VINEdb: a data warehouse for integration and interactive exploration of life science data

Sridhar Hariharaputran[*], Thoralf Töpel, Björn Brockschmidt and Ralf Hofestädt

Bielefeld University, Faculty of Technology, Bioinformatics Department,
PO Box 10 01 31, D-33501 Bielefeld, Germany

Abstract

Control of cell proliferation, differentiation, activation and cell removal is crucial for the development and existence of multi-cellular organisms. Apoptosis, or programmed cell death, is a major control mechanism by which cells die and is also important in controlling cell number and proliferation as part of normal development. Molecular networks that regulate these processes are critical targets for drug development, gene therapy, and metabolic engineering. The molecular interactions involved in this and other processes are analyzed and annotated by experts and stored as data in different databases. The key task is to integrate, manage and visualize these data available from different sources and present them in a user-comprehensible manner.

Here we present VINEdb, a data warehouse developed to interact with and to explore integrated life science data. Extendable open source data warehouse architecture enables platform-independent usability of the web application and the underlying infrastructure. A high degree of transparency and up-to-dateness is ensured by a monitor component to control and update the data from the sources. Furthermore, the system is supported by a visualization component to allow interactive graphical exploration of the integrated data. We will use apoptotic pathway and caspase-3 as a case study to show capability and usability of our approach. VINEdb is available at http://tunicata.techfak.uni-bielefeld.de/VINEdb/.

1 Introduction

Exploration of the enormous amount of data on various molecules and their interactions with their counterparts from day-to-day experiments followed by careful analysis and annotation by experts has paved the way to create large databases. High and low-throughput and post-genome experiments have further contributed to the growth of the databases.

The information that is present in several databases can be brought together through database integration methods. The integrated data can be further applied for a specific purpose that can serve research and community and as well as allow co-operative information sharing in the field of bioinformatics. In systems biology the goal of data integration is to combine information from a number of databases and data sets, which are obtained from both high and low throughput experiments, under one data management scheme. The cumulative information can provide greater biological insight than is possible with individual information sources [1]. The challenge is to capture, model, integrate and analyze these data in a consistent way to provide new and deeper insights into complex biological systems.

The emerging field of integrative bioinformatics provides the essential methods to integrate, manage and analyze the diverse data and allows gaining new insight and a deeper understanding of complex biological systems [2].

[*] Corresponding author, sharihar@techfak.uni-bielefeld.de

Taking advantage of the data stored in heterogeneous biological data can be difficult and time-consuming for various reasons, which has led to the development of automated systems. The integration of heterogeneous databases is an important issue in biological research resulting in the development of several systems and solutions [3]. Graphical representations of the facts are easier to understand than if they were presented as raw data. This is especially true for large datasets or complex situations [4].

The KEGG, OMIM, IntAct, GO and UniProt data that are integrated in this warehouse and other databases store data in different formats such as tables, maps etc. It is also known that these data are related to each other in one way or another and are diverse in nature. We can still integrate the heterogeneous data according to our needs, thus resulting in a new integrated database.

In this paper we introduce VINEdb, a data warehouse for integration and interactive exploration of life science data. The key idea in our paper is not only to develop a data warehouse that integrates and manages diverse data, but also to emphasize the visualization of the integrated data. The advantage of a visualization method is that it provides the user with more information about the data in a comprehensible manner. This method is also effective in determining the relationship between the data of the user’s interest. For example, a gene and its relation to other proteins, enzymes, disease, drugs, compounds, interactors and pathways can be illustrated with an image or two along with the associated information and its source within the data warehouse.

2 Related works

There has been a rapid increase in the volume and number of data resources providing polymorphic views of the same data and often overlap in multiple resources. They are also stored and published in diverse data sources. Each source is distinct in the focus and format [5]. The presence of numerous informational resources on genes, enzymes, pathways etc., raises an acute problem of data integration and suitable access. The integration of these heterogeneous data types is a challenging problem especially in biology, where the number of databases and data types is increasing rapidly [2].

2.1 General integration approaches

The idea of data integration in molecular biology is not a new one. There have been several underlying projects that focused on the challenging problem of interoperability among biological databases. P. Karp first addressed biological database integration in the early nineties [6]. Since then, diverse integration approaches for molecular biological data sources have been developed. These systems are based on different data integration techniques, e.g. text indexing systems (e.g. SRS, BioRS), multi database and federated database systems (e.g. DiscoveryLink, BioKleisli/K2), and data warehouses (e.g. Atlas, BioWarehouse). We will present these approaches using DiscoveryLink and SRS as examples. Afterwards, we will discuss various data warehouse systems.

The DiscoveryLink system [7] was developed by IBM to access multiple heterogeneous data sources through single SQL queries. It is based on federated database techniques, so it requires the development of a global data scheme. DiscoveryLink accesses its original data sources through wrappers and views. Read-only SQL is supported as query language. The system is now part of IBM’s Websphere Information Integrator.

SRS [8] is based on local copies of each integrated data source with a special format that is described by the Icarus language specification. Icarus can help representing the structure of the integrated data source. Through the use of these local copies, SRS is completely materialized. But during this transfer into the new format, no scheme integration is realized. Therefore, the degree of integration can be characterized as loose. SRS runs on a web server and is accessible via any browser. An HTML interface for data queries is provided. Furthermore, the system can be queried by constructing special URLs. But no query languages like SQL or OQL are supported. SRS offers also a C-API. Various output formats are possible (HTML or ACSII text). One problem with the result presentation in SRS is the necessity to parse the outputs for further computer-based processing. The absence of any scheme integration is also disadvantageous for the use of the SRS system.

Text indexing systems and multi/federated database systems provide data from distributed, heterogeneous data sources to the user in a homogenous way. But along with the increasing amount of life science data, there is also a change from pure data management to complex data analysis in bioinformatics. These complex analysis queries are not sufficiently supported by such approaches. However, data warehouse systems provide a global data schema and allow periodical loading of all the data into a central repository. The concept of data warehousing helps to overcome major limitations of the distributed database systems: inconsistency of data and time-consuming or incomplete queries caused by server restrictions [9]. Unlike other database solutions in which the query is not proper or the response is incomplete, the data warehouse approach has several advantages allowing complex querying which can be more time-consuming. There are examples of a data warehouse serving the bioinformatics community at different levels with diverse data. These projects can be roughly separated into two groups: general software infrastructures for further customization within new bioinformatics applications (e.g. Altas, BioWarehouse) and project-oriented data warehouses implementations for particular biological questions (e.g. Systomonas, Columba).

2.2 Bioinformatics data warehouse systems

The Atlas [10] data warehouse stores locally and integrates biological sequences and information of molecular interactions, homology, functional annotation of genes and biological ontologies, thus providing a system for data and for software infrastructure for bioinformatics research and development. BioWarehouse [11] is an open-source toolkit for constructing bioinformatics databases using different database management systems. It facilitates and allows several database integration tasks like comparative analysis and data mining. It integrates its component databases into a representational framework and enables multi-database querying using SQL. But neither system is platform-independent, as each is implemented with different programming languages and requires a time-consuming installation.

Systomonas [12] provides an integrated bioinformatics platform for a systems biology approach and the biology of pseudomonas in infection and biotechnology. Apart from the in-house experimental metabolome, proteome and transcriptome data, it also stores the prediction information of cellular processes such as gene regulatory networks. Columba [4] is an integrated database of proteins, structures and annotations. Both systems are available via a web-based graphical user interface that can be used with any web browser. However, the up-to-dateness of the systems is hard to judge and there is no logging information available on the update process.

Reactome [13] is a knowledge base of biological processes. In the database, the basic unit is a reaction and the reactions are grouped into casual chains to form pathways. It is used to infer equivalent reactions in human and non-human species. Apart from the integrated information, the database is supported by a “Skypainter” tool to colorize the reactions in different ways.

3 Implementation

Based on previous considerations, we designed a platform-independent data warehouse system that integrates heterogeneous data sources into a local database and provides a comprehensible updating strategy to ensure a maximum transparency and up-to-dateness of the integrated data. Beside the common web-based user interface, there is a visualization component that allows interactive graphical exploration of the integrated data. At present, a prototype is implemented and the general availability release will be completed by the end of 2007.

3.1 System architecture

A schematic representation of the VINEdb 4-layer system architecture is shown in Figure 1. The source layer is the basis of the system and contains the data sources OMIM, KEGG, UniProt, IntAct and GO. Parseable flat files are provided by most of the database carriers, whereas Gene Ontology is supplied via a SQL dump file. The different external data sources are controlled by the monitor component of the integration layer. It recognizes changes of the original sources and starts a download of the changed files if necessary.

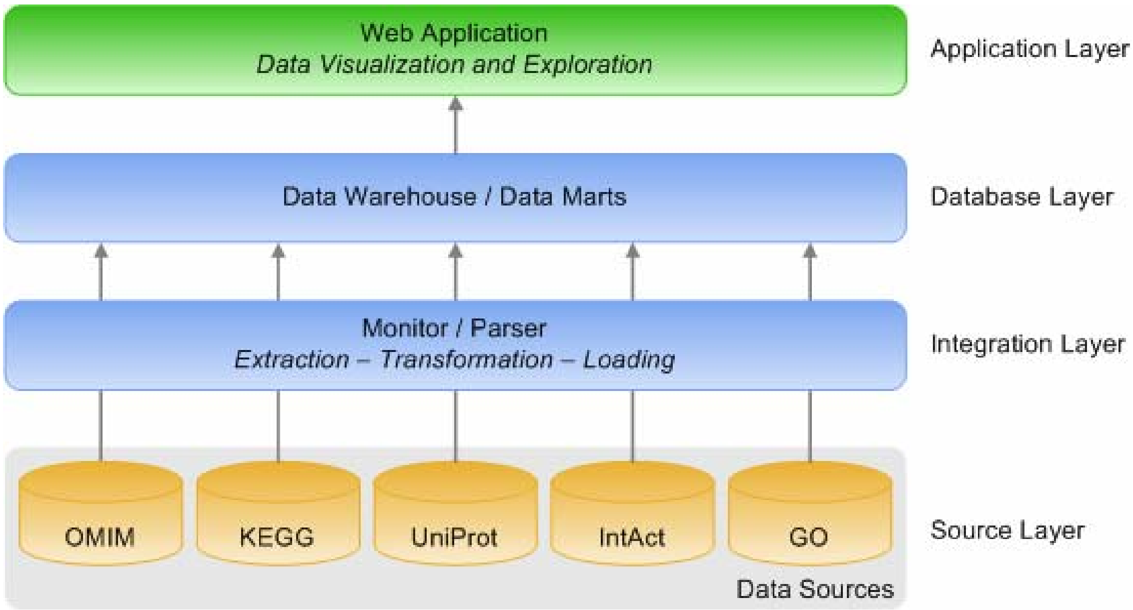


Figure 1: Schematic representation of the VINEdb 4-layer system architecture and the flow of data from the original heterogeneous data sources in the source layer to the web application

When these files are downloaded successfully to the local file system, the parser will be activated to start the ETL process. ETL is a typical data warehouse process and marks the successive steps of extraction, transformation and loading of data. This means that the data is extracted from the original data exchange format, transformed to the target data schema and loaded to the data warehouse. Thereby the export schemata of the data sources are loosely coupled by mapping tables that establish relationships between data entities from different biological aspects. This approach simplifies the maintenance of the data schema and the integration of new data sources. Based on the data loaded to the data warehouse of the database layer, smaller data marts can be constructed for specific analysis applications.

The end-user interacts with the system by a web-based graphical user interface -the web application. As previously described, this web application allows users homogenous access to the integrated data and supports its exploration by interactive visualization of relationships between data entities. The web pages of the system are accessible with any common browser.

Each of these data entities is supported by detailed information and also further referred to the original data source. A search engine facilitates and allows the user to interact at different levels to find the information of their interest.

In VINEdb, interactive visualization and graphical representation are implemented. Creating images at run-time that are dynamic and interactive will provide the user with more advantages and accessibility to explore the integrated data. Interconnected domain information in the system, as shown in Figure 2, allows further exploration. A long list of interactions or a protein pairs table will not suffice to show what happens in the cell. For this purpose, a graphical representation of the facts makes it easier to understand the complex situations or biological complexity. Moreover, the use of graphics suits the human preference for visual perception [14]. A menu at the side panel helps to navigate through the contents of the data warehouse such as pathways, protein, enzymes, disease, drug, etc.

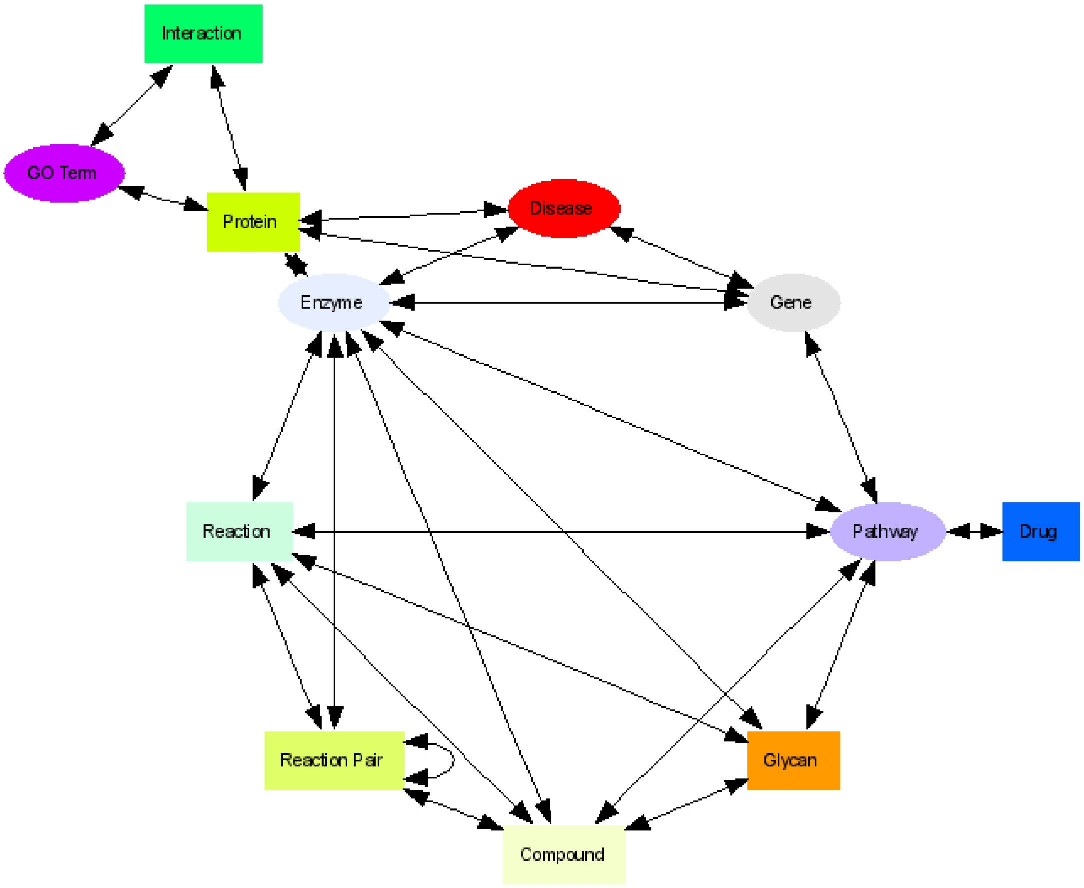


Figure 2: Schematic representation of the relationships between the different biological aspects (domains) within the integrated data of VINEdb

3.2 Implementation details

The core of this data warehouse infrastructure is implemented completely in Java to ensure platform independence of the operating system. Thus it can be used separately from the web interface to integrate several life science data sources into relational databases for individual research. This task is supported by a collection of ready-to-use parsers for standard life science data sources, e.g. Brenda, EMBL, ENZYME, GO, iProClass, KEGG, OMIM, PubChem, Taxonomy, UniProt, SCOP and CATH. Currently, the preferred database management system for the integrated data is MySQL, but an additional persistence layer is under development to enable the use of further relational database management systems, e.g.

Oracle or PostgreSQL. Once a release candidate of the software is finished, it will be available on SourceForge (http://sourceforge.net/projects/biodwh/).

The web-based graphical user interface of VINEdb is implemented with JavaServer Pages and runs on an Apache Tomcat web server. Based on user activity, it carries out the different search, preparation and presentation functions by connecting to the database and generating HTML pages. The graph visualization software Graphviz [15] was used to dynamically create the graphical representations of the relationships between the entities of the data warehouse. Graphviz (http://www.graphviz.org/) is controlled by the DOT language that provides syntax to describe graphs, nodes and edges with additional layout preferences. Thus, a DOT file is generated according to the entities selected by the user and their relationships it is given to the Graphviz software that produces a PNG image file with the graph visualization. Afterwards, this image is embedded in the HTML pages and displayed by the web browser.

4 Application

The huge data from the sources KEGG, GO, UNIPROT and IntAct are quite difficult to handle. At the same time, browsing the web pages for the needed information is time consuming. Therefore, there is a need for a consolidated image that can project or give an overview of the background information. In this section, we have chosen to discuss the highly regulated and complicated apoptotic pathway and the well studied caspase-3 that play a major role in several neurodegenerative disorders like Alzheimer’s and Huntington’s as an example and to show VINEdb’s capability.

Apoptosis, or programmed cell death, is a major control mechanism by which cells die if DNA damage is not repaired. Apoptosis is also important in controlling cell number and proliferation as part of normal development [16]. It is an evolutionary process that is conserved and removes unwanted or damaged cells [17]. Apoptosis occurs through two main pathways. The first, referred to as the extrinsic or cytoplasmic pathway, is triggered through the Fas death receptor, a member of the tumor necrosis factor (TNF) receptor superfamily. The second pathway is the intrinsic or mitochondrial pathway that when stimulated leads to the release of cytochrome-c from the mitochondria and activation of the death signal. Both pathways converge to a final common pathway involving the activation of a cascade of proteases called caspases (**C**ysteine **A**spartate **S**pecific **P**roteases) that cleave regulatory and structural molecules, culminating in the death of the cell [16]. The caspases are classified into as intiator caspases (caspase 8, 9 and etc.) and effector caspases (3, 6, 7 and etc.).

Caspases play role in several neurodegenerative diseases like Huntington’s, Alzheimer’s, diabetes, Parkinson’s, carcinoma and arthritis as shown in Figures 3 and 4. This has been proved from clinical studies and associated literatures. The study led to the development of drugs against caspase-mediated diseases such as those mentioned above including stroke and osteoarthritis using caspase-3 inhibitors. Increased levels of apoptosis and caspase activity are associated with the cellular damage in Alzheimer’s, Parkinson’s and Huntington’s diseases. Caspase-3 is the predominant caspase and has a central role in the cleavage of the amyloid-beta 4A precursor protein, associated with neuronal death in Alzheimer’s disease.

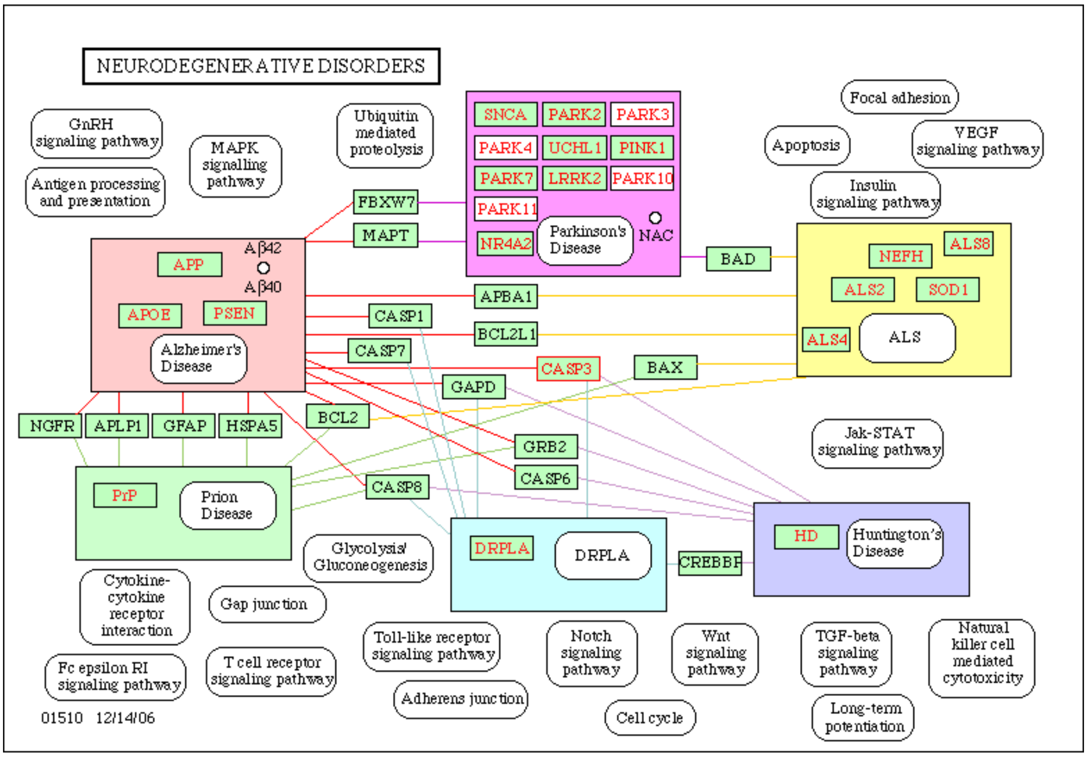
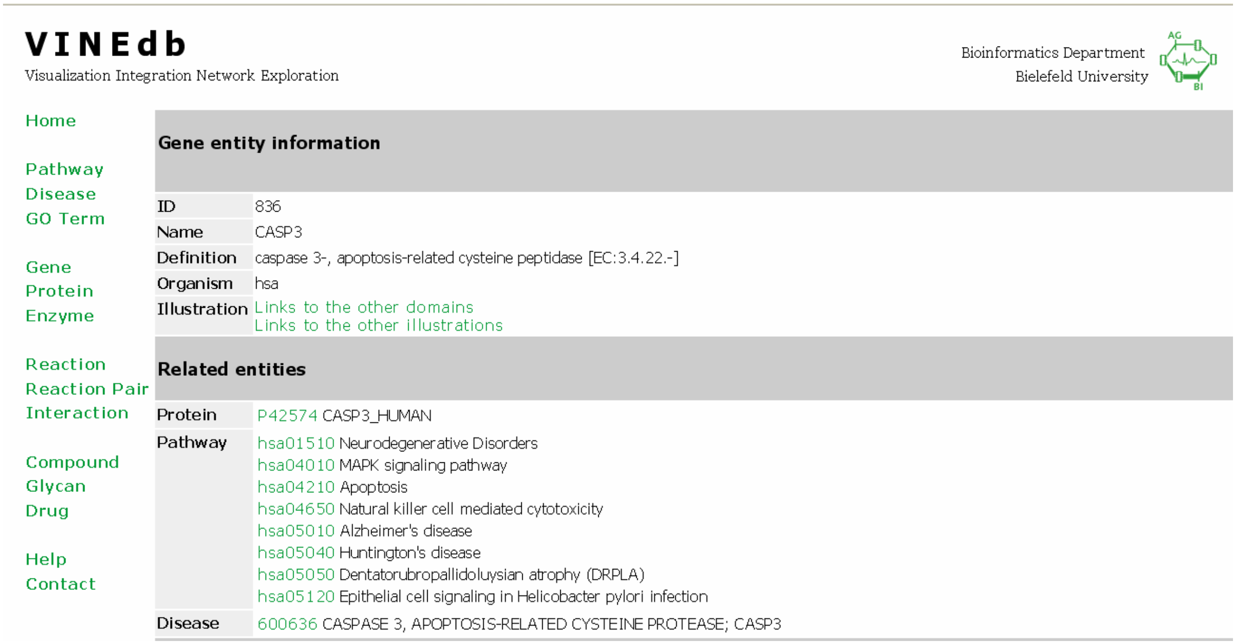


Figure 3: Role of CASP3 in several neurodegenerative diseases as adapted from KEGG



VINEdb

Visualization Integration Network Exploration

Bioinformatics Department
Bielefeld University

Home

Pathway
Disease
GO Term

Gene
Protein
Enzyme

Reaction
Reaction Pair
Interaction

Compound
Glycan
Drug

Help
Contact

Gene entity information

ID	836
Name	CASP3
Definition	caspase 3-, apoptosis-related cysteine peptidase [EC:3.4.22.-]
Organism	hsa
Illustration	Links to the other domains Links to the other illustrations

Related entities

Protein	P42574 CASP3_HUMAN
Pathway	hsa01510 Neurodegenerative Disorders hsa04010 MAPK signaling pathway hsa04210 Apoptosis hsa04650 Natural killer cell mediated cytotoxicity hsa05010 Alzheimer's disease hsa05040 Huntington's disease hsa05050 Dentatorubropallidoluysian atrophy (DRPLA) hsa05120 Epithelial cell signaling in Helicobacter pylori infection
Disease	600636 CASPASE 3, APOPTOSIS-RELATED CYSTEINE PROTEASE; CASP3

Figure 4: CASP3 gene entity information as shown in VINEdb and its role in disease pathways

Alzheimer's disease (AD) is a neurodegenerative disease that is marked by neuronal death, extra cellular senile plaques (SPs) and intracellular neurofibrillary tangles (NFTs). It has been suggested that apoptosis may be one of the mechanisms that leads to neuronal death in Alzheimer's disease [18]. In the case of Alzheimer's disease, there are several genes and pathways that are very well associated as demonstrated in VINEdb, Figure 5 A. This image, generated automatically at run-time, shows the typical representation of the relationships between the data entities of the integrated data. A node specifies the data entities, e.g. a specific pathway or gene. Variations of the nodes in shape and color define the biological aspects within the data warehouse, e.g. a grey ellipse represents a gene. Directed edges notice relationships between the data entities, whereas the edge starts at the node selected before.

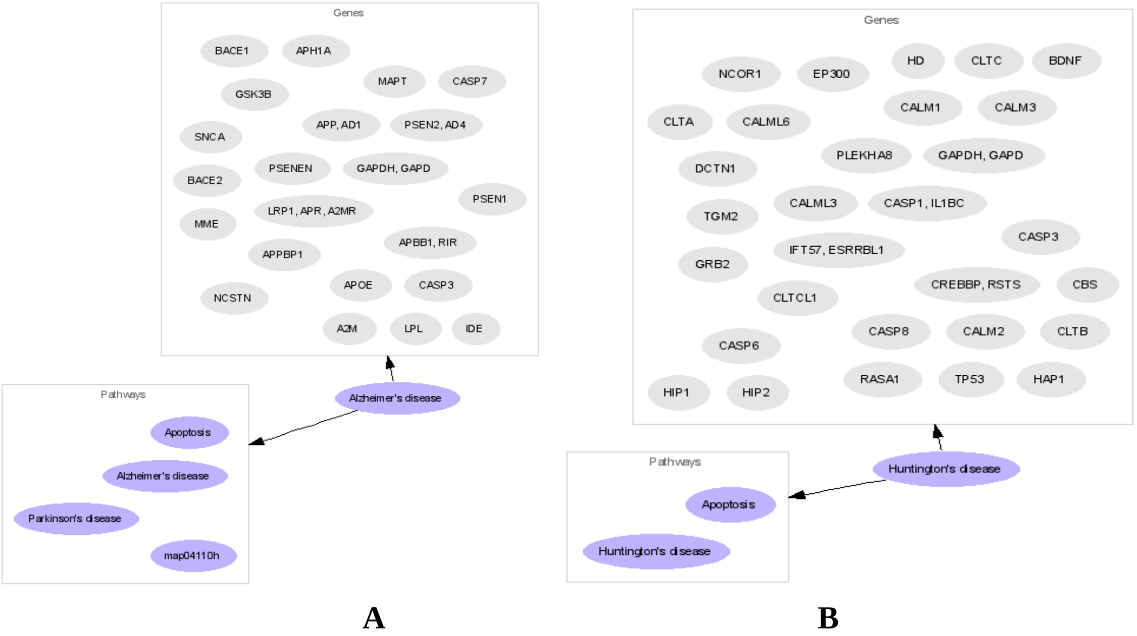


Figure 5: Alzheimer's (A) and Huntington's (B) disease and their relation to other genes and related pathways as shown in VINEdb

Huntington's disease (HD) is an inherited neurodegenerative disease that is characterized by chorea, the movement abnormality consisting of jerky motions, changes in personality, dementia and early death. These symptoms result from the selective death and malfunction of specific neuronal subpopulations in the central nervous system. The disease is a progressive disorder of motor, cognitive and psychiatric disturbances. The gene responsible for HD is the huntingtin gene. The Huntingtin protein contains several cleavage sites for caspase-3. Figure 5 B illustrates how Huntington's disease is associated with different genes and pathways as in VINEdb.

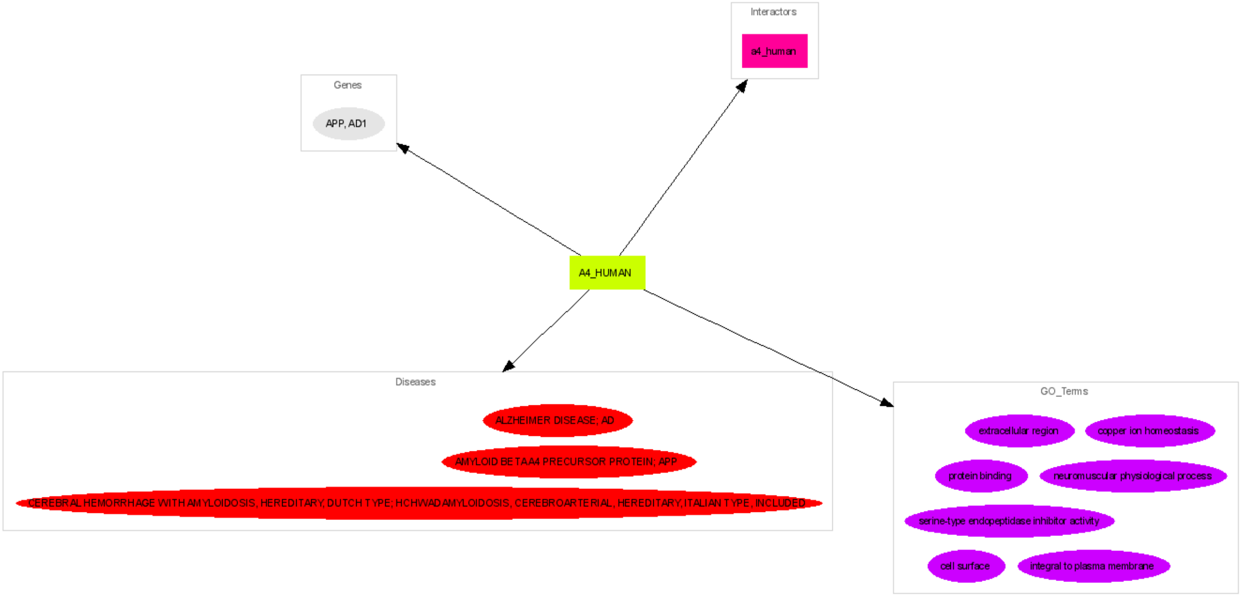

Interactors
a4_human
Genes
APP, AD1
A4_HUMAN
Diseases
ALZHEIMER DISEASE; AD
AMYLOID BETAA4 PRECURSOR PROTEIN; APP
CEREBRAL HEMORRHAGE WITH AMYLOIDOSIS, HEREDITARY, DUTCH TYPE; HCHWAD AMYLOIDOSIS, CEREBROARTERIAL, HEREDITARY, ITALIAN TYPE, INCLUDED
GO_Terms
extracellular region
copper ion homeostasis
protein binding
neuromuscular physiological process
serine-type endopeptidase inhibitor activity
cell surface
integral to plasma membrane

Figure 6: Amyloid beta A4 protein precursor or Alzheimer disease amyloid protein and its relation with other domains as shown in VINEdb

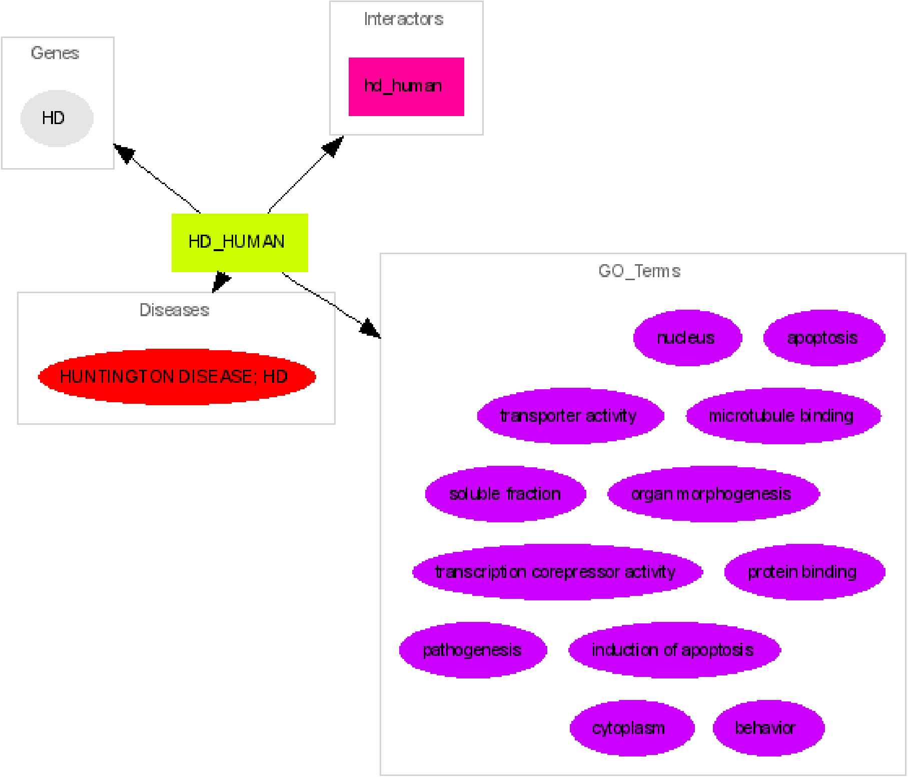


Figure 7: Huntington's disease protein (HD_Human) and its relation with other domains as shown in VINEdb

Figures 5A and B show the diseases and their association with other pathways and several other genes including CASP3. Also, Figures 6 and 7 show how the proteins involved are associated with other diseases, interactors, genes and GO terms. Though the intertwined information which are brought together is vast and complex and derived from several sources like, the disease information from OMIM (Online Mendelian Inheritance in Man), GO terms from Gene Ontology project, protein and interactors information from Uniprot and IntAct, pathways and genes from KEGG they have been illustrated as three simple figures generated on run-time by VINEdb, which can be understood easily by the users. These are further connected to reactions, compounds and to other steps and a link to the source allow drill more detailed information.

5 Summary

The availability of heterogeneous data in different databases is well-known and is stored in different formats. This makes it hard to comprehend the interactions, relationships and underlying complexity between the databases. But the same data, when integrated and presented in an effective way, will allow users to gain more knowledge about a protein and its interacting partners.

For this purpose, we have developed VINEdb, a data warehouse among other functions, enables interactive visualization and graphical representation making it possible to present in one or more images the different domain information. Thus, with a consolidated image the users will be able to explore and understand the complex nature of the molecular interactions involved with ease.

6 Acknowledgements

The work is funded by the DFG Graduate College Bioinformatics at Bielefeld University.

7 References

[1] M Baitaluk, X Qian, S Godbole et al: PathSys: integrating molecular interactions graphs for systems biology, BMC Bioinformatics, 7: 55, 2006.

[2] A Birkland and Golan Yona: BIOZON: a system for unification, management and analysis of heterogeneous biological data, BMC Bioinformatics 7**:**70, 2006.

[3] SL Cao, L Qin, WZ He et al.: Semantic Search among Heterogeneous Biological Databases Based on Gene Ontology, Acta Biochim Biophys Sin (Shanghai). 36(5):365-70, 2004.

[4] S Trißl, K Rother, H Müller et al.: Columba: an integrated database of proteins, structures, and annotations. BMC Bioinformatics, 6:81, 2005.

[5] M Jayapadian, A Chapman, VG Tarcea et al: Michigan Molecular Interactions (MiMI): putting the jigsaw puzzle together, Nucleic Acids Res. Jan; 35 Database issue: D566-71, 2007.

[6] P Karp: A strategy for database interoperation. Journal of Computational Biology, 2(4):573-586, 1995.

[7] LM Haas, PM Schwarz, P. Kodali et al.: DiscoveryLink: A System for Integrated Access to Life Science Data Sources. IBM Systems Journal, 40(2):489-511, 2001.

[8] T Etzold, A Ulyanov, P Argos: SRS: Information Retrieval System for Molecular Biology Data Banks. Methods in Enzymology, 266:114-128, 1996.

[9] M Fischer, QK Thai, M Grieb et al.: DWARF – a data warehouse system for analyzing protein families, BMC Bioinformatics, 7: 495, 2006.

[10] SP Shah, Y Huang, T Xu et al.: Atlas – a data warehouse for integrative bioinformatics. BMC Bioinformatics, 6: 34, 2005.

[11] TJ Lee, Y Pouliot, V Wagner et al.: BioWarehouse: a bioinformatics database warehouse toolkit, BMC Bioinformatics, 7:170, 2006.

[12] C Choi, R Münch, S Leupold et al.: SYSTOMONAS – an integrated database for systems biology analysis of Pseudomonas, Nucleic Acid Research, 35, Database issue, D533- D537, 2007.

[13] GJ Tope, M Gillespie, I Vastrik et al: Reactome: a knowledge database of biological pathways Nucleic Acids Res. 33 (Database issue), 2005.

[14] P Uetz, T Ideker, B Schwikowski: Visualization and integration of protein-protein interactions. In: E Golemis (ed.): Protein-Protein Interactions - A Molecular Cloning Manual. Cold Spring Harbor Laboratory Press, 623-646, 2002.

[15] ER Gansner, SC North: An open graph visualization system and its applications to software engineering. Software - Practice and Experience, 30(11):1203-1233, 2000.

[16] NM Pandya, SM Jain, DD Santani: Apoptosis: A Friend Or Foe? The Internet Journal of Pharmacology 4(2), 2006.

[17] SW Fesik and Y Shi: Controlling the Caspases, Science, 294, 1477-1478, 2001.

[18] JH Su, M Zhao, AJ Anderson et al.: Activated caspase-3 in Alzheimer's and aged control brain: correlation with Alzheimer pathology. Brain Res. 20; 898(2):350-7, 2001.