# MINING BIOMEDICAL LITERATURE TO EXTRACT PHARMACOKINETIC DRUG-DRUG INTERACTIONS

Shreyas Karnik

Submitted to the faculty of the School of Informatics,

in partial fulfillment of the requirements

for the degree of

Master of Science in Bioinformatics,

Indiana University

June 2013

Accepted by the Faculty of Indiana University,

in partial fulfillment of the requirements for the degree of Master of Science

in Bioinformatics

**Master's Thesis**

**Committee**

_____

Lang Li, PhD., Chair

_____

Yunlong Liu, PhD.

_____

Xiaowen Liu, PhD.

Dedicated to *Aai* - *Baba*

## ACKNOWLEDGMENTS

I express my deep gratitude to my advisor Dr. Lang Li for providing me new ideas, inspiration and support that eventually translated to this thesis. During my interactions with Dr. Li I got valuable lessons on how to partake in high quality research, these lessons and close supervision along with critical feedback is something I will remember throughout my career.

I would also like to thank Dr. Yunlong Liu and Dr. Xiaowen Liu for being part of my thesis committee and providing critical feedback on this work.

I am thankful to the faculty and staff of Center for Computational Biology and Bioinformatics and School of Informatics for providing me financial support throughout the graduate program.

I thank members of Li lab Dr. Sara Kay Quinney, Paul, Abhinita, Michael, Guanglong, Jack, Eileen and Santosh for being part of PK corpus creation team (**Chapter 3**) and providing continuous feedback and help during the course of this work.

I am extremely grateful to my friends Yogesh, Rohit, Sarang for making my life at IUPUI memorable.

Special thanks for friends in Indianapolis namely Anagha Tai and Prasad for making me feel at home even I was miles away from it I owe my deepest gratitude to you.

Lastly, no words are adequate to acknowledge support, patience, encouragement and unconditional love of my parents, sisters, and specially my nephew without which none of this was possible.

## ABSTRACT

Polypharmacy is a general clinical practice, there is a high chance that multiple administered drugs will interfere with each other, such phenomenon is called drug-drug interaction (DDI). DDI occurs when drugs administered change each other's pharmacokinetic (PK) or pharmacodynamic (PD) response. DDIs in many ways affect the overall effectiveness of the drug or at some times pose a risk of serious side effects to the patients thus, it becomes very challenging to for the successful drug development and clinical patient care. Biomedical literature is rich source for in-vitro and in-vivo DDI reports and there is growing need to automated methods to extract the DDI related information from unstructured text.

In this work we present an ontology (PK ontology), which defines annotation guidelines for annotation of PK DDI studies. Using the ontology we have put together a corpora of PK DDI studies, which serves as excellent resource for training machine learning, based DDI extraction algorithms. Finally we demonstrate the use of PK ontology and corpora for extracting PK DDIs from biomedical literature using machine learning algorithms.

## TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

# Chapter 1 INTRODUCTION

## 1.1 Motivation

When drugs are introduced in the body by any delivery mechanism two broad classes of effects take place namely: what body does to the drug termed as pharmacokinetics (PK) and what drug does to the body termed as pharmacodynamics (PD). Pharmacokinetics studies drug absorption, disposition, metabolism, excretion, and transportation (ADMET) of the drug whereas pharmacodynamics involves study of binding of the drug to receptors and following the signal cascade towards clinical effect(s) (such as efficacy or off target effects).

Polypharmacy is a general clinical practice. More than 70% of old population (age >65) takes more than 3 medications at the same time in US and some European countries. Given these statistics there is a high chance that given drugs will interfere with each other, such phenomenon is called drug-drug interaction (DDI). Drug-drug interaction (DDI) occurs when drugs administered change each other's PK or PD response.

DDIs are a major cause of morbidity and mortality and lead to increased health care costs [1-4] DDIs total nearly 3% of all hospital admissions and 4.8% of admissions in the elderly. DDIs are also a common cause of medical errors, representing 3% to 5% of all inpatient medication errors. These numbers may actually underestimate the true public health burden of drug interactions as they reflect only well-established DDIs. These DDIs in many ways affect the overall effectiveness of the drug or at some times pose a risk of serious side effects to the patients [5] thus, it becomes very challenging to for the successful drug development and clinical patient care. Regulatory authorities such as the Food and Drug Administration (FDA) and the pharmaceutical companies keep a rigorous tab on the DDIs. Major source of DDI information is the biomedical literature since most of the *in vivo* or *in vitro* DDI research carried is reported in it. Due to the unstructured nature of the free text in the biomedical

literature it is difficult and laborious process to extract and analyze the DDIs from biomedical literature. With the growth of the biomedical literature there is growing need for systems that aim at annotating the DDI information in biomedical literature and information extraction (IE) systems that aim at extracting DDIs, although some efforts are made in this direction there exists significant gap between the resources currently available for annotating and extracting DDIs.

### 1.2 Background

The use of IE systems to extract relationship among biological entities from biomedical literature have been successful to a great extent [6] specifically in the area of protein-protein interaction extraction. Researchers have now started to look at extraction of DDI from biomedical literature some early attempts include retrieval of DDI relevant articles from MEDLINE [7] which forms the basis of IE systems to work upon. There are DDI extraction systems based on mechanism based reasoning approach [8], shallow parsing and linguistic rule based approach [9] and shallow linguistic kernel based method to extract DDI [10].

In this work we focus on development of annotation and extraction tools for PK DDI's by means of creating a PK ontology that forms basis for creation of PK corpus of well annotated *in vitro* and *in vivo* DDI studies. We further demonstrate the usability of the PK DDI corpus to be used as gold standard for developing DDI extraction pipelines. In the forthcoming chapters creation of PK ontology, PK corpus and extraction methodology is discussed in details.

## Chapter 2 CREATION OF PHARMACOKINETICS ONTOLOGY[1]

### 2.1 Introduction

Owing to the gifts of web and modern high throughput experiments growth of biomedical data has been explosive and continues raking up at a very fast pace, as a result we rely more and more on biomedical databases to keep up-to date with the sate of art data. Collection and dissemination of biomedical data is a key factor for the research community [11] this plays a very important role in translational research facilitating translation “bench-side” to “bed-side”.

Ontologies can be defined as collections of formal machine-readable, human-understandable representations of entities, and the relations among those entities, within a defined application domain. Ontologies aid researchers manage information explosion by providing very detailed and precise descriptions of biomedical entities, paving the way for annotating, analyzing and integrating results of biomedical research. Key features of ontologies include reusability and facilitation of heterogeneous data integration [12]. One of the most widely used ontologies in life sciences is Gene Ontology[13].

### 2.2 Need for pharmacokinetics ontology

DDI information is housed in databases like DrugBank [14], DiDB (http://www.druginteractioninfo.org/) and PharmGKB [15] each of these databases have their strengths but there are certain gaps when it comes to content of pharmacokinetic DDI information (*in vitro* and *in vivo*), but to address this issue currently there exists no ontology to define a PK DDI study and its components. This was the motivation of our research group to develop a strong PK ontology, which would eventually translate into information richness in this domain.

[1] This chapter is published as: Wu H-Y, Karnik S, Subhadarshini A, Wang Z, Philips S, Han X, Chiang C, Liu L, Boustani M, Rocha L *et al*: **An integrated pharmacokinetics ontology and corpus for text mining**. *BMC bioinformatics* 2013, **14**(1):35.

### 2.3 Pharmacokinetics ontology

The PK ontology was implemented with Protégé [16] in the Web Ontology Language (OWL format).

Our ontology consists of following components:

- Experiments
- Metabolism
- Transporter
- Drug
- Subject

These components have been summarized in **Table 1** and overview of the ontology is presented in **Figure 1**.

| Categories | Description | Resources |
|---|---|---|
| Pharmacokinetics Experiments | Pharmacokinetics studies and parameters. There are two major categories: *in vitro* experiments and *in vivo* studies. | Manually accumulated from textbooks and literature sources. |
| Transporters | Drug transportation enzymes | http://www.tcdb.org |
| Metabolism Enzymes | Drug metabolism enzymes | http://www.cypalleles.ki.se/ |
| Drugs | Drug names | http://www.drugbank.ca/ |
| Subjects | Subject description for a pharmacokinetics study. It is composed three categories: disease, physiology, and demographics | http://bioportal.bioontology.org/ontologies/42056<br>http://bioportal.bioontology.org/ontologies/39343<br>http://bioportal.bioontology.org/ontologies/42067 |

**Table 1** Components of PK Ontology

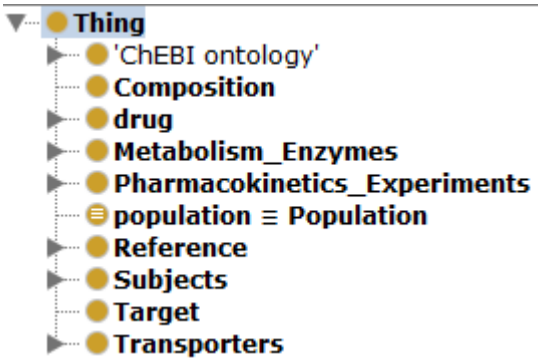


**Figure 1** Overview of PK Ontology

As ontologies support re-use we have re-used some of the existing ontologies in our PK ontology design, and our key contribution in the PK ontology is the definition of the experiments component.

### 2.4 Experiments section of PK ontology

This component describes *in vitro* and *in vivo* PK DDI experiments, experimental setup and the results of results of the same.

#### 2.4.1 *In vitro* PK DDI experiments

According to the FDA guidelines[2] on drug-drug interaction studies a DDI study generally begins with *in vitro* experiments, which deduce that, a drug is inhibitor, inducer or substrate of drug metabolizing enzymes (typically CYP P450 family of oxidative enzymes). The results of *in vitro* studies are valuable in quantitatively assessing the drug-drug interaction potential of an investigational drug and these results serve as decision points for further investigation. In the ontology we gather considerations critical for *in vitro* DDI studies. **Table 2** presents definitions and units of the *in vitro* PK parameters. The PK parameters of the single drug metabolism experiment include Michaelis-Menten constant ($K_m$), maximum velocity of the enzyme activity ($V_{max}$), intrinsic clearance ($CL_{int}$), metabolic ratio, and

[2]http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm292362.pdf

fraction of metabolism by an enzyme ($fm_{enzyme}$) [17]. In the transporter experiment, the PK parameters include apparent permeability (Papp), ratio of the basolateral to apical permeability and apical to basolateral permeability (Re), radioactivity, and uptake volume [18]. There are multiple drug interaction mechanisms: competitive inhibition, non-competitive inhibition, uncompetitive inhibition, mechanism based inhibition, and induction [19]. $IC_{50}$ is the inhibition concentration that inhibits to 50% enzyme activity; it is substrate dependent; and it doesn't imply the inhibition mechanism. $K_i$ is the inhibition rate constant for competitive inhibition, noncompetitive inhibition, and uncompetitive inhibition. It represents the inhibition concentration that inhibits to 50% enzyme activity, and it is substrate concentration independent. $K_{deg}$ is the degradation rate constant for the enzyme. $K_I$ is the concentration of inhibitor associated with half maximal Inactivation in the mechanism based inhibition; and $K_{inact}$ is the maximum degradation rate constant in the presence of a high concentration of inhibitor in the mechanism based inhibition. $E_{max}$ is the maximum induction rate, and $EC_{50}$ is the concentration of inducer that is associated with the half maximal induction.

| **Experiment Types** | **Parameters** | **Description** | **Unit** | **References** |
|---|---|---|---|---|
| Single Drug Metabolism Experiment | $K_m$ | Michaelis-Menten constant. | mg $L^{-1}$ | [17] p28 |
| | $V_{max}$ | Maximum velocity of the enzyme activity. | mg $h^{-1}$ $mg^{-1}$ protein | [17] p19 |
| | $CL_{int}$ | Intrinsic metabolic clearance is defined as ratio of maximum metabolism rate, Vmax, and the Michaelis-Menten constant, | ml $h^{-1}$ $mg^{-1}$ protein | [20] p165 |

| | | | | |
|---|---|---|---|---|
| | | Km. | | |
| | Metabolic ratio | Parent drug/metabolite concentration ratio | NA | |
| | $fm_{enzyme}$ | Fraction of drug systemically available that is converted to a metabolite through a specific enzyme. | NA | [20] xiii |
| Single Drug Transporter Experiment | Papp | The apparent permeability of compounds across the monolayer cells. | cm/sec | [18] |
| | Re | Re is the ratio of basolateral to apical over apical to basolateral. | NA | [18] |
| | Radioactivity | Total radioactivity in plasma and bile samples is measured in a liquid scintillation counter | dpm/mg protein | [18] |
| | Uptake Volume | The amount of radioactivity associated with the cells divided by its concentration in the incubation medium. | ul/mg protein | [18] |
| Drug Interaction Experiment | $IC_{50}$ | Inhibitor concentration that inhibits to 50% of enzyme activity. | mg $L^{-1}$ | |
| | $K_i$ | Inhibition rate constant for competitive inhibition, noncompetitive inhibition, and uncompetitive inhibition. | mg $L^{-1}$ | [17] p103 |
| | $K_{deg}$ | The natural degradation rate constant for the Enzyme. | $h^{-1}$ | [19] |
| | $K_I$ | The concentration of inhibitor associated with half maximal Inactivation in the mechanism based inhibition. | mg $L^{-1}$ | [19] |
| | $K_{inact}$ | The maximum degradation rate constant in the presence of a high concentration of inhibitor in the mechanism based inhibition. | $h^{-1}$ | [19] |

| | | | | |
|---|---|---|---|---|
| | $E_{max}$ | Maximum induction rate | Unit free | [19] |
| | $EC_{50}$ | The concentration of inducer that is associated with the half maximal induction. | mg $L^{-1}$ | [19] |
| Type of Drug Interactions | Competitive inhibition, noncompetitive inhibition, uncompetitive inhibition, mechanism based inhibition, and induction. | Rostami-Hodjegan and Tucker | | |

**Table 2** in vitro PK Parameters

*In vitro* experimental conditions are described in **Table 3**. Metabolism enzyme experiment conditions include buffer, NADPH sources, and protein sources. In particular, protein sources include recombinant enzymes, microsomes, hepatocytes, and etc. Sometimes, genotype information is available for the microsome or hepatocyte samples. Transporter experiment conditions include bi-directional transporter, uptake/efflux, and ATPase. Other factors of *in vitro* experiments include pre-incubation time, incubation time, quantification methods, sample size, and data analysis methods.

| **Experimental Conditions** | **Drugs** | **Substrate, metabolite, and inhibitor/inducer** | | **FDA Drug Interaction Guidelines[2].** |
|---|---|---|---|---|
| Metabolism Enzymes | Buffer | Salt composition | | |
| | | EDTA concentration | | |
| | | MgCl2 concentration Cytochrome b5 concentration | | |
| | NADPH source | Concentration of exogenous NADPH added isocytrate dehydrogenase + NADP | | |
| | protein | Non-recombinant enzymes | Microsomes (human liver microsomes, human intestine microsomes, S9 fraction, cytosol, whole cell lysate, hepatocytes. | |

| | | | | |
|---|---|---|---|---|
| | | Recombinant enzymes | Enzyme name genotype | mg/mL or uM |
| Transporters | Bi-Directional Transport | CHO; Caco-2 cells; HEK-293; Hepa-RG; LLC; LLC-PK1 MDR1 cells; MDCK; MDCK-MDR1 cells; Suspension Hepatocyte | | |
| | Uptake/efflux | tumor cells, cDNA transfected cells, oocytes injected with cRNA of transporters | | |
| | ATPase | membrane vesicles from various tissues or cells expressing P-gp, Reconstituted P-gp | | |
| Other factors | Pre-incubation time | | | |
| | Incubation time | | | |
| | Quantification methods | HPLC/UV, LC/MS/MS, LC/MS, radiographic | | |
| | Sample size | | | |
| | Data Analysis | log-linear regression, plotting; and nonlinear regression | | |

**Table 3** *in vitro* Experiment Conditions

### 2.4.2 *In viv*o PK DDI experiments

*In vivo* PK DDI studies aim at comparing substrate concentrations with and without the interacting drug, these type of studies typically address number of questions of the interaction between two drugs and clinical consequences of the same.

**Table 4** provides compilation of *in vivo* PK parameters based on information summarized from two text-books[20, 21]. There are several main classes of PK parameters. Area under the concentration curve parameters are ($AUC_{inf}$, $AUC_{SS}$, $AUC_t$, AUMC); drug clearance parameters are (CL, $CL_b$, $CL_u$, $CL_H$, $CL_R$, $CL_{po}$, $CL_{IV}$, $CL_{int}$, $CL_{12}$); drug concentration parameters are ($C_{max}$, $C_{SS}$); extraction ratio and bioavailability parameters are (E, $E_H$, F, $F_G$, $F_H$, $F_R$, $f_e$, $f_m$); rate constants include elimination rate constant k, absorption rate constant ka,

urinary excretion rate constant ke, Michaelis-Menten constant Km, distribution rate constants ($k_{12}$, $k_{21}$), and two rate constants in the two-compartment model ($\lambda_1$, $\lambda_2$); blood flow rate (Q, $Q_H$); time parameters ($t_{max}$, $t_{1/2}$); volume distribution parameters (V, $V_b$, $V_1$, $V_2$, $V_{ss}$); maximum rate of metabolism, Vmax; and ratios of PK parameters that present the extend of the drug interaction, (AUCR, CL ratio, Cmax ratio, $C_{ss}$ ratio, $t_{1/2}$ ratio).

| **Category** | **Name** | **Description** | **Unit** | **Reference** |
|---|---|---|---|---|
| PK parameters | $AUC_{inf}$ | Area under the drug concentration time curve. | mg h $L^{-1}$ | [20] p37 |
| | $AUC_{SS}$ | Area under the drug concentration time curve within a dosing curve at steady state. | mg h $L^{-1}$ | [20] pxi |
| | $AUC_t$ | Area under the drug concentration time curve from time 0 to t. | mg h $L^{-1}$ | [20] p37 |
| | AUMC | Area under the first moment of concentration versus time curve. | $mg^2$ h $L^{-2}$ | [20] p486 |
| | AUCR | AUC ratio (drug interaction parameter). | Unit free | |
| | CL | Total clearance is defined as the proportionality factor relating rate of drug elimination to the plasma drug concentration. | ml $h^{-1}$ | [20] p23 |
| | $CL_b$ | Blood clearance is defined as the proportionality factor relating rate of drug elimination to the blood drug concentration. | ml $h^{-1}$ | RT p160 |
| | $CL_u$ | Unbound clearance is defined as the proportionality factor relating rate of drug elimination to the unbounded plasma drug concentration. | ml $h^{-1}$ | [20] p163 |
| | $CL_H$ | Hepatic portion of the total clearance. | ml $h^{-1}$ | [20] p161 |
| | $CL_R$ | Renal portion of the total clearance. | ml $h^{-1}$ | [20] p161 |
| | $CL_{po}$ | Total clearance of drug following an oral dose. | ml $h^{-1}$ | |

| | | | |
|---|---|---|---|
| $CL_{IV}$ | Total clearance of drug following an IV dose. | ml $h^{-1}$ | |
| $CL_{int}$ | Intrinsic metabolic clearance is defined as ratio of maximum metabolism rate, Vmax, and the Michaelis-Menten constant, Km. | ml $h^{-1}$ | [20] p165 |
| $CL_{12}$ | Inter-compartment distribution between the central compartment and the peripheral compartment. | ml $h^{-1}$ | |
| CL ratio | Ratio of the clearance (drug interaction parameter). | Unit free | |
| $C_{max}$ | Highest drug concentration observed in plasma following administration of an extravascular dose. | mg $L^{-1}$ | [20] pxii |
| $C_{max}$ ratio | The ratio of $C_{max}$ (drug interaction parameter). | Unit free | |
| $C_{ss}$ | Concentration of drug in plasma at steady state during a constant rate intravenous infusion. | mg $L^{-1}$ | [20] pxii |
| $C_{ss}$ ratio | The ratio of $C_{ss}$ (drug interaction parameter). | Unit free | |
| E | Extraction ratio is defined as the ratio between blood clearance, $CL_b$, and the blood flow. | Unit free | [20] p159 |
| $E_H$ | Hepatic extraction ratio. | Unit free | [20] p161 |
| F | Bioavailability is defined as the proportion of the drug reaches the systemic blood. | Unit free | [20] p42 |
| $F_G$ | Gut-wall bioavailability. | Unit free | |
| $F_H$ | Hepatic bioavailability. | Unit free | [20] p167 |
| $F_R$ | Renal bioavailability. | Unit free | [20] p170 |
| fe | Fraction of drug systemically available that is excreted unchanged in urine. | Unit free | [20] pxiii |

| | | | |
|---|---|---|---|
| fm | Fraction of drug systemically available that is converted to a metabolite. | Unit free | [20] pxiii |
| fu | Ratio of unbound and total drug concentrations in plasma. | Unit free | [20] pxiii |
| k | Elimination rate constant. | $h^{-1}$ | [20] pxiii |
| $K_{12}$, $k_{21}$ | Distribution rate constants between central compartment and peripheral compartment. | $h^{-1}$ | |
| ka | Absorption rate constant. | $h^{-1}$ | [20] pxiii |
| ke | Urinary excretion rate constant. | $h^{-1}$ | [20] pxiii |
| km | Rate constant for the elimination of a metabolite. | $h^{-1}$ | [20] pxiii |
| Km | Michaelis-Menten constant. | mg $L^{-1}$ | [20] pxiii |
| MRT | Mean time a molecular resides in body. | h | [20] pxiv |
| Q | Blood flow. | L $h^{-1}$ | [20] pxiv |
| $Q_H$ | Hepatic blood flow. | L $h^{-1}$ | [20] pxiv |
| $t_{max}$ | Time at which the highest drug concentration occurs following administration of an extravascular dose. | h | [20] pxiv |
| $t_{1/2}$ | Half-life of the drug disposition. | h | [20] pxiv |
| $t_{1/2}$ ratio | Half-life ratio (drug interaction parameter). | Unit free | |
| $t_{1/2,\alpha}$ | Half-life of the fast phase drug disposition. | h | |
| $t_{1/2,\beta}$ | Half-life of the slow phase drug disposition. | h | |
| V | Volume of distribution based on drug concentration in plasma. | L | [20] pxiv |
| $V_b$ | Volume of distribution based on drug concentration in blood. | L | [20] pxiv |
| $V_1$ | Volume of distribution of the central compartment. | L | [20] pxiv |

| | | | | |
|---|---|---|---|---|
| | $V_2$ | Volume of distribution of the peripheral compartment. | L | |
| | $V_{ss}$ | Volume of distribution under the steady state concentration. | L | [20] pxiv |
| | Vmax | Maximum rate of metabolism by an enzymatically mediated reaction. | mg $h^{-1}$ | [20] pxiv |
| | $\lambda_1$, $\lambda_2$ | Disposition rate constants in a two-compartment model. | $h^{-1}$ | [21] p84 |
| Pharmacokinetics Models | Non-Compartment | Use drug concentration measurements directly to estimate PK parameters, such as AUC, CL, $C_{max}$, $T_{max}$, $t_{1/2}$, F, and V. | | [21] p409 |
| | One Compartment Model | It assumes the whole body is a homogeneous compartment, and the distribution of the drug from the blood to tissue is very fast. It assumes either a first order or a zero order absorption rate and a first order eliminate rate. Its PK parameters include (ka, V, CL, F). | | [20] p34<br>[21] p1 |
| | Two Compartment Model | It assumes the whole body can be divided into two compartments: central compartment (i.e. systemic compartment) and peripheral compartment (i.e. tissue compartment). It assumes either a first order or a zero order absorption rate and a first order eliminate and distribution rates. Its PK parameters include (ka, $V_1$, $V_2$, CL, $CL_{12}$, F). | | [21] p84 |
| Study Designs | Hypothesis | Bioequivalence, drug interaction, pharmacogenetics, and disease conditions. | | |
| | Design | Single arm or multiple arms; cross-over or fixed order design; with or without randomization; with or without stratification; prescreening or no-prescreening; prospective or retrospective studies; and case reports or cohort studies. | | |
| | Sample size | The number of subjects, and the number of plasma or urine samples per subject. | | |
| | Time points | Sampling time points and dosing time points. | | |
| | Sample types | Blood, plasma, and urine. | | |
| | Dose | Subject specific doses. | | |

| | |
|---|---|
| Quantification methods | HPLC/UV, LC/MS/MS, LC/MS, radiographic |

**Table 4** *In vivo* PK studies

It is also shown in **Table 4** that two types of pharmacokinetics models are usually presented in the literature: non-compartment model and one or two-compartment models. There are multiple items need to be considered in an *in vivo* PK study. The hypotheses include the effects of bioequivalence, drug interaction, pharmacogenetics, and disease conditions on a drug's PK. The design strategies are very diverse: single arm or multiple arms, cross-over or fixed order design, with or without randomization, with or without stratification, pre-screening or no-pre-screening based on genetic information, prospective or retrospective studies, and case reports or cohort studies. The sample size includes the number of subjects, and the number of plasma or urine samples per subject. The time points include sampling time points and dosing time points. The sample type includes blood, plasma, and urine. The drug quantification methods include HPLC/UV, LC/MS/MS, LC/MS, and radiography.

CYP450 family enzymes predominantly exist in the gut wall and liver. Transporters are tissue specific. **Table 5** presents the tissue specific transports and their functions. Probe drug is another important concept in the pharmacology research. An enzyme's probe substrate means that this substrate is primarily metabolized or transported by this enzyme. In order to experimentally prove whether a new drug inhibits or induces an enzyme, its probe substrate is always utilized to demonstrate this enzyme's activity before and after inhibition or induction. An enzyme's probe inhibitor or inducer means that it inhibits or induces this enzyme primarily. Similarly, an enzyme's probe inhibitor needs to be utilized if we investigate whether this enzyme metabolizes the drug. **Table 6** presents all the probe inhibitors, inducers, and substrates of CYP enzymes. **Table 7** presents all the probe inhibitors, inducers, and

substrates of the transporters; this information is compiled from the FDA guidelines for DDI studies[2].

| Gene | Aliases | Tissue type | Function |
|---|---|---|---|
| *ABCB1* | P-gp, MDR1 | Intestinal enterocyte, kidney proximal tubule, hepatocyte (canalicular), brain endothelia | Efflux |
| *ABCG2* | BCRP | Intestinal enterocyte, hepatocyte (canalicular), kidney proximal tubule, brain endothelia, placenta, stem cells, mammary gland (lactating) | Efflux |
| *SLCO1B1* | OATP1B1, OATP-C, OATP2, LST-1 | Hepatocyte (sinusoidal) | Uptake |
| *SLCO1B3* | OATP1B3, OATP-8 | Hepatocyte (sinusoidal) | Uptake |
| *SLC22A2* | OCT2 | Kidney proximal tubule | Uptake |
| *SLC22A6* | OAT1 | Kidney proximal tubule, placenta | Uptake |
| *SLC22A8* | OAT3 | Kidney proximal tubule, choroid plexus, brain endothelia | Uptake |

**Table 5** Tissue Specific Transporters

| CYP Enzymes | Inhibitors | Inducers | Substrates |
|---|---|---|---|
| CYP1A2 | Ciprofloxacin, enoxacin, fluvoxamine, Methoxsalen, mexiletine, oral contraceptives, phenylpropanolamine, thiabendazole, vemurafenib, zileuton, acyclovir, allopurinol, caffeine, cimetidine, daidzein, disulfiram, Echinacea, famotidine, norfloxacin, propafenone, propranolol, terbinafine, ticlopidine, verapamil | Montelukast, phenytoin, smokers versus non-smokers, moricizine, omeprazole, phenobarbital | Alosetron, caffeine, duloxetine, melatonin, ramelteon, tacrine, tizanidine, theophylline, tizanidine |
| CYP2B6 | Clopidogrel, ticlopidine prasugrel | Efavirenz, rifampin, nevirapine | Bupropion, efavirenz |
| CYP2C8 | Gemfibrozil, fluvoxamine, ketoconazole, trimethoprim | Rifampin | Repaglinide, Paclitaxel |
| CYP2C9 | Amiodarone, fluconazole, miconazole, oxandrolone, capecitabine, cotrimoxazole, etravirine, fluvastatin, fluvoxamine, metronidazole, sulfinpyrazone, tigecycline, voriconazole, zafirlukast | Carbamazepine, rifampin, aprepitant, bosentan, phenobarbital, St. John’s wort | Celecoxib, Warfarin, phenytoin |
| CYP2C19 | Fluconazole, fluvoxamine, ticlopidine, esomeprazole, fluoxetine, moclobemide, omeprazole, voriconazole, allicin (garlic derivative), armodafinil, carbamazepine, cimetidine, etravirine, human growth hormone (rhGH), felbamate, ketoconazole, oral contraceptives | Rifampin, artemisinin | Clobazam, lansoprazole, omeprazole, Smephenytoin, S-mephenytoin |
| CYP3A | Boceprevir, clarithromycin, conivaptan, grapefruit juice, indinavir, itraconazole, | Avasimibe, carbamazepine, | Alfentanil, aprepitant, budesonide, buspirone, |

| | | | |
|---|---|---|---|
| | ketoconazole, lopinavir/ritonavir, mibefradil, nefazodone, nelfinavir, posaconazole, ritonavir, saquinavir, telaprevir, telithromycin, voriconazole, amprenavir, aprepitant, atazanavir, ciprofloxacin, crizotinib, darunavir/ritonavir, diltiazem, erythromycin, fluconazole, fosamprenavir, grapefruit juice, imatinib, verapamil, alprazolam, amiodarone, amlodipine, atorvastatin, bicalutamide, cilostazol, cimetidine, cyclosporine, fluoxetine, fluvoxamine, ginkgo, goldenseal, isoniazid, lapatinib, nilotinib, oral contraceptives, pazopanib, ranitidine, ranolazine, tipranavir/ritonavir, ticagrelor, zileuton | phenytoin, rifampin, St. John's wort, bosentan, efavirenz, etravirine, modafinil, nafcillin, amprenavir, aprepitant, armodafinil, clobazamechinacea, pioglitazone, prednisone, rufinamide, vemurafenib | conivaptan, darifenacin, darunavir, dasatinib, dronedarone, eletriptan, eplerenone, everolimus, felodipine, indinavir, fluticasone, lopinavir, lovastatin, lurasidone, maraviroc, midazolam, nisoldipine, quetiapine, saquinavir, sildenafil, simvastatin, sirolimus, tolvaptan, tipranavir, triazolam, ticagrelor, vardenafil, Alfentanil, astemizole, cisapride, cyclosporine, dihydroergotamine, ergotamine, fentanyl, pimozide, quinidine, sirolimus, tacrolimus, terfenadine |
| CYP2D6 | Bupropion, fluoxetine, paroxetine, quinidine, cinacalcet, duloxetine, terbinafine,<br><br>amiodarone, celecoxib, clobazam, cimetidine, desvenlafaxine, diltiazem, diphenhydramine, echinacea, escitalopram, febuxostat, gefitinib, hydralazine, hydroxychloroquine, imatinib, methadone, oral contraceptives, pazopanib, propafenone, ranitidine, ritonavir, sertraline, telithromycin, verapamil, vemurafenib | NA | Atomoxetine, desipramine, dextromethorphan, metoprolol, nebivolol, perphenazine, tolterodine, venlafaxine, Thioridazine, pimozide |

**Table 6** in vivo Probe Inhibitors/Inducers/Substrates of CYP Enzymes

| Transporter | Inhibitor | Inducer | Substrate |
|---|---|---|---|
| P-gp | Amiodarone, azithromycin, captopril, carvedilol, clarithromycin, conivaptan, cyclosporine, diltiazem, dronedarone, erythromycin, felodipine, itraconazole, ketoconazole, lopinavir and ritonavir, quercetin, quinidine, ranolazine, ticagrelor, verapamil | Avasimibe, carbamazepine, phenytoin, rifampin, St John's wort, tipranavir/ritonavir | Aliskiren, ambrisentan, colchicine, dabigatran etexilate, digoxin, everolimus, fexofenadine, imatinib, lapatinib, maraviroc, nilotinib, posaconazole, ranolazine, saxagliptin, sirolimus, sitagliptin, talinolol, tolvaptan, topotecan |
| BCRP | Cyclosporine, elacridar (GF120918), eltrombopag, gefitinib | NA | Methotrexate, mitoxantrone, imatinib, irrinotecan, lapatinib, rosuvastatin, sulfasalazine, topotecan |
| OATP1B1 | Atazanavir, cyclosporine, eltrombopag, gemfibrozil, lopinavir, rifampin, ritonavir, saquinavir, tipranavir | NA | Atrasentan, atorvastatin, bosentan, ezetimibe, fluvastatin, glyburide, SN-38 (active metabolite of irinotecan), rosuvastatin, simvastatin acid, pitavastatin, pravastatin, repaglinide, rifampin, valsartan, olmesartan |
| OATP1B3 | Atazanavir, cyclosporine, lopinavir, rifampin, ritonavir, saquinavir | NA | Atorvastatin, rosuvastatin, pitavastatin, telmisartan, valsartan, olmesartan |
| OCT2 | Cimetidine, quinidine | NA | Amantadine, amiloride, cimetidine, dopamine, famotidine, memantine, metformin, pindolol, procainamide, ranitidine, varenicline, oxaliplatin |
| OAT1 | Probenecid | NA | Adefovir, captopril, furosemide, lamivudine, methotrexate, oseltamivir, tenofovir, zalcitabine, zidovudine |

| | | | |
|---|---|---|---|
| OAT3 | Probenecid cimetidine, diclofenac | NA | Acyclovir, bumetanide, ciprofloxacin, famotidine, furosemide, methotrexate, zidovudine, oseltamivir acid, (the active metabolite of oseltamivir), penicillin G, pravastatin, rosuvastatin, sitagliptin |

**Table 7** in vivo Probe Inhibitors/Inducers/Substrates of Selected Transporters

### 2.4.3 Metabolism component

The cytochrome P450 superfamily (officially abbreviated as CYP) is a large and diverse group of enzymes that catalyze the oxidation of organic substances. The substrates of CYP enzymes include metabolic intermediates such as lipids and steroidal hormones, as well as xenobiotic substances such as drugs and other toxic chemicals. CYPs are the major enzymes involved in drug metabolism and bioactivation, accounting for about 75% of the total number of different metabolic reactions [22]. CYP enzyme names and genetic variants were mapped from the Human Cytochrome P450 (CYP) Allele Nomenclature Database (http://www. cypalleles.ki.se/). This site contains the CYP450 genetic mutation effect on the protein sequence and enzyme activity with associated references.

### 2.4.4 Transporters component

Transport Proteins are proteins that serve the function of moving other materials within an organism. Transport proteins are vital to the growth and life of all living things. Transport proteins involved in the movement of ions, small molecules, or macromolecules, such as another protein, across a biological membrane. They are integral membrane proteins; that is they exist within and span the membrane across which they transport substances. Their names and genetic variants were mapped from the Transporter Classification Database (http://www.tcdb.org).

### 2.4.5 Drugs component

Drug names was created using the drug names from DrugBank 3.0 [14]. DrugBank consists of 6,829 drugs that can be grouped into different categories of FDA-approved, FDA approved biotech, nutraceuticals, and experimental drugs. The drug names are mapped to generic names, brand names, and synonyms.

### 2.4.6 Subject component

Subject component includes existing ontologies for human disease ontology (DOID) [23], Suggested Ontology for Pharmacogenomics (SO-Pharm) [24] and mammalian phenotype (MP) [25] from http://bioportal.bioontology.org (see **Table 1**).

## 2.5 Applications of the PK Ontology

To demonstrate utility of the PK Ontology we present 3 case studies in which the ontology was used in annotation.

### 2.5.1 Example 1: An annotated tamoxifen pharmacogenetics study

This example shows how to annotate a pharmacogenetics studies with the PK ontology. We used a published tamoxifen PG study [26]. This PG study investigates the genetic effects (CYP3A4, CPY3A5, CYP2D6, CYP2C9, CYP2B6) on the tamoxifen pharmacokinetics outcome (tamoxifen metabolites) among breast cancer patients. It was a single arm longitudinal study (n = 298), patients took SOLTAMOXTM 20mg/day, and the drug steady state concentration was sampled (1, 4, 8, 12) months after the tamoxifen treatment. The study population was a mixed Caucasian and African American. The key information from this tamoxifen PG trial was extracted as a summary list and the pre-processed information was mapped to the PK ontology Ref. **Figure 2** (under heading Pharmacogenomics Trial). We can

see from the annotation mapping that key information from the study can be easily summarized using the ontology.

| Ontology | Pharmacogenetics Trial | Drug Interaction Trial |
|---|---|---|
| Drugs = SOPHARM_20000 | Tamoxifen (TAM) | Midazolam (MDZ, PO 4mg; IV 0.05mg/kg), Ketoconazole (KTZ, PO, 200, 400 mg) |
| Experiments | | |
| in-vitro | | |
| in-vivo | *in-vivo* | *in-vivo* |
| Analysis_Method | | |
| Assay | HPLC/MS | HPLC/MS |
| Dose | SOLTAMOX™, 20mg/day | MDZ PO, IV; KTZ PO |
| Measurement | month 1, 4, 8, 12 | before and 0.5, 0.75, 1, 2, 4, 6, 9 hrs |
| PK_Parameters | TAM and its metabolites conc | MDZ and KTZ: AUC, AUCR, $t_{1/2}$, and Cmax |
| Pre-dosing_Conditions | | |
| Sample | | |
| Sample_Size | 298 | 24 |
| Sample_Types | Blood | blood |
| Stratification | prior chemo, menopausal | |
| Study_Design | | |
| Bioequivalence_Study | | |
| Dense_Sampling | | |
| Disease-Physiology_PK_Study | | |
| Drug_Interaction_Study | | inhibition |
| Longitudinal | Longitudinal | three-phase crossover |
| Pharmacogenetics_Study | prospective, single arm | prospective, single arm |
| Sparse_Sampling | | |
| Steady_State_Study | steady state | |
| Type_of_PK_Study | | |
| Metabolism | | |
| CYP1_family | | |
| CYP2_family | CYP2D6, 2C9, 2B6 | |
| CYP3_family | CYP3A4/5 | CYP3A4/5 |
| CYP4_family | | |
| CYP_other_families | | |
| Subjects | | |
| Disease = DOID_14974 | breast cancer | healthy volunteers |
| Physiology = MP_0000001 | | |
| Population = SOPHARM_52000 | Caucasian/African American | |
| Target | ESR1/ESR2 | |

**Figure 2** Annotated Pharmacogenomics Study using PK Ontology

### 2.5.2 Example 2 midazolam/ketoconazole drug interaction study

This was a cross-over three-phase drug interaction study [27] (n = 24) between midazolam (MDZ) and ketoconazole (KTZ). Phase I was MDZ alone (IV 0.05 mg/kg and PO 4mg); phase II was MDZ plus KTZ (200mg); and phase III was MDZ plus KTZ (400mg). Genetic variable include CYP3A4 and CYP3A5. The PK outcome is the MDZ AUC ratio before and

after KTZ inhibition. Annotated version of this study is presented in Ref. **Figure 2** (under heading Drug Interaction Trial)

### 2.5.3 Example 3 *in vitro* Pharmacokinetics Study

This was an *in vitro* study [28], which investigated the drug metabolism activities for 3 enzymes, such as CYP3A4, CYP3A5, and CYP3A7 in a recombinant system. Using 10 CYP3A substrates, they compared the relative contribution of 3 enzymes among 10 drug's metabolism. Annotated version of this study is presented in **Figure 3**.

| Ontology | in-vitro study |
|---|---|
| Drugs ≡ SOPHARM_20000 | MDZ, APZ, TZ, CLAR, TAM, DTZ, NIF, BFC, HFC, TEST, E2 |
| Experiments | Compare metabolic capabilities of CYP3A4, 3A5, 3A7 |
| in-vitro | |
| Experimental_Conditions | |
| Buffer | sodium phosphate, NADPH, methanol. |
| NADPH_Source | |
| Other_Information | |
| Data_analysis_method | WinNonlin |
| Dilution | 4 fold, 10% methanol (TZ) |
| Incubation_time | 5 min |
| Microsomal_binding | insect cell (CYP3A) |
| Number_of_replicates | N/A |
| Preincubation_time | 3min; 6 min |
| Quantification_method | HPLC, MS, Fluorimetry |
| kdeg_or_ksyn_of_the_enzyme | CYP3A4/5/7, P450 reductase, b5 |
| Protein | 1mol, 6.6mol, 9mol |
| Protein_Concentration | BD Gentest, PanVera, PanVera |
| Source | |
| Non_Recombinant-Enzymes | CYP3A |
| Recombinant_Enzymes | |
| Inhibitor_or_Inducer | |
| Multi_Drug_Experiments | |
| PK_Parameters | |
| Emax | |
| IC50 | |
| KI | |
| Ki | |
| Kinact | |
| Type_of_Interaction | |
| Single_Drug_Experiments | |
| PK_Paramaters | |
| CLint | CL for individual substrates |
| Km | Km for individual substrates |
| Vmax | Vmax for individual substrates |
| Substrate | MDZ, APZ, TZ, CLAR, TAM, DTZ, NIF, BFC, HFC, TEST, E2 |
| in-vivo | |
| Metabolism | |
| CYP1_family | |
| CYP2_family | |
| CYP3_family | CYP3A4, 3A5, 3A7 |
| CYP4_family | |
| CYP_4_families_other | |

**Figure 3** Annotated in vitro PK study using PK Ontology

## Chapter 3 CREATION OF PHARMACOKINETICS CORPUS[3]

### 3.1 Introduction

With the continuous growth of biomedical literature extracting information from biomedical literature by means of human annotators is a herculean task. Machine learning and NLP methods show tremendous promise in this area to help annotators keep tabs on the collection and summarization of biomedical data from literature that is unstructured.

For successful application of machine learning and NLP methods to automatically extract information from biomedical literature there is need for an annotated corpora. Availability of such corpora makes it feasible to develop algorithms that learn from the corpus and scale across the vast array of biomedical literature.

A well-annotated corpus can be put to use for following tasks in the biomedical domain:

- Named Entity Recognition (Recognition of gene, protein, disease mentions)
- Entity mention normalization. (Gene/protein name normalization)
- Relation Extraction (Extraction of relation between genes/proteins)

In biomedical domain GENIA [29] corpus is one of the most widely used semantically annotated corpus, along with corpora like MedTag[30], PennBioIE (http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2008T20), LINNEAUS [31] which are facilitate training of systems that perform Named Entity Recognition (NER) of various biological entities. Corpora like GNAT [32] are widely used for gene mention normalization. To extract protein-protein interactions (PPI) corpora like BioInfer [33], AIMed (ftp://ftp.cs.utexas.edu/pub/mooney/bio-data/), HPRD50

[3] This chapter is published as: Wu H-Y, Karnik S, Subhadarshini A, Wang Z, Philips S, Han X, Chiang C, Liu L, Boustani M, Rocha L *et al*: **An integrated pharmacokinetics ontology and corpus for text mining**. *BMC bioinformatics* 2013, **14**(1):35.

(http://www.bio.ifi.lmu.de/publications/RelEx/), IEPA (http://class.ee.iastate.edu/berleant/s/IEPA.htm ), LLL (http://genome.jouy.inra.fr/texte/LLLchallenge/) provide annotated PPI data that is used widely to develop PPI extraction methodologies. All these corpora serve as valuable tools for the community.

However, there is a lack of such corpus for the PK DDI domain and this has been our motivation to develop a semantically annotated corpus taking cues from the PK ontology developed in **Chapter 2** as our baseline.

### 3.2 Creation of PK Corpus

Our PK corpus consists of four broad classes of PK studies number of Pubmed abstracts manually annotated for each categories is represented in the parenthesis:

- Clinical PK studies (n = 56)
- Clinical pharmacogenetic studies (n = 57)
- *in vivo* DDI studies (n = 218)
- *in vitro* drug interaction studies (n = 210)

Abstracts of clinical PK studies were selected from previous work from Dr. Li's lab, in which the most popular CYP3A substrate, midazolam was investigated [34]. Clinical pharmacogenetic abstracts were selected based on the most polymorphic CYP enzyme, CYP2D6. The articles for *in vivo* and *in vitro* DDI studies were gathered by querying Pubmed in bulk via. eUtils interface with probe substrates/inhibitors/inducers for metabolism enzymes reported in **Table 6** as query terms. Once abstracts were collected we followed an annotation pipeline where we selected most relevant abstracts for inclusion in the corpus.

Abstracts collected in the previous step were annotated manually by a team of curators which included 3 masters and one Ph.D. students with different training backgrounds: computational science, biological science, and pharmacology respectively. In addition a random subset of 20% of the abstracts that had consistent annotations among four annotators, were double-checked and reviewed by two Ph.D. level scientists having extensive knowledge in pharmacology, drug interactions model based PK. Annotation workflow presented in **Figure 4** was applied to each of the four classes in the PK corpus. We annotated key entities like drug names, enzymes involved in drug metabolism, PK parameters, numerical values, units associated with the PK parameters, DDI mechanisms and change verbs as these components are vital in describing a PK study pertaining to DDI. Guidelines for the annotation of the above listed entities are described as follows:

*Drug Names:*

We used drug generic names from Drug Bank as our standards in tagging drugs in the abstracts. In addition to drug names we also tagged drug metabolites, as these are important in describing a PK DDI study. For tagging the metabolites we used linguistic cues from chemistry like presence of suffixes or prefixes like: *oxi*, *hydroxyl*, *methyl*, *acetyl*, *N-dealkyl*, *N-demethyl*, *nor*, *dihydroxy*, *O-dealkyl*, and *sulfo*. These prefixes and suffixes represents metabolites formed in phase I metabolism (oxidation, reduction, hydrolysis), and phase II of drug metabolism (methylation, sulphation, acetylation, glucuronidation) [35].

*Enzyme Names:*

We tagged all the CYP450 family of enzymes described in human cytochrome P450 allele nomenclature database, http://www. cypalleles.ki.se/. Variations of the enzyme or gene names were considered. We used following regular expression to identify CYP450 names

and variants in the text `(?:cyp|CYP|P450|CYP450)?[0-9][a-zA-Z][0-9]{0,2}(?:\*[0-9]{1,2})?$.`

*PK Parameters:*

We tagged PK parameter and respective units (if present in the abstract) according to **Table 2** and **Table 4**.

*Numerical Data:*

In addition to PK parameters we tagged numerical values associated with them along with any p-values mentioned in the abstracts.

*Mechanisms:*

For tagging the mechanism by which one drug affects other we resorted to use of verbs that are often used to describe DDI and metabolism of drugs. We made use of the following regular expression to tag the mechanisms:

```
inhibit(e(s|d)?|ing|ion(s)?|or)$, catalyz(e(s|d)?|ing)$,
correlat(e(s|d)?|ing|ion(s)?)$, metaboli(z(e(s|d)?|ing)|sm)$,
induc(e(s|d)?|ing|tion(s)?|or)$,
form((s|ed)?|ing|tion(s)?|or)$,
stimulat(e(s|d)?|ing|ion(s)?)$,
activ(e(s)?|(at)(e(s|d)?|ing|ion(s)?))$, and
suppress(e(s|d)?|ing|ion(s)?)$.
```

*Change:*

Numerical data associated with the PK parameters describes quantitative change, to address qualitative change of PK parameters following words were tagged in the corpus:

```
strong(ly)?, moderate(ly)?, high(est)?(er)?, slight(ly)?,
strong(ly)?, moderate(ly)?, slight(ly)?, significant(ly)?,
obvious(ly)?, marked(ly)?, great(ly)?, pronounced(ly)?,
modest(ly)?, probably, may, might, minor, little, negligible,
doesn’t interact, affect((s|ed)?|ing|ion(s)?)?$,
reduc(e(s|d)?|ing|tion(s)?)$, and increas(e(s|d)?|ing)$.
```

After tagging the relevant entities in the abstract we moved to the next step of annotation i.e. sentence level annotation which involved identifying sentence(s) which encompass information key to the DDI study that is the central topic of the abstract. We categorized these sentences into two types namely:

- *Clear DDI Sentence (CDDIS)*: two drug names (or drug-enzyme pair in the *in vitro* study) are in the sentence with a clear interaction statement, i.e. either interaction, or non-interaction, or ambiguous statement (i.e. such as possible or might and etc.).

- *Vague DDI Sentence (VDDIS)*: One drug or enzyme name is missed in the DDI sentence, but it can be inferred from the context. Clear interaction statement also is required.

CDDIS and VDDIS were further distilled into sub-categories as these sentences are of high value for annotation.

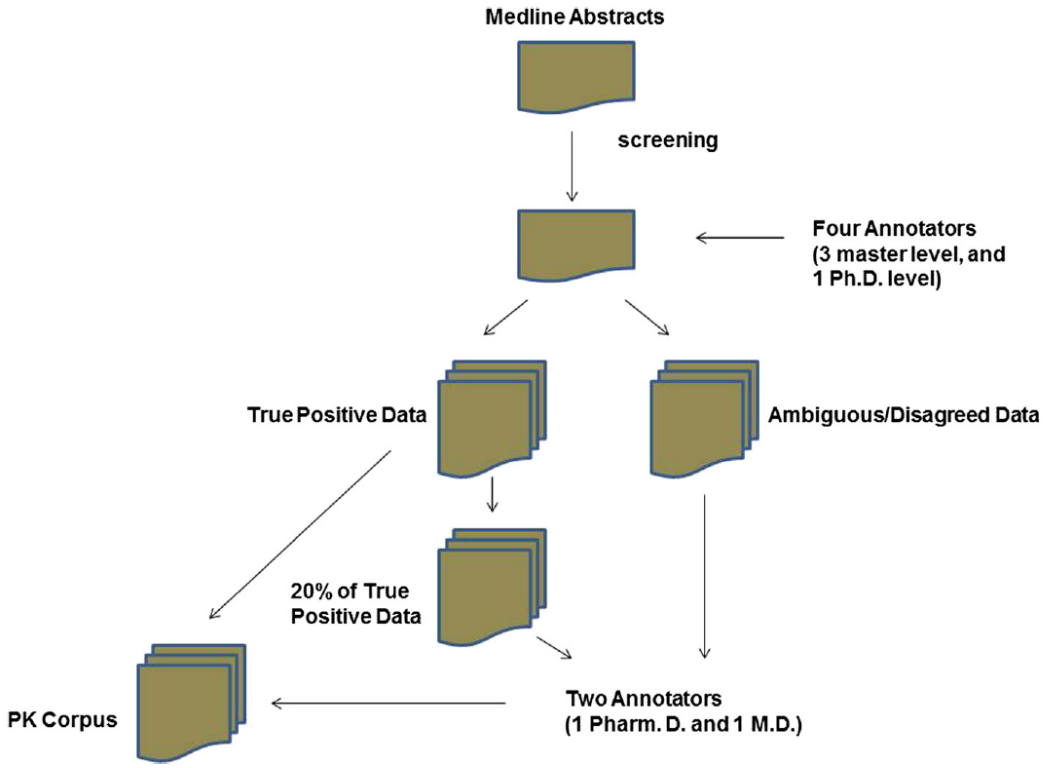


**Figure 4** PK Corpus Annotation Workflow

Due to fundamental difference between *in vivo* DDI studies and *in vitro* DDI studies, their DDI relationships were defined differently. In *in vivo* studies, three types of DDI relationships were defined (Table 8): DDI, ambiguous DDI (ADDI), and non-DDI (NDDI). Four conditions are specified to determine these DDI relationships. Condition 1 (C1) requires that at least one drug or enzyme name has to be contained in the sentence; condition 2 (C2) requires the other interaction drug or enzyme name can be found from the context if it is not from the same sentence; condition 3 (C3) specifies numeric rules to defined the DDI relationships based on the PK parameter changes; and condition 4 (C4) specifies the language expression patterns for DDI relationships. Using the rules summarized in **Table 8**, DDI, ADDI, and NDDI can be defined by C1 ∧ C2 ∧ (C3 ∨ C4). The priority rank of *in vivo* PK parameters is AUC > CL > $t_{1/2}$ > $C_{max}$. In *in vitro* studies, six types of DDI relationships were defined (**Table 8**). DDI, ADDI, NDDI were similar to *in vivo* DDIs, but three more drug-

enzyme relationships were further defined: DEI, ambiguous DEI (ADEI), and non-DDI (NDEI). C1, C2, and C4 remained the same for *in vitro* DDIs. The main difference is in C3, in which either Ki or IC50 (inhibition) or EC50 (induction) were used to defined DDI relationship quantitatively. The priority rank of *in vitro* PK parameters is Ki > IC50. In **Table 9** eight examples of how DDIs or DEIs were determined in the sentences.

| **DDI relationship** | **C1** | **C2** | **C3**** | **C4**** |
|---|---|---|---|---|
| **IN VIVO STUDY** | | | | |
| **DDI** | **Yes** | **Yes** | The PK parameter with the highest priority* must satisfy p-value <0.05 and FC > 1.50 or FC < 0.67 | Significant, obviously, markedly, greatly, pronouncedly and etc. |
| **Ambiguous DDI (ADDI)** | | | The PK parameter with the highest priority* in the conditions of p-value <0.05 but 0.67 < FC < 1.50; or FC >1.50 or FC <0.67, but p-value > 0.05. | Modestly, moderately, probably, may, might, and etc. |
| **Non-DDI (NDDI)** | | | The PK parameter with the highest priority*are in the condition of p-value > 0.05 and 0.67 < FC < 1.50 | Minor significance, slightly, little or negligible effect, doesn't interact etc. |
| **IN VITRO STUDY** | | | | |
| **DDI** | **Yes** | **Yes** | (0< Ki < 10 or 0< EC50 < 10 microM, and p-value <0.05) | Significant, obviously, markedly, greatly, pronouncedly and etc. |
| **DEI** | | | | |
| **Ambiguous DDI (ADDI)** | | | (10 < Ki < 100 or 10 < EC50 < 100 microM, and p-value <0.05 or vice versa) | Modestly, moderately, probably, may, might, and etc. |
| **Ambiguous DEI (ADEI)** | | | | |
| **Non-DDI (NDDI)** | | | (Ki > 100 microM or EC50 > 100 microM, and p-value >0.05) | Minor significance, slightly, little or negligible effect, doesn't interact etc. |
| **Non-DEI (NDEI)** | | | | |

**Table 8** DDI Categories in PK Corpus

Note:

C1: At least one drug or enzyme name has to be contained in the sentence.

C2: Need to label the drug name if it is not from the same sentence.

C3: PK-parameter and value dependent.

C4: Significance statement.

*Priority issue: When C3 and C4 occur and conflict, C3 dominates the sentence.**For the priority of PK parameters: AUC > CL > $t_{1/2}$ > $C_{max;}$; the priority of *in vitro* PK parameters: Ki>IC50.

| PMID | DDI sentence | Relationship and comment |
|---|---|---|
| **20012601** | The pharmacokinetic parameters of *verapamil* were *significantly* altered by the co-administration of *lovastatin* compared to the control. | Because of the words, "significantly", (Verapamil, lovastatin) is a **DDI**. |
| **20209646** | The *clearance* of *mitoxantrone* and *etoposide* was *decreased* by *64%* and *60%*, respectively, when combined with *valspodar*. | Because of the fold changes were less than 0.67, (*mitoxantrone*, *valspodar*.) and (*etoposide*, *valspodar*) are **DDIs**. |
| **20012601** | The *(AUC (0-infinity))* of *norverapamil* and the terminal *half-life* of *verapamil did not significantly changed* with *lovastatin* coadministration. | Because of the words, "not significantly changed", *(verapamil*, *ovastatin*) is a **NDDI**. |
| **17304149** | Compared with placebo, *itraconazole* treatment *significantly increase* the peak plasma concentration (*Cmax*) of paroxetine by 1.3 fold (6.7 2.5 versus 9.0 3.3 ng/mL, P≤0.05) and the area under the plasma concentration-time curve from zero to 48 hours [*AUC(*0–48)] of *paroxetine* by *1.5 fold* (137 73 versus 199 91 ng*h/mL, *P≤0.01*). | AUC has a higher rank than Cmax, and it had a 1.5 fold-change and less than 0.05 p-value, thus, (*itraconazole*, *paroxetine)* is a **DDI**. |
| **13129991** | The mean (SD) *urinary ratio* of *dextromethorphan* to its metabolite was *0.006* (0.010) at baseline and *0.014* (0.025) after *St John's wort* administration *(P=.26)* | The change in PK parameter is more than 1.5 fold but P-value is >0.05. Thus, (dextromethorphan, St John's wort) is an **ADDI**. |
| **19904008** | The obtained results show that *perazine* at its therapeutic concentrations is a *potent inhibitor* of human *CYP1A2*. | Because of words, "potent inhibitor", (perazine, CYP1A2) is a **DEI**. |
| **19230594** | After human hepatocytes were exposed to 10 microM *YM758*, microsomal activity and mRNA level for *CYP1A2* were *not induced* while those for *CYP3A4* were *slightly induced*. | Because of words, "not induced" and "slightly induced", (YM758, CYP1A2) and (YM758, CYP1A2) are **NDEIs**. |
| **19960413** | From these results, *DPT* was characterized to be a competitive *inhibitor* of *CYP2C9* and *CYP3A4*, with *K(i)* values of *3.5* and *10.8 microM* in HLM and *24.9* and *3.5* microM in baculovirus-insect cell-expressed human CYPs, respectively. | Because K was larger than 10microM, (DPT, CYP2C9) and (DPT, CYP3A4) are **ADEIs**. |

**Table 9** DDI Examples from PK Corpus

We calculate Krippendorff's alpha [36] to evaluate the reliability of annotations by our annotators, it serves as a measure of inter-annotator agreement. The frequencies of key entities, DDI sentences, and DDI pairs are presented in **Table 10** along with their

Krippendorff's alphas. Note that the total DDI pairs refer to the total pairs of drugs within a DDI sentence from all DDI sentences.

| Key Terms | Annotation Categories | Frequencies | Krippendorff's alpha |
|---|---|---|---|
| | Drug | 8633 | 0.953 |
| | CYP | 3801 | |
| | PK Parameter | 1508 | |
| | Number | 3042 | |
| | Mechanism | 2732 | |
| | Change | 1828 | |
| | Total words | 97291 | |
| DDI sentences | CDDI sentences | 1191 | 0.921 |
| | VDDI sentences | 120 | |
| | Total sentences | 4724 | |
| DDI Pairs | DDI | 1239 | 0.905 |
| | ADDI | 300 | |
| | NDDI | 294 | |
| | DEI | 565 | |
| | ADEI | 95 | |
| | NDEI | 181 | |
| | Total Drug Pairs | 12399 | |

**Table 10** Annotation Performance Summaries

Our corpus was constructed as follows to be machine-readable: raw abstracts were downloaded from PubMed in XML format. Then XML files were converted into GENIA corpus format following the document type definition (DTD) from the GENIA corpus [29]. The sentence detection in this step is accomplished by using the Perl module *Lingua::EN::Sentence*, from CPAN. Resulting corpus files were then tagged with the prescribed three levels of PK and DDI annotations. Finally, a cascading style sheet (CSS) was used to assign different colors for the entities in the corpus. This feature allows the users to visualize annotated entities. **Figure 5** presents example of *in vivo* DDI abstract from the corpus with the respective color legend. We would like to acknowledge that a DDI Corpus was recently published as part of a text mining competition DDIExtraction 2011

(http://labda.inf.uc3m.es/DDIExtraction2011/dataset.html). The DDIs in this corpus were clinical outcome oriented, not PK oriented. They were extracted from DrugBank, not from PubMed abstracts. Our PK corpus complements to their corpus very well. PK Corpus and PK Ontology described in Chapter 2 and associated data is available for download at http://rweb.biostat.iupui.edu/corpus/ and http://rweb.biostat.iupui.edu/ontology/ respectively.

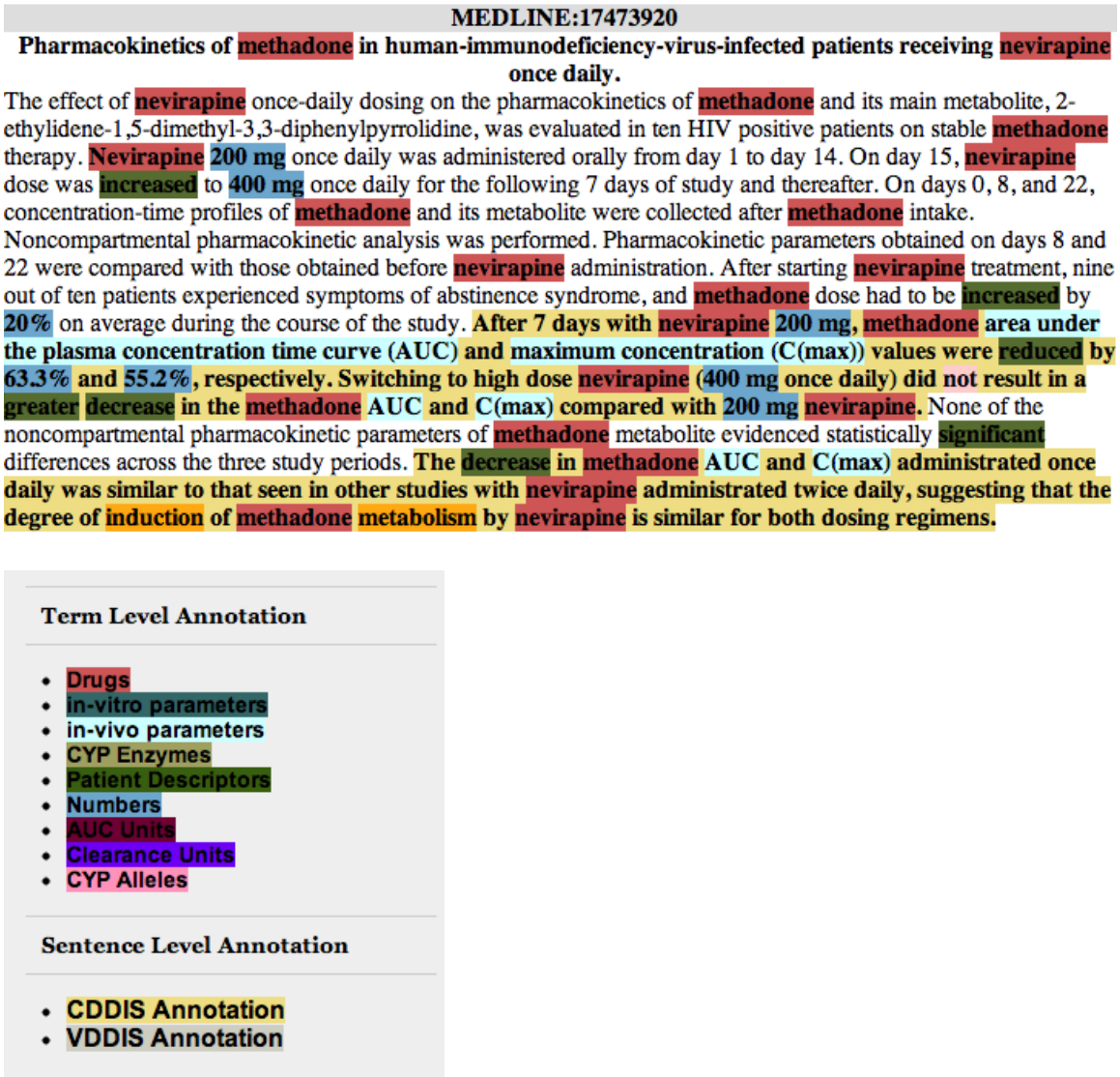


**Figure 5** Visual Example of Annotated *in vivo* DDI abstract

## Chapter 4 EXTRACTION OF DDI PAIRS FROM PK CORPUS[4]

### 4.1 Introduction

We demonstrate the usability of the PK corpus developed in Chapter 3 by utilizing the corpus for training a machine-learning model aimed at extracting DDI pairs from the corpus automatically. We also applied this approach on the DDIExtraction 2011 (http://labda.inf.uc3m.es/DDIExtraction2011/dataset.html) corpus as we participated in the DDIExtraction 2011 competition.

### 4.2 All paths graph kernel

We implemented the approach described by Airola et al. [37] for the DDI extraction. This approach has been previously applied for extraction of protein-protein interactions. Prior to performing DDI extraction, the testing and validation DDI abstracts in our corpus was pre-processed and converted into the unified XML format [37]. Then following steps were performed:

- Drugs were tagged in each of the sentences using dictionary based on DrugBank[14]. This step revised our prescribed drug name annotations in the corpus. One purpose is to reduce the redundant synonymous drug names. The other purpose is only keep the

[4] This chapter is published as: Wu H-Y, Karnik S, Subhadarshini A, Wang Z, Philips S, Han X, Chiang C, Liu L, Boustani M, Rocha L *et al*: **An integrated pharmacokinetics ontology and corpus for text mining**. *BMC bioinformatics* 2013, **14**(1):35. and Karnik S, Subhadarshini A, Wang Z, Rocha LM, Li L: **Extraction of Drug-Drug Interactions Using All Paths Graph Kernel**. In: *Drug-Drug Interaction Extraction (DDIExtraction 2011): 2011*.

parent drugs and remove the drug metabolites from the tagged drug names from our initial corpus, because parent drugs and their metabolites rarely interacts. In addition, enzymes (i.e. CYPs) were also tagged as drugs, since enzyme-drug interactions have been extensively studied and published. The regular expression of enzyme names in our corpus was used to remove the redundant synonymous gene names.

- Each of the sentences was subjected to tokenization, PoS tags and dependency tree generation using the Stanford parser [38].
- $C_2^n$ drug pairs form the tagged drugs in a sentence were generated automatically, and they were assigned with default labels as no-drug interaction. Please note that if a sentence had only one drug name, this sentence did not have DDI. This setup limited us considering only CDDI sentence in our corpus.
- The drug interaction labels were then manually flipped based on their true drug interaction annotations from the corpus. Please note that our corpus had annotated DDIs, ADDIs, NDDIs, DEIs, ADEIs, and NDEIs. Here only DDIs and DEIs were labeled as true DDIs. The other ADDIs, NDDIs, DEIs, and ADEIs were all categorized into the no-drug interactions.

Then sentences were represented with dependency graphs using interacting components (drugs) (**Figure 6**). The graph representation of the sentence was composed of two items: i) One dependency graph structure of the sentence; ii) a sequence of PoS (part-of-speech) tags (which was transformed to a linear order "graph" by connecting the tags with a constant edge weight). We used the Stanford parser [38] to generate the dependency graphs. Airola et al. proposed to combine these two graphs to one weighted, directed graph. This graph was fed into a support vector machine (SVM) for DDI/non-DDI classification. More details about the all paths graph kernel algorithm can be found in [39]. A graphical representation of the approach is presented in **Figure 6**.

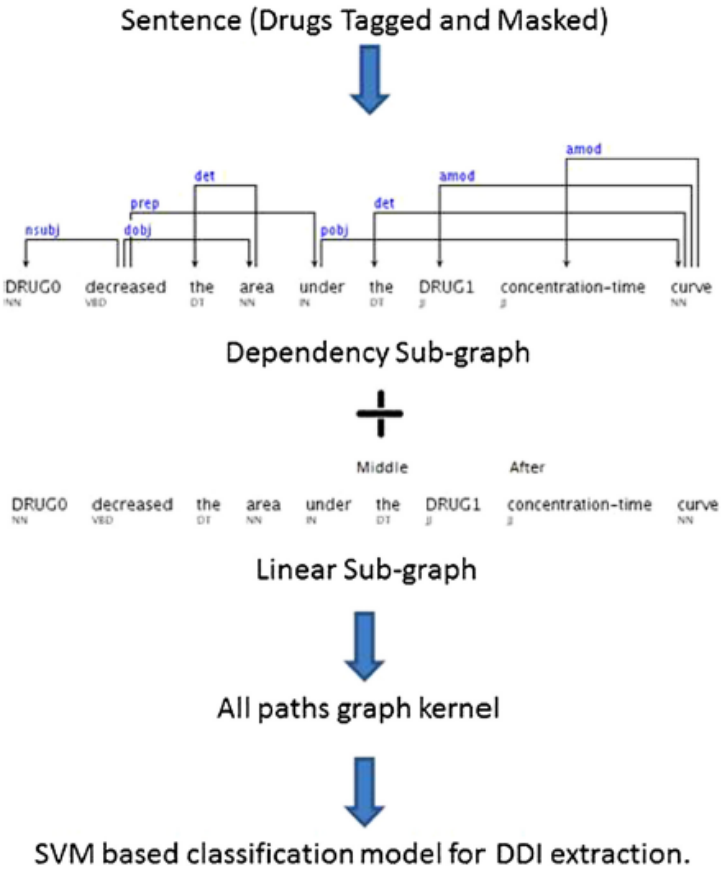


**Figure 6** Summary of All Paths Graph Kernel

DDI extraction was implemented for the *in vitro* and *in vivo* DDI corpora separately we split both the corpora into 70-30 fraction, we kept 30% fraction aside for testing the extraction performance. We also applied this method to the dataset from the DDIExtraction 2011 corpus.

**Table 11** presents the training sample size and testing sample size in all corpus sets and **Table 12** presents the performance of DDI Extraction.

| Dataset | Abstracts | Sentences | DDI Pairs | True DDI Pairs |
|---|---|---|---|---|
| *in vivo* DDI Training | 174 | 2112 | 2024 | 359 |
| *in vivo* DDI Testing | 44 | 545 | 574 | 45 |
| *in vitro* DDI Training | 168 | 1894 | 7122 | 783 |
| in vitro DDI Testing | 42 | 475 | 1542 | 146 |
| DDIExtraction 2011 Training | NA | 3621 | NA | 20888 |
| DDIExtraction 2011 Testing | NA | 1539 | NA | 7036 |

**Table 11** Summary of the Datasets used for DDI Extraction

| Datasets | Precision | Recall | F-measure |
|---|---|---|---|
| *in vivo* DDI Training | 0.67 | 0.78 | 0.72 |
| in vivo DDI Testing | 0.67 | 0.79 | 0.73 |
| *in vitro* DDI Training | 0.51 | 0.59 | 0.55 |
| *in vitro* DDI Testing | 0.47 | 0.58 | 0.52 |
| DDIExtraction 2011 Training | 0.42 | 0.42 | 0.42 |
| DDIExtraction 2011 Testing | 0.14 | 0.12 | 0.17 |

**Table 12** Summary of Performance of DDI Extraction

Error analysis was performed in test data to evaluate the extraction algorithm quality. **Table 13** summarizes the results. Among the known reasons for the false positives and false negatives, the most frequent one is that there are multiple drugs in the sentence, or the sentence is long. The other reasons include that there is no direct DDI relationship between two drugs, but the presence of some words, such as dose, increase, and etc., may lead to a false positive prediction; or DDI is presented in an indirect way; or some NDDI are inferred due to some adjectives (little, minor, negligible).

| No. | Error Categories | Error type | Frequency | | Examples |
|---|---|---|---|---|---|
| | | | In vivo | In vitro | |
| 1 | There are multiple drugs in the sentence, and the sentence is long. | FP | 6 | 34 | PMID: 12426514. In 3 subjects with measurable concentrations in the single-dose study, rifampin significantly decreased the mean maximum plasma concentration (C(max)) and area under the plasma concentration-time curve from 0 to 24 h [AUC(0–24)] of praziquantel by 81% (P <.05) and 85% (P <.01), respectively, whereas rifampin significantly decreased the mean C(max) and AUC(0–24) of praziquantel by 74% (P <.05) and 80% (P <.01), respectively, in 5 subjects with measurable concentrations in the multiple-dose study |
| | | FN | 2 | 17 | PMID: 10608481. Erythromycin and ketoconazole showed a clear inhibitory effect on the 3-hydroxylation of lidocaine at 5 microM of lidocaine (IC50 9.9 microM and 13.9 microM, respectively), but did not show a consistent effect at 800 microM of lidocaine (IC50 >250 microM and 75.0 microM, respectively). |

| | | | | | |
|---|---|---|---|---|---|
| 2 | There is no direct DDI relationship between two drugs, but the presence of some words, such as dose, increase, and etc. may lead to a false positive prediction | FP | 6 | 14 | PMID: 17192504. A significant fraction of patients to be treated with HMR1766 is expected to be maintained on warfarin |
| 3 | DDI is presented in an indirect way. | FN | 2 | 19 | PMID: 11994058. In CYP2D6 poor metabolizers, systemic exposure was greater after chlorpheniramine alone than in extensive metabolizers, and administration of quinidine resulted in a slight increase in CLoral. |
| 4 | Design issue. Some NDDI are inferred due to some adjectives (little, minor, negligible) | FP | 1 | 3 | PMID: 10223772. In contrast,the effect of ranitidine or ebrotidine on CYP3A activity *in vivo* seems to have little clinical significance. |
| 5 | Unknown | FP | 5 | 44 | PMID: 10383922. CYP1A2, CYP2A6, and CYP2E1 activities were not significantly inhibited by azelastine and the two metabolites. |
| | | FN | 6 | 26 | PMID: 10681383. However, the most unusual result was the interaction between testosterone and nifedipine. |

**Table 13** Error Analyses from Test Data

## Chapter 5: CONCLUSIONS AND FUTURE DIRECTIONS

In this work we developed PK ontology and used the annotation guidelines to assemble the PK corpus and used the PK corpus to demonstration application of machine learning and NLP to extract DDI pairs from unstructured text. Our annotation pipeline is very strong thanks to the collective experience of highly experienced team for mentors in Dr Li's group our group used similar annotation technique to demonstrate use of biomedical text to understand DDIs [40].

There are certain areas where we can improve like the performance of extraction of DDI pairs, which advocates for using new methods to make use of high quality PK corpus which will in turn will improve the performance of DDI extraction and facilitate creation of comprehensive PK DDI database which will be an important asset for the research community.

REFERENCES


1. Becker ML, Kallewaard M, Caspers PWJ, Visser LE, Leufkens HGM, Stricker BH: **Hospitalisations and emergency department visits due to drug–drug interactions: a literature review**. *Pharmacoepidemiology and Drug Safety* 2007, **16**(6):641-651.
2. Hamilton RA, Briceland LL, Andritz MH: **Frequency of Hospitalization after Exposure to Known Drug-Drug Interactions in a Medicaid Population**. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* 1998, **18**(5):1112-1120.
3. Jankel C, Fitterman L: **Epidemiology of Drug-Drug Interactions as a Cause of Hospital Admissions**. *Drug-Safety* 1993, **9**(1):51-59.
4. Juurlink DN, Mamdani M, Kopp A, Laupacis A, Redelmeier DA: **Drug-drug interactions among elderly patients hospitalized for drug toxicity**. *JAMA : the journal of the American Medical Association* 2003, **289**(13):1652-1658.
5. Sabers A, Gram L: **Newer anticonvulsants: comparative review of drug interactions and adverse effects**. *Drugs* 2000, **60**(1):23-33.
6. Hirschman L, Yeh A, Blaschke C, Valencia A: **Overview of BioCreAtIvE: critical assessment of information extraction for biology**. *BMC bioinformatics* 2005, **6 Suppl 1**:S1.
7. Duda S, Aliferis C, Miller R, Statnikov A, Johnson K: **Extracting drug-drug interaction articles from MEDLINE to improve the content of drug databases**. *AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium* 2005:216-220.
8. Tari L, Anwar S, Liang S, Cai J, Baral C: **Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism**. *Bioinformatics* 2010, **26**(18):i547-553.
9. Segura-Bedmar I, Martinez P, de Pablo-Sanchez C: **A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents**. *BMC bioinformatics* 2011, **12 Suppl 2**:S1.
10. Segura-Bedmar I, Martínez P, de Pablo-Sánchez C: **Using a shallow linguistic kernel for drug–drug interaction extraction**. *Journal of Biomedical Informatics* 2011, **44**(5):789-804.
11. Payne PRO: **Chapter 1: Biomedical Knowledge Integration**. *PLoS computational biology* 2012, **8**(12):e1002826.
12. Rubin DL, Lewis SE, Mungall CJ, Misra S, Westerfield M, Ashburner M, Sim I, Chute CG, Solbrig H, Storey MA *et al*: **National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge**. *Omics : a journal of integrative biology* 2006, **10**(2):185-198.
13. 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):25-29.
14. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V *et al*: **DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs**. *Nucleic Acids Research* 2011, **39**(suppl 1):D1035-D1041.

15. Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, Altman RB, Klein TE: **PharmGKB: the Pharmacogenetics Knowledge Base**. *Nucleic Acids Research* 2002, **30**(1):163-165.
16. Rubin D, Noy N, Musen M: **Protege: a tool for managing and using terminology in radiology applications**. *J Digit Imaging* 2007, **20**(Suppl 1):34 - 46.
17. Segel H: **Enzyme kinetics – behavior and analysis of rapid equilibrium and steady state enzyme systems**. New York: John Wiley & Sons, Inc; 1975.
18. International Transporter C, Giacomini KM, Huang SM, Tweedie DJ, Benet LZ, Brouwer KL, Chu X, Dahlin A, Evers R, Fischer V *et al*: **Membrane transporters in drug development**. *Nat Rev Drug Discov* 2010, **9**(3):215-236.
19. Rostami-Hodjegan A, Tucker G: **"In silico" simulations to assess the "in vivo" consequences of "in vitro" metabolic drug-drug interactions**. *Drug Discovery Today: Technologies* 2004, **1**:441 - 448.
20. M. R, TN. T: **Clinical pharmacokinetics concept and applications.** London: Lippincott Williams & Wilkins; 1995.
21. Gibaldi M, Perrier D: **Pharmacokinetics**, 2 edn. New York: Marcel Dekker; 1982.
22. Guengerich F: **Cytochrome p450 and chemical toxicology**. *Chemical research in toxicology* 2008, **21**(1):70 - 83.
23. LePendu P, Musen MA, Shah NH: **Enabling enrichment analysis with the Human Disease Ontology**. *J Biomed Inform* 2011, **44 Suppl 1**:S31-38.
24. Coulet A, Smaïl-Tabbone M, Napoli A, Devignes M-D: **Suggested Ontology for Pharmacogenomics (SO-Pharm): Modular Construction and Preliminary Testing**. In: *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*. Edited by Meersman R, Tari Z, Herrero P, vol. 4277: Springer Berlin Heidelberg; 2006: 648-657.
25. Smith C, Goldsmith C-A, Eppig J: **The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information**. *Genome Biology* 2004, **6**(1):R7.
26. Borges S, Desta Z, Jin Y, Faouzi A, Robarge JD, Philips S, Nguyen A, Stearns V, Hayes D, Rae JM *et al*: **Composite functional genetic and comedication CYP2D6 activity score in predicting tamoxifen drug exposure among breast cancer patients**. *J Clin Pharmacol* 2010, **50**(4):450-458.
27. Chien JY, Lucksiri A, Ernest CS, 2nd, Gorski JC, Wrighton SA, Hall SD: **Stochastic prediction of CYP3A-mediated inhibition of midazolam clearance by ketoconazole**. *Drug Metab Dispos* 2006, **34**(7):1208-1219.
28. Williams JA, Ring BJ, Cantrell VE, Jones DR, Eckstein J, Ruterbories K, Hamman MA, Hall SD, Wrighton SA: **Comparative metabolic capabilities of CYP3A4, CYP3A5, and CYP3A7**. *Drug Metab Dispos* 2002, **30**(8):883-891.
29. Kim JD, Ohta T, Tateisi Y, Tsujii J: **GENIA corpus--semantically annotated corpus for bio-textmining**. *Bioinformatics* 2003, **19 Suppl 1**:i180-182.
30. Smith L, Tanabe L, Rindflesch T, Wilbur J: **MedTag: A Collection of Biomedical Annotations**. In: *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics: 2005*. Association for Computational Linguistics: 32-37.

31. Gerner M, Nenadic G, Bergman C: **LINNAEUS: A species name identification system for biomedical literature**. *BMC bioinformatics* 2010, **11**(1):85.
32. Hakenberg J, Gerner M, Haeussler M, Solt I, Plake C, Schroeder M, Gonzalez G, Nenadic G, Bergman CM: **The GNAT library for local and remote gene mention normalization**. *Bioinformatics* 2011, **27**(19):2769-2771.
33. Pyysalo S, Ginter F, Heimonen J, Bjorne J, Boberg J, Jarvinen J, Salakoski T: **BioInfer: a corpus for information extraction in the biomedical domain**. *BMC bioinformatics* 2007, **8**:50.
34. Wang Z, Kim S, Quinney SK, Guo Y, Hall SD, Rocha LM, Li L: **Literature mining on pharmacokinetics numerical data: a feasibility study**. *J Biomed Inform* 2009, **42**(4):726-735.
35. Brunton L, Chabner B, Knollmann B: **Goodman & Gilman's The Pharmacological Basis Of Therapeutics**. New York: McGraw-Hill.
36. Krippendorff K. Thousand Oaks, C: SAGE Publications Inc; 2004.
37. Airola A, Pyysalo S, Bjorne J, Pahikkala T, Ginter F, Salakoski T: **All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning**. *BMC bioinformatics* 2008, **9**(Suppl 11):S2.
38. Marneffe M, Maccartney B, Manning C: **Generating Typed Dependency Parses from Phrase Structure Parses**. In: *Proceedings of LREC-06: 2006*. 449-454.
39. Gärtner T, Flach P, Wrobel S: **On Graph Kernels: Hardness Results and Efficient Alternatives**. In: *Learning Theory and Kernel Machines.* Edited by Schölkopf B, Warmuth M, vol. 2777: Springer Berlin Heidelberg; 2003: 129-143.
40. Duke JD, Han X, Wang Z, Subhadarshini A, Karnik SD, Li X, Hall SD, Jin Y, Callaghan JT, Overhage MJ *et al*: **Literature based drug interaction prediction with clinical assessment using electronic medical records: novel myopathy associated drug interactions**. *PLoS computational biology* 2012, **8**(8):e1002614.

# Shreyas Karnik

## Education

- **Indiana University, School of Informatics** — Indianapolis, IN
  *Masters in Bioinformatics* — *2009 – 2012*
  - GPA: 3.7
- **Dr. D.Y. Patil University** — Pune, India
  *Bachelor of Technology in Bioinformatics* — *2004 – 2008*
  - GPA: 3.62

## Experience

- **ZenDeals** — Redmond, WA
  *Software Engineer* — *July 2012 – Present*
- **Marshfield Clinic, Marshfield WI** — Biomedical Informatics Research Center
  *Programmer/Analyst* — *Jan. 2012 – July 2012*
  - Providing research support to biomedical informatics projects.
- **Center for Computational Biology and Bioinformatics** — Indiana University
  *Scientific Programmer* — *Sept. 2009 – Jan. 2012*
  - Literature mining based on Pharmacokinetics Literature.
- **National Chemical Laboratory** — Pune, India
  *Project Assistant* — *July 2008 - July 2009*
  - Worked on Elucidation of the role of primary structure of proteins on their antimicrobial activity.
  - Project was funded by Department of Science and Technology, Government of India.
  - The system developed as a part of this project is live at CAMP:Collection of Anti-Microbial Peptides
- **Tata Research Development and Design Centre** — Pune, India
  *Project Trainee* — *Jan. 2008 - July 2008*
  - Worked on Analysis of proteins from Visualization and Machine Learning Perspective.

## Publications

1. Wu, H., **Karnik, S.**, Subhadarshini, A., Wang, Z., Philips, S.,Han, X., Chiang, C., Liu, L., Boustani, M., Rocha, L., Quinney, S., and Flockhart, D., and Li, L. An integrated pharmacokinetics ontology and corpus for text mining *BMC Bioinformatics* (2013) doi:10.1186/1471-2105-14-35

2. Duke, J.D., Han, X., Wang, Z., Subhadarshini, A., **Karnik, S.**, Li, X., Hall, S.D., Jin, Y., Callaghan, J.T., Overhage, M.J. and Li L., Literature Based Drug Interaction Prediction with Clinical Assessment Using Electronic Medical Records: Novel Myopathy Associated Drug Interactions *PLoS Computational Biology*, (2012) doi:10.1371/journal.pcbi.1002614

3. Joseph, S., **Karnik, S.**, Nilawe, P., Jayaraman, VK. and Idicula-Thomas, S. ClassAMP: A Prediction Tool for Classification of Antimicrobial Peptides IEEE/ACM *Transactions on Computational Biology and Bioinformatics (TCBB)*, (2012) doi:10.1109/TCBB.2012.8

4. Chowdhary R., Tan S.L., Zhang J., **Karnik S.**, Bajic V.B. and Liu J.S. Context-Specific Protein Network Miner − An Online System for Exploring Context−Specific Protein Interaction Networks from the Literature *PLoS ONE*, 7(4): e34480. (2012) doi:10.1371/journal.pone.0034480

5. **Karnik, S.**, A. Subhadarshini, Z. Wang, L.M. Rocha, and L. Li. Extraction of drug-drug interactions using all paths graph kernel. *In: Drug-Drug Interaction Extraction 2011 (DDIExtraction2011).* September, 7th, 2011, Huelva, Spain, In Press.

6. Kulkarni A., **Karnik, S.** and Angadi S. Analysis of Intrinsically Disordered Regions in Proteins using Recurrence Quantification Analysis *International Journal of Bifurcation and Chaos*, Vol. 21, No. 4 (2011) 1193-1202 doi:10.1142/S0218127411028969

7. Thomas, S., **Karnik, S.**, Barai, R. S., Jayaraman, V. K. and Idicula-Thomas, S, CAMP: a useful resource for research on antimicrobial peptides. *Nucleic Acids Research* 38, D774-D780 (2010) doi: 10.1093/nar/gkp1021.

8. **Karnik, S.**, Prasad, A., Diwevedi, A., Sundararajan, V. and Jayaraman, V. Identification of Defensins Employing Recurrence Quantification Analysis and Random Forest Classifiers in *Pattern Recognition and Machine Intelligence Vol. 5909* Lecture Notes in Computer Science 152-157 (Springer Berlin / Heidelberg, 2009). doi: 10.1007/978-3-642-11164-8_25

9. **Karnik, S.**, Prasad, A., Diwevedi, A., Sundararajan, V. and Jayaraman, V. Identification of N-Glycosylation Sites with Sequence and Structural Features Employing Random Forests in *Pattern Recognition and Machine Intelligence Vol. 5909* Lecture Notes in Computer Science 146-151 (Springer Berlin / Heidelberg, 2009). doi: 10.1007/978-3-642-11164-8_24

## Technical Skills

- Programming Languages
  - Python, Perl, R, Matlab, C/C++, Java
- Special Packages
  - Experience with BioConductor Packages for Bioinformatics.
  - Experience with Data Mining, Text Mining and Machine Learning tools.
- Databases
  - Oracle, MySQL, NoSQL
- Markup Languages
  - LaTeX,HTML,XML

## Awards

- Dean's Award worth 8000 USD form School of Informatics (from 2009 to 2011), Indiana University - Purdue University, Indianapolis.
- Travel award to attend Machine Learning Summer School at Purdue University from the Organizers of the Summer School.