

**PHARMACODYNAMICS MINER: AN AUTOMATED EXTRACTION
OF PHARMACODYNAMIC DRUG INTERACTIONS**

HRISHIKESH A LOKHANDE

Submitted to the faculty of the Bioinformatics Graduate Program

in partial fulfillment of the requirements for the degree

Master of Science in Bioinformatics

in the School of Informatics

Indiana University

MAY 2013

Accepted by the Faculty of Indiana University,
in partial fulfillment of the requirements for the degree of Master of Science
in Bioinformatics

Dr. Lang Li, Ph.D., Chairperson

Master's Thesis
Committee

Dr. Yunlong Liu, Ph.D.

Dr. Xiaowen Liu, Ph.D.

© 2013
HrishikeshArunLokhande
ALL RIGHTS RESERVED

Dedicated to the Almighty

ACKNOWLEDGMENTS

Foremost, I praise GOD, the almighty for this opportunity and for making me capable of proceeding successfully. This thesis exists due to guidance and assistance of many people. I would therefore like to thank all of them.

I owe my sincere gratitude to my thesis advisor Dr. Lang Li, for his encouragement, motivation and immense knowledge. His thoughtful guidance, energy and vision were a great source of inspiration for me. Besides, my advisor, I would also like to thank the rest of committee Dr. Yunlong Liu and Dr. Xiaowen Liu for their encouragement, support and insightful suggestions.

I owe my sincere thanks to School of Informatics at Indiana University Purdue University Indianapolis for providing me an opportunity to pursue a bright carrier in bioinformatics.

It gives me immense pleasure to acknowledge my friends and lab members Akshay, Swapnil, Aniruddha, Aditya, Abdullah, SaiKalyan, Jay, Snehal, Arindhom, Prasad, Sriharsha, Bhargav, Scott, Guanglong, Heng-yi and Zhiping Wang for their valuable comments.

Words are not enough to express my gratitude and affection towards my parents Shri. ArunShripatiLokhande and Smt. JyotiArunLokhande whose support, blessing, love and strength have made me the person, I am today. Last but not the least; I would like to thank my brother Shubham for his support and love.

Absence in this acknowledgement does not mean lack of gratitude.

ABSTRACT OF DISSERTATION

HRISHIKESH LOKHANDE

PHARMACODYNAMICS MINER: AN AUTOMATED EXTRACTION OF PHARMACODYNAMIC DRUG INTERACTIONS

Pharmacodynamics (PD) studies the relationship between drug concentration and drug effect on target sites. This field has recently gained attention as studies involving PD Drug-Drug interactions (DDI) assure discovery of multi-targeted drug agents and novel efficacious drug combinations. A PD drug combination could be synergistic, additive or antagonistic depending upon the summed effect of the drug combination at a target site. The PD literature has grown immensely and most of its knowledge is dispersed across different scientific journals, thus the manual identification of PD DDI is a challenge. In order to support an automated means to extract PD DDI we propose Pharmacodynamics Miner (PD-Miner). PD-Miner is a text-mining tool, which is capable of identifying PD DDI from *in vitro* PD experiments. It is powered by two major features i.e. collection of full text articles and *in vitro* PD ontology. The *in vitro* PD ontology currently has four classes and more than hundred subclasses; based on these classes and subclasses the full text corpus is annotated. The annotated full text corpus forms a database of articles, which can be queried based upon drug keywords and ontology subclasses. Since the ontology covers term and concept meanings, the system is capable of formulating semantic queries. PD-Miner extracts *in vitro* PD DDI based upon references to cell lines and cell phenotypes. The results are in the form of fragments of sentences in which important concepts are visually highlighted. To determine the accuracy of the system, we used a gold standard of 5 expert curated articles.

PD-Miner identified DDI with a recall of 75% and a precision of 46.55%. Along with the development of PD Miner, we also report development of a semantically annotated *in vitro* PD corpus. This corpus includes term and sentence level annotations and serves as a gold standard for future text mining.

TABLE OF CONTENTS

ACKNOWLEDGEMENT	v
ABSTRACT OF DISSERTATION	vi
LIST OF TABLES	x
LIST OF FIGURES	xi
CHAPTER ONE. INTRODUCTION	1
1.1 Background	1
1.2 Objective and contribution	7
1.3 Outline	9
CHAPTER TWO. RELATED WORK	10
2.1 Biological Information retrieval and extraction system	10
2.1.1 Textpresso	11
2.1.2 Pharmpresso	12
2.2 Cell lines	13
2.2.1 Cell line Ontology	13
2.3 Drug Databases	15
2.3.1 DrugBank	15
CHAPTER THREE.METHODS.....	16
3.1 Data collection	16
3.2 Drug dictionary construction	16
3.3 Construction of <i>in-vitro</i> PD ontology	17
3.3.1 Drug interaction classes	18
3.3.2 Drug interaction analysis	19
3.3.3 Cell lines	20
3.3.3.1 Human organ specific	20
3.3.3.2 Animal specific	22
3.3.4 Cell phenotypes	24
CHAPTER FOUR. DATABASE OF ANNOTATED FULL TEXTS	25
4.1 Automated download of free full texts from PubMed.	25
4.1.1 Conversion of PDF to text.	25

4.1.2 Filtering of raw text to titles, abstract and full text	27
4.1.3 Sentence segmentation and tokenization	27
4.2 Sentence level annotation	28
4.2.1 Named entity recognition and annotation with terms of ontology	28
4.2.2 Named entity recognition and annotation with terms of drug dictionary.....	29
CHAPTER FIVE. DEVELOPMENT OF SEMANTICALLY ANNOTATED CORPUS	31
5.1 In vitro PD corpus	31
5.1.1 Corpus annotation	32
5.1.2 PD Tagger	33
5.1.3 Sentence level annotation	34
5.2 GENIA format	34
5.3 Conclusion	35
CHAPTER SIX. ACCESSING THE WEB INTERFACE	36
5.1.1 Home	36
5.1.2 Browse	37
5.1.3 Search	38
5.1.4 Corpus	39
5.1.4 Download	39
CHAPTER SEVEN. RESULTS	40
7.1.1 PD Miner database overview	40
7.1.2 Evaluation of PD Miner	41
7.1.2.1 Error analysis	43
7.2 Categorical searches	44
7.3 Cell line coverage	45
7.4 Corpus and ontology	46
CHAPTER EIGHT. DISCUSSION	48
8.1 Accomplishments	48
8.2 Limitations	50
CHAPTER NINE. REFERENCES	52

LIST OF TABLES

TABLE 1: DRUG DICTIONARY WITH GENERIC NAMES, SYNONYMS, BRAND NAMES	17
TABLE 2: ORGAN WITH THEIR TISSUES, SAMPLE CELL LINE AND CELL LINE COUNT	22
TABLE 3: SHOWING ORGANISM CORRESPONDING CELL LINE AND CELL TYPE.....	22
TABLE 4: SHOWING SAMPLE CELLULAR PHENOTYPES WITH CLASSES AND SUBCLASSES	24
TABLE 5: NUMBER OF INTERACTION, CELL LINES AND CELL PHENOTYPES FROM CORPUS	28
TABLE 6: OVERVIEW OF THE DATABASE	40
TABLE 7: EVALUATION OF PRECISION AND RECALL OF THE SYSTEM.....	42
TABLE 8: ERROR ANALYSIS FOR MISSED DRUG INTERACTIONS	43

LIST OF FIGURES

FIGURE 1: DIFFERENTIATING PK AND PD	2
FIGURE 2:CLO HIERARCHICAL STRUCTURE OF ONTOLOGY TERMS. BLUE BOXES PRESENTS THE EXTERNAL CLASSES WHILE YELLOW INDICATE THE TERMS WITH CLO ID'S	14
FIGURE 3: <i>IN VITRO</i> PD ONTOLOGY, SHOWING DDI CLASSES	18
FIGURE 4: <i>IN VITRO</i> PD ONTOLOGY, SHOWING DDI DETECTION CLASSES	19
FIGURE 5: DATA COLLECTION FOR THE CELL LINE ONTOLOGY CLASS. BA: BIOASSAY ONTOLOGY. AT: AMERICAN TYPE CULTURE COLLECTION. CL: CELL LINE DATABASE. AU: AUSTRALIAN CELLBANK DS: DEUTSCHE SAMMLUNG VON MIKROORGANISMEN UND ZELLKULTUREN. BO: BREAST CELL LINE ONTOLOGY. CORI: CORIELL CELL LINE ONTOLOGY	23
FIGURE 6: IN VITRO PD ONTOLOGY, CELLULAR PROCESS CLASSES	24
FIGURE 7: CONFIGURATION OF PD CORPUS	35
FIGURE 8: BROWSER BASED DISPLAY OF THE CORPUS	35
FIGURE 9: MAIN PAGE OF THE INTERFACE	36
FIGURE 10: ONLINE ACCESS TO THE ONTOLOGY	37
FIGURE 11: CELL LINE SEARCH FOR PD MINER	38
FIGURE 12: ONLINE VIEW OF THE ANNOTATED CORPUS	39
FIGURE 13: SHOWS THE PD MINER PIPELINE FOR DOCUMENT RETRIEVAL AND SENTENCE EXTRACTION	41
FIGURE 14: EVALUATION SCHEMA	42
FIGURE 15: CATEGORICAL SEARCH WITH CELL DEATH AND BRAIN CELL LINE	45
FIGURE 16: CELL LINE SEARCH	46

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

PHARMACOGENETICS

“Pharmacogenetics (PG) is defined as the study of inherited variations in drug responses and metabolism.” The term is derived from ‘Pharmacology’, which studies the mechanism of drug in the body, while ‘Genetics’ takes into account hereditary and variations of the inherited characteristics. PG is highly valuable, especially for clinical studies, it uncovers an individual’s drug response and possible drug side effects based upon the genetic makeup. This knowledge is important to assist the development of new drugs suitable for an individual (Personalized medicine) or to a particular group. Apart from being a key to personalized medicine, PG offers the following potential advantages.

1. Improved drug selection: Every year more than a million individuals die due to adverse drug reaction, while an equal number of individuals are hospitalized. PG can help to predict adverse drug reactions and differentiate individuals likely to have a positive drug reaction, from the pool of other individuals.
2. Potent and target specific drugs: PG could play an important role in developing efficient and target specific drugs, thereby increasing therapeutic effect and reducing damage to nearby cells and organs.

3. Safer doses: Knowledge of genomic variation would enhance the determination of accurate drug dosages, thus decreasing the adverse effect that might arise due to traditional trial and error methods.

Genetic variability is crucial and can affect pharmacokinetics and pharmacodynamics of drug metabolism. "**Pharmacokinetics** (PK) is defined as the study of the time course of drug absorption, distribution, metabolism, and excretion (ADME)". Clinical PK is the study of the relationships between drug dosage regimens and concentration–time profiles [1]. Clinical pharmacokinetics primarily focuses on enhancing the drug efficacy, while reducing the toxicity of patient’s drug therapy. "**Pharmacodynamics** (PD) refers to the study of relationship between drug concentration at the target site, the resulting effect along with the time course, and adverse effect". The effect is usually measured by analyzing the drug binding receptors; binding receptor could be anything neurons in the central nervous systems or cardiac muscles to control the intensity of heart contraction.

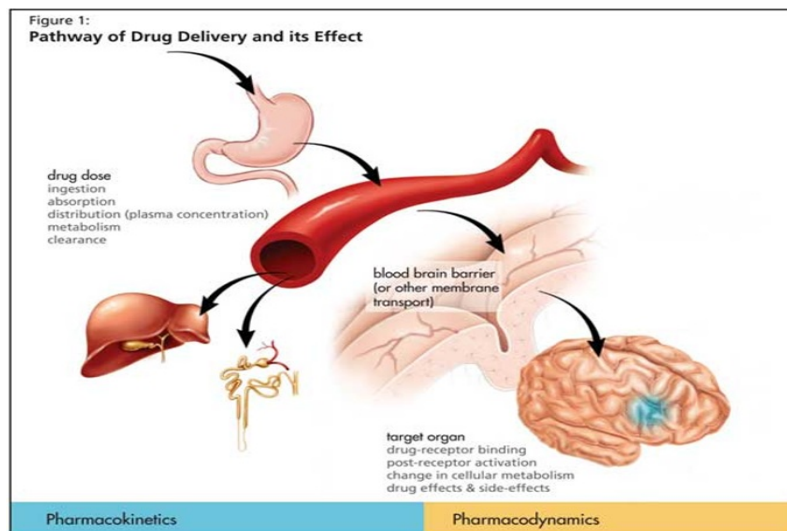


Fig.1 Differentiating PK and PD

DRUG-DRUG INTERACTION

One of the many challenges in drug development and drug administration is accurate identification of drug interactions. The problem of drug interactions arises when multiple drugs are administered together, causing increase or decrease in the activity of one drug by the other drug or drugs. A ***Drug interaction*** can be as defined as a pharmacological or clinical response to the administration or co-exposure of a drug with another substance that modifies the patient's response to the drug [2]. Drug interactions especially adverse drug interactions are responsible for almost 5% of hospitalization in the United States, it is estimated that the total cost associated with each hospitalization is more than \$16000. With the development of new drug the chances of adverse drug interactions grows exponentially as each new drug is added to an individual's regime [3]. The chances of adverse drug reactions are more in elderly population and in patients who consume for than one drug at a time. There are evidences, which suggest that, a clinical drug interactions ranges from 3 to 5% in patients with few medications, but it increases drastically to 20% in patients consuming 10–20 drugs. Adverse drug interaction are complex events and are broadly divided into pharmacokinetic and pharmacodynamics drug interactions.

Pharmacokinetic drug interactions are the drug interactions where one of the drugs alters the rate or activity of absorption, distribution, metabolism, and excretion of the other drug. While, Pharmacodynamic drug interactions are the DDI's where presence of one drug disturbs the functioning of the other drug at a target site. Based upon the resulting of effect

of a drug combination PD-DDI is divided into synergistic, additive, and antagonistic categories.

SYNERGISTIC DRUG INTERACTION

The National Cancer Institute (NCI) defines synergistic drug interaction as the interaction of two or more drugs when their combined effect is greater than the sum of the effects seen when each drug is given alone [4]. Equation wise synergistic drug interaction could be represented as $1(\text{Drug A}) + 1(\text{Drug B}) = 3$. A beneficial synergistic effect occurs when two different types of antibiotics that work in different ways are combined, such as penicillin-G and aminoglycoside antibiotic. This approach is used to cure sub-acute endocarditis. A harmful synergistic effect could be represented by interactions between drugs that depress the CNS, such as morphine and alcohol; this interaction causes additional CNS depression that could be fatal.

ANTAGONISTIC DRUG INTERACTION

Antagonistic effect is the opposite of a synergistic effect. It results in a therapeutic effect that is less than the effect of either drug alone because the second drug either diminishes or cancels the effect of the first drug. As an equation, this concept could be represented as $1(\text{Drug A}) + 1(\text{Drug B}) = 0$. For example, when the heparin antagonist protamine sulfate (A strong basic anticoagulant) is given as an antidote for heparin (A strong acidic anticoagulant) to halt heparin-induced bleeding, a stable salt forms, resulting in loss of anticoagulant activity for both drugs.

ADDITIVE DRUG INTERACTIONS

An additive effect occurs when two or more drugs (in terms of therapeutic effect) are combined and the result is sum of the drug effect. Equation wise additive effect could be represented as $1(\text{Drug A}) + 1(\text{Drug B}) = 2$. Codeine and acetaminophen work differentially to reduce pain. When these two analgesic drugs are combined, the additive effect is better control of pain (compared with that resulting from the use of either drug alone).

TEXT MINING

Text mining can be defined as an automated or semi-automated means for exploiting and analyzing huge sets of textual data to discover knowledge. Text mining is different from data mining as data mining takes into account structured data usually data from databases, while text mining works on unstructured data for example data from websites or scientific articles. Today, almost all-available information is stored in the form of text and hence text mining has more commercial and scientific application than data mining. Text mining is a complex science since most of the inherited information is unstructured and unclear, hence different linguistic algorithms are required before performing actual analysis. Text mining is interdisciplinary and covers information retrieval, information extraction, clustering, web mining etc. Information retrieval and information extraction are the two disciplines that are mostly applied on biological data.

INFORMATION RETRIEVAL

Information retrieval is a science that deals with finding relevant data (mostly documents) from a large collection of unstructured data that satisfies the information need. Today, with the rising information load systems based on information retrieval are widely

used to facilitate retrieval of relevant information from both structured and unstructured (textual) databases. Web based information retrieval engines like Google and PubMed are widely used to access up to date technical and scientific information. Information retrieval systems allow quick retrieval of relevant documents, more search flexibility using Boolean operators and ranking based on relevance of the documents.

INFORMATION EXTRACTION

Information extraction is a science that deals with automated extraction of entities, relationships between entities and attributes that could well describe entities from unstructured data repositories [5]. More precisely, information extraction is a science that derives structured factual information from text. It is tougher than regular keyword searches and requires knowledge and understanding of sub-fields like natural language processing, named entity recognition and shallow linguistics to extract meaning full relationships.

MEASURES DETERMINING ACCURACY

In information retrieval and information extraction precision, recall, and F-measure are measures to determine the accuracy of document retrieval and factual relations respectively. Recall is defined as the number of documents retrieved by the system in response to a query. While precision is defined as, the number of relevant documents retrieved by the system in response to a query or it the system's ability to present only relevant document from the retrieved documents. F-measure is a harmonic mean of precision and recall.

1.2 OBJECTIVES AND CONTRIBUTION

The PD literature has grown enormously over the years, but is dispersed across many journals. Thus, it is challenging to identify important PD parameters and their associations. Drug dosage plays a key role in most PD based studies, identification of optimal drug dosages is important for clinical applications of the drug. Drug interaction is another PD parameter whose identification is crucial to avoid adverse drug reactions [6]. Most PD information is present in descriptive full texts rather than in abstracts. Manual investigation of this information from scientific article is slow and expensive and also depends upon the availability of full text to the researchers. Thus in order to avoid manual data investigation and to facilitate automated data retrieval, text mining system could be applied to PD domain.

This thesis attempts to develop an *in vitro* PD TM system, which is capable of scanning full text articles and abstracts to retrieve PD DDI automatically. Before the development of any actual TM system, a centralized knowledge representation of PD concepts was crucial. Thus, an *in vitro* PD ontology was developed first and then the TM system. The proposed TM system is an information retrieval and extraction system, which can efficiently locate sentences mentioning DDI, drug concentrations, cell lines, and cellular phenotypes. The system is powered by a web interface capable of accepting user input in the form of drug names, DDI types, cell line types, and cellular phenotypes. Presence of ontology surpasses the simple keyword based ability of regular search engine and allows categorical searches. To normalize different drug naming conventions this thesis also

proposes a drug dictionary to capture available generic names, synonym, and brand names of the drugs.

The system processes full text articles and abstracts in which sentences are first split individually and then terms are tokenized to form individual units. The tokenized terms are annotated following entries from both the drug dictionary and the *in vitro*PD ontology. Term annotation with ontology and drug dictionary allows formulation of semantic queries, thereby increasing the precision and recall of the system. The thesis also presents a semantically annotated corpus to serve as a gold standard for future PD text mining.

The contribution of this thesis is threefold:

1. Information retrieval and extraction system: It allows retrieval of relevant articles and extraction of PD DDI, drug concentrations, cell lines, and phenotypes.
2. In vitro PD ontology: The ontology presents drug interaction, cell line, cell phenotypes, and data analysis classes. The ontology is developed with an intention to define the entire *in vitro* PD domain.
3. Semantically annotated corpus: An annotated corpus defining drug interaction pairs, drug interaction type, drug concentrations, cell line type, cell phenotype etc. The term level annotation is automated but the sentence level annotation is manual. At sentence level drug interaction were classified into true, ambiguous and non-DDI.

1.3 OUTLINE

Chapter 2 discusses the related work that covers other text mining system, existing ontologies, and drug repository. Development of the *in vitro* PD ontology is the backbone of the TM system and the annotated corpus, chapter 3 presents the data collection, data integration, and data processing for the ontology. Chapter 4 depicts the steps required for developing the database of annotated full text corpus and algorithms used for entity recognition. Chapter 5 describes the procedure for development of the annotated corpus. Chapter 6 presents an overview of the web tool. The validation and evaluation of the system is described in chapter 7, while chapter 8 concludes the thesis and proposes outlook for future extension and improvements.

CHAPTER 2

RELATED WORK

In this chapter, we present the current state-of-the-art and provide overview of current Drug repositories, biological TMsystems, and bio-ontologies. Finally we explain how the knowledge and information from these Drug repositories, TM system ontologies was gathered to develop new *in vitro* pharmacodynamics ontology and an information retrieval/extraction system.

2.1 BIOLOGICAL INFORMATION RETRIEVAL AND EXTRACTION SYSTEMS

Information retrieval is a science responsible for retrieving a pertinent subset of documents from pool of other documents. Today, the best-known biomedical IR system is PubMed; this IR functions with Boolean operators, vector space model and MeSH terms. The Boolean logic allow users to retrieve relevant documents using Boolean operator like 'AND', 'OR', 'NOT', using Boolean model the search is more customized to get user desired results. While use of Vector space model and MeSH terms assist the scoring and indexing of all terms in PubMed documents respectively. Literature databases like PubMed only contain bibliographic information and abstracts. They often suffer from the constraint of information compression and convolution imposed by a word limit. Thus, access to the full text articles is critical for sufficient coverage of facts and knowledge in the literature and for their retrieval [7]. The section below discusses two full text information retrieval/extraction systems.

2.1.1 TEXTPRESSO

Textpresso [8] is a new text mining system whose capabilities are far beyond the regular keyword based search systems. Textpresso [8] is powered by two major elements viz. collection of full text articles, which are split into individual sentences, and presence of 33 categories of terms in the form of ontology. The ontology categories include classes describing biological concepts e.g. gene, allele, phenotypes etc. and classes that relate two objects e.g. association, regulation etc. Individual terms from the text corpus are annotated with specific XML tags referencing the entries from the classes of the ontology. A search engine is built upon the annotated corpus, which can be queried using combination of keywords and categories, as the ontology covers term meanings it is possible to formulate semantic queries. Access to full text and presence of ontology is responsible for improved precision and recall of the system. Textpresso [8] is built using *Caenorhabditiselegans* literature with an article size of 3,800 full texts and 16,000 abstracts, it also comprises of around 14,500 entries from the ontology. Textpresso [8] is now used a curation tool and a search engine for researchers, it is also extended to other organism-specific corpora of text. Textpresso [8] can be accessed www.textpresso.org.

2.1.2 PHARMSPRESSO

Pharmspresso [9] was built on the lines of textpresso [8], it is a text mining tool designed for extraction of pharmacogenomics concepts and relationships from full text. Pharmacogenomics studies relationship between genetic variations and drug responses, this field is crucial as it plays an important role in defining drugs for an individual or for a subgroup of population (personalized medicine). In order to manage and organize

evergrowingpharmacogenomics literature wealth, pharmspresso [9] was developed. Pharmspresso [9] has the ability to parse text to find references to human genes, polymorphism, drugs and disease and their relationship (Yael Garten et al 2008). Just like textpresso [8], pharmspresso [9] involves the use of full-text articles and ontology and hence it is capable of formulating semantic queries. Currently, pharmspresso [9] is an important curation tool for researchers in the field of pharmacogenomics and in many cases has helped in validating results from literature. Pharmspresso [9] is a free tool and can be downloaded from <http://pharmspresso.stanford.edu>.

Other than textpresso [8] and pharmspresso [9] other text mining tools also offer semantic concept based information retrieval, this tools include Relemed [10], iHop [11] etc. These tools are limited to abstracts and miss relationships that are stated in the full text.

This thesis utilizes the idea from both pharmspresso [9]andtextpresso [8] to build a TM system that could be applied to the pharmacodynamic domain, tofacilitate automated retrieval of sentencesdescribing drug-drug interactions, drug concentrations, cell lines, and possible cell phenotypes.

2.2 CELL LINES

2.2.1 CELL LINE ONTOLOGY (CLO)

A cell line is a colony of cells that is artificially developed and grown under controlled condition [12]. A cell line is typically derived from normal tissues or diseased tissues; it is maintained as both *permanent* and *primary* cell lineages depending upon renewable and non-renewable purposes respectively. Cell lines are common tools for most *in vitro* experiments especially in pharmacology to understand the effect of drugs on certain tissues. Cell lines provide an easy platform for studying cellular mechanisms that may suggest new potential drug targets and, in the case of pathological-derived tissue; it has an interesting application in the evaluation of therapeutic agents that potentially may treat the dysfunction [13].

Cell line ontology (CLO) [12] is a classification of cell lines focusing on permanent cell lines from culture collection. It houses around 36,000 cell line names, 194 cell types, 656 anatomical and 217 organisms. Prior to the development of CLO [12], the same group created Cell line Knowledgebase (CLKB) [14], which is a standardized catalogue of cell line data from commercial repositories like ATCC [15] and HyperCLDB [14]. CLKB [14] was further redesigned to CLO [12] by addition of around 27000 more cell lines from the European bioinformatics institute Coriell catalogue ontology [16] and Bioassay Ontology (BAO) [17]. Apart from cell lines CLO [12] also imported the whole Basic Formal Ontology (BFO) [18] to form the upper level of the ontology and the Relational Ontology (RO) [19] to form its core, external ontologies like NCBI_taxon and Cell Type ontology [20] are also integrated to CLO [12] using Ontofox [21]. Finally, CLO [12] is developed using

protégé(<http://protege.stanford.edu/>) and follows the OBO Foundry principles. Table below provides a tabular and graphical interpretation of ontologies and terms used in CLO [12].

This thesis utilizes data from CLO [12] and combines it with other catalogues like Australian cellbank [30], Breast tissue [31] cell line ontology and Leibniz institute DSMZ-German collection [32] of Micro-organisms and cell cultures.

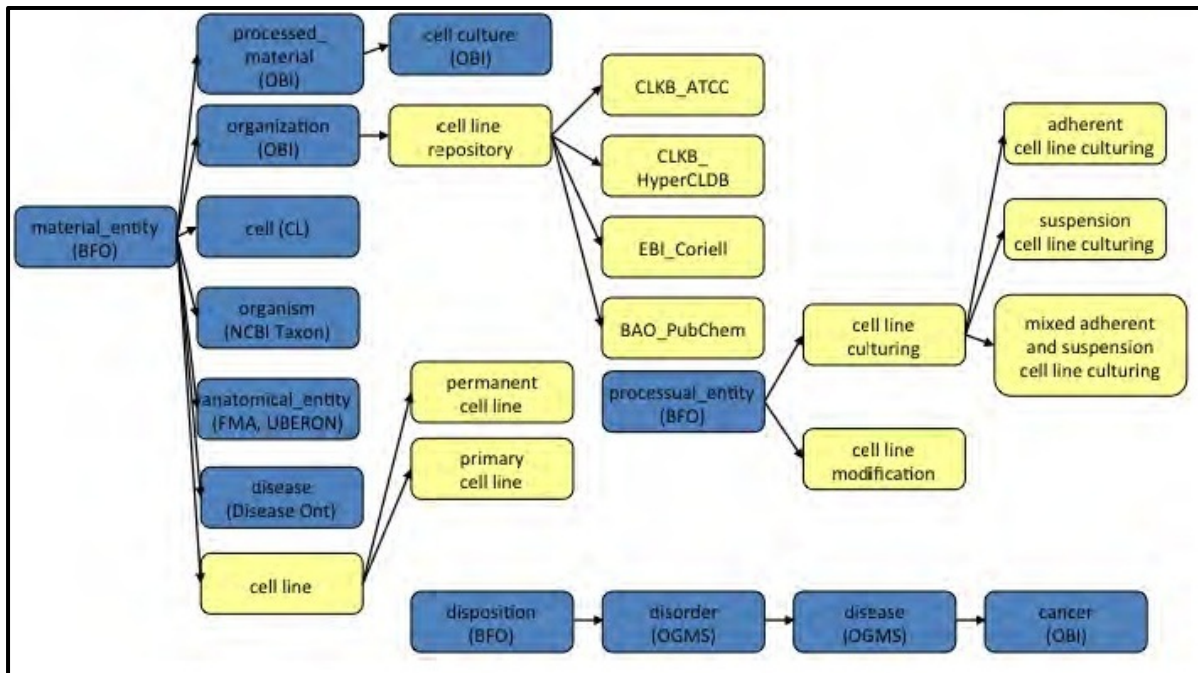


Fig.2 CLO hierarchical structure of ontology terms. Blue boxes presents the External classes while yellow indicate the terms with CLO ID's

2.3 DRUG DATABASES

2.3.1 DRUGBANK

DrugBank [22] is a standard drug repository that includes drug and drug target information. It provides extended information on pharmacokinetics, metabolism, pharmacology, pharmaceuticals, nomenclature, ontology, structure, and action of drugs. It is a unique bioinformatics/chemo-informatics repository with coverage of 6711 drug entries, these entries includes 1447,131, 85 and 5080 FDA-approved drug molecules, FDA-approved biotech drugs, nutraceuticals and experimental drugs respectively. Drug bank [22] provides additional drug attributes in the form of drug cards; each drug card includes more than 80 fields that describe most of the drug/chemical and drug-target/ drug-protein information. For the current study, a drug dictionary was developed from the drug cards downloaded from <http://www.drugbank.ca/downloads>.

CHAPTER 3

METHODS

3.1 DATA COLLECTION

Identification and collection of high quality data was a crucial step for developing the ontology, drug dictionary and the corpus. The datasets were collected from both public and commercial repositories. For developing the drug dictionary, drug cards were downloaded from <http://www.drugbank.ca/downloads> of the DrugBank [22] database. The *in vitro* PD ontology construction included data collection both manually and from repositories. The manual data collection was performed for drug interaction type and method classes of the ontology, while other downloads includes classes from Gene ontology [29] and from repositories like ATCC [15], HyperCLDB [14], Cell line Knowledge base [14], Coriell cell line repository [16], Bioassay ontology [17], Australian cell line repository [30], DSMZ [32] data catalogue and breast tissue ontology [31].

3.2 DRUG DICTIONARY CONSTRUCTION

The biomedical text is highly enriched with drug entities, identification and extraction of these entities has always been a challenge due to different naming conventions used by researchers. To solve the drug convention issue, we propose a drug dictionary that tries to capture most available drug name variants. The proposed drug dictionary is formed by extracting different drug annotations from drug cards of the DrugBank[22] database, drug cards includes information like drug description, generic name, brand name, synonym, IUPAC names, structures etc. In order to develop a standard drug dictionary a Perl script was written to extract drugs generic name, brand name and

synonyms. The drug dictionary includes over 6700 generic names and one or more brand names or drug synonyms per generic name.

Table represents few term of the drug dictionary.

DB_ID	Drug generic name	Synonym	Brand names
DB00091	Cyclosporine	Cyclosporine A	Restasis, Genraf, Neoral, Sandimmune, Sangcya
DB00115	Cyanocobalamin	Vitamin B12, Cyanocob(III)alamin	Bedoz, Cobex, Cobolin-M, Crystamine, Cyomin, Depinar.
DB00104	Octreotide	Octreodiam, Octreodita	Atrigel, Longastatin, Sandostatin.
DB00165	Pyridoxine	Pyridoxol, Vitamin B6	Aderoxine, Alestrol, Becilan, Benadon, Beesix.
DB00153	Ergocalciferol	Vitamin D2	Condol, Calciferol, Crtron, Bucu-D, Crystallina, D-Arthin, D-Tracetten

Table. 1 Drug dictionary with Generic names, synonyms, Brand names

3.3 CONSTRUCTION OF THE *IN VITRO* PD ONTOLOGY

The *in vitro* PD ontology is an effort to represent knowledge as a list of concepts that could describe important components of *in vitro* PD experiments and relations between concepts. The proposed ontology includes classes like drug interaction, analysis methods, cell lines and cell phenotypes. The ontology was developed using protégé and follows the principles of the OBO foundry. Each entry in the ontology has information about the source and definition in case of upper classes. The development of this ontology will open new research avenues for future text mining in this domain.

3.3.1 DRUG INTERACTION CLASSES

Manual investigation of PD related studies and review articles have confirmed that PD drug interactions fall into synergistic, additive, and antagonistic categories. Synergistic interaction is a term used to suggest a drug combination whose effect is greater than expected, antagonistic interaction is a one in which the effect is less than expected; while additive drug interaction is a one in which the effect is just what is expected. The terminology of the drug interaction is consistent among majority of the articles, but few articles have proposed synonyms like positive, supra-additive, potentiation and augmentation for synergistic drug interactions; negative, infra-additive or sub-additive, negative synergy for antagonistic drug interactions; while zero or null interaction for additive drug interactions [23]. Considering the drug interaction types and different naming convention the drug interaction class was developed.

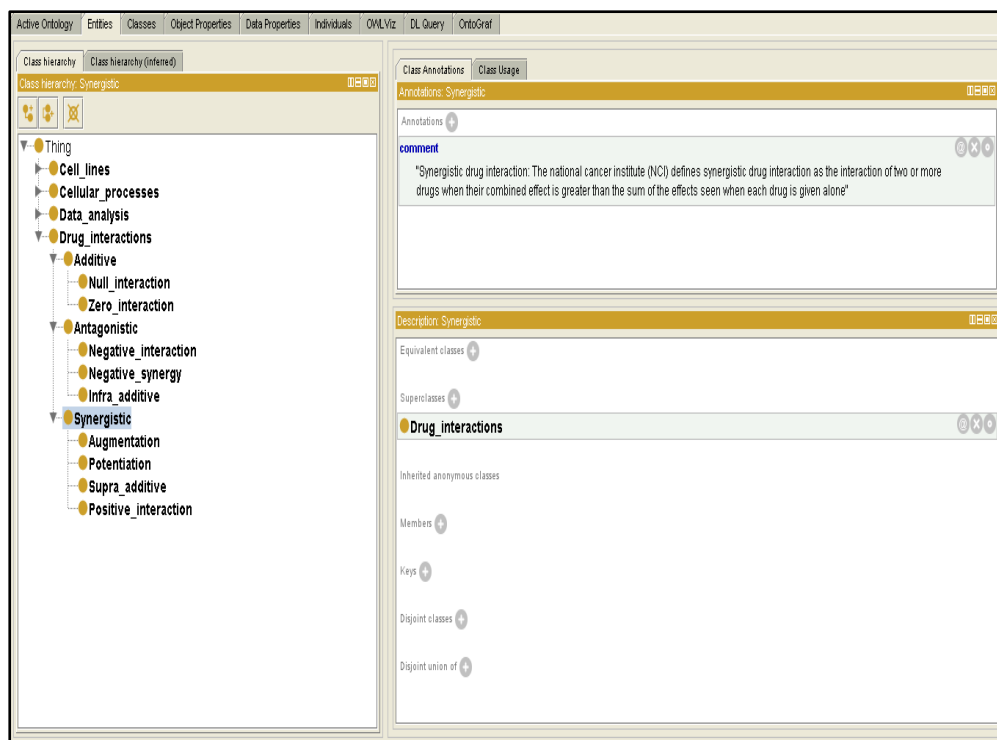


Fig.3 *in vitro* PD ontology, showing DDI classes

3.3.2 DRUG INTERACTION ANALYSIS

Determining synergy, additivity, and antagonism of a drug combination is crucial for detecting efficacy, toxicity, optimal doses and for minimizing the drug resistance. The literature presents different analysis methods for determining drug interactions, many of these methods and algorithms are implemented in the form of software's. For the ontology the analysis methods were added by manual inspection of abstracts and full texts, few of these analysis methods include isobolographic analysis, combination index, median drug effect analysis, Yonetani-Theorell plot, Chou-Talalay, Dixon up-down method, Loewe additivity etc.

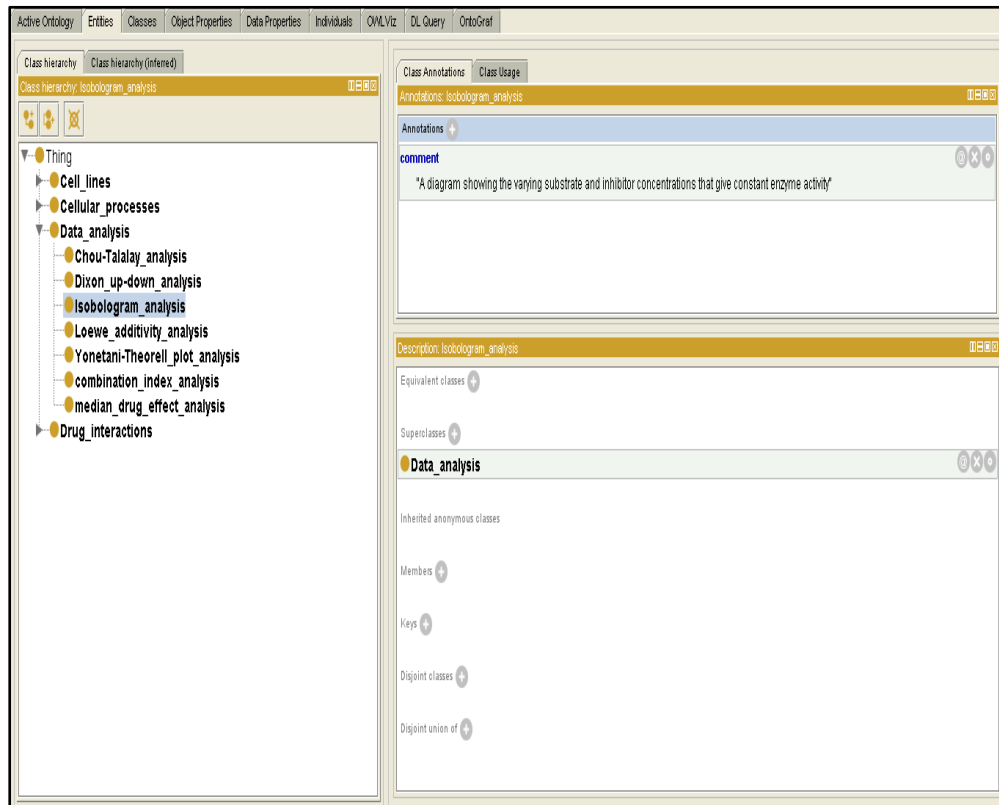


Fig.4 *in vitro* PD ontology, showing DDI detection classes

3.3.3 CELL LINE CLASS

3.3.3.1 HUMAN ORGAN SPECIFIC

Primary tissue cultures are a valuable source for *in vitro* studies, however due to scarcity of suitable samples; phenotypic instability and unavailability of fresh tissue sample for harvesting purposes have hindered the widespread use of tissue cultures. Thus to overcome these limitations cell lines models have been proposed. Cell lines are known to be the back bone of most *in vitro* experiments; literature describes them as a tool or model for analysis of drug metabolism, drug screening and toxicity studies[26][27]. The knowledge of different cell lines is dispersed across different scientific journals; many academic commercial repositories have tried to provide a standard cell line catalogue but have failed to include complete cell line coverage. Recently the cell line ontology (CLO) [12] has tried to capture all the available cell lines and successful to a certain extent, CLO includes roughly around 36,000 cell line entries from providers like Cell line Knowledge Base (CLKB) [14], European Coriell cell line ontology [16] and Bioassay ontology (BAO) [17].

CLO [12] categorizes cell lines in the form of cultures i.e. primary, permanent or continuous cell line cultures. To perform text mining analysis this categorization is irrelevant with respect to annotation and user query, as a user might have a hard time remembering the names of cell lines associated with the above mentioned categories. In order to address this issue, the data from CLO [12] was re-mapped to their corresponding tissues and organs, Since CLO [12] does not provide tissue information the data was recollected from CLKB [14], Coriell ontology [16] and BAO [17]. CLKB, Coriell

ontology[16]and BAO[17] provide organism, tissue and organ information for all cell line entries, if the organ information was not present the tissue was mapped to the organ using tissue ontology [28]as a reference. The human organ specific cell line class with organ specific classes was thus formed by mapping the tissue information to organs.

Additionally the ontology was also enriched with more cell line entries from Australian cell line repository [30], breast tissue cell line ontology [31], and DSMZ [32]. The Australian cell line repository is a national facility that provides cell lines and related services in Australia and other parts of the world. A written permission was obtained for use and downloaded from www.cellbankaustralia.com/Catalogue/default.aspx, this source includes around 34 cell lines. The Deutsche Sammlung von Mikroorganismen und Zellkulturen - DSMZ - (German Collection of Microorganisms and Cell Cultures) [32] is a non- profit organization in Germany responsible for preservation and distribution of human and animal cell lines, plant cell cultures. A cell line catalogue in the form of an excel file was downloaded and includes around 715 cell line entries. Similarly around 180 breast cell lines were extracted from the breast cell line ontology [31]. After combining data from all the sources and removal of possible duplicates, the total number of cell lines the ontology has is around 32,000. (Note: The data supplied by the Coriell ontology had many duplicate entries)

Currently the human organ specific cell line ontology cover 48 organs which includes Lymph , Brain , Blood , Bone , Skin , Breast , Eye , Lung , Liver , Muscle , Kidney , Cervix , Adrenal , Amnion , Urinary Bladder , Umbilical Cord , Uterus , Peritoneum , Heart , Pancreas , Artery , Intestine , Salivary Gland , Vagina , Gonad , Ureter , Urethra , Mouth

,Pelvis , Spleen , Prostate , Placenta , Pharynx , Thymus , Nervous System , Connective Tissue, Thyroid , Stomach , Fetus , Tonsil , Vulva , Embryo , Larynx , Fetus , Vein , Tumor.

Organ name	Tissues present	Sample Cell lines	Total number
Breast	Nipple, Mammary gland	MCF-7, MFM-223	117
Skin	Foreskin, Scalp, Epidermis	HFF1, Detroit 532	2256
Eye	Retina, Cornea, Lens	ARPE-19, HCE-2	8
Kidney	Glomus, Cortex.	Glomotel, HK-2	47

Table. 2 Represents few Organ with their tissues, sample cell line and cell line count

3.3.3. 2 ANIMAL SPECIFIC

All the above mentioned sources also provide information about organisms other than human;these organisms contribute to the animal cell line class. Animals include Camel, Cat, Fish, Chicken, Hamster, Cusimanse, Deer, Dog, Dolphin, Donkey, Drosophila, Duck, Ferret, Fox, Frog, Rat, Monkey, Goat, Goose, Guineapig, Horse, Iguana, Insect, Lizard, Marsupialmouse, Mice, Minipig, Mink, Mongoose, Mosquito, Mouse, Muntjac, Opossum, Orangutang, Oryx, Sheep, Parakeet, Pig, Quail, Rabbit, Raccoon, Snail, and Squirrel.

Organism	Cell line	Cell type
Murine (mouse)	NS20Y, THB-5, H25B10	Hybridoma
Rat	JTC-15, JTC-27, SGE-1	Hepatoma
Mouse	CAB117-12D10, PAb100	Lymphoblast

Table. 3 Showing Organism corresponding cell line and cell type

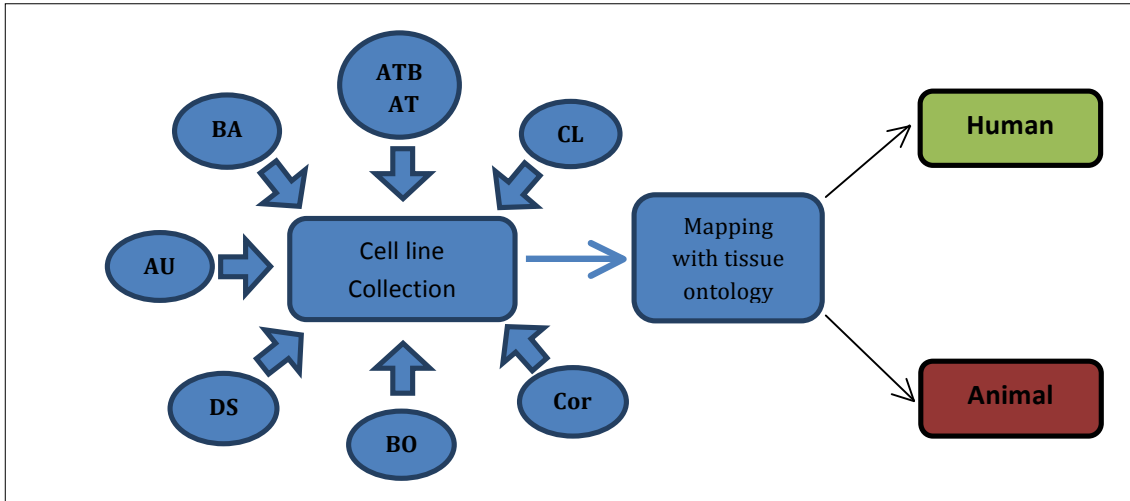


Fig.5 Data collection for the cell line ontology class.

BA: Bioassay ontology. AT: American Type culture collection. CL: Cell line database. AU: Australian cellbank
 DS: Deutsche Sammlung von Mikroorganismen und Zellkulturen. BO: breast cell line ontology.
 Cori: Coriell Cell line ontology.

3.3.4 CELL PHENOTYPES

Drug interactions experimented on cell lines produce a number of cell phenotypes, these phenotypes are important to determine both the drug interaction and optimal drug doses. Our literature investigation has revealed a number of cell phenotypes that includes cell death, cell growth, cell proliferation etc. Many articles represented cell death with terms like apoptosis, programmed cell death, while cell aging with terms like cell senescence. Thus, to ensure that the ontology covers most term synonyms and term variants, few subclasses of the cellular process category of the gene ontology [29] were imported. These subclasses represent cell death, cell aging, cell division, cell cycle etc. which includes terms like apoptosis, necrosis, cornification, senescence, cytokinesis, isotrophic growth respectively. A Perl script was written to extract these classes and retains other information like descriptions, relationships etc.

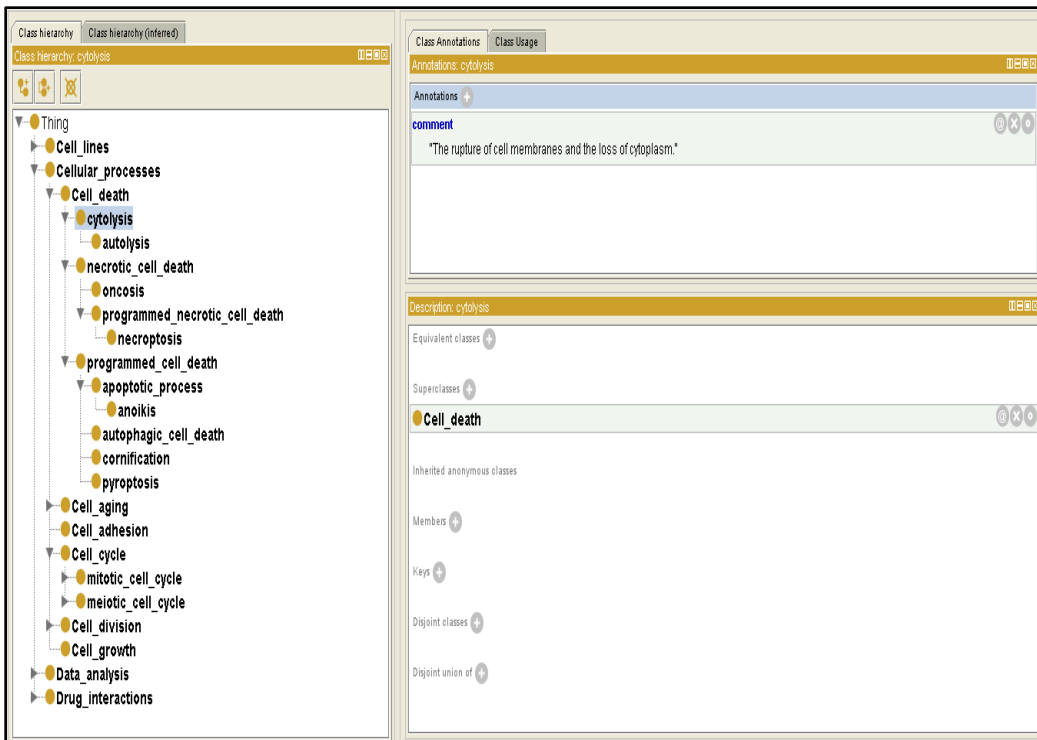


Fig.6 in vitro PD ontology, cellular process classes

Cellular phenotypes	Classes	Subclasses
Cell death	Necrotic cell death	Oncosis, necroptosis
Cell cycle	Meiotic cell cycle	G1 Phase, M Phase
Cell aging	Isotrophic growth	--

Table.4 showing sample cellular phenotypes with classes and subclasses

CHAPTER 4

DATABASE OF ANNOTATED FULL TEXTS

4.1 AUTOMATED DOWNLOAD OF FREE FULL TEXTS FROM PUBMED.

To determine pharmacodynamically relevant articles, PubMed was searched to determine literature reported drug combinations. Combinations of keywords like “synergistic drug combinations”, “synergistic drug interactions”, “additive drug combinations”, “additive drug interactions” etc. were used. These keywords were manually investigated to check the relevancy of the retrieved articles, the above combinations retrieved huge amount of results and hence the manual download was a challenge. To facilitate automated download of articles from PubMed, our in-house PDF tool developed by students at Lang Li Lab was used to download articles. We collected around 502, 147, and 63 articles published in last 5 years that were pharmacodynamically synergistic, additive and antagonistic respectively. Duplicate articles were later removed and a total of 620 articles and 1270 abstracts for which full text articles were not provided were used.

We discuss the necessary data preprocessing from sections 4.1.1 onwards.

4.1.1 CONVERSION OF PDF TO TEXT

Recently PDF's have become a defacto source for information transfer and sharing; they provide high quality graphical display and allow printing information in standard and convenient manner. Most programming languages have developed libraries and tools to parse and process information from PDF documents but the performance of these libraries has always been an issue, also majority of the available tools are either commercial or allow

limited articles to get processed. In order to ensure overall data processing, programming languages offer conversion of PDF's to formats like XML and text. In this study, we have used an open source software package, XPDF (<http://www.foolabs.com/xpdfv>) [33], to convert PDF files to text. The software package is a command line module which was automated using Perl system commands to perform batch conversion of PDFs to text.

```
use strict;

use warnings;

my $execute;

my $prog = "C:\\xpdfbin\\xpdfbin1\\bin64\\pdftotext.exe";

my $Directory="C:\\xpdfbin\\xpdfbin1\\bin64\\Files";

foreach my $fp (glob("$Directory/*.pdf"))
{
    if (-f $prog) # does it exist?
    {
        system("$prog",$fp);
        print "Will perform conversion\n";
    }
    else
    {
        print "$prog doesn't exist.";
    }
}
```

Fig. 6 Perl script to automate the XPDF tool to perform batch conversions

4.1.2 FILTERING OF RAW TEXT TO TITLES, ABSTRACT AND FULL TEXT

The raw text files generated from the PDF's contained additional information like author names, emails, research institutes and finally references for each documents. This information is not important and was excluded prior to data annotation process and sections like titles, abstracts and remaining part of the text excluding references was extracted.

4.1.3 SENTENCE SEGMENTATION AND TOKENIZATION

Sentence segmentation otherwise called sentence boundary detection or sentence boundary recognition is an important step in most text mining and natural language processing operations; it assists to determine how entities in a text are to be processed. Sentence segmentation is the process of determining the longer processing units consisting of one or more words. This task involves identifying sentence boundaries between words in different sentences (<http://comp.mq.edu.au/units/comp348/ch2.pdf>) [33]. In order to achieve a higher accuracy of sentence segmentation Perl module `Lingua::EN::Sentence` was used. `Lingua::EN::Sentence` module is free software package, which was downloaded from the Comprehensive Perl Achieve Network (CPAN). The `Lingua::EN::Sentence` module uses the `get_sentences()` function, which splits text into individual sentences based upon set of rules and in built regular expressions. After sentence segmentation, tokenization is the next process, which is defined as a process of chopping text into individual pieces, called tokens, most tokenization processes involves removal of the punctuation character class. In the present study, tokenization was achieved by adding white spaces across punctuations.

The raw text articles were processed using Perl Lingua::EN::Sentence module and individual sentences were recognized, after segmentation text was chopped into individual tokens. Steps until sentence segmentation and tokenization cover most of the data pre-processing part.

4.2 SENTENCE LEVEL ANNOTATION

The following section covers the procedure for sentence level annotations

4.2.1 NAMED ENTITY RECOGNITION AND ANNOTATION WITH TERMS OF ONTOLOGY

Named entity recognition (NER) is a step in information retrieval that is assists in locating and classifying entities into desired categories. In the presented study, entities were categorized into drug, drug interaction, cell phenotypes, and cell lines types. The identified entities were annotated (marked-up) with specific XML tags. The NER algorithm was build using hashes and necessary regular expression; it can accurately identify and annotate entities along with their variants.

Category	Sample entity	Total number identified
Interaction	Additive, synergistic	2784
Cell lines	Breast (MCF-7), Rat(AR42J)	5399
Phenotypes	Cell death, proliferation	3679

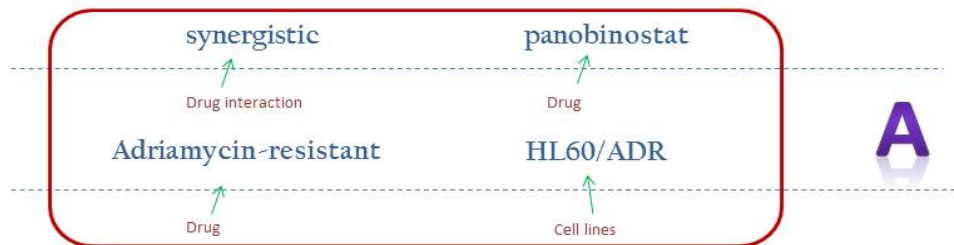
Table.5 shows the number of interaction, cell lines, and cell phenotypes from the text corpus.

4.2.2 NAMED ENTITY RECOGNITION AND ANNOTATION WITH TERMS OF DRUG DICTIONARY

The development of the drug dictionary is an important contribution from this thesis; the proposed dictionary is capable of identifying drug entities not only with their generic names but also with brand names and synonyms. All terms in the text are stemmed to their basic form before processing with the drug dictionary, to achieve an accurate stemming Lingua::Stem algorithm from CPAN was used. The entity recognition algorithm works on stemmed terms and has sets of regular expressions which could easily markup variants of terms.

PubMed: 22863538

“In this study, we investigated the synergistic effects of panobinostat and bortezomib on Adriamycin-resistant HL60/ADR cells and refractory acute myelogenous leukemia (AML) primary cells”



`<drug interaction type = synergistic>synergistic </drug interaction> effects...`

Annotation

`<Sentence id =1>In this study, we investigated the <drug interaction type=
"synergistic">synergistic </drug interaction>effects of <drug name=
"panobinostat"> panobinostat</drug> and <drug name = "bortezomib">
bortezomib</drug> on <drug name = "adriamycin"> adriamycin-resistant</drug>
<cell line type = "Blood"> HL60/ADR</cell line> cells and refractory acute
myelogenous leukemia (AML) primary cells </Sentence>`

The Complete Mark-up or annotation process

The sample sentence “In this study, we investigated the synergistic effects of panobinostat and bortezomib on adriamycin-resistant HL60/ADR cells and refractory acute myelogenous leukemia (AML) primary cells” is taken from **Jiang XJ et. Al** 2012 (Pubmed:22863538)

(A) The Entity recognition algorithm matches terms from the sentence to the terms in the ontology, for every match the term from the sentence is annotated. The annotation is in the form of XML tags with brief information about the tag.

(B) A fully annotated sentence contains tags that classify terms with the respect to categories in the ontology. In many cases the annotation is performed with subclasses of the ontology. E.g. the terms HL60/ADR is annotated with the cell line subclass “Blood”.

CHAPTER 5

DEVELOPMENT OF A SEMANTICALLY ANNOTATED *IN VITRO* PD CORPUS FOR TEXT MINING

Text mining systems have become indispensable for biomedical research, while Natural Language processing (NLP) has raised the TM accuracy and performance. NLP techniques are relatively domain portable but lack of a domain specific reference material has always been a major issue for validation of the TM accuracy. Most reference materials are available in the form of annotated corpuses, which are expert curated and serve as a gold standard for evaluation. Different domain specific corpuses like the GENIA [35] corpus have been suggested in the past and recognized as highly quality reference material. In this chapter we try to take advantage of the *in vitro* PD ontology and the drug dictionary to propose a semantically annotated corpus for PD TM based upon the GENIA guidelines.

5.1 IN VITRO PD CORPUS

The *in vitro* PD corpus is a collection of abstracts extracted from the PubMed database, the current release consists of 150 abstracts that are semantically annotated in reference to the domain ontology and dictionary. The term annotations include drug names (including generic name, brand names, and synonyms), drug concentrations, drug interaction type, cell phenotypes, cell lines, and methods used to determine the drug interactions. Apart from the term level annotations the current release also presents sentence level annotations; the sentence level annotation is a manual process that classifies PD drug interactions into true DDI, ambiguous DDI and non-DDI types.

5.1.1 CORPUS ANNOTATIONS

The corpus annotations are based upon the domain ontology and dictionary. The development of the ontology and the dictionary is already mentioned in chapters 3. The corpus is enriched with important concepts and classes that include drug names, human and animal cell lines, cell phenotypes, drug interaction type, drug concentration (numbers), and DDI identification methods.

- Drug names where extracted from drug cards of the DrugBank[22] 3.0 release. The drug dictionary is based upon this release and annotates most drug names in the corpus.
- Cell lines cover both human and animal cell lines. The human cell lines are categorized organ wise and include 48 organs, while the animal cell lines are categorized based upon animal names and covers animal types.
- Cell phenotypes include classes like cell death, cell growth, cell division, cell proliferation, and cell aging. The annotation schema is based upon above mentioned classes and their respective subclasses.
- Drug Interactions include synergistic, additive, and antagonistic drug interactions with their synonyms like supra-additive, null interaction and infra-additive respectively.
- Drug concentration includes drug dosages and concentration.
- Units includes different units used to measure drug concentrations.
- Methods these include names of most software and data analysis techniques that are used to identify and classify DDI's in synergistic, additive and antagonistic types.

NOTE: All the above terms were stemmed to their basic form using Perl Lingua::Stem module (Downloaded from the CPAN repository) before the entity recognition and annotation procedure.

5.1.2 PD TAGGER

The *in vitro* PD abstracts were annotated using a Perl implemented PD tagger. The tagger tagged the abstract entities following entries in the drug dictionary and the domain ontology. The tagging work flow can be divided into 3 steps: 1) Sentence boundary detection achieved using Lingua::En::Sentence module. This module is written in Perl and performs sentence segmentation with a higher accuracy. 2) Tokenization was based upon simple white space tokenization approach, which allowed separation of words into individual units. 3) Named Entity Recognition (NER) was the step where the actual entity markup was performed. The mark up algorithm uses a set of hashes and regular expression to tag tokenized entities.

The PD tagger tags the following entities from the domain ontology and the drug dictionary: 1) Drug names. 2) Drug interactions. The drug interactions include synergistic, additive, and antagonistic drug interactions. 3) Cell lines. For example breast cell lines like MCF-7, HCC1569, MCF10A, MDA-kb2 etc. 4) Cell phenotypes. For example necrosis, apoptosis, cornification, pyroptosis etc. 5) Number. To represent drug concentration or cell line sample size. 6) Methods. For example Isobolographic analysis, combination index, median drug effect analysis, Yonetani-Theorell plot, Chou-Talalay etc.

The PD tagger processes MEDLINE abstracts and tags entities based upon the domain dictionary and ontology. The mark up is based upon XML tags. The tagger is also

capable of performing part-of-speech tagging based upon `Lingua::EN::Tagger` (<http://search.cpan.org/~acoburn/Lingua-EN-Tagger-0.23/Tagger.pm>) downloaded from CPAN repository.

5.1.3 SENTENCE LEVEL ANNOTATIONS

Sentence level annotation is the next level of annotations performed after term level annotations. Sentence level annotation is a manual process and classifies drug interaction sentences into true DDI, ambiguous DDI, and non DDI.

- *True DDI Sentences* (TDDIS): Two drug names are in the sentence stating clear interaction and its type. (Synergistic, additive and antagonistic).
- *Non DDI Sentences* (NDDIS): Two drug names are in the sentence with clear negation suggesting that drugs do not interact.
- *Ambiguous DDI Sentences* (ADDIS): One drug is absent in the sentence, but DDI could be inferred from the context. A clear interaction statement is essential.

5.2 GENIA FORMAT

The presented corpus is structured based upon the GENIA [] guidelines and uses XML based mark up schema. Each article of the PD corpus contains its MEDLINE ID, title and abstract, sentences in the abstract are segmented into individual sentences. The XML output of each article can be visualized using a cascade style sheet (CSS) in a browser. Each XML tag is color coded with specific entries from the CSS. The resulting corpus configuration is depicted in figure. 7. The browser based visualization is presented in Fig. 8.

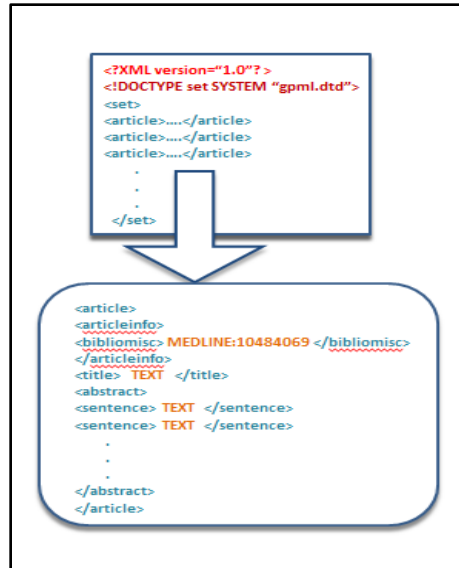


Fig. 7 Configuration of PD corpus

Drugs Cell Lines Interaction Cell Phenotypes
Non-Drug Interaction Sentences Ambiguous-Drug Interaction Sentences
True Drug Interaction Sentences

Medline:10484069

Synergistic effect of estramustine and [3-keto-bmtl]-[val2]-cyclosporine (psc 833) on the inhibition of androgen receptor phosphorylation in lncap cells.

Estramustine phosphate has been used frequently alone or in combination with other drugs for the treatment of hormone-refractory prostate cancer. Estramustine is one of the major active metabolites of estramustine phosphate in vivo. We recently demonstrated that estramustine acts as an androgen **antagonist**, and the combination of estramustine with [3-keto-bmtl]-[val2]-cyclosporine (psc 833) results in **synergistic** cytotoxicity. Unlike other regulators of microtubules, such as paclitaxel, the present study demonstrated that estramustine alone or in combination with psc 833 did not induce bcl-2 phosphorylation in **lncap** cells. **No synergism between estramustine and psc 833 in the induction of bcl-2 phosphorylation was obtained in mcf-7 cells exposed for 16 hr to estramustine [5-15 microm] and psc 833 [5 microm].** A significant **synergistic** antiandrogenic effect as measured by the inhibition of dihydrotestosterone-induced reporter gene luciferase expression in both wild-type and mutated androgen receptor (ar) cDNA-transfected **hela** cells was observed when the cells were exposed to estramustine and psc 833. Treatment of **lncap** cells with estramustine alone (5-15 microm) resulted in a decrease of ar expression and phosphorylation. This effect was enhanced markedly by psc 833. A strong correlation between ar phosphorylation and expression of the ar target gene psa was obtained in dihydrotestosterone-stimulated **lncap** cells. The up-regulated psa expression is a function of the level of the phosphorylated ar ($r = 0.9814$), but not the dephosphorylated form of the receptor protein ($r = 0.4808$). **Thus, our studies suggest that the synergism between estramustine and psc 833 in lncap cells is a consequence of inhibition of ar expression and phosphorylation, thus leading to interruption of ar-mediated gene expression.**

Fig. 8 Browser based display of the corpus

5.3 CONCLUSION

The in vitro PD corpus is highly enriched with PD annotations it also includes sentence boundaries, term boundaries, term classification and sentence level annotation. We hope that researchers utilize this resource in their research and expect to get their valuable feedback for further improvement of the corpus.

CHAPTER 6

ACCESSING THE WEB INTERFACE

The development of the *in vitro* PD miner was powered by providing a user interface; the user interface was developed using HTML, CSS, JavaScript, and PHP and is available at <http://rweb.biostat.iupui.edu/pdcorpus/retrieval/index.html>. The tool is capable of retrieving articles and extracting domain sentences from both full text and abstracts. The following chapter provides details about access and function of different options/tabs on the interface.

6.1 HOME

Pharmacodynamics Miner
A tool to assist automated retrieval and extraction of PD concepts and relations

Home About Browse Search DDI Corpus Download Contact

Enter drug's: Eg: Tamoxifen

Synonyms and brand names

Please select Drug interaction, Cell line type and biological processes from the following options

Center for Computational Biology and Bioinformatics, Indiana University, Indianapolis

Fig.9 Main page of the interface

Fig. 9 provides a description of the main interface, it includes tab options like browse, search, corpus download etc. The main interface provides input options where

user can provide drug names in the form of keywords, while drug interaction type, cell lines and cell phenotypes in the form of categories. As mentioned in last chapters this TM system is powered by drug dictionary and hence drug names can be searched in combination with the brand names and synonyms from the dictionary.

6.2 BROWSE

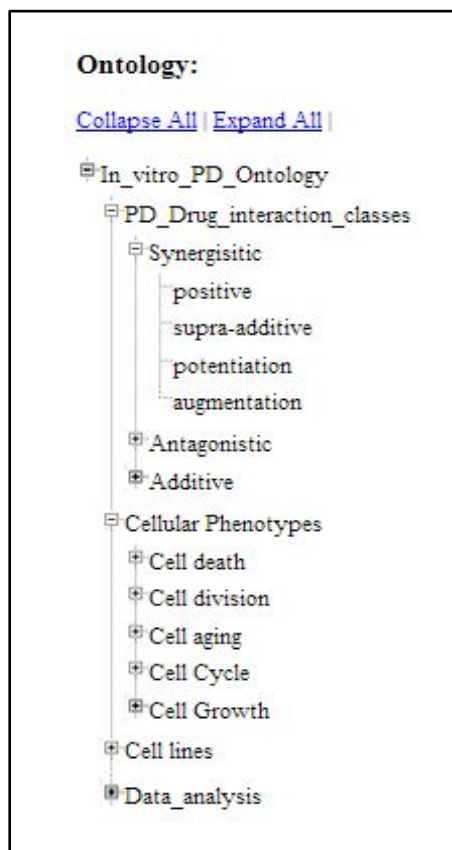
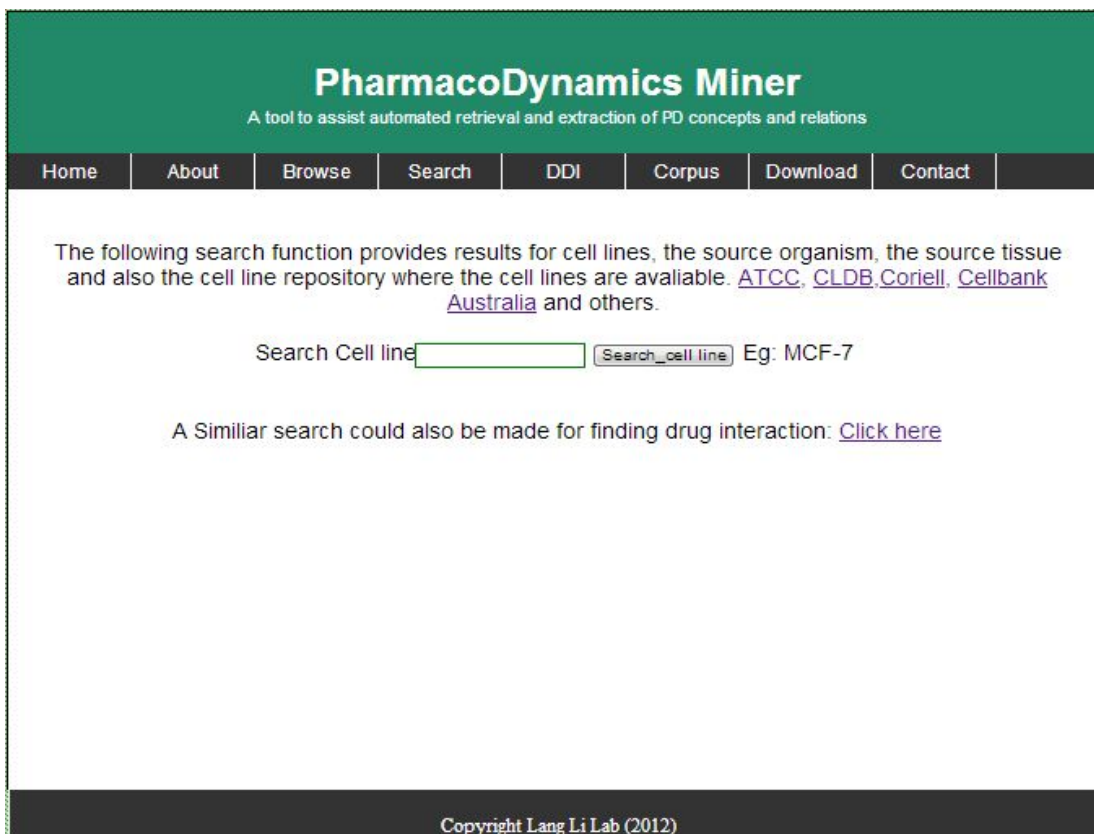


Fig.10 Online access to the ontology

Browse option is a web access to the ontology it is consistent to the ontology developed with protégé. The ontology covers classes like drug interaction, cell lines, cell phenotypes and data analysis methods with their respective subclasses. The web ontology

is developed using JavaScript, HTML and CSS, and can be accessed by clicking the browse button <http://rweb.biostat.iupui.edu/pdcorpus/retrieval/demo/Hello.html>

6.3 SEARCH



The screenshot shows the 'Pharmacodynamics Miner' website. The header is green with the title 'Pharmacodynamics Miner' and the subtitle 'A tool to assist automated retrieval and extraction of PD concepts and relations'. Below the header is a navigation menu with links: Home, About, Browse, Search, DDI, Corpus, Download, and Contact. The main content area has a white background and contains the following text: 'The following search function provides results for cell lines, the source organism, the source tissue and also the cell line repository where the cell lines are available. [ATCC](#), [CLDB](#), [Coriell](#), [Cellbank Australia](#) and others.' Below this text is a search form with the label 'Search Cell line', an input field, a 'Search_cell_line' button, and the example 'Eg: MCF-7'. Below the form is a link: 'A Similiar search could also be made for finding drug interaction: [Click here](#)'. At the bottom of the page is a footer with the text 'Copyright Lang Li Lab (2012)'.

Fig.11 Cell line search for PD Miner

Fig. 11 is the cell line search page, where users can query cell lines. The cell line database is enriched with around 32,000 cell line entries providing information about the cell line name, cell line source, cell subtype and the organism from which the cell line is developed. Sources like American Type Culture Collection (ATCC), Cell line database (CLDB), Coriell cell line ontology, Australian Cellbank, breast tissue cell line ontology and Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ). The cell line search is present at <http://rweb.biostat.iupui.edu/pdcorpus/retrieval/search.php>

6.4 CORPUS

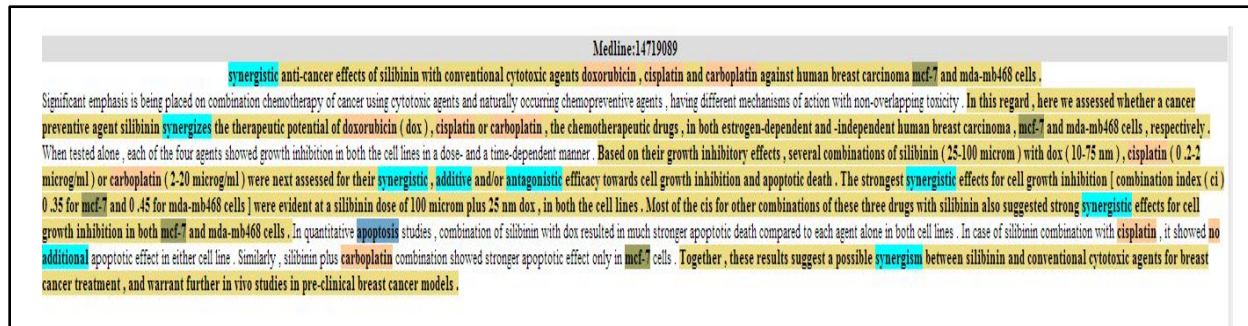


Fig. 12 Online view of the annotated corpus

Fig. 13 provides an overview of the semantically annotated corpus; the data is annotated to XML tags following the PD tagger as explained in chapter 5. Annotations include drug names, DDI, cell lines, cell phenotypes, numbers (to represents concentration) and data analysis technique used. Apart from term level annotation the current corpus release also includes sentence level annotations in the form of true DDI, non-DDI and ambiguous DDI. Every annotation is visually highlighted using CSS. The corpus is available at <http://rweb.biostat.iupui.edu/pdcorpus/retrieval/test.xml>.

6.5 DOWNLOAD

From the download tab information like protégé based ontology, semantic corpus, the tagging/annotation Perl code and Perl packages used for the system can be downloaded.

Chapter 7

RESULTS

7.1.1 PHARMACODYNAMICS MINER DATABASE OVERVIEW

We have developed Pharmacodynamics Miner, a text mining tool capable of retrieving PD articles and extracting PD related facts (sentences) from database of articles. Splitting articles into sentences develops PD miner, and sentences tokenized into words. Each word is then annotated using the eXtensibleMarkupLanguage (XML) following the entries from the drug dictionary and ontology. The markup words are then indexed to facilitate rapid search and retrieval of sentences mentioning the desired terms. PD Miner uses a database of full text articles and abstracts; table A provides a good description of the PD database.

# Full text articles	630
# Abstracts	1732
# Drug terms	12328
# Cell lines	5617
# Cell Phenotypes	3679
# Interactions	2784

Table.6 Overview of the database

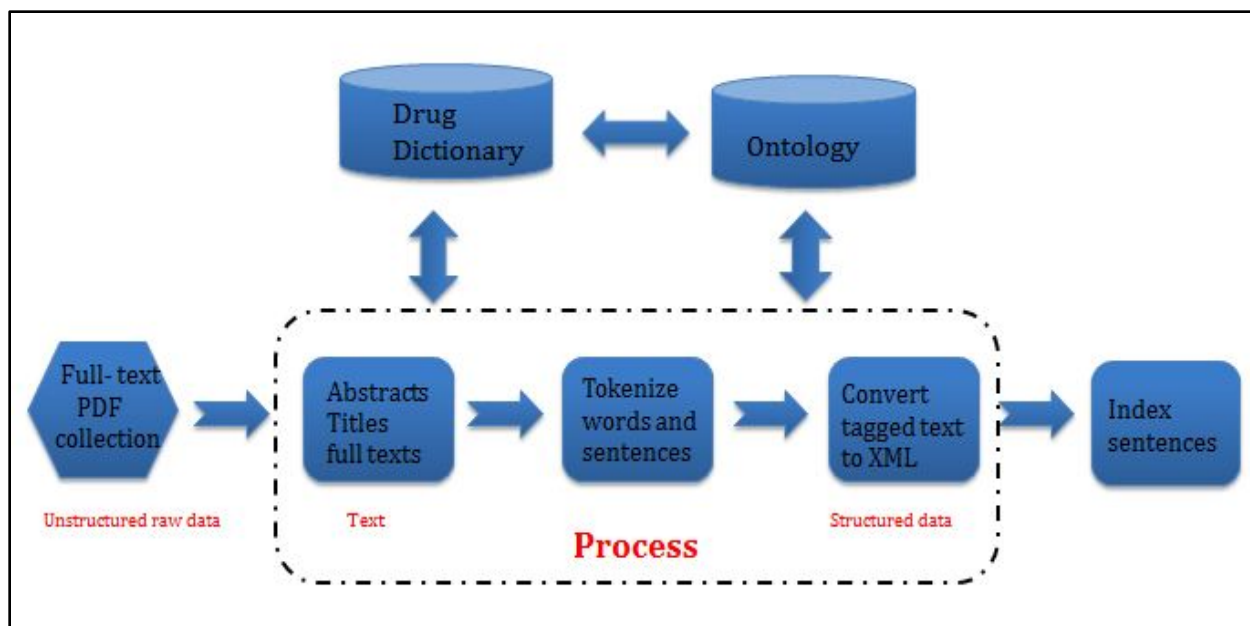


Fig. 13 Shows the PD Miner pipeline for document retrieval and sentence extraction

7.1.2 EVALUATION OF PD MINER

PD Miner was evaluated to determine its ability to extract PD DDI from full text articles and abstracts. Human experts evaluated a total of 5 journal articles and the same were queried using PD Miner. The selected journals include Clinical Cancer Research (18594016), British Journal of cancer (19513066), Breast cancer research (20576088), Anesthesia and analgesia (21474657) and Molecular cancer therapeutics (21646547). At the information extraction step of the PD Miner, DDI sentences were retrieved by the system and inspected by human experts. The comparison revealed that PD Miner identified and extracted DDI with a recall of 75% and a precision of 46.55%. The missed DDI are enlisted in table 9 and an error analysis is presented below.

Fig. 15 provides a description of the evaluation step. Table 8 is a general overview of DDI sentences retrieved by human experts and the system.

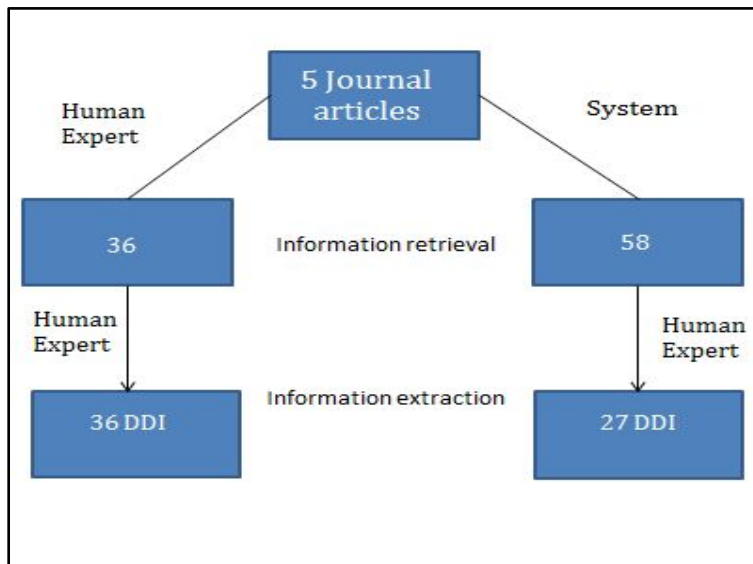


Fig.14 Evaluation schema

MID	True Manual retrieval.	True Retrieved from PD Miner	Total sentences PD Miner	Recall of PD Miner	Precision of PD Miner
19513066	11	9	18	81%	50%
21474657	11	9	12	81%	75%
18594016	4	2	5	50%	40%
21646547	2	2	2	100%	100%
20576088	8	5	21	62.5%	23.8%
Total	36	27	58	75%	46.55%

Table. 7 Evaluation of precision and recall of the system

7.1.2.1 ERROR ANALYSIS

PubMed	Missed drug-drug interaction sentences
19513066	Importantly, combination treatment with doxorubicin resulted in synergistic growth inhibition in all cell lines and blocked the migration and invasion of the highly metastatic, triple-negative MDA-MB-231 cell line.
19513066	The CI values calculated for 50% inhibition of these cells were 0.05 and 0.04, respectively, indicating very strong synergy
18594016	In each case, synergy involved inhibition of the ErbB signaling pathway, which could be shown both in vitro and in vivo in an NSCLC xenograft model
18594016	H1650 cells, the combination showed moderate synergy with an optimum CI of 0.71 at ED90, whereas for H292 cells, the combination showed synergy with an ED50 CI of 0.37, Indicating improved synergism over that in both H358 and H1650 cell lines.
21474657	At the $\alpha_1\beta_2\gamma_2S$ subtype, additive effects were observed irrespective of PKC activation.
21474657	When applied together, their effect was additive only at the $\alpha_1\beta_2\gamma_2S$ GABAA receptor; at the $\alpha_1\beta_2\gamma_2L$ GABAA receptor, their effect was not additive and only became additive when a PKC inhibitor was applied.
21861932	This is slightly greater than an additive effect which would have yielded 3% ($0.42 \times 0.36 \times 0.22 = 0.03$) of the controls (17+/-13 cells compared to 560).
21690228	Using Bonferroni correction for multiple testing, we conclude that for the MDA-MB-231 cell line, The combination is synergistic at (25, 5), (12.5, 2.5), (6.25, 1.25), (3.125, 0.625), (1.56, 0.3125), and (0.78, 0.156).
21690228	There is no evidence of synergy at (50, 10)
21690228	For the MDA-MB-468 line, the combination is synergistic at (25, 5), (12.5, 2.5), (6.25, 1.25), and (3.125, 0.625)

Table.8 Error analysis for missed drug interactions

The retrieval algorithm is implemented in such a way that it extracts sentences based upon co-occurrences of drug interaction type and drug names. The sentence “Importantly, combination treatment with doxorubicin resulted in synergistic growth inhibition in all cell lines and blocked the migration and invasion of the highly metastatic, triple-negative MDA-MB-231 cell line.” (PubMed:19513066), doesn’t mention the name of the second drug thus system fails to identify the sentence as DDI. Combination index (CI) is good measure to determine additivity, synergism and antagonism of a drug combination, in the sentence “The CI values calculated for 50% inhibition of these cells were 0.05 and 0.04, respectively, indicating very strong synergy”, only the CI value and the type of interaction is mentioned but the sentence lacks both interacting drugs and hence is not recognized by the system. Similarly, almost all the remaining sentences miss either one or both drug names, in order to extract and understand such DDI, context of sentences that are above or below is essential. To extract these types of sentences complex NLP system and algorithms are required.

7.2 PD MINER IS CAPABLE OF CATEGORICAL SEARCHES

PD Miner is capable of forming semantic and categorical queries. The categorical facility is very useful over the keyword-based feature especially for scientist and research curators. For E.g. If a scientist needs to query cell death using a keyword approach, the result retrieved would be very limited and might miss terms like necrosis, apoptosis as cell death can also be represented by these terms. In addition, if an individual is interested in querying a breast cancer cell line, he/she may have hard time remembering different cell lines present in that class. An example of categorical searches is performed by querying PD

Miner for drug interaction using cell death and brain cell line as the categories. As seen in fig.3 the system had identified apoptosis as a member of the cell death class and “sk-n-dz” and “sh-sy5y” as member of the brain cell line class. This example strongly supports the idea that an information extraction system that considers word semantics for a query can greatly increase both precision and recall of the retrieval.

Pharmacodynamic Drug Interaction Repository

Home	About	Browse	Search	Corpus	Tutorial	Contact
------	-------	--------	--------	--------	----------	---------

PubMed 21878749

[Sentence 108]discussion the members of the bcl-2 family are the crucial regulators of cell death .15 in our previous studies , we have shown the efficacy of the small molecule **inhibitor** of bcl-2 in combination with **genistein** in induction of **apoptosis** by altering the bax :bcl-2 ratio and activating intrinsic caspase cascade in neuroblastoma cells .16 our current investigation goes one step ahead to demonstrate that 2 ,3-dcpe (a small molecule **inhibitor** of the anti-apoptotic protein bcl-x l) in combination with **4-hpr** (a synthetic retinoid) can work **synergistically** to significantly increase differentiation and **apoptosis** in bcl-x l bountiful neuroblastoma **sk-n-dz** and **sh-sy5y** cells .

Center for Computational Biology and Bioinformatics, Indiana University, Indianapolis

Fig.15 Categorical search with cell death and brain cell line

7.3 COMPREHENSIVE CELL LINE COVERAGE

Apart from being a text mining system, PD Miner also provides a cell line search option. The cell line database houses around 32,000 cell lines extracted from both public and private repositories. As mentioned in earlier chapters the cell line database is enriched

with cell line entities from DSMZ, ATCC, Coriell Ontology, breast tissue cell line ontology etc. The retrieved information for a search is not limited to cell line but also includes the cell type, organism from which the cell line is extracted and the cell line supplier. Fig. 4 describes situation where a user queries the system to retrieve all MCF based breast cell lines. As seen in the figure along with the cell line name information about the organism, tissue source and repository is also provided.

The screenshot shows the 'Pharmacodynamics Miner' web application interface. At the top, there is a green header with the title 'Pharmacodynamics Miner' and the subtitle 'A tool to assist automated retrieval and extraction of PD concepts and relations'. Below the header is a navigation menu with buttons for Home, About, Browse, Search, DDI, Corpus, Download, and Contact. The main content area displays a table with the following data:

Cell_line_name	Organism	Source	Repository
MCF10A	human	mammarygland,epithelial	ATCC: CRL-10317
MCF10F	human	mammarygland,epithelial	ATCC: CRL-10318
MCF-10-2A	human	mammarygland,epithelial	ATCC: CRL-10781
MCF-12A	human	mammarygland,epithelial	ATCC: CRL-10782
MCF-12F	human	mammarygland,epithelial	ATCC: CRL-10783
MCF7	human	pleuraleffusion(metastasis),epithelial	ATCC: HTB-22
MCF7	human,Caucasian	breast	CLDB
MCF-7	human,Caucasian	breast	CLDB
MCF7-382	human,Caucasian	breast	CLDB
MCF7-422	human,Caucasian	breast	CLDB
MCF7-432	human,Caucasian	breast	CLDB
MCF7-488X1	human,Caucasian	breast	CLDB
MCF7-490X1	human,Caucasian	breast	CLDB
MCF7-492X1	human,Caucasian	breast	CLDB

Fig. 16 Cell line search

7.4 CORPUS AND ONTOLOGY

The development of the corpus and the ontology is already explained in the earlier chapters. The corpus was developed as a gold standard for TM based evaluation of the PD

studies, it includes both term and sentence level annotation. Most step of the corpus development is manual. The *in vitro* PD ontology is an effort to standardize PD concepts and knowledge. The ontology was developed using protégé and includes classes and multiple subclasses. Ontology development involved both manual and repository based data collection. Considering the categorical search results an ontology based information retrieval system is crucial for having a wide coverage of terms in the searches.

CHAPTER 8

DISCUSSION

8.1 ACCOMPLISHMENTS

We have developed PD Miner a system that is capable of retrieving information from full text PD articles. As of January 2013, the database includes full text articles published during the last five years. The semantic categories and text markup have introduced a new and a meaning full way to query the biological literature. The evaluation of the system has revealed that presence of full text is important for developing a new bio retrieval system that can present a wide number of biological facts and concepts with a high recall and a satisfying precision, which ultimately could help domain curators and researchers. For researcher and curators presence of such a system offers a great advantage over manual skimming of the text. PD Miner is an important ontology based full text search system, and offer great advantages even with low precision. The current PD Miner algorithm can be improved to achieve a higher precision.

It is well understood that almost all the 630 articles from our database could be well curated using PD Miner, its efficiency is more than human readers. With an increase in the corpus size the tool gets more useful by providing a huge recall rate to the curators. We believe that this tool would be useful for researcher and curators of the PD domain for identifying relevant full text articles and locating sentences that provide information about optimal drug doses, cellular phenotypes, DDI etc., this information is crucial prior to performing an actual PD experiment.

The development of the ontology is also an important contribution and encourages domain expert to use the ontology as a means of a central knowledge representation. Classes like drug interactions, cell lines and cellular phenotypes are primary basis of an *in vitro* PD experiment they cover almost all term semantics and class definitions. The ontology can be further edited or reused to develop any other domain ontology. Along with the ontology the drug dictionary has also played an important role in presenting term semantics and overcoming different drug conventions. It is apparent from the result that both the drug dictionary and the ontology have improved both the precision and recall of the TM system.

The presented corpus is highly enriched with term level annotations like cell lines, drug names, cell phenotypes, DDI types, drug concentrations and POS based annotations, it also cover sentence annotations and divide the DDI into DDI, Non-DDI and ambiguous types. Such annotations are highly significant will performing NLP based operations they provide a good control for entity recognition and extraction. Many NLP studies are performed on corpuses, the accuracy of these NLP algorithms are determined based upon their ability to identify and extract facts and relations. Since the presented corpus is manually annotated it can serve as an important gold standard encouraging more NLP and TM based studies on the PD domain.

8.2 LIMITATIONS

The text mining system is limited in with respect to the coverage of full text article and concepts from the ontology. The current database covers around 630 free full text articles, paid and subscribed journals are not included in the current database. The articles are also limited to last 5 years as the PDF download tool at Lang Li lab is still at infancy, for now it can download articles from limited journals. The ontology covers concepts only for the *in vitro* PD experiments; *in vivo* concepts are not included in the current release. Cell line classes of the ontology still miss some cell lines and did not get annotated during the NER process. Some cell lines are represented by number and not by alphabets or alphanumeric characters. For example the human lymphoblast cell line '36' is falsely identified as concentration (number) than as cell lines. Fortunately the retrieved results from PD Miner are in the form of sentences highlighting important concepts thus a domain expert from the context of the sentence can easily identify and discard such irrelevant hits.

Drug names for annotating the system were downloaded from the Drug bank database. Not all drugs are part of the Drug Bank and hence identification of such drugs is a challenge. Few studies also use natural products like "Curcumin" to study their interaction with standard drugs there is no provision to identify and annotate natural products. In many cases authors tend to use short abbreviations to present a drug instead of its original name. For example the drug Tamoxifen is represented as TM, TAM, TX etc. Such abbreviations are listed in any standard drug abbreviation protocol; it is merely based upon author's convenience. Retrieval of sentences with such abbreviation is challenging.

The articles in the database are stored in the form of text files. The conversion of PDF files to text files was achieved using the XPDF tool (<http://www.foolabs.com/xpdf/>). At many occasions the conversion of PDF to text was problematic due to different template layouts of the articles. Conversion also changes the structured format of the tables, even if the information from the tables is retrieved it becomes difficult to analyze and understand the information.

CHAPTER 9

REFERENCES

1. Alison H. Thomson: **Introduction to Clinical Pharmacokinetics.** [http://group.bmj.com/docs/pdf/4_1_s2.pdf]
2. Charity D. Scripture, William D. Figg: **Drug Interactions in Cancer Therapy.** [<http://www.medscape.com/viewarticle/540883>]
3. Percha B, Garten Y, Altman RB: **Discovery and explanation of drug-drug interactions via text mining.** *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing* 2012:410-421.
4. **National Cancer Institute** [<http://www.cancer.gov/dictionary?cdrid=330184>]
5. Sunita Sarawagi: **Information extraction** [<http://osm.cs.byu.edu/CS652s09/papers/Sarawagi.ieSurvey.pdf>]
6. Jia J, Zhu F, Ma X, Cao Z, Li Y, Chen YZ: **Mechanisms of drug combinations: interaction and network perspectives.** *Nature reviews Drug discovery* 2009, **8**(2):111-128.
7. Blaschke C, Valencia A: **Can bibliographic pointers for known biological data be found automatically? Protein interactions as a case study.** *Comp Funct Genomics* 2: 196-206
8. Muller HM, Kenny EE, Sternberg PW: **Textpresso: an ontology-based information retrieval and extraction system for biological literature.** *PLoS biology* 2004, **2**(11):e309.
9. Garten Y, Altman RB: **Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text.** *BMC bioinformatics* 2009, **10 Suppl 2**:S6.
10. Siadat MS, Shu J, Knaus WA: **Relemed: sentence-level search engine with relevance score for the MEDLINE database of biomedical articles.** *BMC medical informatics and decision making* 2007, **7**:1.

11. **iHOP, a new gene and protein analysis tool.** *Cancer biology & therapy* 2007, **6**(1):7-8.
12. Sarntivijai S, Ade AS, Athey BD, States DJ: **The Cell Line Ontology and its use in tagging cell line names in biomedical text.** *AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium* 2007:1103.5
13. David D. Allen, RaúlCaviedes, Ana María Cárdenas, Takeshi Shimahara, Juan Segura-Aguilar, Pablo A. Caviedes: **Cell lines as in vitro models for drug screening and toxicity studies***DrugDevInd Pharm.* 2005 September; **31**(8): 757–768
14. Sarntivijai S, Ade AS, Athey BD, States DJ: **A bioinformatics analysis of the cell line nomenclature.** *Bioinformatics* 2008, **24**(23):2760-2766.
15. **American Type Culture Collection (ATCC)** [<http://www.atcc.org/>]
16. Chao Pang,TomaszAdamusiak,HelenParkinson,James Malone (2011) **The Coriell Cell Line Ontology: Rapidly Developing Large Ontologies.** *Bio-Ontologies* 2011. <http://bio-ontologies.knowledgeblog.org/13>
17. Visser U, Abeyruwan S, Vempati U, Smith RP, Lemmon V, Schurer SC: **BioAssay Ontology (BAO): a semantic description of bioassays and high-throughput screening results.** *BMC bioinformatics* 2011, **12**:257.
18. Simon J, Dos Santos M, Fielding J, Smith B: **Formal ontology for natural language processing and the integration of biomedical databases.** *International journal of medical informatics* 2006, **75**(3-4):224-231.
19. Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C: **Relations in biomedical ontologies.** *Genome biology* 2005, **6**(5):R46.
20. Bard J, Rhee SY, Ashburner M: **An ontology for cell types.** *Genome biology* 2005, **6**(2):R21.
21. Xiang Z, Courtot M, Brinkman RR, Ruttenberg A, He Y: **OntoFox: web-based support for ontology reuse.** *BMC research notes* 2010, **3**:175.
22. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J: **DrugBank: a comprehensive resource for in silico drug discovery and exploration.** *Nucleic acids research* 2006, **34**(Database issue):D668-672.

23. Lee SI: **Drug interaction: focusing on response surface models.** *Korean journal of anesthesiology* 2010, **58**(5):421-434.
24. Tallarida RJ: **An overview of drug combination analysis with isobolograms.** *The Journal of pharmacology and experimental therapeutics* 2006, **319**(1):1-7.
25. Bijnsdorp IV, Giovannetti E, Peters GJ: **Analysis of drug interactions.** *Methods MolBiol*2011, **731**:421-434
26. Donato MT, Lahoz A, Castell JV, Gomez-Lechon MJ: **Cell lines: a tool for in vitro drug metabolism studies.** *Current drug metabolism* 2008, **9**(1):1-11.
27. Allen DD, Caviedes R, Cárdenas AM, Shimahara T, Segura-Aguilar J, Caviedes PA: **Cell Lines as In Vitro Models for Drug Screening and Toxicity Studies.** *Drug Development and Industrial Pharmacy* 2005, **31**(8):757-768.
28. Gremse M, Chang A, Schomburg I, Grote A, Scheer M, Ebeling C, Schomburg D: **The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources.** *Nucleic acids research* 2011, **39**(suppl 1):D507-D513.
29. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nature genetics* 2000, **25**(1):25-2
30. **Cell bank Australia** [<http://www.cellbankaustralia.com/>]
31. **Breast tissue cell lines**
[<http://bioportal.bioontology.org/ontologies/44178?p=terms>]
32. **DSMZ - Deutsche Sammlung von Mikroorganismen und Zellkulturen**[<http://www.dsmz.de/>]
33. **XPDF** [<http://www.foolabs.com/xpdf/>]
34. David D. Palmer :**Tokenisation and Sentence Segmentation**
[<http://comp.mq.edu.au/units/comp348/ch2.pdf>]
35. Kim, J. D., Ohta, T., Tateisi, Y., and Tsujii, J. "**GENIA corpus—a semantically annotated corpus for bio-textmining.**" *Bioinformatics* 19(Supp 1): i180-182 (2003).

HRISHIKESH LOKHANDE

3711, Gold Street Apt-1, Los Alamos, NM-87544 ♦ Ph-2196449445 ♦ hrishi@lanl.gov

CAREER OBJECTIVE

Seeking a position in the field of bioinformatics where my data analysis, next generation sequencing skills and research experience can be applied to solve interdisciplinary problems associated with biology.

EDUCATION

Master of Science in Bioinformatics Indiana University, School of Informatics, Indianapolis, IN	May 2013 GPA-3.4/4.0
Advanced Diploma in Bioinformatics BhartiVidyapeeth University, India	May 2010 GPA-3.5/4.0
Bachelors of Biotechnology Pune University, India	May 2009 GPA-3.2/4.0

COMPUTATIONAL PROFICIENCY

Languages & Scripts:	R, PERL, PHP, Python, XML, SAS.
Platforms:	UNIX, Windows XP/2000/Vista/7, Linux.
Databases:	MySQL, MySQL workbench, SQL Server, PostgreSQL, MS Access.
Database Tools:	Aqua Data Studio, Toad (MySQL).
Web Development:	HTML, CSS, Drupal.

BIOINFORMATICS PROFICIENCY

Microarray analysis:	MeV, GSEA, Bioconductor.
Database search:	GEO, Drugbank, Entrez, BLAST, GenBank,UCSC Genome browser.
Visualization tools:	Cytoscape, Metacore, GeneGo, Pathway studio,i-GSEA4GWAS.
Ontologies:	Disease ontology, Gene ontology, Cell ontology, Cell line ontology.
Enrichment tools:	DAVID, GeneMAPP, Gene Ontology tools like Protégé and OBO-EDIT2.
Other Analysis:	Pathway analysis, Network generation, Machine learning.
Next Generation seq:	RNA-seq-Atlas, Array Express, Scripture, Galaxy, Cufflinks, SpliceMap.

WORKEXPERIENCE

Bioinformatics Software developer

February 2013 - Present

Los Alamos National Laboratory, Los Alamos, NM

Currently involved in developing software tools for genomic and proteomic analysis of Influenza sequences. Working on maintaining and improving the influenza sequence database

Graduate research assistant

Sept 2011 - Jan2013

Center for Computational Biology and Bioinformatics (CCBB), Indianapolis, IN

Thesis- "Pharmacodynamics Miner: An automated extraction of Pharmacodynamic drug interactions."

Responsibilities:

- Develop pharmacodynamics ontology, classifying drug interaction into synergistic, additive and antagonistic categories.
- Mine literature to find drug interaction following pharmacodynamics ontology.
- Determine cell based phenotypes involved in in-vitro pharmacodynamics experiments,
- Develop a semantic corpus and a keyword based information retrieval system with high precision and recall.
- Develop drug dictionary and organ specific cell line ontology to enhance and improve keyword and category based searches.

Data Analyst/ Scientific Programmer

May 2011- Sept 2011

Indiana University School of Medicine, Indianapolis, IN

Performed data analysis for drug interaction identification and for GWAS study.

Graduate research assistant

Jan 2011- May 2011

Indiana center for Systems Biology and Personalized medicine, Indianapolis, IN

Performed database development, web development and data analysis for "HOMER: a human organ-specific molecular electronic repository" BMC bioinformatics 2011.

CLASS PROJECTS

- Development of Leukemia Database Using web 2.0 technology.
- GPCR Tree Classification using Multilayer Perceptron (Machine learning).
- Development of Pathway database.
- SNP Based pathway analysis of Alzheimer's genome-wide association study.
- Hierarchical access control system for patient privacy in a social health network.