Chapter 3  Development of Plant Protease inhibitor database—PIND

3.1: Abstract

PIND (www.mspind.com) is a manually curated database of Plant PIs, human protease target as well as other protease target in infectious diseases that are deposited in PDB and other heterogeneous databases. It is also a database that allows one to browse, classify and manually visualize structure with the help of local software.

I have developed a new database that collects all Plant PIs data and Protease target into a single, easily accessible public resource. The Plant Protease Inhibitor database (PIND) contains annotated sequences and structure, methodological data for more than 1000 Plant PIs and 400 Protease targets. A user-friendly web interface has been developed that allows powerful searching, browsing and information retrieval, whilst providing links to other protein databases. The database has very tight integration of various bioinformatics proteomic base as well as chemoinformatics data resources. Plant PIs –Protease target will be good resource for both better interpretation of structure and the development of better scoring functions to be used in many drug discovery applications.
3.2 Introduction

The completion of the draft sequence of the human genome has heralded a new era in molecular biology by opening avenues for measurement of a large number of genes or proteins in a single experiment. This has led to a huge paradigm shift for life scientists, from the traditional research limited to experimenting with small, specific, and approximate biological models to the newer paradigm of high throughput experiments. These experiments have the potential to revolutionize academic and industrial research and discovery, with breakthroughs in areas such as identifying genetic causes of disease, predicting an individual's response to drug treatment, identifying biological drug targets, and deepening the basic understanding of evolution and workings of biological organisms. However, their still exist tremendous problems in data and information integration and management before the potential of these technologies can be realized.

The new experimental paradigm has led to an emergent data-centric approach to research in molecular biology. The emphasis is on mining the large amounts of data generated by these high throughput experimental studies to yield specific targets for further research. This approach is complementary to the traditional hypothesis-centered research, where biologists approached their research problems with specific hypotheses, and experiments were conducted to refute or support these. In the traditional hypothesis-centered paradigm, biologists worked within the realm of their domain expertise, where they had sufficient knowledge to develop and test various hypotheses. However, the new data-centric approach is forcing biologists to move out of narrowly focused areas to work more broadly with large numbers of genes or proteins that they are otherwise unfamiliar with, i.e., they are forced to incorporate broader range of scientific knowledge drawn from
disciplines beyond their own. A biologist now has the need to draw insights across multiple data (e.g., comparing proteomic vs. gene expression information) and data types (e.g., comparing information from scientific text, pathway diagrams, and experimental data). Clearly, no one central repository (either technological or human) exists with all the biological knowledge. Thus, the challenge lies in providing biologists with easier access to these highly fragmented and distributed sources of data and information, and a means of viewing, navigating, and synthesizing this information.

A key component of creating the public archive of information is the efficient capture and curation of the data-data processing. Data processing consists of data deposition, annotation and validation. These steps are part of the fully documented and integrated data processing system. This integrated system helps to ensure that the data submitted are consistent with the file which defines data types, enumerate ranges of allowable values where possible and describes allowable relationships between data values.

One major source of knowledge is scientific literature. For generations the research community has represented and shared its knowledge in the form of publications. More recently these are being shared freely in a digital and computationally accessible format (mostly in the form of abstracts, such as the PubMed repository at NCBI (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed) and possibly as full articles in the future). With the explosion in molecular biology literature (for example, there are over 14 million abstracts in PubMed as of March 2007), searching for relevant information is becoming extremely difficult with the user left to manually sift through large number of retrieved documents. Therefore, simple access to digital forms of text
isn't sufficient for analyzing high throughput experimental data and a number of research groups have started to address the problem of automatic information extraction (IE) from text (Collier, 2000; Friedman, 2001; Fukuda, 1998; Humphreys, 2000; Iliopoulos, 2001; Krauthammer, 2000; Ng and Wong, 1999; Palakal, 2002a, 2002b; Park, 2001; Rindflesch, 2000; Sekimizu, 1998; Stephens, 2001; Wong, 2001; Yakushiji, 2001.) However, there are several limitations of automatic text extraction. It behaves as a black box to an end user not knowledgeable in the text extraction domain, its accuracy and efficiency are not easily computable, and it often fails to capture a user's true context (in that its behavior is uniform across all users). Moreover, automated extraction systems are very specific in terms of the problem they are solving. A study of biologists in pharmaceutical industry and academic research labs indicates that most biologists do not completely trust automated text extraction systems, for the above-mentioned reasons (O'Day, 2001).

Manual extraction of information by a user, on the other hand, is true to the user's intent, the user trusts it, and in this method the user has a better understanding of the accuracy of the extracted information. Also, capturing the user's context is implicit with manual extraction. However, manual extraction is tedious and time consuming, and increases in difficulty as data and models grow in complexity. Moreover, it becomes potentially impossible to manually extract information in a high throughput manner. Therefore, in order to overcome the limitations and utilize the advantages of the automated and manual approaches, we propose an interactive, user-guided approach to information search and extraction. The following verbal quote from a biologist in the Text Mining panel discussion at the Pacific Symposium on Biocomputing 2002 (PSB 2002), at Kauai,
Hawaii, succinctly captures the motivation behind our approach. ‘Biologists want interactive tools to help them do their job better and not fully automated tools whose reliability is hard to judge. Thus text mining should not be an end product, but an interacting tool allowing users to extract information they are interested in.’

Finally, extraction of information from text is not an end in itself. Biologists need to relate the extracted information with other sources of information, such as experimental data or diagrammatic biological models. In other words, tools are needed for (i) robust and reliable extraction and representation of information from text, ideally in a high throughput manner; (ii) providing biologists access to the relevant and interesting information based upon the context of their experiments or research; (iii) incorporating the relevant information in experimental data analysis and model generation; and (iv) linking the relevant information, experimental data, and models for reuse, collaborative sharing, and knowledge synthesis. In this manuscript, we describe our efforts to address the above issues via a suite of integrated, interactive, user-guided online software tools. Our approach utilizes some of the more reliable information retrieval and extraction technologies to automatically query multiple information repositories, filter the retrieved results based on the user’s interests, and identify relevant information (what we refer to as potential entities (or concepts) and their interactions (or relations)) in the filtered text corpus. However, rather than completely automate the extraction process, we provide the users interactive tools to guide the extraction process. We have further developed architecture to represent the extracted information in a hierarchical hyper-graph data structure. This architecture, referred to as SAM, allows for easy transformations between
textual, experimental, and diagrammatic biological data by providing means for a standardized and structured representation of information present in these data sources.

I have developed a Plant Protease Inhibitors and proteases target database and data warehouse called PIND. This database locally stores and integrates with biological sequence, molecular interaction, homology information, functional annotation of PIs with CAS. The goal of the database is to provide data, as well as a web based software infrastructure for chemoinformatics and bioinformatics research and development. We have tried to focus Plant Protease Inhibitors as one of the tool for drug designing and discovery.

PIs are of great relevance to biology, medicine and biotechnology. The chemoinformatician often encounter problems in accessing huge amount of widely spread data due to disparate formats, remotely dispersed and varying implementation on different platforms and data models. Deposition of the information related to Plant protease inhibitors and its relevant proteases is growing fast. Recently PIs are being used against human diseases. Our database includes the information related to it. We have tried to correlate PIs with drug development with the help of PIs Card. Plant PIs were linked up to CAS registry. This practical importance creates a need for an integrated source of information about them, and also about their proteases and use of it in CAS code.

The PIND database (www.mspind.com) aims to fill this need. PIND currently integrates twelve different databases, including PDB, KEGG, Swiss-Prot, CATH, Pfam, SCOP, NCBI, MEROPS, UniProt, different plant informative web site, Gene Ontology, enzymes, GenBank, Human Protein Reference Database (HPRD), Database of Interacting
Proteins (DIP), Molecular Interactions database (MINT), HomoloGene and GO. We describe PIND, a public database of PIs and protease available on the web. PIND is built using the digital information of over hundreds of web site and other public source as well as from computational methods, such as prediction with the help of three-dimensional structure.

The PIND system is based on relational data models that we developed for each of the source data types. Data stored within these relational models are managed through structured Query Language (SQL) calls that are implemented in a set of Application Programming Interfaces (API). The API includes Perl, Java Script, HTML and DHTML. With the help of manual information extraction API are constructed database. The retrieval APIs, toolbox and PIs Card applications are critical components that offer end-users for flexible, easy, integrated access to this data. We present cases that use PIND to integrate these sources for PIs annotation and drug designing, which is the prime object of developing this database.

High throughput screening of Plant PIs is being carried out for a member of drug targets. For such studies, it is important to document the availability and location of such Plant PIs in the form of database for such studies. Here we have described a web database containing information (Name, common name, chemistry, Related Protease, use against drug target etc.) about the Plant PIs available on online at present the database is available for free in public domain.

One important goal in chemoinformatics as well as in bioinformatics is to develop new tools to integrate different (heterogeneous) web site databases on the basis of architecture from disparate sources of biological information. With the help of manual and automated
data extraction, we have developed some important tools for PIs data presentation like Family Card, PIs Card, Drug base PIs data warehouse, Organism base, PIs base, Protease base, which are useful for drug design. This whole database is integrated internally as well as externally with cross-links to each other like- Plant protease inhibitors-protease-organism (source of it)- family- literature- PIs SMART CARD -Family info-CAS-ChemDB-ACD- PDB-SCOP-Pfam. This will help in the use and exploitation of the Plant materials for drug designing and discovery.

3.3 Database Description

Fundamentally, PIND is a dual-purpose bioinformatics-chemoinformatics database with a strong focus on quantitative, analytical or molecular-scale information about Plant PIs and Protease targets. Actually it is fusion of interdisciplinary approach. In many respects it combines the data-rich molecular biology content normally found in curated sequence databases such as Swiss-Prot, UniProt, Entrez (Bairoch 2005) with equally rich data found in medicinal chemistry journal, and chemical reference handbooks. By bringing these two disparate types of information together into one unified and freely available resources, we wanted to allow educators and researchers from divers disciplines and backgrounds (academic, industrial, clinical, non-clinical) to conduct the type of in silico learning and discovery that is now routine in the world of genomics and proteomics.

The diversity of data types and the required breadth of domain knowledge, combined with the fact that the data were mostly ‘paper-bound’ made the assembly of PIND both difficult and time-consuming. To compile, confirm and validate this comprehensive collection of data, more than a dozen textbooks, several hundred journal articles, nearly
more than 75 different electronic databases, and at list 30 web-based programs were
individually searched, accessed, compared, written or run over the course of 2 ½ Years. I
got help from different colleague who are working in different areas like archivists,
annotators, Molecular Modeler, pharmacists, a physician and bioinformaticians with
training in computing science and molecular chemistry also in chemoinformatics at IICT,
Hyderabad.

PIND currently contains >600 Plant protease inhibitors, corresponding >400 protease
target and their general name and synonyms. These Plant PI entries were chosen
according to the following rules: the Plant PI must be non-redundant, have a known
structure. To facilitate more target research and exploration, PIND is divided into four
major categories (i) Serine protease inhibitors (ii) cysteine protease inhibitors (iii)
Aspartic protease inhibitors (iv) Metallo protease inhibitors (v) experimental Plant PIs.

PIND is a fully searchable web-enabled resource with many built-in tools and feature for
viewing, sorting and extracting Plant PI and Protease data. Detailed instruction on where
to locate and how to use these browsing / search tools are provided on the PIND
homepage. It also offer general database browsing using the ‘Browse’ and ‘Plant PIs
Browser’ buttons located at the top of each PIND page. To facilitate general browsing,
PIND is divided into synoptic summery tables, which, in turn, are linked, to more
detailed “PLANT PIs CARD” –in analogy to the very successful GeneCards concept. In
addition to the general browsing feature, PIND also provide a more specialized “PLANT
PIs Browser” feature. This is designed for pharmacists, physicians and medicinal
chemists who tend to think of drug.
A key distinguishing feature of PIND from other online Plant PI resources is its extensive support for higher-level database searching and selecting functions. This database mainly focusing Pharma industry base support.

3.4 Architecture

![Diagram showing data sources and PIND database retrieval process]

MANUALLY CURATED PIND DATABASE FROM ABOVE DATA SOURCE WITH THE HELP OF ONLINE AS WELL AS OFFLINE SOFTWARE METHODS

PIND DATABASE

Retrieval

SQL

Application Programming Interface (API)

PERL  JAVA  SCRIPT  HTML  DHTML  ASP
I have proposed a simple architecture model (SAM) to shift through vast volumes of heterogeneous data. We have developed architecture, referred to SAM for qualitative representation of biological information. The goal of SAM is to capture and represent, in a structured manner, information from free form text, experimental data, and diagrammatic models by defining a common hierarchy. While the emphasis of SAM is to provide a means for representing information present in various sources of biological data in an abstract.
Fig 3.3: Unification of all types of data

Fig 3.4: Linking of various database with the help of Manual linking
Fig 3.5: Workflow to find drug target and Drug

3.5 Data Source Format and Size

PIND is a Plant PIs information database system grown not only out of an aggregation of multiple information sources, primarily publicly available repositories e.g. Entrez, PDB, DDBJ, EBI, DrugBank etc. Data is Manually extracted from different online databases. We manually download data and resynchronize the latest updates into PIND. We try to gather information from different data sources also like SMILES, MDL Mol, PDB, Tripose sybyl mol2 and SDF, is easily accomplished using OpenEye software’s OEChem toolkit (www.eyesopen.com), or the open source alternative, OpenBabel. Additional curation and normalization steps are applied to the data as it is inserted.

3.6 Database Schema
The basic database schema is relationally organized on the bases of hierarchy and enforced the uniqueness. There is many-to-many relationship between sources. The database schema contains primary tables for sources, molecular descriptors and annotations.

3.7 Implementation

The database is implemented using the leading relational database for initially with MS Access. Web interface and tools are delivered using the Microsoft window 2000 server. Many of the basic application tools, scripts and web interface are written in ASP (Active server page), while computationally intensive modules are written in CGI-PERL. PERL has convenient interface to important package. Several basic algorithms, which are needed for data processing, we develop at our lab.

3.8 Web Interface

The current paradigm for data output and presentation to the user is the Internet / Web interface. This paradigm is the best solution to date to issues associated with a "federation" of databases: the Web interface is basically with back end server means database and front-end software independent. I had used MS Access database system with ASP database connectivity for Web interfaces

I have developed Front end with the help of HTML, DHTML, JAVA SCRIPT, PERL SERIPT, CGI-PERL. Dream weaver I used for designing Home Page, which is giving, automated designing facility with HTML and other scripting.
The Web interface however is still cumbersome and not always satisfactory. Web design is always an issue, and it is difficult at times to program what one wants, for example, appropriate links cannot always be programmed. Links change and new appropriate Web sites become available at an increasing rate, presenting updating and maintenance issues. 3D protein structure representations mainly use web browser plugin modules, such as RasMol or Chime. These modules permit several representations of the protein structure properties, as well as both transnational and rotational motion capability, slab cutthroughs, etc.

### 3.9 Data Integration

PIND is implemented on MS Access. It integrates data from the different twelve data sources shown in table . The data from the original sources are available in different formats, such as flat files, database dump files, or pure HTML pages.

**Table 3.1 : Data sources integrated in PIND**

<table>
<thead>
<tr>
<th>Sr.No.</th>
<th>Source</th>
<th>Web site link</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PDB</td>
<td><a href="http://www.rcsb.org/pdb">http://www.rcsb.org/pdb</a></td>
</tr>
<tr>
<td>2</td>
<td>SCOP</td>
<td><a href="http://scop.berkeley.edu">http://scop.berkeley.edu</a></td>
</tr>
<tr>
<td>3</td>
<td>CATH</td>
<td><a href="http://www.biochem.ucl.ac.uk/bsm/cath">http://www.biochem.ucl.ac.uk/bsm/cath</a></td>
</tr>
<tr>
<td>4</td>
<td>DSSP</td>
<td>Computed</td>
</tr>
<tr>
<td>7</td>
<td>KEGG</td>
<td><a href="http://www.genome.jp/kegg">http://www.genome.jp/kegg</a></td>
</tr>
</tbody>
</table>
3.10 Utility

PIND is relational, integrated database of Plant Protease inhibitors and Protease as a drug target which very important in Drug development and designing. In Pharmaceutical industries it is very essential. This is one of strategy for development of drugs. I tried to provide one integrated information resources for chemoinformatician as well as computer aided drug designer and also for synthetic drug designer.

3.11 Discussion

The most frequent approach to the interconnection of data on Plant Protease Inhibitors and Proteases that are spread over multiple original data sources is the usage of hyperlinks. Examples are Plant PIs and Merops. This methods is well suited for browsing of single entries, but as soon as it come to handling sets of objects, following many hyperlinks becomes a tedious and time consuming task. Efficient handling of sets can only be achieved if data are physically integrated into a single system. In the genomic and proteomic world there are different databases. NCBI focus on providing sequence, literature and some structure not full-fledged information. EBI and DDBJ are doing also
same. If we are looking in proteomic database like 3DinSight focus on visualization of sequence feature. iProClass concentrate on protein sequence and integrates 50 different databases using so-called ‘rich links’. Finally BioMolQuest integrates in total four data sources. But nobody is provided drug and drug target base data integration or database here I had tried to develop one of the rich proteomic base drug and drug target manually curated database.

PIND currently integrates more than 20 data sources concerned with different aspects of genomic and proteomic according to the Protease as a drug target and Plant protease inhibitors is one of the aspects for drug. PIND doesn’t store all information initially but after primary architecture of it, I am going to develop one of automated system for this.

An important design principle of PIND is that it never mixes data from different sources into a single table. Each data source is considered as a dimension in which PDB entries, compounds and chains are annotated. We call this approach multidimensional data integration (chaudhuri 1997), which is inspired by data warehouse design, where as facts that e.g. sales, are described by dimensions, such as store, product, or customer (chaudhuri 1997). The resulting database schema is called star schema in correspondence with the visual appearance. I also use a star schema like structure with the tables holding information from Plant PIs in center of a set of tables containing the data from other data sources.

I strongly believe that merging data from different databases is counterproductive for the biologist because it blurs important difference. On the other hand, keeping data separated inevitably leads to a certain degree of semantic redundancy, i.e. different schema
elements provide the same type of information. For instance, functional annotation of proteins is encoded both in Swiss-Prot keywords and Gene Ontology terms. “TIM barrels” are annotated in CATH, SCOP, and PDB annotation itself. But this redundancy does not originate from data duplication, but rather from evidence obtained independently by different people or by different experiments. These evidences are important in their own right.

I believe that the advantages of our approach prevail for mainly two reasons:

- Users recognize the origin of the data they query and obtain as result. In my experience, biologists often have their favorite set of database, where they know about the pitfalls and peculiarities. By keeping data separated, personal performance or differences in trust in particular databases can be expressed and the results can be judged based on prior experience.

- Subtle differences in the semantic of fields of different databases are conserved. For instance, both Swiss-Prot keywords and GO annotations express functional annotation. However, the process of creating this annotation is quite different, and it is often meaningful to discriminate between the two.

Furthermore, separating data and software for the different data sources greatly simplifies system maintenance. Changes to data sources, only affect a well-defined part of the schema and of the web interface.

Some people are using ‘reverse star schema’ to connect genes with different types of information, such as genomic position, transcription factors, or expression data. The data...
are queried through a generic web interface, which also allow source-specific queries and their combinations. Conceptually PIND is working on this. PIND is directly designed for handling annotations of genomic and proteomic data, which has advantages in term of result.

3.12 Conclusion

PIND will prove to be very useful for number of tasks in drug designing and discovery research. Generating information, which previously required days of manual browsing, now only take a few mouse clicks, or an SQL query. Once the Plant PIs and Protease is obtained, there are many other programs to compare sequences and structures. PIND’s future development will be further concentration on automated data curation and comparison. The next data sources to be integrated are those covering protein domains and motifs, i.e. InterPro (Apweiler 2001) and its relatives. In the long run, I will push PIND towards a medical orientation. Obvious candidates for being integrated are literature abstracts from Medline and the OMIN database (McKusick 1998). The LIGAND database (Goto 2002) will provide information about small molecules interacting with proteins to use PIND for the prediction of drug target sides. Moving towards medical data is a natural next step since much of structural research, including my own is concerned with drug development.