis/diseases co-occurrences and Appendix B Table 51 shows Mycobacterium tuberculosis/genes co-occurrences based on the Swissprot filter server.

5.8 Listeria Monocytogenes

5.8.1 Drug Time Analysis of Listeria Monocytogenes

Listeria infections are associated with a high mortality rate, and thus effective antibiotic treatment is essential. Although a variety of antibiotics have activity against the organism, ampicillin alone or in combination with gentamicin remains the treatment of choice (Temple and Nahata 2000). The first strains of Listeria Monocytogenes resistant to antibiotics were reported in 1988 (Poros-Gluchowska and Markiewicz 2003).

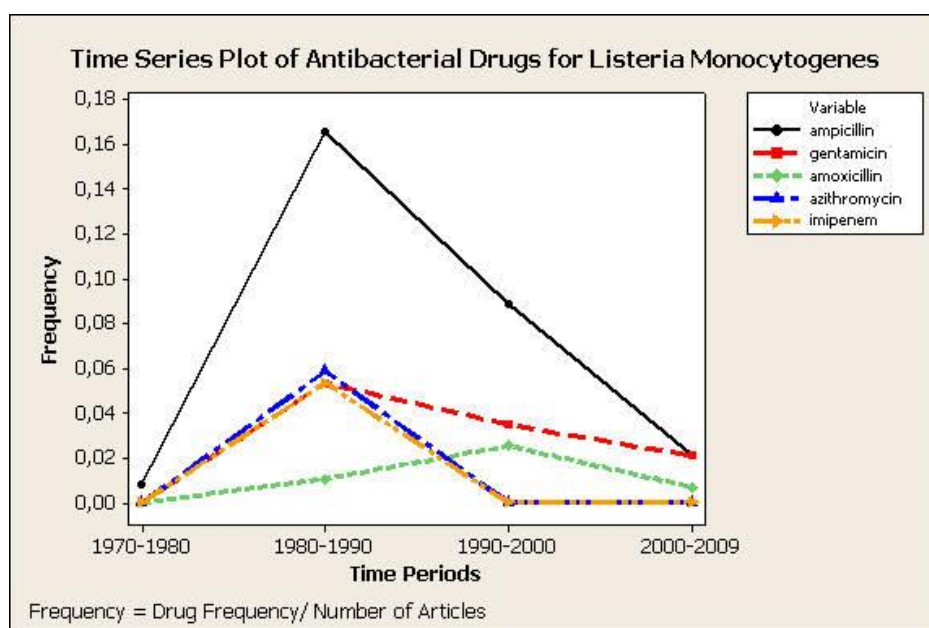


Figure 29: Time series plot of antibacterial drugs for Listeria Monocytogenes

Figure 29 shows time series plot of antibacterial drugs for Listeria Monocytogenes. In the figure, after 1980-1990, the frequencies of both ampicillin and gentamicin have significantly decreased. This can indicate the resistance of these drugs.

5.8.2 Relation Extraction of Listeria Monocytogenes

Appendix B Table 52 shows Listeria Monocytogenes/drugs co-occurrences based on the Drugbank filter server and Appendix B Figure 54 shows time series plot for most ranked drugs in both co-occurrence and PMI list. Appendix B Table 53 shows Listeria Monocytogenes/diseases co-occurrences based on the UMLS disease filter server and

Appendix B Table 54 shows *Listeria Monocytogenes*/genes co-occurrences based on the Swissprot filter server.

5.9 Klebsiella Pneumonia

5.9.1 Drug Time Analysis of Klebsiella Pneumonia

The prevalence of infections caused by *Klebsiella Pneumonia* approaches 50% in some countries, with particularly high rates in Eastern Europe and Latin America. The treatment of these infections is difficult because the organisms are frequently resistant to multiple antibiotics (Paterson et al. 2004).

(Paterson, et al. 2004) study, they found that 47.2% were resistant to piperacillin-tazobactam, 70.8% were resistant to gentamicin, and 19.4% were resistant to ciprofloxacin.

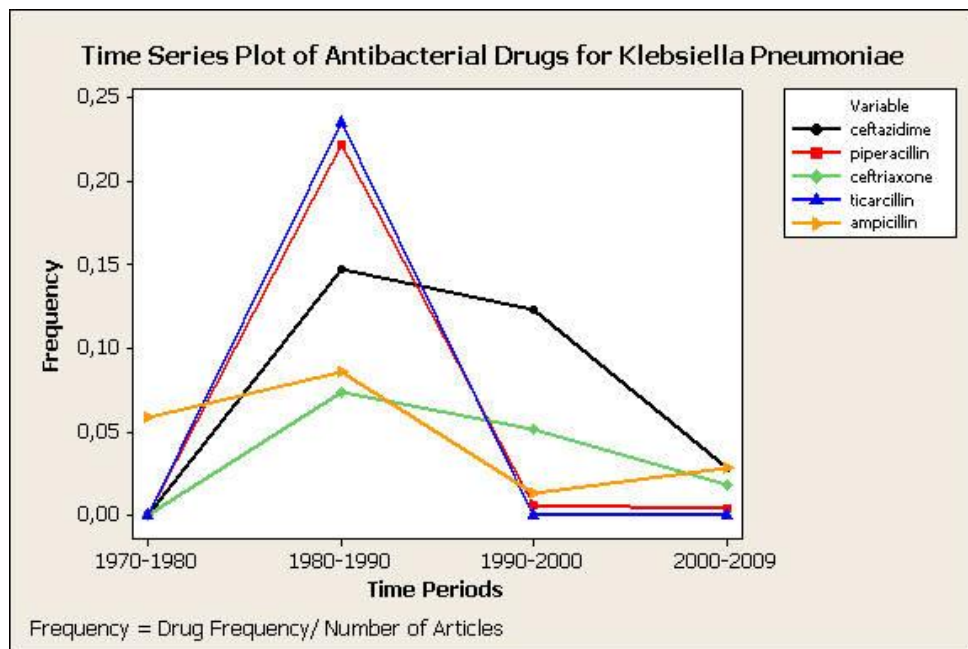


Figure 30: Time series plot of antibacterial drugs for *Klebsiella Pneumoniae*

To examine temporal trends in ceftazidime resistance, susceptibility data reported to the National Nosocomial Infections Surveillance system during 1987-1991 in the USA were

analyzed among nosocomial *Enterobacter* species, *Klebsiella Pneumoniae* and *Pseudomonas Aeruginosa*. Linear increases in resistance were observed for *Enterobacter* species and *Klebsiella Pneumoniae* (Burwen, Banerjee and Gaynes 1994) (Burwen et al. 1994).

Figure 30 shows time series plot of antibacterial drugs for *Klebsiella Pneumoniae*. According to figure, after 1980-1990 time period, big decreases have been seen in the frequencies of some drugs. This can indicate the beginning of drug resistance.

5.9.2 Relation Extraction of *Klebsiella Pneumoniae*

Appendix B Table 55 shows *Klebsiella Pneumoniae*/drugs co-occurrences based on the Drugbank filter server and Appendix B Figure 55 shows time series plot for most ranked drugs in both co-occurrence and PMI list. Appendix B Table 56 shows *Klebsiella Pneumoniae*/diseases co-occurrences based on the UMLS disease filter server and Appendix B Table 57 shows *Klebsiella Pneumoniae*/genes co-occurrences based on the Swissprot filter server.

5.10 *Pseudomonas Aeruginosa*

5.10.1 Drug Time Analysis of *Pseudomonas Aeruginosa*

Pseudomonas Aeruginosa is a serious and life-threatening infection. Reported mortality rates vary significantly from 20%-70% depending on patient-and infection-related factors (Zelenitsky et al. 2003).

It has been the generally accepted practice to treat *Pseudomonas* bacteraemia with the combination of an antipseudomonal penicillin, plus an aminoglycoside. Until 1976, the aminoglycosides, gentamicin, tobramycin, and the carboxypenicillin carbenicilli, were the only systemic antibiotics suitable for the treatment of infections. It is known that the organism can acquire or develop resistance is not however well defined and it may be that it is extensively reported rather than widespread. Additionally there may be considerable geographic variation in the incidence of resistance (Williams et al. 1984).

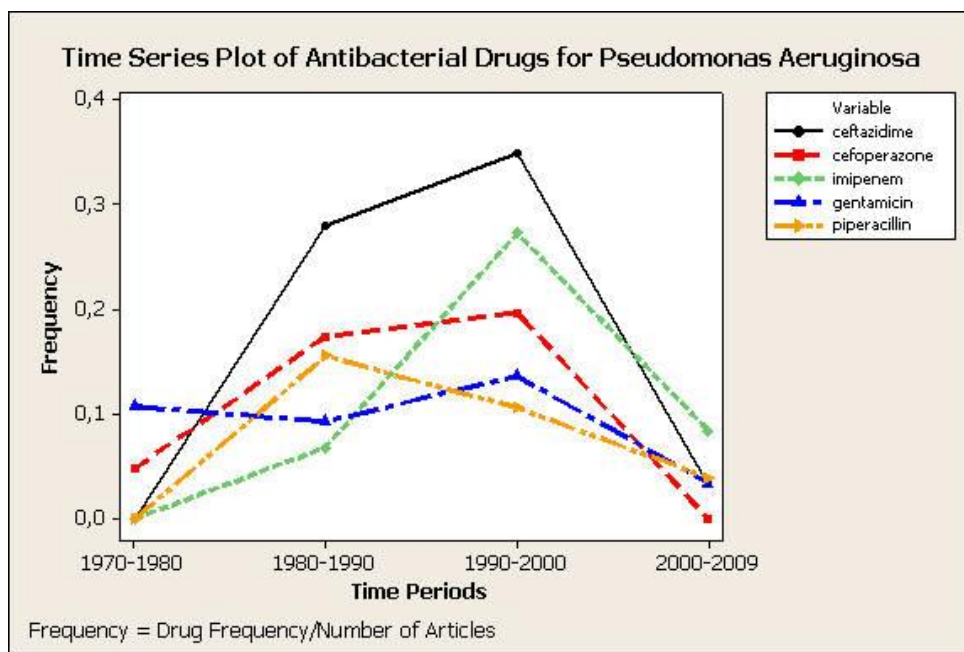


Figure 31: Time series plot of antibacterial drugs for Pseudomonas Aeruginosa

(Bert and LambertZechovsky 1997) investigated the susceptibility to some agents of Pseudomonas Aeruginosa during the period from 1989-1996 in a French hospital. Amikacin and ceftazidime were detected as the most frequently active agents. Figure 31 shows a time series plot of antibacterial drugs for Pseudomonas Aeruginosa. According to the figure, the frequency of ceftazidime has increased until 1990-2000. After this period, the sharp decrease has been seen.

Gentamicin was introduced in 1963 and in 1982, (Williams et al. 1984) performed a national survey of antibiotic resistance in Pseudomonas Aeruginosa and they detected %5.5 the resistance frequency of gentamicin. Figure 31 shows the decreases in its frequency after 1970-1980 and 1990-2000.

5.10.2 Relation Extraction of Pseudomonas Aeruginosa

Appendix B Table 58 shows Pseudomonas Aeruginosa/drugs co-occurrences based on the Drugbank filter server and Appendix B Figure 56 shows time series plot for most ranked drugs in both co-occurrence and PMI list. Appendix B Table 59 shows Pseudomonas Aeruginosa/diseases co-occurrences based on the UMLS disease filter

server and Appendix B Table 60 shows *Pseudomonas Aeruginosa*/genes co-occurrences based on the Swissprot filter server.

5.11 Streptococcus Pneumoniae

5.11.1 Drug Time Analysis of Streptococcus Pneumoniae

Streptococcus Pneumoniae causes various human infections such as meningitis, septicemia, otitis media, sinusitis, and pneumonia. Antibiotic resistance has already been reported with increasing frequency worldwide and is spreading (Erdem and Pahsa 2005). Figure 32 shows time series plot of antibacterial drugs for *Streptococcus Pneumoniae*. According to the figure, the frequencies of some drugs such as azithromycin and trovafloxacin have decreased after 1990-2000 time period.

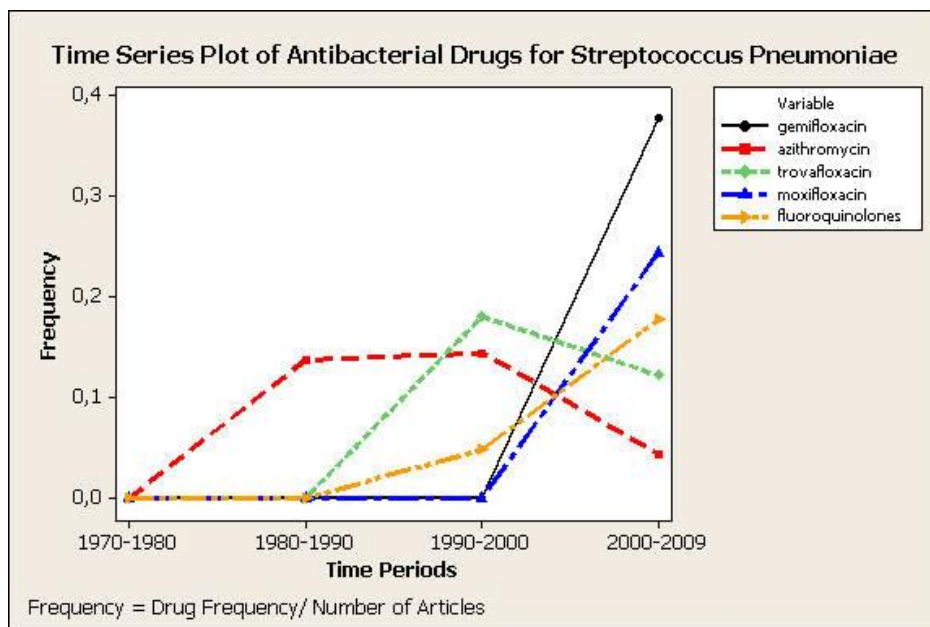


Figure 32: Time series plot of antibacterial drugs for *Streptococcus Pneumoniae*

The fluoroquinolone gemifloxacin has recently been approved for the treatment of acute bacterial exacerbations of chronic bronchitis and mild community acquired pneumonia, including that caused by multidrug-resistant *Streptococcus Pneumoniae* (File and Tillotson 2004). In the figure, it has been seen in 2000-2009 time period and has high frequency. Moxifloxacin is established quinolones, being available as i. v. formulation

since 2002. It has also significant increases after this period (Kruse and Stahlmann 2006).

5.11.2 Relation Extraction of Streptococcus Pneumoniae

Appendix B Table 61 shows Streptococcus Pneumoniae/drugs co-occurrences based on the Drugbank filter server and Appendix B Figure 57 shows a time series plot for most ranked drugs in both co-occurrence and PMI list. Appendix B Table 62 shows Streptococcus Pneumoniae/diseases co-occurrences based on the UMLS disease filter server and Appendix B Table 63 shows Streptococcus pneumoniae/genes co-occurrences based on the Swissprot filter server.

5.12 General Distribution of Drugs in Main Classes

Figure 33, 34, 35 and 36 show distribution of antiinfective, analgesic, antibacterial, and antiinflammatory drugs in all of time periods respectively. These charts provide general view of main classes of drugs mentioned in the articles for liver specific bacteria.

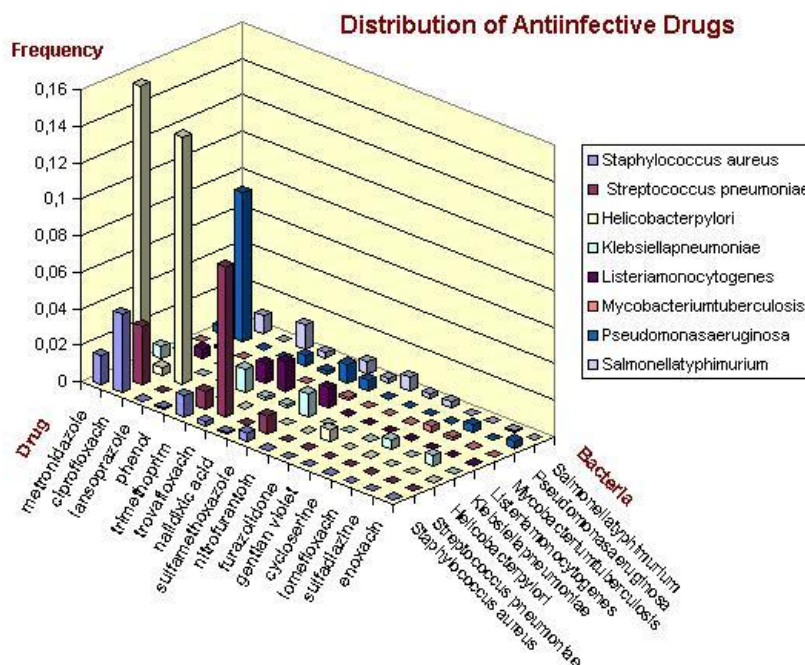


Figure 33: Distribution of anti-infective drugs for bacteria

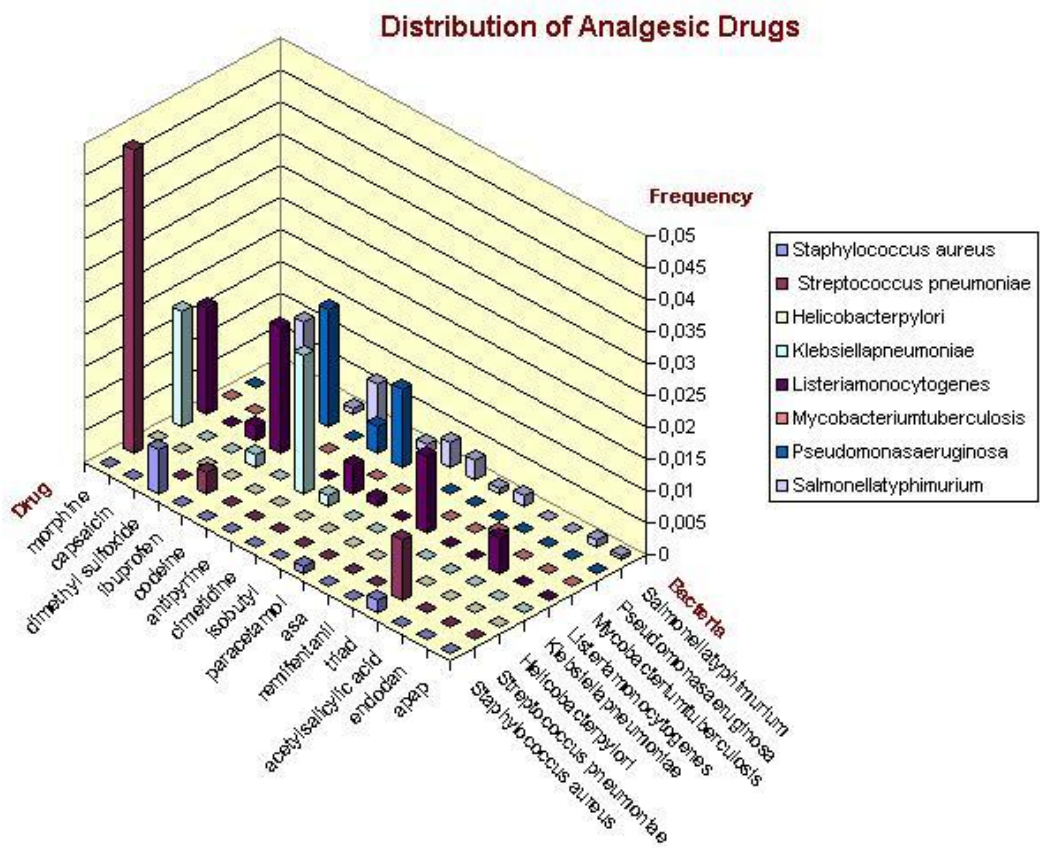


Figure 34: Distribution of analgesic drugs for bacteria

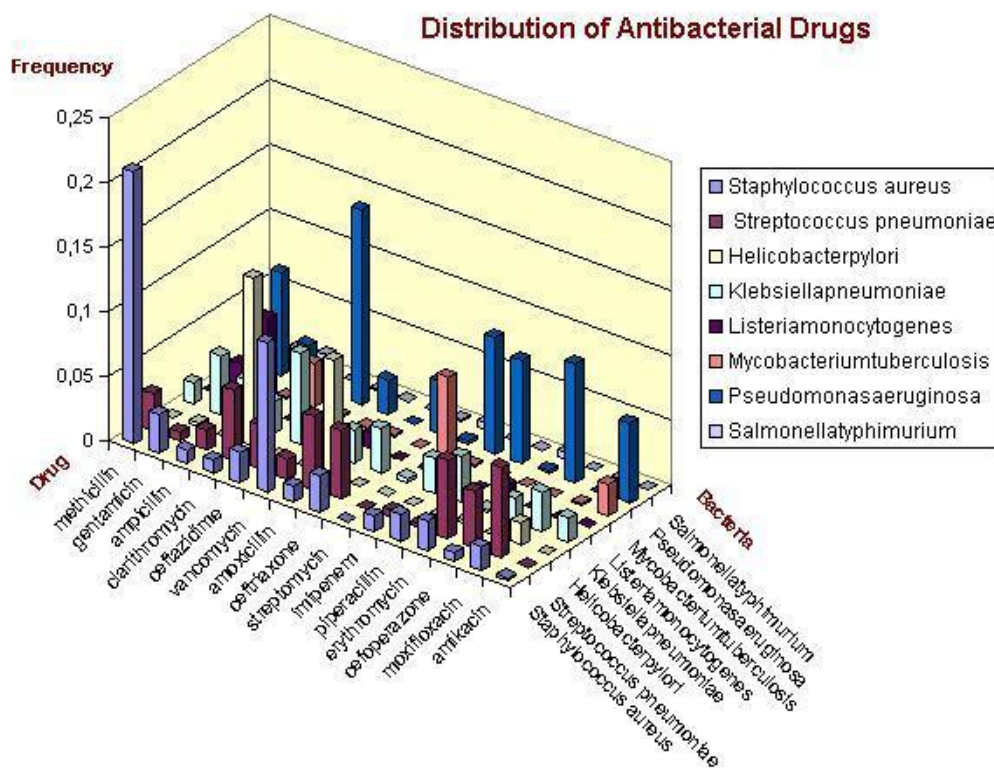


Figure 35: Distribution of antibacterial drugs for bacteria

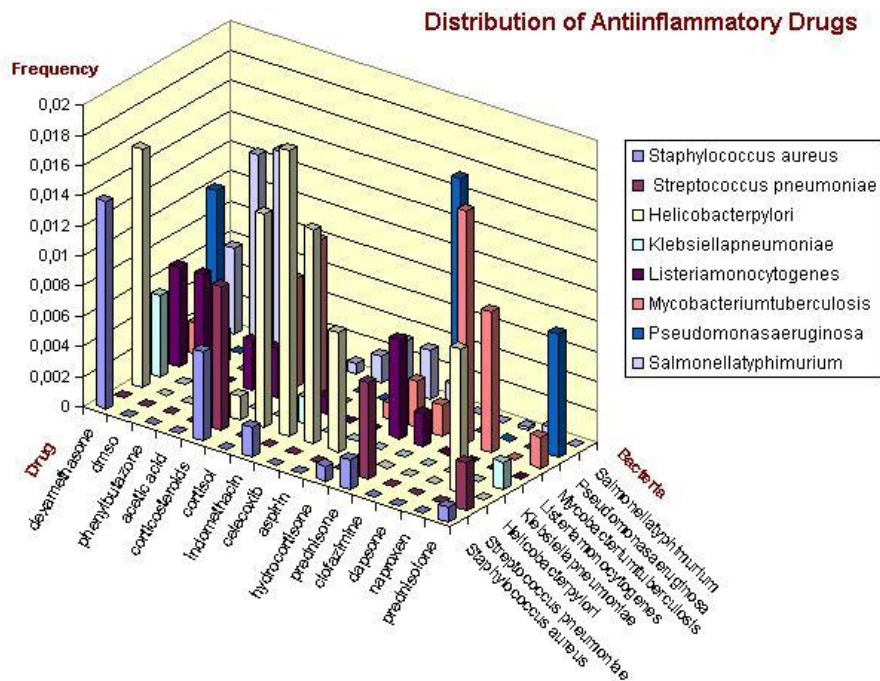


Figure 36: Distribution of anti-inflammatory drugs

5.13 Clustering Analysis of Bacteria

Hierarchical clustering analysis was performed on the basis of time periods for bacteria in the similar way to parasites. Although data set consists of many different classes of drugs, only the most frequent fifteen antibacterial and five antibacterial drugs that are mainly used for treatment of bacteria were selected for clustering analysis. Figure 37, 38, 39, 40 and 41 show heatmaps of clustering in different time periods and combination of these time periods.

The following findings could be identified in the analysis of the clustering results and the interpretation of the heatmaps:

- Anti-bacterial treatment of Helicobacter Pylori did not start before 1990, since bacterial infection was not known to form the cause of ulcers before this time. Thereafter (1990 - 2009), several antibacterial drugs have been used for the treatment of Helicobacter pylori.

- Patients with an infection by Mycobacterium Tuberculosis (M.T.) receive their own kind of treatment, which shifted in 2000 to 2009 to anti-inflammatory drugs treatment. We can expect that the treatment of M.T. infections did not change in the past, but that the treatment made use of anti-inflammatory drugs to improve the side effects induced by the infectious disease.
- Staphylococcus Aureus and Streptococcus Pneumoniae cluster together in our analysis. They share the same profile of drugs in their treatment: mainly penicillin-like antibiotic drugs or replacements of penicillin-like antibiotic drugs with drugs that have stronger antibiotic effects. In the case of Staphylococcus Aureus, the drugs vancomycin and linezolid have been used in recent years. We assume that this shift in this treatment was necessary due to the development of resistencies against penicillin-like drugs in this species. The results for Streptococcus Pneumoniae are not as clear.
- From 2000 to 2009, increasingly anti-inflammatory drugs are used in the treatment of diseases. According to our cluster analysis, it seems to be that selected species show similar profiles in the use of antibacterial versus anti-inflammatory treatments.

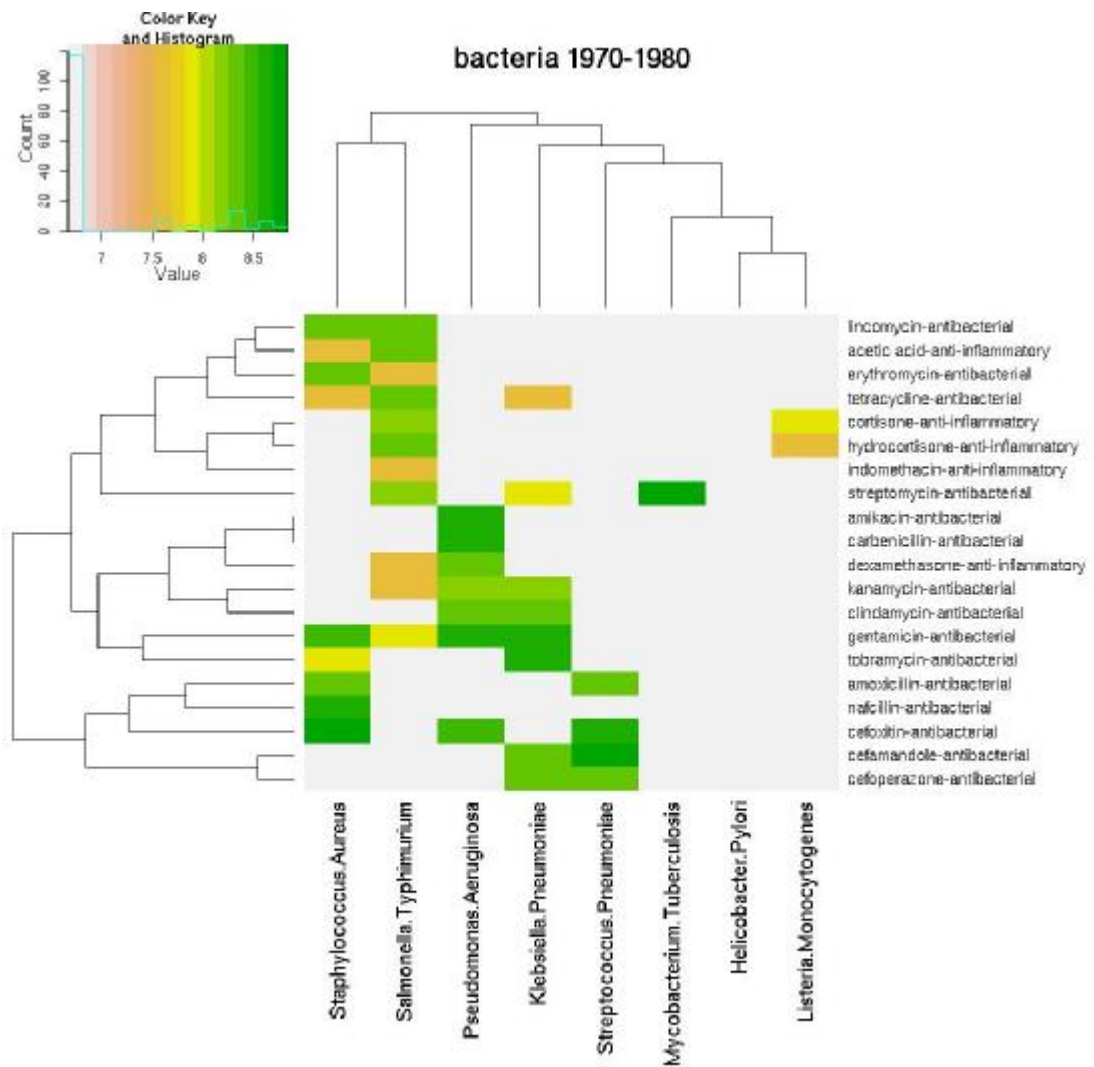


Figure 37: Drug heatmap of bacteria for 1970-1980 time period

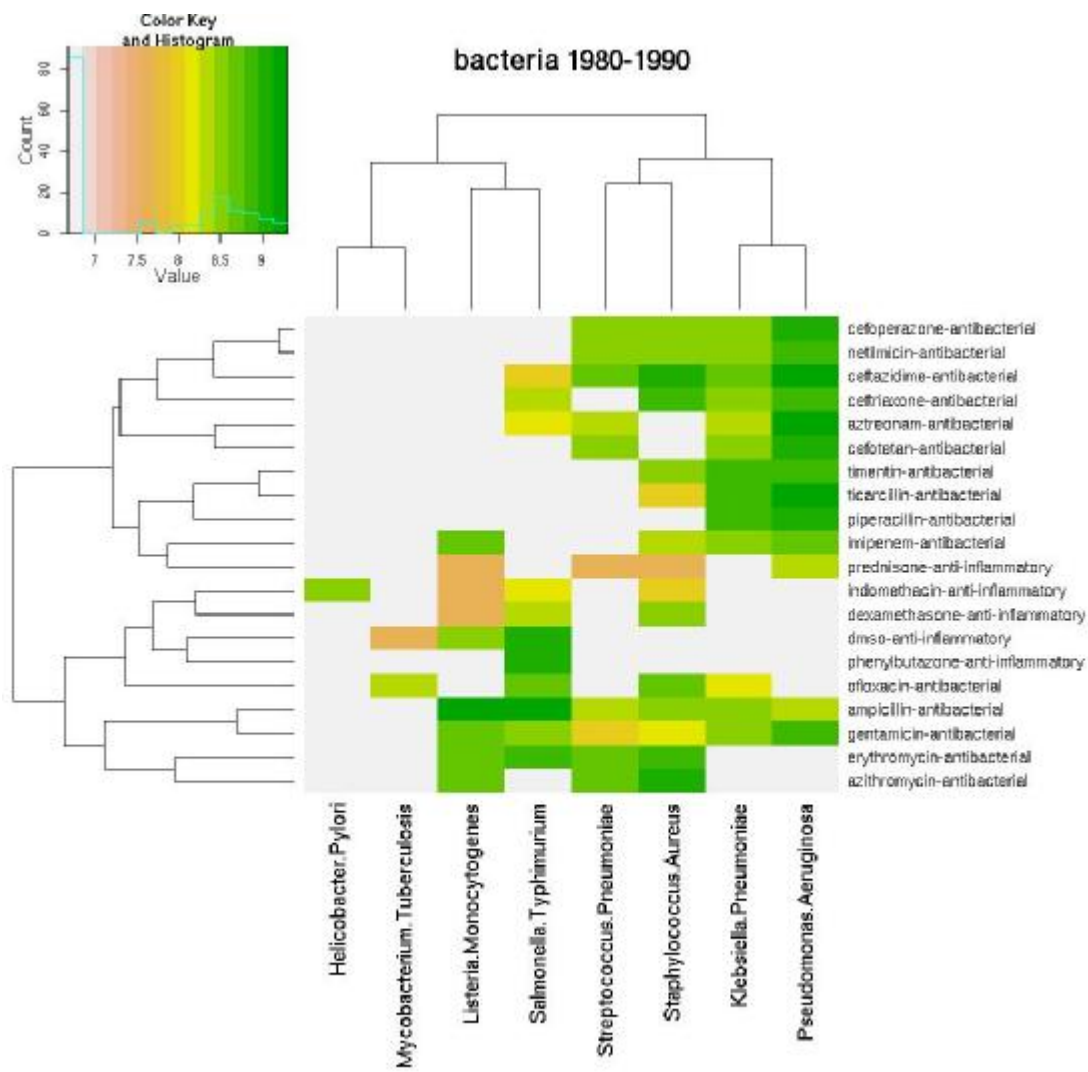


Figure 38: Drug heatmap of bacteria for 1980-1990 time period

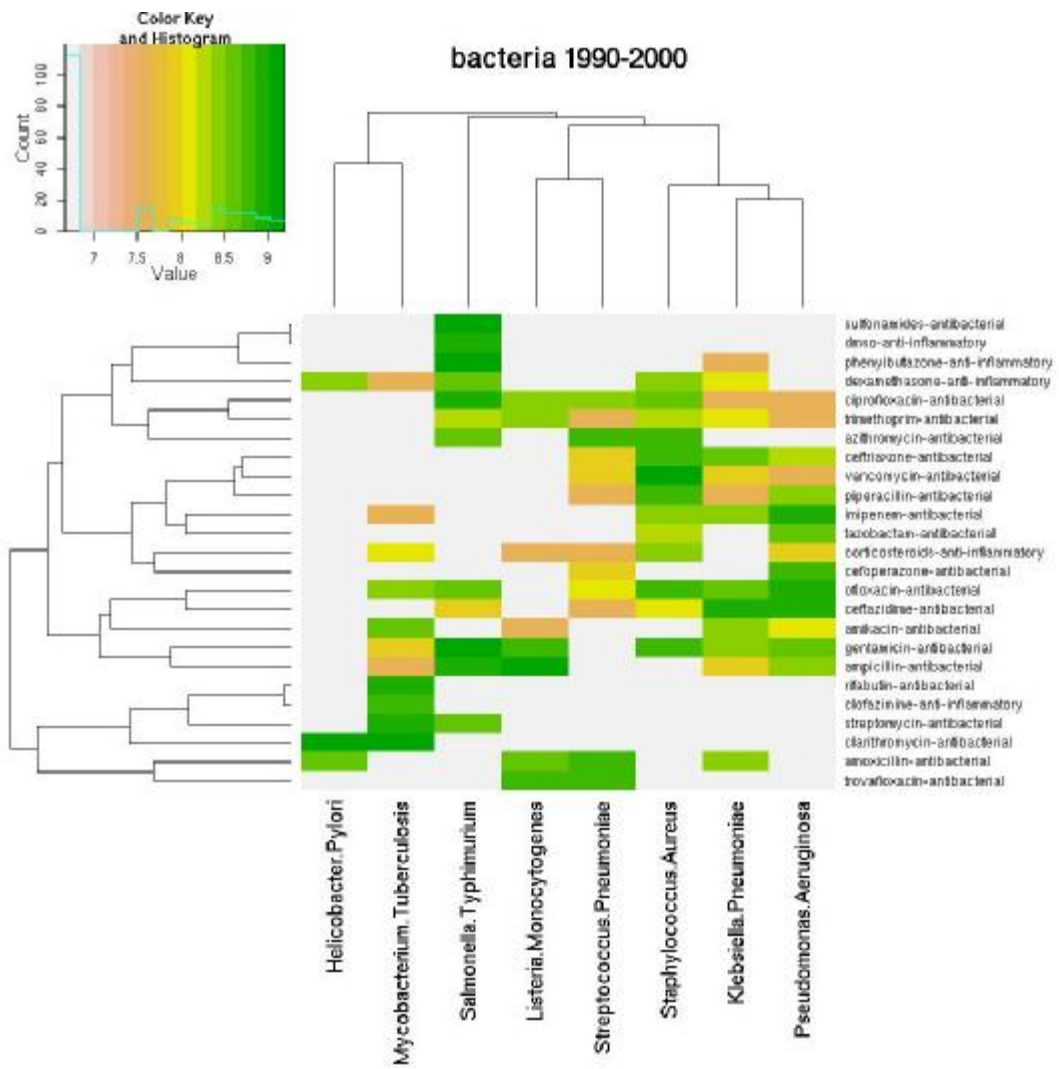


Figure 39: Drug heatmap of bacteria for 1990-2000 time period

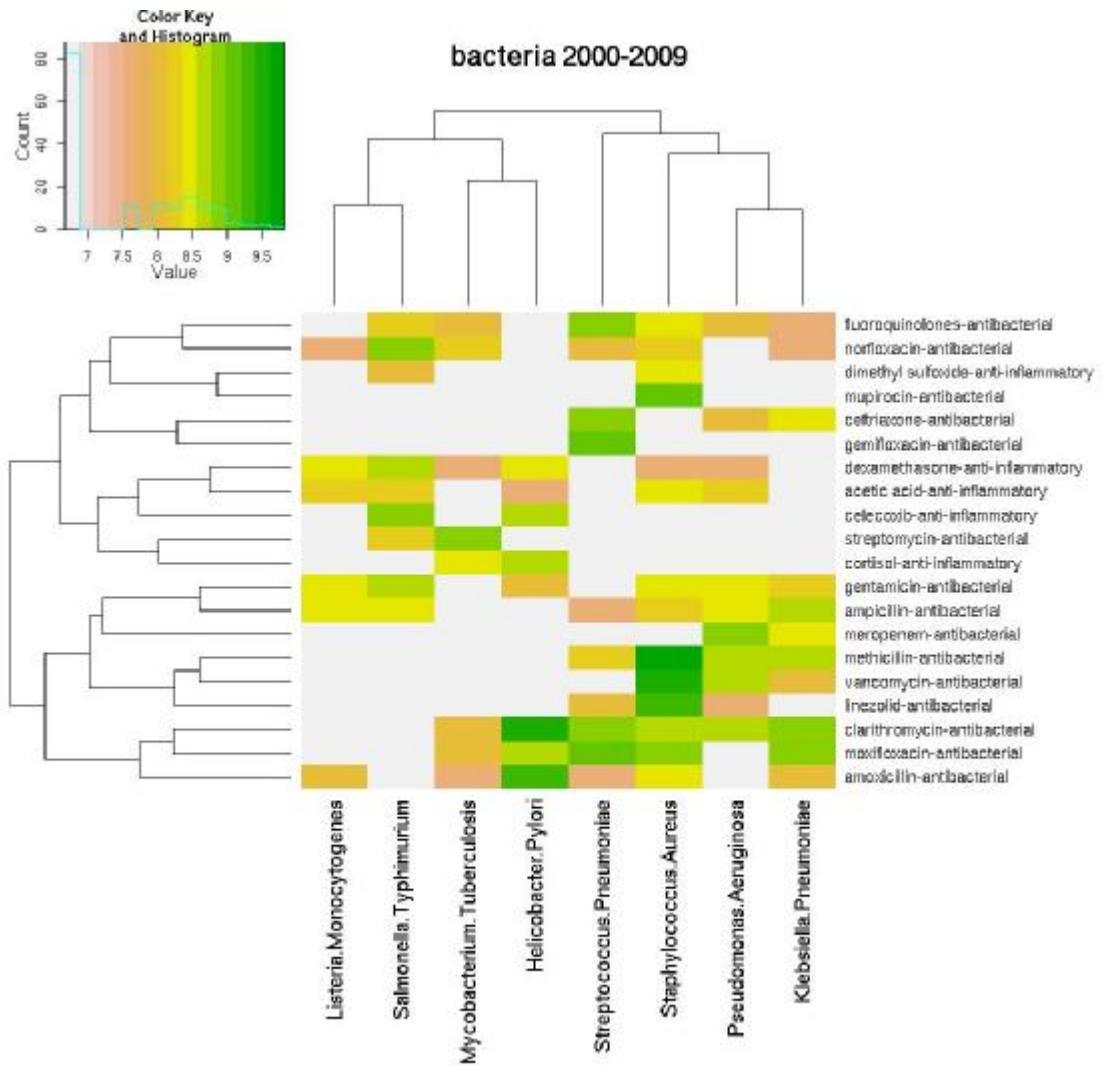


Figure 40: Drug heatmap of bacteria for 2000-2009 time period

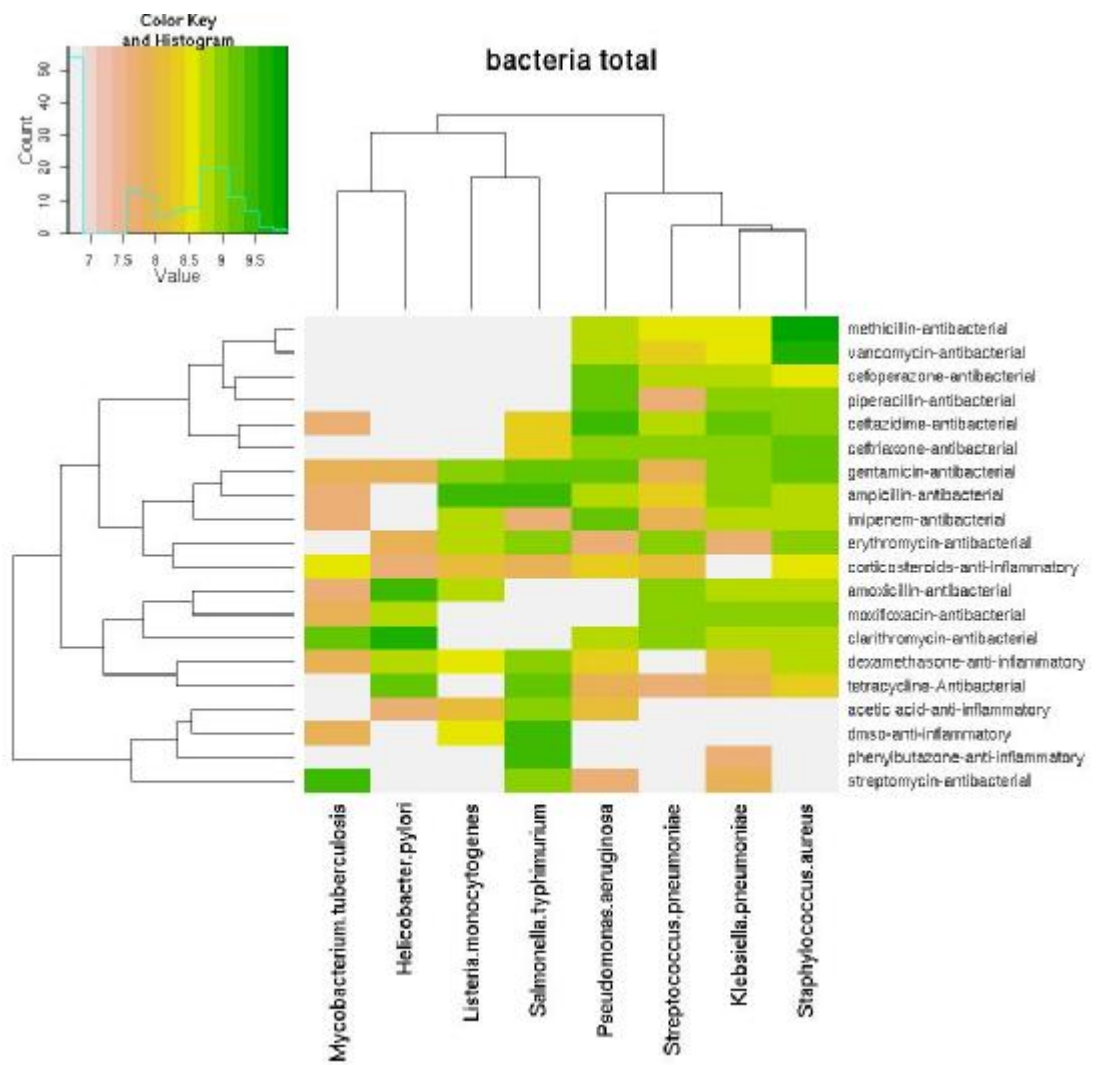


Figure 41: Total drug heatmap of bacteria

CHAPTER 6

6 DISCUSSION AND CONCLUSIONS

6.1 Discussion

Infections with parasites are important causes of morbidity and mortality worldwide (Wood 1996). The control of parasitic disease requires a complex interplay of activities in the fields of public health, education, politics and medical science. Advances in science, especially in the field of parasite genomics and its attended technology, have opened up possibilities for new drugs.

According to our results that we got in this study, there are no big changes between 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. Aside from antimalarial drugs, ivermectin, which was developed as a veterinary product, nitazoxanide which is a new broad-spectrum antiparasitic agent, no major human agent has been introduced to market for decades (White, 2004).

In antiparasitic and antibacterial drug market, following pharmaceutical companies are commonly: Merck&Co, Johnson&Johnson, Pfizer, GlaxoSmithKline, Bristol-Myers Squibb, Aventis, Pharmacia, Novartis, F. Hoffmann-La Roche, AstraZeneca, Abbott Laboratories, Wyeth, Eli Lilly and Co, Schering-Plough, Bayer, Targanta, Forest

Laboratories, Toyoma, Arpida and ActivBiotics (Spellberg et al. 2004) ((Projan and Bradford 2007).

Drug resistance and economic issues are main problems for the fight against parasitic diseases. Sometimes parasites develop resistance to a drug, and combining the drug with an appropriate partner drug may in some cases effectively reduce drug pressure and allow for the reemergence of the susceptible forms of the parasite. For example, chloroquine, an inexpensive, safe and initially highly effective drug was the cornerstone of the effort to eradicate Malaria in 1950s and 1960s. Within 10 years, chloroquine resistance arose and reached high levels in the face of unacceptably high failure rates, and the this drug was switched to more effective drugs (Laufer and Plowe 2004).

In today's global economy, the escalating costs of drug discovery programs make research and development difficult to promote and sustain (Watkins 2003). Many drugs still effective against parasitic diseases are either no longer available, no longer manufactured, or are in danger of being pulled from the market because they are not economically viable. Although parasites cause significant diseases, these diseases are most commonly found in developing countries that do not have the money to effectively support production of these drugs.

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 (Renslo and McKerrow 2006).

The protection of proprietary rights and the recovery of investments are also important issues to 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. A sales decline of over 50% is expected within the first few months

of generic entry. Moreover, unfair competition and counterfeit products are not common (Trouiller and Olliario 1999).

Lastly, regulatory requirements have a considerable impact on the length and costs of the process and hence, on the ultimate market price of the product. Paradoxically, increasingly demanding standards favor the larger wealthy companies, which are those least interested in tropical diseases. Nevertheless, dossiers do not necessarily undergo the same level of review the world over, sometimes because of bare-bones health budgets, and sometimes owing to a misconception of the regulatory process (Trouiller and Olliario 1999).

When meeting the challenge of drug discovery and development for parasitic diseases, how can drugs be discovered, developed and delivered? Several unusual approaches are being undertaken. Funding for target discovery in parasites have been sustained at a basic level by government funding agencies, especially in the United States and Europe. The World Health Organization (WHO) has also been a consistent partner through its Special Programme for Research and Training in Tropical Diseases (TDRs). More recently, numerous major philanthropies have begun to support portions of the drug pipeline. One promising (and in some cases already successful) approach has been to create academic or nonprofit drug development centers that are either staffed by scientists recruited from the drug industry or directly interfaced with industrial partners. This has significantly changed the landscape of antiparasitic drug development, so much that the hit-to-lead and even lead optimization work can conceivably be done outside industry (Renslo and McKerrow 2006).

There are unique aspects of antiparasitic drug discovery with respect to the selection of development candidates as well. Cost of goods is a crucial issue because the drugs will have to be cheap to produce and distribute. The drugs should be orally available given that other routes of administration are problematic in rural settings. With respect to safety, a greater tolerance for adverse effects may be acceptable given the short course of therapy, the seriousness of the condition and the poor safety profiles of current therapies (Renslo and McKerrow 2006).

Similar to parasites, antimicrobial resistance for bacteria is threatening the management of infections such as pneumonia, tuberculosis, malaria and AIDS. During the past 10-15, antibiotic-resistant organisms have steadily increased, and now present a threat to disease management. In the past, resistance could be handled by development of new drugs active against resistant microbes. However, the pharmaceutical industry has reduced its research efforts in infections; genomics has not delivered the anticipated novel therapeutics; new regulatory requirements have increased costs; antibiotic use in common infections e.g., bronchitis and sinusitis is questioned; and, compared with other drugs, return on investments is lower for microbials (Norrby et al. 2005). From both a clinician's and patient's perspective, antibiotic resistance is an issue which continues to pose a significant threat. From the perspective of patients, headlines such as 'The revenge of the killer microbes' only add to their anxiety (Lee 2008).

Anti-infective agents differ from many other drugs in that treatment is normally given for a short time. By contrast, with drugs used to treat hyperlipidaemia, hypertension and diabetes mellitus, for example, treatment of infections is rarely life long. This short treatment period makes anti-infective drugs more susceptible to competition, the return per treatment course is limited, and the need for industry representatives increases. Marketing efforts generally continue during the entire life span of these drugs (Norrby et al. 2005).

Another problem for industry is that if it is able to achieve high sales figures, the result is likely to be more rapid emergence of resistance, which would have an effect on future sales (Norrby et al. 2005).

6.2 Future Work

In our study, apart from other analyses, we applied statistical co-occurrence analysis to extract hidden knowledge in the abstracts and got some relationship between entities such as specific parasite and disease or specific parasite and drug. Medical experts can make connections between seemingly unrelated facts and generate new ideas and hypotheses and discover previously unknown information. Some parasites can cause

specific diseases. For example, *Opisthorchis Viverrini* is reported as a Group I carcinogen, despite its widespread prevalence (Sripa et al. 2007). *Clonorchis Sinensis* has also been shown to be associated with cholangiocarcinoma through geographic and experimental studies (Sithithaworn et al. 1993) but it has not been proved yet. According to our results, new associations can be found based on most frequent diseases related a specific parasite or bacterium. In the similar way, some associations between most frequent drugs-parasite or bacterium, disease-parasite or bacterium and genes-parasite or bacterium can be investigated to discover hidden knowledge. On the other hand, some associations are only statistically meaningful. Medical experts can do clinical studies or research to evaluate such patterns.

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 nowadays and soon they will be made public and shared. As future work, it would be of interest to develop an efficient way to analyze full text papers to compare the results of abstracts (Vlachos 2007).

6.3 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. We introduced a reference model of knowledge processing in biomedical literature for specific medical problems. We worked with medical doctors and we considered their needs, as well as their point of view. Liver specific parasites and bacteria were selected for analysis. The combination of data mining and text mining techniques were used to get some facts hidden in MEDLINE articles. Drugs based on specific time periods, diseases and gene names were extracted from the articles. Named entity recognition and statistical co-occurrence analysis were applied on the data respectively. Some drugs, diseases and genes were found more related to specific parasite and bacterium. Apart from these applications, time based analysis and hierarchical clustering techniques were used for advanced drug analysis. Time based analysis provided the historical evaluation of treatment for clinicians and

pharmaceutical industry. Hierarchical clustering revealed that some treatments of parasites or bacteria are similar and the others are different. We investigated the reasons for the challenge of drug discovery and development for parasitic and bacterial diseases. Drug resistance and economic issues are main problems for the treatment of both parasites and bacteria. Apart from these problems, the protection of proprietary rights and regulatory requirements are other concerns. We believe that our results will make an important contribution to medical research and clinical studies.

REFERENCES

- Bansal, D., N. Malla & R. C. Mahajan (2006) Drug resistance in amoebiasis. *Indian Journal of Medical Research*, 123, 115-118.
- Beckstead, J. W. (2002) Using hierarchical cluster analysis in nursing research. *Western Journal of Nursing Research*, 24, 307-319.
- Bert, F. & N. LambertZechovsky (1997) Antibiotic resistance patterns in *Pseudomonas aeruginosa*: an 8-year surveillance study in a French hospital. *International Journal of Antimicrobial Agents*, 9, 107-112.
- Bodenreider, O., S. Ananiadou & J. Mc Naught. 2006. *Text mining for biology and biomedicine*. Artech House.
- Brennan, G. P., I. Fairweather, A. Trudgett, E. Hoey, McCoy, M. McConville, M. Meaney, M. Robinson, N. McFerran, L. Ryan, C. Lanusse, L. Mottier, L. Alvarez, H. Solana, G. Virkel & P. M. Brophy (2007) Understanding triclabendazole resistance. *Experimental and Molecular Pathology*, 82, 104-109.
- Brunetti, E. (2008) Echinococcus hydatid cyst: treatment & medication. *Medscape's Continually Updated Clinical Reference*.
- Burwen, D. R., S. N. Banerjee & R. P. Gaynes (1994) CEFTAZIDIME RESISTANCE AMONG SELECTED NOSOCOMIAL GRAM-NEGATIVE BACILLI IN THE UNITED-STATES. *Journal of Infectious Diseases*, 170, 1622-1625.
- Caffrey, C. R. (2007) Chemotherapy of schistosomiasis: present and future. *Curr Opin Chem Biol*, 11, 433-9.
- Cao, H., M. Markatou, G. B. Melton, M. F. Chiang & G. Hripcsak (2005) Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics. *AMIA Annu Symp Proc*, 106-10.
- Chen, E. S., G. Hripcsak, H. Xu, M. Markatou & C. Friedman (2008) Automated acquisition of disease-drug knowledge from biomedical and clinical documents: An initial study. *Journal of the American Medical Informatics Association*, 15, 87-98.

- Chen, K. L., Yeh, C. N., Lee, H.C. (2003) Analysis of the trend of drug resistance of *Salmonella typhimurium* in Taiwan, 1991-2001. *Epidemiology Bulletin*, 19, 291-307.
- Cioli, D. & L. Pica-Mattocchia (2003) Praziquantel. *Parasitology Research*, 90, S3-S9.
- Cohen, A. M. & W. R. Hersh (2005) A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6, 57-71.
- Cook, G. C. (1991) ANTHELMINTIC AGENTS - SOME RECENT DEVELOPMENTS AND THEIR CLINICAL-APPLICATION. *Postgraduate Medical Journal*, 67, 16-22.
- Craig, P. (2003) *Echinococcus multilocularis*. *Current Opinion in Infectious Diseases*, 16, 437-444.
- Danilchanka, O., C. Mailaender & M. Niederweis (2008) Identification of a novel multidrug efflux pump of *Mycobacterium tuberculosis*. *Antimicrobial Agents and Chemotherapy*, 52, 2503-2511.
- Davila-Gutierrez, C. E., C. Vasquez, B. Trujillo-Hernandez & M. Huerta (2002) Nitazoxanide compared with quinamide and mebendazole in the treatment of helminthic infections and intestinal protozoa in children. *American Journal of Tropical Medicine and Hygiene*, 66, 251-254.
- de Bruijn, B. & J. Martin. 2002. Getting to the (c)ore of knowledge: mining biomedical literature. In *Symposium on Natural Language Processing for Biomedical Applications*, 7-18. Nicosia, Cyprus.
- Erdem, H. & A. Pahsa (2005) Antibiotic resistance in pathogenic *Streptococcus pneumoniae* isolates in Turkey. *Journal of Chemotherapy*, 17, 25-30.
- Fairweather, I., Boray, J.,C. (1999) ScienceDirect - The Veterinary Journal : Fasciolicides: Efficacy, Actions, Resistance and its Management. *The Veterinary Journal*, 158, 81-112.
- Fallon, P. G., L. F. Tao, M. M. Ismail & J. L. Bennett (1996) Schistosome resistance to praziquantel: Fact or artifact? *Parasitology Today*, 12, 316-320.
- Feldman, R. & J. Sanger. 2007. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press.
- File, T. M., Jr. & G. S. Tillotson (2004) Gemifloxacin: a new, potent fluoroquinolone for the therapy of lower respiratory tract infections. *Expert Rev Anti Infect Ther*, 2, 831-43.

- Geary, T. G. (2005) Ivermectin 20 years on: maturation of a wonder drug. *Trends in Parasitology*, 21, 530-532.
- Gryseels, B., A. Mbaye, S. J. De Vlas, F. F. Stelma, F. Guisse, L. Van Lieshout, D. Faye, M. Diop, A. Ly, L. A. Tchuem-Tchuente, D. Engels & K. Polman. 2000. Are poor responses to praziquantel for the treatment of *Schistosoma mansoni* infections in Senegal due to resistance? An overview of the evidence. In *Colloquium on Moving Target: Parasites, Resistance and Access to Drugs*, 864-873. Antwerp, Belgium.
- Gryseels, B., K. Polman, J. Clerinx & L. Kestens (2006) Human schistosomiasis. *Lancet*, 368, 1106-1118.
- Harder, A. (2002) Chemotherapeutic approaches to trematodes (except schistosomes) and cestodes: current level of knowledge and outlook. *Parasitology Research*, 88, 587-590.
- Haux, R. 1997. Aims and tasks of medical informatics. In *Scientific Colloquium Held on the Occasion of the 60th Birthday of Claus O Kohler*, 9-20. Heidelberg, Germany.
- Helms, M., S. Ethelberg, K. Molbak & D. T. S. Grp (2005) International Salmonella Typhimurium DT104 infections, 1992-2001. *Emerging Infectious Diseases*, 11, 859-867.
- Hidaka, T. (1999) [Current status and perspectives on the development of rifamycin derivative antibiotics]. *Kekkaku*, 74, 53-61.
- Hsueh, P. R., Y. C. Liu, J. So, C. Y. Liu, P. C. Yang & K. T. Luh (2006) Mycobacterium tuberculosis in Taiwan. *J Infect*, 52, 77-85.
- Hu, X. 2006. Semantic Text mining and its application in biomedical domain. Drexel University.
- Jimeno Yepes, A. 2008. Study of named entity recognition in biomedicine: towards the refinement of ontologies.
- Kaewpitoon, N., S. Kaewpitoon & P. Pengsaa (2008) Opisthorchiasis in Thailand: Review and current status. *World Journal of Gastroenterology*, 14, 2297-2302.
- Keiser, J., J. Chollet, S. H. Xiao, J. Y. Mei, P. Y. Jiao, J. Utzinger & M. Tanner (2009) Mefloquine-An Aminoalcohol with Promising Antischistosomal Properties in Mice. *Plos Neglected Tropical Diseases*, 3.

- Kern, P. (2003) Echinococcus granulosus infection: clinical presentation, medical treatment and outcome. *Langenbecks Archives of Surgery*, 388, 413-420.
- Krauthammer, M. & G. Nenadic (2004) Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37, 512-526.
- Kruse, M. & R. Stahlmann (2006) [New antibiotics for the therapy of respiratory tract infections]. *Pneumologie*, 60, 417-27.
- Laufer, M. K. & C. V. Plowe (2004) Withdrawing antimalarial drugs: impact on parasite resistance and implications for malaria treatment policies. *Drug Resistance Updates*, 7, 279-288.
- Lee, C. (2008) Therapeutic challenges in the era of antibiotic resistance. *International Journal of Antimicrobial Agents*, 32, S197-S199.
- Leroy, G., H. Chen & J. D. Martinez (2003) A shallow parser based on closed-class words to capture relations in biomedical text. *J Biomed Inform*, 36, 145-58.
- Liu, X. L. (1996) Antiparasitic Drugs. *The England Journal of Medicine*, 334, 1178-1184.
- M., H. S. 2006. Cluster Analysis. Department of Geology, University of Georgia, Athens, GA 30602-2501.
- Marcos, L. A., A. Terashima & E. Gotuzzo (2008) Update on hepatobiliary flukes: fascioliasis, opisthorchiasis and clonorchiasis. *Current Opinion in Infectious Diseases*, 21, 523-530.
- Megraud, F. (2004) H pylori antibiotic resistance: prevalence, importance, and advances in testing. *Gut*, 53, 1374-84.
- Michel, M. & L. Gutmann (1997) Methicillin-resistant Staphylococcus aureus and vancomycin-resistant enterococci: Therapeutic realities and possibilities. *Lancet*, 349, 1901-1906.
- Monaghan, R. L. & J. F. Barrett (2006) Antibacterial drug discovery - Then, now and the genomics future. *Biochemical Pharmacology*, 71, 901-909.
- Mooney R., J., Nahm U.,Y. 2005. Text mining with information extraction. In *Multilingualism and electronic language management, International MIDP Colloquium*, 141-160.
- Mortele, K. J., E. Segatto & P. R. Ros (2004) The infected liver: Radiologic-pathologic correlation. *Radiographics*, 24, 937-955.

- Nenadic, G., I. Spasic & S. Ananiadou (2003) Terminology-driven mining of biomedical literature. *Bioinformatics*, 19, 938-943.
- Ng, D. C., S. Y. Kwok, Y. Cheng, C. C. Chung & M. K. Li (2006) Colonic amoebic abscess mimicking carcinoma of the colon. *Hong Kong Med J*, 12, 71-3.
- Nordmann, P., T. Naas, N. Fortineau & L. Poirel (2007) Superbugs in the coming new decade; multidrug resistance and prospects for treatment of *Staphylococcus aureus*, *Enterococcus* spp. and *Pseudomonas aeruginosa* in 2010. *Current Opinion in Microbiology*, 10, 436-440.
- Norrby, S. R., C. E. Nord, R. Finch & I. European Soc Clinical Microbiol (2005) Lack of development of new antimicrobial drugs: a potential serious threat to public health. *Lancet Infectious Diseases*, 5, 115-119.
- Ofori-Adjei, D., A. N. O. Dodoo, A. Appiah-Danquah, M. Couper, P. Centre for Tropical Clinical, U. o. G. M. S. A. G. Therapeutics & W. H. O. G. S. Psm (2008) A review of the safety of niclosamide, pyrantel, triclabendazole and oxamniquine. *Risk & Safety in Medicine*, 20.
- Paterson, D. L., W. C. Ko, A. Von Gottberg, S. Mohapatra, J. M. Casellas, H. Goossens, L. Mulazimoglu, G. Trenholme, K. P. Klugman, R. A. Bonomo, L. B. Rice, M. M. Wagener, J. G. McCormack & V. L. Yu (2004) Antibiotic therapy for *Klebsiella pneumoniae* bacteremia: Implications of production of extended-spectrum beta-lactamases. *Clinical Infectious Diseases*, 39, 31-37.
- Petric, I., Urbanic, T., Cestnik, B. (2007) Discovering hidden knowledge from biomedical literature. *Informatica*, 31, 15-20.
- Poros-Gluchowska, J. & Z. Markiewicz (2003) Antimicrobial resistance of *Listeria monocytogenes*. *Acta Microbiol Pol*, 52, 113-29.
- Projan, S. J. & P. A. Bradford (2007) Late stage antibacterial drugs in the clinical pipeline. *Current Opinion in Microbiology*, 10, 441-446.
- Rebholz-Schuhmann, D., M. Arregui, S. Gaudan, H. Kirsch & A. Jimeno (2008) Text processing through web services: calling Whatizit. *Bioinformatics*, 24, 296-298.
- Rebholz-Schuhmann, D., H. Kirsch, M. Arregui, S. Gaudan, M. Riethoven & P. Stoehr (2007) EBIMed - text crunching to gather facts for proteins from MEDLINE. *Bioinformatics*, 23, E237-E244.
- Rebholz-Schuhmann, D., H. Kirsch, S. Gaudan & M. Arregui, Nenadic, G. 2006. Annotation and disambiguation of semantic types in biomedical text: a cascaded approach to named entity recognition. In *Workshop on multidimensional markup with XML (XML NLP), EACL 2006, Trento, Italy*.

- Rebholz-Schuhmann D., A. M., Yepes, A., :J., Kirsch, H., Nenadic, G.,. 2007. Automatic text analysis based on Web services. In *Handout for the ISMB tutorial, Vienna*.
- Redfearn, J., National Text Mining Centre. (2008) Text mining. *JISC theme(s):e-Research*.
- Renslo, A. R. & J. H. McKerrow (2006) Drug discovery and development for neglected parasitic diseases. *Nature Chemical Biology*, 2, 701-710.
- Richeldi, L., M. Covi, G. Ferrara, F. Franco, P. Vailati, E. Meschiari, L. M. Fabbri & G. Velluti (2002) Clinical use of Levofloxacin in the long-term treatment of drug resistant tuberculosis. *Monaldi Arch Chest Dis*, 57, 39-43.
- Rim, H. J. (1984) THERAPY OF FLUKE INFECTIONS IN THE PAST - A REVIEW. *Arzneimittel-Forschung/Drug Research*, 34-2, 1127-1129.
- Rindfleisch, T., Fiszman, M., Libbus, B. 2006. Semantic interpretation for the biomedical research literature. 399-422. Springer.
- Saric, J., H. Engelken & U. Reyle (2008) Discovering biomedical knowledge from the literature. *Methods in molecular biology*, 484, 415-433.
- Scherf, M., A. Epple & T. Werner (2005) The next generation of literature analysis: Integration of genomic analysis into text mining. *Briefings in Bioinformatics*, 6, 287-297.
- Shaikenov, B. S. 2006. Distribution and ecology of *Echinococcus multilocularis* in Central Asia. In *International Symposium on Taeniasis/Cysticercosis and Echinococcosis with Focus on Asia and the Pacific*, S213-S219. Asahikawa, JAPAN.
- Shu-Hua, X. (2005) Development of antischistosomal drugs in China, with particular consideration to praziquantel and the artemisinin. *Acta Tropica*, 96, 135-167.
- Sithithaworn, P., M. R. Haswellelkins, P. Mairiang, S. Satarug, E. Mairiang, V. Vatanasapt & D. B. Elkins. 1993. PARASITE-ASSOCIATED MORBIDITY - LIVER FLUKE INFECTION AND BILE-DUCT CANCER IN NORTHEAST THAILAND. In *1993 Conference of the Australian-Society-for-Parasitology*, 833-843. Heron Isl, Australia.
- Spellberg, B., J. H. Powers, E. P. Brass, L. G. Miller & J. E. Edwards (2004) Trends in antimicrobial drug development: Implications for the future. *Clinical Infectious Diseases*, 38, 1279-1286.

- Sripa, B., S. Kaewkes, P. Sithithaworn, E. Mairiang, T. Laha, M. Smout, C. Pairojkul, V. Bhudhisawasdi, S. Tesana, B. Thinkamrop, J. M. Bethony, A. Loukas & P. J. Brindley (2007) Liver fluke induces cholangiocarcinoma. *Plos Medicine*, 4, 1148-1155.
- Stamatakis, M., C. Sargedi, C. Stefanaki, C. Safioleas, I. Matthaipoulou & M. Safioleas (2009) Anthelmintic treatment: An adjuvant therapeutic strategy against *Echinococcus granulosus*. *Parasitology International*, 58, 115-120.
- Teggi, A., M. G. Lastilla & F. Derosa (1993) THERAPY OF HUMAN HYDATID-DISEASE WITH MEBENDAZOLE AND ALBENDAZOLE. *Antimicrobial Agents and Chemotherapy*, 37, 1679-1684.
- Temple, M. E. & M. C. Nahata (2000) Treatment of listeriosis. *Annals of Pharmacotherapy*, 34, 656-661.
- Threlfall, E. J., M. Day, E. de Pinna, A. Charlett & K. L. Goodyear (2006) Assessment of factors contributing to changes in the incidence of antimicrobial drug resistance in *Salmonella enterica* serotypes Enteritidis and Typhimurium from humans in England and Wales in 2000, 2002 and 2004. *International Journal of Antimicrobial Agents*, 28, 389-395.
- Trouiller, P. & P. L. Olliaro (1999) Drug development output: what proportion for tropical diseases? *Lancet*, 354, 164-164.
- Tsuruoka, Y., J. Tsujii & S. Ananiadou (2008) FACTA: a text search engine for finding associated biomedical concepts. *Bioinformatics*, 24, 2559-2560.
- Uramoto, N., H. Matsuzawa, T. Nagano, A. Murakami, H. Takeuchi & K. Takeda (2004) A text-mining system for knowledge discovery from Biomedical Documents. *Ibm Systems Journal*, 43, 516-533.
- Vipin, K. 2006. Introduction to Data Mining. Addison-Wesley.
- Vlachos, A. 2007. Evaluating and combining biomedical named entity recognition systems. In *In Poster Proceedings of BioNLP at ACL, Prague*.
- Watkins, B. M. (2003) Drugs for the control of parasitic diseases: current status and development. *Trends in Parasitology*, 19, 477-478.
- White, C. A., Jr. (2004) Nitazoxanide: a new broad spectrum antiparasitic agent. *Expert Rev Anti Infect Ther*, 2, 43-9.
- Williams, R. J., M. A. Lindridge, A. A. Said, D. M. Livermore & J. D. Williams (1984) NATIONAL SURVEY OF ANTIBIOTIC-RESISTANCE IN *PSEUDOMONAS-AERUGINOSA*. *Journal of Antimicrobial Chemotherapy*, 14, 9-16.

- Wintermeyer, S. M., S. M. AbdelRahman & M. C. Nahata (1996) Dirithromycin: A new macrolide. *Annals of Pharmacotherapy*, 30, 1141-1149.
- Wishart, D. S., C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang & J. Woolsey (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, 34, D668-D672.
- Wood, A. J. J. (1996) Drug therapy. *The England Journal of Medicine.*, 1178-1184.
- Zelenitsky, S. A., G. K. M. Harding, S. Sun, K. Ubhi & R. E. Ariano (2003) Treatment and outcome of *Pseudomonas aeruginosa* bacteraemia: an antibiotic pharmacodynamic analysis. *Journal of Antimicrobial Chemotherapy*, 52, 668-674.
- Zhou, W., N. R. Smalheiser & C. Yu (2006) A tutorial on information retrieval: basic terms and concepts. *J Biomed Discov Collab*, 1, 2.

7 APPENDICES

8 A. RELATION EXTRACTION FOR PARASITES

Table 16: Fasciola hepatica/drugs co-occurrences based on the Drugbank filter server

Rank	Drug	Co-occurrence Frequency	PMI
1	triclabendazole	148	13.367
2	albendazole	47	10.951
3	ivermectin	35	12.346
4	praziquantel	26	10.033
5	antipyrene	19	8.116
6	levorphanol	11	9.894
7	IL-2	11	6.280
8	fructose	9	5.042
9	vitamin e	8	5.095
10	caffeine	7	7.166

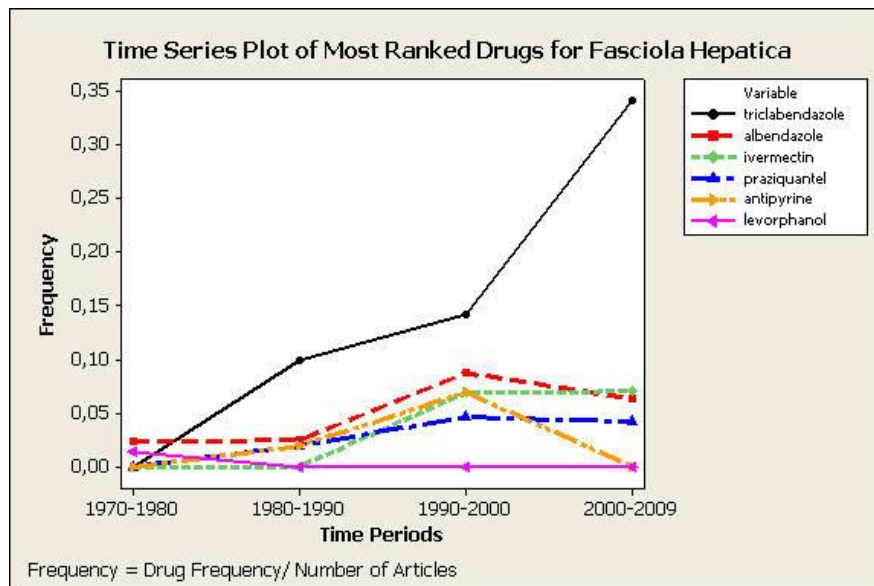


Figure 42: Time series plot of most ranked drugs for Fasciola Hepatica

Table 17: Fasciola Hepatica/diseases co-occurrences based on the UMLS disease filter server

Rank	Disease	Co-occurrence Frequency	PMI
1	cholangiocarcinoma	24	6.943
2	mass	17	2.984
3	cirrhosis	15	1.857
4	echinococcus	12	7.847
5	hepatitis	12	1.825
6	tumor	12	0.549
7	schistosomiasis	10	5.772
8	cholelithiasis	10	5.245
9	abscess	10	4.466
10	mono	9	6.311

Table 18: Fasciola Hepatica/genes co-occurrences based on the Swissprot filter server

Rank	Gene	Co-occurrence Frequency	PMI
1	rats	271	4.653
2	had	160	2.107
3	rat	108	3.258
4	gst	39	8.006
5	adenylate cyclase	38	8.742
6	cathepsin l	37	11.557
7	phospho fructokinase	35	9.250
8	pain	33	6.207
9	proteases	32	5.279
10	who	30	3.831

Table 19: Schistosoma Mansoni/drugs co-occurrences based on Drugbank filter server

Rank	Drug	Co-occurrence Frequency	PMI
1	praziquantel	374	12.963
2	IL-2	138	9.0128
3	oxamniquine	118	11.302
4	interferon	52	5.5805
5	interferon gamma	36	6.5778
6	colchicine	33	8.1394
7	deet	25	8.6021
8	cyclosporine	20	5.5706
9	indomethacin	19	7.3264
10	polymyxin b	14	9.8205

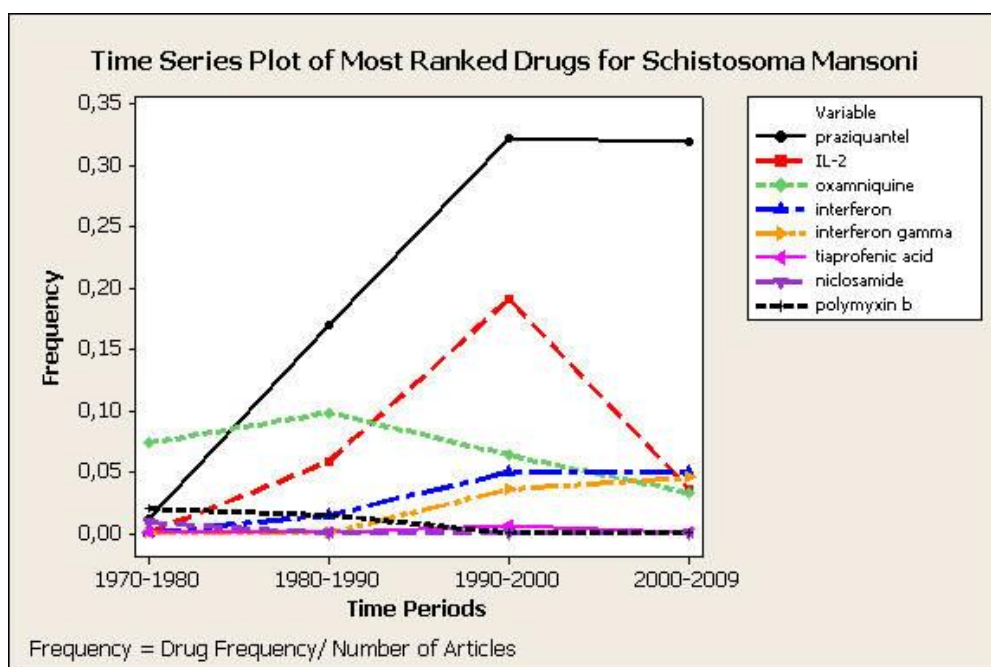


Figure 43: Time series plot of most ranked drugs for Schistosoma Mansoni

Table 20: Schistosoma Mansoni/diseases co-occurrences based on the UMLS disease filter server

Rank	Diseases	Co-occurrence Frequency	PMI
1	schistosomiasis	862	11.285
2	hypertension	90	5.885
3	hepatitis c	61	5.085
4	liver disease	50	0.652
5	adhesion	43	6.428
6	hepatitis	39	2.608
7	hepatitis b	35	4.126
8	malaria	33	7.562
9	tumor	29	0.905
10	mass	26	2.681

Table 21: Schistosoma Mansoni/genes co-occurrences based on the Swissprot filter server

Rank	Gene	Co-occurrence Frequency	PMI
1	had	268	1.935
2	il-4	227	10.836
3	ifn-gamma	176	9.388
4	sea	155	10.095
5	th1	112	10.284
6	cd4	112	8.269
7	il-10	107	9.372
8	il-5	96	11.992
9	il-2	83	8.270
10	il-13	82	12.267

Table 22: Schistosoma Japonicum/drugs co-occurrences based on the Drugbank filter server

Rank	Drug	Co-occurrence Frequency	PMI
1	praziquantel	96	12.958
2	IL-2	44	9.320
3	colchicine	22	9.511
4	vitamin-a	19	7.281
5	interferon	14	5.644
6	sma	9	8.854
7	rosiglitazone	9	9.791
8	interferon-gamma	7	6.172
9	hyaluronic acid	6	7.818
10	heparin	4	5.515

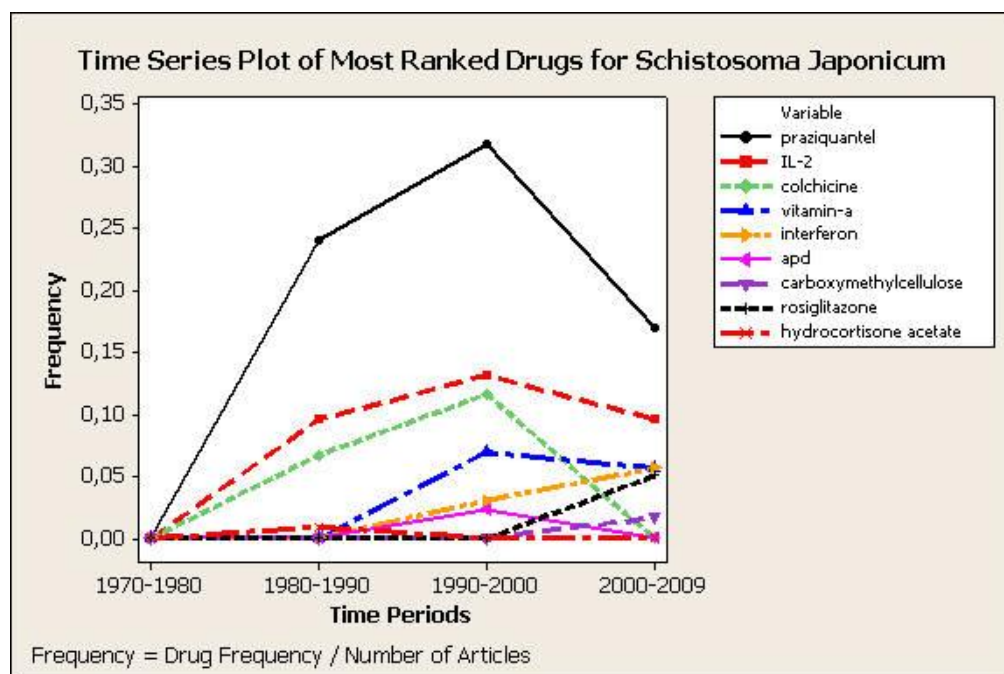


Figure 44: Time series plot of most ranked drugs for Schistosoma Japonicum

Table 23: Schistosoma Japonicum/diseases co-occurrences based on the UMLS disease filter server

Rank	Disease	Co-occurrence Frequency	PMI
1	schistosomiasis	313	11.810
2	liver cancer	24	17.101
3	hypertension	24	13.951
4	cirrhosis	18	16.311
5	tumor	15	17.297
6	arf	14	8.409
7	liver disease	13	18.336
8	cancer	13	17.248
9	hepatocellular carcinoma	13	15.507
10	mass	13	15.364

Table 24: Schistosoma Japonicum/genes co-occurrences based on the Swissprot server

Rank	Gene	Co-occurrence Frequency	PMI
1	pigs	168	8.347
2	had	94	2.380
3	ifn-gamma	58	9.743
4	sea	53	10.503
5	il-4	52	10.667
6	sj23	47	17.855
7	il-2	36	9.021
8	art	35	11.301
9	gst	32	8.760
10	pcr	31	5.692

Table 25: Entamoeba Histolytica/drugs co-occurrences based on the Drugbank filter server

Rank	Drug	Co-occurrence Frequency	PMI
1	metronidazole	172	11.721
2	luminal	18	4.706
3	chloroquine	17	7.996
4	interferon	17	4.655
5	niacin	12	8.343
6	IL-2	10	5.913
7	nitazoxanide	9	9.467
8	beta2	9	8.635
9	indomethacin	7	6.573
10	histamine	7	5.924

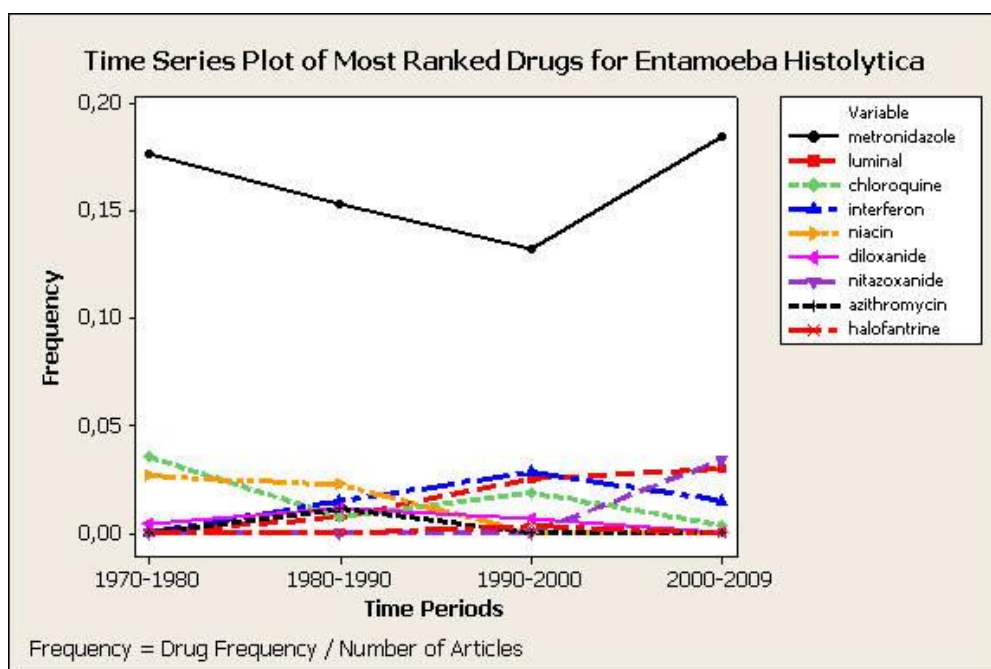


Figure 45: Time series plot of most ranked drugs for Entamoeba Histolytica

Table 26: Entamoeba Histolytica/diseases co-occurrences based on the UMLS disease server

Rank	Disease	Co-occurrence Frequency	PMI
1	liver abscess	516	13.111
2	amebiasis	435	11.689
3	amebic liver abscess	376	11.586
4	abscess	281	13.117
5	colitis	134	10.983
6	hepatic amebiasis	69	11.592
7	adhesion	40	12.342
8	intestinal amebiasis	39	9.416
9	amebic dysentery	36	9.157
10	hepatitis	22	16.021

Table 27: Entamoeba Histolytica/genes co-occurrences based on the Swissprot filter server

Rank	Gene	Co-occurrence Frequency	PMI
1	lectin	242	9.683
2	had	227	2.382
3	pcr	138	6.577
4	gal	92	10.412
5	who	83	5.069
6	srehp	70	16.161
7	pain	56	6.740
8	cysteine proteinase	31	7.494
9	pigs	28	4.493
10	cysteine proteinases	27	7.374

Table 28: Echinococcus Granulosus/drugs co-occurrences based on the Drugbank filter server

Rank	Drugs	Co-occurrence Frequency	PMI
1	mebendazole	106	13.321
2	albendazole	60	13.538
3	nitazoxanide	7	10.530
4	tranexamic acid	7	4.951
5	praziquantel	6	9.115
6	triclabendazole	4	7.926
7	diphtheria	4	6.018
8	PAS	4	3.994
9	levamisole	3	5.106
10	tuberculin	2	8.480

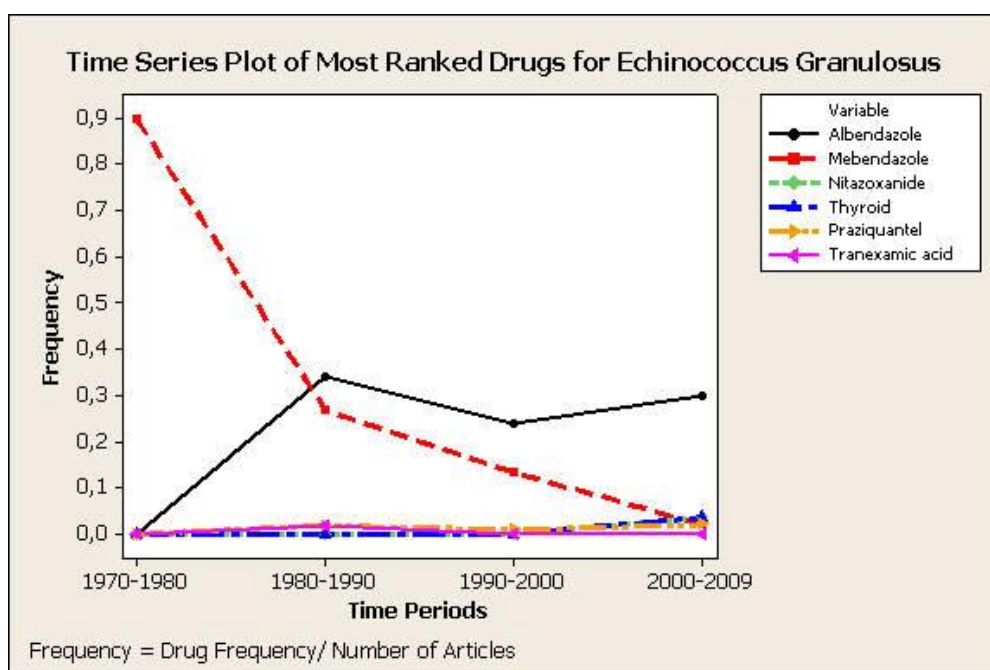


Figure 46: Time series plot of most ranked drugs for Echinococcus Granulosus

Table 29: Echinococcus Granulosus/diseases co-occurrences based on the UMLS disease filter server

Rank	Disease	Co-occurrence Frequency	PMI
1	echinococcus	518	9.999
2	hydatidosis	156	11.972
3	mass	32	15.364
4	tapeworm	30	12.106
5	liver disease	14	18.336
6	tapeworms	11	9.662
7	anaphylaxis	11	9.031
8	tumor	7	17.297
9	pneumonia	7	11.541
10	chf	7	7.392

Table 30: Echinococcus Granulosus/genes co-occurrences based on the Swissprot filter server

Rank	Genes	Co-occurrence Frequency	PMI
1	had	145	3.162
2	who	37	5.330
3	pain	31	7.313
4	hooks	22	14.010
5	pigs	22	5.571
6	gal	20	9.636
7	hcf	18	16.306
8	pcr	18	5.064
9	she	14	7.380
10	opn	13	12.699

Table 31: Echinococcus Multilocularis/drugs co-occurrences based on the Drugbank filter server

Rank	Drug	Co-occurrence Frequency	PMI
1	albendazole	85	13.807
2	mebendazole	47	13.990
3	interferon	10	6.120
4	pas	8	9.731
5	IL-2	7	7.630
7	cimetidine	4	8.447
6	interferon gamma	4	6.326
7	cortisone	3	7.162
8	doxorubicin	3	5.916
9	nitazoxanide	2	9.528
10	praziquantel	2	8.335

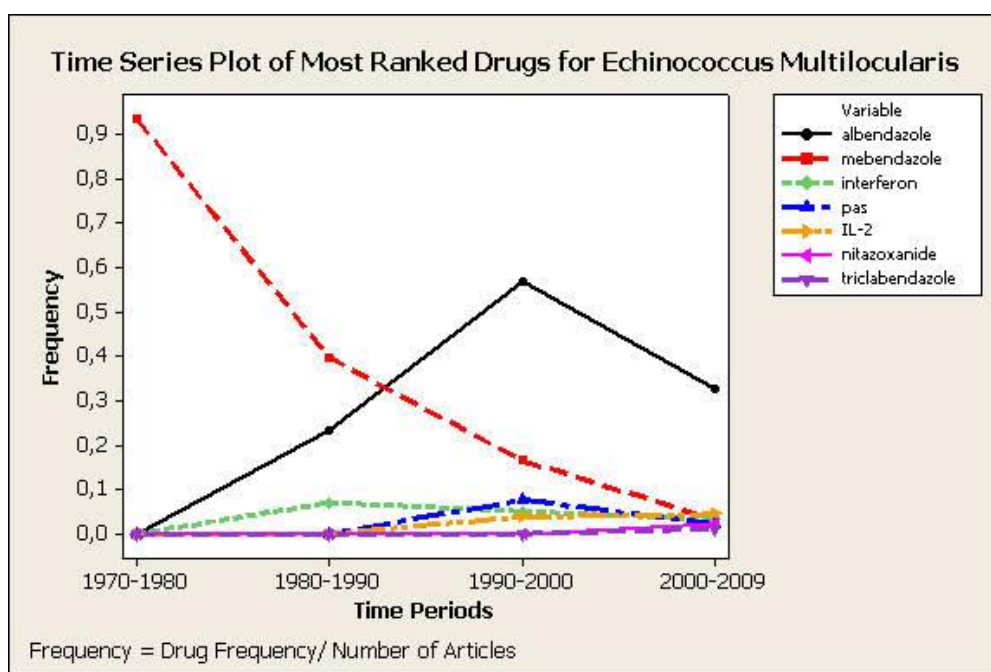


Figure 47: Time series plot of most ranked drugs for Echinococcus Multilocularis

Table 32: Echinococcus Multilocularis/diseases co-occurrences based on the UMLS disease filter server

Rank	Disease	Co-occurrence Frequency	PMI
1	echinococcus	384	9.999
2	tapeworm	27	12.106
3	mass	19	15.364
4	tumor	15	17.297
5	hydatidosis	10	11.972
6	liver cancer	8	17.101
7	liver disease	6	18.336
8	cancer	5	17.248
9	aids	5	11.458
10	carcinoma	4	16.145

Table 33: Echinococcus Multilocularis/genes co-occurrences based on the Swissprot filter server

Rank	Gene	Co-occurrence Frequency	PMI
1	had	54	2.542
2	pcr	29	6.557
3	who	18	5.095
4	cd4	16	8.380
5	red	16	7.179
6	il-10	14	9.356
7	rats	12	2.157
8	inos	10	9.606
9	ifn-gamma	9	8.017
10	tnf-alpha	9	6.424

Table 34: Clonorchis Sinensis /drugs co-occurrences based on the Drugbank Filter Server

Rank	Drugs	Co-occurrence Frequency	PMI
1	praziquantel	41	8.928
2	IL-2	7	11.439
3	pgi	4	5.700
4	albendazole	4	8.864
5	acetone	2	10.633
6	prednisolone	2	11.533
7	choline	2	11.960
8	gentian violet	1	5.087
9	chloroquine	1	10.123
10	hydroxypropyl methylcellulose	1	10.254

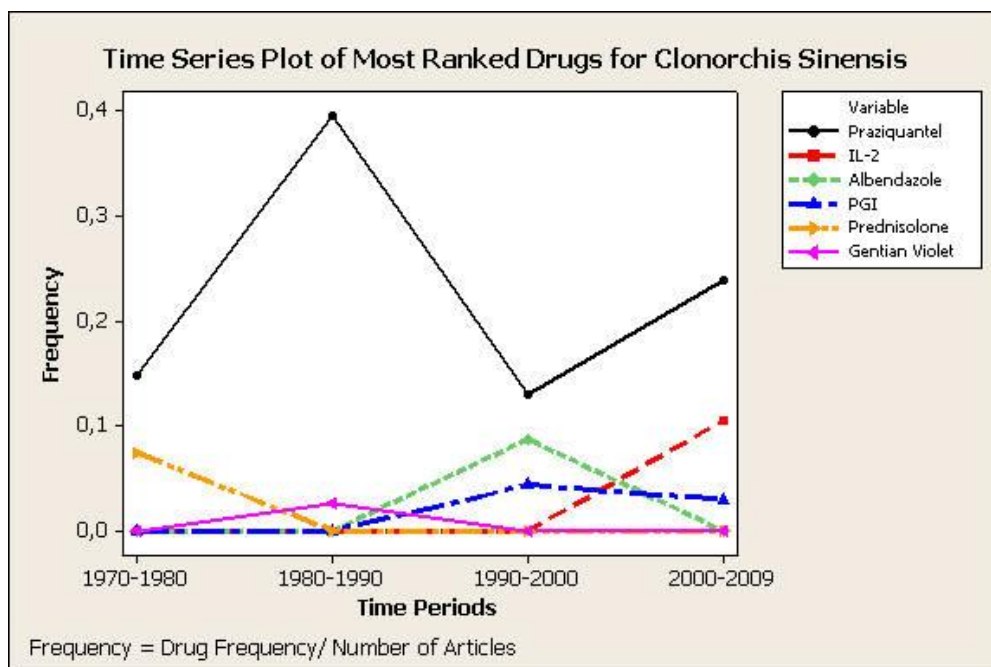


Figure 48: Time series plot of most ranked drugs for Clonorchis Sinensis

Table 35: Clonorchis Sinensis/diseases co-occurrences based on the UMLS disease filter server

Rank	Diseases	Co-occurrence Frequency	PMI
1	cholangiocarcinoma	108	11.478
2	cholangitis	26	8.892
3	tumor	24	3.914
4	carcinoma	17	4.568
5	hepatocellular carcinoma	14	4.926
6	liver cancer	14	3.332
7	cancer	14	3.185
8	tumors	13	3.150
9	cirrhosis	11	3.775
10	hepatitis b	9	5.448

Table 36: Clonorchis Sinensis/genes co-occurrences based on the Swissprot filter server

Rank	Genes	Co-occurrence Frequency	PMI
1	fish	54	8.788
2	had	41	2.508
3	rats	36	4.105
4	who	21	5.681
5	dmn	19	11.768
6	hcc	16	6.883
7	fabp	13	11.263
8	pcr	13	5.763
9	cox-2	12	11.016
10	hbv	11	7.021

Table 37: Opisthorchis Viverrini/drugs co-occurrences based on the Drugbank filter server

Rank	Drug	Co-occurrence Frequency	PMI
1	praziquantel	94	13.994
2	cholera vaccine	14	9.544
3	ivermectin	6	10.088
4	gentian violet	6	7.025
5	albendazole	4	7.835
6	rifampicin	4	13.238
7	magnesium sulfate	4	11.322
8	chenodeoxycholic acid	2	4.707
9	nicotine	1	6.677
10	absorbic acid	1	6.244

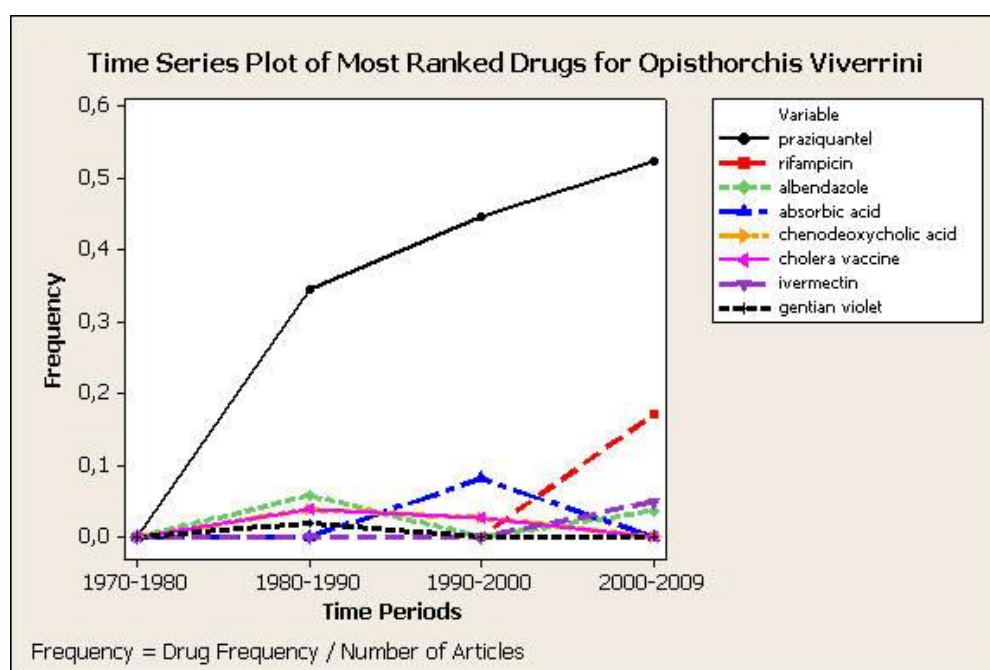


Figure 49: Time series plot of most ranked drugs for Opisthorchis Viverrini

Table 38: Opisthorchis Viverrini/diseases co-occurrences based on the UMLS disease filter server

Rank	Disease	Co-occurrence Frequency	PMI
1	cholangiocarcinoma	23	11.903
2	cancer	5	17.248
3	tumor	4	17.297
4	tumors	3	17.176
5	carcinoma	3	16.145
6	cholangitis	3	12.435
7	cholelithiasis	2	12.338
8	cholangiocarcinomas	2	11.966
9	cancers	1	17.151
10	liver cancers	1	16.978

Table 39: Opisthorchis Viverrini/genes co-occurrences based on Swissprot filter server

Rank	Gene	Co-occurrence Frequency	PMI
1	pcr	21	6.196
2	fish	14	6.581
3	cca	11	12.400
4	aep	10	13.934
5	kras	6	14.094
6	01-Apr	5	16.689
7	tp53	5	11.700
8	timps	4	11.737
9	alp	4	7.825
10	mmmps-2	3	15.630

9 B. RELATION EXTRACTION FOR BACTERIA

Table 40: Salmonella Typhimurium/drugs co-occurrences based on the Drugbank filter server

Rank	Drug	Co-occurrence Frequency	PMI
1	phenobarbital	265	6.811
2	thyroid	165	6.177
3	acetone	143	8.653
4	cyclophosphamide	114	7.321
5	carisoprodol	110	14.586
6	interferon	106	5.386
7	vitamin a	89	6.334
8	methylcellulose	82	10.467
9	azt	73	10.977
10	sodium fluoride	55	9.197

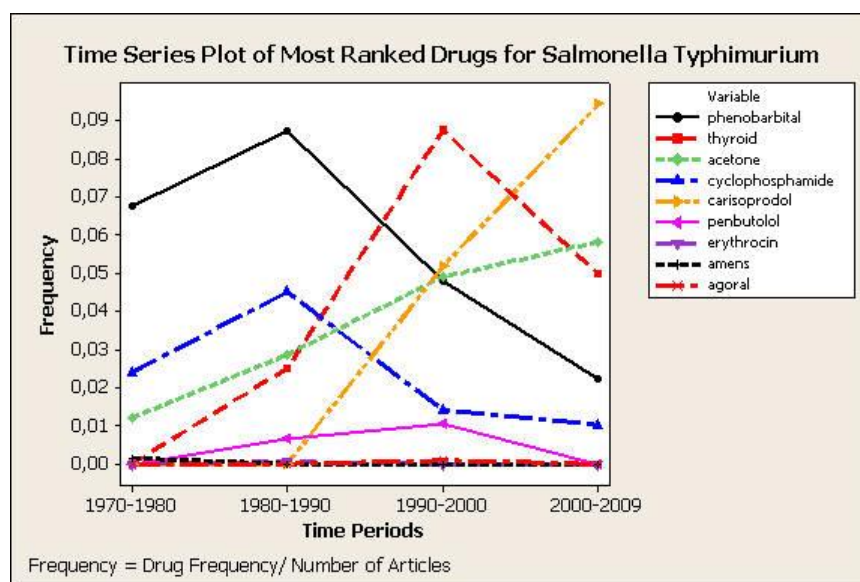


Figure 50: Time series plot of most ranked drugs for Salmonella Typhimurium

Table 41: Salmonella Typhimurium/diseases co-occurrences based on the UMLS disease filter

Rank	Disease	Co-occurrence Frequency	PMI
1	neoplasms	565	4.119
2	consumption	363	6.416
3	tumor	298	3.042
4	carcinoma	262	4.006
5	carcinomas	255	3.923
6	tumors	240	2.850
7	cancer	223	2.672
8	lymphoma	211	6.843
9	salmonellosis	119	11.399
10	mass	119	3.650

Table 42: Salmonella Typhimurium/genes co-occurrences based on the Swissprot filter server

Rank	Genes/Protein	Co-occurrence Frequency	PMI
1	rats	5786	6.935
2	rat	2286	5.528
3	had	796	2.288
4	fed	363	6.052
5	trp	332	11.155
6	and 1	286	2.102
7	red	277	7.157
8	cas	252	10.268
9	dyes	217	6.420
10	mice	195	3.343

Table 43: Staphylococcus Aureus/drugs co-occurrences based on the Drugbank filter server

Rank	Drug	Co-occurrence Frequency	PMI
1	methicillin	215	14.341
2	vancomycin	118	13.171
3	linezolid	53	14.930
4	ciprofloxacin	44	10.998
5	rifampin	40	9.154
6	azithromycin	33	12.778
7	gentamicin	31	10.114
8	ceftriaxone	29	11.827
9	il-2	29	7.521
10	urea	25	5.198

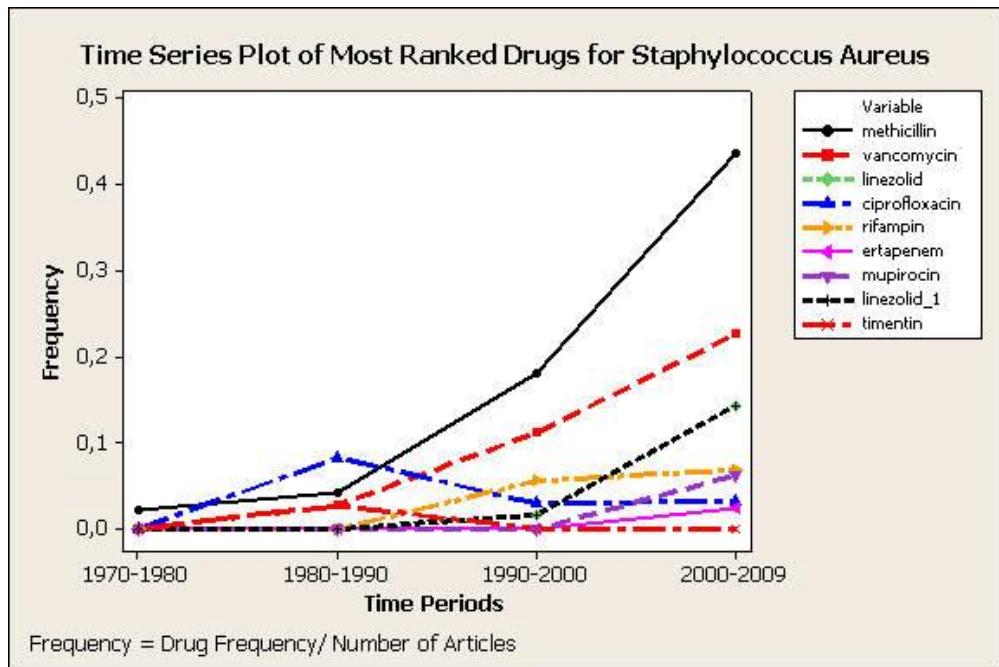


Figure 51: Time series plot of most ranked drugs for Staphylococcus Aureus

Table 44: Staphylococcus Aureus/diseases co-occurrences based on the UMLS disease filter server

Rank	Disease	Co-occurrence Frequency	PMI
1	sepsis	180	8.511
2	abscess	115	7.833
3	pneumonia	108	9.317
4	cirrhosis	89	4.267
5	tumor	83	3.179
6	endocarditis	71	11.190
7	liver abscess	68	7.081
8	liver disease	60	1.673
9	peritonitis	56	8.262
10	ascites	49	6.288

Table 45: Staphylococcus Aureus/genes co-occurrences based on the Swissprot filter server

Rank	Gene	Co-occurrence Frequency	PMI
1	had	383	3.214
2	rat	308	4.618
3	rats	183	3.934
4	who	131	5.805
5	seb	97	15.008
6	protein a	74	10.917
7	insulin	69	5.529
8	tnf-alpha	63	7.077
9	v8 protease	61	12.879
10	ifn-gamma	56	8.501

Table 46: Helicobacter Pylori /drugs co-occurrences based on the Drugbank filter server

Rank	Drug	Co-occurrence Frequency	PMI
1	metronidazole	98	11.671
2	urea	95	7.810
3	omeprazole	94	12.512
4	lansoprazole	86	14.712
5	clarithromycin	83	13.717
6	amoxicillin	57	12.632
7	tetracycline	36	10.114
8	pantoprazole	33	14.056
9	rabeprazole	31	14.538
10	ranitidine	26	11.204

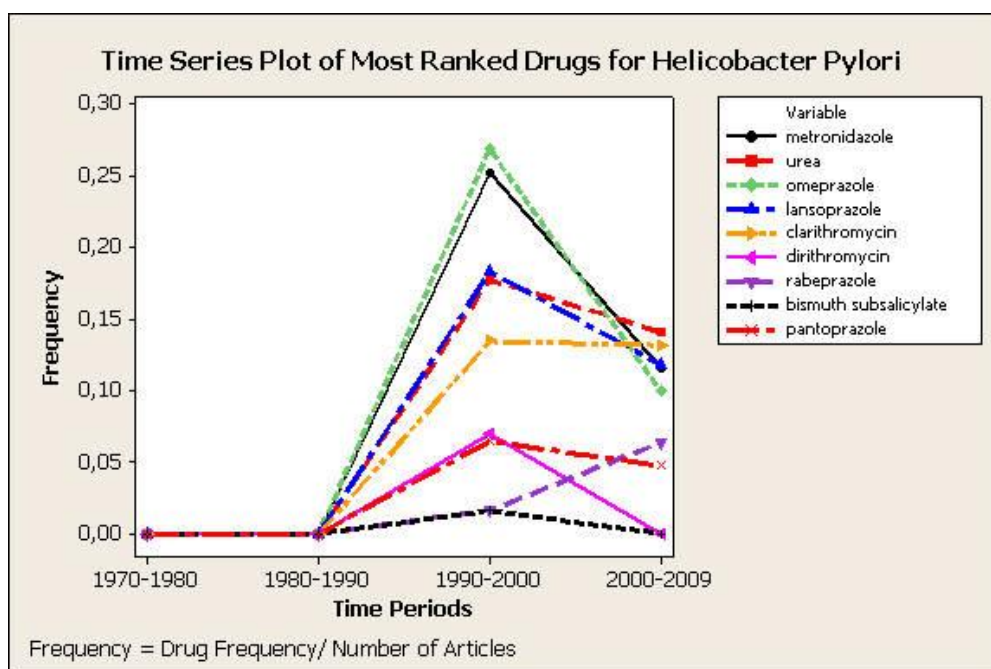


Figure 52: Time series plot of most ranked drugs for Helicobacter Pylori

Table 47: Helicobacter Pylori /diseases co-occurrences based on the UMLS disease filter server

Rank	Disease	Co-occurrence Frequency	PMI
1	helicobacter pylori	861	15.326
2	gastritis	282	13.255
3	cirrhosis	245	6.413
4	peptic ulcer	189	10.977
5	liver cirrhosis	142	5.766
6	lymphoma	127	8.778
7	liver disease	110	3.232
8	cancer	109	4.306
9	hepatic encephalopathy	84	8.518
10	duodenal ulcer	83	11.337

Table 48: Helicobacter Pylori /genes co-occurrences based on the Swissprot filter server

Rank	Gene	Co-occurrence Frequency	PMI
1	had	227	3.144
2	urease	118	13.941
3	caga	117	16.172
4	who	111	6.251
5	pcr	110	7.011
6	cyp2c19	94	11.886
7	hcc	89	7.527
8	proton pump	63	9.320
9	pbc	48	9.336
10	vaca	46	15.147

Table 49: Mycobacterium Tuberculosis/drugs co-occurrences based on the Drugbank filter server

Rank	Drug	Co-occurrence Frequency	PMI
1	rifampicin	199	11.183
2	isoniazid	183	11.162
3	tuberculin	75	12.023
4	inh	70	12.143
5	ethambutol	69	12.221
6	streptomycin	60	11.022
7	pyrazinamide	59	11.827
8	interferon	50	6.353
9	interferon gamma	43	7.662
10	lam	40	10.018

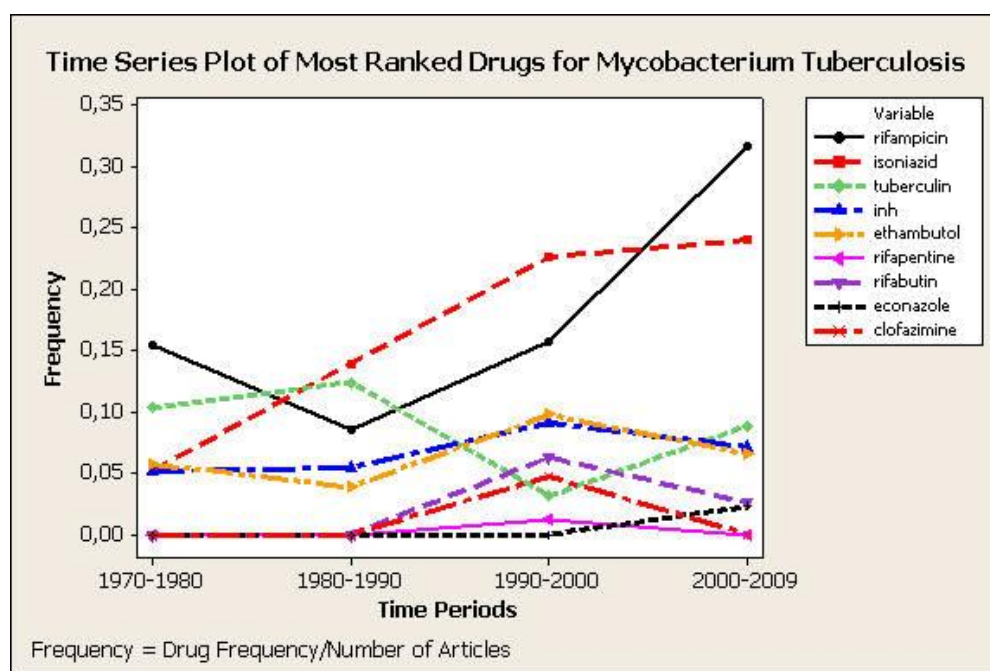


Figure 53: Time series plot of most ranked drugs for Mycobacterium Tuberculosis

Table 50: Mycobacterium Tuberculosis/diseases co-occurrences occurrences based on the UMLS disease filter server

Rank	Disease	Co-occurrence Frequency	PMI
1	tb	302	13.462
2	aids	150	9.941
3	pulmonary tuberculosis	127	10.645
4	miliary tuberculosis	71	12.033
5	hepatitis	60	4.058
6	tumor	56	2.682
7	leprosy	47	11.096
8	pneumonia	37	7.841
9	hiv infection	36	7.201
10	disseminated tuberculosis	35	11.635

Table 51: Mycobacterium tuberculosis/genes co-occurrences based on the Swissprot filter server

Rank	Gene	Co-occurrence Frequency	PMI
1	had	305	2.955
2	ifn-gamma	178	10.239
3	tnf	120	8.520
4	who	115	5.687
5	pcr	106	6.343
6	pigs	92	6.356
7	cd4	87	8.738
8	rats	73	2.678
9	th1	57	10.144
10	rin	52	14.521

Table 52: Listeria Monocytogenes/drugs co-occurrences based on the Drugbank filter server

Rank	Drug	Co-occurrence Frequency	PMI
1	interferon	82	7.156
2	ampicillin	66	10.774
3	interferon gamma	39	7.611
4	gentamicin	27	10.074
5	cyclosporine	22	6.629
6	beta 2	20	4.259
7	cyclophosphamide	20	6.951
8	ibuprofen	18	9.879
9	il 2	16	6.826
10	vitamin a	16	5.999

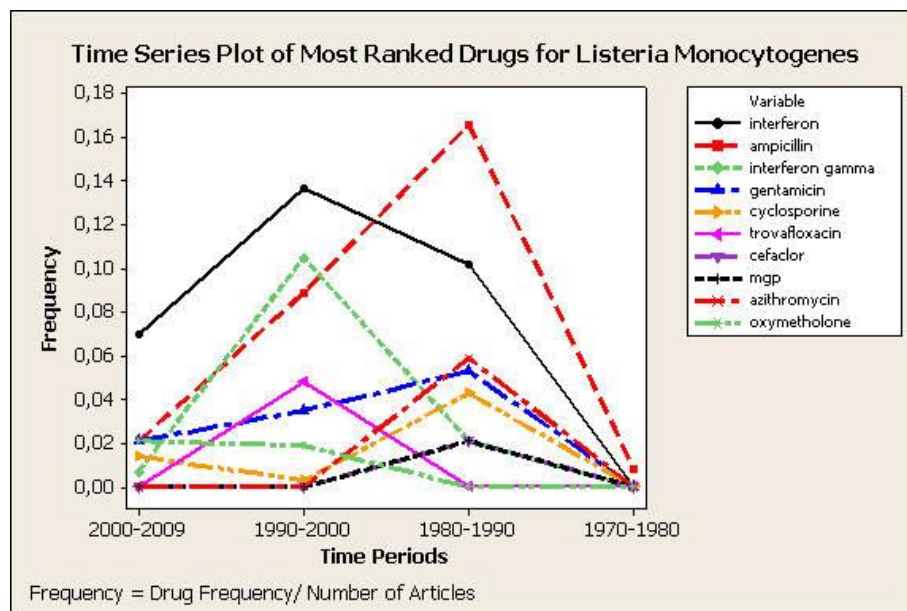


Figure 54: Time series plot of most ranked drugs for Listeria Monocytogenes

Table 53: Listeria Monocytogenes/diseases co-occurrences based on the UMLS disease filter server

Rank	Disease	Co-occurrence Frequency	PMI
1	listeriosis	368	13.083
2	tumor	138	4.072
3	peritonitis	64	8.614
4	meningitis	51	10.180
5	adhesion	32	6.915
6	hepatitis	29	3.098
7	cirrhosis	28	2.758
8	liver abscess	20	5.474
9	cancer	18	1.182
10	consumption	17	4.140

Table 54: Listeria Monocytogenes/genes co-occurrences based on the Swissprot filter server

Rank	Gene	Co-occurrence Frequency	PMI
1	ifn-gamma	362	11.352
2	had	174	2.234
3	delta	156	8.995
4	cd4	137	9.482
5	il-6	96	8.711
6	rats	94	3.131
7	tnf-alpha	92	7.782
8	tnf	89	8.178
9	fed	73	5.878
10	il-4	64	9.933

Table 55: Klebsiella Pneumoniae/drugs co-occurrences based on the Drugbank filter server

Rank	Drug	Co-occurrence Frequency	PMI
1	kp	50	13.283
2	ceftazidime	39	13.391
3	gentamicin	25	10.683
4	piperacillin	20	12.883
5	ticarcillin	19	14.651
6	ceftriaxone	19	12.096
7	ampicillin	19	9.698
8	moxifloxacin	17	13.866
9	timentin	16	15.988
10	tobramycin	15	12.310

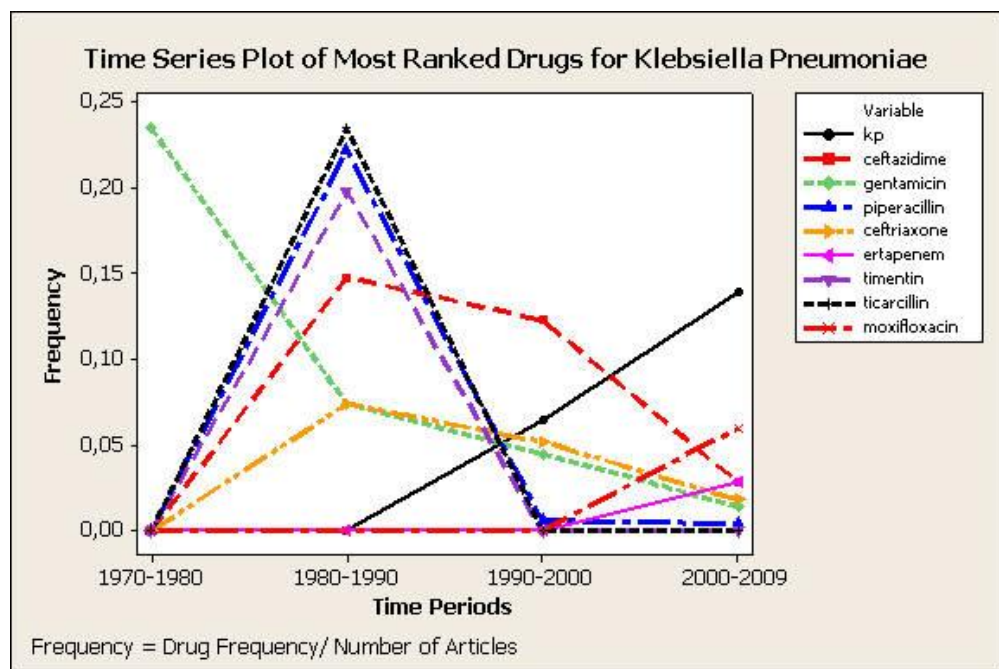


Figure 55: Time series plot of most ranked drugs for Klebsiella Pneumoniae

Table 56: Klebsiella Pneumoniae/diseases co-occurrences based on the UMLS disease filter server

Rank	Disease	Co-occurrence Frequency	PMI
1	liver abscess	383	10.454
2	abscess	219	9.642
3	pyogenic liver abscess	191	12.814
4	endophthalmitis	134	15.484
5	diabetes mellitus	90	7.699
6	peritonitis	84	9.726
7	pneumonia	81	9.781
8	meningitis	71	11.378
9	sepsis	71	8.048
10	cirrhosis	49	4.285

Table 57: Klebsiella Pneumoniae/genes co-occurrences based on the Swissprot filter server

Rank	Gene	Co-occurrence Frequency	PMI
1	had	238	3.407
2	rats	83	3.672
3	who	81	5.991
4	gas	70	6.638
5	tnf-alpha	59	7.862
6	sbp	53	12.180
7	pain	47	7.444
8	pla	38	10.720
9	cps	31	11.920
10	rat	30	2.137

Table 58: Pseudomonas Aeruginosa/drugs co-occurrences based on the Drugbank filter server

Rank	Drug	Co-occurrence Frequency	PMI
1	ceftazidime	74	14.495
2	cefoperazone	45	14.233
3	imipenem	44	13.745
4	ciprofloxacin	40	11.920
5	piperacillin	39	14.027
6	gentamicin	39	11.505
7	cyclophosphamide	39	8.815
8	aztreonam	38	14.989
9	cilastatin	34	14.348
10	ticarcillin	33	15.627

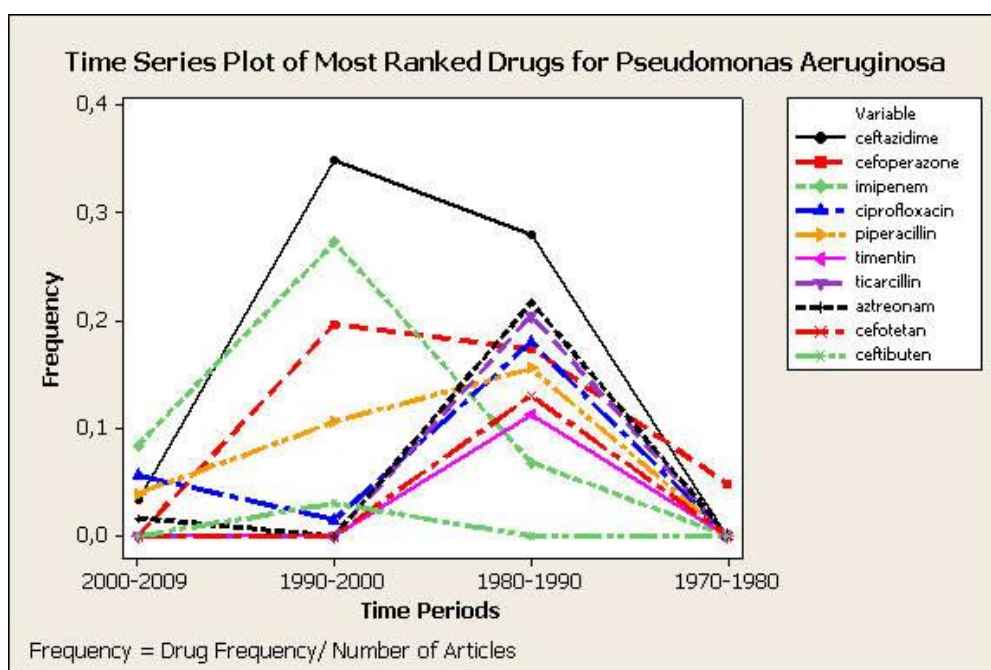


Figure 56: Time series plot of most ranked drugs for Pseudomonas Aeruginosa

Table 59: Pseudomonas Aeruginosa/diseases co-occurrences based on the UMLS disease filter server

Rank	Disease	Co-occurrence Frequency	PMI
1	sepsis	202	9.737
2	pneumonia	113	10.441
3	cystic fibrosis	83	11.068
4	tumor	52	3.564
5	abscess	41	7.405
6	pa	28	2.791
7	septicemia	24	6.583
8	cholangitis	23	7.252
9	liver abscess	23	6.576
10	liver cirrhosis	22	3.450

Table 60: Pseudomonas Aeruginosa/genes co-occurrences based on the Swissprot filter server

Rank	Gene	Co-occurrence Frequency	PMI
1	had	247	3.641
2	rats	131	4.511
3	who	88	6.290
4	rat	64	3.411
5	pea	58	13.225
6	tnf-alpha	58	8.017
7	pao1	51	17.140
8	plus	43	7.405
9	pe38	36	16.753
10	tnf	35	7.732

Table 61: Streptococcus Pneumoniae/drugs co-occurrences based on the Drugbank filter server

Rank	Drugs	Co-occurrence Frequency	PMI
1	azithromycin	27	14.178
2	trovafloxacin	26	15.176
3	moxifloxacin	22	14.972
4	fluoroquinolones	20	10.704
5	erythromycin	19	9.810
6	amoxicillin	18	11.973
7	ceftriaxone	17	12.746
8	clarithromycin	17	12.434
9	cefamandole	15	12.977
10	morphine	15	9.097

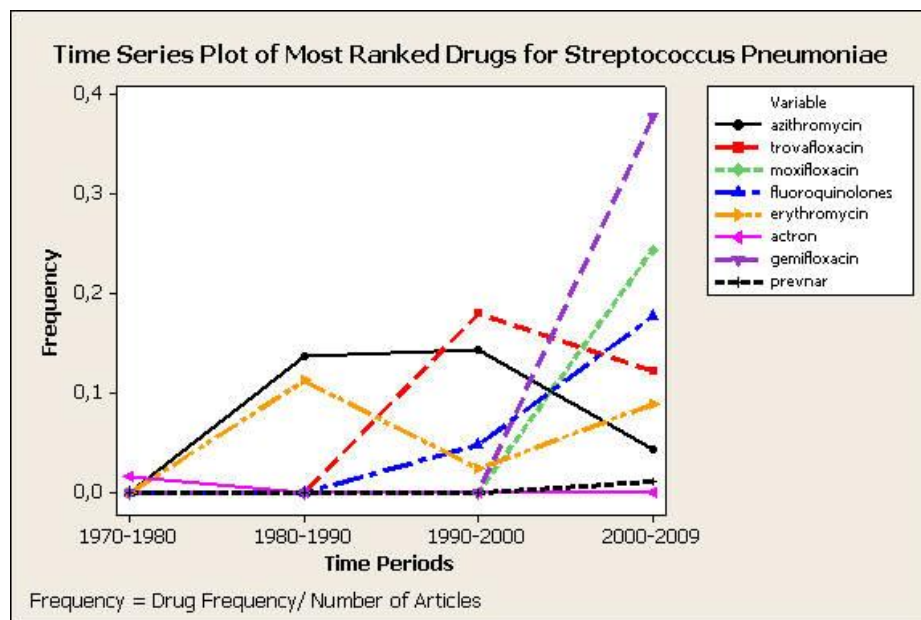


Figure 57: Time series plot of most ranked drugs for Streptococcus Pneumoniae

Table 62: Streptococcus Pneumoniae/diseases co-occurrences based on the UMLS disease filter server

Rank	Disease	Co-occurrence Frequency	PMI
1	pneumonia	125	11.217
2	peritonitis	80	10.465
3	cirrhosis	64	5.481
4	sepsis	38	7.956
5	abscess	37	7.887
6	liver cirrhosis	35	4.750
7	meningitis	32	11.038
8	ascites	29	7.220
9	liver disease	26	2.155
10	community-acquired pneumonia	19	13.491

Table 63: Streptococcus pneumoniae/genes co-occurrences based on the Swissprot filter server

Rank	Gene	Co-occurrence Frequency	PMI
1	pneumolysin	19	18.046
2	cbpa	4	16.798
3	sbem	8	16.213
4	irak-4	8	16.213
5	accd	1	15.798
6	peptide deformylase	2	15.213
7	slpi	9	15.061
8	ermB	1	14.798
9	iga1 protease	1	14.798
10	aspm	2	14.798

10 CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Pınar, Yıldırım
Nationality: Turkish (TC)
Date and Place of Birth: 30 June 1967, Erzincan
Marital Status: Single
Phone: +90 542 237 04 22
Fax: +90 312 284 80 43
email: pinar@cankaya.edu.tr

EDUCATION

Degree	Institution	Year of Graduation
MS	Akdeniz University Medical Informatics and Biostatistics	1995
BS	Yıldız Technical University Electronics and Communications	1989
High School	Antalya High School Antalya	1984

WORK EXPERIENCE

Year	Place	Enrollment
2007- Present	Çankaya University Department of Computer Engineering	Ph.D Fellow
2004- 2007	Hacettepe University Hospitals Computer Center	System Manager
1991- 2004	Akdeniz University Computer Center	System Analyst
1990- 1991	Bilisim Inc	IT Specialist

FOREIGN LANGUAGES

Advanced English

PUBLICATIONS

1. BAŞKURT O.K., ZAYİM N., YILDIRIM P., SAKA O., "Interactive Medical Education System" ACIME (Academic Computing in Macintosh Environment), Akdeniz University, Antalya, Turkey 1995, Page 13-19.

2. YILDIRIM P., ARIÖZ U., "Analysis of Picture Archiving and Communications System " Medical Informatics 2005, Antalya, Turkey 17–20 Nov 2005, Page 153–159.

3. ARIÖZ U., YILDIRIM P., KAYMAKOĞLU B., ERDEM A., "Clinical Document Architecture(CDA)", Medical Informatics 2005, Antalya, Turkey 17-20 Nov 2005, Page 163.
4. GÜLKESEN K.H., ERDEM A., KAYMAKOĞLU B., ARIÖZ B., YILDIRIM P., SAKA O., "Typographic Errors Variants in Pathology Reports", Medical Informatics 2005, Antalya, Turkey 17-20 Nov 2005, Page 145-152.
5. YILDIRIM P., ÖZTANER S. M., GÜLKESEN K.H., "Survey on Radiologists' View of PACS", Medical Informatics 2006, Antalya, Turkey 16-19 Nov 2006, Page 24-28.
6. YILDIRIM P., BAKIR A., "Interoperability of IT Systems in Hacettepe University Hospitals", Medical Informatics 2006, Antalya, Turkey, 16-19 Nov 2006, Page 103-109.
7. YILDIRIM P., BAKIR A., BİRİNCİ S., "Teleradiology Application in Hacettepe University Hospitals", Medical Informatics 2007, Antalya, Turkey 15-18 Nov 2007, Page 24-29.
8. YILDIRIM P., ULUDAĞ M., GÖRÜR A., "Data Mining in Hospital Information Systems", Akademik Bilişim 2008, Çanakkale Onsekiz Mart Üniversitesi 30 Ocak-01 Şubat 2008, Page 106.
9. YILDIRIM P., ULUDAĞ M., GÖRÜR A., "Data Mining in Hospital Information Systems", Hastane ve Yasam Dergisi(Akademik ve Aktüel Tıp Dergisi) Mart 2008, No 31, ISSN 1305-516X, Page 96-100.
10. YILDIRIM P., MUMCUOĞLU Ü.E., TOLUN M.R., ÖZMEN M.N., ULUDAĞ M., "Multi-Relational Data Mining Approach for Effective Data Mining in Clinical Databases", Medical Informatics 2008, Antalya, Turkey 13-16 Nov 2008, Page 59-67.
11. YILDIRIM P., BAKIR A., "Interoperability between various IT Systems in Hacettepe University Hospitals in Turkey", International HL7 Interoperability Conference IHIC 2006, Cologne Germany, 24-25 August 2006 Available at <http://ihic.hl7.de/proceedings.html>.
12. YILDIRIM P., BAKIR A., BİRİNCİ S., "Web Based Teleradiology in Hacettepe University Hospitals in Turkey", In Proceedings of 12th International Symposium for Health Information Management Research, Sheffield, UK, 18-20 July 2007, Page 109-118.
13. YILDIRIM P., TOLUN M.R., "Induction for Radiology Patients", e-health 08, City University, London, UK, 8-9 September 2008, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Vol 0001, ISSN 1867-8211, Springer Berlin Heidelberg, Page 208-220.

14. YILDIRIM P., JIMENO YEPES A.J., SAKA O., REBHOLZ SCHUHMANN D., TURKMIA' 09 Proceedings, "Biyomedikal Metinlerde Bilgi Keşfi", VI. Ulusal Tıp Bilişimi Kongresi Bildirileri, ENMI (European Notes in Medical Informatics) Vol V, No 1, 2009, ISSN 1861-3179, p 323.
15. YILDIRIM P., ÇEKEN Ç., SAKA O., "Disease Drug Relationships for Vasculitis Diseases", 5th International Symposium on Health Informatics and Bioinformatics, HIBIT'10, April 20-22, 2010, Antalya, Turkey(IEEE Xplore).
16. YILDIRIM P., TOLUN R. MEHMET,"Biyomedikal Metin Madenciliği", 3. Mühendislik ve Teknoloji Sempozyumu, 29-30 Nisan 2010, Çankaya Üniversitesi, Ankara. Sayfa 150-155.
17. YILDIRIM P., ÇEKEN Ç., HASSANPOUR R., TOLUN M.R., "Prediction of Similarities among Rheumatic Diseases", Journal of Medical Systems, Science Citation Index Expanded(accepted).
18. YILDIRIM P., ÇEKEN K., SAKA O., "Knowledge discovery for the treatment of bacteria affecting liver", Turkish Journal of Medical Sciences, Science Citation Index Expanded (accepted).
19. YILDIRIM P., ÇEKEN Ç., ÇEKEN K., TOLUN M.R., "Cluster analysis for vasculitis diseases", Second International Conference Networked Digital Technologies 2010, Prague, Czech Republic, July 7-9, 2010, Part II, CCIS 88, pp.36-45, Springer-Verlag Berlin Heidelberg.