DESIGN, DEVELOPMENT AND IMPLEMENTATION OF TOOLS IN DRUG DISCOVERY

Cheemakurthi, Usha Deepika

Submitted to the faculty of the School of Informatics
in partial fulfillment of the requirements
for the degree of
Master of Science in Chemical Informatics,
Indiana University

DECEMBER, 2007

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

Master's Thesis Committee

Dr. Kelsey Forsythe, Ph D, Chair

Dr. Malika Mahoui, Ph D

Dr. David Wild, Ph D

Dedicated to my husband

TABLE OF CONTENTS

APPENDIX

ACKNOWLEDGEMENTS

My sincerest thanks to Dr. Kelsey Forsythe for his guidance and encouragement throughout my research. I am grateful to the members of my thesis committee, Dr.Malika Mahoui and Dr. David Wild, for their support, valuable suggestions and continued guidance throughout the project. I would also like to acknowledge the faculty and staff of the Indiana University School of Informatics at IUPUI, Indianapolis and Dr. William Scott from the chemistry department for their support.

I would like to extend my thanks to Deepthi Jonnala, Amey Godse, Nithin Cheemalavagupalli and Ashwin Selka Padmanabhan for their input and help.

Very special thanks to my husband, Sri ParthaSarathi, who has always stood by my side. I sincerely thank him for his constant support and encouragement in achieving this important milestone in my life. I thank my parents for their guidance and assistance in all endeavors throughout my whole life.

The errors and inconsistencies remain my own.

ABSTRACT

Usha D Cheemakurthi

DESIGN, DEVELOPMENT AND IMPLEMENTATION OF TOOLS IN DRUG DISCOVERY

The main focus of our work is to develop, apply and assess cheminformatics tools and methods. In particular, we focus on the following three areas: Integration of open source tools with application to drug discovery, usability studies to assess the efficacy of these software tools and finally, developing novel techniques for database query.

Rapid globalization in the present time has sparked a need in the scientific community to interact with each other at an economic and a fast pace. This is achieved by developing and sharing open source databases using World Wide Web. A web based open source database application has been developed to incorporate freeware from varied sources. The deployment of developed database and user interface in a university lab setting is discussed.

To aid in connecting the end user and the software tools, usability studies are necessary. These studies communicate the end users' needs and desires, resulting in a user-friendly and more powerful interactive software packages. Usability studies were conducted on developed database student application and on different drawing packages to determine their effectiveness.

Developing new and interactive search engines to query publicly available databases helps researchers work more efficiently. The huge volume of data available and its heterogeneous nature presents issues related to querying, integration and presentation.

In aiding the retrieval process, an innovative multi faceted classification system, called ChemFacets, is developed. This system provides dynamic categorization of large result sets retrieved from multiple databases.

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Background

Chemical informatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization and use of chemical information[1]. This mostly applies to handling of information related to molecules and their reactions. It has wide applications in the pharmaceutical field, drug and chemical industries, from managing the intellectual property to enabling theoretical and experimental studies of chemical entities. Most of the chemical informatics research operations include creating or managing lists of databases, searching/browsing the databases, drawing/manipulating structures.

1.1.1 Distributed Drug Discovery

The need for innovative and inexpensive drugs to treat diseases in the developing world is self evident. In the developed world, drug discovery has been fueled by the pharmaceutical industry, where economic incentives have financed the expensive equipment and procedures currently required. Unfortunately, because the world burden of disease is disproportionately focused in poor nations, there is no economic incentive for the pharmaceutical industry to discover drugs for the diseases of the developing world[2]. In order to reduce the cost of drug discovery, the concept of Distributed Drug Discovery (DDD) has been proposed.

The fundamental principle of Distributed Drug Discovery is to breakdown large research problems into manageable smaller units and carry out necessary research at multiple academic sites simultaneously using simple, inexpensive equipment and procedures developed for the core drug discovery disciplines: computational chemistry, synthetic chemistry and biochemical screening. The coordinated and combined results of these “distributed” resources inexpensively accelerate the identification of leads in the early stages of the drug discovery process. Figure 1 illustrates the schema for solving large problems through a ‘distributed’ approach.

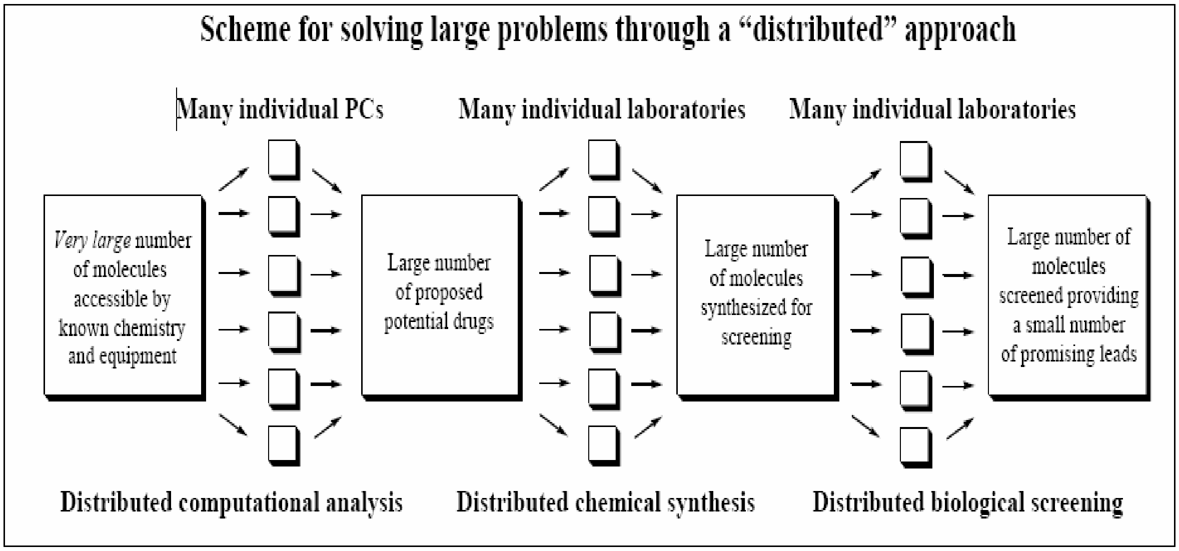

Scheme for solving large problems through a “distributed” approach
Many individual PCs
Many individual laboratories
Many individual laboratories
Very large number of molecules accessible by known chemistry and equipment
Large number of proposed potential drugs
Large number of molecules synthesized for screening
Large number of molecules screened providing a small number of promising leads
Distributed computational analysis
Distributed chemical synthesis
Distributed biological screening

Figure 1: Scheme for solving large problems through a distributed approach[2]

Significant progress towards the goal of distributed discovery has been made in the computational chemistry area, where inexpensive personal computers and free software can expedite the discovery of drugs on a global scale. For example Grid.org describes how a “screen saver surrogate” process permits the computational screening of thirty-five million potential drug molecules for their ability to interact with several protein targets that could lead to drugs to treat small pox.

The distributed chemical synthesis operation consists of dividing potential drug lead candidates, for example 5000, into batches of compounds for synthesis in 100 different academic laboratories, fifty compounds per laboratory throughout the world.

The chemistry labs at Indiana University-Purdue University Indianapolis have developed novel chemistry and equipment to meet the requirements of ease-of-use and low cost synthetic components. This preliminary work is being done by Dr. William Scott and Dr. Martin O'Donnell in collaboration with University of Barcelona, Moscow and School of Pharmacy in Lublin, Poland to demonstrate the feasibility of a globally distributed project. The pooled molecular products of these local efforts could be tested by globally distributing screening effort. The summation of these efforts, networked across the three core disciplines, becomes a powerful globally distributed drug lead discovery process and solution.

In this process, a requirement for the success of the distributed approach is the availability of appropriate information technology to coordinate the dissemination, acquisition and analysis of the data from all the three core disciplines.

1.1.2 Chemical Facets

Chemists rely on both online and paper-based chemical databases to conduct their research. Online computer database systems require computer software to search, retrieve and/or analyze the information in the particular database. Some databases are task oriented, whereas some require authentication. Nevertheless, each database has its specific features and potential for supporting existing research investigations. There is a need to integrate these databases to enhance the search process. Chemical facets play an important role in such an integration environment. A facet is a method of classification which groups together the results which have the same value for a particular category.

For example, wines from French, Italian and Portuguese can be grouped together into, what is called, regional wines.

1.1.3 Usability Study

Usability studies are a principal means of determining whether a system meets its intended purpose. A system is a collection of interacting and interdependent parts that are organized to meet some purpose[3]. The need for usability testing especially arises when a system is designed for use by non-developers. The system could be a service, a collection of computer programs, a diagram, piece of text, a methodology, a set of instructions, or a dataset. Usability testing measures the extent to which the intended user can meet the set goals using the system being tested i.e., it measures the usefulness of the system. The subject in the usability study is a potential user of the system who participates in a usability test. The underlying model for virtually all usability tests is that real users carry out real work with a system (product).

If usability testing uncovers difficulties, such as people having difficulty understanding instructions, manipulating parts, or interpreting feedback, then developers can improve the design and test the system again. During usability testing the aim is to observe people using the product in as realistic a situation as possible, to discover errors and areas of improvement. Usability testing involves watching people trying to use something for its intended purpose. Several other test instruments such as scripted instructions, paper prototypes, and pre- and post-test questionnaires are also used to gather feedback on the product being tested.

1.2 Importance of Subject

1.2.1 Chemical Information Management

Chemical information management is at the core of drug discovery. Pharmacological experiments and optimization workflow must track chemical information. Though there has been considerable privatized effort developing information technology for chemistry, it has not been disseminated for the wider scientific community. The tools that have been developed in cheminformatics are often too expensive for academic research groups and small biotech companies to use. Additionally, some of the tools commercially available may not meet the required standards or satisfy all the needs of a particular organization. As a result, companies and research groups are left developing their own proprietary software package.

The packages thus developed must be self sufficient, customizable and expandable. The structure of the package and the tables should be self explanatory. Open standards need to be used in storing and manipulating the data. Accessibility to the database needs to be at a global level. Flexibility in querying the data by varying the properties including the structure of the compound is to be incorporated in the package. Though substructure search, etc. could pose challenging problems, the package must have the ability to support standard file formats and standard view formats. For reaction libraries, the user should be able to model several reactions to obtain required combinatorial library of products.

The packages should also provide the developers with the flexibility of adding calculators, plug-ins, etc. on demand. The user should be able to retrieve the output in the requested format and different search engines could be incorporated.

Scientists have developed different database applications using open source databases such as such as MySQL (My Structured Query Language), PostgreSQL (Postgres Structured Query Language) and integrating some of the freely available chemical drawing packages including Open Eye scientific software, Open Babel, Marvin, Java Molecule Editor[5, 6], etc.

1.2.2 Chemical Information Retrieval

There are mainly three types of chemical information: bibliographic data, non-bibliographic data and structural data. The first type[7] includes chemical abstracts; such databases contain textual information, citations, abstracts, but not factual data. The second type[7] includes numeric data or factual data such as infrared data, boiling points, spectral data, etc. The third type[7] has the chemical structure data in the computer readable form. With such huge data and diverse databases, the user is required to go through several databases to search and retrieve information. For example, a chemistry lab technician is in charge of purchasing a chemical for their lab experiment. In order to gather information about the chemical, he/she will have to go a database to find its physical and chemical properties, another database to find its drug facts when used in the experiment, another database to find the suppliers of the chemical and another to order it. In such research process the chances of skipping important information or a data source is likely furthermore this research process is time consuming. This results in loss of information. The need arises for such a tool which can query diverse data sources; integrate multiple data sources and present voluminous and diverse results just in one step process. For example, at Amazon[8] online shopping, the user is initially given a choice of browsing categories. Once a category is finalized, sub categories choice like the manufacturer choice, price choice is given to the user and this process continues until the user is satisfied with the product he wants to purchase. This hierarchy follows until the user gets the desired product. This tells us that a particular category can be viewed in many different ways. Such classification is called faceted classification. Similar kind of classification and hierarchy is a posing challenge in the scientific environment.

1.2.3 Usability Analysis

Usability testing helps the developer and the user in understanding and developing a product. The important concept in usability testing is that the users are asked to do something realistic with a product, and to do enough of it to approximate the experience they would have with the real product in the real world[9]. In usability studies one should define the questions the study should answer. Defining different kinds of users and the tasks to be assigned comes next. Agreeing on a set of usability measures with good validity for the users, their tasks, and the environment where the study is conducted is necessary. Then a detailed study design should be developed which includes gathering information about the product, along with providing a way to deliver data-driven answers to specific questions about the product.

2. BACKGROUND AND LITERATURE SURVEY

2.1 Work in Polymer Laboratories

Polymer Laboratories has launched a new resin for the solid-phase synthesis of resin-bound unnatural amino acids, one of the most commonly used intermediates in combinatorial chemistry, peptides as well as peptidomimetics. One of the principles behind ‘Distributed Drug Discovery’ is that many potential drug molecules can be produced by simple techniques, using inexpensive equipment, readily available starting materials and robust synthetic procedures. [10]

2.2 Schools Malaria Project

The *Schools Malaria Project* brings together students with university researchers in the hunt for a new anti-malaria drug. The Escience group at the University of Southampton School of Chemistry developed this project. The design challenge being offered to students is to use a distributed drug search and selection system to design potential anti-malaria drugs. The system will be accessed via a Web interface. This e-science project will display the results of the trials in an accessible manner, giving students an opportunity for discussion and debate both with peers and with the university contacts. The project has been implemented by using distributed computing techniques, spread over a network of machines that cross institutional boundaries forming a grid. This will provide access to greater computing power and allows a much more complex and detailed formulation of the drug design problem to be tackled for research, teaching, and learning[11].

2.3 Grid.org

The Web site Grid.org[12] describes how a "screen saver surrogate" process permits the computational screening of thirty five million potential drug molecules for their ability to interact with several protein targets that could lead to drugs for treating smallpox. According to the project, though small pox was eliminated from the world in 1977 by a World Health Organization campaign, stocks of the variola virus exist and its use as a weapon of bioterrorism remains a frightening possibility. Since the vaccination ended in 1972 and the world population is highly susceptible, the availability of drugs to counter the virus would be a major defense. By blocking a particular molecular target, this would prevent the ravages of an infection. Grid computing is used to screen millions of potential anti-smallpox drugs against this target. This project harnesses millions of computers belonging to people in over two hundred countries. Similarly, the concept behind Distributed drug discovery is carrying out the research in multiple academic sites globally to maximize the likelihood of lead target identification.

2.4 ChemAxon Tools

ChemAxon's Marvin tools provide enough functionality to build a compact and efficient database system and are used in several systems.

ChemAxon[13] is a leader in providing Java based chemical software development platforms for the biotechnology and pharmaceutical industries. By focusing upon active interaction, it creates cross platform solutions to power modern cheminformatics and chemical communication. ChemAxon supplies products which have a broad range of functionality. It provides various applications like file conversion, internet/intranet structure editing, viewing, structural prediction and search, Java toolkit platform development, library enumeration, virtual screening, drug design and much more. The applications provided by ChemAxon are platform independent, database independent and browser independent.

They are built on a strong scientific base with the ease of implementation and partnered with an open licensing position. It supports sophisticated virtual chemistry technology with high performance and capacity. It allows broad range of format support with customizable chemistry knowledge bases. Multiple deployment formats are available and include java applets, java beans, java web start, plug-ins and java server pages the end user can implement in his own applications. Implementation support is available via email or support forum with a fast response – maximum 24 hours. Detailed documents are available online and extensive help is bundled within the software. There are many freely available implementations from the site itself. They are freely available for academic licensing and have flexible pricing models like annual licensing, fixed fees for commercial firms. Figure 2 explains the functionality of products provided by ChemAxon. Briefly, Marvin supports the functionality of drawing and viewing the structures and JChem supports substructure search, superstructure search, exact search, perfect search, similarity search with different interfaces.

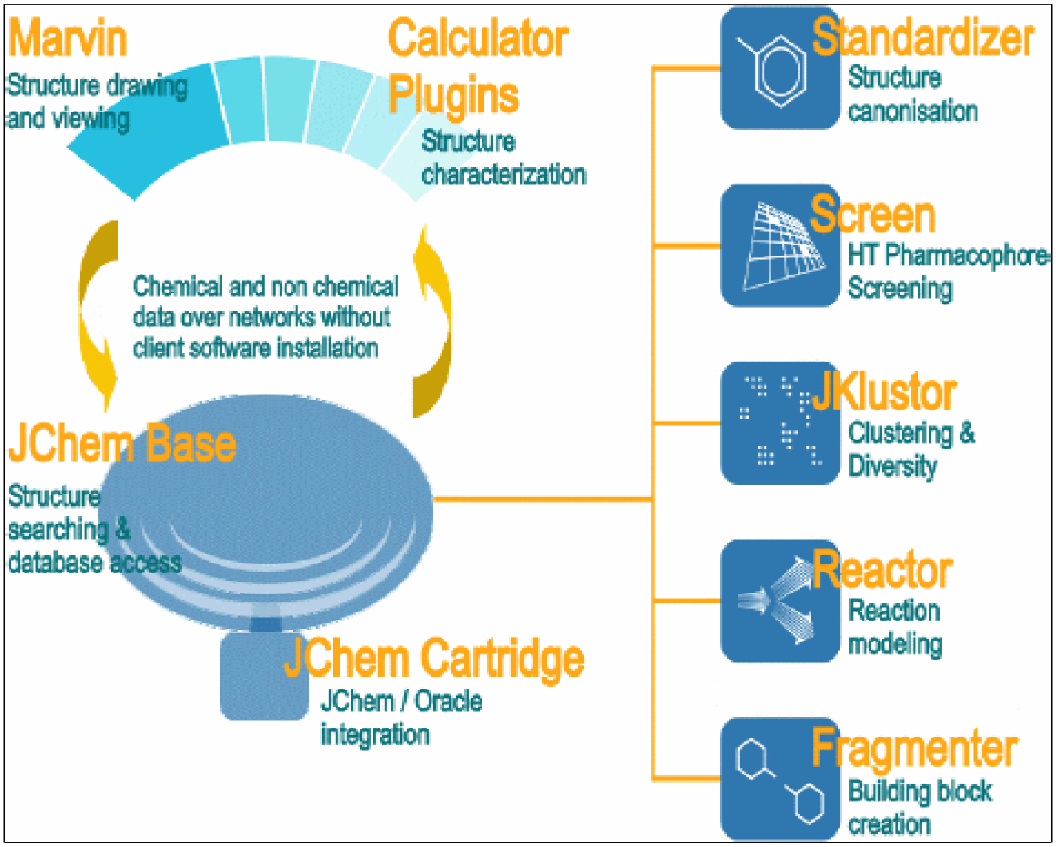

Marvin
Structure drawing and viewing
Calculator Plugins
Structure characterization
Chemical and non chemical data over networks without client software installation
JChem Base
Structure searching & database access
JChem Cartridge
JChem / Oracle integration
Standardizer
Structure canonisation
Screen
HT Pharmacophore-Screening
JKlustor
Clustering & Diversity
Reactor
Reaction modeling
Fragmenter
Building block creation

Figure 2: Broad product range of ChemAxon technology

2.4.1 Real World Chemistry

According to Joe Mernagh[14] in the real world, laboratory based drug efforts, on their own, are insufficient to deal with the wealth of potentially drug-like targets. Drug lead discovery, both rational and high throughput, is playing a greater role in the discovery process; however the demands of setting up the technology are considerable, even for a major pharmaceutical company. Only virtual techniques, computationally filtering, highly profiled candidate molecules for laboratory testing, can address the bottleneck. According to Mernagh, virtual discovery is cheap with immediate, rapid experiments covering all accessible chemistry space. There is no shelf-life or repeatability issues nor physical infrastructure involved in the virtual discovery This concept aims to provide a sophisticated in-silico drug discovery framework available online. This will allow the user to create virtual library from selected reactions and input data followed by the series of desired experiments like cluster analysis, screening, etc. to identify the drug-like candidates. ChemAxon's discovery tools are used to create the virtual library.

2.4.2. Neogenesis

At Neogenesis, a complete system for storing and retrieving chemical structures accurately is being developed. ChemAxon's Marvin is used within Oracle to ensure uniqueness and to calculate molecular properties of the components. The Graphical User Interface (GUI) developed to interact with the system relied on Marvin components[15].

2.4.3. Genomics Institute of the Novartis Research Foundation (GNF)

The informatics team at Genomics Institute of the Novartis Research Foundation (GNF) has been developing a comprehensive lead discovery database (LDDB) in order to support an aggressive drug discovery portfolio. The chemical structure searching capability of LDDB has been supported by Marvin and Daylight cartridge[16].

2.5 BioFacets

BioFacets is an integration system for Web-based biological databases. The distinctive feature in BioFacets is its innovative multi-faceted classification system which provides dynamic categorization of the large result sets retrieved from multiple databases.

The results, thus obtained, are viewable from different orthogonal views or facets. This provides researchers with a search/browse interface wherein they begin with a search, and use various facets recursively to browse through and narrow the result set. The system also features a meta-search engine that facilitates efficient search across multiple online biological databases, which are integrated using the wrapper-mediator approach. Query optimization and cache management provide improved system performance.

3. PROPOSAL

Researchers recognized a need for a refined database with advanced search tools which enable the collection of large amounts of data from students from different locations and educational institutions throughout the world. This problem translated into the development of a Distributed Drug Discovery database. It enables research scholars to interactively search the database for past results and input current research results within minutes.

3.1 Proposed Solution

To address the above problem the proposed solution was divided into three parts:

1. The development of a Distributed Drug Discovery (DDD) Database and web-based user interface to enable data entry/searching of the database and a workflow environment to support pipelining of tasks involving the Distributed Drug Discovery database. This web application includes a structure applet drawing for entry and querying the database.

2. Designing and implementing usability studies, both in the context of the Distributed Drug Discovery project and general case of chemical drawing packages.

3. Adapting the faceted classification[17] to integrate open source chemical databases.

3.2 Materials

PostgreSQL[18] (Postgres Structured Query language) is an open source database system available freely to everyone. When compared to MySQL[19] PostgreSQL supports foreign key, views, stored procedures, triggers, full joins, constraints, cursors and complex queries[20, 21]. PostgreSQL is used to host the database. Java application is used to build the front end user interface using Tomcat server and JDBC (Java Database Connectivity) driver. ChemAxon's Marvin tools are used to build a structure applet and reaction program.

A web survey is used to analyze the usability of the web application designed for Distributed Drug Discovery database. A series of tasks for the usability analysis of chemical drawing packages are deployed.

The facet based integration system uses XML (Extensible Markup Language) technology to facilitate the querying and integration of open source chemical databases like NCBI PubChem[22], DrugBank[23], and Chemical Entities of Biological Interest (ChEBI) [24]. XML is a simple, very flexible text format. It is a markup language for documents containing structured information. Structured information contains both content (i.e., data) and some indication of what role that content plays (i.e., metadata). It plays an important role in the exchange of variety of data on the Web and elsewhere.

4. METHODS

4.1 Database Design

Enterprise statement:

In order to collect the application requirements with respect to the data and the tasks it needs to support, a series of interviews were conducted with the leader of the DDD project, Dr. William Scott. From these interviews, a list of specific main objects of data containers that need to be provided were identified. For each of the main object, a list of descriptive attributes was determined.

The actual experimental procedure starts with collection of reagents (alkalyting reagents and acylating reagents) from Aldrich catalog. By using enumeration software developed by ChemAxon Reactor, a theoretical library of products can be created called virtual library. Then, by testing a select group of reagents in the lab, a theoretical rehearsed library of products is created which is a subset of the large virtual library but is guided by the select group tested. The students perform lab experiments on the reagents to obtain a synthesized library which is a subset of rehearsed library. These experiments are performed by students at different organizations.

To illustrate the role of rehearsed library and its relationship with the virtual and synthesized library, consider the following example.

Example: The reagents that are involved in the reaction are R1 =200 and R2 = 50.

A Virtual library can be generated by all the possible combination of the reagents i.e., 200 * 50 = 10,000 compounds.

For a rehearsed library, in step one a representative from R1 is chosen and reacted with each of the R2 reagents to conduct 50 experiments.

In step 2, a representative from R2 is chosen and reacted with each of the R1 reagents to conduct 200 experiments. Consider the situation of 40 successful experiments from part one and 80 successes in part two.

The resulting rehearsed library is made up of 80 * 40 = 3,200structures. Based on these results, students perform the 3200 experiments in the laboratories which leads to a synthesized library.

The information that needs to be collected about the reagents is their structure, molecular weight, formula, CAS number, SMILES, PubChem number, source from which it is extracted and its role in the reaction. The different reactions performed by the enumeration software are stored in the reaction table. For the products; the structure, formula and molecular weight information is collected. From the experimental work the information about the products purity, the lot number of the student or the team and the vial number are recorded.

Based on the above statements, the following tables were created: reagents table, reaction table, virtual library table, rehearsed library table, synthesized library table, source table and organization table. Figure 3 represents the entity-relationship diagram.

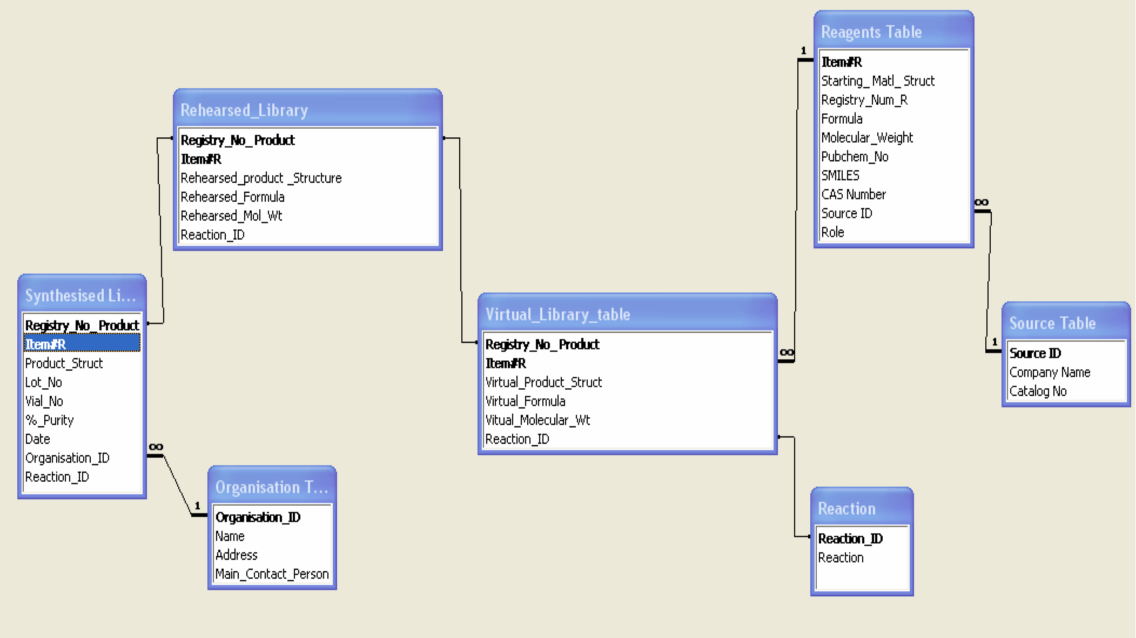


Figure 3: Entity Relationship diagram

The reagents table stores information about the starting material information. Table 1 represents the reagent table. Item No R which is an internal identifier is the primary key for the table. This table stores the information about the starting material structure, registry number of the reagents, their formula and molecular weight, the PubChem number, SMILES notation, CAS registry number, the role of the reagent (alkylating or acylating).

Table 1: Reagents Table

Item No R	Starting Material Structure	Registry # R	Formula	Mol wt	PubChem No	SMILES	CAS No	Source ID	Role

The source table stores information about the source id, company name and the catalog number where the source id is the primary key. Table 2 represents the source table.

Table 2: Source Table

Source ID	Company Name	Catalog Number

The reaction table stores the information of the reactions performed using the enumeration software Marvin reactor. At present it contains a reaction_id and the reaction in SMARTS notation. Table 3 represents the reaction table.

Table 3: Reaction Table

Item No. R	Reaction_ID	Reaction

The virtual library table stores the information about the item number of reagents which is a foreign key linking to the reagents table, the registry number of the products which is a primary key, the product structure, its formula and molecular weight. This is the combinatorial library of structures which is generated by enumeration software packages such as ChemAxon's Marvin reactor[25].

Table 4: Virtual Library Table

Item No.R	Registry # P	Virtual Product Structure	Virtual Formula	Virtual Mol.Wt	Reaction_ID

The rehearsed library table consists of the information about the item number of reagents, registry number of the product, rehearsed product structure, its formula and molecular weight. Table 5 represents the rehearsed library.

Table 5: Rehearsed Library Table

Item No R	Registry # P	Rehearsed Product Structure	Rehearsed Formula	Rehearsed Mol.Wt	Reaction_ID

The synthesized library table consists of the experimental information recorded by the students in teams. Here, the Lot number and vial number are the main primary keys. Lot number represents the team of the students and vial number the product stored. It also stores information about the registry number of the product, item number of the reagent, product structure, date of the experiment and the organization id foreign key to the organization table where the experiment has been conducted. Table 6 represents the synthesized library.

Table 6: Synthesized Library Table

Registry# P	Item No. R	Product Structure	Lot #	Vial #	%Purity	Date	Organization ID	Reaction_ ID

The organization table consists of the information of school or organization where the experiment is performed. Table 7 represents the organization table.

Table 7: Organization Table

Organization ID	Name	Address	Main_Contact_Person

4.2 Web Application

To facilitate the retrieval and populating the designed database, an interactive web based communication tool is proposed. This tool is to use JSP developing environment which is supported by any PC working windows platform.

The proposed web application is intended to be used by administrators, professors, researchers and students who participate in the DDD project. The main aim of web application is to enable the user to input, update and search information in the database. In order to avoid manipulation of the data in the database, a hierarchical user authentication is been proposed. The administrator who is on the top level in the hierarchy is supposed to have all rights to access the database. He/she would have all permissions to make changes to the database both at data level and at the design level including making changes to the structure of tables. The professor who comes next in the hierarchy is supposed to have both read and write permissions to the database. The students who come third in the hierarchy have read only permissions to the database. On the other hand, a student workbook is being proposed as part of functionality requirement for the students to enter the experimental information done by them in the laboratories. The idea behind designing student workbook is to be able to enter complete accurate experimental information into the database once the administrator reviews the students work. The student workbook is intended for the students to maintain their lab experimental work electronically.

In the present project, both administrator and professor are at the same hierarchical level. The functionality of the present interface has two aspects: Administrator and Student. The administrator is given privileges of storing the information of the reagents used in the experiment, the products generated both virtually and synthetically along with the information of the source of the reagent and the organization in which the products are being made.

The flow diagram in Figure 4 represents functionality of the web application interface.

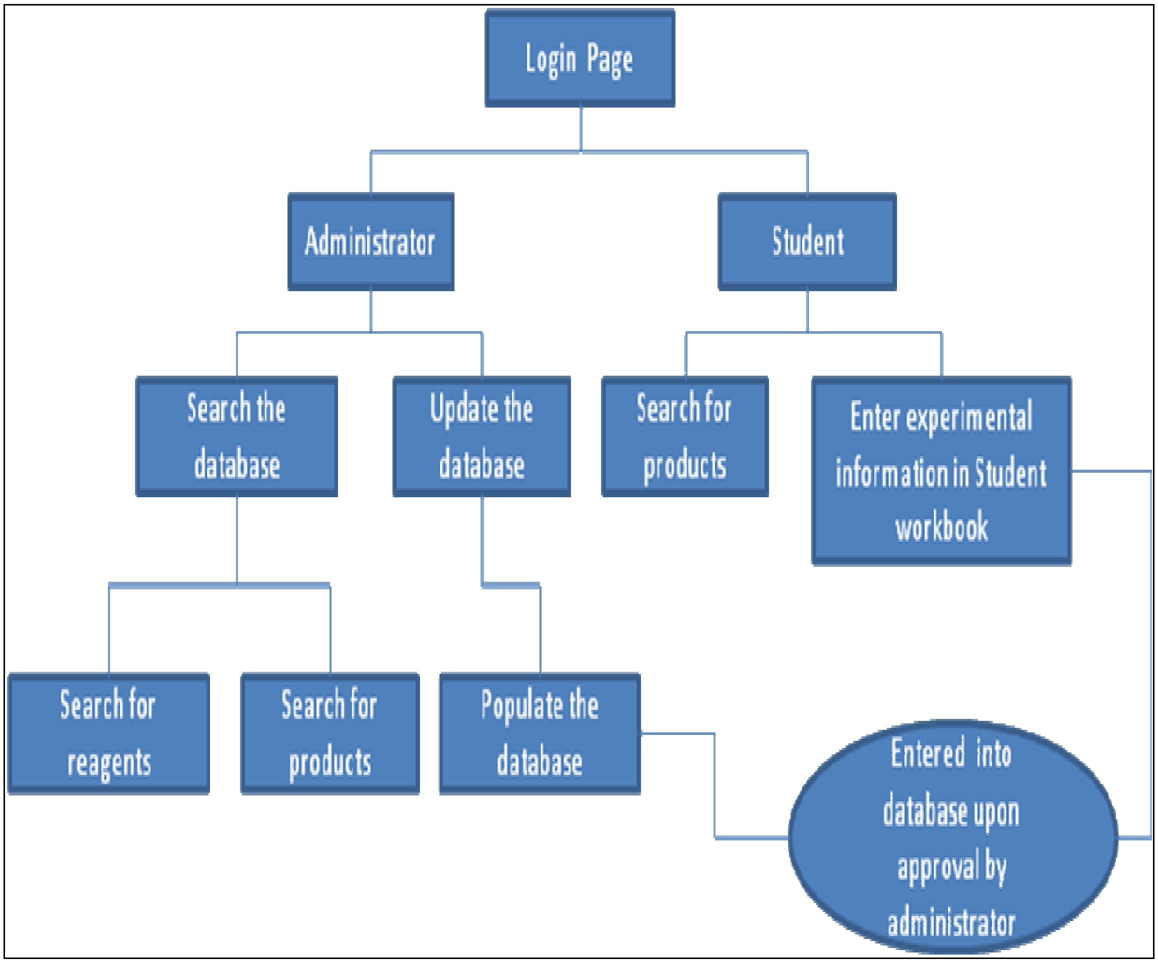


Figure 4: Flow chart of the Web Application Interface

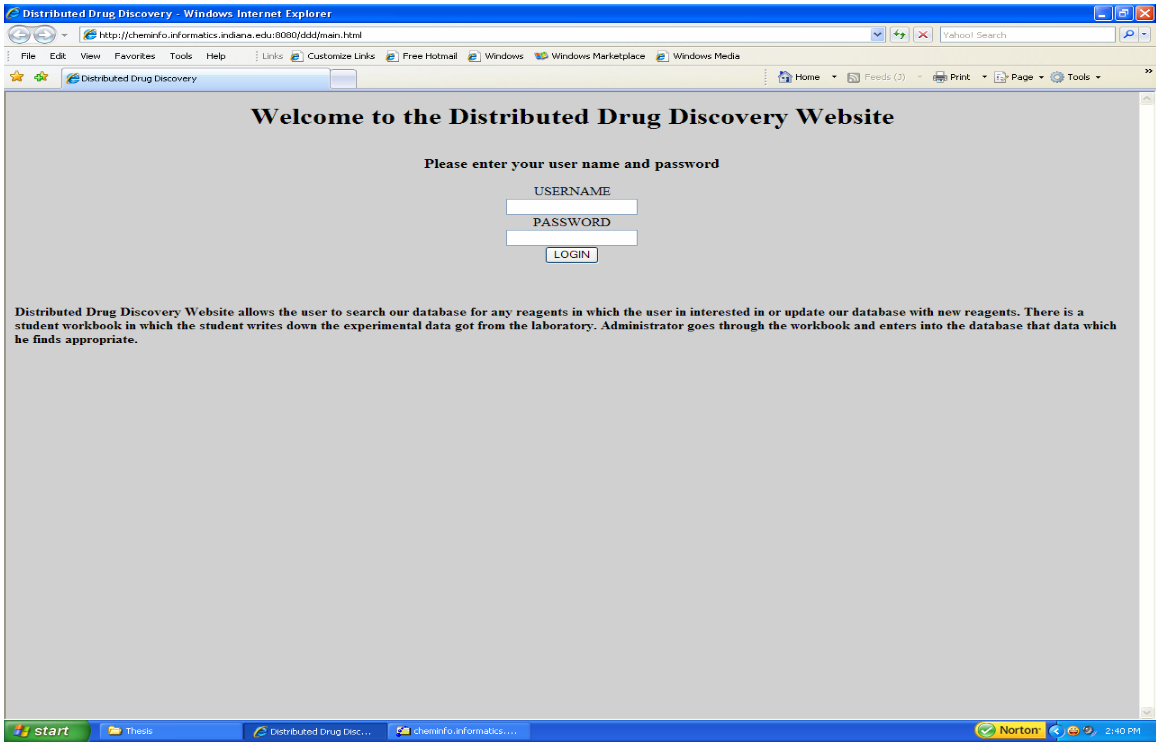
To avoid any malicious input of the data into the database, the main page starts with a Login page where the user is asked to enter a username and password that is unique to him/her. Then he/she is given a choice of either inputting the data into the database or search for the already available data. Then a choice of searching either for the reagents or the products is also given. The queries for searching either the products or reactants are being built by taking into account of the SMILES format of the chemical structure as the chemical structure is unique for that particular compound.

In order to be able to draw the chemical structures, the Marvin Sketch application is included in the web based application. This makes it easy for the user to draw the structure and copy it as SMILES and paste it in the text box provided.

4.2.1. Administrative Interface – 1

The following screen shots demonstrate the step-by-step procedure of the interface.

When the user would like to use the application he/she needs to be authenticated. Figure 5 demonstrates the authentication page. A separate table is being created in the database to enter the usernames and password. In this application, predefined usernames and password are given to the administrator and the student. If the entered information is correct it takes the user to next page of searching or updating the database. Else it would take him to a page where it asks him to retry entering correct information.

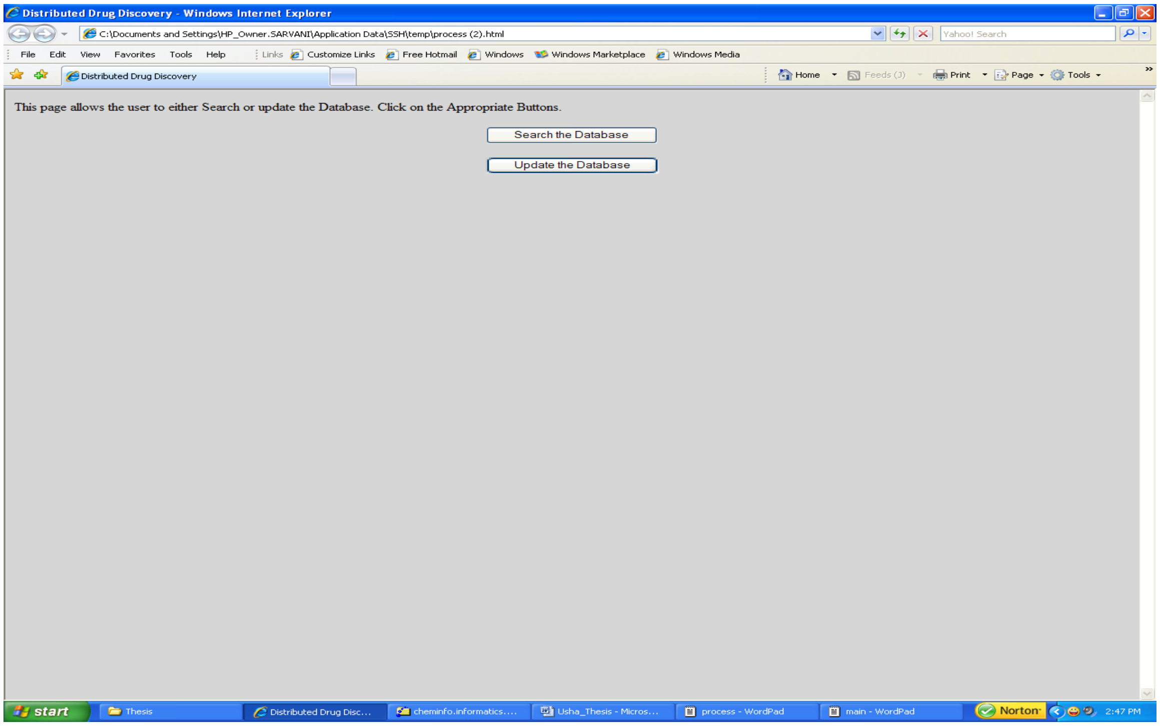

Distributed Drug Discovery - Windows Internet Explorer
http://cheminfo.informatics.indiana.edu:8080/ddd/main.html
USERNAME
PASSWORD
LOGIN

Welcome to the Distributed Drug Discovery Website

Please enter your user name and password

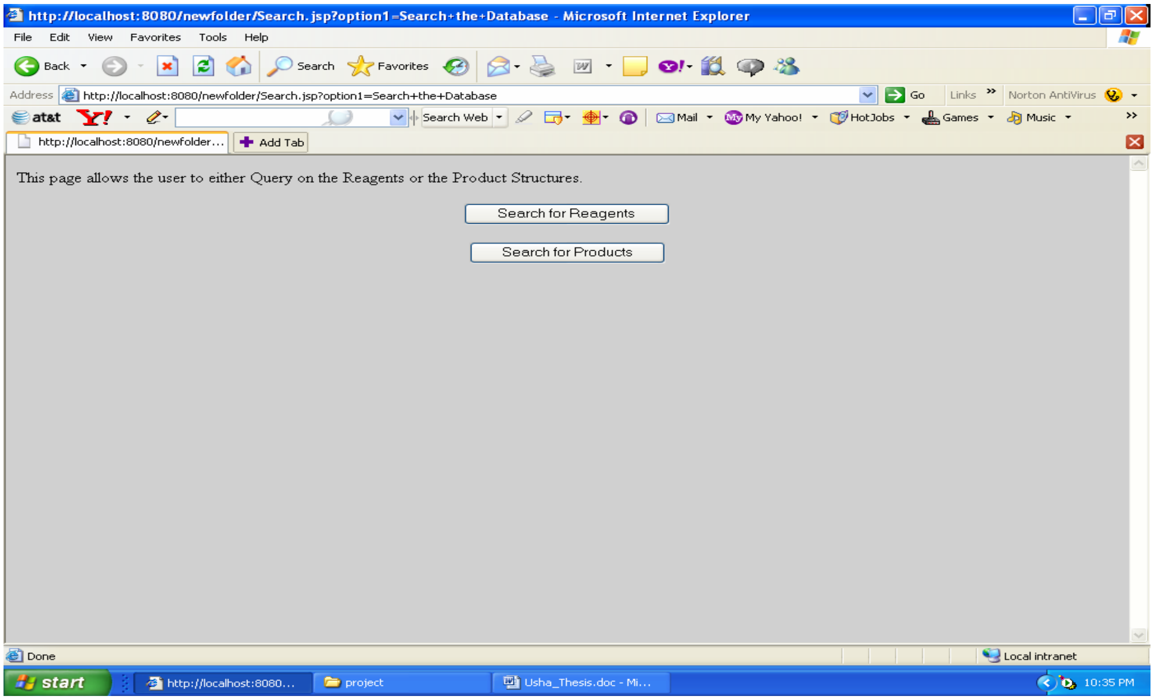
Distributed Drug Discovery Website allows the user to search our database for any reagents in which the user in interested in or update our database with new reagents. There is a student workbook in which the student writes down the experimental data got from the laboratory. Administrator goes through the workbook and enters into the database that data which he finds appropriate.

Figure 5: The Login page.


Distributed Drug Discovery - Windows Internet Explorer
C:\Documents and Settings\HP_Owner.SARVANI\Application Data\SSH\temp\process (2).html
Search the Database
Update the Database

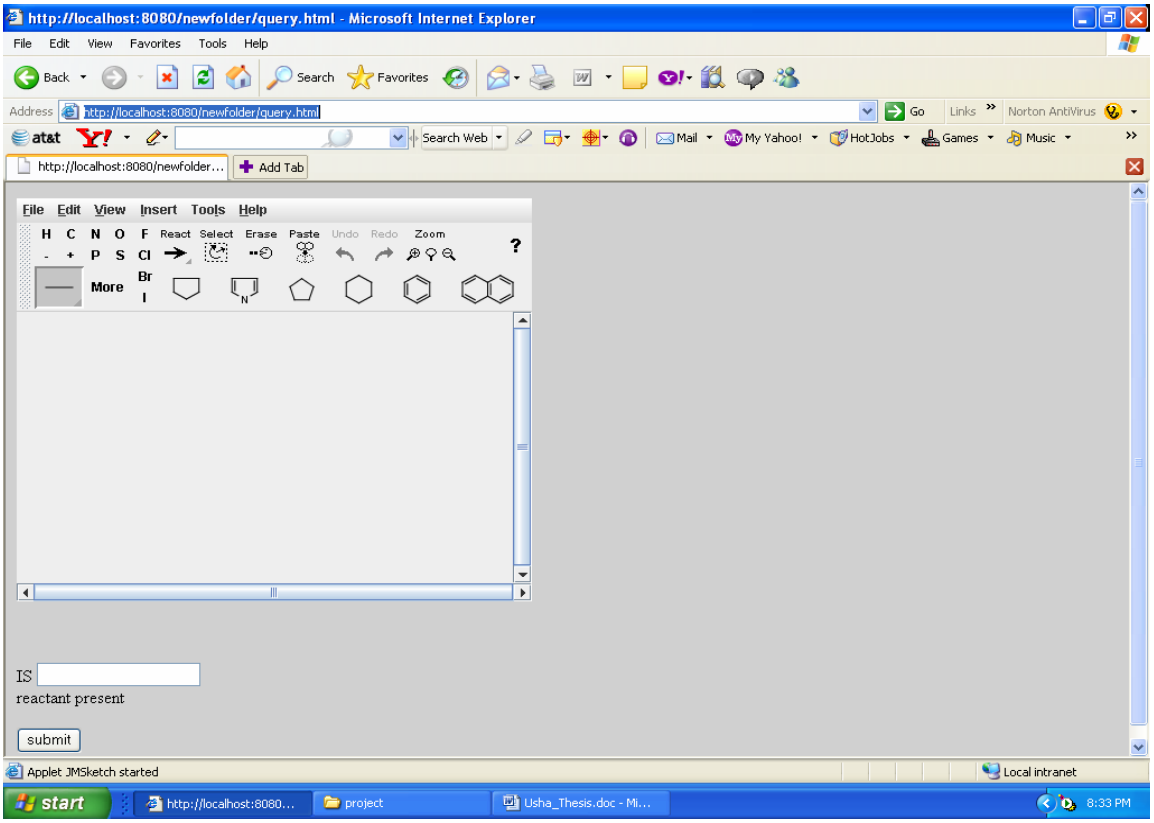
This page allows the user to either Search or update the Database. Click on the Appropriate Buttons.

Figure 6: Option of Searching or Updating the Database


http://localhost:8080/newfolder/Search.jsp?option1=Search+the+Database - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://localhost:8080/newfolder/Search.jsp?option1=Search+the+Database
Search for Reagents
Search for Products
Done
Local intranet
start
10:35 PM

This page allows the user to either Query on the Reagents or the Product Structures.

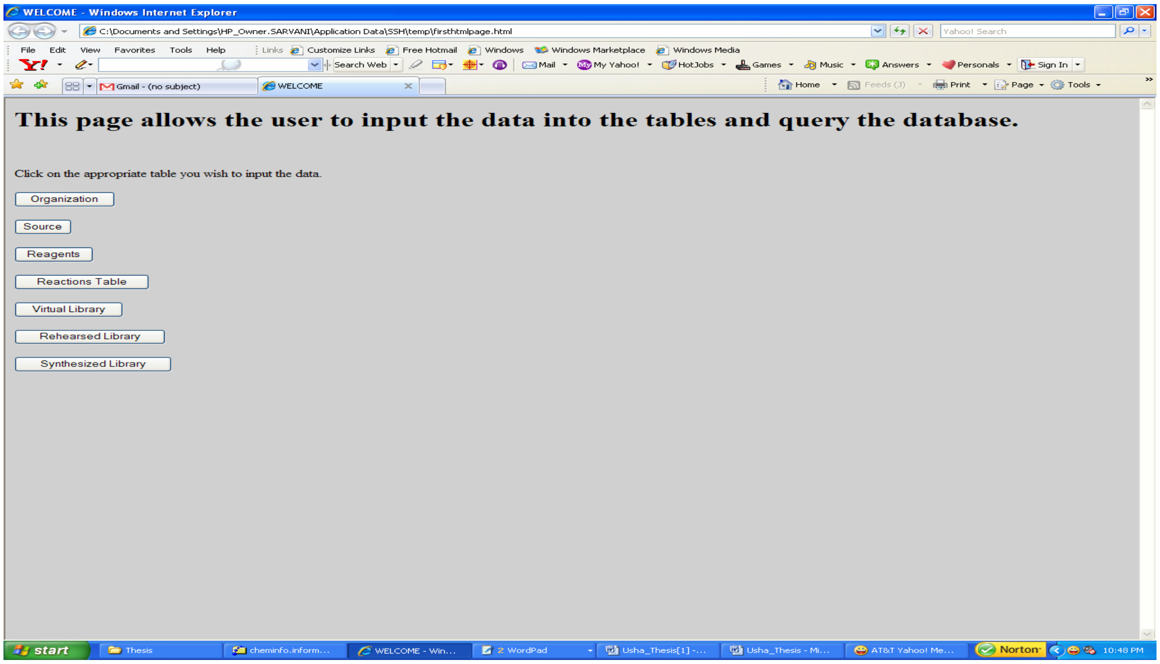
Figure 7: Option of Searching for reagents or products


http://localhost:8080/newfolder/query.html - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://localhost:8080/newfolder/query.html
File Edit View Insert Tools Help
H C N O F React Select Erase Paste Undo Redo Zoom
- + P S Cl
More Br I
submit
Local intranet
start
8:33 PM

IS □
reactant present

Applet JMSketch started

Figure 8: Searching for reactants



This page allows the user to input the data into the tables and query the database.

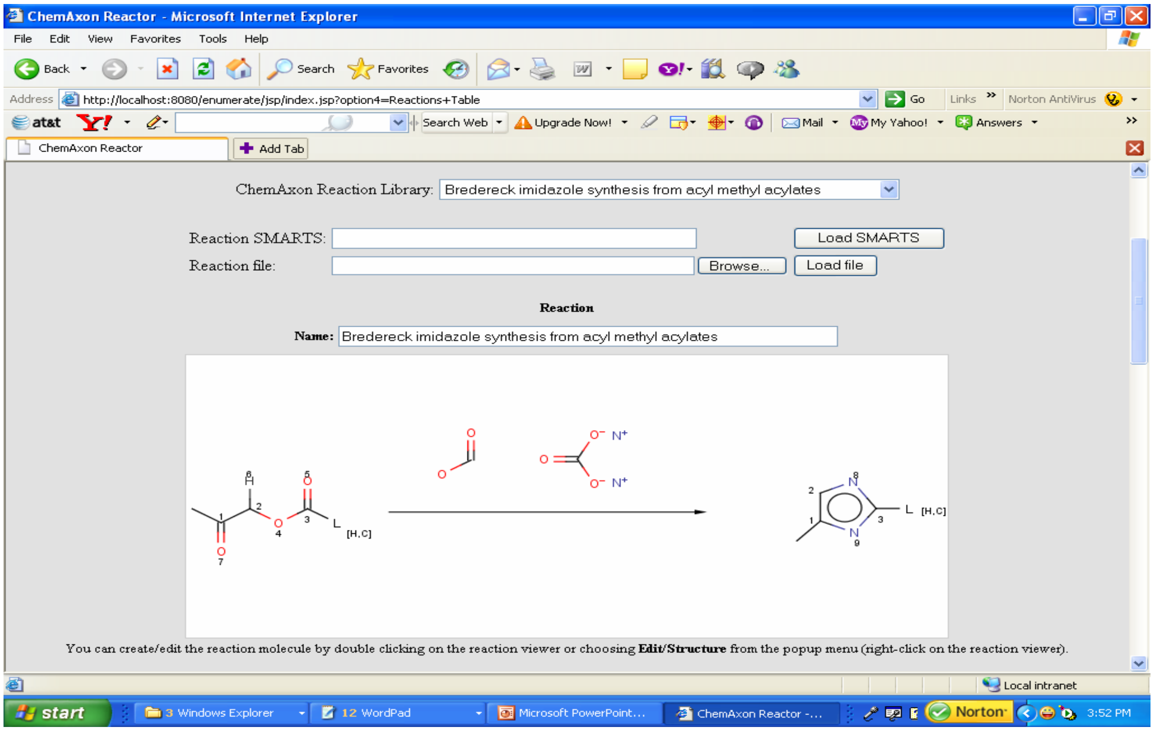
Click on the appropriate table you wish to input the data.

WELCOME - Windows Internet Explorer
C:\Documents and Settings\HP_Owner.SARVANI\Application Data\SSH\temp\firsthtmlpage.html
Organization
Source
Reagents
Reactions Table
Virtual Library
Rehearsed Library
Synthesized Library
start
10:48 PM

Figure 9: Updating the database

Once the administrator decides what he/she wants to do with the database, he/she can update the database or enter the new data into the database. In order to enter or search for the chemical structure information, one needs to draw the structure in the Marvin Sketch provided and copy – paste it as SMILES (Simplified molecular input line entry specification) in the text box provided. The user is given the privilege of storing all the information about the compounds they are working with including the source from where the raw material is purchased, the organization in which the products are made and all the properties associated with the chemical like their molecular weight, molecular formula, CAS (chemical abstract services) number, PubChem number, etc.

The GUI of the ChemAxon reactor is demonstrated in Figure 10, Figure 11 and Figure 12. The first page allows the user to select a predefined reaction from the reaction library or allows one to draw of one’s own choice.



ChemAxon Reactor - Microsoft Internet Explorer

Address: http://localhost:8080/enumerate/jsp/index.jsp?option4=Reactions+Table

ChemAxon Reaction Library: Bredereck imidazole synthesis from acyl methyl acylates

Reaction SMARTS:

Reaction file:

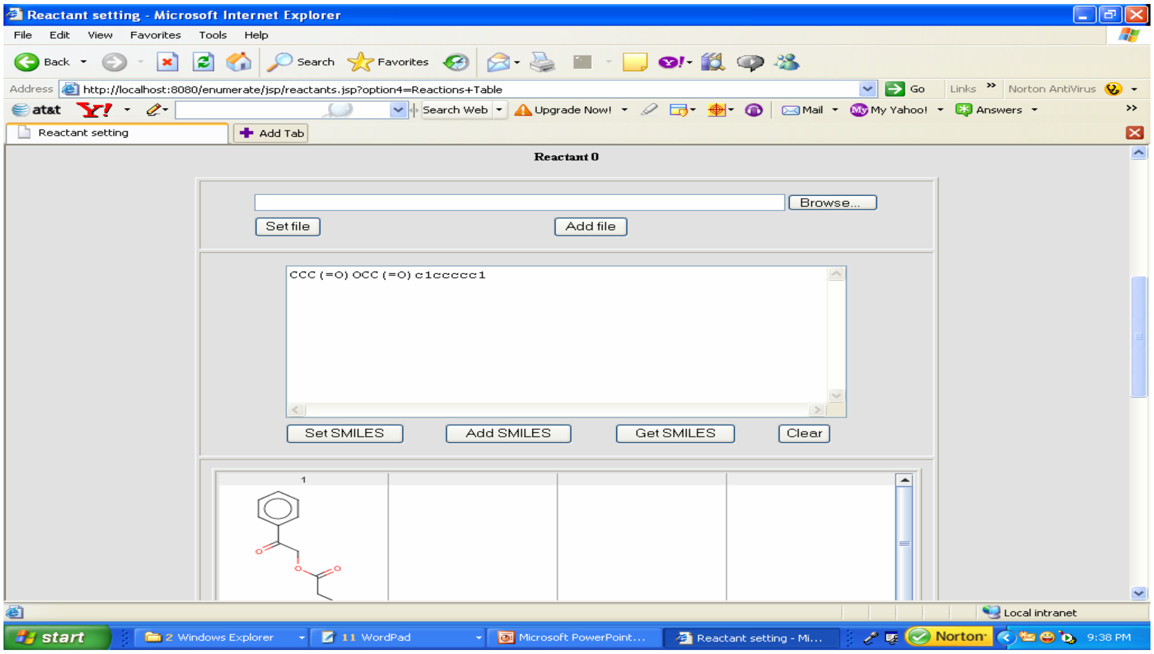
Reaction

Name: Bredereck imidazole synthesis from acyl methyl acylates

You can create/edit the reaction molecule by double clicking on the reaction viewer or choosing **Edit/Structure** from the popup menu (right-click on the reaction viewer).

Figure 10: Selecting a reaction scheme

Once the desired reaction scheme is decided, the second page allows mentioning the reactants participating in the reaction and the third page displays the products.



Reactant setting - Microsoft Internet Explorer

Address: http://localhost:8080/enumerate/jsp/reactants.jsp?option4=Reactions+Table

Reactant 0

```
CCC(=O)OCC(=O)c1ccccc1
```

Figure 11: Entering the reactants participating in the reaction

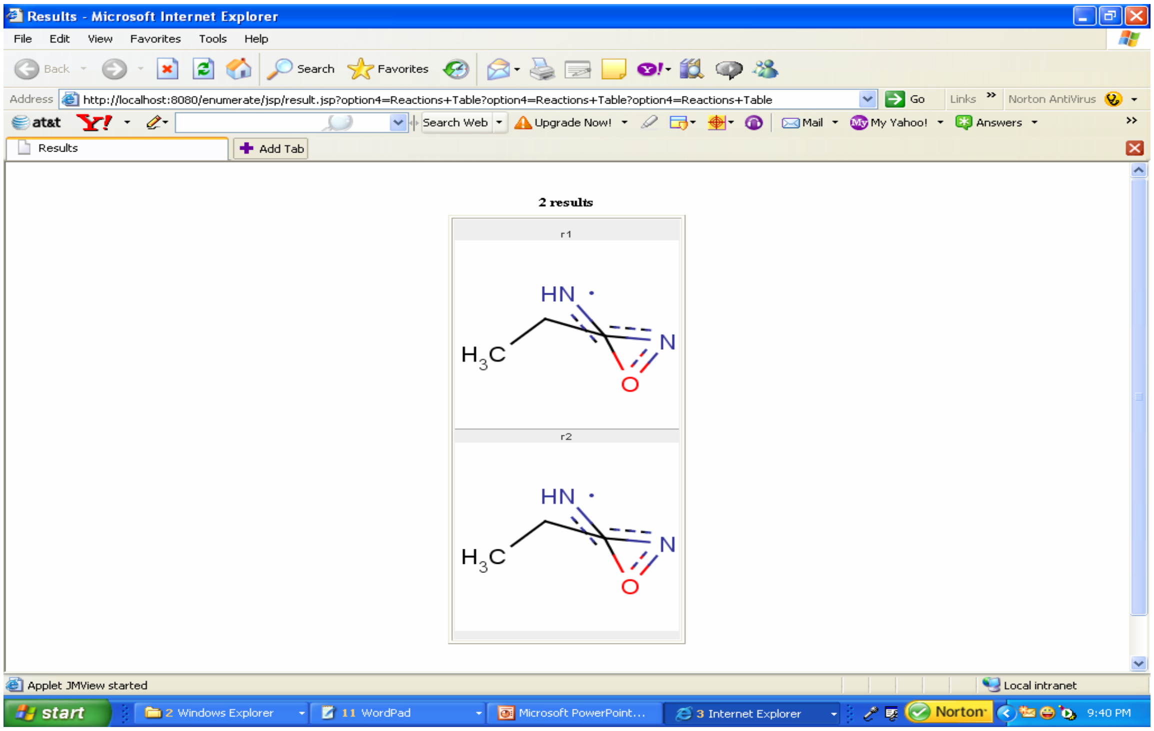

Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://localhost:8080/enumerate/jsp/result.jsp?option4=Reactions+Table?option4=Reactions+Table?option4=Reactions+Table
Go Links Norton AntiVirus
Results
Add Tab
2 results
r1
HN
H3C
N
O
r2
HN
H3C
N
O
Applet JMView started
Local intranet
start
2 Windows Explorer
11 WordPad
Microsoft PowerPoint...
3 Internet Explorer
Norton
9:40 PM

Figure 12: The products

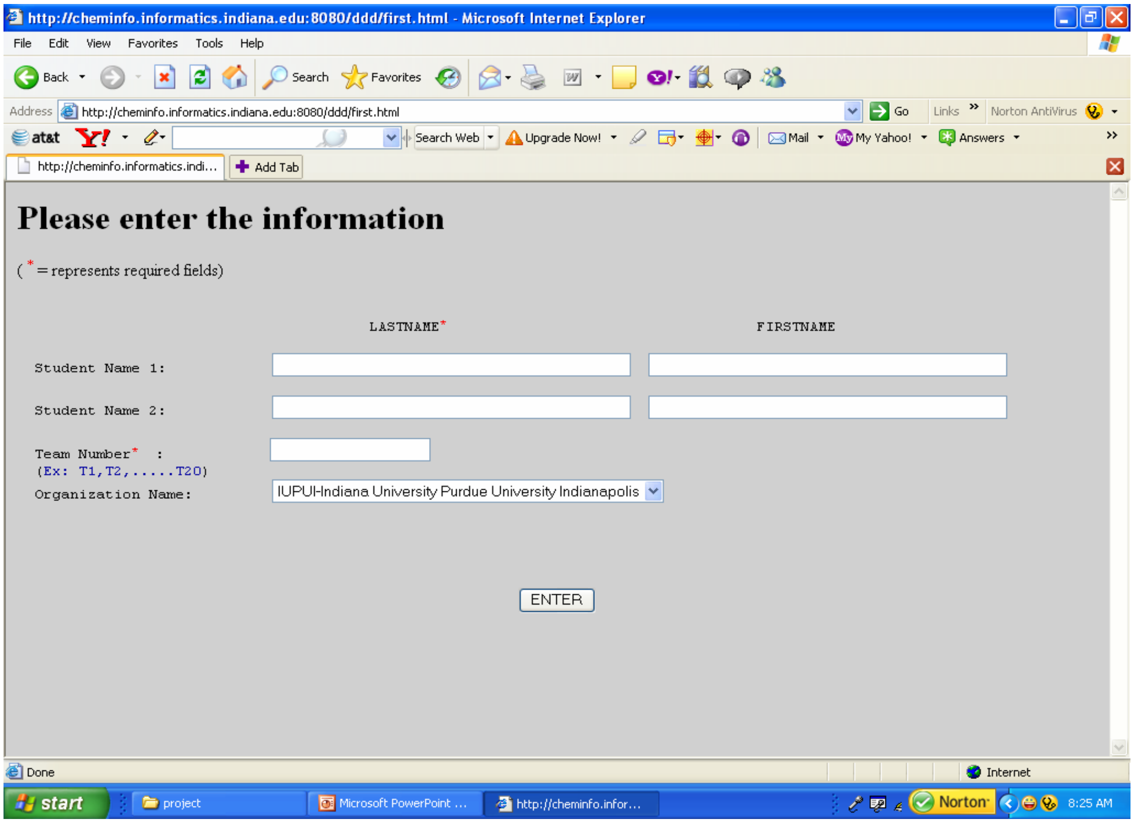
4.2.2. Student WorkBook

4.2.2.1 Student View

The student interface is designed to enter the experimental information from the laboratory into the synthesized table. The main purpose of designing the student interface, instead of entering the experimental information using the administrative interface, is to avoid any malicious input of data by the students. Also, the students may or may not have the complete experimental information available, which would result in incomplete data input.

In order to achieve this task, two tables are created in the database. One table enables the storage of student information along with the team they are working. The second table stores the information of the experimental results. The first table stores information about the student's last name, first name and the team number. Last name and team number together forms the primary key for this table. In the second table, experimental information like starting material structures, product structure, mass and weight of the product, crude percent yield, weight of purified product, purified percent yield, predicted mass, retention time, hplc purity, retention time and hplc of the unknown products is recorded.

To keep track of the experimental information entered by the students, the first page is designed to store the name of the students and the team number involved in the experiments as shown in figure 13. The student's last name and their team numbers are unique keys to restore the information entered and also to keep track of the information record.



http://cheminfo.informatics.indiana.edu:8080/ddd/first.html - Microsoft Internet Explorer

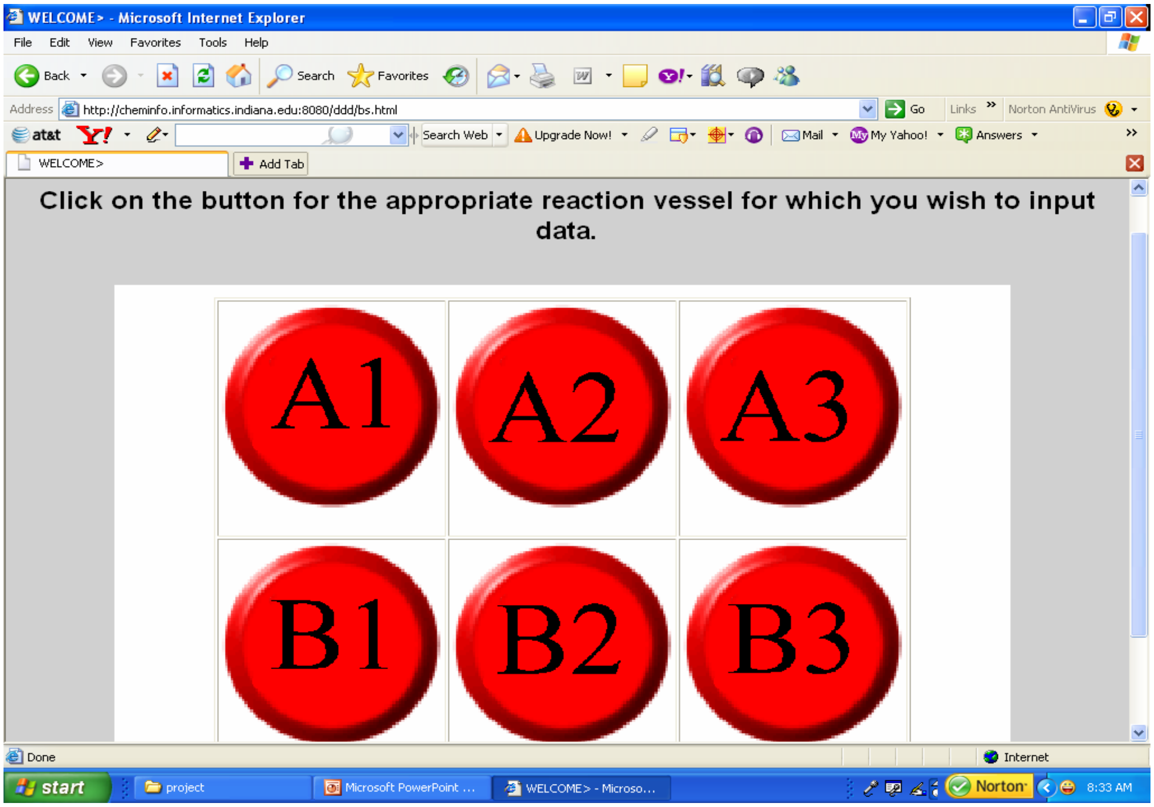
Please enter the information

(* = represents required fields)

	LASTNAME*	FIRSTNAME
Student Name 1:		
Student Name 2:		
Team Number* : (Ex: T1,T2,.....T20)		
Organization Name:	IUPUI-Indiana University Purdue University Indianapolis	

Figure 13: Student entry information

When the student enters the required information, the query behind would check if the person is already present in the database. If it is a new entry into the database, the user is welcomed to the student workbook. The information about the user will be updated in the database. It enables the user to enter the experimental information in the student workbook. The experiment entry page is designed as a “Bill board” (a replica of the experimental kit used in the laboratory) represented as in Figure -14.

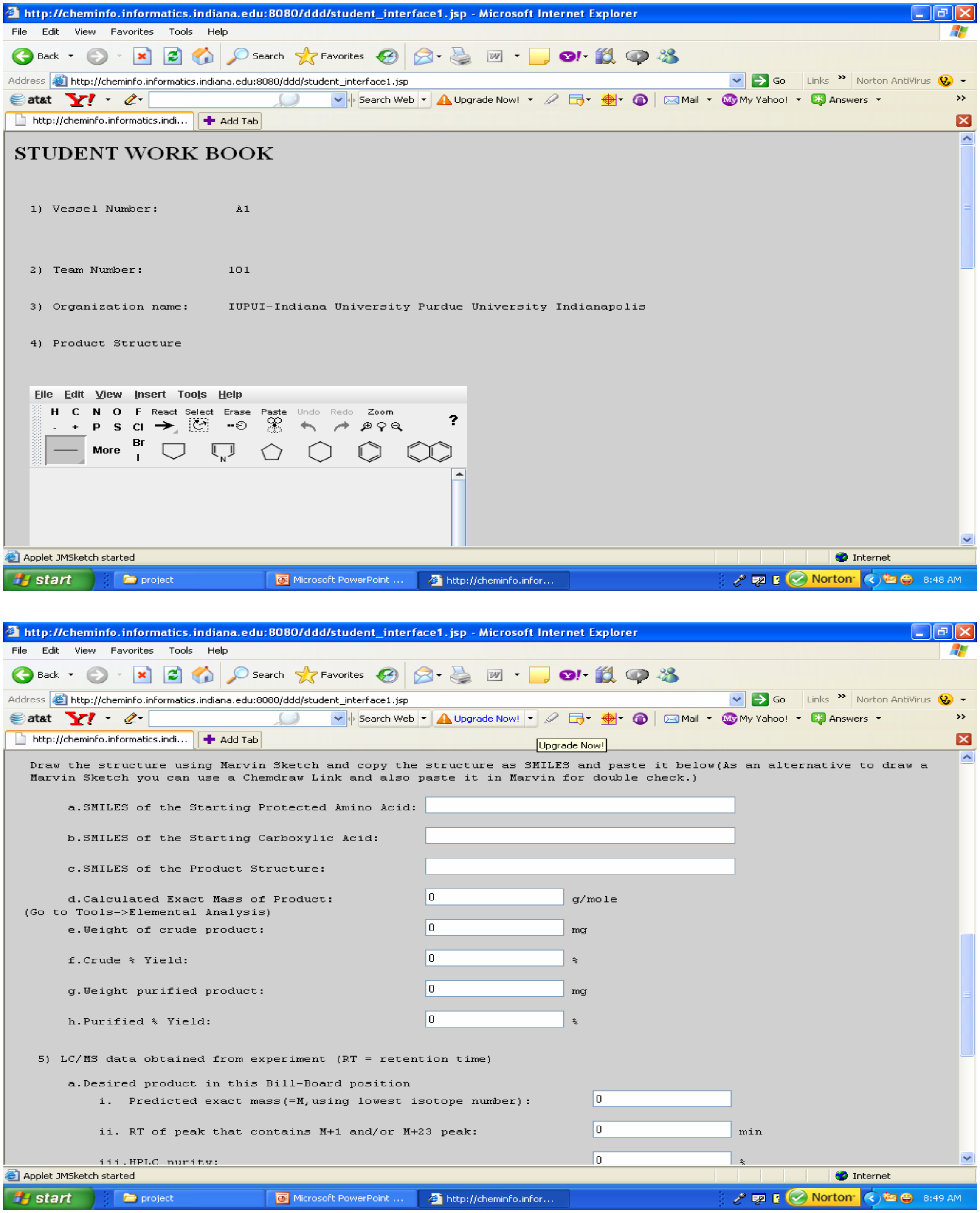

WELCOME> - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://cheminfo.informatics.indiana.edu:8080/ddd/bs.html
Go Links Norton AntiVirus
at&t Search Web Upgrade Now! Mail My Yahoo! Answers
WELCOME> Add Tab

Click on the button for the appropriate reaction vessel for which you wish to input data.

A1
A2
A3
B1
B2
B3
Done
Internet
start project Microsoft PowerPoint ... WELCOME> - Microso... Norton 8:33 AM

Figure 14: Experimental Setup page

“A1, A2, A3, B1, B2, B3” buttons represent the experimental setup of the vessels (Bill Board) used in the experiment. The students need to keep track of the information in all these vessels. In order to avoid confusion for the students while entering the data, the experimental set up page is designed as the one used in the real laboratory. When the student clicks on a desired button, he/she is taken into a new page where he/she enters the experimental information as shown in Figure 15. In this way the student can enter the information correctly.


http://cheminfo.informatics.indiana.edu:8080/ddd/student_interface1.jsp - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://cheminfo.informatics.indiana.edu:8080/ddd/student_interface1.jsp
Go Links Norton AntiVirus
at&t Search Web Upgrade Now! Mail My Yahoo! Answers
http://cheminfo.informatics.indi... Add Tab
File Edit View Insert Tools Help
H C N O F React Select Erase Paste Undo Redo Zoom
- + P S Cl Br I More ?
Applet JMSketch started
Internet
start project Microsoft PowerPoint ... http://cheminfo.infor... Norton 8:48 AM
Upgrade Now!
8:49 AM

STUDENT WORK BOOK

1) Vessel Number: A1

2) Team Number: 101

3) Organization name: IUPUI-Indiana University Purdue University Indianapolis

4) Product Structure

Draw the structure using Marvin Sketch and copy the structure as SMILES and paste it below(As an alternative to draw a Marvin Sketch you can use a Chemdraw Link and also paste it in Marvin for double check.)

a.SMILES of the Starting Protected Amino Acid:

b.SMILES of the Starting Carboxylic Acid:

c.SMILES of the Product Structure:

d.Calculated Exact Mass of Product: 0 g/mole
(Go to Tools->Elemental Analysis)

e.Weight of crude product: 0 mg

f.Crude % Yield: 0 %

g.Weight purified product: 0 mg

h.Purified % Yield: 0 %

5) LC/MS data obtained from experiment (RT = retention time)

a.Desired product in this Bill-Board position

i. Predicted exact mass(=M,using lowest isotope number): 0

ii. RT of peak that contains M+1 and/or M+23 peak: 0 min

iii.HPLC purity: 0 %

Figure 15: The Experiment Information page

The student interface keeps track of the reagents structure, product structure, the vessel number, the team number, mass of the product, percentage yield of the product, weight of the purified product, and LC/MS (liquid chromatography mass spectroscopy) data of the product.

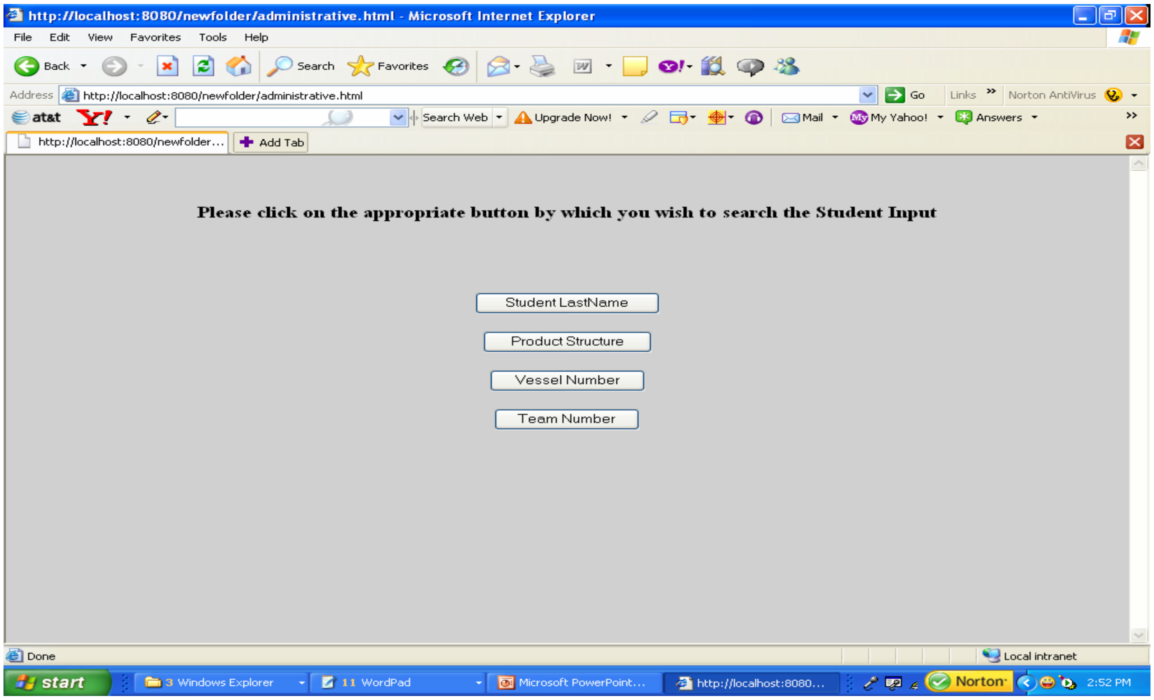
If the student does not have the complete experimental information, he can always come back and enter his information (last name and team number). The query checks if the person is already present in the database. If the student is already in the database, he/she will be notified that the name is already present in the dataset and takes him to the buttons page. When he/she clicks on the desired button, a page is displayed which contains the information already entered by the student and allows him to modify data that will be automatically updated in the database.

The student interface aids in entering the new information, modifying the information, deleting the wrong ones and providing with necessary instructions on each page. Marvin Sketch is embedded to facilitate in drawing the chemical structure and enter the structural information in the SMILES format.

4.2.3 Instructor View

Once all the experiments are done, and the students enter the information into the student workbook, the administrator needs to check if the information entered is correct or not. To perform this task, another interface is being created where the administrator can check the students entered information.

The administrator is given a choice of searching the student input by queries based on student last name, product structure, team number and vessel number as these all terms are unique in the student's entry. The interface is as shown in Figure 16.



http://localhost:8080/newfolder/administrative.html - Microsoft Internet Explorer

Please click on the appropriate button by which you wish to search the Student Input

Figure 16: Administrative interface

Once the administrator clicks on the desired button, he is taken into the appropriate page where he can enter the query. One more advantage of this interface is that the administrator can grade the students based on experimental information input. Once the administrator checks and validates the information entered by the students, he can enter the correct information into the original synthesized library table.

4.3. Usability Study

4.3.1. Web Application for Student Interface

Once the database is created along with the interfaces, it is desired to check the usefulness of the interface. The student interface is made available to the students to enter the experimental information. Once the students have used the interface to enter their data, they all are requested to participate in a usability survey to give feedback on the design of the interface.

In order to conduct this study, a web survey option is chosen. A web survey is defined as an online survey software tool which can easily aid in designing, creating and emailing a survey. An online survey option is chosen since the survey is to facilitate access to the survey by user. Online surveys play an important role in conducting a survey on a website or an interface. Online surveys are often easily and quickly assembled with minimal cost or sometimes free. Distribution of the survey is very broad and quick with an assumption that one has an up-to-date list of email addresses. Online surveys can usually produce high response rates since there would be a direct link to the survey in the email announcement. It is also easy to share the survey results with others via a direct link to the survey tool. Data is captured electronically so no manual data entry is necessary.

The survey tool was provided by stellarsurvey.com and used for conducting the online survey of the student interface. The main concepts that are considered for analyzing are the usability of the website, the use of Marvin Sketch in the application, comfort using a web base data entry rather than a paper based entry, and the overall impression of the interface.

When analyzing the website usability, the sub-concepts that are considered are its user-friendly nature, the instructions provided in using the website, the legibility of the instructions, navigation through the site and the size of the fields. A ranking scale of very good-good-average-poor-very poor is used in order to provide easy answering to the survey.

In analyzing the Marvin Sketch, the sub-concepts that are considered are its usefulness, complex nature in using the program and the overall impression of Marvin Sketch. A ranking scale of Useless to Very useful is used in rating the usefulness nature of Marvin Sketch. A ranking scale of Simple/Easy to Use to Very confusing is used in rating the complexity while using Marvin Sketch. A ranking scale of Irritating to Pleasing is used in rating the Overall impression of Marvin Sketch.

Until the interface is made available to the students, all the experimental and theoretical data was entered into a lab notebook. So, a survey on finding out the advantageousness on using a computer entry rather than a paper entry is desired. A ranking scale of Disadvantage-No Difference-Highly Advantageous is used to find out the advantage of using a computer rather than a paper entry. Also a ranking scale of Very Easy-Easy-Average-Hard-Very Difficult is used to find out how easy was it to update and edit the information entered into the database.

Finally, in order to know the overall performance of the interface, the sub-concepts that are considered are the website usefulness, complexity and the overall impression. A ranking scale of varying from useless to very useful is used to rate the usefulness of the website. A ranking scale of simple/easy to use to very confusing is used to rate the complex nature of the interface. A ranking scale of irritating to pleasing is used to rate the overall impression of the website. On a final note a textbox is provided to suggest any further modifications/comments on the interface.

Once the survey is defined, it is being distributed to all the students who used the survey through their email. This survey is anonymous as no information about the participants to the survey is being recorded. Once the students received the email, they could click on the survey link provided, complete the survey and submit their responses.

4.3.2. Chemical Drawing Packages

Several Chemical Informatics software programs or drawing packages are freely available for both commercial and academic applications. These programs are designed with the idea to assist researchers in various research areas. But it is not clear whether they provide sources to be able to easily follow and understand the software to all classes of users. This is meant to say that in a pharmaceutical company or in an academic institution there may be users with good background in chemistry or programming or both or none and are willing to use the software. None of these categories users know whether the software provides all the resources for them to complete the task unless they work on it.

The goal of this study is to test the usability nature on four types of chemical informatics drawing packages. The programs that are chosen for the study are ChemDraw, ISIS Draw, Marvin Sketch and Java Molecular Editor (JME).

Usability Subjects

The subjects chosen for the study range from novice to expert in the field. This is usually done in the usability studies to ensure that all classes of expertise are comfortable with the software being tested. To do so, help of 8 participants with different strengths and backgrounds is enlisted. They are as follows: three Chemistry Majors, two Programming + Chemistry background, one Human Computer Interaction person, one Biology with good Chemistry background and an Application user who used several drawing packages like AutoCAD, Catia etc.

Program acquisition

All the programs are downloaded on a laptop computer. To obtain a two week free academic trial of a fully functional version of ChemOffice Ultra 9.0., registration was required at the www.cambridgesoft.com. The ISIS draw is downloaded from the "MDL" website http://www.mdli.com/. Marvin Sketch is downloaded from http://www.chemaxon.com/marvin/do-download.html. Java Molecular Editor is obtained from Peter Ertl, who designed this software as chemical informatics software at Novartis, through email.

All the four programs are saved under the name Usability Study folder on the laptop and observed the order as shown in the following Table. Both the test giver and the participant knew on which program he/she is working on.

Table 8: Order of the software tested.

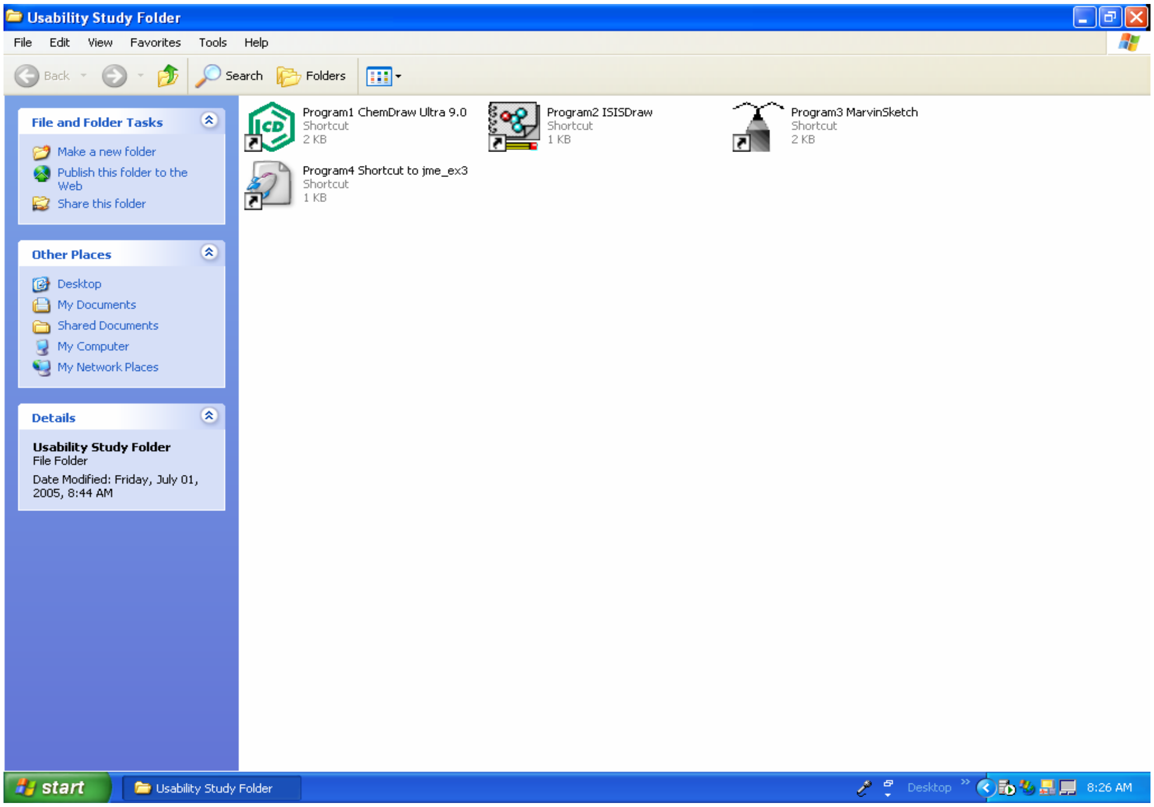
Sl. no	Title of the Program	Company
Program 1	ChemDraw	CambridgeSoft
Program 2	ISISDraw	MDL
Program 3	Marvin Sketch	ChemAxon
Program 4	JME Molecular Editor	Novartis

Task Definition

The next step is to decide the tasks to be given to the participants. Keeping in mind that the study is not to test the participants but the software, three tasks of drawing molecules were designed by increasing their complexity one after another. An official package containing a series of steps to be followed was given to each participant. The package can be found in Appendix A. It starts with a study consent form which tells the participants about the study and records their willingness for the study. This is followed by information which clearly states to the participants that it is the software being tested and not their aptitude. It also discusses the procedure and the tasks to be performed for each of the software applications.

Experimental Set-Up

Short cuts to all four programs are arranged under the name “Usability Study Folder” to avoid any confusion to the participants. The study folder looked as shown in Figure 17.



Program1 ChemDraw Ultra 9.0
Shortcut
2 KB

Program2 ISISDraw
Shortcut
1 KB

Program3 MarvinSketch
Shortcut
2 KB

Program4 Shortcut to jme_ex3
Shortcut
1 KB

Usability Study Folder
File Folder
Date Modified: Friday, July 01, 2005, 8:44 AM

Figure 17: The Usability Folder

The study is initiated by handing the usability study packet to the participants. Analysis report is being maintained on the participants approach in fulfilling the tasks. The participants were to pretend as if no one is observing them and it is an exam and so should not ask questions while trying to complete the tasks. At the end of the study each participant is asked to fill out a study survey shown at the end of Appendix-A. The survey is comprised of a satisfaction poll measuring how irritating/pleasing, complicated/straightforward, useless/useful each software seemed to them. It also allows them to provide any comments/suggestions about the tasks given to them along with the four programs.

4.4 Chemical Facets

4.4.1 Online Chemical Databases

The directory of databases is increasing in number each year in the field of science, technology, business and other areas. In 1980 there were about 500 computer readable databases and the number increased to 2900 databases by 1986 from about 450 different sources[7]. In chemistry itself there are around 100 open source databases like NCBI PubChem, Chemical entities of biological entities (ChEBI). In many cases the same database is available from more than one source[8]. Most of the available chemical databases are proprietary like Scifinder Scholar, CambridgeSoft, etc.

With such huge data and diverse databases, the user is required to go through several databases to search and retrieve information. In this process the chances of skipping important information or data source is likely and of course time consuming. This results in loss of information, less efficacy and efficiency in data searching. The need arises for such a tool which can query diverse data sources; integrate multiple data sources and present voluminous and diverse results just in one step process.

4.4.2 Facets

A faceted classification system facilitates the assignment of multiple classifications to a single record, allowing searching and browsing through several classes. A facet is a method of classification. It groups the results which have the same value for a particular category and provides a view of the result set classified according to each category. FacetMap[26] provides automated tools to develop faceted classification systems. The Flamenco[27] search tool is an apt example. Similar type of classification called BioFacets[17] has been proposed towards integration of biological databases at IUPUI, Indianapolis. This is a new integration system for web based biological databases. The main feature of the system is the proposal of a new faceted classification system that allows dynamic categorization of large amounts of data records retrieved from the searched databases. A wrapper-mediator approach[17] is adopted in the integration of the databases. In this approach, wrappers specific to each remote data source are applied to extract and translate data from the remote source to the integrated system. BioFacets provides wrappers only for HTTP based biological search databases. The first task of the wrapper is to convert the user query into a database specific query which is achieved by construction of the query URL specific to the user query. The second task of the wrapper is converting query results into an internal format via a set of extraction rules. The function of mediator[17] in the wrapper-mediator approach is processing of results returned by the wrappers. The mediator performs the task of assigning values to each record and performing classification accordingly. This is known as a faceted classification system.

BioFacets is an existing framework[17] that enables a user to

1) Define a faceted scheme

2) Integrate the databases using the defined faceted scheme, for biological databases.

The same principle and framework can be applied to Chemical databases. The main aim of Chemical Facets is to define the faceted scheme and integrate the databases using the defined faceted scheme. This framework applies for only HTTP based chemical search databases.

Two types of facets are used: static facets and dynamic facets. Static facets are facets for which the value assigned to a record is determined without knowledge of the record, usually using the information about the database which the record belongs to. For example, the static facet "Data type" will take a fixed value from the set (Compound data, Bioassay data, Substance data, chemical data). Dynamic facets is one for which a value is assigned by using the record. There are two methods by which values are assigned to facets. In the first method, the value of a facet may correspond to the value of one or more fields in a record. In the second method, the value is extracted from a third party database.

Facets are also hierarchical or non-hierarchical facets. Non-hierarchical facets are single layered. They are usually static facets i.e., fixed facets. Ex: data type, data source. Hierarchical facets take a value represented by a path through a hierarchical tree. These may be static or dynamic. In static hierarchical facets, the hierarchy is defined first and then the facet value is obtained. In dynamic hierarchical facets, the facets values are assigned first and then the hierarchy is obtained.

Thus finally a facet value is specified in three ways. The static facet value is assigned using the fixed value rule while the dynamic facet value is assigned using either a field value or a lookup value rule. The rules are written using EXtended Markup Language (XML) known as database schemas. Database schemas play an important role in querying and extracting the data sources.

4.4.3 Chemical Facets Design

4.4.3.1 Chemical Databases Considered For Study And Their Characteristics:

The databases that are considered for study are: PubChem Substance, PubChem Compound, PubChem BioAssay, Chemical Entities of Biological Interest (ChEBI), ChemExper: Directory of chemicals, The UC Irvine ChemDB, DrugBank. All the databases are freely available through internet. The first three databases of PubChem are components of NIH[28]'s (National Institute of Health) Molecular Libraries Roadmap Initiative. PubChem provides information on the biological activities of small molecules. It is integrated with Entrez[29,] NCBI (National Center for Biological Information)'s primary search engine, and also provides compound neighboring, sub/superstructure, similarity structure, bioactivity data, and other searching features. ChEBI is a freely available dictionary of molecular entities focused on 'small' chemical compounds. ChemExper is a freely accessible database of chemicals over internet. ChemDB is a database of chemicals of various criteria. DrugBank is a database comprising of drug information of chemical compounds. Databases comprising of common and diverse properties of chemicals have been chosen for the integration. The intention of such integration system is to ease the researchers' effort to be able to get diverse information from a single interface.

PubChem

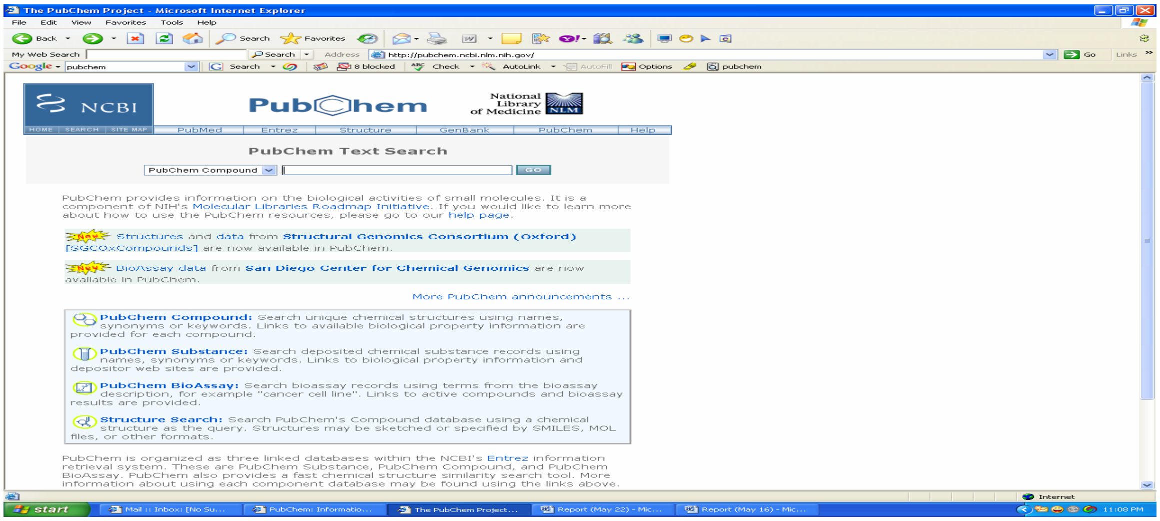
PubChem is a database of chemical molecules. It contains mostly small molecules with a molecular mass below 500.

It consists of three databases:

1. PubChem Compound with 5.2 million entries, contains pure and characterized chemical compounds.
2 PubChem Substance with 7.7 million entries, contains mixtures, extracts, complexes and uncharacterized substances.
3. PubChem BioAssay contains bioactivity results from 176 high throughput screening programs with several million values.

PubChem consists of a molecule editor to allow structure search on the database. Searching these databases provide a broad range of properties including chemical structure, name fragments, chemical formula, molecular weight, XLogP, hydrogen bond donor and acceptor count, bioactivity, SMILES and InChi strings, links to structurally related compounds and other NCBI databases like PubMed.

The home page of PubChem is represented as shown in figure 18:


The PubChem Project - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
My Web Search Search Address http://pubchem.ncbi.nlm.nih.gov/
Go Links
Google pubchem Search 8 blocked Check AutoLink AutoFill Options pubchem
NCBI
PubChem
National Library of Medicine NLM
HOME SEARCH SITE MAP PubMed Entrez Structure GenBank PubChem Help
PubChem Text Search
PubChem Compound
GO
New
New
More PubChem announcements ...
Internet
start
Mail :: Inbox: [No Su...
PubChem: Informatio...
The PubChem Project...
Report (May 22) - Mic...
Report (May 16) - Mic...
11:08 PM

PubChem provides information on the biological activities of small molecules. It is a component of NIH's Molecular Libraries Roadmap Initiative. If you would like to learn more about how to use the PubChem resources, please go to our help page.

Structures and data from Structural Genomics Consortium (Oxford) [SGCOxCompounds] are now available in PubChem.

BioAssay data from San Diego Center for Chemical Genomics are now available in PubChem.

PubChem Compound: Search unique chemical structures using names, synonyms or keywords. Links to available biological property information are provided for each compound.

PubChem Substance: Search deposited chemical substance records using names, synonyms or keywords. Links to biological property information and depositor web sites are provided.

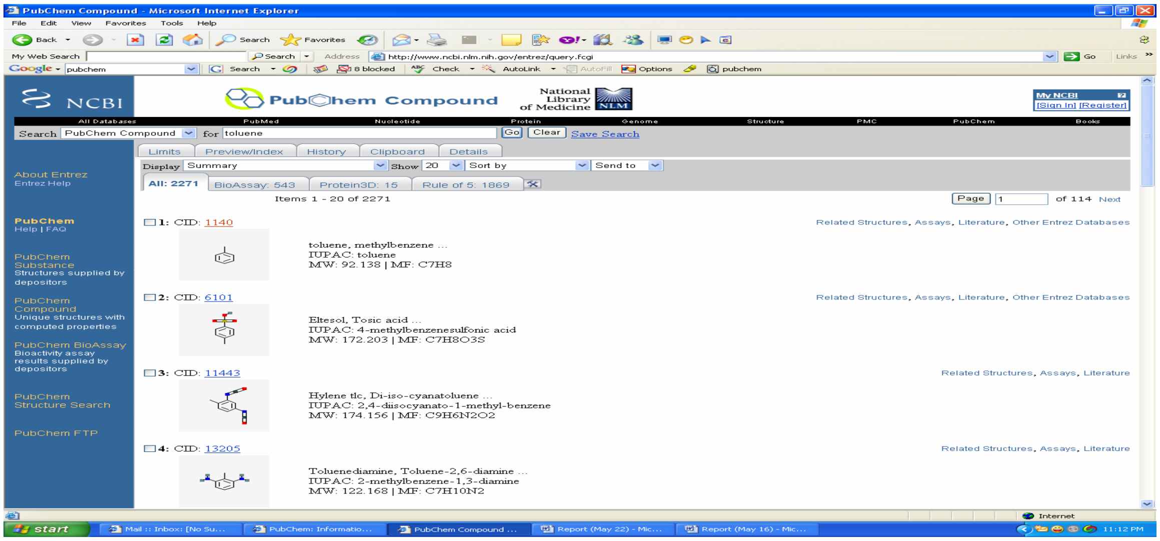
PubChem BioAssay: Search bioassay records using terms from the bioassay description, for example "cancer cell line". Links to active compounds and bioassay results are provided.

Structure Search: Search PubChem's Compound database using a chemical structure as the query. Structures may be sketched or specified by SMILES, MOL files, or other formats.

PubChem is organized as three linked databases within the NCBI's Entrez information retrieval system. These are PubChem Substance, PubChem Compound, and PubChem BioAssay. PubChem also provides a fast chemical structure similarity search tool. More information about using each component database may be found using the links above.

Figure 18: PubChem Page

By default PubChem Compound option is automatically selected. The PubChem Compound Database contains validated chemical depiction information that is provided to describe substances in PubChem Substance. For example, when a query is being processed, a resultant page would give the summary of hits of the compounds corresponding to the query as shown in the figure 19.



PubChem Compound - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
My Web Search Search Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi Go Links
Google pubchem Search 8 blocked Check AutoLink AutoFill Options pubchem
NCBI
PubChem Compound
National Library of Medicine NLM
My NCBI
[Sign In] [Register]
All Databases PubMed Nucleotide Protein Genome Structure PMC PubChem Books
Go Clear Save Search
Limits Preview/Index History Clipboard Details
Send to
Page
Next
About Entrez
Entrez Help
PubChem
Help | FAQ
PubChem Substance
Structures supplied by depositors
PubChem Compound
Unique structures with computed properties
PubChem BioAssay
Bioactivity assay results supplied by depositors
PubChem Structure Search
PubChem FTP
Internet
start Mail :: Inbox: [No Su... PubChem: Informatio... PubChem Compound ... Report (May 22) - Mic... Report (May 16) - Mic... 11:12 PM

Search PubChem Compound for toluene

Display Summary Show 20 Sort by

All: 2271 BioAssay: 543 Protein3D: 15 Rule of 5: 1869

Items 1 - 20 of 2271 Page 1 of 114

□1: CID: 1140 Related Structures, Assays, Literature, Other Entrez Databases
toluene, methylbenzene ...
IUPAC: toluene
MW: 92.138 | MF: C7H8

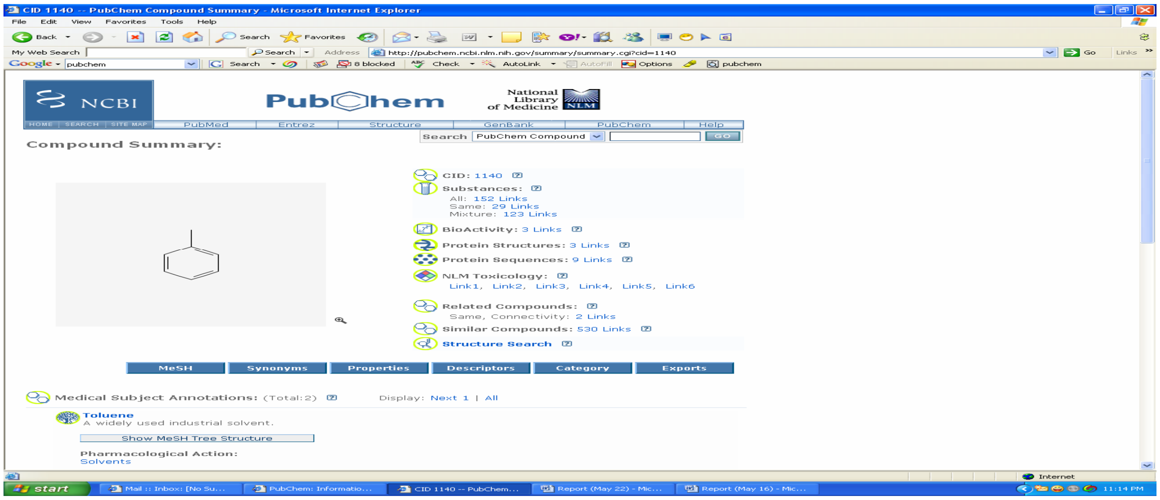
□2: CID: 6101 Related Structures, Assays, Literature, Other Entrez Databases
Eltesol, Tosic acid ...
IUPAC: 4-methylbenzenesulfonic acid
MW: 172.203 | MF: C7H8O3S

□3: CID: 11443 Related Structures, Assays, Literature
Hylene tlc, Di-iso-cyanatoluene ...
IUPAC: 2,4-diisocyanato-1-methyl-benzene
MW: 174.156 | MF: C9H6N2O2

□4: CID: 13205 Related Structures, Assays, Literature
Toluenediamine, Toluene-2,6-diamine ...
IUPAC: 2-methylbenzene-1,3-diamine
MW: 122.168 | MF: C7H10N2

Figure 19: PubChem Compound Records Result page

When one record is selected, the compound Summary is displayed as shown in Figure 20.



CID 1140 -- PubChem Compound Summary - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
My Web Search Search Address http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=1140 Go Links
Google pubchem Search 8 blocked Check AutoLink AutoFill Options pubchem
NCBI
PubChem
National Library of Medicine NLM
HOME SEARCH SITE MAP PubMed Entrez Structure GenBank PubChem Help
GO
MeSH Synonyms Properties Descriptors Category Exports
Show MeSH Tree Structure
Internet
start Mail :: Inbox: [No Su... PubChem: Informatio... CID 1140 -- PubChem... Report (May 22) - Mic... Report (May 16) - Mic... 11:14 PM

Search PubChem Compound

Compound Summary:

CID: 1140
Substances:
All: 152 Links
Same: 29 Links
Mixture: 123 Links
BioActivity: 3 Links
Protein Structures: 3 Links
Protein Sequences: 9 Links
NLM Toxicology:
Link1, Link2, Link3, Link4, Link5, Link6
Related Compounds:
Same, Connectivity: 2 Links
Similar Compounds: 530 Links
Structure Search

Medical Subject Annotations: (Total:2) Display: Next 1 | All

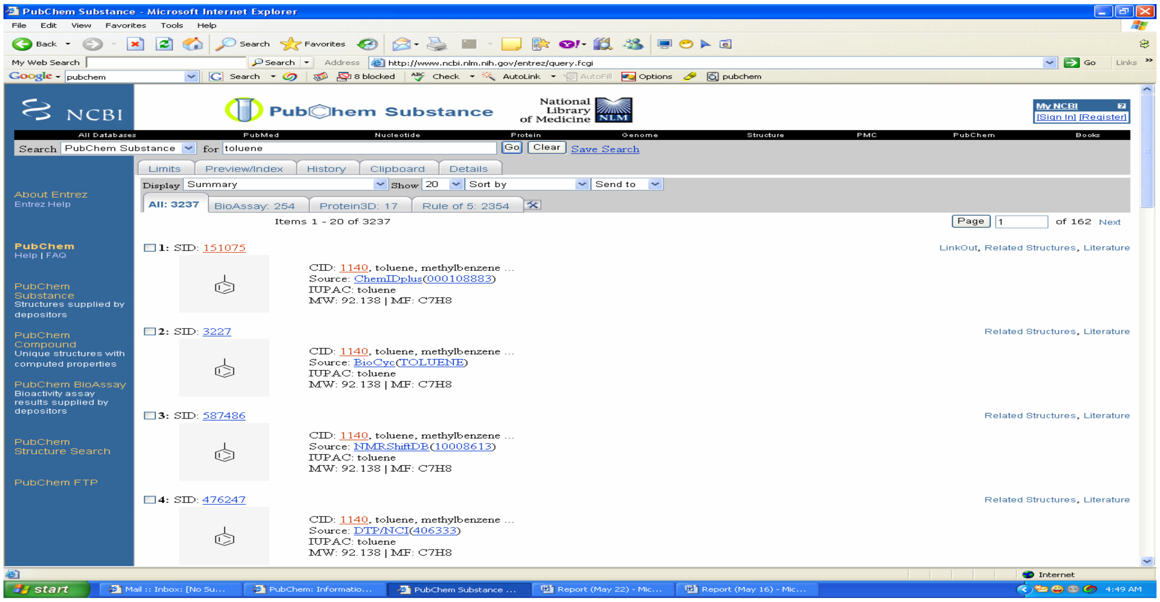
Toluene
A widely used industrial solvent.

Pharmacological Action:
Solvents

Figure 20: PubChem Compound Result Summary Page

This page consists of the information of the particular compound selected. This displays the compound ID for that compound which is unique to that database along with the links to Substances, BioActivity, Protein Structures, Protein Sequences, NLM Toxicology, Related Compounds, Similar Compounds. It also provides information about the Medical Subject Annotations (MeSH), Pharmacological Action and Depositor-Supplied Synonyms. It also displays the computed properties from the structure like Molecular weight, Molecular Formula, XLogP, Hydrogen Bond Donor Count, Rotatable Bond Count and descriptors computed from structurelike IUPAC name, Canonical SMILES, InChI (IUPAC International Chemistry Identifier), etc.

The PubChem substance database contains chemical structures, synonyms, registration IDs, description, database cross-reference links to PubMed, protein 3D structures, and biological screening results. When PubChem Substance is selected and the query “Toluene” is processed, page of records appearsas shown in Figure 21:


PubChem Substance - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
My Web Search Search Address
Go Links
Google Search 8 blocked Check AutoLink AutoFill Options pubchem
NCBI
PubChem Substance
National Library of Medicine NLM
My NCBI
[Sign In] [Register]
All Databases PubMed Nucleotide Protein Genome Structure PMC PubChem Books
Go Clear Save Search
Limits Preview/Index History Clipboard Details
Send to
About Entrez
Entrez Help
PubChem
Help | FAQ
PubChem Substance
Structures supplied by depositors
PubChem Compound
Unique structures with computed properties
PubChem BioAssay
Bioactivity assay results supplied by depositors
PubChem Structure Search
PubChem FTP
LinkOut, Related Structures, Literature
Related Structures, Literature
Related Structures, Literature
Related Structures, Literature
Internet
start
Mail :: Inbox: [No Su...
PubChem: Informatio...
PubChem Substance ...
Report (May 22) - Mic...
Report (May 16) - Mic...
4:49 AM

Address: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

Search PubChem Substance for toluene

Display Summary Show 20 Sort by

All: 3237 BioAssay: 254 Protein3D: 17 Rule of 5: 2354

Items 1 - 20 of 3237

Page 1 of 162 Next

□ 1: SID: 151075
CID: 1140, toluene, methylbenzene ...
Source: ChemIDplus(000108883)
IUPAC: toluene
MW: 92.138 | MF: C7H8

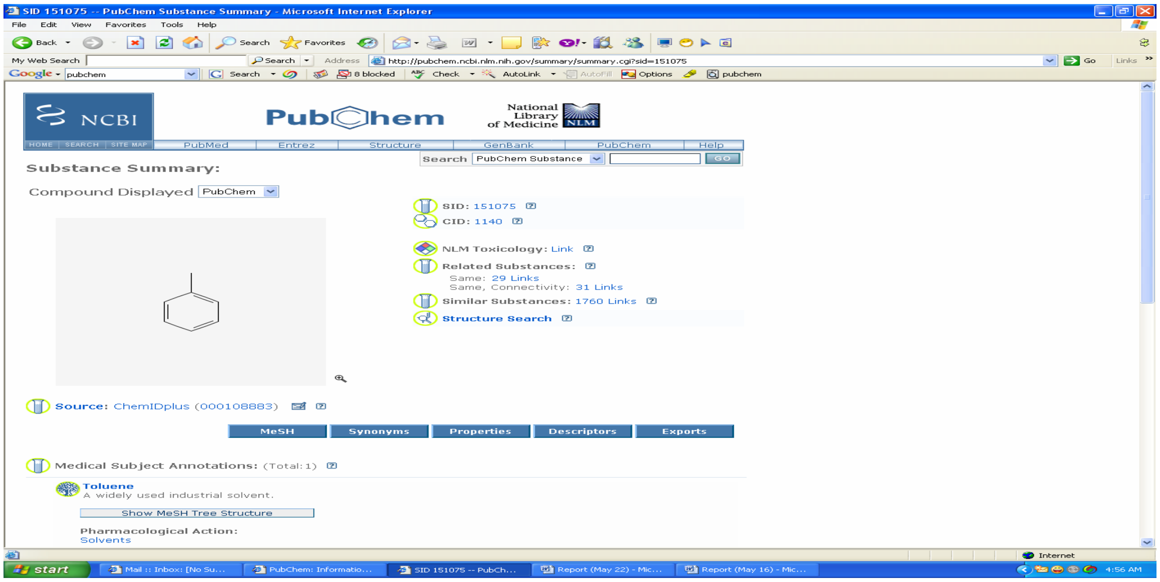
□ 2: SID: 3227
CID: 1140, toluene, methylbenzene ...
Source: BioCyc(TOLUENE)
IUPAC: toluene
MW: 92.138 | MF: C7H8

□ 3: SID: 587486
CID: 1140, toluene, methylbenzene ...
Source: NMRShiftDB(10008613)
IUPAC: toluene
MW: 92.138 | MF: C7H8

□ 4: SID: 476247
CID: 1140, toluene, methylbenzene ...
Source: DTP/NCI(406333)
IUPAC: toluene
MW: 92.138 | MF: C7H8

Figure 21: PubChem Substance Records Result page

When a particular record is selected, the following page of information appears as shown in Figure 22.



SID 151075 -- PubChem Substance Summary - Microsoft Internet Explorer

Address http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?sid=151075

Search PubChem Substance

Substance Summary:

Compound Displayed PubChem

SID: 151075

CID: 1140

NLM Toxicology: Link

Related Substances:
Same: 29 Links
Same, Connectivity: 31 Links

Similar Substances: 1760 Links

Structure Search

Source: ChemIDplus (000108883)

Medical Subject Annotations: (Total: 1)

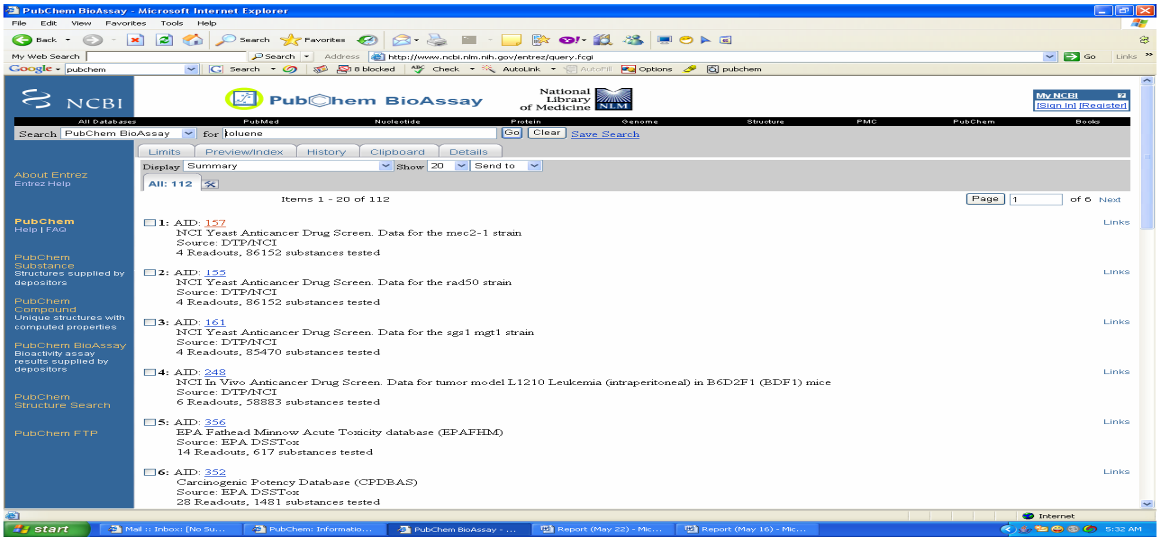
Toluene
A widely used industrial solvent.

Pharmacological Action:
Solvents

Figure 22: PubChem Substance Summary Result Page

The information displayed consists of the substance summary like the substance ID, corresponding compound ID, links to NLM toxicology, related substances, similar substances, etc. It also displays information about the Medical Subject Annotations, Pharmacological Action along with the Depositor-Supplied Synonyms. This displays properties computed from structure like the Molecular Weight, Molecular Formula, XLogP, Hydrogen Bond Donor Count, Hydrogen Bond Acceptor Count, Rotatable Bond Count, etc. along with the descriptors computed from structures like the IUPAC Name, Canonical SMILES, InChI, etc.

The PubChem BioAssay Database contains bioactivity screens of chemical substances described in PubChem Substance. It provides searchable descriptions of each bioassay, including descriptions of the conditions and readouts specific to a screening protocol. When PubChem BioAssay is selected and a query for “Toluene” is processed, the result page summarizing the records is depicted as in figure 23:


PubChem BioAssay - Microsoft Internet Explorer
File Edit View Favorites Tools Help
NCBI
PubChem BioAssay
National Library of Medicine NLM
My NCBI [Sign In] [Register]
All Databases PubMed Nucleotide Protein Genome Structure PMC PubChem Books
Go Clear Save Search
Limits Preview/Index History Clipboard Details
About Entrez
Entrez Help
PubChem
Help | FAQ
PubChem Substance
Structures supplied by depositors
PubChem Compound
Unique structures with computed properties
PubChem BioAssay
Bioactivity assay results supplied by depositors
PubChem Structure Search
PubChem FTP
start
Mail :: Inbox: [No Su...
PubChem: Informatio...
PubChem BioAssay - ...
Report (May 22) - Mic...
Report (May 16) - Mic...
Internet
5:32 AM

Address: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

Search PubChem BioAssay for toluene

Display Summary Show 20 Send to

All: 112

Items 1 - 20 of 112 Page 1 of 6 Next

□ 1: AID: 157 Links
NCI Yeast Anticancer Drug Screen. Data for the mec2-1 strain
Source: DTP/NCI
4 Readouts, 86152 substances tested

□ 2: AID: 155 Links
NCI Yeast Anticancer Drug Screen. Data for the rad50 strain
Source: DTP/NCI
4 Readouts, 86152 substances tested

□ 3: AID: 161 Links
NCI Yeast Anticancer Drug Screen. Data for the sgs1 mgt1 strain
Source: DTP/NCI
4 Readouts, 85470 substances tested

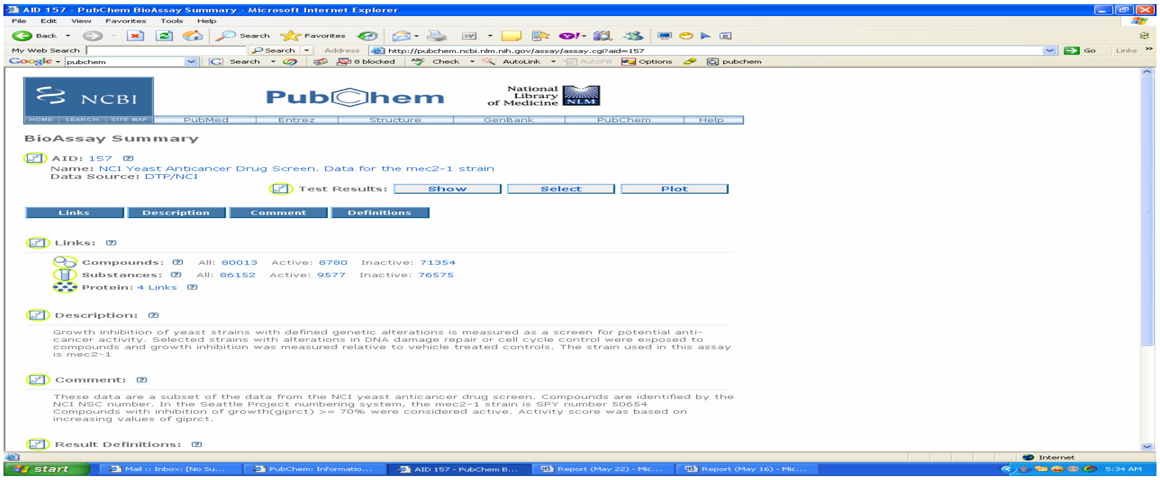
□ 4: AID: 248 Links
NCI In Vivo Anticancer Drug Screen. Data for tumor model L1210 Leukemia (intraperitoneal) in B6D2F1 (BDF1) mice
Source: DTP/NCI
6 Readouts, 58883 substances tested

□ 5: AID: 356 Links
EPA Fathead Minnow Acute Toxicity database (EPAFHM)
Source: EPA DSSTox
14 Readouts, 617 substances tested

□ 6: AID: 352 Links
Carcinogenic Potency Database (CPDBAS)
Source: EPA DSSTox
28 Readouts, 1481 substances tested

Figure 23: PubChem BioAssay Records Result page

When a record is selected the following information of that particular record is obtained as in figure 24.


AID 157 - PubChem BioAssay Summary - Microsoft Internet Explorer
File Edit View Favorites Tools Help
NCBI
HOME SEARCH SITE MAP
PubChem
National Library of Medicine NLM
PubMed Entrez Structure GenBank PubChem Help
Show Select Plot
Links Description Comment Definitions
start
Mail :: Inbox: [No Su...
PubChem: Informatio...
AID 157 - PubChem B...
Report (May 22) - Mic...
Report (May 16) - Mic...
Internet
5:34 AM

Address: http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=157

BioAssay Summary

AID: 157
Name: NCI Yeast Anticancer Drug Screen. Data for the mec2-1 strain
Data Source: DTP/NCI

Test Results:

Links:
Compounds: All: 80013 Active: 8780 Inactive: 71354
Substances: All: 86152 Active: 9577 Inactive: 76575
Protein: 4 Links

Description:
Growth inhibition of yeast strains with defined genetic alterations is measured as a screen for potential anti-cancer activity. Selected strains with alterations in DNA damage repair or cell cycle control were exposed to compounds and growth inhibition was measured relative to vehicle treated controls. The strain used in this assay is mec2-1

Comment:
These data are a subset of the data from the NCI yeast anticancer drug screen. Compounds are identified by the NCI NSC number. In the Seattle Project numbering system, the mec2-1 strain is SPY number 50654 Compounds with inhibition of growth(giprct) >= 70% were considered active. Activity score was based on increasing values of giprct.

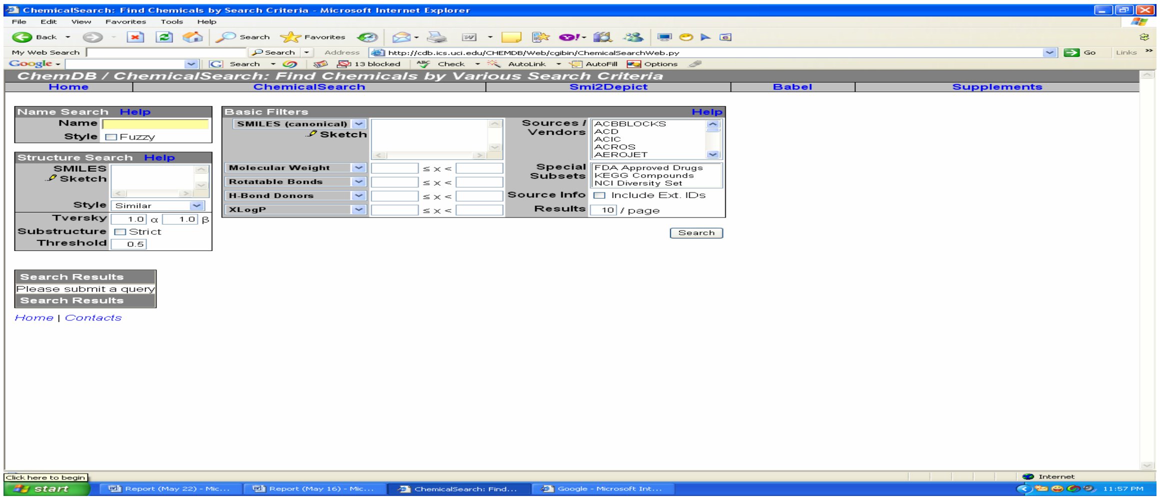
Result Definitions:

Figure:24: PubChem BioAssay Summary Result Page

The information consists of the Bio-Assay ID, Name, Data Source and links to the Compounds, Substances, and Protein along with the description of the data, comment on the data, and the result definitions.

CHEMDB: The UC Irvine ChemDB (http://cdb.ics.uci.edu/CHEMDB/Web/)

ChemDB is a public database of small molecules and related chemoinformatics resources. ChemDB is built using the digital catalogs of over hundred vendors and other public sources and is annotated with information derived from these sources as well as from computational methods, such as predicted solubility and 3D structure. The database contains approximately 4.1 M commercially available compounds, 8.2 M counting isomers. The page where the query is entered is depicted in figure 25.


ChemicalSearch: Find Chemicals by Search Criteria - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
My Web Search Search Address
Google Search 13 blocked Check AutoLink AutoFill Options
Home ChemicalSearch Smi2Depict Babel Supplements
Click here to begin
start Report (May 22) - Mic... Report (May 16) - Mic... ChemicalSearch: Find... Google - Microsoft Int...
Internet
11:57 PM

http://cdb.ics.uci.edu/CHEMDB/Web/cgibin/ChemicalSearchWeb.py

ChemDB / ChemicalSearch: Find Chemicals by Various Search Criteria

Name Search Help
Name
Style □ Fuzzy

Structure Search Help
SMILES Sketch
Style Similar
Tversky 1.0 α 1.0 β
Substructure □ Strict
Threshold 0.5

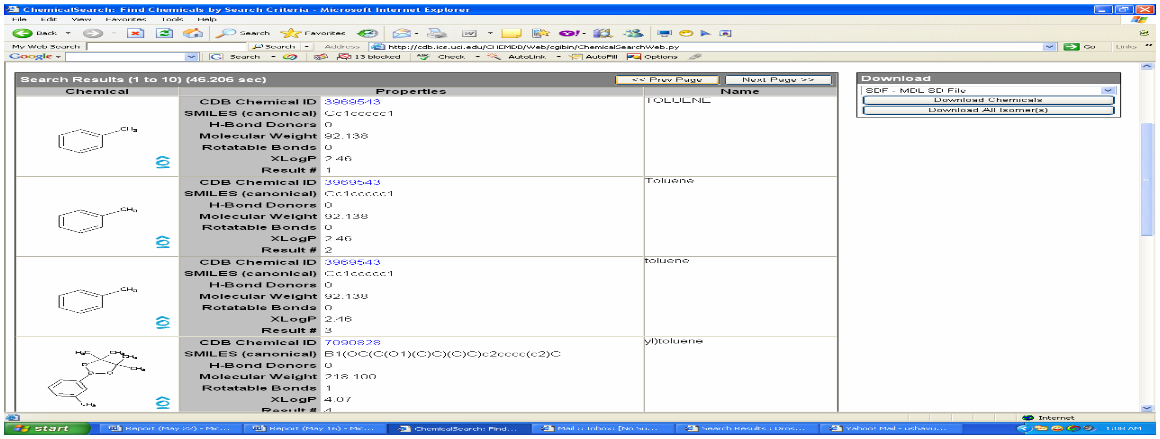
Basic Filters Help
SMILES (canonical) Sketch
Molecular Weight ≤ x <
Rotatable Bonds ≤ x <
H-Bond Donors ≤ x <
XLogP ≤ x <
Sources / Vendors: ACBBLOCKS, ACD, ACIC, ACROS, AEROJET
Special Subsets: FDA Approved Drugs, KEGG Compounds, NCI Diversity Set
Source Info □ Include Ext. IDs
Results 10 / page
Search

Search Results
Please submit a query
Search Results

Home | Contacts

Figure 25: ChemDB query input page

When searched for “toluene” the result page is depicted as in figure 26:


ChemicalSearch: Find Chemicals by Search Criteria - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
My Web Search Search Address
Google Search 13 blocked Check AutoLink AutoFill Options
CH3
H3C
CH3
O
B
start Report (May 22) - Mic... Report (May 16) - Mic... ChemicalSearch: Find... Mail :: Inbox: [No Su... Search Results : Dros... Yahoo! Mail - ushavu...
Internet
1:08 AM

http://cdb.ics.uci.edu/CHEMDB/Web/cgibin/ChemicalSearchWeb.py

Search Results (1 to 10) (46.206 sec) << Prev Page Next Page >>

Chemical	Properties		Name
	CDB Chemical ID	3969543	TOLUENE
	SMILES (canonical)	Cc1ccccc1	
	H-Bond Donors	0	
	Molecular Weight	92.138	
	Rotatable Bonds	0	
	XLogP	2.46	
	Result #	1	
	CDB Chemical ID	3969543	Toluene
	SMILES (canonical)	Cc1ccccc1	
	H-Bond Donors	0	
	Molecular Weight	92.138	
	Rotatable Bonds	0	
	XLogP	2.46	
	Result #	2	
	CDB Chemical ID	3969543	toluene
	SMILES (canonical)	Cc1ccccc1	
	H-Bond Donors	0	
	Molecular Weight	92.138	
	Rotatable Bonds	0	
	XLogP	2.46	
	Result #	3	
	CDB Chemical ID	7090828	yl)toluene
	SMILES (canonical)	B1(OC(C(O1)(C)C)(C)C)c2cccc(c2)C	
	H-Bond Donors	0	
	Molecular Weight	218.100	
	Rotatable Bonds	1	
	XLogP	4.07	
	Result #	4	

Download
SDF - MDL SD File
Download Chemicals
Download All Isomer(s)

Figure 26: ChemDB Record Result page

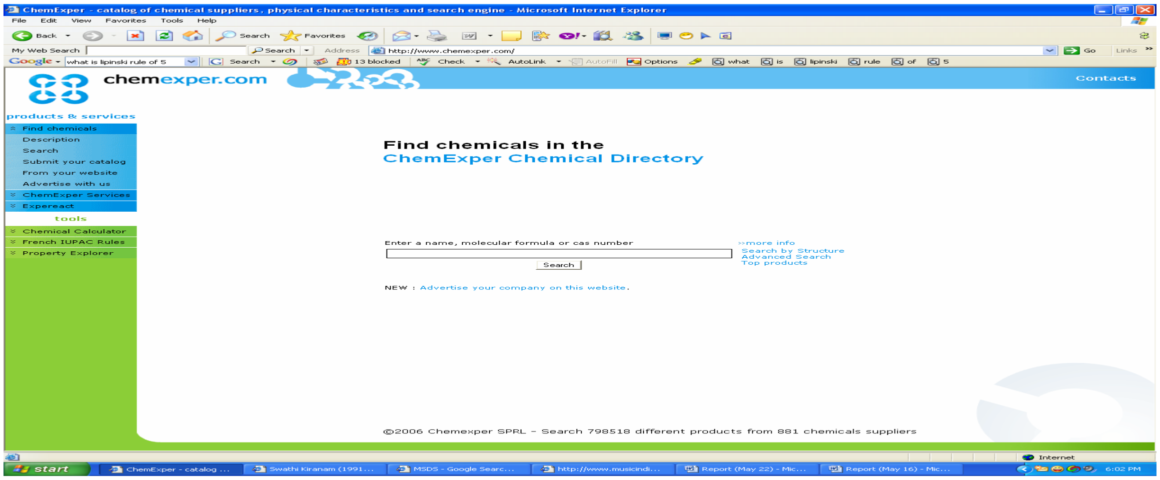
This gives information of all possible structures of toluene and also other structures which contain toluene as part of them. For each structure displayed, it provides information about the CDB Chemical ID, SMILES, H-Bond Donors, Molecular weight, Rotatable bonds, XLogP , etc.

CHEMEXPER (http://www.chemexper.com)

This is a freely accessible database of chemicals over the internet. This database contains chemicals with their physical characteristics. Everybody can submit chemical information and retrieve information with a web browser.

The ChemExper Chemical Directory displays information about chemicals (physical and chemical characteristics, structure, MSDS and more.). This directory contains over 200,000 different chemicals, MSDS and over 10,000 IR spectra. The directory can be searched by registry number, molecular formula, chemical name or synonyms as well as by physical and chemical characteristics and combinations of those data. The ChemExper Chemical Directory may also be searched by substructure.

Figure 27 represents the home page of ChemExper Chemical Directory



Find chemicals in the
ChemExper Chemical Directory

Enter a name, molecular formula or cas number

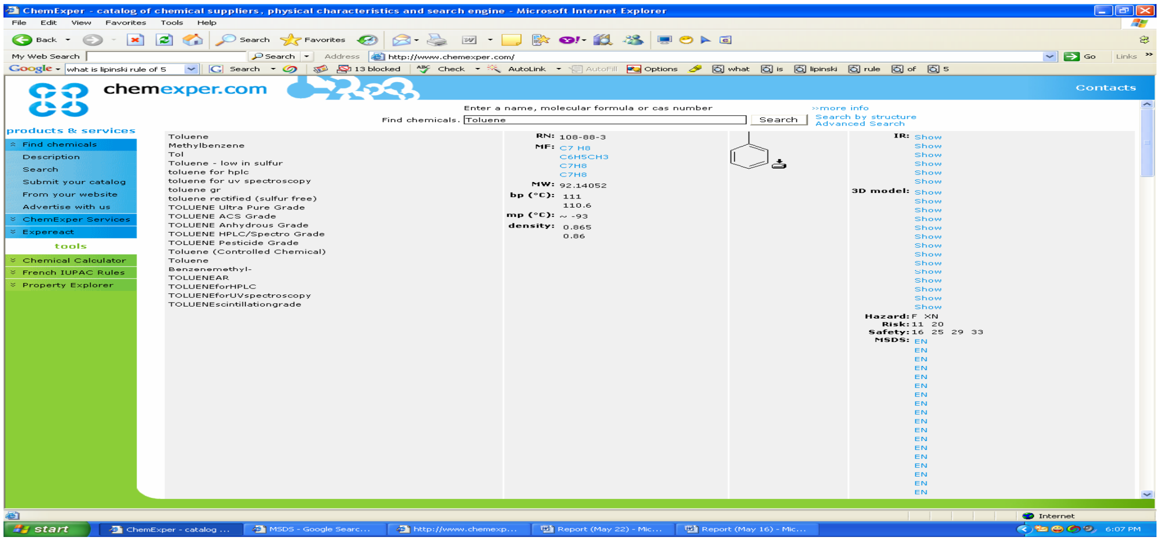
NEW : Advertise your company on this website.

©2006 Chemexper SPRL - Search 798518 different products from 881 chemicals suppliers

Figure 27: ChemExper query input Page

This page is the fastest and the easiest way to search in the database. The user needs to enter a value/keyword and click on the search button. The program will then try to find corresponding products by looking for molecular formula, registry number (CAS number), product name and synonyms as well as a supplier catalog number.

Figure 28 represents the result page when searched for “Toluene”.



Enter a name, molecular formula or cas number

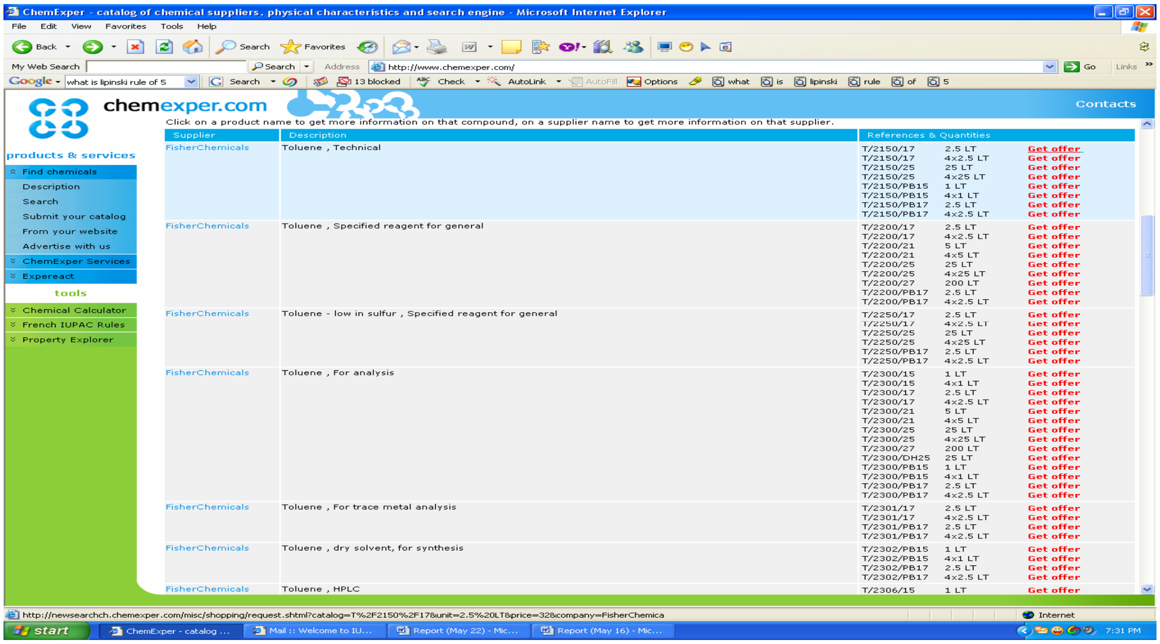
Find chemicals. Toluene Search

>>more info
Search by structure
Advanced Search

Toluene
Methylbenzene
Tol
Toluene - low in sulfur
toluene for hplc
toluene for uv spectroscopy
toluene gr
toluene rectified (sulfur free)
TOLUENE Ultra Pure Grade
TOLUENE ACS Grade
TOLUENE Anhydrous Grade
TOLUENE HPLC/Spectro Grade
TOLUENE Pesticide Grade
Toluene (Controlled Chemical)
Toluene
Benzenemethyl-
TOLUENEAR
TOLUENEforHPLC
TOLUENEforUVspectroscopy
TOLUENEscintillationgrade

RN: 108-88-3
MF: C7 H8
C6H5CH3
C7H8
C7H8
MW: 92.14052
bp (°C): 111
110.6
mp (°C): ~ -93
density: 0.865
0.86

IR: Show
Show
Show
Show
Show
Show
3D model: Show
Show
Show
Show
Show
Show
Show
Show
Show
Show
Show
Show
Show
Show
Hazard: F XN
Risk: 11 20
Safety: 16 25 29 33
MSDS: EN
EN
EN
EN
EN
EN
EN
EN
EN
EN
EN
EN
EN
EN
EN
EN
EN
EN



Click on a product name to get more information on that compound, on a supplier name to get more information on that supplier.

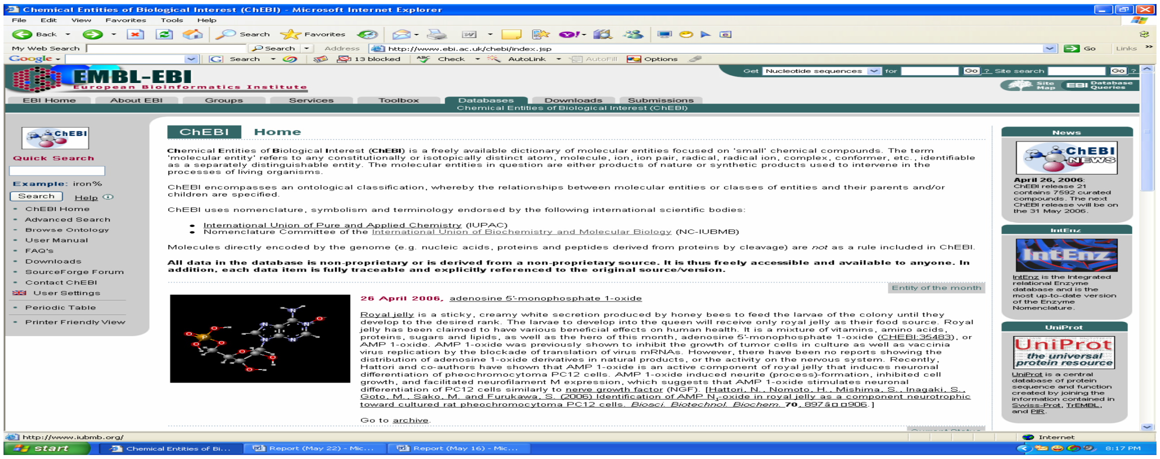
Supplier	Description	References & Quantities		
FisherChemicals	Toluene , Technical	T/2150/17	2.5 LT	Get offer
		T/2150/17	4x2.5 LT	Get offer
		T/2150/25	25 LT	Get offer
		T/2150/25	4x25 LT	Get offer
		T/2150/PB15	1 LT	Get offer
		T/2150/PB15	4x1 LT	Get offer
		T/2150/PB17	2.5 LT	Get offer
		T/2150/PB17	4x2.5 LT	Get offer
FisherChemicals	Toluene , Specified reagent for general	T/2200/17	2.5 LT	Get offer
		T/2200/17	4x2.5 LT	Get offer
		T/2200/21	5 LT	Get offer
		T/2200/21	4x5 LT	Get offer
		T/2200/25	25 LT	Get offer
		T/2200/25	4x25 LT	Get offer
		T/2200/27	200 LT	Get offer
		T/2200/PB17	2.5 LT	Get offer
		T/2200/PB17	4x2.5 LT	Get offer
FisherChemicals	Toluene - low in sulfur , Specified reagent for general	T/2250/17	2.5 LT	Get offer
		T/2250/17	4x2.5 LT	Get offer
		T/2250/25	25 LT	Get offer
		T/2250/25	4x25 LT	Get offer
		T/2250/PB17	2.5 LT	Get offer
		T/2250/PB17	4x2.5 LT	Get offer
FisherChemicals	Toluene , For analysis	T/2300/15	1 LT	Get offer
		T/2300/15	4x1 LT	Get offer
		T/2300/17	2.5 LT	Get offer
		T/2300/17	4x2.5 LT	Get offer
		T/2300/21	5 LT	Get offer
		T/2300/21	4x5 LT	Get offer
		T/2300/25	25 LT	Get offer
		T/2300/25	4x25 LT	Get offer
		T/2300/27	200 LT	Get offer
		T/2300/DH25	25 LT	Get offer
		T/2300/PB15	1 LT	Get offer
		T/2300/PB15	4x1 LT	Get offer
		T/2300/PB17	2.5 LT	Get offer
		T/2300/PB17	4x2.5 LT	Get offer
FisherChemicals	Toluene , For trace metal analysis	T/2301/17	2.5 LT	Get offer
		T/2301/17	4x2.5 LT	Get offer
		T/2301/PB17	2.5 LT	Get offer
		T/2301/PB17	4x2.5 LT	Get offer
FisherChemicals	Toluene , dry solvent, for synthesis	T/2302/PB15	1 LT	Get offer
		T/2302/PB15	4x1 LT	Get offer
		T/2302/PB17	2.5 LT	Get offer
		T/2302/PB17	4x2.5 LT	Get offer
FisherChemicals	Toluene , HPLC	T/2306/15	1 LT	Get offer

http://newsearchch.chemexper.com/misc/shopping/request.shtml?catalog=T%2F2150%2F17&unit=2.5%20LT&price=32&company=FisherChemica

Figure 28: ChemExper Records Result page

ChEBI, Chemical Entities Of Biological Interest (EMBL-EBI, European Bioinformatics Institute) (http://www.ebi.ac.uk/chebi/index.jsp)

Chemical Entities of Biological Interest is a freely available dictionary of molecular entities focused on small chemical compounds. The term 'molecular entity' refers to any constitutionally or isotopically distinct atom, molecule, ion, ion pair, radical, radical ion, complex, conformer, etc., identifiable as a separately distinguishable entity. The home page of this database where the query is entered is shown in figure 29.


Chemical Entities of Biological Interest (ChEBI) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.ebi.ac.uk/chebi/index.jsp
Go Links
13 blocked Check AutoLink AutoFill Options
EMBL-EBI
European Bioinformatics Institute
Get Nucleotide sequences for Go ? Site search Go ?
Site Map EBI Database Queries
EBI Home About EBI Groups Services Toolbox Databases Downloads Submissions
Chemical Entities of Biological Interest (ChEBI)
ChEBI
Quick Search
Example: iron%
Search Help
ChEBI Home
Advanced Search
Browse Ontology
User Manual
FAQ's
Downloads
SourceForge Forum
Contact ChEBI
User Settings
Periodic Table
Printer Friendly View
Entity of the month
News
IntEnz
UniProt
UniProt the universal protein resource
http://www.iubmb.org/
Internet
start
Chemical Entities of Bi...
Report (May 22) - Mic...
Report (May 16) - Mic...
8:17 PM

ChEBI Home

Chemical Entities of Biological Interest (ChEBI) is a freely available dictionary of molecular entities focused on 'small' chemical compounds. The term 'molecular entity' refers to any constitutionally or isotopically distinct atom, molecule, ion, ion pair, radical, radical ion, complex, conformer, etc., identifiable as a separately distinguishable entity. The molecular entities in question are either products of nature or synthetic products used to intervene in the processes of living organisms.

ChEBI encompasses an ontological classification, whereby the relationships between molecular entities or classes of entities and their parents and/or children are specified.

ChEBI uses nomenclature, symbolism and terminology endorsed by the following international scientific bodies:

- International Union of Pure and Applied Chemistry (IUPAC)
- Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB)

Molecules directly encoded by the genome (e.g. nucleic acids, proteins and peptides derived from proteins by cleavage) are *not* as a rule included in ChEBI.

All data in the database is non-proprietary or is derived from a non-proprietary source. It is thus freely accessible and available to anyone. In addition, each data item is fully traceable and explicitly referenced to the original source/version.

26 April 2006, adenosine 5'-monophosphate 1-oxide

Royal jelly is a sticky, creamy white secretion produced by honey bees to feed the larvae of the colony until they develop to the desired rank. The larvae to develop into the queen will receive only royal jelly as their food source. Royal jelly has been claimed to have various beneficial effects on human health. It is a mixture of vitamins, amino acids, proteins, sugars and lipids, as well as the hero of this month, adenosine 5'-monophosphate 1-oxide (CHEBI:35483), or AMP 1-oxide. AMP 1-oxide was previously shown to inhibit the growth of tumor cells in culture as well as vaccinia virus replication by the blockade of translation of virus mRNAs. However, there have been no reports showing the distribution of adenosine 1-oxide derivatives in natural products, or the activity on the nervous system. Recently, Hattori and co-authors have shown that AMP 1-oxide is an active component of royal jelly that induces neuronal differentiation of pheochromocytoma PC12 cells. AMP 1-oxide induced neurite (process)-formation, inhibited cell growth, and facilitated neurofilament M expression, which suggests that AMP 1-oxide stimulates neuronal differentiation of PC12 cells similarly to nerve growth factor (NGF). [Hattori, N., Nomoto, H., Mishima, S., Inagaki, S., Goto, M., Sako, M. and Furukawa, S. (2006) Identification of AMP N_1-oxide in royal jelly as a component neurotrophic toward cultured rat pheochromocytoma PC12 cells. *Biosci. Biotechnol. Biochem.* **70**, 897â□□906.]

Go to archive.

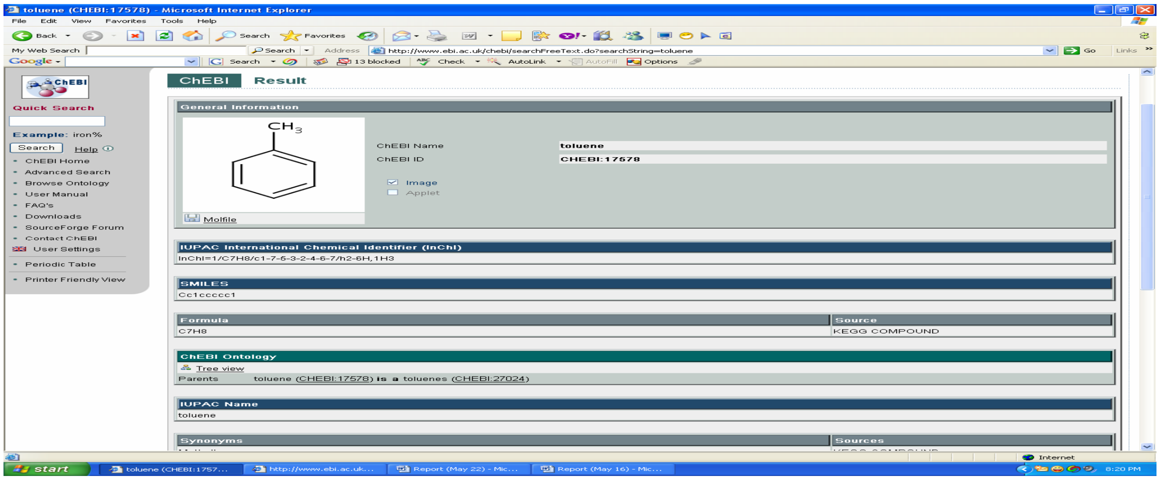
April 26, 2006: ChEBI release 21 contains 7592 curated compounds. The next ChEBI release will be on the 31 May 2006.

IntEnz is the Integrated relational Enzyme database and is the most up-to-date version of the Enzyme Nomenclature.

UniProt is a central database of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR.

Figure 29: ChEBI query input page

When the query "toluene" is processed, the result page is depicted as in figure 30.



ChEBI Result

General Information

ChEBI Name toluene

ChEBI ID CHEBI:17578

☑ Image
☐ Applet

IUPAC International Chemical Identifier (InChI)

InChI=1/C7H8/c1-7-5-3-2-4-6-7/h2-6H,1H3

SMILES

Cc1ccccc1

Formula	Source
C7H8	KEGG COMPOUND

ChEBI Ontology

Parents toluene (CHEBI:17578) **is a** toluenes (CHEBI:27024)

IUPAC Name

toluene

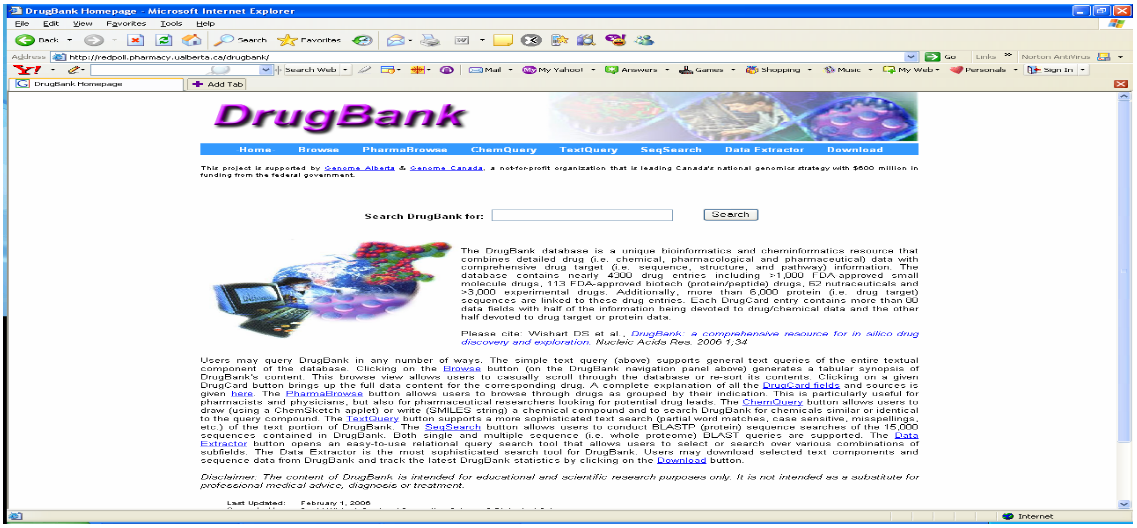
Synonyms	Sources

Figure 30: ChEBI Record Result page

The result page provides general information of the chemical entered like its Structure, ChEBI Name, ChEBI ID, IUPAC International Chemical Identifier (InChI), SMILES, Formula, IUPAC Name, Synonyms, Database Links, Registry Numbers and the Linked Uniprot proteins associated with this compound.

DRUGBANK (http://redpoll.pharmacy.ualberta.ca/drugbank/)

The DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information. The database contains nearly 4300 drug entries including >1,000 FDA-approved small molecule drugs, 113 FDA-approved biotech (protein/peptide) drugs, 62 nutraceuticals and >3,000 experimental drugs. Additionally, more than 6,000 protein (i.e. drug target) sequences are linked to these drug entries. Each DrugCard entry contains more than 80 data fields with half of the information being devoted to drug/chemical data and the other half devoted to drug target or protein data. The homepage is represented in figure 31.


DrugBank Homepage - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://redpoll.pharmacy.ualberta.ca/drugbank/
Go Links Norton AntiVirus
Search Web Mail My Yahoo! Answers Games Shopping Music My Web Personals Sign In
DrugBank Homepage
Add Tab
DrugBank
-Home- Browse PharmaBrowse ChemQuery TextQuery SeqSearch Data Extractor Download
Internet

This project is supported by Genome Alberta & Genome Canada, a not-for-profit organization that is leading Canada's national genomics strategy with $600 million in funding from the federal government.

Search DrugBank for: Search

The DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information. The database contains nearly 4300 drug entries including >1,000 FDA-approved small molecule drugs, 113 FDA-approved biotech (protein/peptide) drugs, 62 nutraceuticals and >3,000 experimental drugs. Additionally, more than 6,000 protein (i.e. drug target) sequences are linked to these drug entries. Each DrugCard entry contains more than 80 data fields with half of the information being devoted to drug/chemical data and the other half devoted to drug target or protein data.

Please cite: Wishart DS et al., *DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006 1;34*

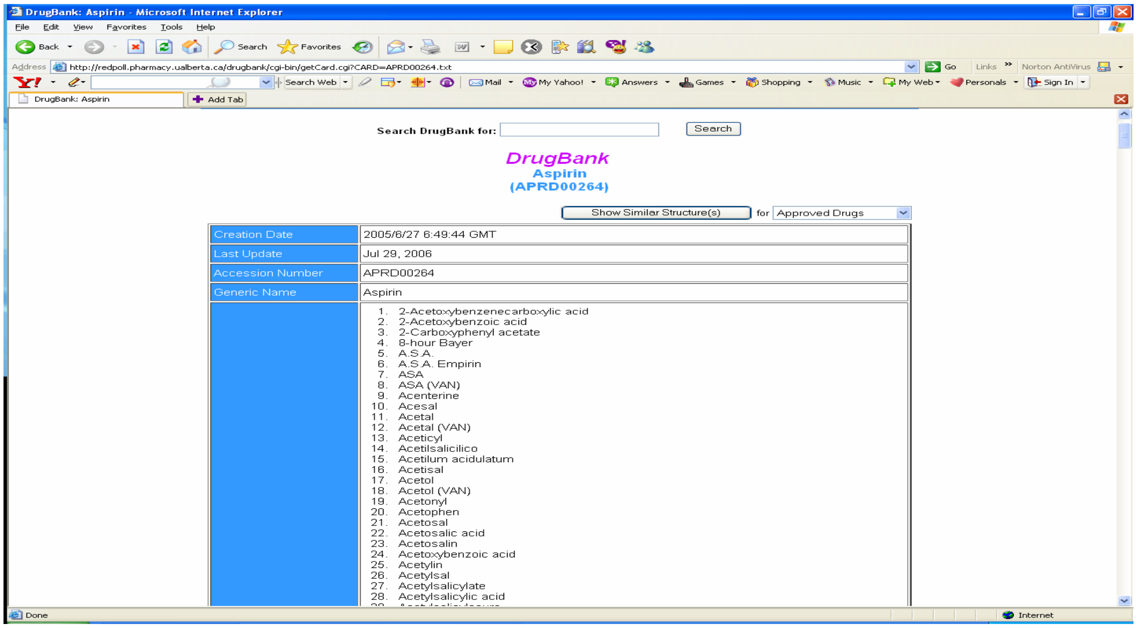
Users may query DrugBank in any number of ways. The simple text query (above) supports general text queries of the entire textual component of the database. Clicking on the Browse button (on the DrugBank navigation panel above) generates a tabular synopsis of DrugBank's content. This browse view allows users to casually scroll through the database or re-sort its contents. Clicking on a given DrugCard button brings up the full data content for the corresponding drug. A complete explanation of all the DrugCard fields and sources is given here. The PharmaBrowse button allows users to browse through drugs as grouped by their indication. This is particularly useful for pharmacists and physicians, but also for pharmaceutical researchers looking for potential drug leads. The ChemQuery button allows users to draw (using a ChemSketch applet) or write (SMILES string) a chemical compound and to search DrugBank for chemicals similar or identical to the query compound. The TextQuery button supports a more sophisticated text search (partial word matches, case sensitive, misspellings, etc.) of the text portion of DrugBank. The SeqSearch button allows users to conduct BLASTP (protein) sequence searches of the 15,000 sequences contained in DrugBank. Both single and multiple sequence (i.e. whole proteome) BLAST queries are supported. The Data Extractor button opens an easy-to-use relational query search tool that allows users to select or search over various combinations of subfields. The Data Extractor is the most sophisticated search tool for DrugBank. Users may download selected text components and sequence data from DrugBank and track the latest DrugBank statistics by clicking on the Download button.

Disclaimer: The content of DrugBank is intended for educational and scientific research purposes only. It is not intended as a substitute for professional medical advice, diagnosis or treatment.

Last Updated: February 1, 2006

Figure 31: DrugBank Query Input page

The result page when a query is processed is shown as in figure 32



DrugBank
Aspirin
(APRD00264)

Creation Date	2005/6/27 6:49:44 GMT
Last Update	Jul 29, 2006
Accession Number	APRD00264
Generic Name	Aspirin
	1. 2-Acetoxybenzenecarboxylic acid 2. 2-Acetoxybenzoic acid 3. 2-Carboxyphenyl acetate 4. 8-hour Bayer 5. A.S.A. 6. A.S.A. Empirin 7. ASA 8. ASA (VAN) 9. Acenterine 10. Acesal 11. Acetal 12. Acetal (VAN) 13. Aceticyl 14. Acetilsalicilico 15. Acetilum acidulatum 16. Acetisal 17. Acetol 18. Acetol (VAN) 19. Acetonyl 20. Acetophen 21. Acetosal 22. Acetosalic acid 23. Acetosalin 24. Acetoxybenzoic acid 25. Acetylin 26. Acetylsal 27. Acetylsalicylate 28. Acetylsalicylic acid

Figure 32: DrugBank Record Result page

The information displayed consists of Creation date of the page, Last update of the information, Accession Number, Generic name, Synonyms, Brand name mixtures, Chemical structure, CAS Registry number, InChi Identifier, KEGG Compound ID, PubChem ID, ChEBI ID, HET ID, Molecular weight, LogP, pKa, NMR Spectrum, etc.

Summary of Databases Descriptors:

The summary of all the descriptors displayed in these databases is shown in the follow figure 33. With such huge data in these databases and the diversity of the descriptors, one would certainly appreciate the availability of common interface which would bring all these databases onto one roof. Chemical facets would help in achieving such goal.

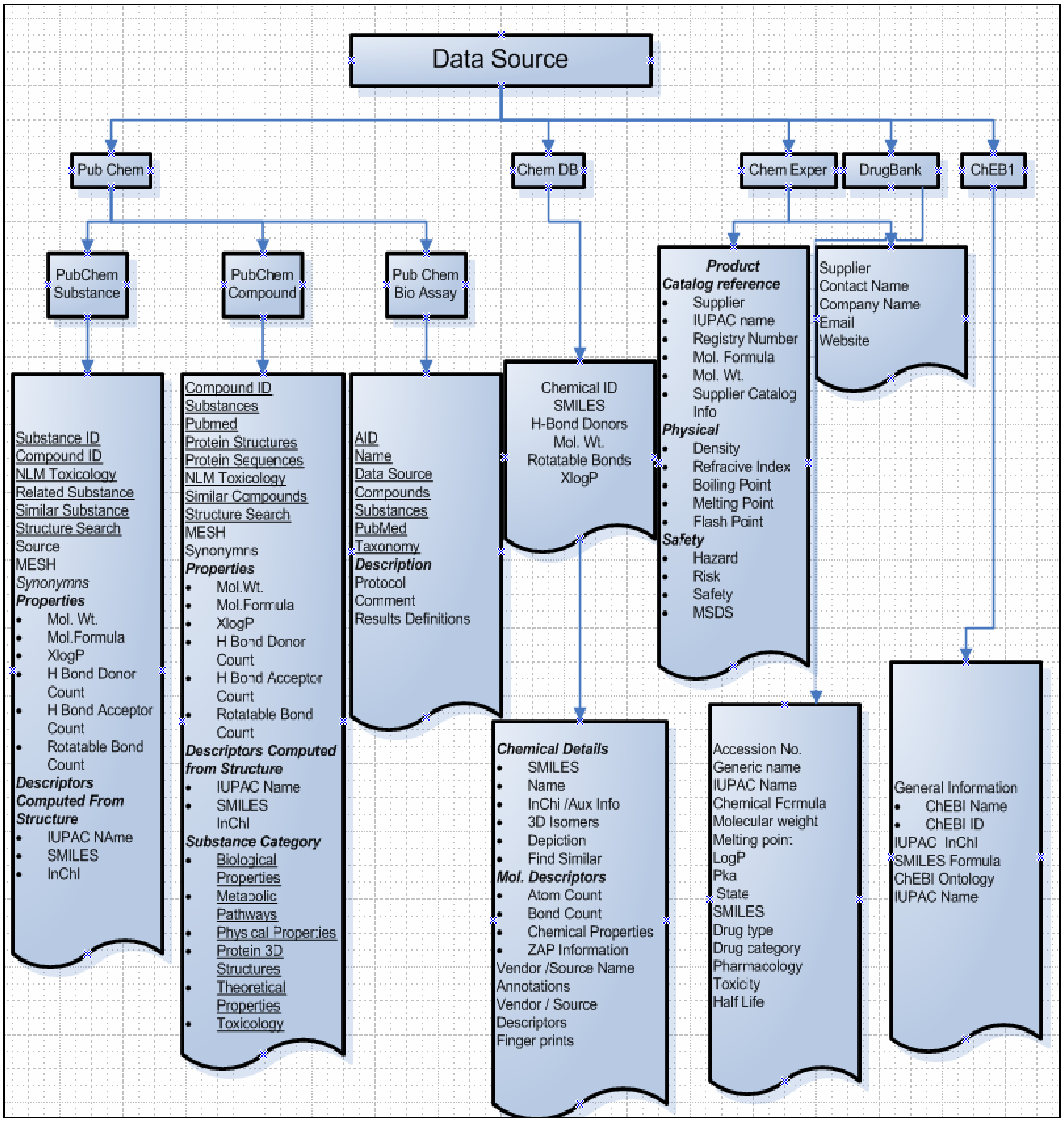


Figure 33: Characteristics of databases

4.4.3.2 Chemical Facets Design

The databases that were started with for the design are PubChem Substance, PubChem Compound, PubChem Bioassay, ChemExper, ChemDB. The main facet that was considered was the "*data type*". This facet indicates the type of data described by each database. For example compound data, substance data, bioassay data, or the drug data, etc. In the process of writing the XML from these databases, it is being studied that

ChemDB and ChemExper databases do not support the query building URL that is being used by the code. It is being studied that they are not an http supported sites. For example querying NCBI PubChem substance, the query URL is http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pcsubstance&cmd=search&term=methotrexate. This query url is constructed by joining the base url "http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pcsubstance&cmd=search&term=" and the keyword "methotrexate". Such type of query could not be processed in databases: ChemExper, ChemDB. The facet based integration system only supports HTTP based databases and the above databases are not HTTP based. Then the databases that were considered for the study are ChEBI and DrugBank. These two databases allowed the feasibility of building the query url.

The second facet that was taken into account was the "*data source*". This facet represented the source of the database itself.

The third facet that was considered is the '*property*' facet. This is a grouped facet. Molecular weight values, xlogP values, hydrogen bond donor count values, hydrogen bond acceptor count values are grouped under this facet. Using these facets, the records that have a particular range of molecular weights or xlogP values are grouped together and the results are displayed. This represents a single layered hierarchical facet.

The literature facet is being added from the Pubmed database. The XML for this database is already being written for biological databases and is directly applied the chemical facets.

Two types of database extraction rules are written: Summary extraction rules and extended extraction rules. In the summary extraction rules, the result set page after a query is being processed is extracted for each database. In the extended extraction rules, the required fields are being extracted for each record. These are used to develop the facets. Figure 34 summarizes the design of the facets.

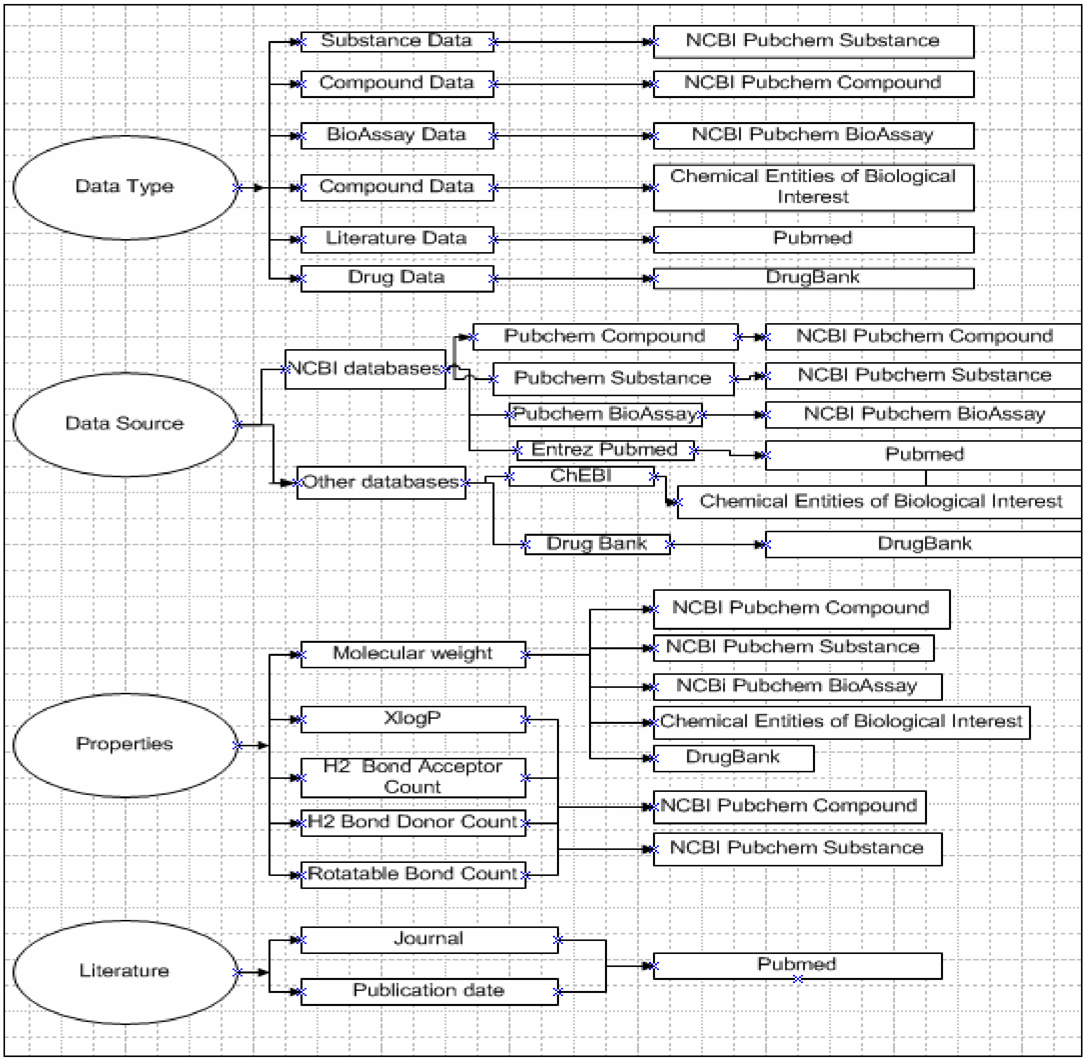


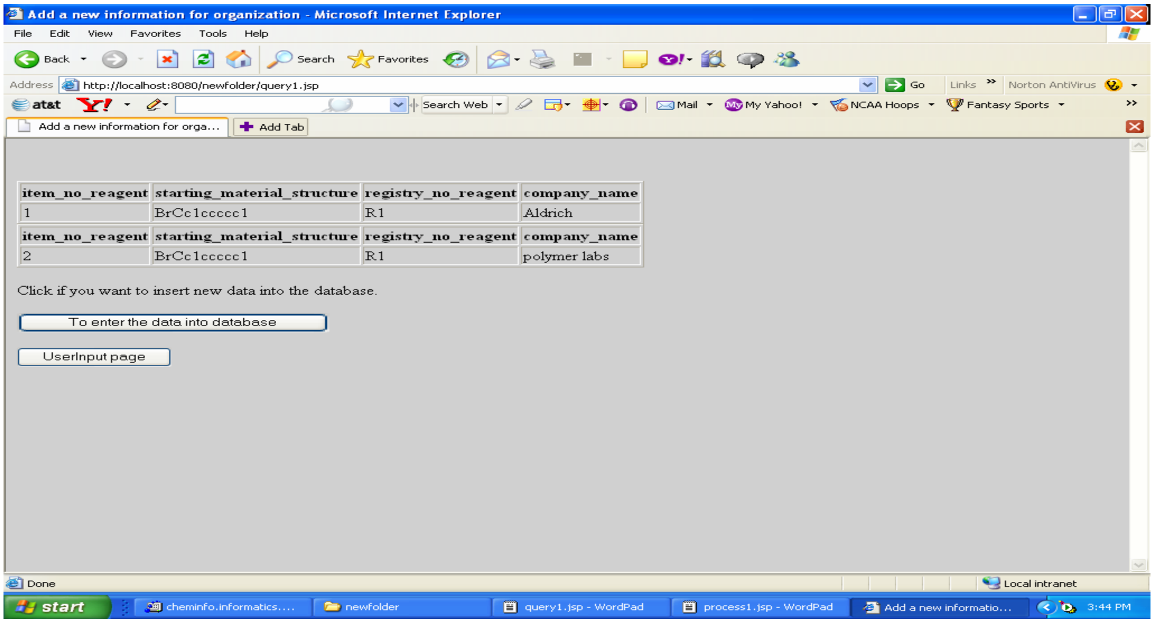
Figure 34: Chemical Facets Design

5 RESULTS

Distributed Drug Discovery

The distributed Drug Discovery database is being developed mainly for undergraduate chemistry laboratories. This is to store and retrieve the data associated with their experimental results in the laboratory.

The main web application enables the administrator to enter the information of the reagents used for the experiment. In order to avoid duplication of reagents, the administrator was given accessibility to check the information of the reagents already entered into the database along with the source where the particular reagent was purchased. Figure 35 illustrates the information retrieved from the database about a reagent.



item_no_reagent	starting_material_structure	registry_no_reagent	company_name
1	BrCc1ccccc1	R1	Aldrich
item_no_reagent	starting_material_structure	registry_no_reagent	company_name
2	BrCc1ccccc1	R1	polymer labs

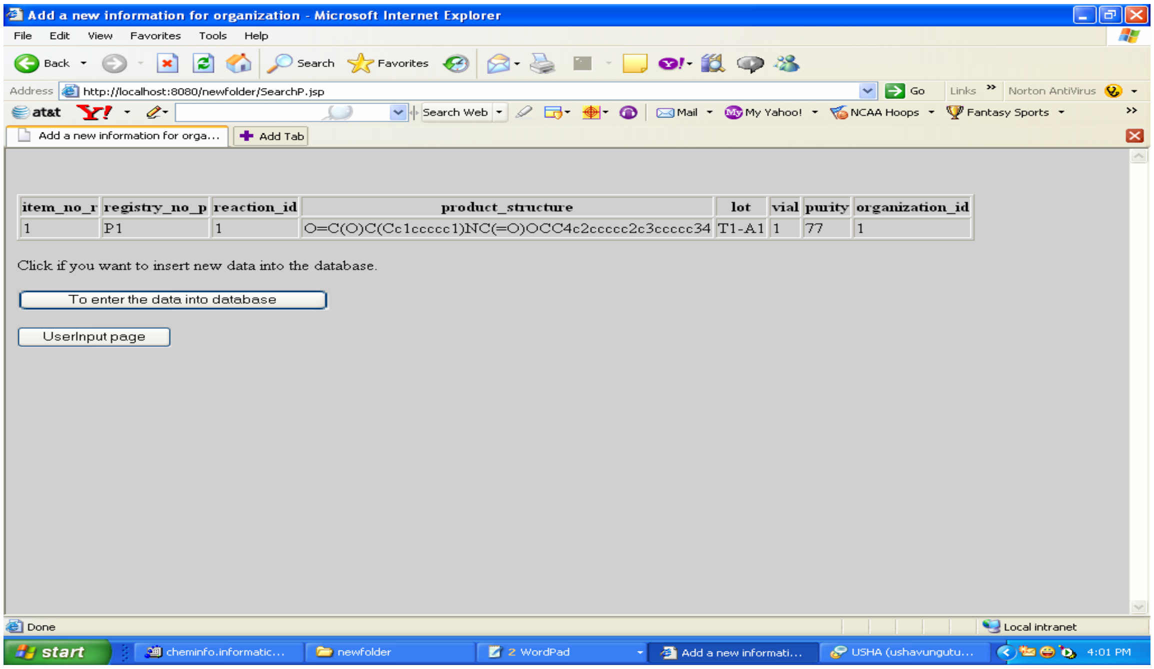
Click if you want to insert new data into the database.

Figure 35: Reagent Retrieval Information

The information that was stored in the database along with the reagents name are its registry number that is unique to the database, molecular formula, molecular weight, the

PubChem number, chemical abstract services number, the source from where the reagent was purchased, the SMILES format and its role in the experiment..

Similarly the administrator can also verify the Product information already present in the database and avoid synthesis of similar products by the students. Figure 36 illustrates the data of products retrieved from the database.


Add a new information for organization - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://localhost:8080/newfolder/SearchP.jsp
Go Links Norton AntiVirus
at&t Search Web Mail My Yahoo! NCAA Hoops Fantasy Sports
Add a new information for orga... Add Tab
Done Local intranet
start cheminfo.informatic... newfolder 2 WordPad Add a new informati... USHA (ushavungutu... 4:01 PM

item_no_r	registry_no_p	reaction_id	product_structure	lot	vial	purity	organization_id
1	P1	1	O=C(O)C(Cc1ccccc1)NC(=O)OCC4c2ccccc2c3ccccc34	T1-A1	1	77	1

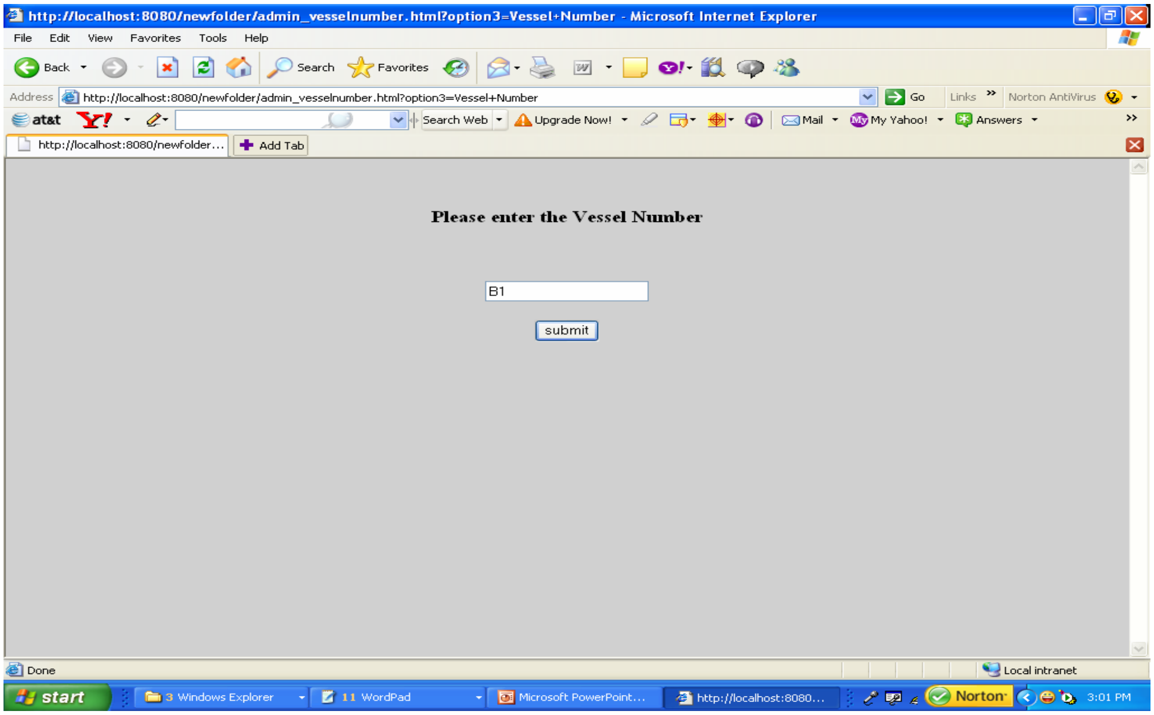
Click if you want to insert new data into the database.

To enter the data into database

UserInput page

Figure 36: Product Retrieval Information

The student interface developed aids in entering the experimental information by the students. The Administrator interface-2 developed is used in validating the results of the student's input. Several searching mechanisms were incorporated to make it easier for the administrator to check the results. The administrator can check the results based on the structure of the product or the student last names, or the vessel number used in the reaction or the team number. This can be utilized for grading the students work. The results obtained when the database is queried for student input based on the vessel number are illustrated in the Figure-37 and Figure-38.



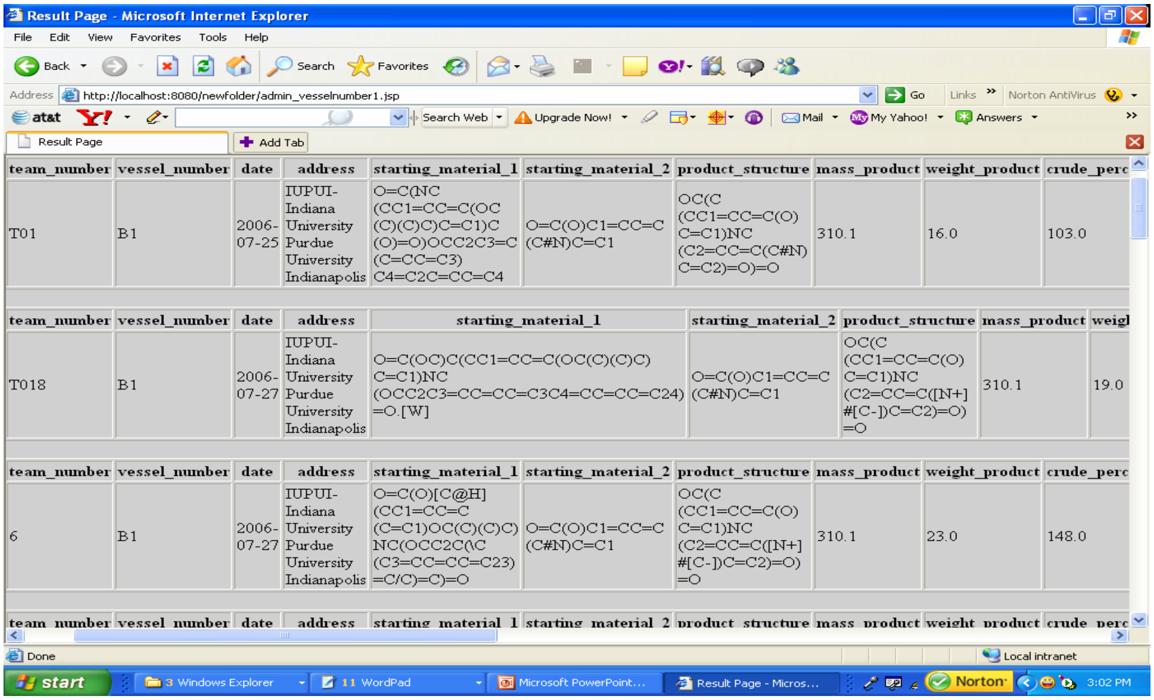
http://localhost:8080/newfolder/admin_vesselnumber.html?option3=Vessel+Number - Microsoft Internet Explorer

Please enter the Vessel Number

B1

submit

Figure 37: Query based on Vessel number



Result Page - Microsoft Internet Explorer

Address http://localhost:8080/newfolder/admin_vesselnumber1.jsp

team_number	vessel_number	date	address	starting_material_1	starting_material_2	product_structure	mass_product	weight_product	crude_perc
T01	B1	2006-07-25	IUPUI-Indiana University Purdue University Indianapolis	O=C(NC(CC1=CC=C(OC(C)(C)C)C=C1)C(O)=O)OCC2C3=C(C=CC=C3)C4=C2C=CC=C4	O=C(O)C1=CC=C(C#N)C=C1	OC(C(CC1=CC=C(O)C=C1)NC(C2=CC=C(C#N)C=C2)=O)=O	310.1	16.0	103.0

team_number	vessel_number	date	address	starting_material_1	starting_material_2	product_structure	mass_product	weigl
T018	B1	2006-07-27	IUPUI-Indiana University Purdue University Indianapolis	O=C(OC)C(CC1=CC=C(OC(C)(C)C)C=C1)NC(OCC2C3=CC=CC=C3C4=CC=CC=C24)=O.[W]	O=C(O)C1=CC=C(C#N)C=C1	OC(C(CC1=CC=C(O)C=C1)NC(C2=CC=C([N+]#[C-])C=C2)=O)=O	310.1	19.0

team_number	vessel_number	date	address	starting_material_1	starting_material_2	product_structure	mass_product	weight_product	crude_perc
6	B1	2006-07-27	IUPUI-Indiana University Purdue University Indianapolis	O=C(O)[C@H](CC1=CC=C(C=C1)OC(C)(C)C)NC(OCC2C(\C(C3=CC=CC=C23)=C/C)=C)=O	O=C(O)C1=CC=C(C#N)C=C1	OC(C(CC1=CC=C(O)C=C1)NC(C2=CC=C([N+]#[C-])C=C2)=O)=O	310.1	23.0	148.0

team_number	vessel_number	date	address	starting_material_1	starting_material_2	product_structure	mass_product	weight_product	crude_perc

Figure 38: Results of the Student Input.

Thus the Distributed Drug Discovery database helps the researchers to store and retrieve the theoretical and experimental information. This acts as a tool for their research process. This database was developed using open source databases, open source cheminformatics software programs which can be accessed through the World Wide Web. To support such databases, many open source software programs would be developed or improvements would be done on the existing packages.

Usability Study

Web Application for Student Interface

The survey for the student interface was published under the name 'CombiChem Database Survey'. The total responses received for the study was 9. The average response time is defined as "the sum of all (end time – start time) for each respondent divided by number of responses" The calculated average response time was 7 minutes and 39 seconds.

Based on the ranking of different characteristics, a 'Response Average' was calculated for each characteristic. Response Average, by definition is: The calculation of Matrix Rating Scale involving weighed average of question results. This calculation of weighed average excludes N/A columns. Weights are assigned as follows depending on the criteria: Column 1 has weight of 1; column 2 has weight of 2, etc. The following formula is used to calculate the Response Average:

Response Average = Sum of each (column weight x number of responses for that column) / (total responses – skipped respondents – responses in N/A column)

For example, if the first row contains the following results:

Column 1: 3 responses

Column 2: 4 responses

Column 3: 2 responses

Column N/A: 1 response, the response average is calculated as follows:

Response Average = (1x3 + 2x4 + 3x2) / (10 – 1) = 1.889

Calculations

1. Table 9 represents the Response Average of Database Website Usability:

Table 9: Response Average of Database Website Usability

Characterize the website?						
	Very Good	Good	Average	Poor	Very Poor	**Response Average**
User-Friendly?	22.20%(2)	44.40%(4)	33.30%(3)	0.00%(0)	0.00%(0)	3.889
Instructions?	11.10%(1)	66.70%(6)	22.20%(2)	000%(0)	0.00%(0)	3.889
Legibility?	25.00%(2)	62.50%(5)	12.50%(1)	0.00%(0)	0.00%(0)	4.125
Navigation?	0.00%(0)	75.00%(6)	25.00%(2)	0.00%(0)	0.00%(0)	3.25
Size of fields?	37.50%(3)	50.00%(4)	12.50%(1)	0.00%(0)	0.00%(0)	4.25
				Total Respondents		9

The weights for the cells to characterize the website were given as Very Good -5, Good – 4, Average -3, Poor – 2, Very Poor -1.

The response average for the first characteristic: User – friendly is calculated as follows:

Response Average = (2x5 + 4x4 + 3x3 + 0x2 + 0x1) / (9) = 3.889.

The response average for the second characteristic: Instructions is calculated as follows:

Response Average = (1x5 + 6x4 + 2x3 + 0x2 + 0x1) / (9) = 3.889.

The response average for the third characteristic: Legibility is calculated as follows:

Response Average = (2x5 + 5x4 + 1x3 + 0x2 + 0x1) / (9-1) = 4.125.

The response average for the fourth characteristic: Navigation is calculated as follows:

Response Average = (0x5 + 6x4 + 2x3 + 0x2 + 0x1) / (9-1) = 3.25.

The response average for the fifth characteristic: Size of fields is calculated as follows:

Response Average = (3x5 + 4x4 + 1x3 +0x2 + 0x1) / (9-1) = 4.25.

The response average results to characterize the website lie on a more than average scale. The users were comfortable in understanding the instructions in completing the tasks.

2. Table 10 below represents the response average for the Marvin Sketch.

Table 10: Response Average of Marvin Sketch

Marvin Sketch										
	Useless								Very Useful	**Response Average**
Usefulness	0.00% (0)	0.00% (0)	12.50% (1)	12.50% (1)	0.00% (0)	25.00% (2)	25.00% (2)	25.00% (2)	0.00% (0)	6.125
	Very Confusing								Simple/ Easy to Use	
Complexity	0.00% (0)	0.00% (0)	12.50% (1)	25.00% (2)	0.00% (0)	12.50% (1)	25.00% (2)	25.00% (2)	0.00% (0)	5.875
	Irritating								Pleasing	
Overall Impression	0.00% (0)	0.00% (0)	12.50% (1)	0.00% (0)	12.50% (1)	12.50% (1)	37.50% (3)	25.00% (2)	0.00% (0)	6.375
									Total Respondents	8
									Skipped this question	1

The usability of Marvin Sketch is characterized based on the Usefulness of the program, Complexity in learning and using the software and the overall impression of the software. Out of the 9 respondents, one respondent skipped this question.

The usefulness of Marvin Sketch is ranked from useless to very useful with weights from useless - 1 to Very useful - 9. The response average is calculated as follows:

Response Average = (1x0 + 2x0+ 3x1 + 4x1 + 5x0 + 6x2 + 7x2 + 8x2 + 9x0) / (9-1) = 6.125.

The complexity of Marvin Sketch is ranked from very confusing to simple/easy to use with weights from Very confusing - 1 to Simple/Easy to use – 9. The response average is calculated as follows:

Response Average = (1x0 + 2x0 + 3x1 + 4x2 + 5x0 + 6x1 + 7x2 + 8x2 + 9x0) / (9-1) = 5.875.

The overall impression of Marvin Sketch is ranked from irritating to pleasing with weights from Irritating – 1 to Pleasing - 9. The response average is calculated as follows:

Response Average = (1x0 + 2x0 + 3x1 + 4x0 + 5x1 + 6x1 + 7x3 + 8x2 + 9x0) / (9-1) = 6.375.

The response average results for the usability of Marvin Sketch in various categories like on a slightly above average scale. Since this a new program, the users might have taken some extra time to understand and properly use the program.

4. Table 11 below represents the response average for the database usability.

A study was conducted to know the ease of data manipulation using the website provided. This is done by rating the advantageousness of the computer entry rather than a paper entry and manipulation of the information in the database. Out of 9 respondents, one respondent skipped this question.

The advantages of computer verses paper entry is ranked from Disadvantage to Highly Advantageous with weights from Disadvantage – 1 to Highly Advantageous – 5.

Table 11: Response Average of Database Usability

Database						
Computer versus Paper entry						
	Disadvantage		No Difference		Highly Advantageous	**Response Average**
Did you find it advantageous to use computer rather than paper entry	0.00% (0)	0.00% (0)	25.00% (2)	50.00% (4)	25.00% (2)	4
Database Entry						
	Very Difficult	Hard	Average	Easy	Very Easy	
Updating/editing information	0.00% (0)	0.00% (0)	50.00% (4)	25.00% (2)	25.00% (2)	3.75
					Total Respondents	8
					Skipped this question	1

The response average is calculated as follows:

Response Average = (1x0 + 2x0+ 3x2 + 4x4 + 5x2) / (9-1) = 4

The ease of updating/editing the database entry is ranked from Very Difficult to Very Easy with weights Very Difficult – 1 to Very Easy – 5. The response average is calculated as follows:

Response Average = (1x0 + 2x0 + 3x4 + 4x2 + 5x2) / (9-1) = 3.75

The response average results for comparing the advantages of computer based entry to a paper based entry fall on an above average scale. This tells that the users are much comfortable of the idea of a computer based entry. They found the computer based/website entry advantageous. They were comfortable in updating and editing the information they wanted.

5. Table 12 below represents the response average for the Overall Impression of the website.

Table 12: Response Average of Overall Impression of Website

Overall Database / Website										
How would you rate the database/website in terms of:										
	Useless								Very Useful	**Response Average**
Usefulness	0.00% (0)	0.00% (0)	12.50% (1)	12.50% (1)	0.00% (0)	37.50% (3)	25.00% (2)	12.50% (1)	0.00% (0)	5.875
	Very Confusing								Simple/ Easy to Use	
Complexity	0.00% (0)	0.00% (0)	0.00% (0)	0.00% (0)	12.50% (1)	12.50% (1)	37.50% (3)	25.00% (2)	12.50% (1)	7.125
	Irritating								Pleasing	
Overall Impression	0.00% (0)	0.00% (0)	0.00% (0)	37.50% (3)	12.50% (1)	0.00% (0)	12.50% (1)	25.00% (2)	12.50% (1)	6.125
								Total Respondents		8
								Skipped this question		1

This study represents the overall impression of the database/website in terms of its usefulness, complexity, etc. This study is intended to know the usefulness of the idea of such databases / websites. Out of 9 respondents, one respondent skipped this question.

The usefulness of the website is ranked from useless to very useful with weights being Useless – 1 to Very useful – 9. The response average is calculated as follows: Response Average = (1x0 + 2x0 + 3x1 + 4x1 + 5x0 + 6x3 + 7x2 + 8x1 + 9x0) / (9-1) = 5.875.

The complexity in using the website is ranked from Very Confusing to Simple/Easy to use with weights being Very confusing – 1 to Simple/easy to use – 9. The response average is calculated as follows:

Response Average = (1x0 + 2x0 + 3x0 + 4x0 + 5x1 + 6x1 + 7x3 + 8x2 + 9x1) / (9-1) = 7.125.

The overall impression of the website is ranked from irritating to pleasing with weights from irritating – 1 to pleasing – 9. The response average is calculated as follows:

Response Average = (1x0 + 2x0 + 3x0 + 4x3 + 5x1 + 6x0 + 7x1 + 8x2 + 9x1) / (9-1) = 6.125.

The users were impressed by the overall performance of the student interface; however there were some suggestions on appearance of the HTML of the website. They said, the website was functionally straightforward. It required some color and fanciness on the appearance. Based on the results and the calculations, the users were comfortable in the migration of paper based entry to a computer based entry. They were comfortable in using the Marvin Sketch software. They found the instructions given on the website appropriate and user-friendly. Also, they found it easy to update the information of the incomplete experiments. They found it simple to use and easy to understand.

Chemical Drawing Packages

Based on the tasks given to the participants, the observations on participants to complete the given tasks were recorded. As part of the results, the satisfaction survey results of all the participants were taken into consideration and measured. To perform this task, the averages of the scores for each of the three criteria for each of the programs were calculated.

The average score is categorized as follows:

- 80-100 – working well, maybe some minor improvements can be made.
- 60-80 – acceptable, but minor to moderate usability problems could make some users reject the software, and others become frustrated.
- Below 60 – major usability problem, most people are unable to complete their task in a satisfactory manner.

ChemDraw

The following table represents the survey results of each of the participant for each program in terms of the following categories: Irritating / Pleasing, Complicated / Straightforward, Useless / Useful along with the average of responses.

Table 13: Survey Results of ChemDraw

Program 1: ChemDraw			
	Irritating / Pleasing	**Complicated / Straightforward**	**Useless / Useful**
Participant 1	40	40	80
Participant 2	100	100	100
Participant 3	90	100	100
Participant 4	45	45	75
Participant 5	60	30	30
Participant 6	20	40	40
Participant 7	50	40	70
Participant 8	45	45	45
Average	**56.25**	**55**	**67.5**

From the above data, the irritating/pleasing score falls under the category of major usability problem i.e., most people are unable to complete their tasks in a satisfactory manner. Since the value is 4.75 points less than the acceptable value i.e., 60 there may be some minor changes that can bring a big difference.

The complicated/straightforward score also falls under the category of major usability problem but as score value is 5 points less than the acceptable value, there may be some minor changes that can bring the software into acceptable category which makes it less complicated and straightforward.

The useless/useful scored highest falling under the acceptable category, but minor to moderate usability problems could make some users reject the software, and others become frustrated. The users realized that it is useful software but requires some major improvements to come under the 'working well' category.

ISIS Draw

Table 14 represents the survey results of each of the participant for each program in terms of the following categories: Irritating / Pleasing, Complicated / Straightforward, Useless / Useful along with the average of responses.

Table 14: Survey Results of ISIS Draw

Program 2: ISIS Draw			
	Irritating / Pleasing	**Complicated / Straightforward**	**Useless / Useful**
Participant 1	10	10	10
Participant 2	70	80	100
Participant 3	10	10	10
Participant 4	25	35	75
Participant 5	50	30	30

Participant 6	20	40	40
Participant 7	60	60	70
Participant 8	35	45	45
Average	**35**	**38.75**	**47.5**

From the above data, the irritating/pleasing score falls under the category of major usability problem i.e., most people are unable to complete their tasks in a satisfactory manner. Major changes should be made to this software to make it pleasing to the users.

The complicated/straightforward score also falls under the category of major usability problem which should be thoroughly looked into. Even the users who had once used this before, but a lesser version are not satisfied and unable to complete the tasks.

The useless/useful average score falls under the category of major usability problem. On the whole, this software requires more work to be put into to be able to be accepted by the users.

Marvin Sketch

Table 15 represents the survey results of each of the participant for each program in terms of the following categories: Irritating / Pleasing, Complicated / Straightforward, Useless / Useful along with the average of responses.

From the data below, the irritating/pleasing, complicated/straightforward, useless/useful scores fall under the acceptable category, but minor to moderate usability problems could make some users reject the software, and others become frustrated. This software ranked second best among all the four software programs

Table 15: Survey Results of Marvin Sketch

Program 3: Marvin Sketch			
	Irritating / Pleasing	**Complicated / Straightforward**	**Useless / Useful**
Participant 1	80	80	80
Participant 2	10	80	100
Participant 3	80	70	70
Participant 4	85	85	85
Participant 5	60	50	50
Participant 6	0	10	10
Participant 7	20	40	80
Participant 8	85	65	65
Average	**60**	**60**	**67.5**

Java Molecule Editor

Table 16 represents the survey results of each of the participant for each program in terms of the following categories: Irritating / Pleasing, Complicated / Straightforward, Useless / Useful along with the average of responses.

Table 16: Survey Results of Java Molecule Editor

Program 4: Java Molecule Editor			
	Irritating / Pleasing	**Complicated / Straightforward**	**Useless / Useful**
Participant 1	100	100	100
Participant 2	40	80	70
Participant 3	80	100	70
Participant 4	85	85	75
Participant 5	70	60	70
Participant 6	90	80	90
Participant 7	60	30	60
Participant 8	35	45	45
Average	**70**	**71.25**	**72.5**

From the above data, the irritating/pleasing, complicated/straightforward, useless/useful scores fall under the acceptable category. This software ranked the best among all the four software programs but could not make to the ‘Working well’ category. Since all the three average scores are more than 70, with some sensible changes could be on the top pleasing all categories of users.

Observations about the programs

Below is the analysis of each program individually along with their strengths and weaknesses.

Program 1: ChemDraw

This program has a tool bar which has an option for aromatics, bonds, text, wedge, orbitals, etc but it lacks in not displaying either the periodic table or the commonly and importantly used elements on the tool bar. The text box provided shows intelligence in recognizing the written molecule. For example when the text box is opened and CH3 is typed (in capitals) it would change the 3 into subscript as CH_3. Also the program showed intelligence in letting the user know whether the structure is correct or not or whether it satisfies the valency or not by a red box around the typed molecule/atom. The help files provided by the program would be helpful to the user in getting started or verifying what he/she has done. The program also allows the user to know what he has selected by its highlighting feature. It has the option of zooming in and out which helps the user in drawing complicated structures. This program requires some understanding of the software before the user is comfortable in drawing the structures.

Program 2: ISIS Draw

This program has tool bar which consists of the aromatics, atoms, text box, bonds, wedge, eraser, etc. But all these icons are not clear to the users. There are two options namely text box and atom box. Some users used one while the others used the other. The users who used the text box were able to produce the same structure as given in the task but was not the one which usually obtained when drawn by the atom box which contained the structure of the molecule. For example if we choose to draw a structure CH_3-----CH_3, using both the options the one with the atom box would yield correct results while the one with the text box yields incorrect results. The users get confused by the options. The single bond icon also posed difficulties as many did not understand how they can join two groups and how to convert a single bond to a double bond. If pressed hard on the icon with the right click one can get double bond, triple bond options and this goes the same with other icons also to get similar properties. But the user who does not know about this feature and who cannot draw double bond by clicking on the bond twice would end up in trouble. The help files for the ISIS Draw are not so user friendly. Surprisingly, the users who have used this software earlier for their class assignments were not comfortable to complete the tasks.

Program 3: Marvin Sketch

This program has some nice features which when understood by the users would be the best program they have used. The first feature which attracts as well as confuses the user is the ‘highlight-icon-before-clicking’. This feature would help the users while drawing the bonds, structure, etc. For example when a hexagon is chosen and the user wants to attach a pentagon to it, when he brings the pentagon near hexagon and orient it he can observe the direction of orientation where he needs it to be attached. Also whatever icon is chosen is followed by the cursor making him/her sure what they have chosen. On the tool bar it has options of commonly used elements along with the aromatics. But the bonds option is not clear to the users as it does not have separate icons for double, triple bonds. And some users get confused when clicking the single bond icon twice or thrice to get double or triple bonds and some may even choose the Cis/Trans double/triple bonds to complete the tasks which is not correct. The software has a nice feature of zooming in and out. Also the option of orientation of the ring structure provides ease to the users to orient the molecule even before clicking on the window. The basic help files provided adds extra ease and comfort for the novice users.

Program 4: JME

This program is a good program for quick and simple structure drawing. It has the essential elements along with the ring structures and all the bonds on the tool bar. If a user wishes to have other elements other than the one on the tool bar, he can use the option ‘X’ to enter the element name he/she wishes to have and clicking on the screen would give the required. It has the option of undo and delete but if the user accidentally presses ‘CLR’ (clear) he/she should have to start it all over again. The software does not have pull down menus or other help files which makes the user a bit irritable. One example where the user gets confused is the disappearance of Carbon. It is quite hard for the users to understand that it does not appear on the screen. Also the user gets frustrated when he first tries to put all the elements and then attach them with the bonds. The software does not support this feature. The user is supposed to first click on a bond before he clicks on the second element.

Ranking based on my observations:

Based on my analysis of the programs and observing the participant performance, Marvin Sketch is the easiest because of its simple tool bar, highlighting before clicking features, its capability of letting the user know where a structure or a bond can be attached by its highlighting circles and orientation features. Though some of these features may confuse some users at first glance, once he/she gets acquainted with the software are sure of appreciating these features. Next the easiest one is JME. Though this is the simplest among all the four programs, the user needs to develop a special style of structure drawing and understand the software. This software is very good for simple structure drawing since it does not exhibit many menus like the other programs did. Next easiest one is the ChemDraw. It has good features embedded in it but it is not that user friendly. It has a good number sub menus which when are understood by the user would help them to construct the structures with ease. The most difficult program was ISIS Draw. It did not provide much flexibility during the construction of structures. It did not allow some users to draw a single bond when they wanted to join two atoms and did not allow them to orient the rings in the required fashion.

5.3 Chemical Facets

The code for the user interface to represent chemical facets is being borrowed from the BioFacets at IUPUI, Indianapolis. The XML files written for chemical facets are integrated in the code. The mechanism begins with a search tool and then provides a searching/browsing to narrow the result set. The user interface is being supported by a web interface. The Figure 39, Figure40, Figure 41 represents the home page for the chemical facets.

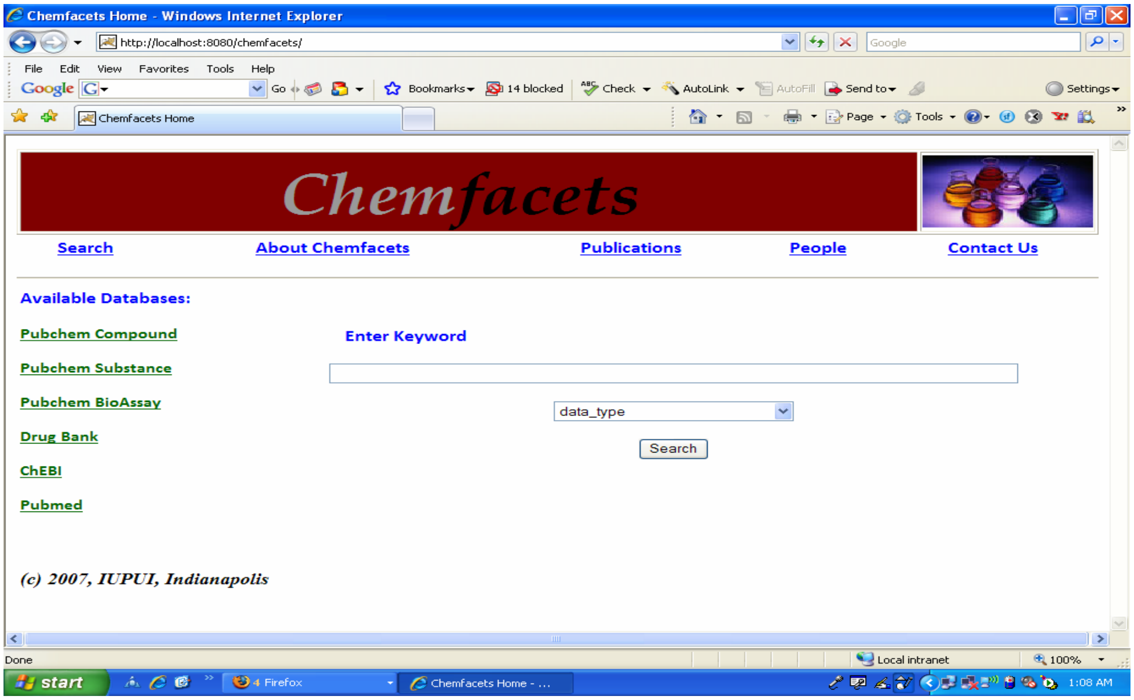


Figure 39: ChemFacets Home Page

This page enables the user to view the different databases integrated into the system. If the users wishes to chose the databases and search them separately he is allowed to follow the database link he desires.

A simple form enables the user to input the query based on different facets. The facets are listed in the drop-down menu below the query input. Once the query is submitted, the results from all these databases are integrated and returned to the user based on the facet chosen.

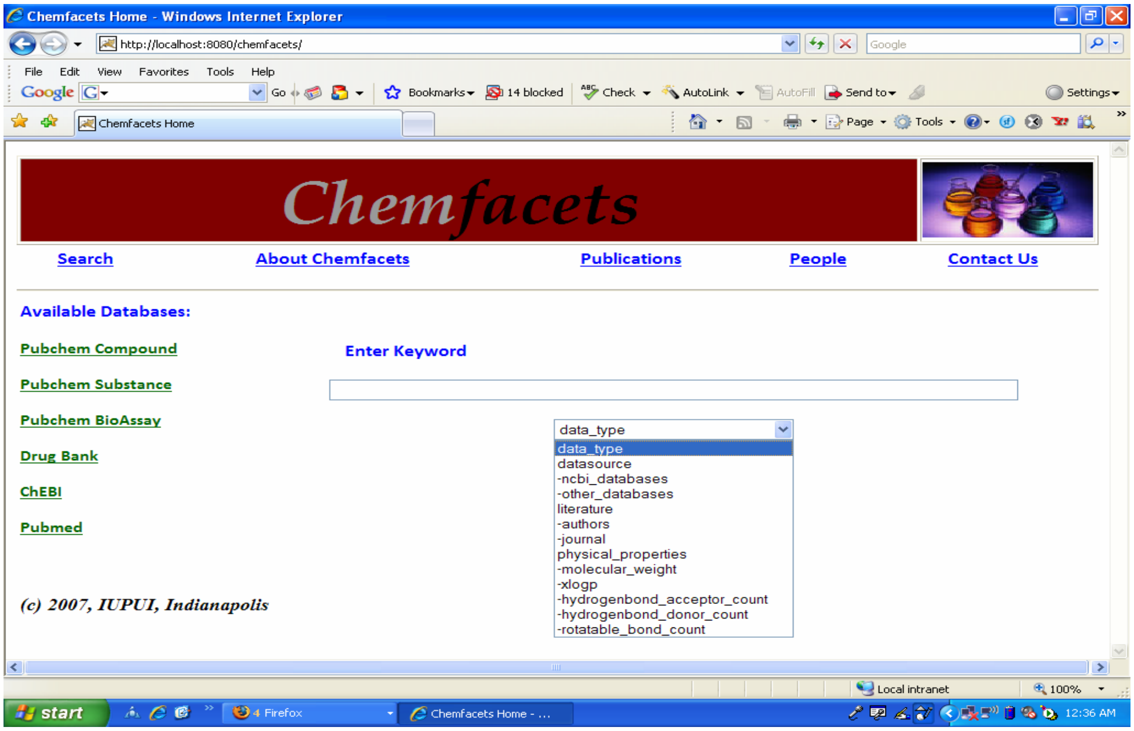
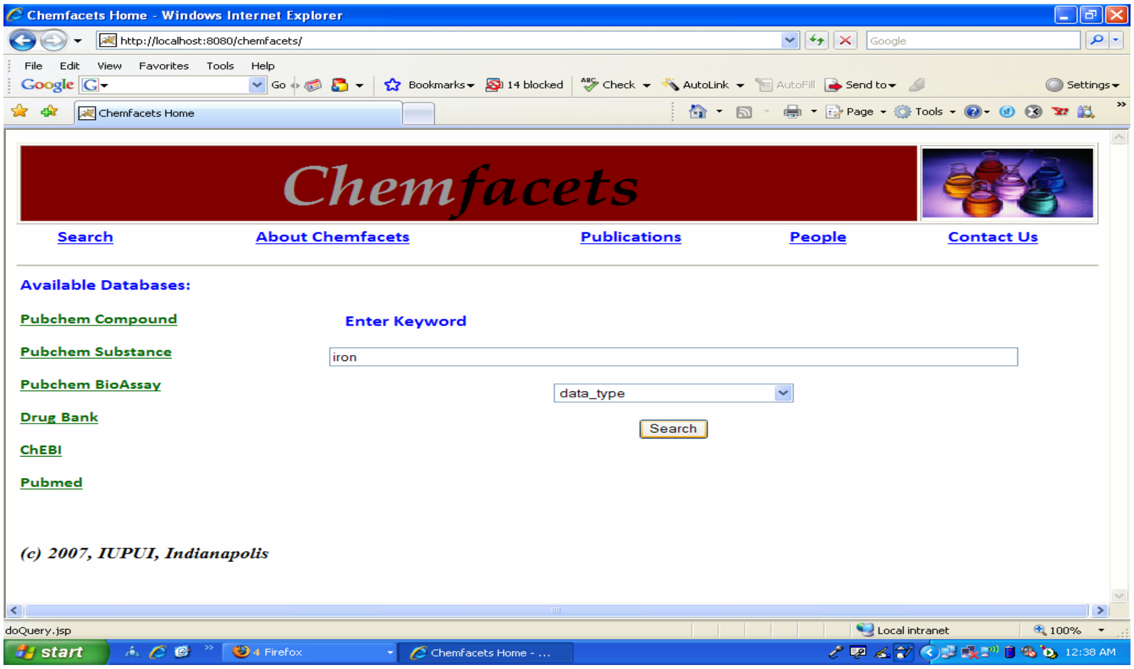


Figure 40: ChemFacets Home Page



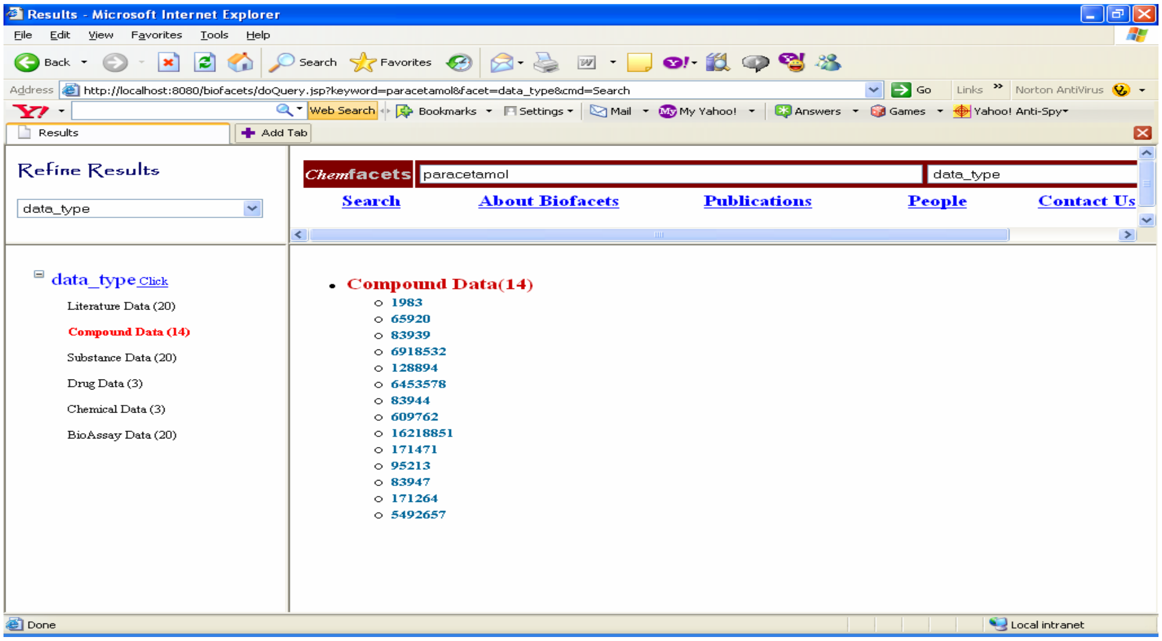
Enter Keyword

iron

data_type

Figure 41: ChemFacets Home Page

Figure 42 and Figure 43 illustrates the results of records obtained by the data type facet for compound data.

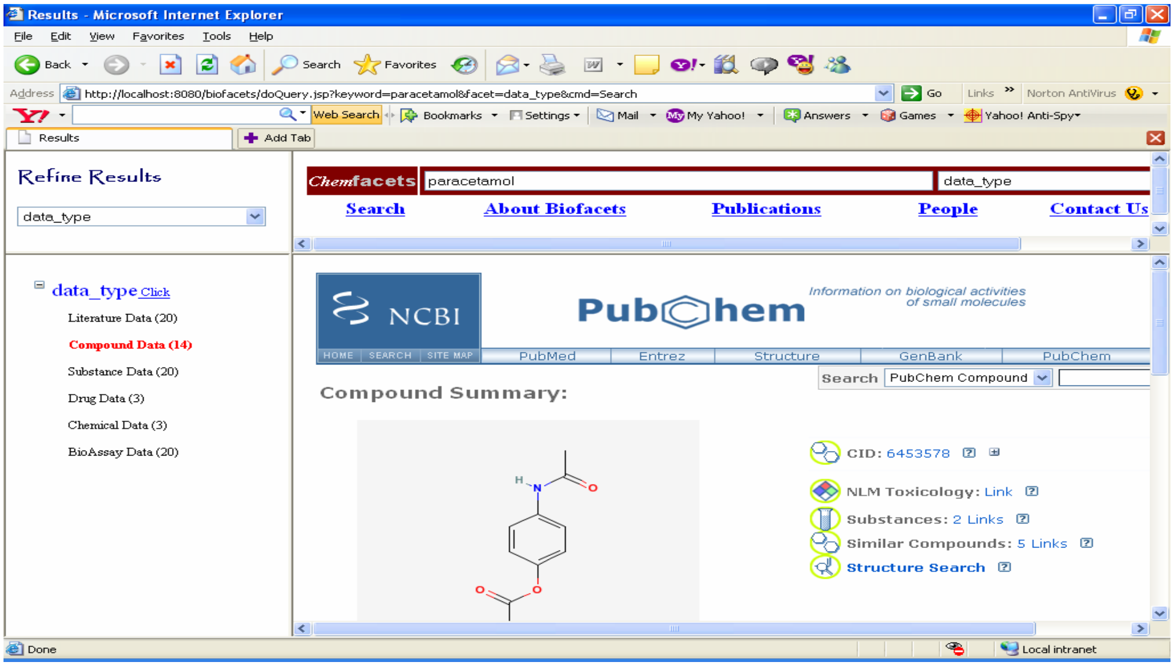

Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://localhost:8080/biofacets/doQuery.jsp?keyword=paracetamol&facet=data_type&cmd=Search
Go Links Norton AntiVirus
Web Search Bookmarks Settings Mail My Yahoo! Answers Games Yahoo! Anti-Spy
Results Add Tab
Refine Results
data_type
Chemfacets
Search About Biofacets Publications People Contact Us
Done Local intranet

Search: paracetamol — data_type

- data_type Click
  - Literature Data (20)
  - **Compound Data (14)**
  - Substance Data (20)
  - Drug Data (3)
  - Chemical Data (3)
  - BioAssay Data (20)

- **Compound Data(14)**
  - 1983
  - 65920
  - 83939
  - 6918532
  - 128894
  - 6453578
  - 83944
  - 609762
  - 16218851
  - 171471
  - 95213
  - 83947
  - 171264
  - 5492657

Figure 42: Results of ChemFacets


Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://localhost:8080/biofacets/doQuery.jsp?keyword=paracetamol&facet=data_type&cmd=Search
Go Links Norton AntiVirus
Web Search Bookmarks Settings Mail My Yahoo! Answers Games Yahoo! Anti-Spy
Results Add Tab
Refine Results
data_type
Chemfacets
Search About Biofacets Publications People Contact Us
NCBI
HOME SEARCH SITE MAP
PubChem
PubMed Entrez Structure GenBank PubChem
Search PubChem Compound
H N O O O
Done Local intranet

Search: paracetamol — data_type

- data_type Click
  - Literature Data (20)
  - **Compound Data (14)**
  - Substance Data (20)
  - Drug Data (3)
  - Chemical Data (3)
  - BioAssay Data (20)

Information on biological activities of small molecules

Compound Summary:

CID: 6453578

NLM Toxicology: Link

Substances: 2 Links

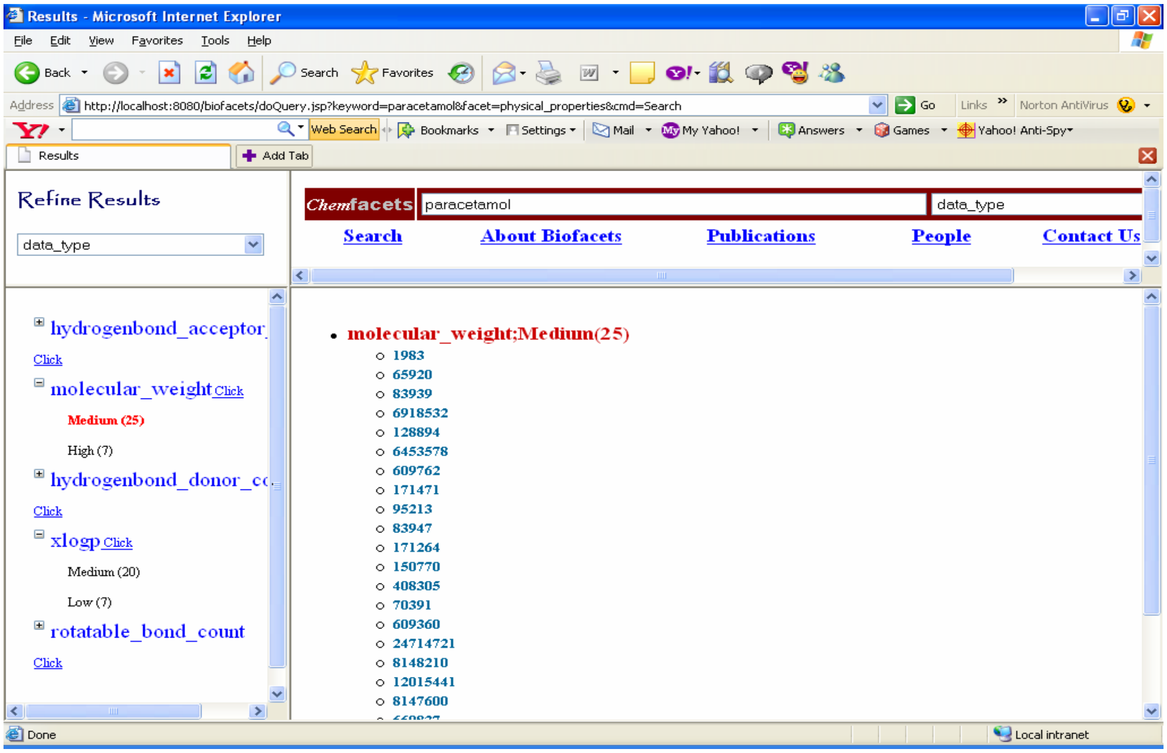
Similar Compounds: 5 Links

Structure Search

Figure 43: Results of ChemFacets

All the records are being displayed from these different databases. The interface provides the user with a simultaneous search/browse mechanism. The search bar at the top may be used to submit a fresh query. The user may also narrow the result set by selecting a desired facet value. From the obtained data set, the user can select the desired record to be directed to the original database.

Figure 44 illustrates the concept of grouped facets. The records are classified based on a set of grouped results. For example, all records having a medium molecular weight are grouped under one category, Low. Similarly compounds having a range of xlogp's are grouped under one category. Likewise each physical property facet is divided into three categories namely low, medium and high.



Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address Go Links Norton AntiVirus
Web Search Bookmarks Settings Mail My Yahoo! Answers Games Yahoo! Anti-Spy
Results
Add Tab
Refine Results
Chemfacets
Search
About Biofacets
Publications
People
Contact Us
Done
Local intranet

Address: http://localhost:8080/biofacets/doQuery.jsp?keyword=paracetamol&facet=physical_properties&cmd=Search

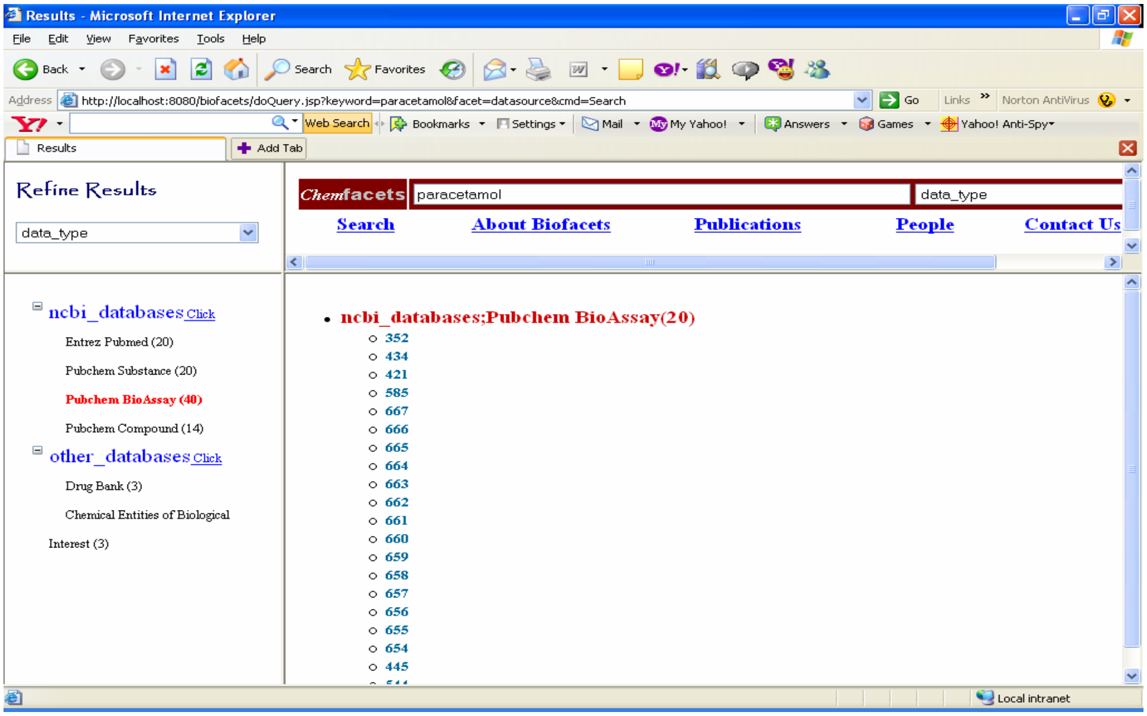
Search box: paracetamol

Selected: data_type

- hydrogenbond_acceptor Click
- molecular_weight Click
  - Medium (25)
  - High (7)
- hydrogenbond_donor_c Click
- xlogp Click
  - Medium (20)
  - Low (7)
- rotatable_bond_count Click

- molecular_weight;Medium(25)
  - 1983
  - 65920
  - 83939
  - 6918532
  - 128894
  - 6453578
  - 609762
  - 171471
  - 95213
  - 83947
  - 171264
  - 150770
  - 408305
  - 70391
  - 609360
  - 24714721
  - 8148210
  - 12015441
  - 8147600

Figure 44: Grouped Facets


Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address
Go Links Norton AntiVirus
Web Search Bookmarks Settings Mail My Yahoo! Answers Games Yahoo! Anti-Spy
Results
Add Tab
Local intranet

http://localhost:8080/biofacets/doQuery.jsp?keyword=paracetamol&facet=datasource&cmd=Search

Refine Results

data_type

Chemfacets paracetamol data_type

Search About Biofacets Publications People Contact Us

- ncbi_databases Click
  - Entrez Pubmed (20)
  - Pubchem Substance (20)
  - **Pubchem BioAssay (40)**
  - Pubchem Compound (14)
- other_databases Click
  - Drug Bank (3)
  - Chemical Entities of Biological Interest (3)

- **ncbi_databases;Pubchem BioAssay(20)**
  - 352
  - 434
  - 421
  - 585
  - 667
  - 666
  - 665
  - 664
  - 663
  - 662
  - 661
  - 660
  - 659
  - 658
  - 657
  - 656
  - 655
  - 654
  - 445

Figure 45: Data Source Results

Figure 45 illustrates the records obtained based on data source facet. Thus the users have the ease of switching between different databases and explore data collectively. It allows them to view a majority of relevant results keeping in mind the themes or facets present.

6 CONCLUSION

Overview

In this research, three different tools were designed, developed and implemented in aiding researchers in the field of drug discovery and development. The research helps in the idea of utilizing open source (freely available) tools.

The first tool, Distributed drug discovery database aids the researchers in keeping track of the theoretical and the experimental data. This is being developed using the open source database, PostgreSQL, ChemAxon's tools, Java and Tomcat.

The second study, Usability Analysis is an implementation tool which helps researchers in designing new products and makes considerable changes to the existing products based on the users' comments and suggestions. This is a powerful analysis tool to test and analyze the usefulness of a particular product.

The third study is designed to help researchers query multiple databases with query browsing and results refinement approach. The faceted classification, implemented via the classification rules helped in classifying the chemical data associating with different characteristics or facets.

Future Work

As part of future work, scalability and efficiency are the main areas of focus. The Drug Discovery Database application should be extended to other countries (Moscow, Lublin) that are interested in participating in the research and store their data in the database. Efficient search mechanisms and automated support should be integrated in both the Drug Discovery database and Chemical Facets design in order to make a robust system. An in depth Usability Study design should be formulated which can help the developers in designing packages that are more user friendly and more error free.

7 REFERENCES

1. Andrew R. Leach, Valerie J. Gillet (2003). An Introduction to Chemoinformatics. Netherlands, Kluwer Academic Publishers.
2. An internal report given by Dr.Scott
3. Usability Testing, http://www.cu.edu/irm/stds/usability/
4. Structure Based Search, http://webbook.nist.gov/chemistry/str-app.shtml
5. Chemical Structure Drawing, http://changbioscience.com/mis/chemdraw.html
6. ChemExper – catalog of chemicals, suppliers, physical characteristics and search engine, http://chemexper.com/
7. Online Chemical Information, http://www.hellers.com/steve/resume/p100.html
8. Amazon, www.amazon.com
9. Usability Testing, http://en.wikipedia.org/wiki/Usability_testing
10. Drug Discovery News: New Products, http://www.drugdiscoverynews.com/index.php?newsarticle=454.
11. A computer aided drug discovery system for chemistry teaching, http://pubs.acs.org/cgi-bin/abstract.cgi/jcisd8/asap/abs/ci050383q.html.
12. A Distributed Drug Discovery Concept to Search for Developing World Disease Drug Leads, https://oncourse.iu.edu/access/content/group/15c9b083-8c80-4004-8066-5a0521cd91c7/Material%20Given%20by%20Dr.Scott/William%20Scott%20D3%208_26_05_02.pdf.
13. Structural Search using ChemAxon tools, http://www.chemaxon.com/forum/download582.ppt
14. Real world chemistry, http://www.chemaxon.com/forum/download389.ppt

15. Developing a compact chemical Database system using Marvin tools, http://www.chemaxon.com/forum/download415.pdf

16. Virtual Classrooms and E-Learning: Bringing Cheminformatics Training Into Academic and Industrial Settings, TJ O'Donnell, John MacCuish and Norah MacCuish http://www.chemaxon.com/forum/viewpost2300.html

17. M. Mahoui, Z. B. Miled, A. Godse, H. Kulkarni, N. Li, "BioFacets: Faceted Classification for Biological Information. IEEE Computer Society, 2006.

18. PostgreSQL: The world's most advanced open source database, http://www.postgresql.org/.

19. MySQL AB: The world's most popular open source database, http://www.mysql.com/

20. MySQL vs PostgreSQL, http://articles.techrepublic.com.com/5100-22-1050671.html.

21. Mysql vs postgres – GEANT2-JRA1 Wiki, http://monstera.man.poznan.pl/wiki/index.php/Mysql_vs_postgres.

22. The PubChem Project, http://pubchem.ncbi.nlm.nih.gov.

23. DrugBank Homepage, http://redpoll.pharmacy.ualberta.ca/drugbank/.

24. Chemical Entities of Biological Interest (ChEBI), http://ebi.ac.uk/chebi.

25. ChemAxon, www.chemaxon.com.

26. FacetMap, http://facetmap.com/.

27. Flamenco, http://bailando.sims.berkeley.edu/flamenc.html.

28. National Institute of Health, http://www.nih.gov.

29. Database, http://www.ncbi.nlm.nih.gov.gov/Database/.

ADDITIONAL REFERENCES

1. Configuring and Using Apache Tomcat, http://www.coreservlets.com/Apache-Tomcat-Tutorial/

2. JSP Tutorial, http://www.fing.edu.uy/~ruggia/tecn/J2EE/VisualBuilder_JSP_tutorial.pdf

3. BEA Product documentation, http://www.weblogic.com/docs51/intro/intro_jdbc.html

4. Java Input and Output, http://www.cs.wisc.edu/~cs302/io/JavaIO.html

5. Forms, Populating a drop down menu with info from database, http://www.powerasp.com/content/code-snippets/forms-populate-drop-down-menu.asp

6. Chemical Databasing, http://www.acdlabs.com/download/catalogs/dbcat.pdf

7. Java Technology Forums, Executing DOS commands in Java, http://forum.java.sun.com/thread.jspa?threadID=523745&start=15&tstart=0

8. Java Technology Forums, Executing DOS commands in JSP, http://forum.java.sun.com/thread.jspa?threadID=640785&messageID=4132747

9. New Address creation using ExecuteUpdate, Database, JSP, http://www.java2s.com/Code/Java/JSP/NewAddressCreationusingexecuteUpdate.htm

10. Accelrys, http://www.accelrys.com/technologies/informatics/cheminformatics/

11.Factual Databases in Chemistry, http://www.hellers.com/steve/resume/p98.html.

12. http://www.chemaxon.com/conf/Integrating_chemaxon_technology_into_your_End_User_Applications.pdf

APPENDIX – A

STUDY #05-10124

INDIANA UNIVERSITY - BLOOMINGTON

INFORMED CONSENT STATEMENT

USABILITY OF CHEMICAL STRUCTURE DRAWING TOOLS

You are invited to participate in a research study. The purpose of this study is to compare the usability of several chemical structure drawing tools.

INFORMATION

During this usability study, you will be asked to perform a small number of chemical structure drawing tasks using one or more chemical structure drawing packages. The purpose of this is to measure how effectively the tools have been designed, not to assess your performance. 5-10 participants will be recruited in total. An observer will take notes during the study, but no personal information (including your name) will be preserved after the study is complete. The notes taken will be observations about how you interacted with the tool and how it responded. The study will take less than one hour. If you wish to pause for a break or stop the session at any time you may do so.

BENEFITS

By participating in this study, you will help provide valuable feedback on the design of structure drawing tools.

CONFIDENTIALITY

Your name and the fact that you are participating in this study will be treated confidentially. After the test, no documentation will be kept with your name on or other personal identifiable details.

RISK

There are no foreseeable risks associated with this research.

CONTACT

If you have questions at any time about the study or the procedures, you may contact the researcher,

Dr. David Wild, at 1900 East Tenth Street, Bloomington, IN, (812) 856-1848, djwild@indiana.edu.

If you feel you have not been treated according to the descriptions in this form, or your rights as a participant in research have been violated during the course of this project, you may contact the office for the Indiana University Bloomington Human Subjects Committee, Carmichael Center L03, 530 E. Kirkwood Ave., Bloomington, IN 47408, 812/855-3067, or by e-mail at iub_hsc@indiana.edu.

PARTICIPATION

Your participation in this study is voluntary, you may refuse to participate without penalty. If you decide to participate, you may withdraw from the study at anytime without penalty and without loss of benefits to which you are otherwise entitled. If you withdraw from the study before data collection is completed your data will be returned to you or destroyed.

CONSENT

I have read the consent form, have had questions answered to my satisfaction, acknowledge receiving a copy of this form, and agree to take part in the study.

VOLUNTEER: ____________________ INVESTIGATOR: ___________________

DATE: ___ / ____ / ____ DATE: ___ / ____ / ____

Information Sheet date June 6th, 2005

(1 of 1 pages)

2D Structure Drawing Software Usability Study

Thank you for agreeing to participate in this study. The goal of the project is to determine the relative usability of four different free 2D structure drawing software.

Disclaimers

1. Please be aware and understand that it is the system and the specific applications that are being tested, not your aptitude as a user.
2. Your name will only be divulged to my supervisor, Dr. David Wild and will be coded for my presentation.
3. A test supervisor will be in the room behind you taking notes. Please pretend he/she is not there and do not ask him/her for any personal opinions.

Set-up

You will note that a desktop folder entitled "Usability Study" is open and there are four icons, titled Program 1, Program 2, Program 3, and Program 4 in the folder. Please complete each of the three tasks listed below with all four programs before moving on to the next task. If you feel like you have exhausted all your resources and you cannot figure out how to complete the task, please remark "Concede" to the test supervisor and move on to the next program or next task.

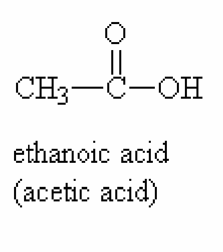
At the completion of the study, please fill out a short satisfaction survey and list any comments or suggestions you may have. If you conceded any of the tasks on any of the programs, please use this space to explain why you could not complete that task.

Please ask the test supervisor any questions at this time.

If you have no further questions, please turn to next page and begin with Task 1.

Task 1: Acetic acid (ethanoic acid) CH_3COOH

Please draw Acetic acid(shown below), with each program.

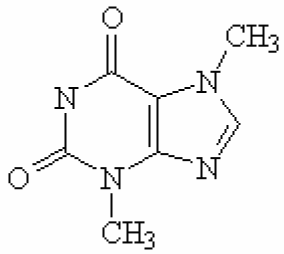


When you think you are done with this task for each program, please remark to the test supervisor that you are moving to the next task.

You may use the space below to note any comments that may be useful to the test supervisor.

Task 2: Theobromine $C_7H_8N_4O_2$

Please draw Theobromine (shown below) with each program.



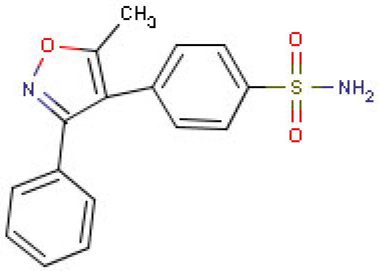
Theobromine
3,7-dihydro-3,7-dimethyl-1H-purine-2,6-dione

When you think you are done with this task for each program, please remark to the test supervisor that you are moving to the next task.

You may use the space below to note any comments that may be useful to the test supervisor.

Task 3: Valdecoxib $C_{16}H_{14}N_2O_3S$

Please draw Valdecoxib (shown below) with each program.



When you think you are done with this task for each program, please remark to the test supervisor that you are moving to the next task.

You may use the space below to note any comments that may be useful to the test supervisor.

SURVEY AND COMMENTS

Thank you for participating in my Usability Study of 2Dchemical drawing software. Please complete the following satisfaction survey by simply placing an "X" on the line corresponding to how irritating/pleasing, complicated/straightforward, and useless/useful you felt each software was. Please then note any comments/suggestions you have about the programs. Please also mention if you have worked with any of the programs before.

Program 1

Irritating |—|—|—|—|—|—|—|—|—|—| Pleasing

Complicated |—|—|—|—|—|—|—|—|—|—| Straightforward

Useless |—|—|—|—|—|—|—|—|—|—| Useful

Comments and Suggestions

Program 2

Irritating |—|—|—|—|—|—|—|—|—|—| Pleasing

Complicated |—|—|—|—|—|—|—|—|—|—| Straightforward

Useless |—|—|—|—|—|—|—|—|—|—| Useful

Comments and Suggestions

Program 3

Irritating |—|—|—|—|—|—|—|—|—|—| Pleasing

Complicated |—|—|—|—|—|—|—|—|—|—| Straightforward

Useless |—|—|—|—|—|—|—|—|—|—| Useful

Comments and Suggestions

Program 4

Irritating	├─┼─┼─┼─┼─┼─┼─┼─┼─┼─┤	Pleasing
Complicated	├─┼─┼─┼─┼─┼─┼─┼─┼─┼─┤	Straightforward
Useless	├─┼─┼─┼─┼─┼─┼─┼─┼─┼─┤	Useful

Comments and Suggestions

APPENDIX – B

Installation of PostgreSQL 7.4

PostgreSQL is usually run on a UNIX-compatible platform. The software packages that are required for building PostgreSQL are GNU make, ISO/ANSI C compiler, gzip and GNU Readline library. For this project, PostgreSQL installed on Cheminfo server located at Indiana University, Bloomington is used. To install this, one needs to be the administrator. The sources to download the file are obtained from ftp://ftp.postgresql.org/pub/source/v7.4.12/postgresql-7.4.12.tar.gz. The file is unpacked using the following commands: gunzip postgresql-7.4.12.tar.gz
tar xf postgresql-7.4.12.tar.

This would create a directory named "postgresql-7.4.12" under the current directory with the PostgreSQL sources. The rest of the installation is done by changing into that directory. The first step in the installation is to configure the source tree and choosing the options as per requirements. This is done by the following command: ./configure. This script runs a number of tests to guess values for various system dependent variables and detect some quirks of the operating system. All the files are installed under /usr/local/pgsql by default.

The second step is to Build. The command used is: gmake. This command usually takes from 5 minutes to half an hour to run depending on the hardware. After the command is executed, the last line displayed appears as follows: "All of PostgreSQL is successfully made. Ready to install."

The third step is to run the regression tests using the following command: gmake check. Regression tests are a test suite to verify that PostgreSQL runs on the machine on which it is being installed in the way the developers expected it to be. This command is usually done as an unprivileged user and not at the root.

The fourth step is to install the files by typing the following command: gmake install. This would install the files into the directories that were specified in step one. This command is usually run as a root user, usually to have write permissions in that area.

Then the following commands are run to create a database cluster i.e., to initialize database storage area on disk. A database cluster, is a collection of databases, is accessible by a single instance of running database server. A directory is created at the root and the owner is specified as the PostgreSQL user. This is done by the following commands:

```
 root# mkdir /usr/local/pgsql/data
root# chown postgres /usr/local/pgsql/data
root# su postgres
postgres$ initdb -D /usr/local/pgsql/data
```

Then the database server is restarted and a database named “gnova” is created using the following commands:

```
 /usr/local/pgsql/bin/postmaster -D
/usr/local/pgsql/data >logfile 2>&1 &
/usr/local/pgsql/bin/createdb gnova
/usr/local/pgsql/bin/psql gnova
```

After the successful installation of PostgreSQl, the users are created with read and write permissions as per required.

APPENDIX-C

Installation of Tomcat Server, JDBC driver:

There are GUI’s available for the PostgreSQL database but only at an administrator level. The available GUI’s are free only for the first 30 days and then require a licensing fee. These do not allow a common user to browse the internet and search the database nor allow them to do on any computer. They work on a particular computer only on which they are installed. The GUI’s would be expensive to afford.

To make the database accessible to everyone who uses it having an Internet explorer, the client side application is built using JSP pages running on Tomcat server.

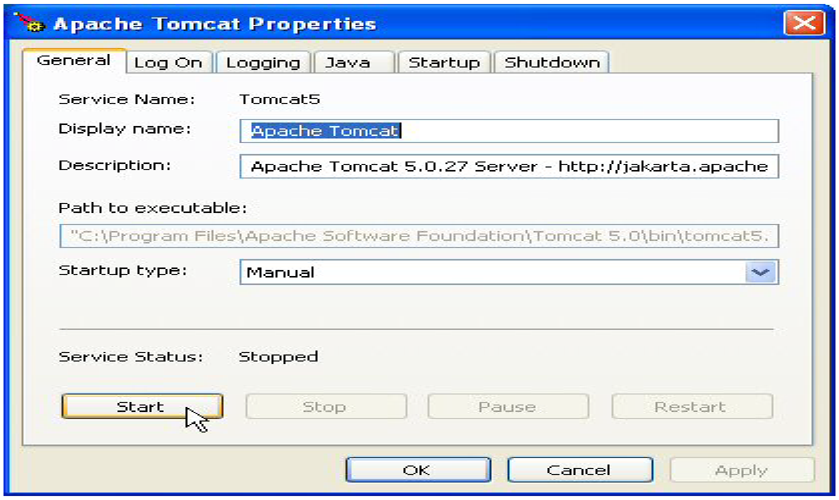
The following steps involved the set up procedure.

Step 1. JAVA installation: Full JDK (J2SE Development Kit) is downloaded and java is installed. To confirm that Java installation and the PATH are configured properly, the following commands: “java –version” and “javac –help” are ran in a DOS (Disk operating system) window. A result appeared both the times and not an error about an unknown command.

Step 2. JDBC Driver Setup: The driver is downloaded and installed from http://jdbc.postgresql.org/. It is stored in the following folder: “C:\Tomcat 5.5\webapps\example\WEB-INF\lib”. The path is set with an environment variable “JAVA_HOME” and the path as “C:\program files\Java\jdk 1.5.0_06\bin”.

Step 3. Configure Tomcat: The Jakarta Tomcat software is downloaded from: http://jakarta.apache.org/site/binindex.cgi#tomcat. For Windows there is a .exe installer.

To start running the Tomcat, one needs to go to the Start Menu, select Apache Tomcat 5.5 and select Monitor Tomcat and then select Start or Stop as shown in the following figure 43.



Service Name: Tomcat5

Display name: Apache Tomcat

Description: Apache Tomcat 5.0.27 Server - http://jakarta.apache

Path to executable:

"C:\Program Files\Apache Software Foundation\Tomcat 5.0\bin\tomcat5.

Startup type: Manual

Service Status: Stopped

PostgreSQL database configuration settings to access the remote server are made from the operating system by adding the port number and allowing TCP connections as follows:

For Windows XP: Windows Firewall -> Exceptions -> Add Port -> Name: Postgre, Port number: 5432 TCP.

JSP programs:

All the JSP programs and html files are stored in a new folder in the Tomcat webapps directory. Two more directories named classes and lib are created in the following folder in which the class files are placed.

For running the program, the tomcat server needs to be started and then the web address should be given in the Internet Explorer as http://localhost:8080/example/name of the file. The following JSP code is used to connect to the database on the cheminfo server.

```
try {
    Class.forName ("org.postgresql.Driver");
   }
   catch (Exception e) {
    out.println ("Error occurred” + e);
   }
   try {
     conn =
DriverManager.getConnection("jdbc:postgresql://cheminfo.informatics.indiana.edu:5432/
gnova","cicc2","linux271");
   }
   catch (SQLException e) {
      out.println("Error occurred " + e);
   }
```

APPENDIX-D

Example of Extraction Rules

1. PubChem Compound Summary Extraction Rules:

```
<?xml version="1.0" encoding="UTF-8"?>
<!--
url:http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pccompound&cmd=sea
rch&term=chlorine
-->
<database name="NCBI_PubChem_Compound">
    <prefetch_rule>
        <base_url></base_url>
        <append_url_field>extended_record_link</append_url_field>
    </prefetch_rule>
    <field name="summery_record_list" save_value="false" minOccurs="0"
maxOccurs="1" multi_value="false" multi_page="false">
        <extraction_rules>
            <ld>Items+of+&lt;table+&gt;</ld>
            <rd>Items</rd>
        </extraction_rules>
        <next_page>
            <extraction_rules>
                <ld></ld>
                <rd></rd>
            </extraction_rules>
        </next_page>
        <field name="record" save_value="false" minOccurs="0"
maxOccurs="unbounded" multi_value="false">
            <extraction_rules>
                <ld>&lt;input</ld>
                <rd>&lt;/dl&gt;&lt;dl&gt;&lt;dt&gt;</rd>
            </extraction_rules>
            <field name="extended_record_link" save_value="true"
minOccurs="1" maxOccurs="1" multi_value="false">
                <extraction_rules>
                    <ld>a href="</ld>
                    <rd>"&gt;</rd>
                </extraction_rules>
            </field>
            <field name="pubchem_compound_id" save_value="true"
minOccurs="1" maxOccurs="1" multi_value="false" displayAsName="true">
                <extraction_rules>
                    <ld>"&gt;</ld>
                    <rd>&lt;/a&gt;</rd>
                </extraction_rules>
            </field>
            <field name="compound_name" save_value="true" minOccurs="1"
maxOccurs="1" multi_value="false">
                <extraction_rules>
                    <ld>&lt;td&gt;&lt;dt&gt;&lt;dd&gt;</ld>
                    <rd>&lt;/dd&gt;</rd>
                </extraction_rules>
            </field>
            <field name="IUPAC_name" save_value="true" minOccurs="0"
maxOccurs="1" multi_value="false">
                <extraction_rules>
                    <ld>&lt;dd&gt;IUPAC:</ld>
                    <rd>&lt;/dd&gt;</rd>
                </extraction_rules>
            </field>
            <field name="properties" save_value="true" minOccurs="1"
maxOccurs="1" multi_value="false">
                <extraction_rules>
                    <ld>&lt;dd&gt;</ld>
                    <rd>&lt;/dd&gt;</rd>
                </extraction_rules>
            </field>
        </field>
     </field>
</database>
```

2. PubChem Compound Extended Extraction Rules

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="yes"?>
<database name="NCBI_PubChem_Compound">
    <!--
        Given the xml file containing the extracted summery info, the
prefetch rule tells us how to get the full records. This information
tells us what base URL to use and which field gives us the remaining
part of the URL. The code will fetch each of this URLs and append them
in a file.
    -->
    <prefetch_rule>
        <base_url></base_url>
        <append_url_field>extended_record_link</append_url_field>
    </prefetch_rule>

    <field name="record" save_value="false" minOccurs="1" maxOccurs="1"
multi_value="false">
        <extraction_rules>
            <ld>&lt;b&gt;Compound Summary:+&lt;/b&gt;</ld>
            <rd></rd>
        </extraction_rules>

            <field name="pubchem_compound_id" save_value="true"
minOccurs="0" maxOccurs="1" multi_value="false">
                <extraction_rules>
                    <ld>CID:</ld>
                    <rd>&lt;/a&gt;</rd>
                </extraction_rules>
            </field>
            <field name="medical_subject_annotation" save_value="true"
minOccurs="0" maxOccurs="1" multi_value="false">
                <extraction_rules>
                    <ld>&lt;b&gt;Medical Subject Annotations:
&lt;/b&gt;</ld>                    <rd>&lt;br&gt;&lt;/div&gt;</rd>
                </extraction_rules>
            </field>
            <field name="molecular_weight" save_value="true" 
minOccurs="0" maxOccurs="1" multi_value="false">
                <extraction_rules>
                    <ld>molecular weight: &lt;/b&gt;</ld>
                    <rd>&lt;br&gt;</rd>
                </extraction_rules>
            </field>
            <field name="molecular_formula" save_value="true" 
minOccurs="0" maxOccurs="1" multi_value="false">
                <extraction_rules>
                    <ld>Molecular Formula: &lt;/b&gt;</ld>
                    <rd>&lt;br&gt;</rd>
                </extraction_rules>
            </field>
            <field name="xlogp" save_value="true" minOccurs="0" 
maxOccurs="1" multi_value="false">
                <extraction_rules>
                    <ld>XLogP: &lt;/b&gt;</ld>
                    <rd>&lt;br&gt;</rd>
                </extraction_rules>
            </field>
            <field name="hydrogen_bond_donor_count" save_value="true" 
minOccurs="0" maxOccurs="1" multi_value="false">
                <extraction_rules>
                    <ld>Hydrogen Bond Donor Count: &lt;/d&gt;</ld>
<rd>&lt;br&gt;</rd>
                </extraction_rules>
            </field>
            <field name="hydrogen_bond_acceptor_count" 
save_value="true" minOccurs="0" maxOccurs="1" multi_value="false">
                <extraction_rules>
                    <ld>Hydrogen Bond Acceptor Count: &lt;/b&gt;</ld>
<rd>&lt;br&gt;</rd>
                </extraction_rules>
            </field>
            <field name="rotatable_bond_count" save_value="true" 
minOccurs="0" maxOccurs="1" multi_value="false">
                <extraction_rules>
                    <ld>Rotatable Bond Count: &lt;/b&gt;</ld>
                    <rd>&lt;br&gt;</rd>
                </extraction_rules>
            </field>
            <field name="iupac_name" save_value="true" minOccurs="0" 
maxOccurs="1" multi_value="false">
                <extraction_rules>
                    <ld>IUPAC Name: &lt;/b&gt;</ld>
                    <rd>&lt;br&gt;</rd>
                </extraction_rules>
            </field>
            <field name="canonical_smiles" save_value="true" 
minOccurs="0" maxOccurs="1" multi_value="false">
                <extraction_rules>
                    <ld>Canonical SMILES: &lt;/b&gt;</ld>
                    <rd>&lt;br&gt;</rd>
                </extraction_rules>
            </field>
            <field name="iupac_international_chemical_identifier" 
save_value="true" minOccurs="0" maxOccurs="1" multi_value="false">
                <extraction_rules>
                    <ld>InChI: &lt;/b&gt;</ld>
                    <rd>&lt;/a&gt;</rd>
                </extraction_rules>
            </field>
    </field>
</database>
```

CURRICULUM VITAE

USHA DEEPIKA CHEEMAKURTHI

38812 Polo Club Dr. # 201
Farmington Hills, MI 48335

Email**:** ushavungutur@yahoo.com
Phone: (248)478-6402, (317)828-7539

OBJECTIVE: Seeking a challenging software engineer position in QA/development primarily in areas of testing, development and database management.

SUMMARY

- Experienced in development of front-end and back-end applications using Java, JSP 1.2, PostgreSQL, JDBC, XML and HTML.
- Experienced in prototype design and documentation for open source databases.
- Experienced in working with Relational Databases like PostgreSQL, Oracle 9i/8i and MS Access.
- Experienced in working with Application and Web Servers Apache Tomcat 5.1.
- Knowledge in tools like ERWIN, Altova XML spy, CVS, Bugzilla.
- Good analytical skills and problem solving skills.
- Excellent communication skills.

PROFESSIONAL SKILLS

Languages	SQL, XML, Java.
Technologies	JSP1.2, JDBC, XML
Web Server	Tomcat 5.1
Data Bases	MySQL, Oracle 9i/8i, MS Access, PostgreSQL
Operating Systems	Windows, Linux, MS-DOS
Chemical Databases	Scifinder, PubChem, DrugBank, ChEBI, ChemTK
Biological Databases	NCBI, OMIM, Pubmed

Tools MATLAB 7, ERWIN, CVS, Bugzilla, AltovaXML spy,

WORK EXPERIENCE / PROJECTS UNDERTAKEN

University Of Michigan, Ann Arbor, MI Jan 2007- Present

Project: Database Development Role: Computer Research Specialist

Responsibilities:

- Development of database using MySQL.
- Integration of database with Matlab and Miner3D.
- Construction of 3D images from 2D scans.
- Working with Visualization tools.

Indiana University Purdue University Indianapolis, IN Jan 2006 - Present

Project: Distributed Drug Discovery (DDD) Role: Database Developer

Responsibilities:

- Development of database for Distributed Drug Discovery Project, using PostgreSQL.
- User interface website development using apache tomcat, html and java.
- Conducted usability survey to improve the design of user interface.
- Developed specific retrieval tools using ChemAxon software inputs to the database.

Indiana University Purdue University Indianapolis, IN May 2006 – Aug 2006

Project: Chemical Facets Role: Facets and Schema Developer

Responsibilities:

- Developed Chemical facets for open source databases.

Indiana University Purdue University Indianapolis, IN May 2005 – Aug 2005

Project: Usability Study Role: Usability Analysis

Responsibilities:

- Conducted usability Study on different chemical informatics drawing tools.

Indiana University Purdue University Indianapolis, IN Jan 2005 – May 2005

Project: SIBIOS Role: XML Schemas

Responsibilities:

- Developed XML schemas for "Center of Excellence for Computational Diagnostics" for open source biological databases.

Indian Institute of Chemical Technology (IICT), Hyderabad Jan 2002 – May 2002

Project: Study of Advanced Bio Reactors

- Study of Up flow Anaerobic Sludge Blanket Reactors, Anaerobic Fluidized Bed Reactors, and Anaerobic Fixed Film Reactors

EDUCATION

- **MS Chemical Informatics** (2004-Present) Indiana University Purdue University Indianapolis (IUPUI), Indianapolis, Indiana, U.S.A. (G.P.A 3.777/4.0)
- **B Tech. Chemical Engineering** (April 2002), J.N.T.University, A.P, India

Affiliations:

- Chemical Engineering Students Association of India.
- Indian Institute of Chemical Engineer

References upon request