

## CHAPTER ONE: INTRODUCTION

### 1.1 Motivation

Data integration, data mining and methods for computational analysis facilitate biological research by providing easy access to a wealth of biological information. While individual databases focus on certain areas of biological knowledge, integrated and comprehensive use of multiple datasets may provide new insights into research and discovery. Biological pathway databases are of special interest since they link together molecular entities with metabolic, transduction and regulatory events, and integration may work to expand analytical power.

### 1.2 Introduction to Pathways

The study of bio-molecular pathways is essential in systems biology [1]. A pathway refers to a series of biochemical reactions which are linked by having the product of one reaction be either a reactant of a subsequent reaction, or an enzyme that catalyzes a subsequent reaction. There are three major classes of pathways: metabolic pathways, signal transduction pathways, and gene regulatory pathways. Each pathway connection can be characterized as the collection of component molecules (DNA, genes, proteins, snRNAs, metabolites, and drug compounds) and component molecule reaction/interactions. Metabolic pathways usually consist of a series of chemical reactions that provide basic biochemical functions to maintain metabolite/protein synthesis and energy metabolisms in cells. Signal transduction pathways act to send signals between cellular locations. For example, signal transductions are found to occur from cell membrane to cytoplasm and from cytoplasm to nucleus. Gene regulatory pathways are responsible for converting genetic information into proteins (gene products)

and controlling when and how this information is released in response to intracellular signals. Understanding what these pathways are and how they relate to each other represents a major step towards simulating biology in silico and devising engineering solutions to treat complex human diseases.

While there are more than 196 online pathway databases of different coverage and quality as of September 2007 according to Pathguide [2], our knowledge of human pathways is still quite incomplete. For example, BioCarta [3], an open-access expert-curated pathway illustration database for humans, contains 254 pathways, 2,308 proteins, 205 compounds, and 880 complexes, and 3,064 interactions as of its June 2004 release. The scale of these counts is comparable to those reported in other popular curated human pathway databases such as KEGG [4] and Reactome[5]. The limited content of these databases suggests that there is still a large gap with at least an order of magnitude difference between the “annotated” portion of human pathways and the “un-annotated” portion of human pathways.

In order to make a full account of annotated pathways across the existing set of pathway databases, a key step is to develop tools and resources to accomplish the integration and analysis. Challenges include the matching of pathway molecular entity names, interaction/reaction relationships from heterogeneous pathway databases, and the “high level noise” inherent in pathway data generated from text mining or computational predictions. For example, in the NCI-Nature curated database [6], component and interaction data is provided, whereas pathway regulation data are not reported. The commercially available pathway software, Protein Lounge [7] , only provide users with pathway component lists in searchable text whereas pathway interaction and regulation

information is embedded in pathway image files and is non-searchable. These database resources also represent pathway entities (molecular entities and interaction/regulations) in different formats, often incompatible with PSI-MI or BioPAX standard pathway exchange formats, except for a few recent tools such as cPATH [8]. Even when the formats can be managed and merged at the syntactic level, semantic level incompatibility still exists. For example, pathway molecules can be often represented both in theory and in practice with any type of public database identifiers, e.g., NCBI Gene ID, Ensembl Database ID, Gene Symbol, NCBI GB Accession Number, SwissProt ID, UniProt Database ID, and IPI number. The heterogeneity of the pathway data sources, incomplete coverage of each pathway database, and syntactic as well as semantic level incompatibility among these pathway databases, have all contributed to the current lack of high-coverage integration of pathway data for large-scale systems biology studies where coverage is essential.

In this work, I aim to develop a semantically integrated comprehensive human pathway database supporting a searchable web interface. This database, the Human Protein Database (HPD), was developed to collect pathway data from multiple quality pathway sources. My particular focus was to prioritize data collection for human signaling pathways. The pathway databases that were chosen for integration are NCI-Nature Curated data [6], Biocarta [3], Protein Lounge [7] and Pathway Studio [9]. A comprehensive entity-relationship (ER) data model was built with data warehousing techniques to facilitate the semantic-level integration of data. With HPD, I have constructed a high-coverage human pathway database with integrated information of molecules, complexes, regulation relationships of molecules, and reactions involved in

signaling and regulatory pathways. HPD provides an integrated view of current pathway data from both annotated and predicted resources – 1,895 pathways and 10,631 molecular entities. Furthermore, I analyzed similar pathways sharing components, and developed methods for merging similar pathways. The usability and feasibility of HPD is validated by case studies on Alzheimer’s disease. A prototype web interface of the HPD found online at <http://discover.uits.indiana.edu:8340/pathway3>.

### 1.3 Contributions of the Thesis

There are two major contributions of this thesis:

#### 1. Technical Contributions

- A. Present a novel database framework enabling an integrated view of pathway data in web-based environment to provide a view of:
  - Proteins, complexes, compounds and reactions involved in the pathway,
  - Pathway diagram,
  - Kinase – disease associations and
  - Effect of environmental factors on pathways.
- B. Develop a prototype web interface to query the integrated database.
- C. Analyze the merging of similar pathways based on associated gene/protein identifiers to better provide the user with comprehensive and non-redundant information.

## 2. Analytical Contributions

- A. Pilot the discovery of potential biomarkers by integrating pathway data with network analysis and gene expression data.

### 1.4 Organization of the Thesis

This thesis is divided into six major chapters. Chapter Two of the thesis is a background and literature review about signaling pathways. Also describes and compares existing signaling pathway databases and source database statistics. Chapter Three describes the methodology of constructing HPD including the road map, architecture, data integration, user interface and method for merging together similar pathways. Chapter Four summarizes the results. Chapter Five presents the case study 1 about Alzheimer's disease and cases study 2 about Tumor Necrosis Factor-r- and Interleukin-1-Induced Cellular Responses. Finally, Chapter Six concludes the thesis with a discussion.

## CHAPTER TWO: BACKGROUND

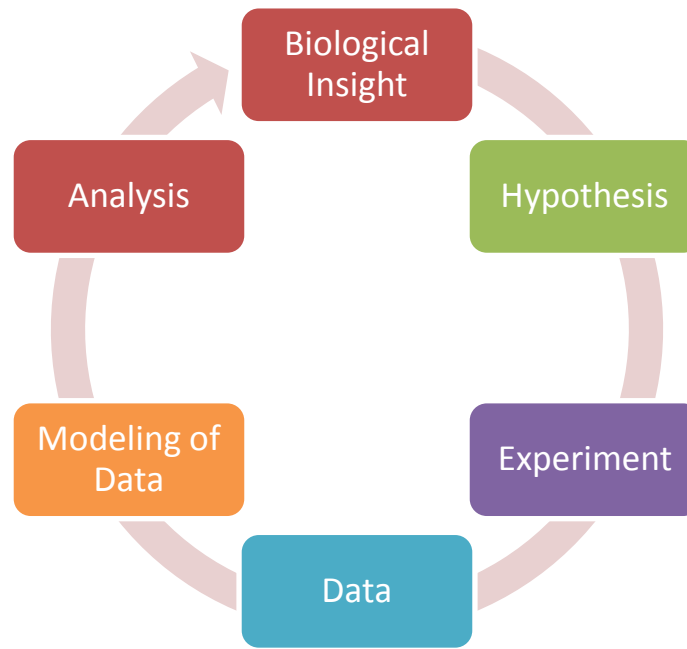
This chapter introduces topics that are essential towards a better understanding of the problem domain and the solutions proposed in this thesis. This chapter starts with systems biology and its importance in research and discovery. The latter part of this chapter introduces pathway databases and ends with a problem definition and related research questions.

### 2.1 Systems Biology

Recent advancements in data analysis and approaches for generating hypotheses in the biological domain are leading to more insights into research and discovery. A traditional approach to studying biology and human health is to investigate individual proteins and genes one at a time, to understand their functionality and their contribution towards a specific functional aspect of the organism. This leads to a limited understanding of how the human body operates, and how we can best predict, prevent, or remedy potential health problems. We have had limited success in curing complex diseases such as cancer, HIV, and diabetes, and investigative approaches are being changed to investigate the behavior and relationships of the many elements in a biological system. Collection and comparison of biology entities requires the ability to analyze different data sets as an integrated view. It aims to support the functional genomics, proteomics and other systems-based data sets through global examination of networks of genes, proteins, metabolites, cells, and tissues.

Systems biology is the study of an organism based on integrated and interacting network analyses of genes, proteins and biochemical reactions. Systems biology can be viewed as a cyclical process consisting of laboratory experiments, data generation from

experiments and the collection of data, followed by data analysis leading to biological insight and finally to hypothesis and further experimentation. The systems biology process is illustrated in Figure 2.1.



**Figure 2.1 Systems Biology - A cyclical process.**

The study of systems biology is aided by advances in new research technologies for:

- **data generation** - microarrays, DNA sequencing, quantitative proteomics, and mass spectrometry analysis;
- storing and distributing massive amounts of data through internet;
- **extracting** useful information from the data - laboratory information management systems (LIMS), bioinformatics pipelines, database frameworks; and
- **analysis and data visualization** - Cytoscape, ProteoLens.

## 2.2 Literature Review

Signaling pathways represent the cascade of information from plasma membrane to nucleus in response to an extracellular stimulus. In general, extracellular signaling molecules bind to specific intracellular receptors and initiate the signaling pathway. Here we have taken the NF- $\kappa$ B pathway as an example to explain the signaling pathway. NF- $\kappa$ B pathway activation is shown in Figure 2.2. Inflammatory signals, mainly TNF $\alpha$ , IL-1 or Toll, bind to their corresponding receptors and lead to activation of TAK1 by recruiting receptor-associated proteins, such as MyD88 and IRAK for IL-1R/TLR, TRADD and RIP1 for TNF receptor. In turn, these associated proteins recruit TRAF2 or TRAF6, both of which are recruited possibly through a non-destructive G76-K63 polyubiquitin chain-dependent mechanism (Ub<sub>63</sub>).

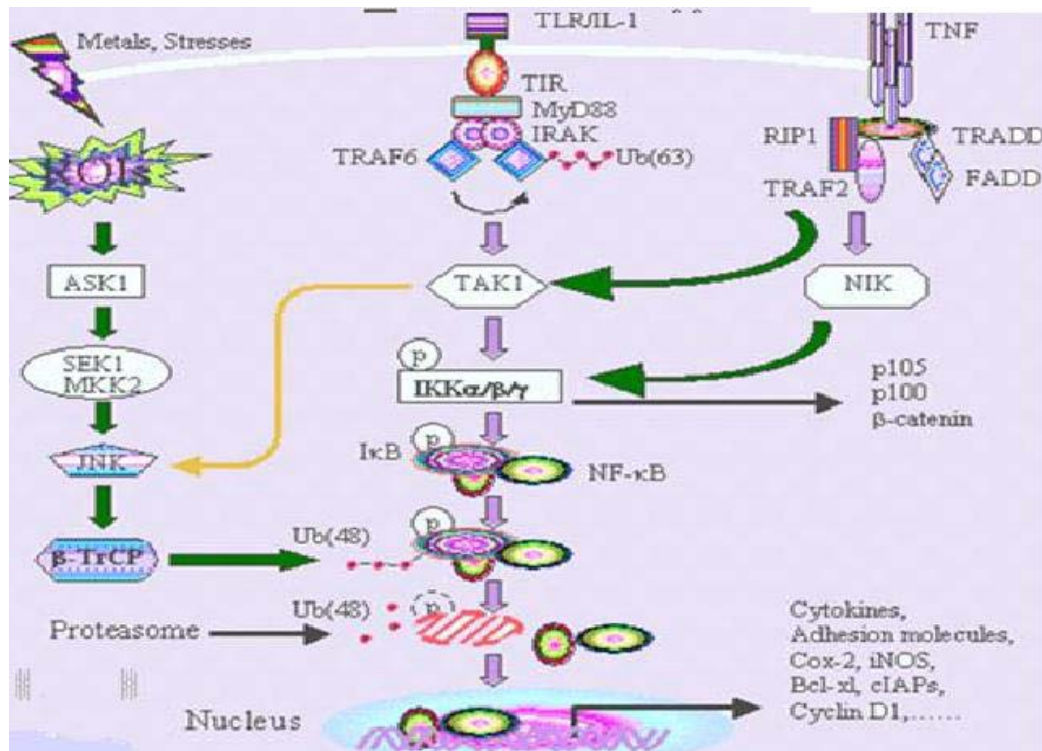


Figure 2.2 Simplified signal transduction pathways of NF- $\kappa$ B activation [10].



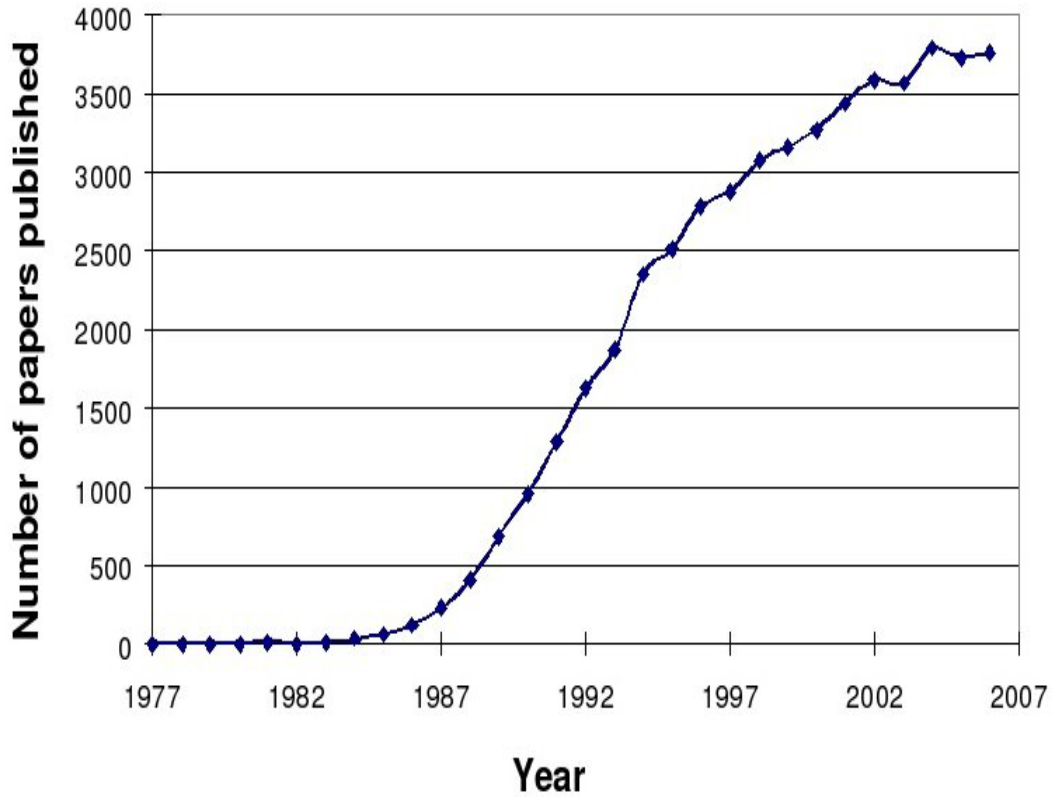
Activated TAK1 or other MAPKKK family kinases, such as NIK and MEKK1, phosphorylate and activate IKK complexes which phosphorylate the I $\kappa$ B protein. Phosphorylated I $\kappa$ B proteins are recognized and modified by the G76-K48 polyubiquitin chain (Ub48) via the SCF- $\beta$ -TrCP complex. This leads to the proteasome-mediated degradation of I $\kappa$ Bs. Stress signals result in sequential activation of ASK1, SEK1 and JNK. Activated JNK induces the accumulation of  $\beta$ -TrCP protein, which facilitates the ubiquitination process of I $\kappa$ B proteins contributing to the activation of NF- $\kappa$ B. The next section is about the growth in the signal transduction publications and data.

#### Growth of Data for Signal Transduction

The earliest published article recorded in the MEDLINE database containing the term "signal transduction" was published in 1972 [11]. Before 1977, most published articles had the term "signal transmission" or "sensory transduction" [12, 13]. After 1977, published articles began to appear with the specific term "signal transduction" , and in 1979 this specific term appears within a paper title [14, 15]. In 1980, there was a review article by Rodbell [16, 17] with the extensive use of the term signal transduction. The total number of papers published in each year since 1977 with the term signal transduction in the title or abstract section are plotted in Figure 2.2. These numbers were extracted based on the papers contained within the MEDLINE database. The total number of scientific papers related to signal transduction published from 1st Jan 1977 up to the 31st December 2007 was 48,377 of which 11,211 were reviews.

In the early 1990s, the research papers directly addressing signal transduction processes began to appear in large numbers in the scientific literature. There are a number

of landmark or important discoveries in the field of signal transduction, such as the link made by Rodbell between metabolic regulation and the activity of GTP and GTP-binding proteins [16]. Most of our current knowledge of signal transduction is as a result of numerous contributions made to the field over many years by different research groups. As seen from the Figure 2.3, there is an exponential growth from the early 1990s.



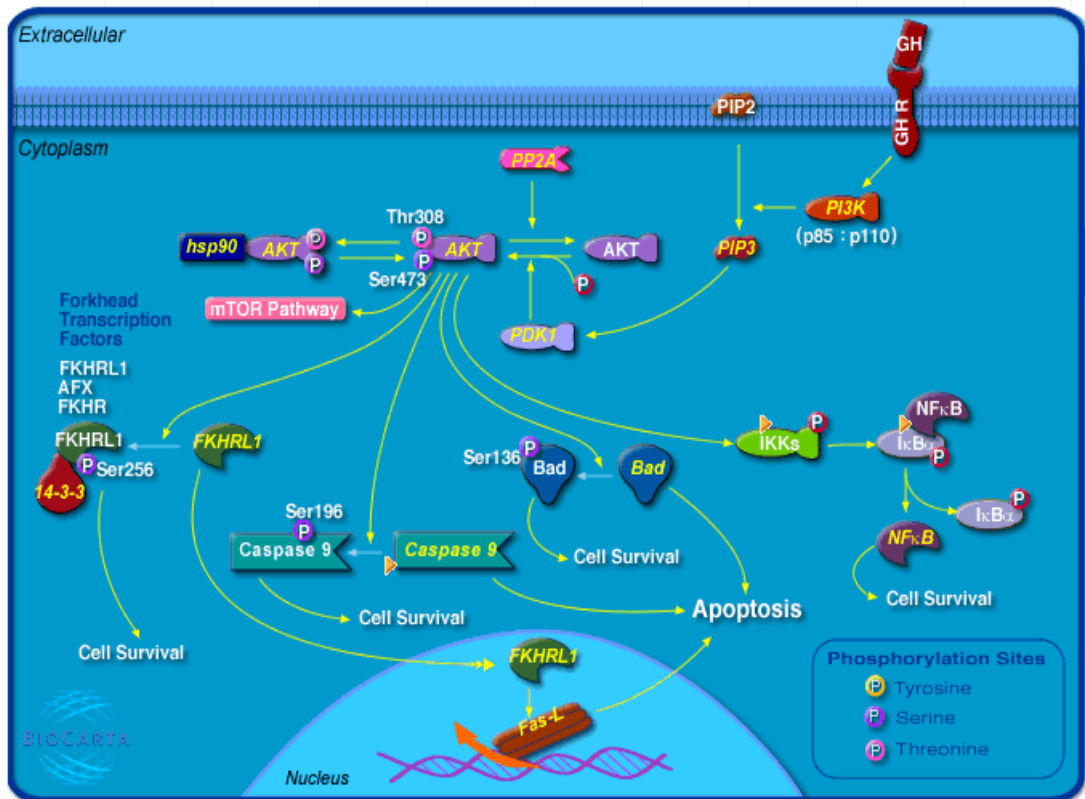
**Figure 2.3 Signal transduction publications graph.**

More publications can generally indicate more data, and the amount of data for signal transduction pathway analysis is reflected by the presence of 41 different signal transduction pathway databases. In the next section we talk about the existing pathway databases.



Biocarta

Biocarta represents how proteins interact in dynamic graphical models and also gives pathway descriptions along with references. As such, there is no downloadable data from Biocarta. But the Pathway Interaction Database (PID) contains a June 2004 snapshot of pathway data at the BioCarta web site [3] without additional expert review. Pathway molecules are annotated by NCBI Gene ID without associated post-translational modifications. There are 254 signaling pathways in Biocarta.

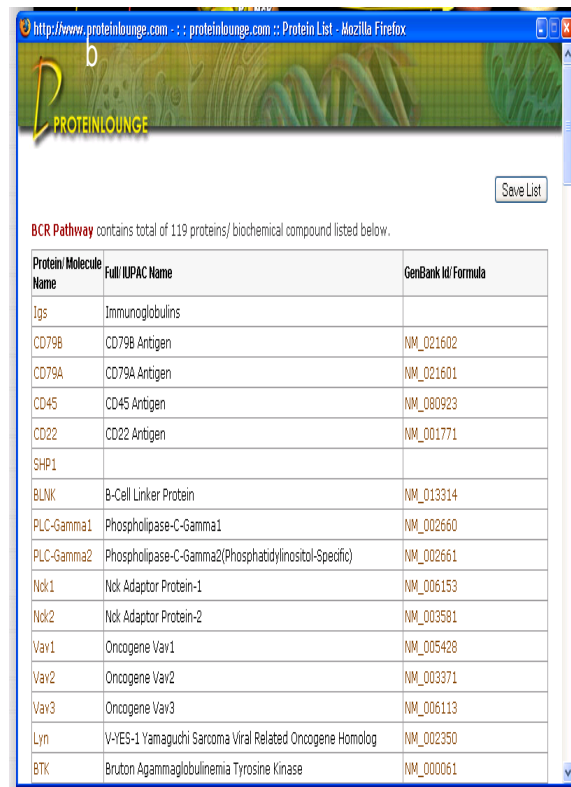


**Figure 2.5 Schematic diagram of pathway in Biocarta.**

Browsing Biocarta and NCI-Nature curated data pathways is similar on the Nature PID web site. In Biocarta, these pathways can be browsed through clickable images (Figure 2.5).

## Protein Lounge

Protein Lounge [7] is a commercially available database of curated signaling and metabolic pathways for all organisms. Pathway data was collected from a licensed version of the Protein Lounge database to Indiana University. The pathways include detailed pathway items and pathway diagram drawings. Pathway molecules derived from Protein Lounge are identified by NCBI GB or GI accession numbers. The effect of some environmental factors like UV, radiation and stress on pathways is also collected. There are 427 signaling pathways in Protein Lounge. Screenshots of Protein Lounge pathways and pathway items are shown in Figure 2.6(a) and Figure 2.6(b) respectively. However, these items do not provide downloadable data.



The screenshot shows a web browser window with the URL <http://www.proteinlounge.com>. The page title is "Protein List - Mozilla Firefox". The main content area displays the "BCR Pathway" which contains a total of 119 proteins/biochemical compounds. A table lists the following items:

Protein/Molecule Name	Full IUPAC Name	GenBank ID/Formula
Igs	Immunoglobulins	
CD79B	CD79B Antigen	NM_021602
CD79A	CD79A Antigen	NM_021601
CD45	CD45 Antigen	NM_080923
CD22	CD22 Antigen	NM_001771
SHP1		
BLNK	B-Cell Linker Protein	NM_013314
PLC-Gamma1	Phospholipase-C-Gamma1	NM_002660
PLC-Gamma2	Phospholipase-C-Gamma2(Phosphatidylinositol-Specific)	NM_002661
Nck1	Nck Adaptor Protein-1	NM_006153
Nck2	Nck Adaptor Protein-2	NM_003581
Vav1	Oncogene Vav1	NM_005428
Vav2	Oncogene Vav2	NM_003371
Vav3	Oncogene Vav3	NM_006113
Lyn	V-YES-1 Yamaguchi Sarcoma Viral Related Oncogene Homolog	NM_002350
BTK	Bruton Agammaglobulinemia Tyrosine Kinase	NM_000061

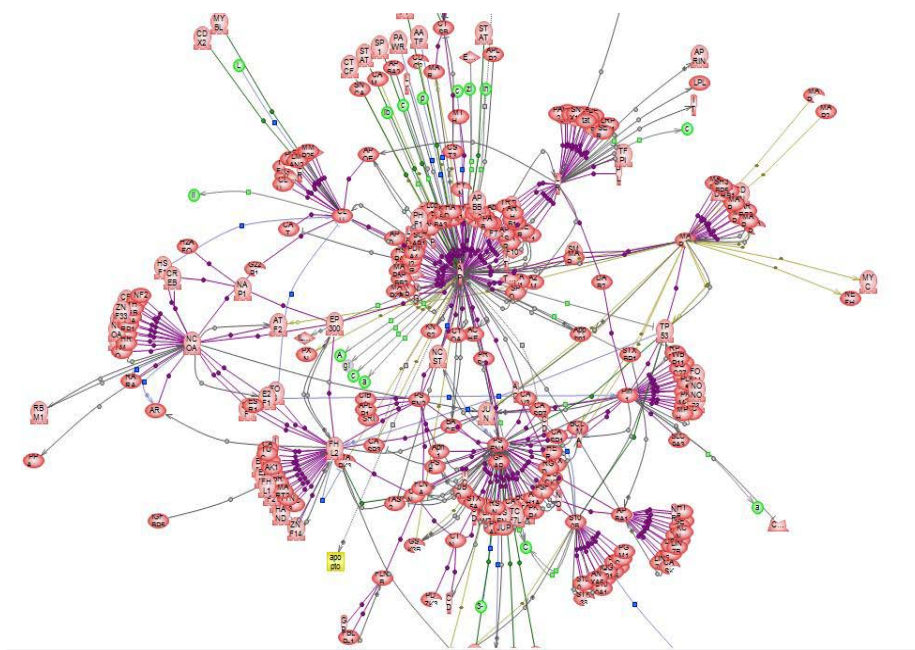
**Figure 2.6 (a) Schematic diagram of pathway in protein lounge. (b) Pathway items.**

## Pathway Studio/ResNet

Pathway Studio/Resnet is commercial software developed by Ariadne Genomics, Inc. ResNet is a pathway database extracted from PubMed using Med Scan and natural language processing tools from Ariadne Genomics. Simple interaction types between molecules, such as regulation, expression, transport and protein modification, are given.

In the Resnet Database:

- entities are linked together by relations;
- both entities and relations are annotated with properties; and
- content and scope of the database can be extended by the user.



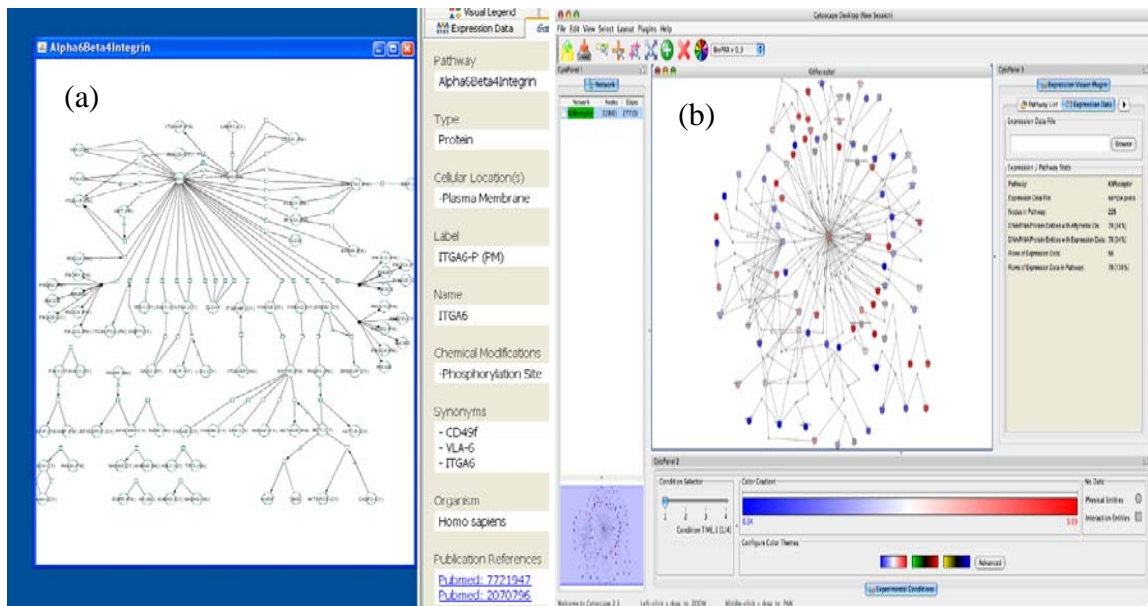
**Figure 2.7 Pathway generated using pathway studio.**

Pathway data generated using pathway studio is shown in Figure 2.7. A trial version of pathway studio, ResNet 4.0, was downloaded for this project during April 2007. There are about 1,132 pathways present in ResNet[18]. This trial version data has partial information about the pathways.

## The Cancer Cell Map

Cancer Cell Map [19] defines a pathway as “a collection of all genes/proteins that have been described as pathway members in any publication and all the interactions between them that can be found described in the literature.” Cancer Cell Map is a collection-selected set of human cancer focused pathways. Cancer Cell Map curated the pathways which are of scientific interest to research laboratories at Memorial Sloan-Kettering Cancer Center.

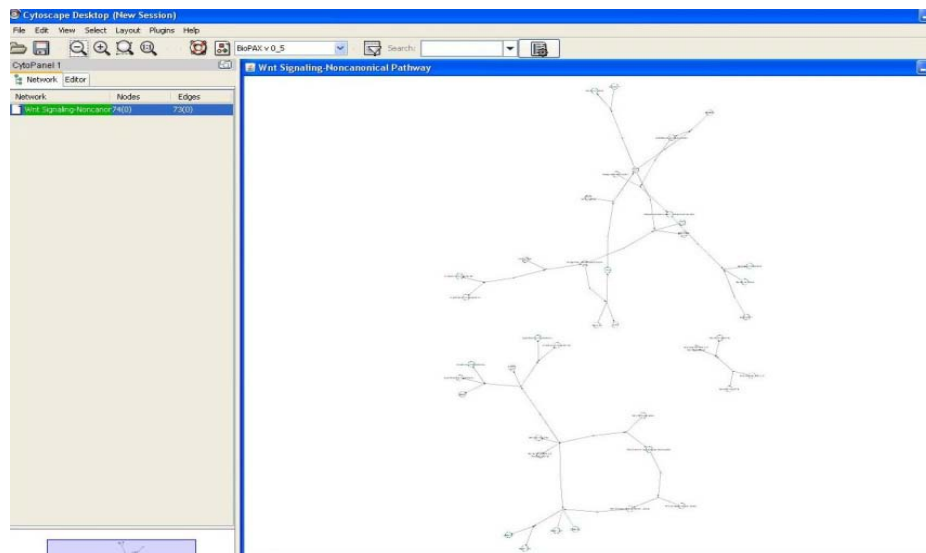
Biologists can view the pathways using Cytoscape (Figure 2.8a) and view gene expression data on any pathway (Figure 2.8b). Each pathway contains proteins and their cellular locations as well as different types of physical interactions, such as molecular interaction, biochemical reaction, catalysis and transport, post-translational protein modifications, original citations, experimental evidence, and links to other databases. Only some information is available in the downloaded BioPAX files.



**Figure 2.8 (a) Cytoscape view of Pathway. (b) Expression Data view on Wnt pathway using expression viewer software.**

## Pathway Commons

Pathway Commons [20] provides access to biological pathway information collected from public pathway databases, which we can browse or search. Currently, Pathway Commons has 994 pathways including 12,550 interactions from 9 different organisms. Pathway Commons currently contains Cancer Cell Map, Humancyc, NCI/Nature pathway interaction database and Reactome. Biologists can access biological pathway information collected from public pathway databases, which we can browse or search. Pathways include biochemical reactions, complex assembly, transport and catalysis events, and physical interactions involving proteins, DNA, RNA, small molecules and complexes. Pathways can be browsed by gene name, gene identifier or pathway name, respective examples are p53, P38398 and mTOR. Also, searching can be restricted to specific data sources or specific organisms. NCI/Nature curated database is taken as an example (Figure 2.9). For this source, it lists the biochemical reactions, complex assemblies, transport reactions, and molecules involved in the pathway.

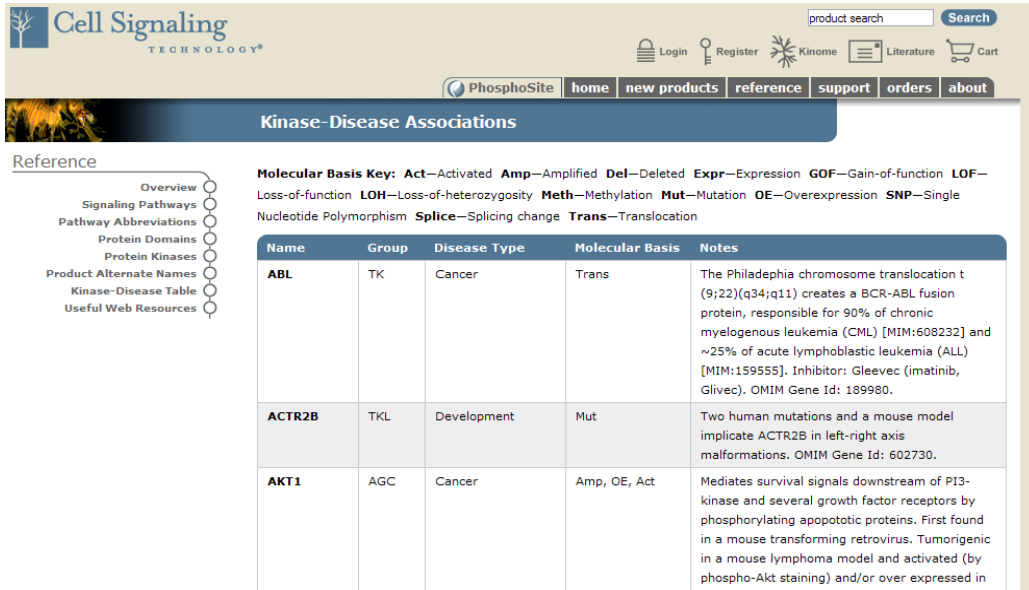


**Figure 2.9 Cytoscape view of Pathway in Pathway commons.**



## Cell Signaling Technology

Cell Signaling Technology, Inc. (CST) [21], is a commercial site committed to developing innovative new research tools to help define the mechanisms underlying cell function and disease. It is a small pathway portal for showcasing Cell Signaling Technology products. CST has 17 signaling pathways and pathway diagrams that are clickable and link to more information about each protein and the commercial products that are available for that protein. They also have kinase-disease association data, which provides how the kinases involved in the different disease types and also provides the molecular basis for the disease along with some notes with references (Figure 2.10) which support the kinase-disease association.



Reference

- Overview
- Signaling Pathways
- Pathway Abbreviations
- Protein Domains
- Protein Kinases
- Product Alternate Names
- Kinase-Disease Table
- Useful Web Resources

**Molecular Basis Key:** Act—Activated Amp—Amplified Del—Deleted Expr—Expression GOF—Gain-of-function LOF—Loss-of-function LOH—Loss-of-heterozygosity Meth—Methylation Mut—Mutation OE—Overexpression SNP—Single Nucleotide Polymorphism Splice—Splicing change Trans—Translocation

Name	Group	Disease Type	Molecular Basis	Notes
ABL	TK	Cancer	Trans	The Philadelphia chromosome translocation t(9;22)(q34;q11) creates a BCR-ABL fusion protein, responsible for 90% of chronic myelogenous leukemia (CML) [MIM:608232] and ~25% of acute lymphoblastic leukemia (ALL) [MIM:159555]. Inhibitor: Gleevec (imatinib, Glivec). OMIM Gene Id: 189980.
ACTR2B	TKL	Development	Mut	Two human mutations and a mouse model implicate ACTR2B in left-right axis malformations. OMIM Gene Id: 602730.
AKT1	AGC	Cancer	Amp, OE, Act	Mediates survival signals downstream of PI3-kinase and several growth factor receptors by phosphorylating apoptotic proteins. First found in a mouse transforming retrovirus. Tumorigenic in a mouse lymphoma model and activated (by phospho-Akt staining) and/or over expressed in

**Figure 2.10** Snapshot of kinase - disease associations from cell signaling technology.

## SPAD

The Signaling Pathway Database (SPAD) [22] is an integrated database for genetic information and signal transduction systems. They have developed an integrated

database SPAD to understand the overview of signaling. They divided signaling pathways into four categories based on extracellular signal molecules: growth factor, cytokine, hormone and stress that initiate the intracellular signaling pathway. SPAD provides clickable pathway maps.

### Reactome

Reactome [5] is an online bioinformatics database of human pathways - DNA replication, transcription, translation, the cell cycle, metabolism, and signaling cascades - and can be browsed to the molecular details of the signaling cascade. The information in Reactome pathways is curated from the published research literature by expert biologists. Reactome contains 1,115 pathways. The basic unit of the Reactome database is a reaction. Reactions are then grouped into pathways. Figure 2.11 shows the reaction page from Reactome. Reactome can infer equivalent reactions in multiple non-human species.

**Diagram**

Assembly of the destruction complex (Homo sapiens)

Association of beta-catenin with the destruction complex

**Details**

**Assembly of the destruction complex**

<b>Stable identifier</b>	REACT_10134.1
<b>Author</b>	Kimeiman, D., 2007-04-03
<b>Reviewed</b>	Pagano, M., 2007-04-27

The exact composition of the destruction complex is not known. A number of components appear to form a core complex, while others may associate with the complex transiently when a Wnt signal is present (reviewed in Kimeiman and Xu, 2006). The core components include Axin, glycogen synthase kinase 3 (GSK-3), Casein Kinase 1 (CK1) alpha, beta-catenin, Protein phosphatase 2A (PP2A) and Adenomatous Polyposis Coil (APC). CK1 epsilon, Diversin and PP1 may also be components of the complex. [Kimeiman & Xu 2006]

<b>Input (present at start of reaction)</b>	GSK3B [cytosol] APC [cytosol] CK1alpha [cytosol] Axin [cytosol] PP2A [cytosol]
<b>Output (present at end of reaction)</b>	Axin:GSK3:CK1alpha:APC:PP2A complex [cytosol]
<b>Following event(s)</b>	Association of beta-catenin with the destruction complex [Homo sapiens]
<b>Organism</b>	Homo sapiens
<b>Cellular compartment</b>	cytosol

**References**

Seeling, JM, Miller, JR, Gil, R, Moon, RT, White, R, Virshup, DM Regulation of beta-catenin signaling by the 856 subunit of protein phosphatase 2A **1999 Science** [PubMed](#)

Dajani, R, Fraser, E, Roe, SM, Yeo, M, Good, VM, Thompson, V, Dale, TC, Pearl, LH Structural basis for recruitment of glycogen synthase kinase 3beta to the axin/APC scaffold complex **2003 EMBO J** [PubMed](#)

**Participating molecules**

- APC [cytosol]
- APC\_1 [cytosol]
- APC\_2 [cytosol]
- Axin [cytosol]
- CK1alpha [cytosol]
- GSK3B [cytosol]
- GSK3B\_1 [cytosol]
- GSK3B\_2 [cytosol]
- PP2A regulatory subunit B [cytosol]
- PP2A catalytic subunit C [cytosol]
- ...

List all 11 participating molecules

**Figure 2.11** Screen shot of Reactome reaction page.

Features such as availability, download format, details such as small molecules, genes/proteins, and interactions/reactions documented in pathway databases were surveyed. Table 2.1 lays out a comparison of different pathway databases

**Table 2.1 Comparison of pathway databases**

Database	URL	Content/ type(s) of data	Interaction Statistics	Availa- bility	Features	Data download /format
The Cancer Cell Map	<a href="http://cancer.cellmap.org/cellmap/home.do">http://cancer.cellmap.org/cellmap/home.do</a>	Human-focused cellular pathways implicated in cancer.	6 pathways Genes /Proteins:479 Interactions/ Reactions: 1092	Free to all users	Molecules reactions cytoscape view	BioPAX
Protein Lounge	<a href="http://www.proteinlounge.com">http://www.proteinlounge.com</a>	curated pathway clickable images. Pathways are available for many organisms.	254-signaling pathways	Free for IUPUI users	Pathway items clickable image	No data download
Pathway Interaction Database (PID)	<a href="http://hpid.org">http://hpid.org</a>	signaling pathways, assembled by NCICB staff from publicly available sources of information (mainly KEGG and BioCarta). Pathway Diagrams	Free to all users	Free to all users	Clickable image	XML
Signaling Pathway Database (SPAD)	<a href="http://www.grt.kyushu-u.ac.jp/eny-doc/">http://www.grt.kyushu-u.ac.jp/eny-doc/</a>	Signaling Pathways, Pathway Diagrams	-	Free to all users	Clickable Image	No data download
Cell Signaling Technology Pathway Database (CST)	<a href="http://www.cellsignal.com/">http://www.cellsignal.com/</a>	Contains pathway diagrams that are clickable and link to more information about each protein and	17 Pathways, Experimental, Predicted	Free to all users	Clickable image	No data download

		the commercial products that are available for that protein. kinase - disease associations				
Reactome Knowledge Base	<a href="http://www.reactome.org">http://www.reactome.org</a>	Metabolic Pathways, Signaling Pathways	1115-pathways	Free to all users	Reactions	BioPAX, SBML
Biocarta	<a href="http://www.biocarta.com/genes/index.asp">http://www.biocarta.com/genes/index.asp</a>	Pathway Diagrams	315 -pathways	Free to all users	Clickable image	
Kyoto Encyclopedia of Genes and Genomes (KEGG)	<a href="http://www.genome.ad.jp/kegg/">http://www.genome.ad.jp/kegg/</a>	Pathway Diagrams	Small Molecules: 13463; 6461 Pathways:	Free to all users	Clickable image	BioPAX
ResNet	<a href="http://www.ariadnegenomics.com/products/resnet.html">http://www.ariadnegenomics.com/products/resnet.html</a>	Protein-Protein Interactions, Signaling Pathways.	Genes / Proteins: 15000 Interactions / Reactions: 192 Pathways: 506	License purchase required	Molecules reactions, regulations	-
<b>Pathway Commons</b>	<a href="http://www.pathwaycommons.org/pathwaycommons/home.do">http://www.pathwaycommons.org/pathwaycommons/home.do</a>	Protein-Protein Interactions, biological Pathways.	<b>Pathways: 994</b> <b>Physical Entities: 16,933</b> <b>Interactions: 12,550</b> <b>Organisms: 9</b>	Free to all users	Molecules, reactions	BioPAX

Based on the above review we chose the NCI-Nature curated database, Biocarta, Protein Lounge and Resnet, to integrate and develop a comprehensive pathway database. Next, we compared the features between the selected databases in detail (Table 2.2).

As we see in Table 2.2, NCI-Nature curated data and Biocarta do not provide pathway items, categorization of molecules, or pathway reactions at the interface level, but do provide this data in raw form. Contrastingly, Protein Lounge provides the pathway items without any further categorization of molecules and does not provide pathway reactions. Protein Lounge is the only source which provides the effect of environmental

**Table 2.2 Feature Comparisons of Human Signaling Pathway Databases**

<b>Entity</b> \ <b>Database</b>	<b>NCI-Nature Curated data</b>	<b>Biocarta</b>	<b>Protein Lounge</b>	<b>Resnet</b>
<i>Pathway items list</i>	-*	-*	+	+
<i>Pathway description</i>	-	+	+	-
<i>Categorization of molecules</i>	-*	-*	-	+
<i>Pathway reactions</i>	-*	-*	-	+
<i>Effect of environmental factors</i>	-	-	+	-
<i>Similar pathways</i>	-	-	+	+
<i>Pathway output format</i>	GIF XML BioPAX SVG	Clickable image	Clickable image along with pathway items list	Clickable image, EXCEL

\* Not available from interface, but this data is available for download from PID.

factors on the pathways. Most of these sources are clickable images. Only Pathway Studio/Resnet provides an option for Excel export. NCI-Nature curated data and Biocarta data from PID provides XML, GIF, BioPAX and SVG output formats of the pathway.

**Table 2.3 Comparison of source data statistics**

<b>Entity</b> \ <b>Database</b>	<b>NCI-Nature Curated data</b>	<b>Biocarta</b>	<b>Protein Lounge</b>	<b>Resnet</b>
pathways	32	254	464	1259
Molecules	1595	3407		15974
Complex	23	65	-	135
Interactions	1306	3064	-	21454
Data Format	XML	XML	HTML	EXCEL

We compared the source data statistics in Table 2.3. Based on this data, we can say that these databases differ greatly with respect to the coverage of signaling pathways.

## 2.4 Problem Statement and Research Question

Existing pathway databases define pathways by different levels of details such as proteins, compounds, complexes and their cellular locations, biochemical reaction, catalysis, complex assembly and transport, post-translational modifications, and links to other databases (e.g., of protein sequence annotation). Also, the coverage with respect to the number of pathways is the main problem with the signaling pathway databases, so integration of signaling pathway databases may be essential to provide comprehensive information about the pathways. The question is how best we can integrate, organize and represent the data of human signaling pathways and annotation information computationally to extract new biological knowledge (e.g., validating disease biomarkers in molecular network context and identifying better drug targets).

Pathway databases have different coverage for specific pathways and may have different pathway names. In order to address this problem, we came up with the concept of merging similar pathways based on gene/protein identifiers to better provide the user with a more comprehensive treatment of the information.

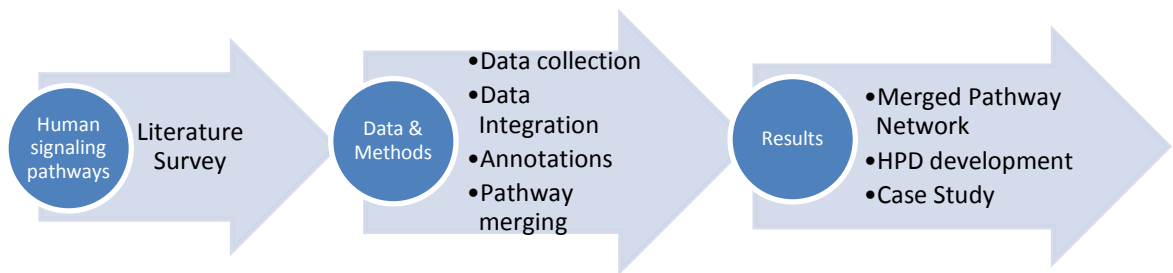
## CHAPTER THREE: ARCHITECTURAL APPROACH TO HUMAN PATHWAY DATABASE

The previous chapter introduced related work and presented possible research questions that can be answered by integration of pathway databases. It gave a brief overview of the pathway databases that are used for data integration with a comparison of content and features. This chapter details the method road map, architecture of the HPD database, and framework of the data integration system describing the various components that contribute to the integration. The latter part of this chapter describes the method for analyzing pathway merging. The motivation of this work came from the heterogeneity in pathways, including observed differences between pathway boundaries for similar pathways, between the pathway databases.

### 3.1 Approach to Pathway Data Integration

#### 3.1.1 Roadmap of Methods Used

The Method Roadmap of current research work shown in Figure 3.1 explains the overview of tasks like data collection, data integration, methods and results of current thesis work. This provides for a quick understanding of the thesis project.



**Figure 3.1 Method Roadmap of the thesis.**

### 3.1.2 Goals and Challenges of Integration

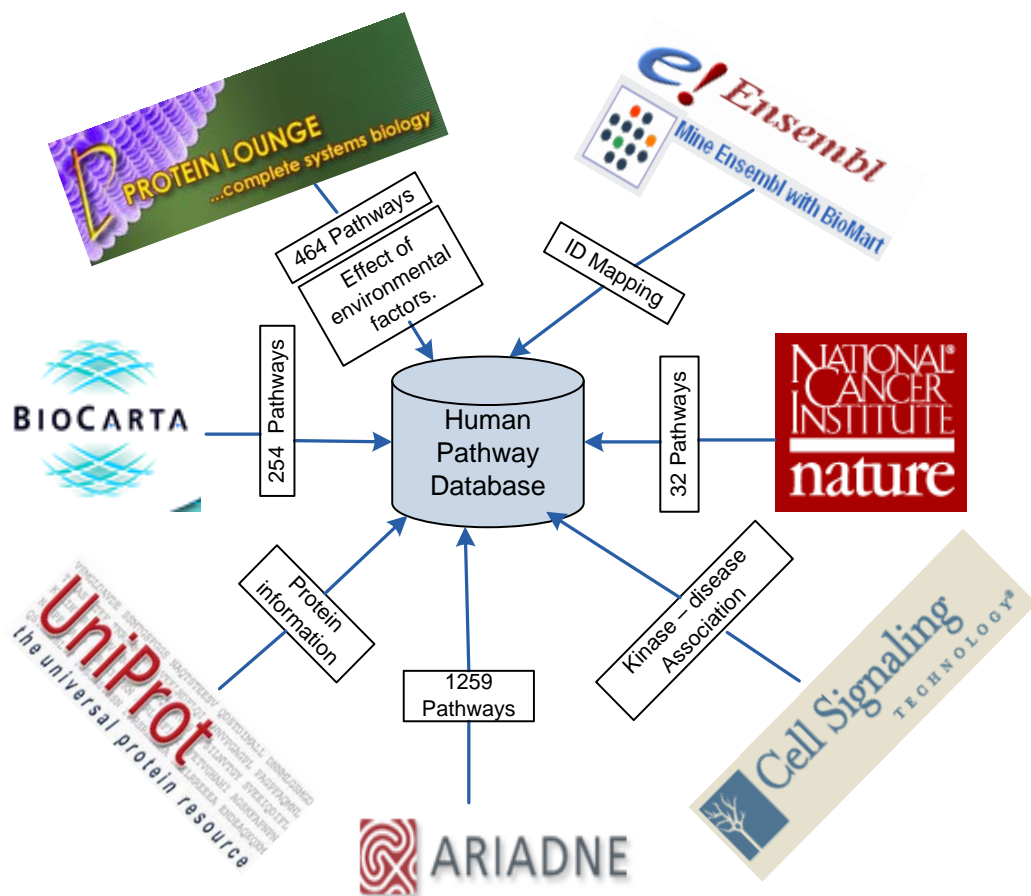
In this section, I describe integration of pathway data from heterogeneous data sources into a unified platform, the Human Pathway Database (HPD) using a data warehouse approach.

Advances in new research technologies for data generation, for example microarrays, DNA sequencing, quantitative proteomics and mass spectrometry analysis have generated massive amounts of data. The often accelerating increase in publication for different areas of biological research has also provided copious amounts of biological data as indicated by Figure 2.2 of the previous chapter. There is also increasing heterogeneity in the data especially in biology because of the distribution of data in various sources and lack of standard exchange protocols and controlled vocabularies to describing data structures and semantics (bioinformatics system integration). In order to make a decision based on available data, data integration works to provide a scientist with a complete depiction of the system. While integration is important for most application domains for providing comprehensive view of the domain, it is a challenging task. This is due to heterogeneity of the different data sources i.e. rich diversity in data, differences in curation, multiple sources of similar data with different identifying systems, and different coverage.

The objective of systems biology research is to come up with new hypotheses by analyzing experimental data. For example, microarray gene expression data against publicly available annotation databases may provide better insight into research and discovery. Cross analysis of the data across domain boundaries to generate a biological hypothesis is a general approach in systems biology studies. Systems biology is the study



of an organism, and can provide an integrated and interacting network view of genes, proteins and biochemical reactions. Solving a biological question based on systems biology studies requires data from different sources such as annotation databases, literature findings and experimental data. The aim of this research work is to provide the researcher with comprehensive information from data sources with differences in coverage of pathway content and designations of pathway boundaries. Figure 3.2 illustrates the number of pathway databases that are integrated in this project in order to perform pathway analysis to generate pathway network views overlaid with kinase – disease annotations and the effects of environmental factors.



**Figure 3.2 Pathway data integration and annotations.**

## Data integration challenges

Heterogeneity of databases can generally introduces four different types of challenges, and these four types of challenges were encountered with the HPD integration.

### *Syntactic Heterogeneity:*

Syntactic heterogeneity occurs when different names are used to refer to the same entity such as can occur from the use of synonyms and the use of different identifiers to identify the attributes. For example, in pathway biology, one database identifies molecular entities with the UniProt ID system and the other database identifies with the Refseq ID system.

### *Semantic Heterogeneity:*

Semantic heterogeneity arises when the same term refers to different entities in different contexts. For example, this was found for instances with the same name for the protein and gene. APP is a synonym for Amyloid beta A4 protein precursor and the gene name for the APP protein is also represented as APP.

### *Data Model Heterogeneity:*

This is a common type of heterogeneity encountered during a data integration process. This heterogeneity refers to the differences in formats. For example, NCI is downloaded in XML format and Protein Lounge data is obtained from HTML format. Data model heterogeneity can also refer to the data coming from different types of databases.

### *Schematic Heterogeneity:*

Schematic heterogeneity refers to the databases where data is represented in structurally different forms although syntactically similar. We found this type of

heterogeneity between the NCI – Nature curated database and the Biocarta database with respect to the Molecule entity. Figure 3.3 represents the NFATc entity in NCI – Nature curated structure of Molecule entity and Figure 3.4 represents the corresponding entity in Biocarta. NCI – Nature curated data represents the NFATc under the Molecule entity as a protein whereas Biocarta represents the NFATc under the Molecule entity as a protein family.

```
<Molecule molecule_type="protein" id="201227">
  <Name name_type="UP" value="095644"/>
  <Name name_type="AS" value="NFATc"/>
</Molecule>
```

**Figure 3.3 Structure of NFATc entity in NCI – Nature curated data.**

```
<Molecule molecule_type="protein" id="100197">
  <Name name_type="AS" value="NFATc"/>
  <Family family_molecule_idref="100197" member_molecule_idref="100198"/>
</Molecule>
<Molecule molecule_type="protein" id="100198">
  <Name name_type="LL" value="4772"/>
  <Name name_type="OF" value="NFATC1"/>
</Molecule>
```

**Figure 3.4 Structure of NFATc entity in Biocarta.**

Most of the above heterogeneities were taken care of during the development of HPD. A data warehouse approach was used to integrate the data.

## 3.2 Pathway Data Warehouse and Database Application

### 3.2.1 Strategy

Data warehousing is a popular approach for data integration where data from different sources are extracted, transformed, and loaded into source and queried with a single schema.

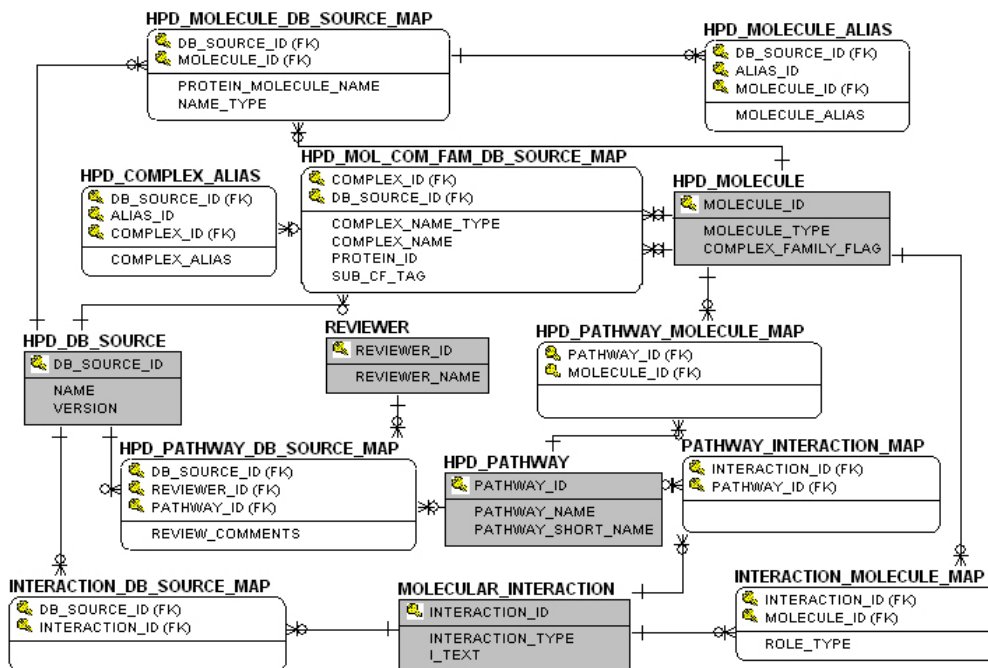
The data were extracted from four different sources: NCI-Nature Curated data [6], Biocarta [3], Protein Lounge [7], and Pathway Studio [9]. The data are in varying formats. Figure 4.3 shows the framework graphically. In the next few sections, a step-by-step methodology is shown for the development of this framework and the various tools that could assist in the development.

There are two main steps in the data warehousing approach.

1. Development of a unified data model.
2. Development of software programs.

### 3.2.1.1 Development of a Unified Data Model

Based on our experience working with pathway data and lessons learned from pathway XML data representation standards such as SBML [23] and BioPAX [24], we developed a pathway entity-relationship (ER) data model for HPD (Figure 3.5).



**Figure 3.5 ER diagram for the HPD.**

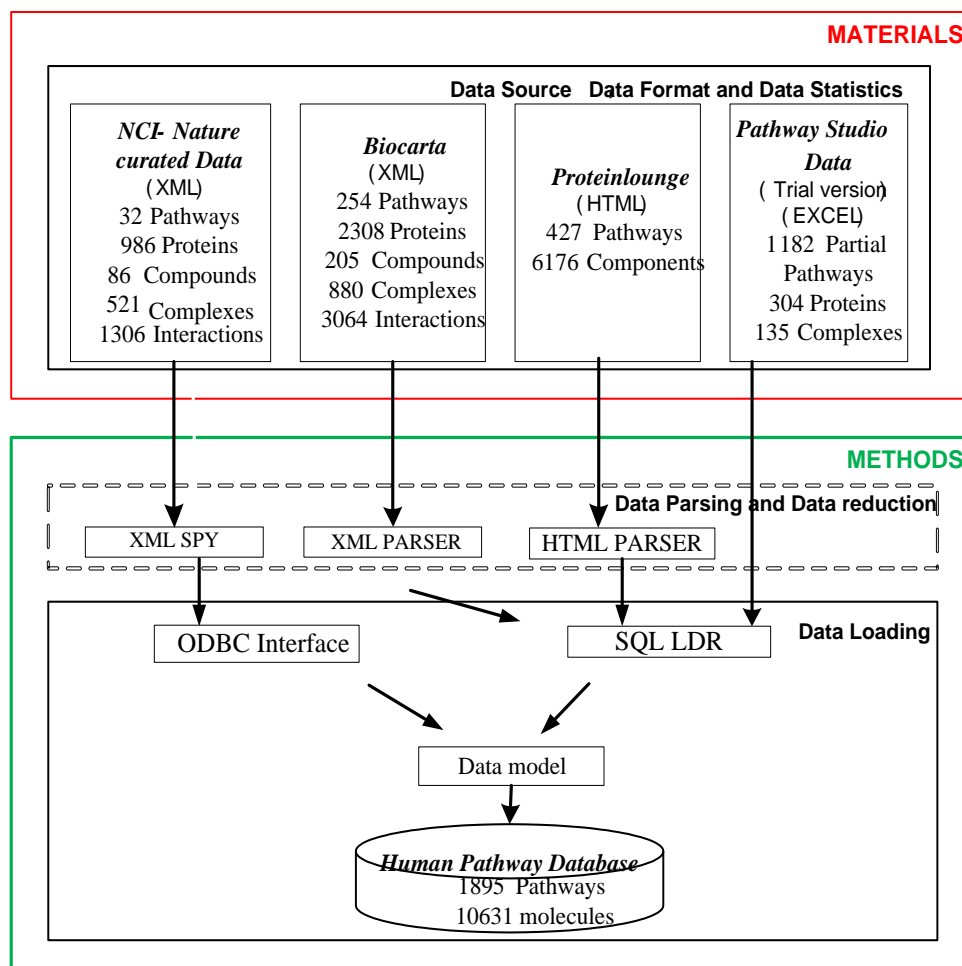
The data model in Figure 3.5 unifies the representation of all integrated pathway entities, including molecules, complexes, compounds, regulatory relationships of molecules, and reactions involved in signaling and regulatory pathways.

The data model allows flexible representation and management data on all types of pathway “HPD Molecules”, many-to-many relationships among them as “Molecular-Interactions”, multiple names and aliases from different database sources for each HPD molecule, and current/future review comments about each “HPD Pathway”. The semantics of relationships among different entities are represented in an Entity-Relationship (ER) data model.

#### 3.2.1.2 Development of Software Programs

In this section, we talk about a series of software programs that were developed to get the data from source databases and transform the data such that it can fit into the unified data model and then loaded into the warehouse. Framework and database statistics are specified in detail in Figure 3.6.

A data parsing step was required to convert data into a tab-delimited format. Altova XML spy was used to parse the XML data from NCI-Nature Curated data and Biocarta data into a relational format. While validating the parsed data, I found errors with the Biocarta data parsing, the I wrote a Perl script to parse the Biocarta data. Protein Lounge does not provide options for data download, so the HTML pages related to human signaling pathways were accessed by passing the pathway name variable into the URL, and subsequently parsed. Pathway studio data was downloaded in EXCEL format.



**Figure 3.6 Overview of pathway data integration process.**

Except for Protein Lounge, the other databases selected for integration (NCI, Biocarta, and Pathway Studio) consist of only human pathways. As this work focuses on humans, steps were taken to limit the data to be integrated to *Homo sapiens* data. The data from Protein Lounge had data from 14 different organism categories encompassing fungi and other diverse types of unicellular and multi cellular eukaryotes as well as bacteria. To collect only human pathways, we have taken the following two steps with Protein Lounge:

1. A Perl program was written to extract only human pathways, where it rejects the pathways which have other organisms' categorical names in the pathway name, e.g., cAMP signaling in *S.Cerevisiae*.
2. We mapped the proteins involved in pathways to human UniProt identifiers to collect human signaling pathways.

Vigorous data cleansing to remove potential errors and duplications was performed in the transforming of data to fit into the data model. Oracle's SQL Loader was used to load the final data into tables. As the data files and tables to be loaded are created, control files were created for each table to control data loading from the corresponding data file. The SQL Loader was executed to read the control file and to load the data. A log file was produced that describes what happened and describes any errors that may have occurred. An ODBC-managed database loaded data from access to Oracle.

The data warehouse of pathway information was developed on BIO10G2, an Oracle database available on the libra45 server, and this serves as a maintained platform for future pathway analysis studies within the laboratory. The warehouse can be used as a 'one-stop shop' for answering any of the questions that the source databases can handle, as well as those that require integrating knowledge that the individual sources do not have. Overall, it was more difficult to create a data warehouse than might be initially anticipated because of the heterogeneity between the databases, inconsistency between the databases, and issues with mappings between common identifiers.

### 3.2.2 Query Processing

The query performance was optimized by creating indexes for all the required attributes over different tables in the database. An index optimizes the query performance by ordering rows to speed access.

In order to process the queries from the front end, PHP's OCI Extension module was used to connect Oracle to PHP because the OCI extension module is optimized and provides more options such as CLOBs, BLOBs, BFILEs, ROWIDs than the alternative ORA Extension module.

### 3.2.3 Architecture of Web Application

User interface-flow diagrams [25] were used to represent the overview of the user interface of HPD database. A high-level overview and architectural approach was implemented to understand the complete user interface for this system. Factors like simplicity, usability, clarity, and speed have been considered during the design of the HPD website. The HPD website is available at <http://discover.uits.indiana.edu:8340/pathway2/>.

The HPD database is a typical 3-tier web application, an application program that is organized into three major parts, each of which is distributed to a different place in a network. The 3-tier application uses the client/server computing model (Figure 3.7). The three parts are:

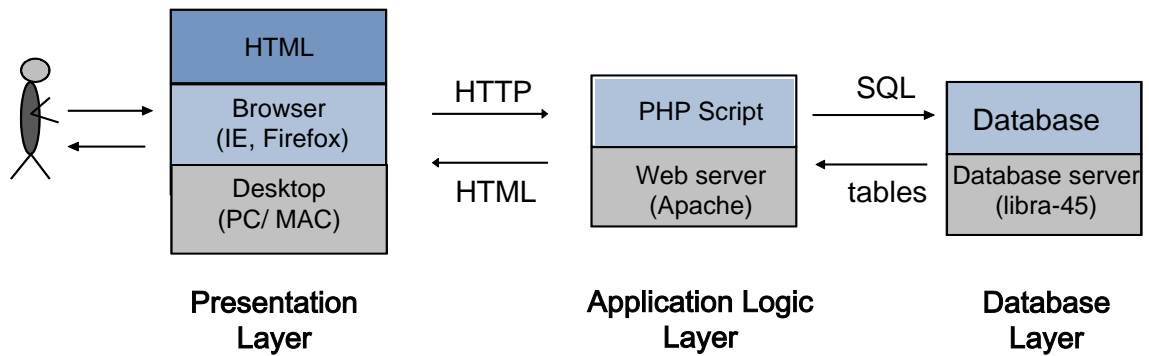
- The presentation layer,
- The application logic layer,
- The database layer.



In the next few paragraphs, I describe these layers in detail. Figure 4.2 presents an overview of the structure and technologies of HPD’s hardware and software architectures.

Presentation Layer

The application presentation layer provides the graphical user interface. It is responsible for receiving inputs, presenting data, and controlling the user interface. The application presentation layer receives an HTTP request and returns a response in the form of an HTML document. HTML was used for defining content structures and images on the web page.



**Figure 3.7 3-Tier Architecture of HPD.**

Application Logic Layer

The business logic exists in this layer and it is located on a local area network server. The business logic of the application decides if all conditions are met and implements use case scenarios. It processes requests according to the business rules: for example, deciding whether to reject input data or to send it to the database. In this sense, the business logic acts as the server for client requests from the user interface. In turn, it determines what data is needed (and where it is located) and acts as a client in relation to

a third tier of programming. The functionality of the program is found in the application layer, and application layer functionality was developed with PHP.

#### Database Layer

This layer includes the database and manages the persistence of application information. It is powered by an Oracle relational database server. Functions are used to execute database server-side processes related to data integrity. Queries are used for presenting data to applications. The data was stored in database tables.

PHP provides two extension modules that can be used to connect to Oracle:

- (1) Oracle functions (ORA), and
- (2) Oracle Call-Interface functions (OCI).

The OCI Extension module was used to connect to Oracle using PHP since it is optimized. A 3-tier architecture was chosen because of its flexibility, maintainability, reusability, scalability and reliability [26].

#### 3.2.4 Pathway Mergability

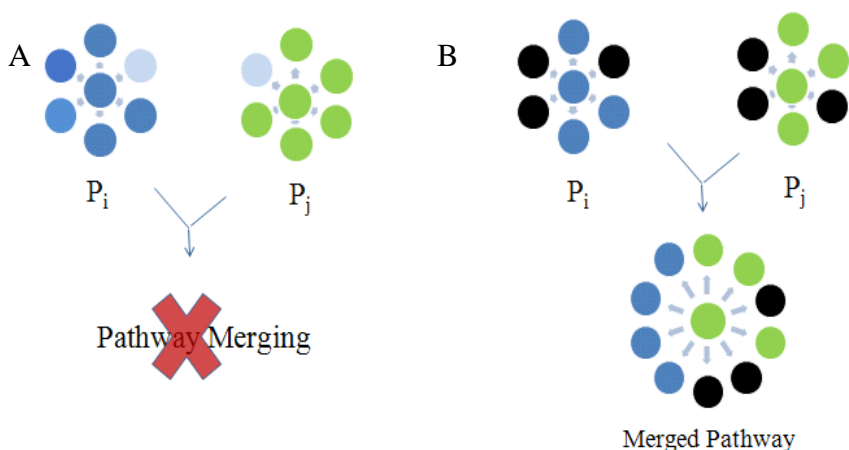
It is observed that there is much inconsistency with respect to the pathway boundary and the component information between the pathway databases. The same pathway appears in different databases which are inconsistent and each source has different names for the same pathway. Since anecdotal examination shows pathway boundaries to be vague, the merging of similar pathways can help provide more comprehensive information of the pathway in a non-redundant fashion. Towards this, we came up with a 'Pathway Mergability' concept.

We analyzed similar pathways sharing components and developed methods for merging similar pathways. To assess how many pathways could be merged, we used

clustering techniques to identify similar pathways based on pathway components (instead of by names). In a pathway clustering process, since a pathway is represented as a set of molecules, we define the similarity score  $S_{i,j}$  of two different pathways as shown in Equation 1.

$$S_{i,j} = \alpha \times \frac{|P_i \cap P_j|}{|P_i \cup P_j|} + (1 - \alpha) \times \frac{|P_i \cap P_j|}{\min\{|P_i|, |P_j|\}} \quad (1)$$

In Equation 1,  $P_i$  and  $P_j$  denote two different pathways. Their intersection  $P_i \cap P_j$  denotes a common set of molecules that are mapped with the same UniProt ID, and their union  $P_i \cup P_j$  is calculated as  $|P_i| + |P_j| - |P_i \cap P_j|$ . Here  $\alpha$  is a weight coefficient, which we use to take into account the varying degree of contributions from calculations based on the union and the overlap. Based on our experiments to observe the effects of different  $\alpha$  values on  $S_{i,j}$ , we determined that when  $\alpha = 0.8$ , the score  $S_{i,j}$  distribution of all the pathways is closest to a Poisson distribution. Therefore, we set  $\alpha = 0.8$  for the rest of analysis performed in this work.



**Figure 3.8 Pathway mergability examples (A) Example 1 (B) Example 2.**

We define the similarity as Eq. (1) with the condition as  $\{S_{ij} \geq 0.2, \text{ and } |P_i \cap P_j| \geq 2\}$ [43]. Figure 3.8 gives contrasting examples of how this mergability concept works. In Example 1 (Figure 3.8 a), pathway merging is not possible since there is only one common element between the pathways. Whereas in Example 2 (Figure 3.8 b), there are three common elements between the pathways (black nodes), and this successfully meets the pathway merging condition.

In the next chapter we talk about more examples and results based on these concepts.

## CHAPTER FOUR: RESULTS

The previous chapter detailed the methodology for the HPD database including the road map of development, architecture, and components and approaches associated with data integration, including the methods for pathway mergability. This chapter summarizes the results, overlapping analysis, and merging of pathways based on the methods described in the previous chapter. This chapter also presents analyses of how single proteins can be involved in multiple pathways as a way of characterizing the general significance of different proteins to the overall cellular network. The final part of this chapter describes the HPD web interface.

### 4.1 Overlap of Pathways among Source Databases

We performed a source pathway database overlapping analysis by applying a pathway merging condition  $\{S_{i,j} \geq 0.2, \text{ and } |P_i \cap P_j| > 2\}$  as described in Section 3.2.4 with results shown in Table 4.1.

**Table 4.1 Pairwise comparisons of pathways between the data sources**

<b>Source</b>	<b>Source</b>	<b>Protein Lounge</b>	<b>NCI -Nature Curated data</b>	<b>Biocarta</b>	<b>Resnet</b>
<b>Protein Lounge (427)</b>		-	7	88	104
<b>NCI -Nature curated data (32)</b>		4	-	12	8
<b>Biocarta (254)</b>		65	22	-	112
<b>Resnet (1182)</b>		210	31	700	-

Table 4.1 compares the number of overlapping pathways between each of the four data sources used for HPD. For instance, 4 pathways from NCI -Nature curated data

overlap with 7 pathways from Protein Lounge, and 65 pathways from the Biocarta data overlap with 88 pathways from Protein Lounge.

Table 4.2 shows the number and names of distinct pathways from each of the four data sources that overlap with all of the other data sources.

**Table 4.2 Number Of Pathways overlapping between all the data sources**

Source	Number of pathways overlap	Pathway Name
Protein Lounge	3	<ul style="list-style-type: none"> <li>• HGF Pathway</li> <li>• Apoptotic Pathways Triggered By HIV1</li> <li>• TWEAK Pathway</li> </ul>
NCI	2	<ul style="list-style-type: none"> <li>• hiv-1 nef: negative effector of fas and tnf</li> <li>• signaling pathways activated by hepatocyte growth factor receptor (c-met)</li> </ul>
Biocarta	12	<ul style="list-style-type: none"> <li>• inhibition of cellular proliferation by gleevec</li> <li>• fas signaling pathway (cd95)</li> <li>• ceramide signaling pathway</li> <li>• tnfr1 signaling pathway</li> <li>• caspase cascade in apoptosis</li> <li>• il-2 receptor beta chain in t cell activation</li> <li>• fc epsilon receptor signaling in mast cells</li> <li>• pdgf signaling pathway</li> <li>• induction of apoptosis through dr3 and dr4/5 death receptors</li> <li>• hiv-1 nef: negative effector of fas and tnf</li> <li>• tnf/stress related signaling</li> <li>• sodd/tnfr1 signaling pathway</li> </ul>
Resnet	17	<ul style="list-style-type: none"> <li>• DR3 and DR4-5 pathways</li> <li>• Caspases</li> <li>• 4-1BB ligand</li> <li>• IL23</li> <li>• GITR Ligand</li> <li>• CD30L</li> <li>• CD27L</li> <li>• KLRG1 -&gt; MYC signaling pathway</li> <li>• SEMA6B -&gt; FOS signaling pathway</li> </ul>

		<ul style="list-style-type: none"> <li>• EPOR -&gt; VAV1 signaling pathway</li> <li>• EPOR -&gt; STAT1 signaling pathway</li> <li>• EPOR -&gt; CEBPA signaling pathway</li> <li>• GRIN1 -&gt; FOS signaling pathway</li> <li>• GRM2 -&gt; FOS signaling pathway</li> <li>• GRM5 -&gt; FOS signaling pathway</li> <li>• TNFSF10-TNFRSF1A</li> <li>• TNFSF4-TNFRSF9</li> </ul>
--	--	--

A pathway that overlaps among Protein Lounge, NCI-Nature curated data and Biocarta data is the “Apoptotic Pathways Triggered By HIV1” pathway from Protein Lounge, the “HIV-1 nef: negative effector of fas and tnf” pathway from the NCI-Nature curated database, and the “hiv-1 nef: negative effector of fas and tnf” pathway from the Biocarta database. The two pathways share a similarity score of 0.3 (30% of molecules from each pathway are identical). This similarity score is low because NCI-Nature curated data categorizes molecules into proteins, protein complexes, metabolic compounds, and RNA, whereas Protein Lounge contains only protein molecules. Even though these pathways are both related to apoptosis induced by HIV-1, the NCI-Nature curated pathway does not have as many details on the apoptosis pathway as the Protein Lounge data source. Protein Lounge describes the comparable pathway as having a Nef protein that down regulates cell surface expression of the primary HIV1 receptor CD4 by increasing endocytosis of cell surface CD4, and further implicates enzymatic activity of ASK1-induced apoptosis and apoptosis-suppressing activity of MKK7 and JNK. The omission of important details on molecular mechanisms of Nef and other regulatory mechanisms of apoptosis may be perplexing to users who choose to query only one of the available sources of pathway data and, in this instance, be particularly the case for users

who know that Nef may accelerate the development of AIDS in HIV-infected individuals [27]. Therefore, integrated HPD pathways could provide additional insights to researchers studying pathway data.

#### 4.2 Pathway-Spanning Proteins

Extensive proteins may be shared across different pathways in HPD. **Pathway-spanning proteins** may be considered to be those proteins that are shared among multiple HPD pathways. Table 4.3 shows the results from querying the database to find out the proteins that span across pathways and identifies the top 10 pathway-spanning proteins. Several proteins in the list, including MK01\_HUMAN, RAF1\_HUMAN, and MK03\_HUMAN, are kinases involved in the MAPK/ERK pathway - a signal transduction pathway that couples intracellular responses to the binding of growth factors to cell surface receptors. Activation of this pathway promotes cell division. AKT1\_HUMAN plays an important role in glucose transport. MK08\_HUMAN and MK14\_HUMAN responds to activation by environmental stress and pro-inflammatory cytokines by phosphorylating a number of transcription factors. FOS\_HUMAN is a proto-oncogene protein that plays an important role in cell proliferation and differentiation. These pathway-spanning proteins are universal “control proteins” that regulate many aspects of cell energy metabolism, growth, differentiation, and emergency response.



**Table 4.3 Top ten proteins involved in the pathways**

<b>UniProt ID (Protein Identifier)</b>	<b>Protein Description</b>	<b>No. of Pathways involved</b>	<b>Function</b>
MK01_HUMAN	Mitogen-activated protein kinase 1	786	<ul style="list-style-type: none"> <li>• Initiation and regulation of meiosis, mitosis, and post mitotic functions in differentiated cells.</li> </ul>
RAF1_HUMAN	RAF proto-oncogene serine/threonine-protein kinase	568	<ul style="list-style-type: none"> <li>• Involved in the transduction of mitogenic signals from the cell membrane to the nucleus.</li> </ul>
AKT1_HUMAN	Protein kinase B	566	<ul style="list-style-type: none"> <li>• Plays a role in glucose transport.</li> <li>• Mediates the anti apoptotic effects.</li> <li>• Mediates insulin-stimulated protein synthesis.</li> <li>• Promotes glycogen synthesis by mediating the insulin-induced activation of glycogen synthase.</li> </ul>
MK03_HUMAN	Mitogen-activated protein kinase 3	532	<ul style="list-style-type: none"> <li>• Involved in both the initiation and regulation of meiosis, mitosis, and post mitotic functions in differentiated cell.</li> </ul>
MK08_HUMAN	Mitogen-activated protein kinase 8	452	<ul style="list-style-type: none"> <li>• Responds to activation by environmental</li> </ul>

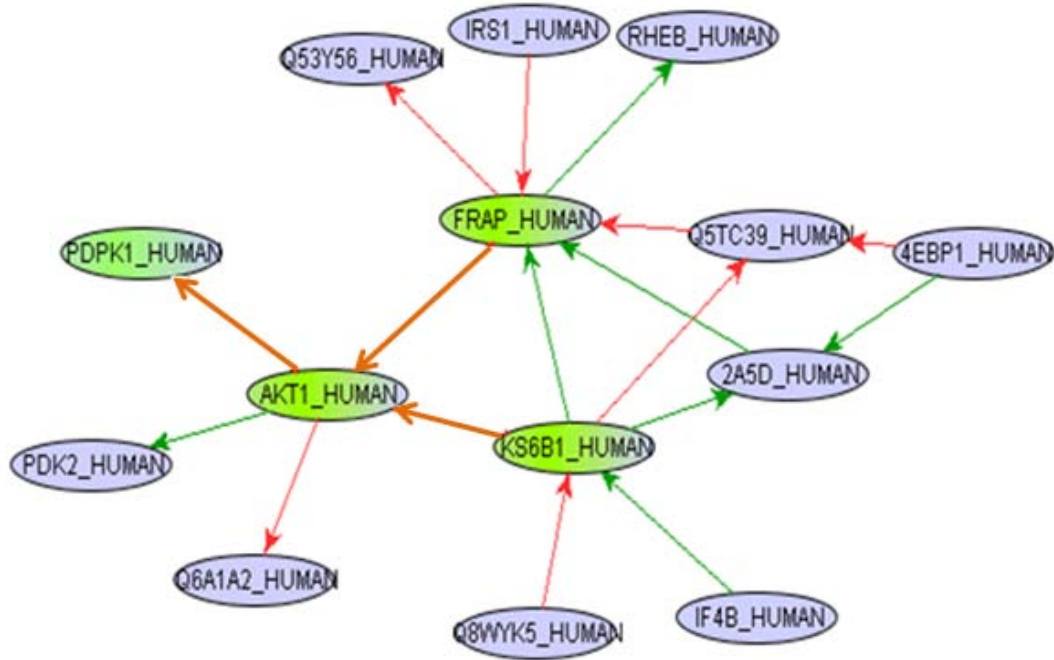
			stress and pro-inflammatory cytokines by phosphorylating a number of transcription factors
SHC1_HUMAN	SHC-transforming protein 1	428	<ul style="list-style-type: none"> <li>• Signaling adapter that couples activated growth factor receptors to signaling pathway.</li> <li>• Cytoplasmic propagation of mitogenic signals.</li> <li>• Involved in signal transduction pathways that regulate the cellular response to oxidative stress and life span.</li> </ul>
SRC_HUMAN	Proto-oncogene tyrosine-protein kinase Src	413	<ul style="list-style-type: none"> <li>• The kinases c-Src (Giepmans et al. 2001; Sorgen et al. 2004), play an essential role in the phosphorylation of Cx which leads to its degradation.</li> <li>• c-Src appears to associate with and phosphorylate Cx43 leading to closure of gap junctions.</li> </ul>
MK14_HUMAN	Mitogen-activated protein kinase 14	403	<ul style="list-style-type: none"> <li>• Responds to activation by environmental stress, pro-inflammatory cytokines and lipopolysaccharide (LPS) by phosphorylating a number of</li> </ul>

			transcription factors.
MP2K1_HUMAN	MAP kinase kinase 1	388	<ul style="list-style-type: none"> <li>• Catalyzes the phosphorylation of MAP kinases. Activates ERK1 and ERK2 MAP kinases.</li> </ul>
FOS_HUMAN	Proto-oncogene protein c-fos	361	<ul style="list-style-type: none"> <li>• Has a critical function in regulating the development of cells destined to form and maintain the skeleton.</li> <li>• Have an important role in signal transduction, cell proliferation and differentiation.</li> </ul>

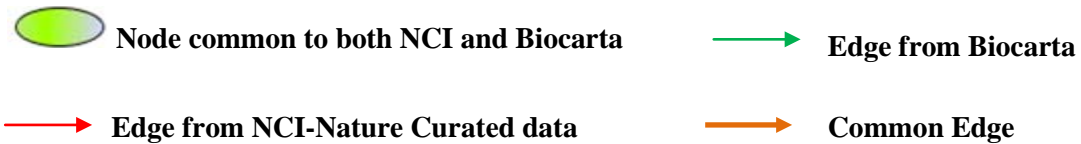
#### 4.3 Pathway mergability

We have taken mTOR Pathway as an example to show the merged pathway. The merged pathway is shown in Figure 4.1. The merged pathway is generated using data from both NCI-Nature curated data and Biocarta data with the condition as  $\{S_{i,j} 0.2, \text{ and } |P_i \cap P_j| > 2\}$ . The mergability scoring of the mTOR pathway is 0.3 and  $|P_i \cap P_j|$  is 4. The mTOR pathway regulates skeletal muscle atrophy and hypertrophy. As shown in Figure 4.1, the nodes that are common to both NCI-Nature curated data and Biocarta data are shown in green and the edges that are common to both databases are shown in orange. The edges from the NCI-Nature curated data are shown in red and the edges from Biocarta are shown in green. By merging the pathways, we found a protein 2A5D\_HUMAN that is not present in NCI-Nature curated data, and a protein IRS1\_HUMAN that is not present in Biocarta. This is evidence for how pathway

mergability can help provide comprehensive information along with the non-redundancy for pathway analysis.



**Figure 4.1 Merged Pathway Network of mTOR Pathway.**



4.4 Web Interface

In this work, the aim is to develop an integrated comprehensive human pathway database resource providing an integrated view of current pathway data from both annotated and predicted resources with the total number of:

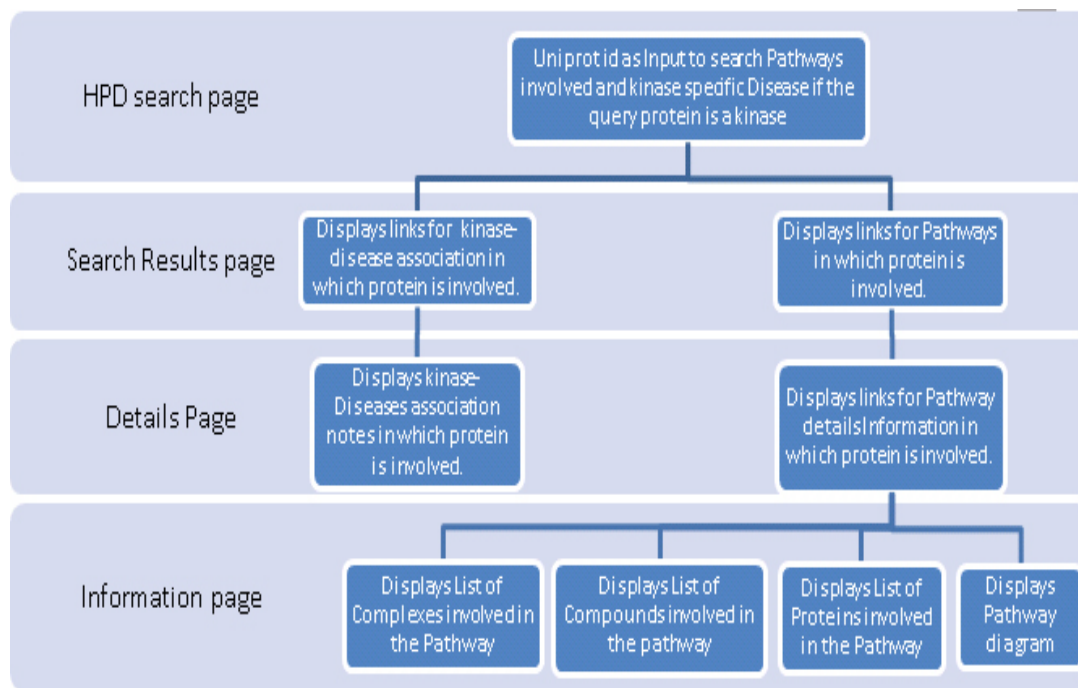
Pathways	1,895
Molecular entities (proteins/complexes/compounds)	10,631
Reactions	4,370
Kinase –disease associations	149

The diversity of pathway data was taken into account before designing the comprehensive database system. With HPD, this high-coverage human pathway database act to integrate information of proteins, complexes, compounds and reactions involved in signaling pathways along with kinase-disease annotations. The interface was designed to be user-friendly. The database is publicly available and can be accessed within the Discovery Informatics and Computing Group website at

<http://discover.uits.indiana.edu:8340/pathway2/>

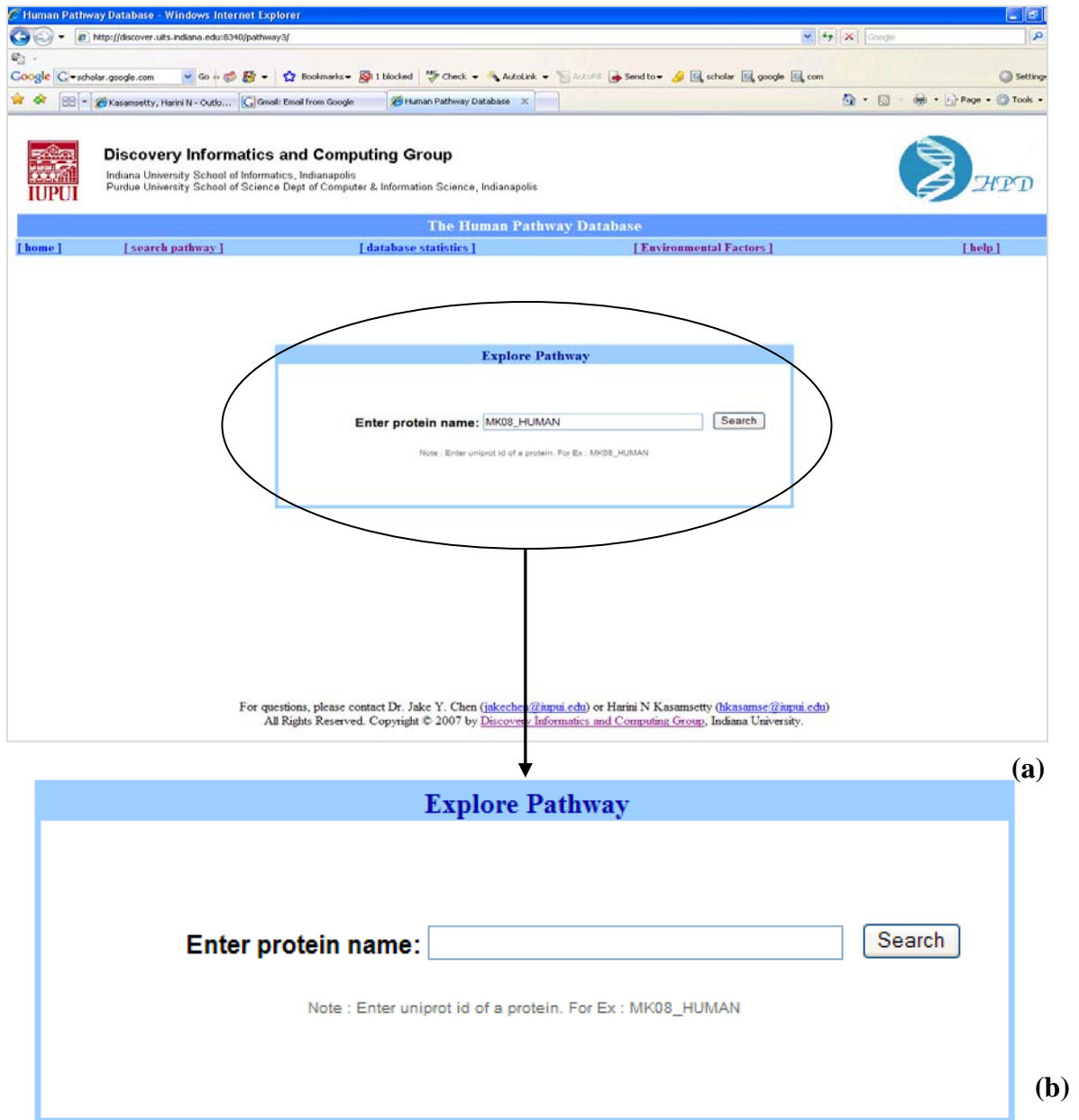
Over view of HPD User Interface

The following figure 3.8 gives an overview of HPD website where it shows the work flow of the HPD webservice.



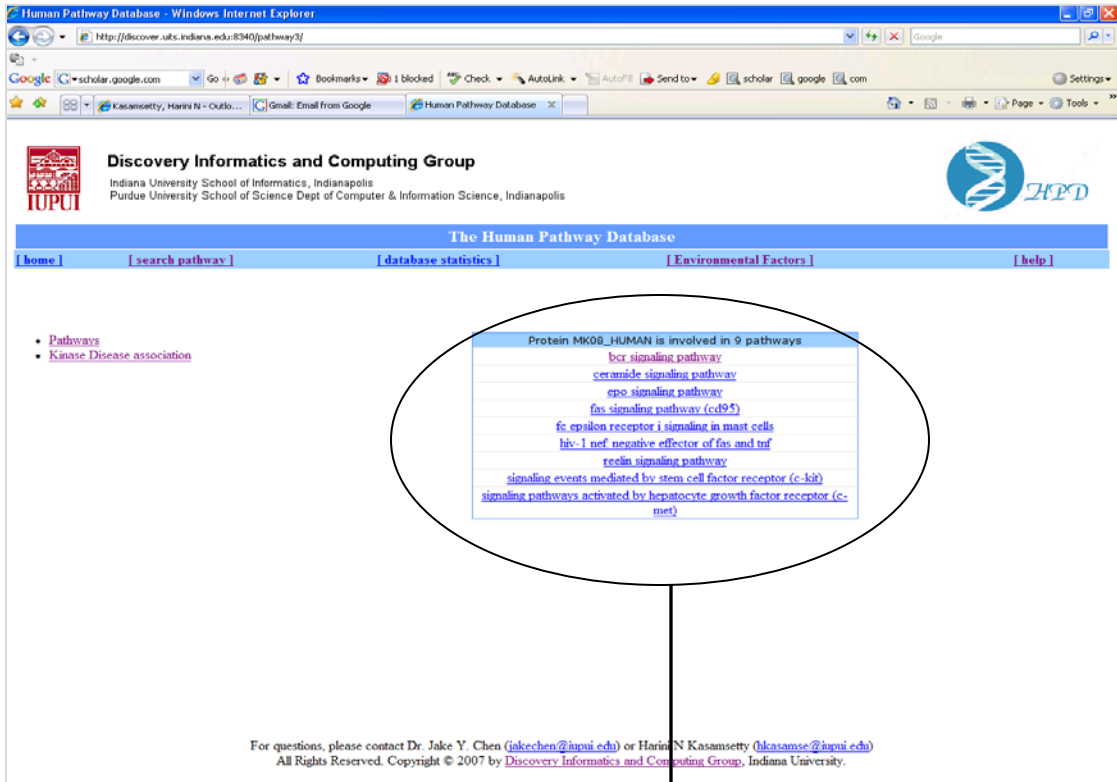
**Figure 4.2 Flow diagram of HPD User Interface.**

The web page for pathway searching has been designed to be as simple to use as possible. Figure 4.3 shows a screenshot of the pathway search interface, indicating where a protein is to be keyed in as a UniProt identifier.



**Figure 4.3 (a) Screenshot of the search page. Figure 4.3 (b) Zoomed view from oval region in 4.3(a).**

The results of a protein name search “MK08\_HUMAN” in HPD are shown in Figure 4.4 and consist of a list of curated and predicted pathways and a link for kinase-disease association.



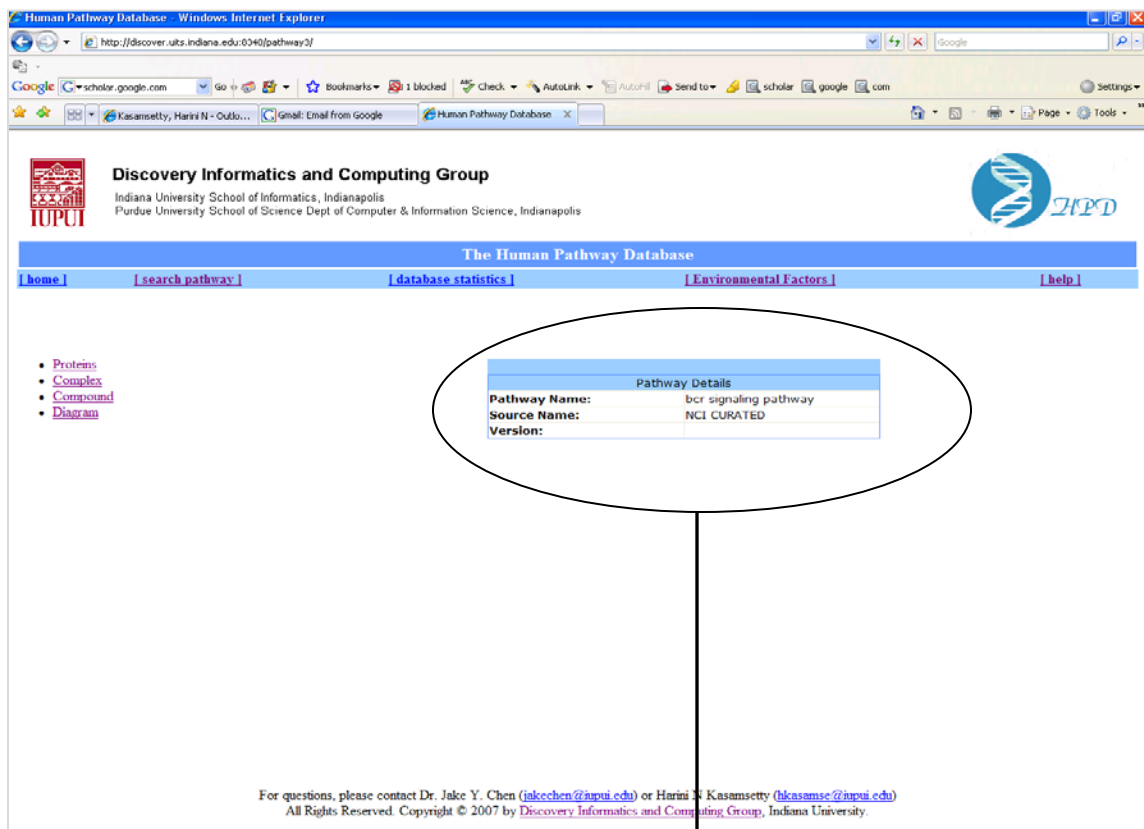
(a)

Protein MK08_HUMAN is involved in 9 pathways
<a href="#">bcr signaling pathway</a>
<a href="#">ceramide signaling pathway</a>
<a href="#">epo signaling pathway</a>
<a href="#">fas signaling pathway (cd95)</a>
<a href="#">fc epsilon receptor i signaling in mast cells</a>
<a href="#">hiv-1 nef: negative effector of fas and tnfr</a>
<a href="#">reelin signaling pathway</a>
<a href="#">signaling events mediated by stem cell factor receptor (c-kit)</a>
<a href="#">signaling pathways activated by hepatocyte growth factor receptor (c-met)</a>

(b)

**Figure 4.4 (a) Screenshot of the pathway list page. (b) Zoomed view from oval region in (a).**

The user can find more details about the pathways that are involved by selecting the pathway name from the list. Detailed information about pathways starts with a page like that shown in Figure 4.5 that links the user to a listing of pathway proteins (Figure 4.6), listings of complexes or compounds, and a diagram of the pathway (Figure 4.7). In Figure 4.6, proteins are further linked to UniProt entries, giving detailed information about the proteins.



(a)

Pathway Details	
<b>Pathway Name:</b>	bcr signaling pathway
<b>Source Name:</b>	NCI CURATED
<b>Version:</b>	

(b)

**Figure 4.5 (a) Screenshot of the pathway details page. (b) Zoomed view from oval region in (a).**



Human Pathway Database - Windows Internet Explorer

http://discover.uts.indiana.edu:8340/pathway3/

Discovery Informatics and Computing Group  
 Indiana University School of Informatics, Indianapolis  
 Purdue University School of Science Dept of Computer & Information Science, Indianapolis

The Human Pathway Database

home | search pathway | database statistics | Environmental Factors | help

- Proteins
- Complex
- Compound
- Diagram

Number Of Proteins in the Pathway: 68

Protein Name	UniprotId
MEKK1	<a href="#">M3K1 HUMAN</a>
PI3K catalytic alpha polypeptide	<a href="#">PK3CA HUMAN</a>
PI3K regulatory subunit polypeptide 1	<a href="#">P85A HUMAN</a>
SYK	<a href="#">KSYK HUMAN</a>
Lyn	<a href="#">LYN HUMAN</a>
Btk	<a href="#">BTK HUMAN</a>
Grb2	<a href="#">GRB2 HUMAN</a>
SHC	<a href="#">SHC1 HUMAN</a>
RAF1	<a href="#">RAF1 HUMAN</a>
Fos	<a href="#">FOS HUMAN</a>
Jun	<a href="#">AP1 HUMAN</a>
MEK1	<a href="#">MP2K1 HUMAN</a>
Erk1	<a href="#">MK03 HUMAN</a>
SHIP	<a href="#">O00145 HUMAN</a>
AKT1	<a href="#">AKT1 HUMAN</a>
p62DOK	<a href="#">DOK1 HUMAN</a>
p120GAP	<a href="#">RASA1 HUMAN</a>
IKK-alpha	<a href="#">IKKA HUMAN</a>

For questions, please contact Dr. Jake Y. Chen ([jakechen@iupui.edu](mailto:jakechen@iupui.edu)) or Harini N Kasamsetty ([hkasamse@iupui.edu](mailto:hkasamse@iupui.edu))  
 All Rights Reserved. Copyright © 2007 by Discovery Informatics and Computing Group, Indiana University.

Number Of Proteins in the Pathway: 68

Protein Name	UniprotId
MEKK1	<a href="#">M3K1 HUMAN</a>
PI3K catalytic alpha polypeptide	<a href="#">PK3CA HUMAN</a>
PI3K regulatory subunit polypeptide 1	<a href="#">P85A HUMAN</a>
SYK	<a href="#">KSYK HUMAN</a>
Lyn	<a href="#">LYN HUMAN</a>
Btk	<a href="#">BTK HUMAN</a>
Grb2	<a href="#">GRB2 HUMAN</a>
SHC	<a href="#">SHC1 HUMAN</a>
RAF1	<a href="#">RAF1 HUMAN</a>
Fos	<a href="#">FOS HUMAN</a>
Jun	<a href="#">AP1 HUMAN</a>
MEK1	<a href="#">MP2K1 HUMAN</a>
Erk1	<a href="#">MK03 HUMAN</a>
SHIP	<a href="#">O00145 HUMAN</a>
AKT1	<a href="#">AKT1 HUMAN</a>
p62DOK	<a href="#">DOK1 HUMAN</a>
p120GAP	<a href="#">RASA1 HUMAN</a>
IKK-alpha	<a href="#">IKKA HUMAN</a>

(a)

(b)

Figure 4.6 (a) Screenshot of proteins involved in the pathway page. (b) Zoomed view from oval region in (a).

Discovery Informatics and Computing Group  
Indiana University School of Informatics, Indianapolis  
Purdue University School of Science Dept of Computer & Information Science, Indianapolis

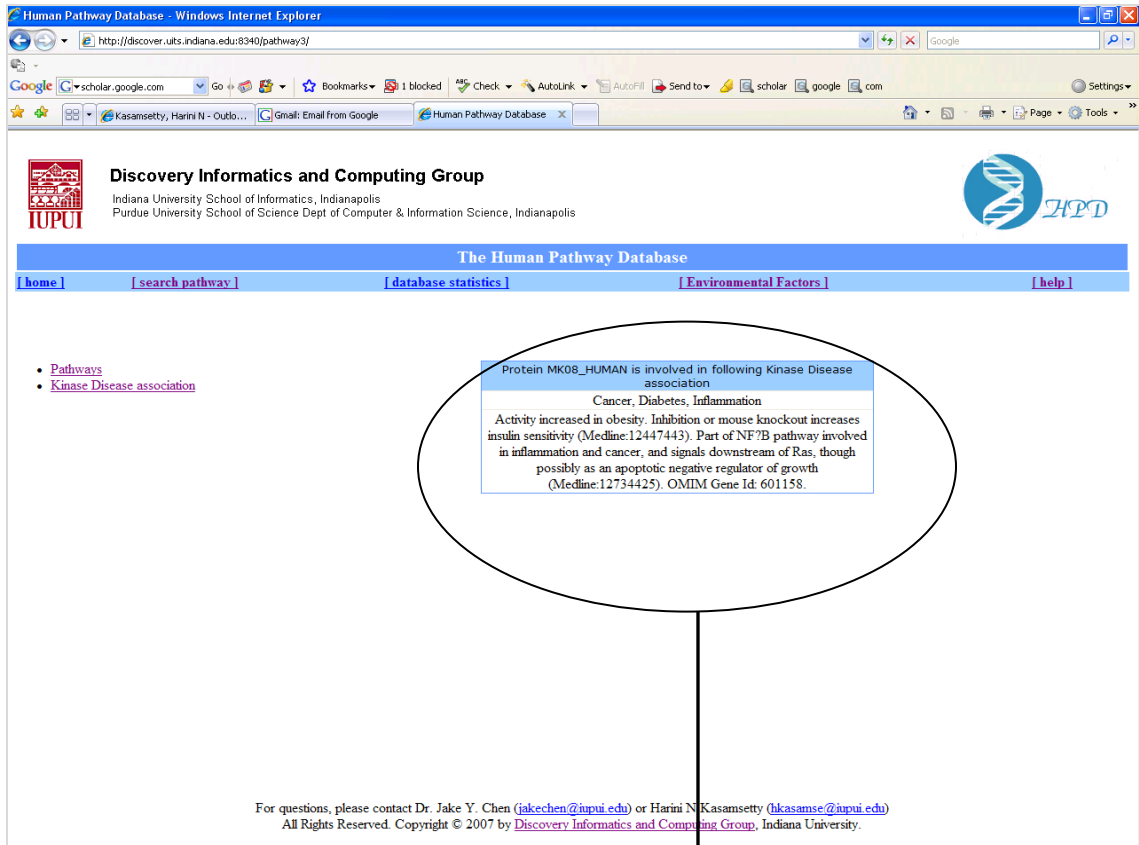
The Human Pathway Database

Proteins  
Complex  
Compound  
Diagram

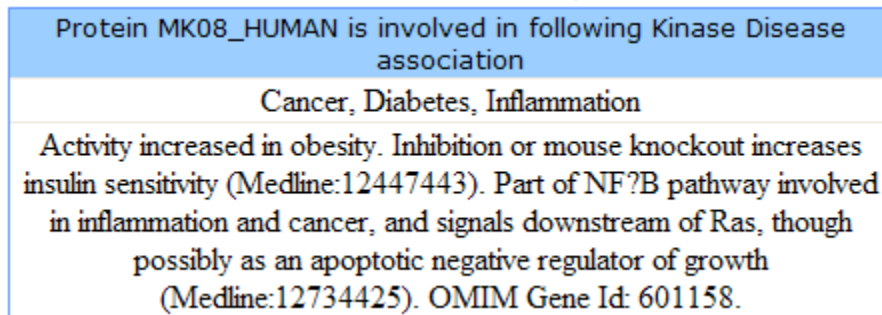
For questions, please contact Dr. Jake Y. Chen ([jakechen@iupui.edu](mailto:jakechen@iupui.edu)) or Harini N Kasamsetty ([hkasamse@iupui.edu](mailto:hkasamse@iupui.edu))  
All Rights Reserved. Copyright © 2007 by Discovery Informatics and Computing Group, Indiana University.

**Figure 4.7 Screenshot of the pathway diagram page.**

The kinase-disease association link takes the user to the details and references about the involved diseases as shown in Figure 4.8.



(a)

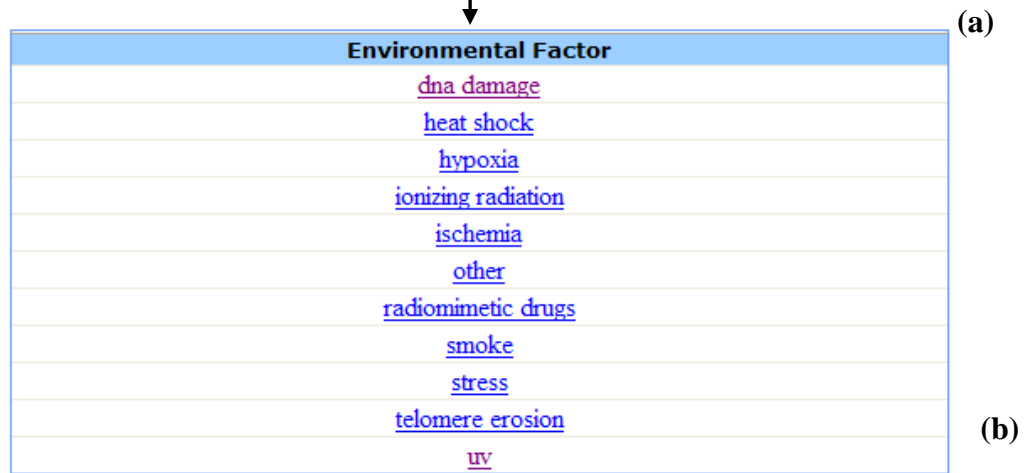
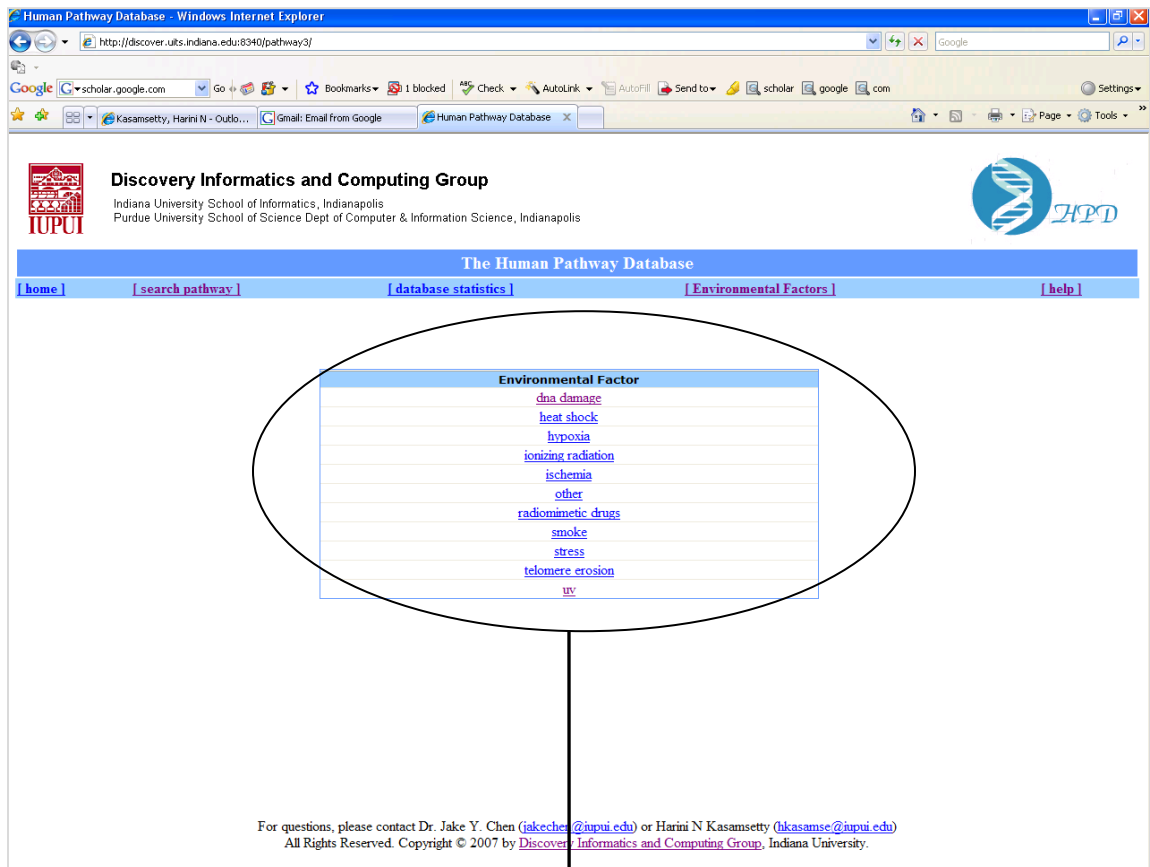


(b)

**Figure 4.8 (a) Screenshot of the kinase-disease association page. (b) Zoomed view from oval region in (a).**

The effect of environmental factors on pathways has also been integrated into HPD. Pathways are grouped by environmental factors, and the environmental factors tab on the HPD interface lists the most common environmental factors that affect the

pathways (Figure 4.9). Selection of an environmental factor lists the pathways that are affected (Figure 4.10).



**Figure 4.9 (a) Screenshot of list of environmental factors. (b) Zoomed view from oval region in (a).**

Human Pathway Database - Windows Internet Explorer

http://discover.uts.indiana.edu:8340/pathway3/

Discovery Informatics and Computing Group  
 Indiana University School of Informatics, Indianapolis  
 Purdue University School of Science Dept of Computer & Information Science, Indianapolis

The Human Pathway Database

[ home ] [ search pathway ] [ database statistics ] [ Environmental Factors ] [ help ]

10 Pathway(s) found

Pathway Name
<a href="#">14-3-3 and Cell Cycle Regulation</a>
<a href="#">Ceramide Pathway</a>
<a href="#">JNK pathway</a>
<a href="#">NF-KappaB (p50/p65) Pathway</a>
<a href="#">Repair of Thymine Dimers</a>
<a href="#">UVA-Induced MAPK Signaling</a>
<a href="#">UVB-Induced MAPK Signaling</a>
<a href="#">UVC-Induced MAPK Signaling</a>
<a href="#">p38 signaling</a>
<a href="#">p53 signaling</a>

For questions, please contact Dr. Jake Y. Chen ([jakechen@iupui.edu](mailto:jakechen@iupui.edu)) or Harini N Kasamsetty ([hkasamse@iupui.edu](mailto:hkasamse@iupui.edu))  
 All Rights Reserved. Copyright © 2007 by [Discovery Informatics and Computing Group](#), Indiana University.

(a)

10 Pathway(s) found
Pathway Name
<a href="#">14-3-3 and Cell Cycle Regulation</a>
<a href="#">Ceramide Pathway</a>
<a href="#">JNK pathway</a>
<a href="#">NF-KappaB (p50/p65) Pathway</a>
<a href="#">Repair of Thymine Dimers</a>
<a href="#">UVA-Induced MAPK Signaling</a>
<a href="#">UVB-Induced MAPK Signaling</a>
<a href="#">UVC-Induced MAPK Signaling</a>
<a href="#">p38 signaling</a>
<a href="#">p53 signaling</a>

(b)

Figure 4.10 (a) Screenshot of list of pathways affected by environmental factors. (b)

Zoomed view from oval region in (a).

Further selections on listed pathways provide information about the effects that are due to the environmental factor. For example, when the environmental factor UV is selected, the HPD interface lists 10 pathways that are affected by UV. Further selection of the ‘p53 signaling’ pathway lists the effects of ‘apoptosis’ and ‘angiogenesis inhibition’.

Table 4.4 presents a comparison of interface features of HPD with other sources. As we can see in Table 4.4, the most unique features of HPD are kinase-disease associations and the effect of environmental factors when compared to other source databases. HPD is designed to be both a resource for the laboratory scientist to explore known and predicted pathways, and to facilitate bioinformatics initiatives exploring pathway networks.

**Table 4.4 Comparison of HPD database features with the other sources**

<b>Source</b> <b>Feature</b>	<b>NCI- Nature Curated data</b>	<b>BIOCARTA</b>	<b>PROTEIN LOUNGE</b>	<b>RESNET</b>	<b>HPD</b>
List of pathway items			✓	✓	✓
Categorization of pathway items (proteins, complexes and compounds interface)				✓	✓
Pathway diagram	✓	✓	✓	✓	✓
Kinase- disease Annotation					✓
Effect of environmental factors on the pathways			✓		✓

## CHAPTER FIVE: CASE STUDIES

### 5.1 Case Study 1: Alzheimer's Disease

This chapter is a case study of Alzheimer's disease that starts with expanding a seed list of Alzheimer's disease-proteins with the pathway data from HPD. The expanded set of proteins is then expanded further with protein – protein interactions from the HAPPI database. The analysis then proceeds to overlay the twice-expanded set with gene expression data. A network analysis that includes visualization of integrated networks is then conducted to find significant proteins related to Alzheimer's disease. This work represents an approach to a common challenge in the field of biology where new opportunities for discovery are arrived at by integration of publicly available databases.

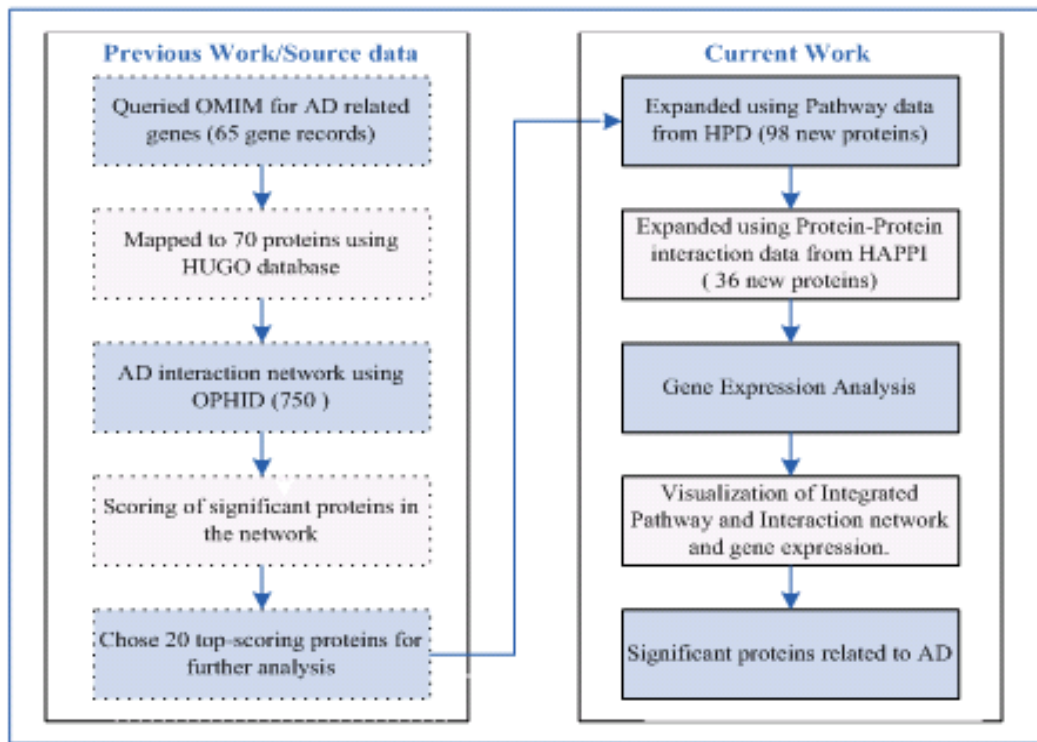
#### 5.1.1. Introduction

A flowchart for the case study of Alzheimer's disease proteins using the integrated HPD is shown in Figure 6.1. Alzheimer's disease is a progressive and fatal neurodegenerative disease and is the most common form of dementia. Today it is the seventh-leading cause of death in the United States [28]. More than 5 million Americans have been diagnosed with Alzheimer's disease, and the number of diagnoses is expected to quadruple in the next 40 years.

AD destroys brain cells and causes problems with memory, thinking and behavior. Plaques and tangles are the most common abnormalities associated with the disease and are prime suspects in the damaging and killing of nerve cells. Plaques are deposits of beta-amyloid protein fragments that build up between nerve cells. Tangles are twisted fibers of tau protein that form inside dying cells.

Even healthy persons develop plaques and tangles with age, but those with

Alzheimer's disease tend to develop a greater degree of plaques and tangles. In the initial stages of Alzheimer's disease, plaques and tangles tend to develop in areas important in learning and memory and then spread to other regions. There is not yet a cure for Alzheimer's disease, however patients with Alzheimer's disease can get temporary relief from a few drugs. For example, Tocopherol is a drug for Alzheimer's disease that acts as an antioxidant and delays the damage to the nerve cells, and this inspires the question of whether there may be a better way to treat the disease, delay its onset, or prevent it from developing.



**Figure 5.1 Flow Chart for case study.**

In this study, I have made an attempt to build a network of linking between proteins with both pathway data and protein interaction data for a seed list of proteins directly implicated with Alzheimer's disease. Gene expression analysis is also used to



enrich this network and visualization is used to provide comparative views of new information arising from this overall effort shown in Figure 5.1.

### 5.1.2. Constructing Alzheimer's disease integrated network

#### 5.1.2.1 Data Set of Alzheimer's disease Related Genes

The 20 top scoring proteins related to Alzheimer's disease (AD) were obtained from previous work [29]. Initially, the genes related to Alzheimer's disease were retrieved from the OMIM database [3] in which the "description" field contains the term "Alzheimer" by performing a search. 65 OMIM gene records were retrieved. The genes were then mapped to their protein identifiers. 70 Alzheimer's disease-related proteins were obtained after mapping gene symbols to protein SwissProt IDs. The increase in protein count from gene record count is due to one-to-many mappings between a gene and its multiple splice variant forms at the protein level.

In the work by [2], the Online Predicted Human Interaction Database (OPHID) [30] was used to expand the initial set based on protein interaction data. Protein interacting pairs were drawn such that at least one member of the pair belongs to the seed-AD-set. This set of interacting pairs was called the Alzheimer's disease-interaction-set. The proteins that were expanded from the initial seed-AD-set by new proteins involved in the Alzheimer's disease interaction set were denoted as the enriched-AD-set (a superset of seed-AD-set). The Alzheimer's disease-interaction-set contains 775 human protein interactions and the enriched-AD-set contains 657 human proteins identified by SwissProt IDs. Next, the relevance score was calculated for each protein in the enriched Alzheimer's disease-set [29]. We selected 20 top-scoring proteins for this case study (see Table 5.1.).

**Table 5.1 20 top-scoring proteins of Alzheimer’s disease**

<b>UniProt id</b>	<b>Rank</b>
A4_HUMAN	1
LRP1_HUMAN	2
PSN1_HUMAN	3
PIN1_HUMAN	4
FHL2_HUMAN	5
PSN2_HUMAN	6
NP1L1_HUMAN	7
S100B_HUMAN	8
CDK5_HUMAN	9
NOG1_HUMAN	10
CLUS_HUMAN	11
NCOA6_HUMAN	12
CATB_HUMAN	13
ARLY_HUMAN	14
FLNB_HUMAN	15
CTND2_HUMAN	16
APBA1_HUMAN	17
C1TC_HUMAN	18
ODO2_HUMAN	19
MK10_HUMAN	20

5.1.2.2 Generation of Pathway Network for Alzheimer’s disease proteins

The Alzheimer’s disease pathway reaction network was generated by expanding the 20 top-scoring Alzheimer’s disease related proteins with pathway reactions from HPD database such that at least one member of the 20 top-scoring proteins belongs to a pathway in HPD. I chose the NCI-Nature curated database and the Biocarta database to generate a pathway reaction network since these two databases provide directed reactions for the pathways. Of 20 top-scoring proteins, 4 proteins are found in 16 pathways, such that at least one member of the 20 top-scoring proteins belongs to a pathway (Table.5.2). To explain the significance of these pathways in Alzheimer’s disease, I chose an example

of selected presenilin action in the Notch and Wnt signaling pathway. This pathway plays a major role in brain development. Studies have established several characteristics of this pathway such as how 1) PS proteins function as components of Notch signal transduction, 2)  $\beta$ -catenin and GSK-3 $\beta$  are transducers of the Wnt signaling pathway, and 3)  $\beta$ -catenin and GSK-3 $\beta$  are connected through the Dishevelled (Dvl) protein, a known transducer of the Wnt pathway [31].

**Table.5.2 Pathways involved using 20 top-scoring proteins**

Pathway_Name	Source
caspase cascade in apoptosis	NCI- Nature curated Data
lissencephaly gene (lis1) in neuronal migration and development	NCI- Nature curated Data
pdgfrb signaling pathway	NCI- Nature curated Data
presenilin action in notch and wnt signaling	NCI- Nature curated Data
reelin signaling pathway	NCI- Nature curated Data
role of hdac class iii	NCI- Nature curated Data
bioactive peptide induced signaling pathway	Biocarta
deregulation of cdk5 in alzheimers disease	Biocarta
fosb gene expression and drug abuse	Biocarta
hiv-1 nef: negative effector of fas and tnf	Biocarta
How progesterone initiates the oocyte maturation	Biocarta
lissencephaly gene (lis1) in neuronal migration and development	Biocarta
mapkinase signaling pathway	Biocarta
phosphorylation of mek1 by cdk5/p35 down regulates the map kinase pathway	Biocarta
rac1 cell motility signaling pathway	Biocarta
regulation of ck1/cdk5 by type 1 glutamate receptors	Biocarta

The 20 top-scoring proteins are expanded to 127 proteins using the pathway data. Among these 127 proteins, I found 98 new proteins that are not present in the seed set and enriched-AD-set (Table 5.3).

**Table.5.3 List of new proteins obtained after expansion using pathway data**

UNIPROT ID	ACC1	HGNC SYMBOL
2A5D_HUMAN	Q14738	PPP2R5D
ACTS_HUMAN	P68133	(null)
ADCY1_HUMAN	Q08828	ADCY1
AG22_HUMAN	P50052	AGTR2
AKT1_HUMAN	P31749	AKT1
ANGT_HUMAN	P01019	AGT
AP2A_HUMAN	P05549	TFAP2A
APAF_HUMAN	O14727	APAF1
APC_HUMAN	P25054	APC
AXN1_HUMAN	O15169	AXIN1
BID_HUMAN	P55957	BID
CAP1_HUMAN	Q01518	CAP1
CASP2_HUMAN	P42575	CASP2
CASP6_HUMAN	P55212	CASP6
CASP9_HUMAN	P55211	CASP9
CASPA_HUMAN	Q92851	CASP10
COF2_HUMAN	Q9Y281	CFL2
CRKL_HUMAN	P46109	CRKL
CYC_HUMAN	P99999	(null)
DAXX_HUMAN	Q9UER7	DAXX
DBLOH_HUMAN	Q9NR28	DIABLO
DVL1_HUMAN	O14640	DVL1
DYN2_HUMAN	P50570	DNM2
EGR1_HUMAN	P18146	EGR1
ERBB4_HUMAN	Q15303	ERBB4
FADD_HUMAN	Q13158	FADD
FAK1_HUMAN	Q05397	PTK2
FAK2_HUMAN	Q14289	PTK2B
FBW1A_HUMAN	Q9Y297	BTRC
GAB1_HUMAN	Q13480	GAB1
GAS2_HUMAN	O43903	GAS2
GDIS_HUMAN	P52566	ARHGDIB

GNA11_HUMAN	P29992	GNA11
GRAB_HUMAN	P10144	GZMB
H4_HUMAN	P62805	(null)
HIF1A_HUMAN	Q16665	HIF1A
IKKA_HUMAN	O15111	CHUK
IKKB_HUMAN	O14920	IKBKB
IPPD_HUMAN	Q9UD71	PPP1R1B
K1C18_HUMAN	P05783	KRT18
KC1A_HUMAN	P48729	CSNK1A1
KC1D_HUMAN	P48730	CSNK1D
KS6A1_HUMAN	Q15418	RPS6KA1
LIMK1_HUMAN	P53667	LIMK1
LIS1_HUMAN	P43034	PAFAH1B1
M3K13_HUMAN	O43283	MAP3K13
M3K14_HUMAN	Q99558	MAP3K14
M3K1_HUMAN	Q13233	MAP3K1
M3K5_HUMAN	Q99683	MAP3K5
M3K7_HUMAN	O43318	MAP3K7
MAP1B_HUMAN	P46821	MAP1B
MK03_HUMAN	P27361	MAPK3
MK07_HUMAN	Q13164	MAPK7
MK14_HUMAN	Q16539	MAPK14
MLRV_HUMAN	P10916	MYL2
MP2K1_HUMAN	Q02750	MAP2K1
MP2K2_HUMAN	P36507	MAP2K2
MP2K3_HUMAN	P46734	MAP2K3
MP2K4_HUMAN	P45985	MAP2K4
MP2K5_HUMAN	Q13163	MAP2K5
MP2K7_HUMAN	O14733	MAP2K7
MYLK_HUMAN	Q15746	MYLK
MYOD1_HUMAN	P15172	MYOD1
MYPT2_HUMAN	O60237	PPP1R12B
NLK_HUMAN	Q9UBE8	NLK
NUMA1_HUMAN	Q14980	NUMA1
O00211_HUMAN	O00211	(null)
P2BA_HUMAN	Q08209	PPP3CA
P85A_HUMAN	P27986	PIK3R1
PAK1_HUMAN	Q13153	PAK1
PARP1_HUMAN	P09874	PARP1
PCAF_HUMAN	Q92831	PCAF

PERF_HUMAN	P14222	PRF1
PGFRB_HUMAN	P09619	PDGFRB
PK3CA_HUMAN	P42336	PIK3CA
PLCG1_HUMAN	P19174	PLCG1
Q53ZZ1_HUMAN	Q53ZZ1	FASLG
Q59FU8_HUMAN	Q59FU8	FAS
Q5TC39_HUMAN	Q5TC39	(null)
RAF1_HUMAN	P04049	RAF1
RASH_HUMAN	P01112	HRAS
RELN_HUMAN	P78509	RELN
RPGF1_HUMAN	Q13905	RAPGEF1
SATB1_HUMAN	Q01826	SATB1
SIRT1_HUMAN	Q96EB6	(null)
SOS1_HUMAN	Q07889	SOS1
SPTA2_HUMAN	Q13813	SPTAN1
SRBP1_HUMAN	P36956	SREBF1
TAB1_HUMAN	Q15750	MAP3K7IP1
TNFA_HUMAN	P01375	TNF
VAV2_HUMAN	P52735	VAV2
VIME_HUMAN	P08670	VIM

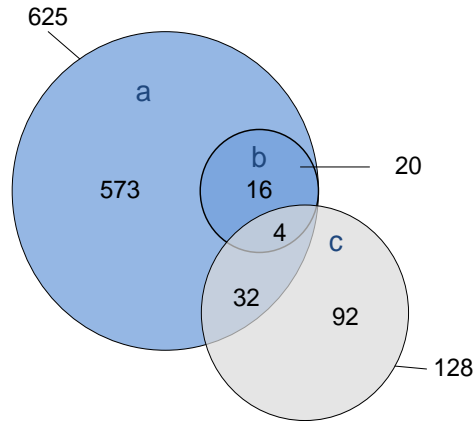
These new proteins are related to Alzheimer's disease. We further explored the pathway network and analyzed it using gene expression data from gene expression profiles.

### 5.1.2.3 Further expansion using HAPPI data

I expanded the above pathway network using protein-protein interaction data from HAPPI database such that one member of the pathway reaction pair belongs to a protein-protein interaction in HAPPI database. The protein-protein interaction data from HAPPI database enabled further expansion of the network to encompass 149 proteins.

#### 5.1.2.4 Overlap of Proteins among Networks

Overlapping analysis was used to show those proteins that overlap between the OPHID and HPD expanded networks as is shown in Figure 5.2.



- [a] 65 AD gene records from OMIM expanded with OPHID to 625 proteins
- [b] The top 20 scoring proteins from OPHID expansion [2]
- [c] 4 out of the top 20 scoring proteins expanded using HPD.

**Figure.5.2 Overlap of proteins between the networks.**

The above figure represents the overlap of proteins between the OPHID and pathway networks. 36 proteins overlap between OPHID and HPD expansion networks. Four of these 36 proteins, A4\_HUMAN, CDK5\_HUMAN, PIN1\_HUMAN, and PSN1\_HUMAN, are also present in the set of 20 top scoring proteins [29]. Three of these four proteins, A4\_HUMAN, PSN1\_HUMAN, and PIN1\_HUMAN, are present in the top 4 list.

#### 5.1.3 Gene expression analysis

The Alzheimer's disease gene expression data obtained from a published expression microarray data set, derived from a microarray analysis of brain tissues from 31 individuals, including 9 healthy individuals, 7 incipient Alzheimer's disease patients, 8 moderate Alzheimer's disease patients, and 7 severe Alzheimer's disease patients [32].

The gene expression value for each gene is calculated from gene-mapped probe sets identified by its AFF\_ID [32] and contains a single gene expression value. I mapped each probe set gene expression value to a gene. The statistical average is taken to represent the aggregated expression value if multiple probe sets map to one gene.

To discover Alzheimer’s disease related proteins which may act as biomarkers or drug targets in the Alzheimer’s disease pathway network, I calculated the differential expression levels (as fold changes) for each gene [33]. This analysis involves the following steps:

1. Calculate the average gene expression for each group, for the genes from pathway and protein-protein interaction expansion list.
2. Calculate relative gene expression for the pairs of Alzheimer’s disease patient groups (incipient, moderate, and severe) with the normal control group.

Relative gene expression values are calculated according to standard gene expression analysis conventions using the following formula [33]:

$$\text{Re Exp}(pro\_id) = \begin{cases} \frac{\text{Exp2}(pro\_id)}{\text{Exp1}(pro\_id)}, & \text{Exp2}(pro\_id) \geq \text{Exp1}(pro\_id) \\ -\frac{\text{Exp1}(pro\_id)}{\text{Exp2}(pro\_id)}, & \text{Exp2}(pro\_id) < \text{Exp1}(pro\_id) \end{cases}$$

ReExp (pro\_id) represents the differential gene expression ratio for the diseased stage versus the normal control condition for a given protein with pro\_id as the gene identifier, Exp1 (pro\_id) is the absolute gene expression value for the same protein under condition 1, and Exp2 (pro\_id) is the absolute gene expression value for the same protein under condition 2. I chose a threshold of 1.5 to filter significant differential gene expression values due to natural variability of gene expressions. Tables 6.4 & 6.5 list proteins/genes



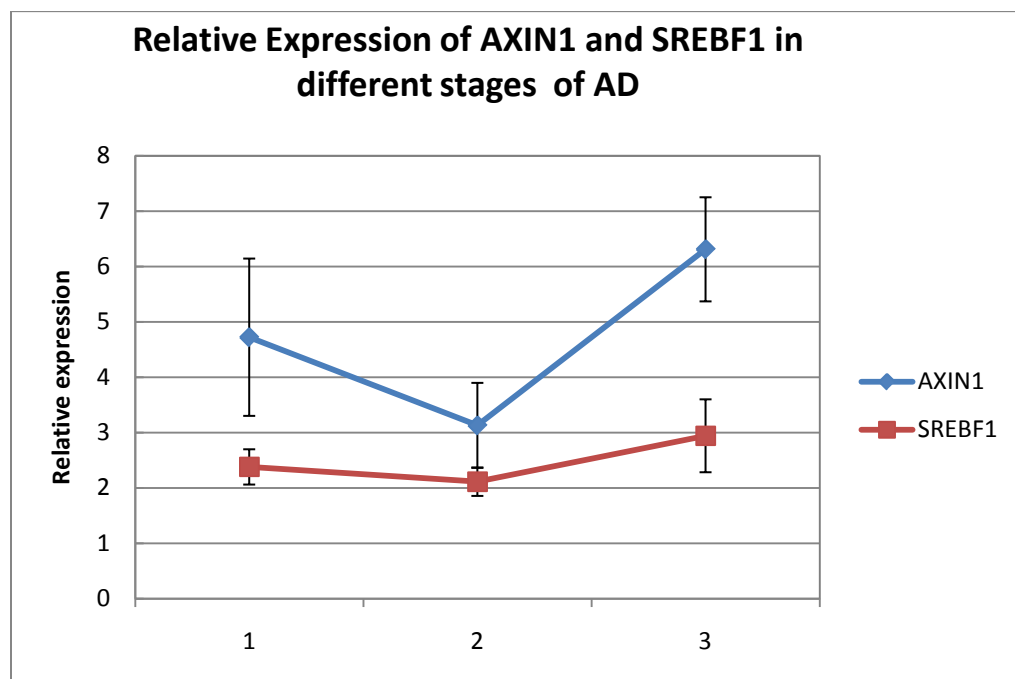
(here, I refer to the two distinct molecular entities interchangeably, because I use a standard ID mapping table available from the UniProt database [34] and can map between genes identified by standard gene symbols and corresponding proteins identified by unique UniProt identifiers) that are up-regulated and down-regulated with the relative gene expressions in pathway networks respectively.

**Table 5.4 List of up-regulated proteins in integrated network data sources**

<b>Gene Symbol</b>	<b>UniProt Id</b>	<b>Relative expression</b>	<b>Disease state</b>	<b>Source</b>
ARHGDIB	GDIS_HUMAN	1.517	Incipient	HPD
<b>AXIN1</b>	<b>AXN1_HUMAN</b>	<b>4.722</b>	<b>Incipient</b>	<b>HPD</b>
<b>AXIN1</b>	<b>AXN1_HUMAN</b>	<b>3.576</b>	<b>Moderate</b>	<b>HPD</b>
<b>AXIN1</b>	<b>AXN1_HUMAN</b>	<b>6.311</b>	<b>Severe</b>	<b>HPD</b>
BID	BID_HUMAN	1.507	Moderate	HPD
KAPCB	PRKACB	1.593	Moderate	HAPPI
KAPCG	PRKACG	1.556	Incipient	HAPPI
KAPCG	PRKACG	2.878	Moderate	HAPPI
KAPCG	PRKACG	2.364	Severe	HAPPI
KNG1	KNG1	1.536	Severe	HAPPI
MAP2K1	MP2K1_HUMAN	1.547	Incipient	HPD
MAP2K2	MP2K2_HUMAN	1.528	Moderate	HPD
MAP2K2	MP2K2_HUMAN	1.611	Severe	HPD
MAPK3	MK03_HUMAN	1.536	Moderate	HPD
PAK1	PAK1_HUMAN	2.431	Incipient	HPD
PAK1	PAK1_HUMAN	1.938	Severe	HPD
RELN	RELN_HUMAN	1.56	Moderate	HPD
SATB1	SATB1_HUMAN	1.606	Moderate	HPD
<b>SREBF1</b>	<b>SRBP1_HUMAN</b>	<b>2.378</b>	<b>Incipient</b>	<b>HPD</b>
<b>SREBF1</b>	<b>SRBP1_HUMAN</b>	<b>2.413</b>	<b>Moderate</b>	<b>HPD</b>
<b>SREBF1</b>	<b>SRBP1_HUMAN</b>	<b>2.941</b>	<b>Severe</b>	<b>HPD</b>
VAV2	VAV2_HUMAN	2.262	Incipient	HPD
VAV2	VAV2_HUMAN	2.443	Moderate	HPD

In the above table (Table 5.4), we listed all the genes that are up regulated at different stages (incipient, moderate and severe) of Alzheimer’s disease. We found

AXIN1 and SREBF1 genes have significant up regulation in Alzheimer's disease i.e. relative expression is 2 fold difference when compared to normal patients. A graph is plotted against relative expression of AXIN1 and SREBF1 and different stages of Alzheimer's disease patients with sample size of 7 or 8 for incipient, moderate, severe patients as shown in Figure 5.3. There is significant increase in relative expression from moderate to severe. We studied these two proteins for the validation based on the already existing work.



**Figure.5.3 Relative expressions of AXIN1 and SREBF1 in different stages of AD and Standard error bars are shown.**

AXIN1 is known as the Axis inhibition protein. Substitution of certain residues on the conserved region of one of the variants, the *AXIN1 D545E* protein, leads to the binding of  $\beta$ -catenin and negatively regulates the Wnt signaling pathway by interacting with GSK-3 $\beta$  and  $\beta$ -catenin; mediating the signal from GSK-3beta to beta-catenin

ultimately leads to beta-catenin degradation [35]. This substitution has been described in patients with sporadic medulloblastomas [36] and the variant *AXINI G700S* substitution leads to binding to the catalytic subunit of protein phosphatase 2A [37]. The variant *AXINI G700S* has been described in patients with both hepatoblastoma and hepatocellular carcinoma.

The SREBF1 transcription factor controls the expression of most enzymes of cholesterol synthesis. It has been associated with atherosclerosis or high cholesterol levels. Polymorphisms within the SREBF1 in Alzheimer's disease are also likely to affect cholesterol and lipoprotein status in the periphery, via diverse metabolic pathways, and in so doing may well contribute to atherosclerotic pathology [38]. Also, the SREBF1 transcription factor is found to be up-regulated in Alzheimer's disease.

**Table.5.5 List of down-regulated proteins in pathway network**

<b>Gene Symbol</b>	<b>Uni Prot Id</b>	<b>Relative expression</b>	<b>Disease state</b>	<b>Data set</b>
AKT1	AKT1_HUMAN	-1.542	Severe	HPD
DOCK3	DOCK3	-1.978	Severe	HAPPI
GD	SERPINE2	-1.635	Severe	HAPPI
MYL2	MLRV_HUMAN	-1.708	Incipient	HPD
MYL2	MLRV_HUMAN	-1.602	Moderate	HPD
MYOD1	MYOD1_HUMAN	-2.024	Moderate	HPD
PCAF	PCAF_HUMAN	-1.569	Incipient	HPD
PTK2B	FAK2_HUMAN	-1.912	Severe	HPD

Table 5.5 shows the genes that are down-regulated at different stages of Alzheimer's disease. PCAF and MYOD1 form a complex with SIRT1. SIRT1 activation extends lifespan and promotes longevity and healthy aging in a variety of species, potentially delaying the onset of age-related neurodegenerative disorders. In mammalian

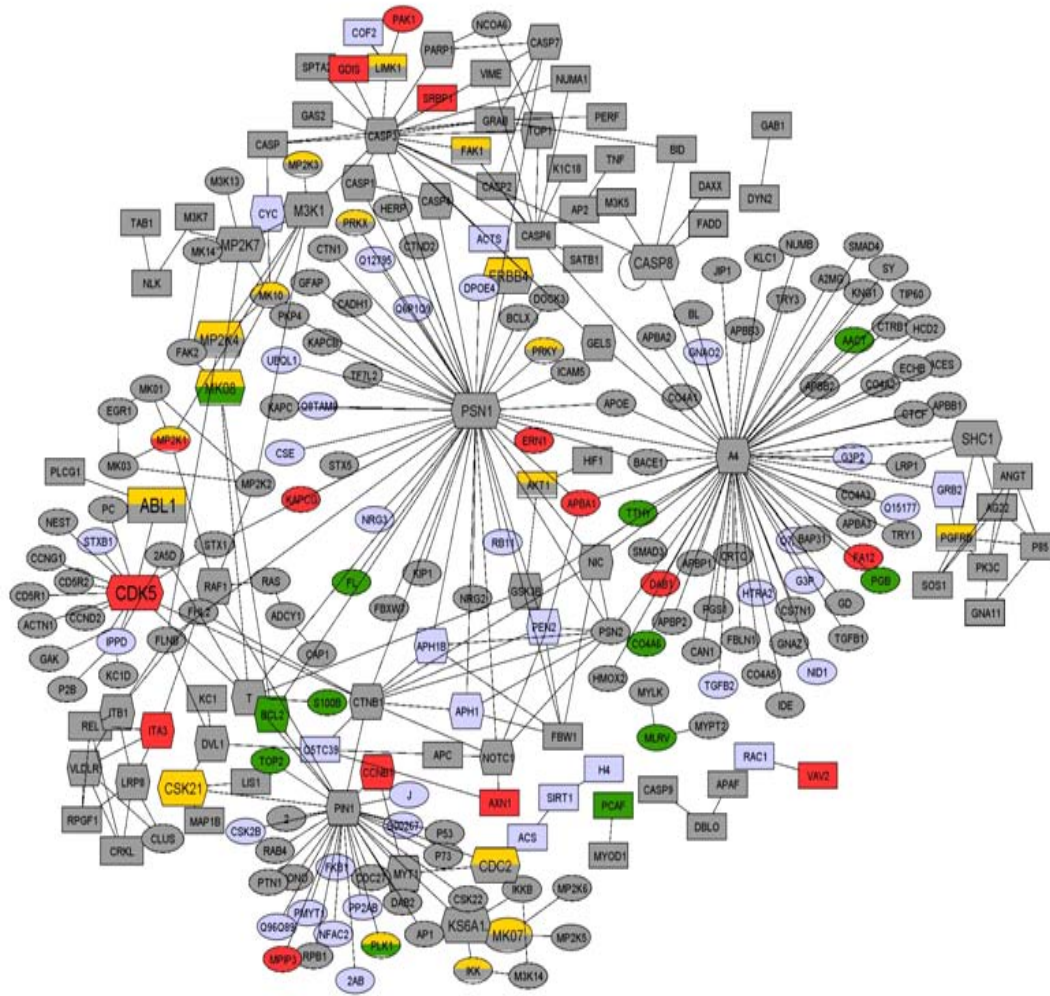
systems, sirtuin activators protect against axonal degeneration, poly-glutamine toxicity and microglia-mediated amyloid beta toxicity. It suggests there to be potential therapeutic value of sirtuins in patients with neurodegenerative diseases, such as Alzheimer's disease. In this regard, PCAF and MYOD1 may be involved in the activation of SIRT1.

This approach helps in eliminating non specific genes. Usually in microarray analysis there are lots of genes which are up and down regulated. It is difficult to find those genes which have some biological significance and there may be lots of up regulated genes which may be false positive. So in this approach we try to integrate the pathway network to find those specific genes which are also up regulated in the microarray data. So in this way we increase the specificity of finding genes related to Alzheimer's disease.

#### 5.1.3.1 Visualization

Using the ProteoLens visualization tool [39], I show how the integrated pathway and protein interaction network expands from the seed list of 20 top scoring Alzheimer disease proteins. A 1.5 cut off for relative gene expression is used for incipient, moderate and severe patients as shown for the comparisons of patient groups in the Figures 5.4, 5.5 and 5.6. Figure 5.7 shows a single network from data based on a 1.3 cut off for relative gene expression for all the patients (incipient, moderate and severe).

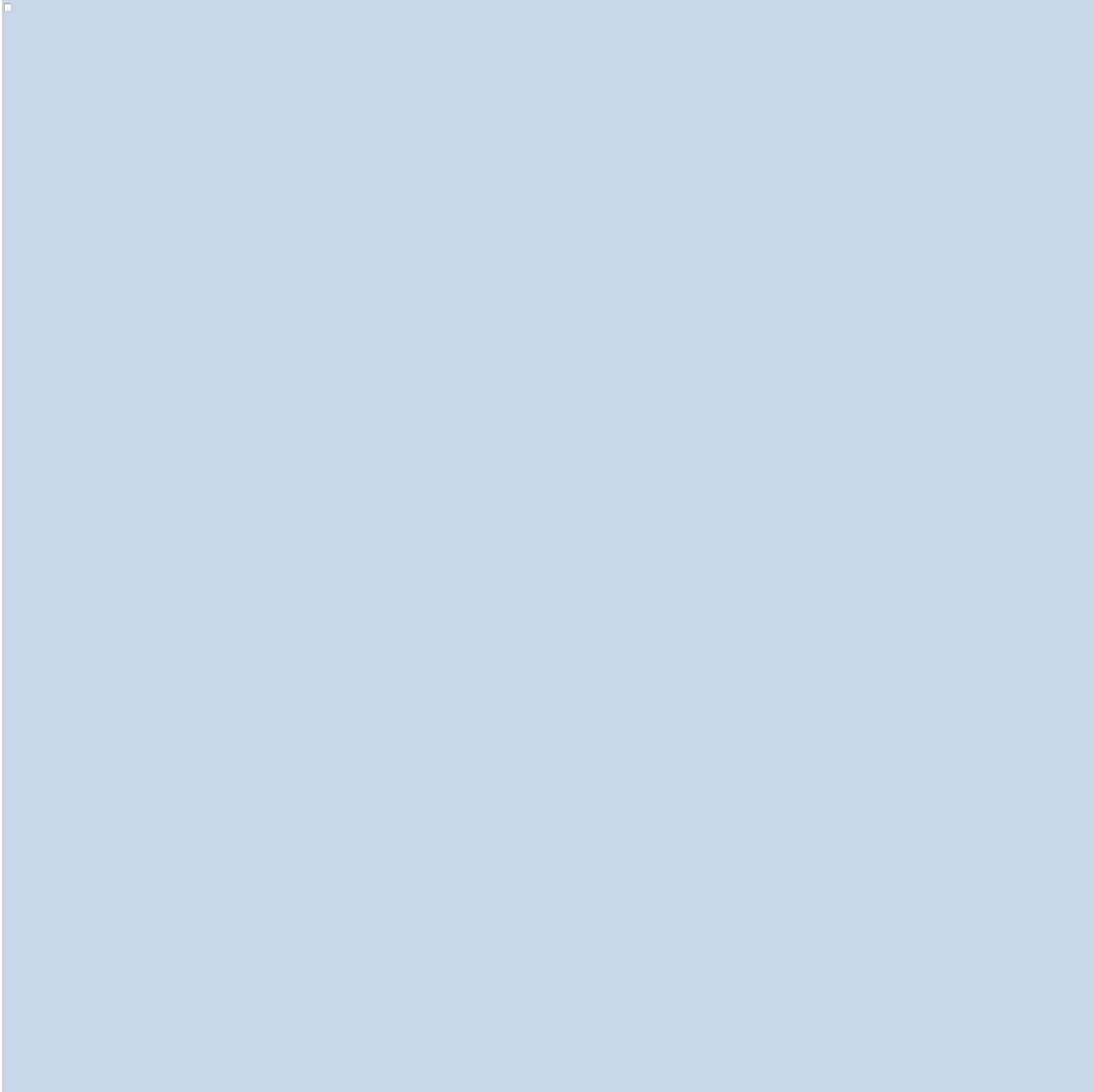
The integrated network with gene expression values of normal vs incipient patients is shown in Figure 5.4.



**Figure 5.4 Integrated pathway and interaction network expanded from 20 top-scoring AD proteins\*, overlaid with gene expression data of incipient patients. Protein node size as shown in proportion to their degree of connectivity in the network. Color legend for expression values and kinase- disease association and shape of the nodes is as follows.**



The integrated network with gene expression values of normal versus moderate patients is shown in Figure 5.5.



**Figure 5.5 Integrated pathway and interaction network** expanded from 20 top-scoring AD proteins\*, overlaid with gene expression data of moderate patients. Protein node size as shown in proportion to their degree of connectivity in the network. Color legend for expression values and kinase - disease association and shape of the nodes is as follows.



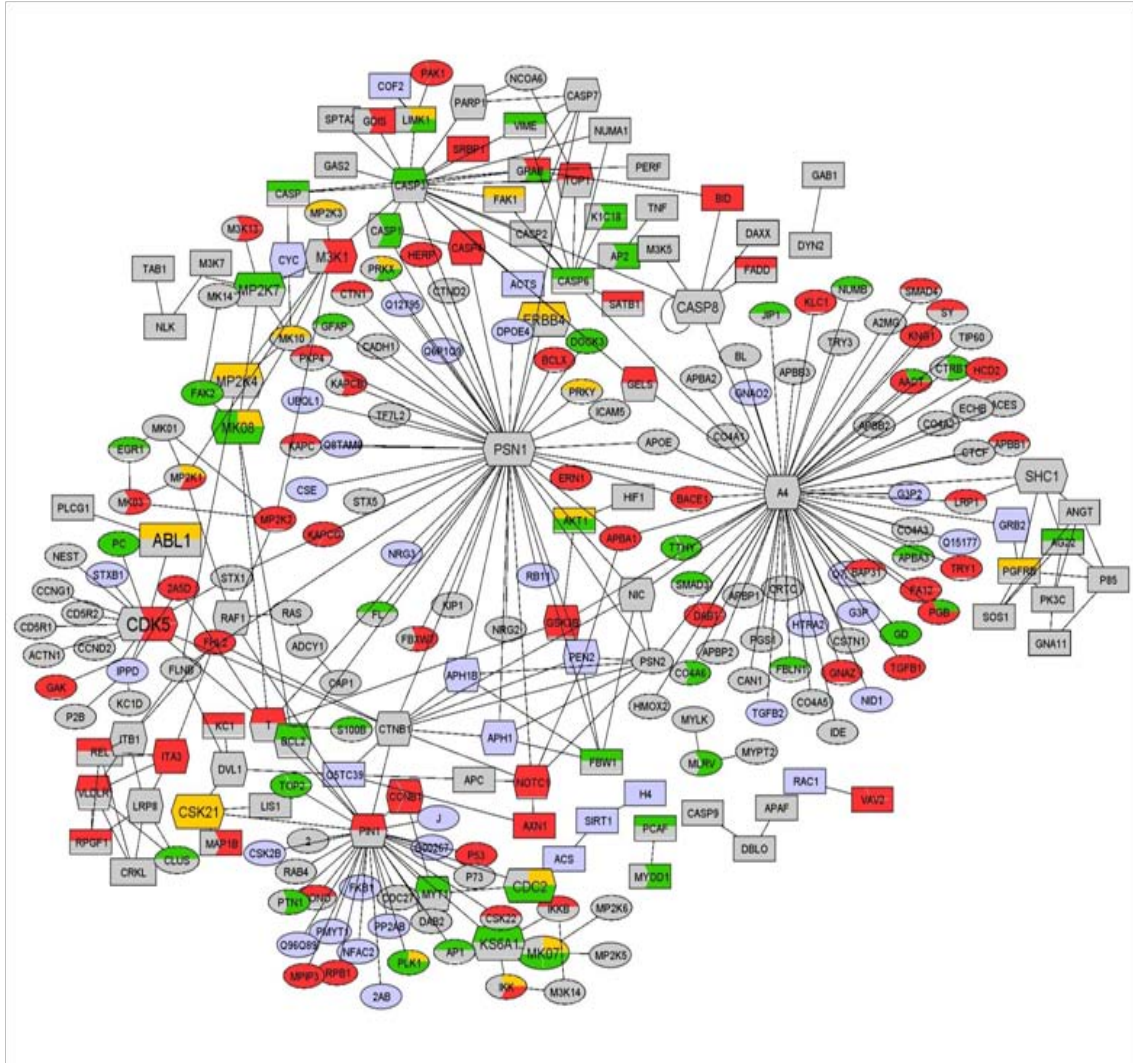
The integrated network with gene expression values of normal versus severe patients is shown in Figure 5.6.

□

**Figure 5.6 Integrated pathway and interaction network expanded from 20 top-scoring AD proteins\*, overlaid with gene expression data of severe patients. Protein node size as shown in proportion to their degree of connectivity in the network. Color legend for expression values and kinase - disease association and shape of the nodes is as follows.**



The integrated network with gene expression values of normal versus all patients is shown in Figure 5.7.



**Figure 5.7** Integrated pathway and interaction network expanded from 20 top-scoring AD proteins\*, overlaid with gene expression data of all(incipient, moderate, severe) patients. Protein node size as shown in proportion to their degree of connectivity in the network. Color legend for expression values and kinase - disease association and shape of the nodes is as follows.

- |  |   |  |   |
|--|---|--|---|
|  | Protein from Protein-Protein interactions   |  | Protein from Pathways                             |
|  | Protein from both pathways and interactions |  | Down regulated protein                            |
|  | Kinase-disease association protein          |  | Protein with no significant change in expression. |
|  |   |  | Up regulated protein                              |



All proteins are shown as nodes. Proteins from pathways are represented as rectangle shaped nodes, proteins from protein-protein interactions are represented as oval shaped nodes, and proteins in both pathways and interactions are represented as hexagon shaped nodes. Expression values up and down are represented using red (up) and green (down). The kinase-disease association nodes are represented using a yellow color. The size of the node is proportional to the number of degree connectivity associated with each protein, counting both pathway and interaction connections. In Figure 6.6, there is an overlap between up-regulated, down-regulated, and “no change” proteins. This is because of different expression values between different stages of Alzheimer’s disease in patients. The overall outcome to this case study of pathway analysis using HPD was the finding of AXIN1 and SRBPF1 proteins to be of potential significance in Alzheimer’s disease.

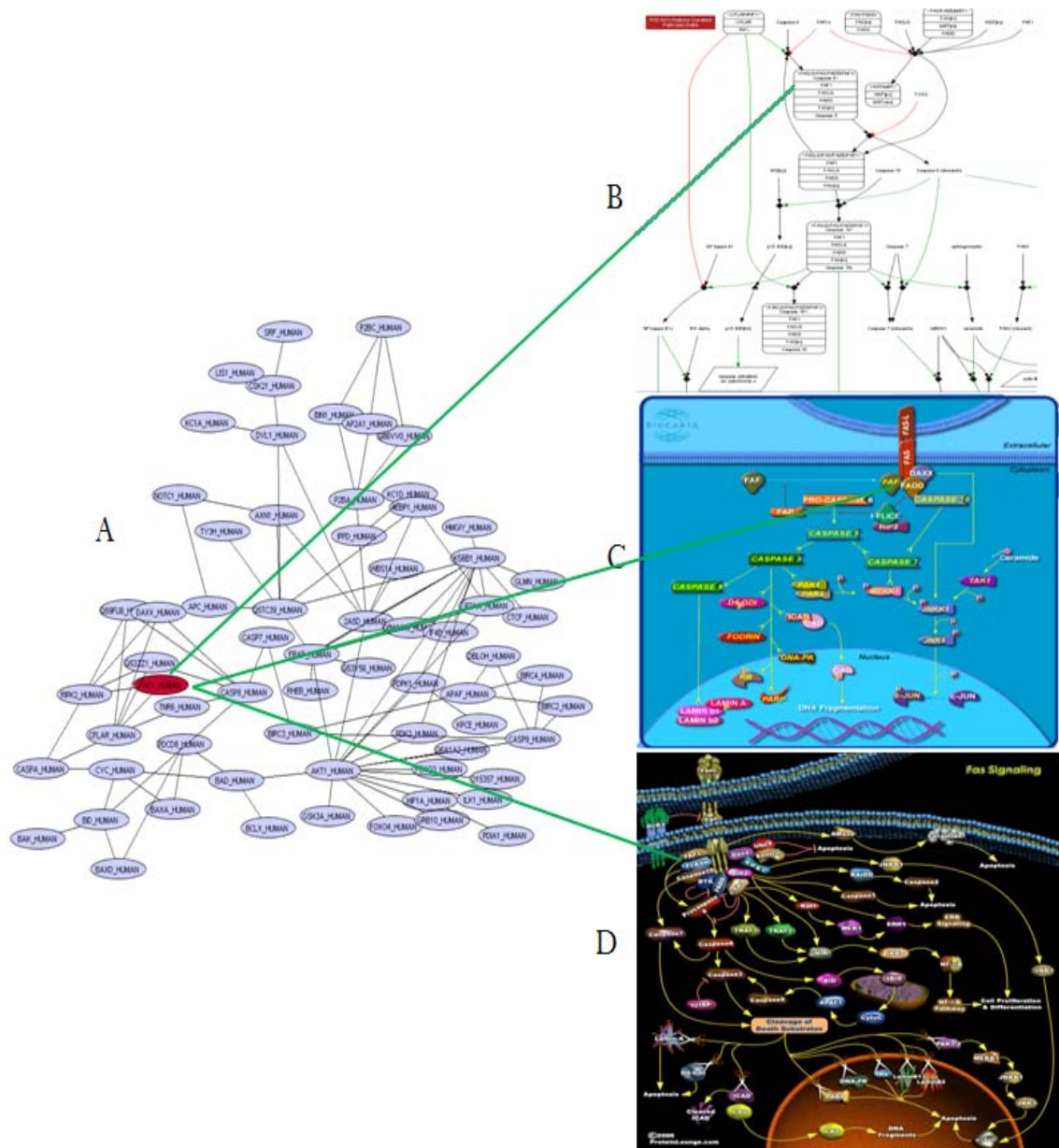
## 5.2 Case study 2: Tumor Necrosis Factor- $\alpha$ and Interleukin-1-Induced Cellular Responses: integrating with pathway information

Tumor Necrosis Factor- $\alpha$  (TNF $\alpha$ ) and Interleukin-1 (IL-1) are pro-inflammatory cytokines which mediate the innate immune response [40]. Dysregulation of innate immune response may contribute to chronic inflammatory diseases such as arthritis [41], diabetes and cancer. The expression of these cytokines are mediated by similar transcription factors. However, TNF $\alpha$  and IL-1 receptor differ in their sensitivities to a known initiator (lipopolysaccharide, LPS) of the innate immune response in knock-out mice. The contrasting responses to LPS indicate that TNF $\alpha$  and IL-1 regulate different processes. A large-scale proteomic analysis of TNF $\alpha$ - and IL-1-induced responses was performed to identify processes uniquely regulated by TNF $\alpha$  and IL-1 by integration with pathway data.

### 5.2.1 Generation and Analysis of Pathway Network

The proteins that changed significantly ( $p < 0.05$ ) upon TNF $\alpha$  or IL-1 treatment [42] were mapped to UniProt IDs using the mapping data downloaded from Ensembl using their BioMart tool. Using these proteins, we generated a human pathway network using the HPD (Human Pathway Database), which is an integrated human signaling pathway database developed in-house. This database has comprehensive information about human signaling pathways. We used pathway data from NCI-Nature curated database and Biocarta since they provide not just the content of a pathway, but also the ordering of molecular content within each pathway. After integrating protein expression data, we found FAF1, which is a splice isoform of Fas Associated Factor 1, to be

expressed high both in TNF treated and IL-1 treated cells. FAF1 was originally identified as a protein that binds to the cytoplasmic tail of the Fas protein.



**Figure 5.8 (A) Network generated using proteolens (Partial) (B) NCI- Nature curated database – Fas signaling Pathway (C) Biocarta - Fas signaling pathway (D) Protein Lounge – Fas signaling pathway**

When immune cells are activated, especially T-cells, they either survive or they die. For example, when T-cells are treated with the protein IL-2, they survive. When

these cells are treated with IL-2, this activates the IL-2 receptor, leading to the activation of the transcription factor NF- $\kappa$ B and also the activation of the IL I $\kappa$ B kinase complex (IKK $\alpha$ , IKK $\beta$  and IKK $\gamma$ ). IKK $\alpha$  and/or IKK $\beta$  will phosphorylate a protein, I $\kappa$ B $\alpha$ , which is bound to the transcription factor NF- $\kappa$ B. Phosphorylation of I $\kappa$ B $\alpha$  targets it for ubiquitin dependent degradation by the proteasome. Removal of I $\kappa$ B $\alpha$  allows NF- $\kappa$ B to go nuclear and activate genes, like the inhibitors of apoptosis that prevent cell death and therefore promote cell survival. When Fas Ligand binds to the Fas receptor on T-cells, it promotes cell death by activating proteases like the caspases.

The pathway network (Figure 5.8 a) is visualized from data from HPD, and is generated with ProteoLens with the highlighting of the FAF1 protein in red color; Fas signaling pathway diagrams from source databases are shown in Figures 5.8 b-d. As shown in the diagrams, when Fas ligand binds to Fas, it also leads to the activation of FAF1 (mechanism unknown). But the activated FAF1 binds to IKK $\beta$  and prevents it from phosphorylating I $\kappa$ B $\alpha$  - this prevents NF- $\kappa$ B signal transmission to the nucleus and activates survival signals. TNF and IL-1 also lead to an increase in FAF1 protein levels even though both TNF and IL-1 activate NF- $\kappa$ B activity. Too much NF- $\kappa$ B activity is will lead to a chronic inflammatory response. Activation of FAF1, after the initial activation of NF- $\kappa$ B, will therefore prevent the continued activation of NF- $\kappa$ B.

In general the outcome to this second case study has been that pathway knowledge from HPD helps to better integrate molecular signals with physiological responses. Both case studies suggest the potential for a comprehensive pathway data

source to link meaningfully to analyses such as those involving protein interactions, gene expression, and cellular physiology.

## CHAPTER SIX: CONCLUSION

### 6.1 Conclusion

The publication of the draft human genome consisting of 30,000 genes is just the beginning of genome biology. The complexity and wealth of molecular and cellular function of proteins in biological processes can be approached with biological networks such as protein-protein interaction networks and pathway networks. Hence, pathway databases documenting pathway information are necessary tools for network biology.

HPD, the Human Pathway Database, is the most comprehensive data warehouse of 1,895 human signaling pathways, 10,473 molecular entities and 22,974 biological reactions and has unique annotations with information on kinase-disease associations and perturbation effects of different environmental factors which helps in network biology based systems biology studies. First, this integrated pathway database provides a single relational database platform providing users with integrated views of pathway data and also the data warehouse approach enhances access for the end user. Second, the kinase-disease annotation provides notes and references for how the kinase is involved in a disease. Third, information related to effects of environmental factors gives information on pathway perturbations associated with exposure of cells to different types of environmental conditions. Overall, the HPD acts as a resource for building up pathway networks and gives basic information to researchers when they encounter proteins of interest related to signaling pathways.

The potential for merging similar pathways—“pathway mergability” —was created based on gene/protein identifiers, and is a novel method that we developed in this study to provide comprehensive information since there are differences in pathway

boundaries and pathway content for 100% of pair-wise comparisons of signal transduction pathways between any 2 of the 4 databases in our study. An ideal purpose for pathway merging is to provide non-redundant pathways.

We successfully developed a systems biology approach to analyze the proteomics and genomics data by coupling this data with pathway data. We applied this approach to two case studies: Alzheimer's disease study, and Tumor Necrosis Factor alpha- and Interleukin-1-Induced Cellular Responses. In the Alzheimer's disease case study, the proteins involved in Alzheimer's disease are categorized as the set of seed proteins. We constructed a pathway network based on the seed proteins. Then, we analyzed the pathway network by coupling with gene expression data to find significant proteins related to Alzheimer's disease. We performed gene expression analyses to find significantly enriched over-expressed and under-expressed proteins. We found AXIN1 and SREBF1 proteins to have differences in gene expression from incipient to severe patients indicating some potential consideration of them as biomarkers for Alzheimer's disease. ProteoLens was used to visualize the networks. In the Tumor Necrosis Factor alpha- and Interleukin-1-Induced Cellular Responses case study, we have taken the statistically significant proteins to expand using the pathway data. Then, we coupled with protein expression analysis. This shows that over-expression of FAF1 protein in both TNF $\alpha$  treated and IL-1 treated cells. This helps to better integrate the molecular signal with physiological response. These findings in the two case studies strongly support the future developing of a framework for evaluating functional genomics and proteomics data using pathway networks.

## 6.2 Discussion

Comprehensive access to databases in the field of bioinformatics is very essential, especially for pathway analysis. Pathway databases have been growing exponentially, for example, the NCI/Nature curated database adds pathways frequently. So, there is a need to keep HPD up to date. There are about 41 signaling pathway databases. Of these, 20 are pathway diagrams and we looked only at four databases. Expanding HPD by integration with other databases like Reactome will lead to even more comprehensive analyses.

Protein Lounge pathway items are not categorized, so there is a need to develop strategies for categorizing these pathway items, for example, by using UniProt and Chemical Accession identifier systems. Also, Protein Lounge provides reaction information in their pathway descriptions. So there is a need for developing text mining techniques to mine the reaction data.

Mapping the proteins involved in the pathways to UniProt IDs is required for pathway mergability. So far, we have achieved 80% mapping to UniProt IDs. With the increase in biological data, we may in the future find better data sets for mapping. With the pathway mergability concept, the similarity score is computed based on the similar proteins present between pathways. In order to increase the specificity of this scoring function, we can also consider the compounds that are involved in the pathway by mapping them to standard chemical accession identifiers.

This database can be extended by integrating the protein compound interaction data, which may lead to insights in drug target discovery. The architectural approach may also be extended to an organism like the mouse (*Mus musculus*) in order to integrate gene



regulation data as there can be more data available in organisms other than humans. The analyses in this thesis suggest that developing pathway analysis tools based on the HPD may potentially involve utilities for producing graphical views in order to accelerate analysis.

## REFERENCES

1. Cary, M.P., G.D. Bader, and C. Sander, *Pathway information for systems biology*. FEBS Lett, 2005. **579**(8): p. 1815-20.
2. Bader, G.D., M.P. Cary, and C. Sander, *Pathguide: a pathway resource list*. Nucleic Acids Res, 2006. **34**(Database issue): p. D504-6.
3. <http://biocarta.com>, *Biocarta*.
4. Kanehisa, M., et al., *The KEGG resource for deciphering the genome*. Nucleic Acids Res, 2004. **32**(Database issue): p. D277-80.
5. Vastrik, I., et al., *Reactome: a knowledge base of biologic pathways and processes*. Genome Biology, 2007. **8**(3): p. R39.
6. <http://pid.nci.nih.gov>, *National Cancer Institute Center for Bioinformatics 2005 Pathway Interaction Database*.
7. <http://proteinlounge.com>, *Systems Biology Database*.
8. Cerami, E.G., et al., *cPath: open source software for collecting, storing, and querying biological pathways*. BMC Bioinformatics, 2006. **7**: p. 497.
9. Nikitin, A., et al., *Pathway studio—the analysis and navigation of molecular networks*. Bioinformatics, 2003. **19**(16): p. 2155-2157.
10. Fei Chen, J.B., Laurence M. Demers, and Xianglin Shi, *Upstream Signal Transduction of NF-kB Activation*. Atlas of Genetics and Cytogenetics in Oncology and Haematology 2001. [atlasgeneticsoncology.org/Deep/NFKBID20033.html](http://atlasgeneticsoncology.org/Deep/NFKBID20033.html).
11. Rensing, L.I.J.B., *Periodic geophysical and biological signals as Zeitgeber and exogenous inducers in animal organisms*. . Int. J. Biometeorol. , 1972. **16**: **Suppl:113-125**. .
12. J., T., *Davis-1961 revisited. Signal transmission in the cochlear hair cell-nerve junction*. Arch Otolaryngol., 1975. **101** (9): p. 528-535.
13. Ashcroft SJ, C.J., Crossley PC. , *The effect of N-acylglucosamines on the biosynthesis and secretion of insulin in the rat*. Biochem. J., 1976. **154** (3): p. 701-707.
14. Kenny JJ, M.-M.O.e.a., *Lipid synthesis: an indicator of antigen-induced signal transduction in antigen-binding cells*. J. Immunol, 1979. **112** (4): p. 1278-1284.
15. E., H., *What does Halobacterium tell us about photoreception?* Biophys. Struct. Mech., 1977. **3** (1): p. 69-77.
16. Rodbell, M., *The role of hormone receptors and GTP-regulatory proteins in membrane transduction*. Nature, 1980. **284** (5751): p. 17-22.
17. Gomperts, B.K., IM. Tatham, PER. , *Signal transduction*. Academic Press, 2002.
18. <http://www.ariadnegenomics.com/products/resnet.html>, *Resnet*.
19. <http://cancer.cellmap.org>, *Cancer Cell Map*.
20. <http://www.pathwaycommons.org>, *Pathway Commons*.
21. [http://www.cellsignal.com/reference/kinase\\_disease.html](http://www.cellsignal.com/reference/kinase_disease.html), *Cell Signaling Technology*.
22. <http://www.grt.kyushu-u.ac.jp/eny-doc>, *Signalng Pathway Database*.

23. Hucka, M., et al., *The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models*. Bioinformatics, 2003. **19**(4): p. 524-31.
24. Luciano, J.S., *PAX of mind for pathway researchers*. Drug Discov Today, 2005. **10**(13): p. 937-42.
25. S.W., A., *Agile model driven development with UML2, The Object Primer Third Edition, 2004, Chapter 6*. 2004.
26. Jain P., K.M., Parameswaran K. , <http://posa3.org/workshops/ThreeTierPatterns/>.
27. Blagoveshchenskaya, A.D., et al., *HIV-1 Nef Downregulates MHC-I by a PACS-1-and PI3K-Regulated ARF6 Endocytic Pathway*. Cell, 2002. **111**(6): p. 853-866.
28. [http://alz.org/alzheimers\\_disease\\_what\\_is\\_alzheimers.asp](http://alz.org/alzheimers_disease_what_is_alzheimers.asp), *what is alzheimers*.
29. Chen, J.Y., C. Shen, and A.Y. Sivachenko, *Mining Alzheimer disease relevant proteins from integrated protein interactome data*. Pac Symp Biocomput, 2006. **11**: p. 367.
30. Brown, K.R. and I. Jurisica, *Online Predicted Human Interaction Database*. Bioinformatics, 2005. **21**(9): p. 2076-2082.
31. De Ferrari, G.V. and N.C. Inestrosa, *Wnt signaling function in Alzheimer's disease*. Brain Res Brain Res Rev, 2000. **33**(1): p. 1-12.
32. Wu, C. and D.W. Nebert, *Update on genome completion and annotations: Protein Information Resource*. Hum Genomics, 2004. **1**(3): p. 229-33.
33. You, Q., Shiao-fen Fang, and Jake Y. Chen, *GeneTerrain: Visual Exploration of Differential Gene Expression Profiles Organized in Native Biomolecular Interaction Networks*. . Information Visualization, (Accepted with revision), 2007.
34. Apweiler, R., et al., *UniProt: the Universal Protein knowledgebase*. Nucleic Acids Research.
35. Ikeda, S., et al., *Axin, a negative regulator of the Wnt signaling pathway, forms a complex with GSK-3 beta and beta-catenin and promotes GSK-3 beta-dependent phosphorylation of beta-catenin*. The EMBO Journal, 1998. **17**: p. 1371-1384.
36. Dahmen, R.P., et al., *Deletions of AXIN1, a Component of the WNT/wingless Pathway, in Sporadic Medulloblastomas 1*. 2001, AACR. p. 7039-7043.
37. Kawahara, K., et al., *Down-regulation of  $\beta$ -Catenin by the Colorectal Tumor Suppressor APC Requires Association with Axin and  $\beta$ -Catenin*. Journal of Biological Chemistry, 2000. **275**(12): p. 8369-8374.
38. Carter, C.J., *Convergence of genes implicated in Alzheimer's disease on the cerebral cholesterol shuttle: APP, cholesterol, lipoproteins, and atherosclerosis*. Neurochem Int, 2006.
39. Sivachenko, A., J. Chen, and C. Martin, *ProteoLens: A Visual Data Mining Platform for Exploring Biological Networks*. Bioinformatics, 2005.
40. Dinarello, C.A., *Proinflammatory Cytokines\**. 2000, Am Coll Chest Phys. p. 503-508.
41. Cohen, S.B., *The use of anakinra, an interleukin-1 receptor antagonist, in the treatment of rheumatoid arthritis*. Rheumatic Disease Clinics of North America, 2004. **30**(2): p. 365-380.
42. Ott, L.W., Katheryn A. Resing, Alecia W. Sizemore, Joshua W. Heyen, Ross R. Cocklin, Nathan M. Pedrick, H. Cary Woods, Jake Y. Chen, Mark G. Goebel,

- Frank A. Witzmann, and Maureen A. Harrington, *Tumor Necrosis Factor-alpha and Interleukin-1 Induced Cellular Responses: Coupling Proteomic and Genomic Information*. Journal of Proteome Research, 2007. **6**(6): p. 2176-2185.
43. Kasamsetty, Harini, Xiaogang Wu, and **Jake Y. Chen** (2008) An Integrative Human Pathway Database for Systems Biology Applications. Proceedings of the 23<sup>rd</sup> Annual ACM Symposium on Applied Computing.