# CHAPTER 3 MATERIALS AND METHODOLOGY

## 3.1 Datasets

In order to drive the experiments, various datasets were collected to construct the database. Database for ADRs took advantage of five data sources which were applied for six purposes. Firstly, The Medical Dictionary for Regulatory Activities (MedDRA) was conducted for retrieving ADRs terminology. Second, data from HUGO Gene Nomenclature Committee (HGNC) were applied for the conversion of protein nomenclature. Third, the information of drugs and their targets were collected from DrugBank in which drugs were additionally given the probabilistic score for ADRs class using Weka. The fourth information was relationships between ADRs and drugs that had been recorded in Canada Vigilance Adverse Reaction Online Database. Lastly, PubMed and Taverna were utilized for finding the relationships between ADRs and proteins. Figure 3.1 was simplified five data sources and six purposes. Moreover, details of datasets were described below.

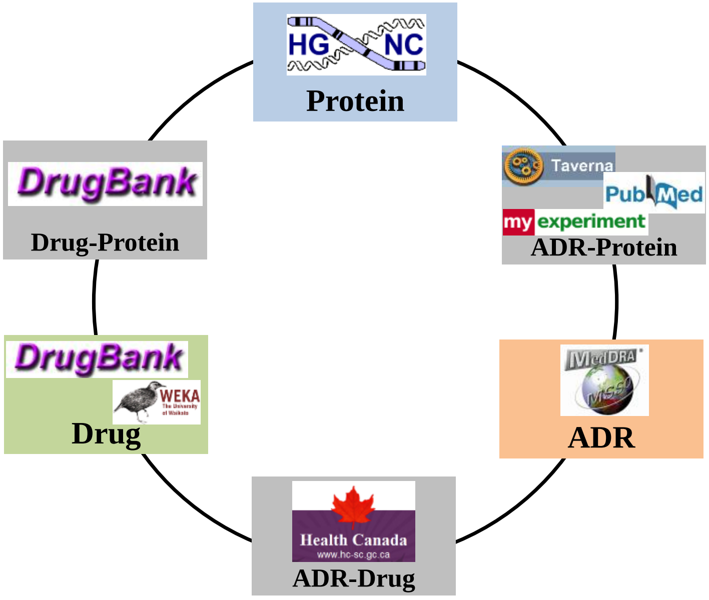


**Figure 3.1** Five data sources and six purposes that were utilized in the database for ADRs

### 3.1.1 The Medical Dictionary for Regulatory Activities (MedDRA)

MedDRA (International Federation of Pharmaceutical Manufacturers and Associations, 2010) is a pragmatic, clinically validated terminology that applies to all phases of drug development, excluding animal toxicology. In addition, it is the adverse event classification dictionary approved by the International Conference on Harmonization (ICH). Standardized MedDRA queries are developed to facilitate retrieval of MedDRA-coded data as a first step in investigating drug safety issues in pharmacovigilance and clinical development. The developers of the terminology design a structure that promotes specific and comprehensive data entry and flexible data retrieval. Figure 3.2 represented the hierarchical structure of the terminology.

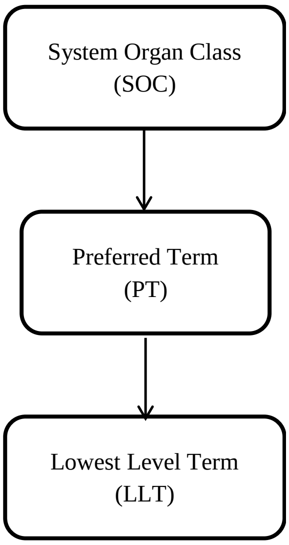


**Figure 3.2** Structural Hierarchy of the MedDRA Terminology

The latest version of MedDRA was 13.1 that was released on September 2010. The Maintenance and Support Services Organization (MSSO), the repository, maintainer, and distributor of MedDRA, also provided a desktop browser at no additional cost to subscribers. In addition, subscriber can access to an online web-based. A username and password was also required to access the web-based browser.

#### 3.1.1.1 System Organ Class (SOC)

A SOC is the highest level of the hierarchy that provides the broadest concept for data. SOCs are grouped by etiology, manifestation site, and purpose. There were 26 SOCs in MedDRA version 13.1. Code, name, and abbreviation of each SOC was shown in Table 3.1

#### 3.1.1.2 Preferred Term (PT)

A PT is a distinct descriptor or single medical concept. PT should be unambiguous and as specific and self-descriptive as possible. In MedDRA version 13.1, there were 18,919 preferred terms.

#### 3.1.1.3 Lowest Level Term (LLT)

LLT sets up the lowest level of the terminology. Each LLT is linked to only one PT. LLT has any of the following relationships to their parent PT: synonyms, lexical variants, quasi-synonyms, sub-element, or identical LLT. The LLT level plays an important role in facilitating the transfer of historical data because many of the terms from other terminologies incorporated, are represented at this level. In MedDRA version 13.1, there are 68,661 lowest level terms.

**Table 3.1** Twenty-six System Organ Classes in MedDRA version 13.1

| SOC Code | SOC Name | SOC Abbreviation |
|---|---|---|
| 10005329 | Blood and lymphatic system disorders | Blood |
| 10007541 | Cardiac disorders | Card |
| 10010331 | Congenital, familial and genetic disorders | Cong |
| 10013993 | Ear and labyrinth disorders | Ear |
| 10014698 | Endocrine disorders | Endo |
| 10015919 | Eye disorders | Eye |
| 10017947 | Gastrointestinal disorders | Gastr |
| 10018065 | General disorders and administration site conditions | Genrl |
| 10019805 | Hepatobiliary disorders | Hepat |
| 10021428 | Immune system disorders | Immun |
| 10021881 | Infections and infestations | Infec |
| 10022117 | Injury, poisoning and procedural complications | Inj&P |
| 10022891 | Investigations | Inv |
| 10027433 | Metabolism and nutrition disorders | Metab |
| 10028395 | Musculoskeletal and connective tissue disorders | Musc |
| 10029104 | Neoplasms benign, malignant and unspecified (incl cysts and polyps) | Neopl |
| 10029205 | Nervous system disorders | Nerv |
| 10036585 | Pregnancy, puerperium and perinatal conditions | Preg |
| 10037175 | Psychiatric disorders | Psych |
| 10038359 | Renal and urinary disorders | Renal |
| 10038604 | Reproductive system and breast disorders | Repro |
| 10038738 | Respiratory, thoracic and mediastinal disorders | Resp |
| 10040785 | Skin and subcutaneous tissue disorders | Skin |
| 10041244 | Social circumstances | SocCi |
| 10042613 | Surgical and medical procedures | Surg |
| 10047065 | Vascular disorders | Vasc |

In the task, MedDRA database provided the information about ADR terminology. It assisted the data from Canada Vigilance Adverse Reaction Online Database to link ADR and drug information. Database for ADRs applied SOC in the process of drug scoring for possible ADR class, PT as the beginning of ADR-protein relationship discovery using literature mining tool, and LLT for complete ADR term exploration in the database.

### 3.1.2 HUGO Gene Nomenclature Committee (HGNC)

The problems of nomenclature in human genetics were recognized since 1960s. The full guidelines for human gene nomenclature were presented at the Edinburgh Human Genome Meeting in 1979. Since then many attempts have continued to compromise between convenience and simplicity required for the use of human gene nomenclature. Human Gene Organization (HUGO) is the host to approve a gene name and symbol (short-form abbreviation) and store in HUGO Gene Nomenclature Committee (HGNC) database (HGNC and Wellcome Trust, 2010). For over thirty years, HGNC intended to approve gene symbols and corresponding gene name (Seal, et al., 2010). HGNC is an excellent resource to convert gene coded protein into the same standard. As for October 2010, the information of protein-coding gene was 19,361 records that supported the conversion of proteins information in the database.

### 3.1.3 DrugBank

From the previous reviews about database in pharmacology in Chapter 2, DrugBank was thus an outstanding database that should be applied as a resource for drugs and drug-target protein information (Yildirim, et al., 2007). As of information retrieval on October 2010, the last information of DrugBank was 4,774 in total number of drugs and 8,507 in association between drugs and their targets. Moreover, drugs were later given the predictive score for ADR class using data mining approach.

### 3.1.4 Canada Vigilance Adverse Reaction Online Database

Canada Vigilance Adverse Reaction Online Database (Health Canada and Government of Canada, 2010) is the program that is managed by MedEffect™ Canada for improving health product safety. Its reported information is about suspected adverse reactions or side effects to drugs and other health products, for instance, prescription and non-prescription medication, natural health products, and radiopharmaceuticals. This database does not include preventative vaccines, blood, blood components, medical devices, and cosmetics. Consumers, patients and health professionals can report adverse reactions to the Canada Vigilance Program via mail, report online, fax, or telephone (toll-free).

The restrictions of Canada Vigilance Adverse Reaction Online Database are the data cannot be used on its own to evaluate a health product's safety profile or calculate the incidence of an adverse even. There is no assurance that the reported events are actually due to the health product and there is not warrant that all adverse event reports are received by MedEffect™ Canada. Because of the limitations of the data, a thoughtful decision has to arise to use this information.

As of September 2010, data of Canada Vigilance Adverse Reaction Online Database covered time period from 1965 to March 31, 2010. It is updated four times a year. The data can be downloaded as zip file format and be extracted to flat file format. The flat file format comprises of 12 files: reports, drug product, drug product ingredients, reactions, outcome, gender, report feature, report type, seriousness, source, report links, and report drug. For extracting ADRs-drug relationship, database for ADRs only employed reactions and report drug file that provided the information about the reaction terms and drug associated with report.

### 3.1.5 PubMed

Competent access to information is needed for life science research (Krallinger, et al., 2008). The online scientific literature collections participate in the initial stage of experiment designing to the final interpretation and also communication. PubMed is the largest biomedical and other life science online database which faces double-exponential growth of citations and abstracts. It collects over 19 million citations and abstracts (Yu, et al., 2007; Lok, 2010) that are explored more than 70 million times per month. The content of PubMed is accessible through Entrez, a text-based search and retrieval system. Database for ADRs took advantage of literature mining in this centralized literature repository for ADR and protein relationship.

## 3.2 Applications and Tools

### 3.2.1 Weka

Weka (Witten and Frank, 2005) is a collection of state-of-the-art machine learning software written in Java, developed at the University of Waikato, New Zealand. The workbench includes methods for all the standard data mining problems: regression, classification, clustering, association rule mining and attribute selection. Weka is free software available under the general public license. Weka provides an easy graphical user interface that can be used to process even large datasets. The learning methods called classifiers were applied for the projects to make the predictive score of drugs for ADR class.

### 3.2.2 Literature Mining Tools

To extract the relationships between ADRs and proteins, literature mining tools were applied for the experiments. There were two literature mining approaches that were qualified to the research.

#### 3.2.2.1 Bio-NLP resources database (BioNLPdb)

Literature mining tools were selected from the depository; Bio-NLP resources database (BioNLPdb). This depository provided a compendium of IR, IE, text mining, and literature processing applications. BioNLPdb is especially developed for providing an access to information in life sciences and biomedical literature. Additionally, links to some relevant scientific literature repositories and search engines are provided. As October 2010, BioNLPdb reserved 150 applications which can be accessed at http://zope.bioinfo.cnio.es/bionlp_tools. The two most impressive tools from this depository were EAGLi and FACTA.

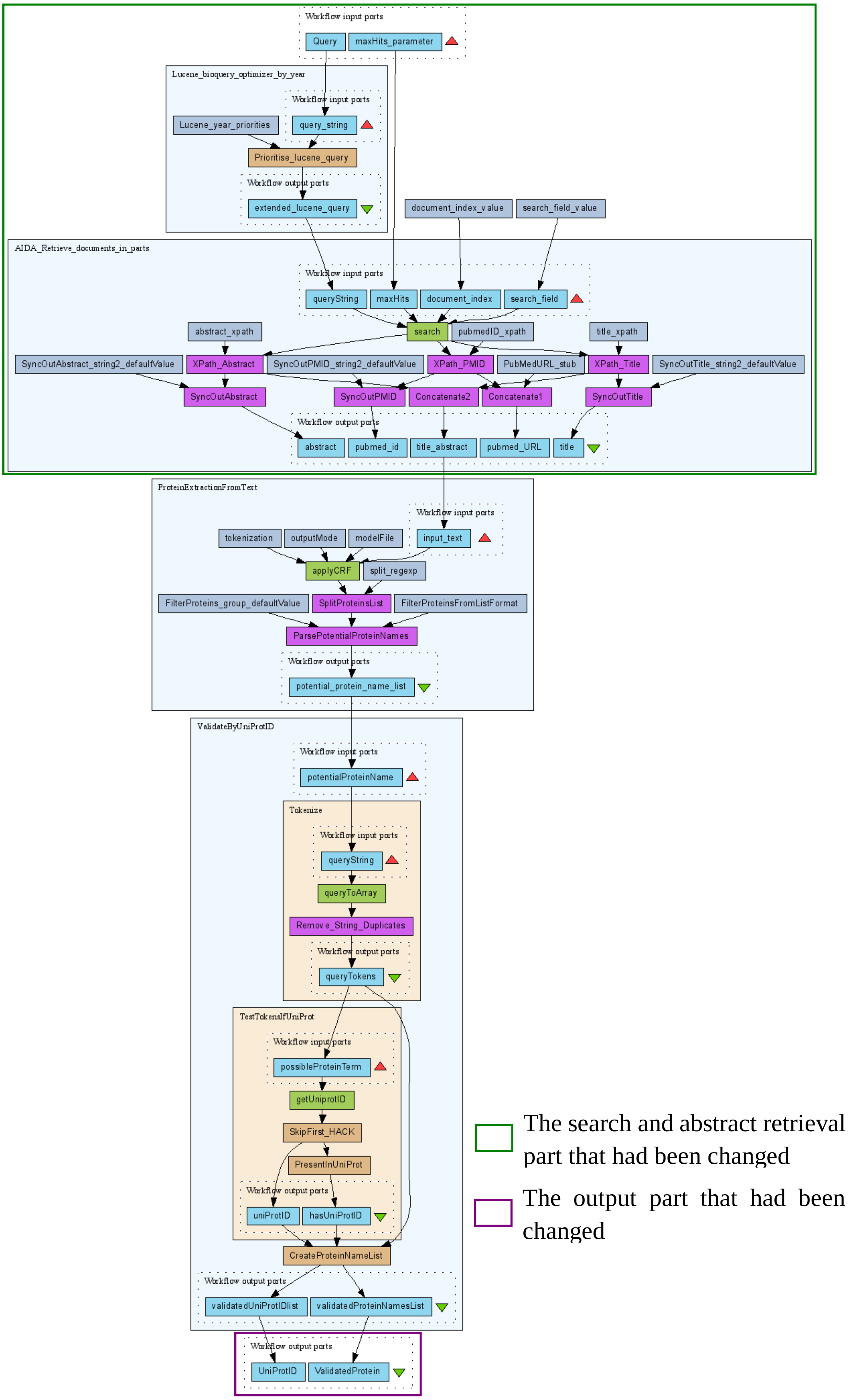
EAGLi and FACTA were easy to use via their friendly user interfaces and their rapid for returning the result. Unfortunately, these tools were difficult to deal with a large amount of data. Their input queries and output results needed to be ‘cut and paste’ which were laborious to repeat and examine. Finally, the usage of EAGLi and FACTA had been dripped.

### 3.2.2.2 Literature Mining with Taverna

Taverna is a part of myGrid project that has developed a tool for the composition and enactment of bioinformatics workflows for the life sciences community (Oinn, et al., 2004; Hull, et al., 2006; Lanzen and Oinn, 2008). The Taverna suite is written in Java. The tool implements a workbench application which presents a graphical user interface for the composition of workflows. The graphical of processors transforms a set of data inputs into a set of data outputs. Each step within a workflow represents one atomic task. It allows users to construct workflows or pipelines of services for performing different analyses. By integrating many several molecular biology tools and databases available on the web, especially web services, these high-level workflows can generate different resources into a single analysis. The application of Taverna offers an environment to access web services, without technical knowledge of web services or programming. This means Taverna allows users who are not necessarily expert programmers to design, execute, and share workflows of web services.

Another benefit of using Taverna is the user community, for instance, myExperiment. The myExperiment (Goble, et al., 2010) enables users to discover, manipulate, and distribute scientific workflows which can be reused and repurposed to other specific requirements. It is brought by a join team from the universities of Southampton, Manchester and Oxford which can be accessed at http://www.myexperiment.org/. Since its declaration in 2007 to 2010, myExperiment had over 3,500 registered users and contained more than 1,000 workflows. The most attractive workflow of literature mining for finding proteins associated ADR in PubMed was BioAID_ProteinDiscovery.

BioAID_ProteinDiscovery was created by Dr. Marco Roos and his colleagues from University of Amsterdam, Netherlands. The original workflow was illustrated in Figure 3.3. It was available at http://www.myexperiment.org/workflows/74.html and version 3 was employed in the task. BioAID_ProteinDiscovery retrieved documents from Medline based on a user query. Then, protein names were extracted by protein entity recognition and were filtered by checking if there existed a valid UniProt ID for the given protein name. For testing this research, BioAID_ProteinDiscovery had been modified to be the workflow in Figure 3.4. It had been changed the search and abstract retrieval part to be from PubMed and output part as shown in Figure 3.3 and Figure 3.4 with green and purple circles, respectively. The workflow was composed of various colors of atomic task. Light blue color and dark blue color represented input/output and constant, respectively. While purple atomic tasks were local services, the greens stood for web service description language (WSDL) that needed the internet connection to do the assignments. Brown fragments symbolized Beanchell, a Java-like scripting language.



🟩 The search and abstract retrieval part that had been changed

🟪 The output part that had been changed

**Figure 3.3** The original BioAID_ProteinDiscovery workflow

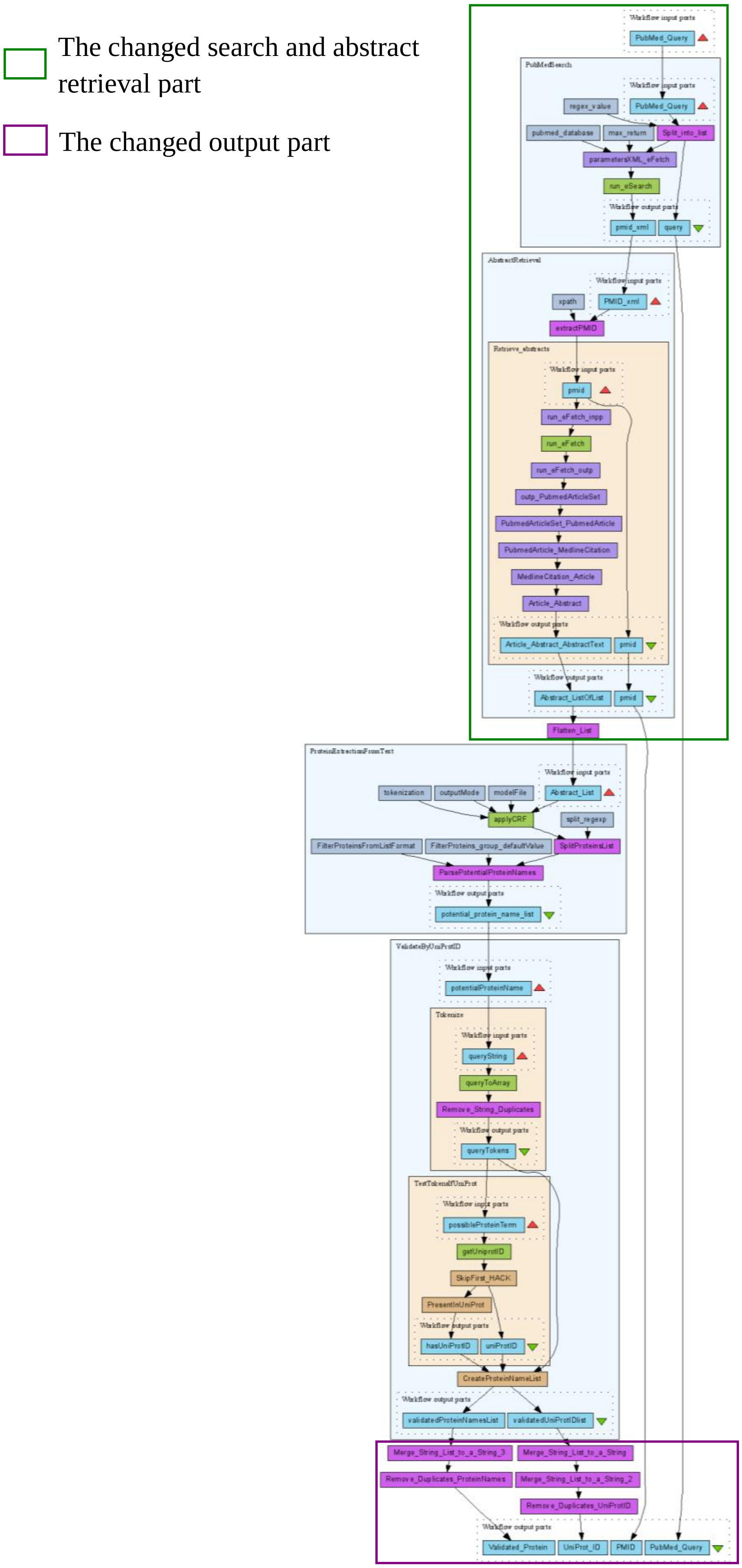


**Figure 3.4** The Modified BioAID_ProteinDiscovery workflow

The results from modified BioAID_ProteinDiscovery were saved in text file format that were collected in different directories of output types. Then the saved files were manipulated with program written by python language, shown in Appendix A, to transform into table format.

### 3.2.3 MySQL Database

MySQL (Hinz, et al., 2010) is the most popular Open Source SQL relational database management system that runs as a server providing multi-user access to a number of databases. Free-software projects that require a full-featured database management system often use MySQL. This program is also used in many high-profiles, large-scale WWW products. MySQL code uses C and C++. MySQL works on many different system platforms, including Linux and Microsoft Windows. The MySQL server and official libraries are mostly implemented in ANSI C/ANSI C++.

For MySQL Database installation, AppServ package was employed which. It was an easier way to install Apache, PHP, MySQL, and phpMyAdmin than installed each component individually. The employment version of AppServe was 2.5.10 for Windows that comprised of Apache Web Server version 2.2.8, PHP Script Language version 5.2.6, MySQL Database version 5.0.51b, and phpMyAdmin Database Manager version 2.10.3. The procedure for installing AppServ and applying in this research was explained in Appendix B.

## 3.3 Methodology

This research study proposed to construct the database that contained the relationships between ADR, drug, and protein. The workflow was illustrated in Figure 3.5.

### 3.3.1 Data Type and Collection

As details in the datasets section, the workflow of this research started from collection of datasets. ADR terms were derived from MedDRA. DrugBank and HGNC were exercised to gather drug and protein information, respectively. The associations between ADR and drug came from Canada Vigilance Adverse Reaction Online Database while relations of ADR and protein were mined from PubMed using Modified BioAID_ProteinDiscovery workflow in Taverna Workbench 2.1.0, described in topic 3.2.2.2. At last, drug and target protein information was received from data extraction function in DrugBank, the same source as drug data. In summary, all datasets were downloaded from the corresponding data sources, except ADR-protein associations. The purposes and file usage of each dataset were summarized in Table 3.2.

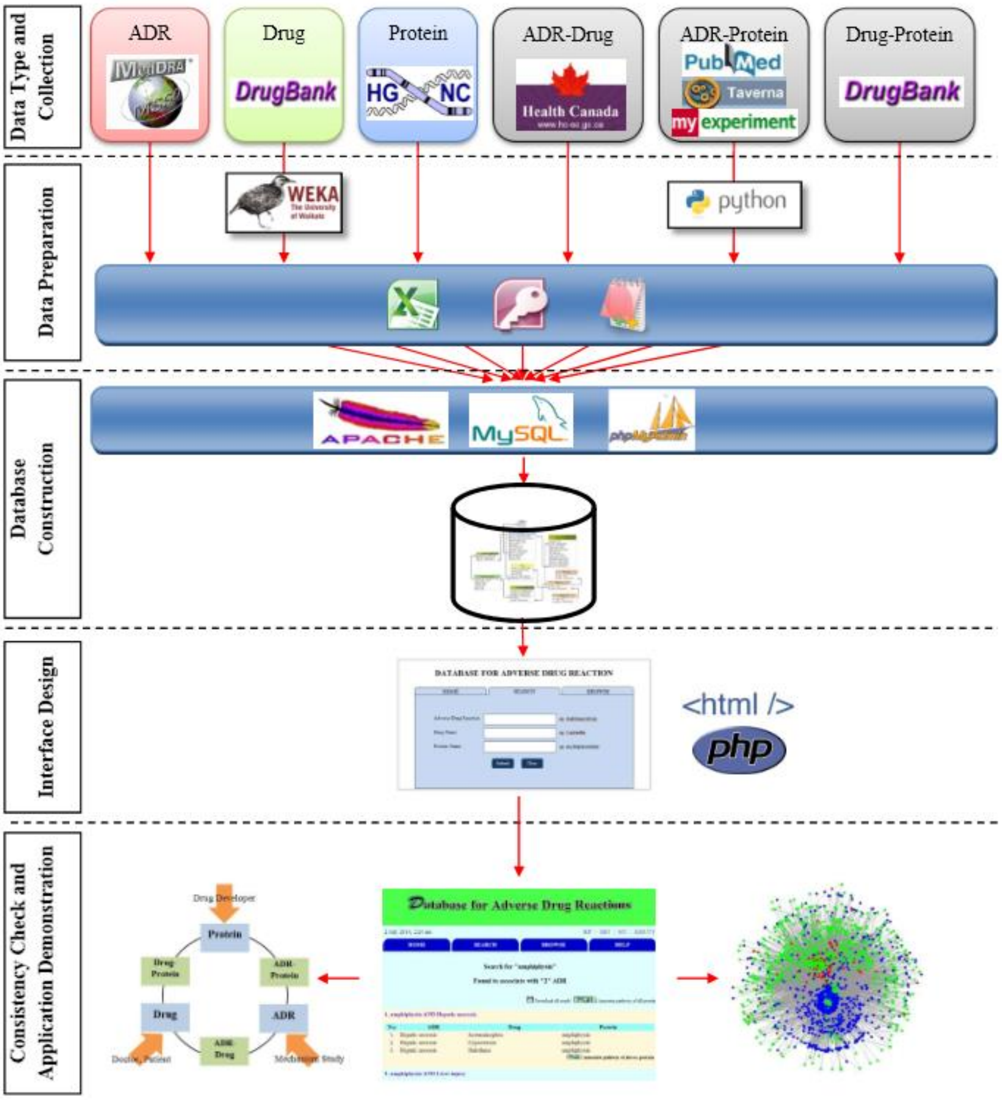


**Figure 3.5** Work process diagram representing the overall methodology

**Table 3.2** The summarized of each dataset

| **Data Source** | **File** | **Purpose** |
|---|---|---|
| MedDRA | llt | Lowest level terms of ADR |
| | pt | Preferred terms of ADR |
| | soc | System organ classes of ADR |
| DrugBank | drugbank_mapping | Information of drug |
| | drugbank_target | Information of drug and target protein |
| HGNC | hugo_nomenclature | Protein detail and conversion |
| Canada Vigilance Adverse Reaction Online Database | reaction | Information of ADR and case report |
| | report_drug | Information of drug and case report |
| PubMed/Taverna | taverna_mining | ADR and protein associations |

### 3.3.2 Data Preparation

The datasets that were collected from previous mention in topic 3.3.1 had to be prepared before applying to construct the database. The objectives of this step were removal duplicates data and editing text before transfer into database. Microsoft Excel, Microsoft Access and EditPlus were utilized for these purposes. In addition, information of drug and ADR-protein associations were experienced data mining and python programming, respectively. The details of data preparation were described below.

#### 3.3.2.1 ADR

MedDRA is designed as hierarchical pattern that promotes specific and comprehensive data entry and flexible data retrieval. This supports in creating and populating a relational database. The data preparation of ADR was pruning unnecessary fields which only remained the desired information and structure for hierarchy. The rest of the fields were presented in database construction section.

#### 3.3.2.2 Protein

Protein nomenclature was downloaded from HGNC. The applied information was HGNC ID, approved symbol, approved name, previous symbol, aliases, chromosome, enzyme ID, and Entrez ID. Human Protein Reference Database (HPRD) was also implemented to give additional fields. It provided three attributes of protein grouping by gene ontology: molecular function term, biological process term, and cellular component term. The information from HGNC and HPRD were connected by Entrez ID.

#### 3.3.2.3 Drug and Drug-Protein

DrugBank served Data Extractor, an easy-to-use search tool that allowed users to select or search over various combinations of subfields. The extracted data was saved in comma-separated values (CSV) format. Only interested fields of drug and drug-protein information from DrugBank were shown in Table 3.3.

**Table 3.3** The selected fields of drug and drug-protein information

| Information | Field Name |
|---|---|
| Drug | DrugBank ID, Generic name, Category, Wikipedia link, Drug group, Brand name, ATC code, Caco2 permeability, Dosage route, pKa, Half life, LogP (experimental), LogS (experimental), Protein binding, Route of elimination, Toxicity |
| Drug-Protein | DrugBank ID, HGNC ID, HPRD ID |

For the most efficient database operation, drugs were scored the possibility to cause ADR class by giving the point named predictive score. This score was calculated from quantitative structure-property of the drug using data mining (Ivanciuc, 2008). Weka (Witten and Frank, 2005) was applied as a data mining tool. From the determination by expert, nine quantitative structure-properties were selected for predicting drug toxicity. They were ATC code, Caco2 permeability, dosage route, pKa, half life, LogP (experimental), LogS (experimental), protein binding, and route of elimination. ATC code was diminished the length to remain only first level that indicated fourteen anatomical main classification. Drug dosage route was signified that it was enteral and/or parenteral. Drug half life was converted in minutes. Route of elimination and drug toxicity were read and encoded. After data preparation, Weka built classification models using ten folds cross-validation method which was the standard way of predicting technique. Original specifications of J48, IBk, MultilayerPerceptron, and SMO were utilized in Weka for classification of decision tree, K-nearest-neighbor, neural network, and SVM, respectively. The models were compared the correctness, precision, and recall. In addition, the predicted probabilistic scores were checked with experimental data before the best model was chosen to forecast the possible points of ADR class.

### 3.3.2.4 ADR and Drug Relationships

Two files from Canada Vigilance Adverse Reaction Online Database were applied. First was the reaction file. It was erased the unspecified ADR records and linked adverse reaction term to obtain PT code from MedDRA. Second, the report_drug file was also accompanied by removing the undefined drug product identifier and coupling to DrugBank ID. Additionally, the later file was deposed the duplicate AER number to against the incidence of one reaction from many drugs or many reactions from many drugs. At last, the revised reaction file and report_drug file were joined by AER number. These operations produced the associations of ADR and drug which one drug caused one ADR or one drug induced many ADRs.

### 3.3.2.5 ADR–Protein Associations

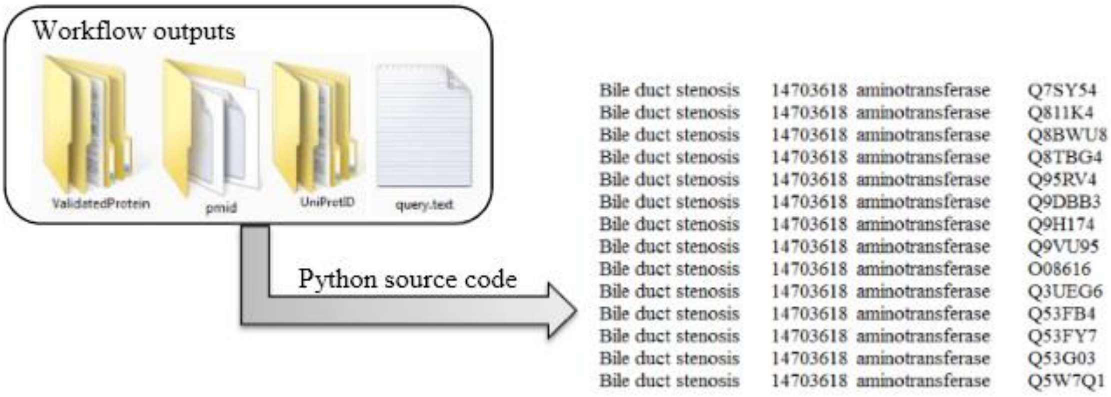
A program, written by Python language as mentioned in 3.2.2.2 (the source code and example were in Appendix A), was exploited to manipulate the outputs from literature mining using Modified BioAID_ProteinDiscovery workflow in Taverna. Illustration of manipulation was shown in Figure 3.6. Results were modified to be the table format that contained ADR term, PMID, UniProt ID and protein name. In order to indicate the relevance of extracted information between ADR and discovered protein which were observed together, pointwise mutual information (PMI) was applied. PMI can be used to measure the strength of association between ADR and protein (Church and Hanks, 1990). It compared the word bigram occurrence of ADR and protein to function of the unigram frequencies of the individual words. PMI is defined as

$$PMI \ = \ \log_2\left(\frac{P(xy)}{P(x) \cdot P(y)}\right)$$

where *P(x)* is the probability of the documents that match the ADR, *P(y)* is the probability of the documents that contain the protein, and *P(xy)* is the probability of the documents that match the ADR and contain the protein. PMI can be either positive or negative within these following bounds:

$$-\infty \leq PMI(x;y) \leq \min\,[-\log_2 P(x),\ -\log_2 P(y)].$$

PMI is zero when ADR and protein occur independently, $P(xy) = P(x)\cdot P(y)$, whereas it reaches maximum value when these two terms are perfect co-occurrence, $P(xy) = P(x)\cdot P(y)$. As PMI calculation, the frequency counts of co-occurrence between ADR and protein were done in Microsoft Excel. ADR and protein were counted for achieve articles in PubMed using Taverna. The frequency counts of co-occurrence between ADR and protein association, ADR, and protein were then calculated for *P(xy)*, *P(x)*, and *P(y)*, respectively. The processed data was also connected with PT code from MedDRA and HGNC ID for protein nomenclature. Finally, circumspect technique, the comparative of known or previous studies, was introduced to prove the consistency of associations between ADR and protein.



```
Bile duct stenosis   14703618 aminotransferase   Q7SY54
Bile duct stenosis   14703618 aminotransferase   Q811K4
Bile duct stenosis   14703618 aminotransferase   Q8BWU8
Bile duct stenosis   14703618 aminotransferase   Q8TBG4
Bile duct stenosis   14703618 aminotransferase   Q95RV4
Bile duct stenosis   14703618 aminotransferase   Q9DBB3
Bile duct stenosis   14703618 aminotransferase   Q9H174
Bile duct stenosis   14703618 aminotransferase   Q9VU95
Bile duct stenosis   14703618 aminotransferase   O08616
Bile duct stenosis   14703618 aminotransferase   Q3UEG6
Bile duct stenosis   14703618 aminotransferase   Q53FB4
Bile duct stenosis   14703618 aminotransferase   Q53FY7
Bile duct stenosis   14703618 aminotransferase   Q53G03
Bile duct stenosis   14703618 aminotransferase   Q5W7Q1
```

**Figure 3.6** Illustration of manipulating outputs from Modified BioAID_ProteinDiscovery workflow in Taverna by program written in python language (example was clarified in Appendix A)

### 3.3.2.6 ADR-Drug-Protein Associations

In order to assemble the relationships between ADR, drug, and protein, all pairs of association data were joined together on a share component. The relations of ADR-drug from Canada Vigilance Adverse Reaction Online Database were connected to drug-target protein information collected from DrugBank on drug element. The data linkages were done only on code or ID for reducing data redundancies. These concepts of information connection were also applied for linking ADR-drug to ADR-protein and ADR-protein to drug-protein. Then the three sets of ADR-drug-protein associations were merged into one and were removed duplicate associations. Because of the accumulation of ADR-drug-protein associations into a table, this took out overlapping associations from the procedure of connection and provided the unambiguous accession for the database.

### 3.3.3 Database Construction

Database construction would be conducted for relational database of adverse drug reaction. There were nine tables in the database for ADRs. The type and explanation of attributes in each table were clarified in data dictionary, Appendix C. The entity relationship (ER) diagram that presented how tables were connected to another was illustrated in Figure 3.7. Three tables which were meddra_llt, meddra_pt, and meddra_soc were donated for ADRs information. While table named protein contained protein information, drugbank_mapping table had information of drug. Three tables were association datasets of ADR-drug, ADR-protein, and drug-protein which were canada_adr_drug, taverna_adr_protein, and drugbank_target, respectively. Lastly, center table had ADR-drug-protein relationships from joining association data process.

### 3.3.4 Interface Design

To value the constructed database, user interface design was introduced to interact with users. All of the interface design was written in HTML for displaying the appearance on web browser and PHP for interacting with the database. The user-friendly interfaces were created for flexible information investigation, simple result pages, and user guide. Two pages were designed for interactive information examination. There were search page and browse page that surveyed the name of keyword and the type of searching term, respectively. The browse page explored SOC for ADR, first level of ATC classification for drug, and enzyme reaction for protein. Two result pages were provided for user. The first page primarily showed the searching word and major interesting entity that could be expanded to observe the complete associations of ADR-drug-protein. The second page was displayed the details of each element in ADR-drug-protein association.

### 3.3.5 Consistency Check

Database for ADRs was constructed from integrating three association data. Even though it was considered to have the power to generate an effective knowledge, different types of association data still had various degrees of reliability. Drug and target protein relations from DrugBank had excellent correctness because they was acquired through primary literature sources, checked by experts, edited and entered manually (Wishart, et al., 2008). ADR-drug associations came from reported information which was filtered only suspected adverse reaction or side effects to a drug. While ADR-protein that was connected by literature mining tool was set against the previous studies, the finest technique to check the consistency of the database was also comparison with known research

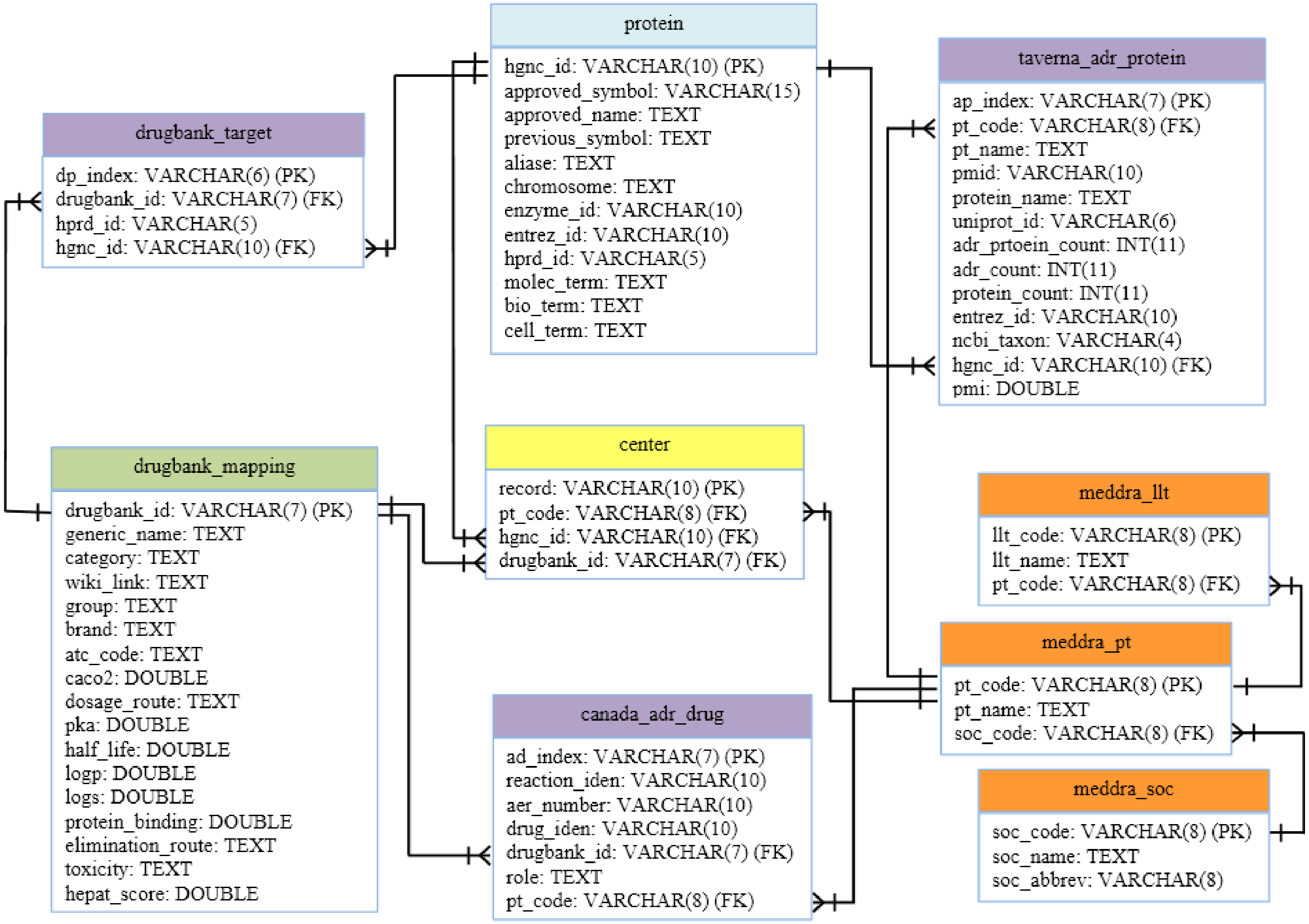


**Figure 3.7** ER diagram of constructed database (type and explanation of attributes were in Appendix C)

### 3.3.6 Application Demonstration

The constructed database for ADRs comprised of ADR terms, proteins, and drugs which entities were associated to others. From user interface design, different users had divergent purposes to use the database for ADRs. These variations were expressed through their inquiries which were demonstrated in Figure 3.8. Doctors and patients required to notice about drug could possibly cause any ADR. While drug developers made their questions that interesting protein could be a target of what drug or associated to what ADR, they also desired to recognize that their medication had even more target protein. Another advantage of the database for ADRs was for mechanism study. This was an impressive benefit because most ADRs did not know mechanism. Pathway annotation was therefore a brilliant approach to identify or reveal the underlying mechanism of ADRs. Related proteins of ADR that were the result from the database were mapped into Biocarta and KEGG pathway using functional annotation of Database for Annotation, Visualization and Integrated Discovery (DAVID) 6.7. DAVID (Huang, et al., 2007; Huang, et al., 2009) had the power to analyze a large gene list in a single space which was applied as gene-annotation knowledgebase in this study. It could be accessed freely at http://david.abcc.ncifcrf.gov/home.jsp. The results were functional annotation pathway of the input proteins which starred the mapped proteins in a graphical format and provided the link to Biocarta and KEGG website for detail.

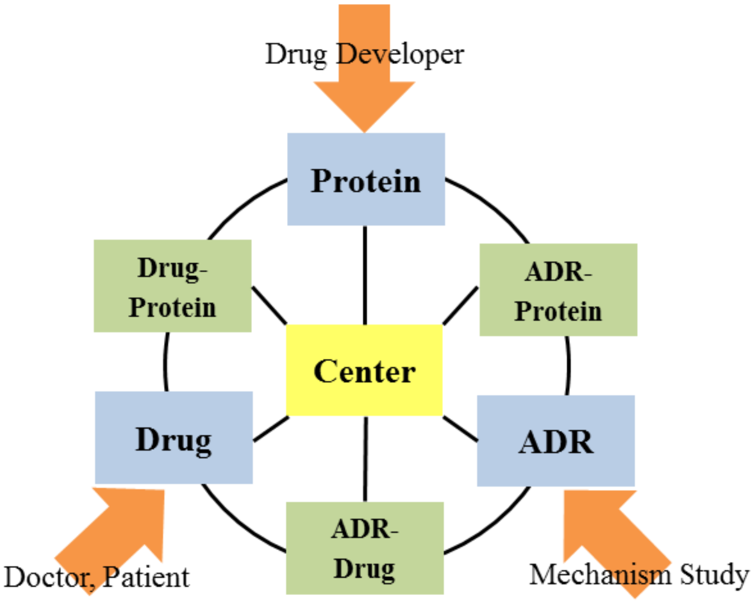


**Figure 3.8** Different users of the database for ADRs