A Prototype of the HumanVirus Interactome Resource(HVIR)

© Alex Pothen, Mohammad Zubair, Kurt Maly,
Computer Science, Old Dominion University
Chris Osgood,
Biological Sciences, Old Dominion University
Oliver John Semmes,
Microbiology and Molecular Cell Biology, Eastern Virginia Medical School
Emad Ramadan, Mahmoud Abu-elela and Praveen Namburi
Computer Science, Old Dominion University

Abstract

Documenting the interaction of HTLV virus proteins with those of the host cell is crucial to understanding the process of the virus replication and pathogenesis, and provides an essential foundation for the development of safe and effective therapeutic treatments. Although numerous interactions have been reported in the scientific literature and various databases there is currently no method for efficiently accessing this information. In this paper we report on a project to design and implement mechanisms to extract and harvest this information on a continuous basis to compile a comprehensive, up-to-date digital library of the described interactions between HTLV and cellular proteins. We have added a visualization service to the digital library that allows researchers to view the interaction network and manipulate it to narrow the choices of future experiments that will validate hypotheses about the various biological processes.

1 Introduction

Consider the following scenario: A new virus with the potential of causing a human pandemic, e.g., the bird-flu virus H5N1, is identified, and it is necessary to create a a vaccine or an anti-viral therapy quickly. To do so, virologists need to determine the viral proteins responsible for infection and the human proteins they interact with. This requires mining the existing literature and databases for proteins involved in viral attachment and transcription from known viruses of the same family. Unfortunately there are currently no electronic tools and proteomic repositories to assist virologists in this task, and they have to resort to time-consuming manual searches.

Proceedings of the 8th National Russian Conference on Digital Libraries RCDL2006, Suzdal, Russia, 2006

In this work, we have proposed to build the Human Virus Interactome Resource (HVIR), the first digital library containing interactions between proteins from viruses and humans, protein co-localization, and other data from proteomics experiments, for multiple virus pathogens. Our long-term vision for HVIR is to provide virologists with proteomic data mined dynamically from the literature and predictions based on this data in an easily visualized and easily accessible manner. Many research issues need to be solved to develop automated processes for accomplishing this vision. The HVIR will provide virologists the ability to design experiments to detect how the human protein interactome is perturbed on viral infection, and help in the timely development of therapies.

The objectives of this project are:

- Build a prototype of HVIR and demonstrate its effectiveness in generating testable hypotheses through experiments guided by protein interactions predicted by HVIR.
- Identify issues that impact the expansion of the pilot program to a multi-virus platform for identifying human-viral protein interactions.

There are several reasons why this work is timely. An imminent problem facing the NIH is the bird flu virus H5N1, prevalent now in 12 countries including Canada; it has caused 61 human deaths as of mid Nov. 2005. A flu pandemic, like the 1918 outbreak that caused about 50 million deaths, could be caused by about ten mutations in one viral gene in the H5N1 avian flu strain [3]. The Hampton Roads region, with Navy personnel returning from different parts of the world, is particularly vulnerable. A second reason for the timeliness of HVIR is the lack of a proof-reading mechanism in RNA viruses that causes them to be a million-fold more likely to undergo mutations than DNA viruses. Hence the genomes of RNA viruses are rapidly changing. A digital library accessible through the Web, capable

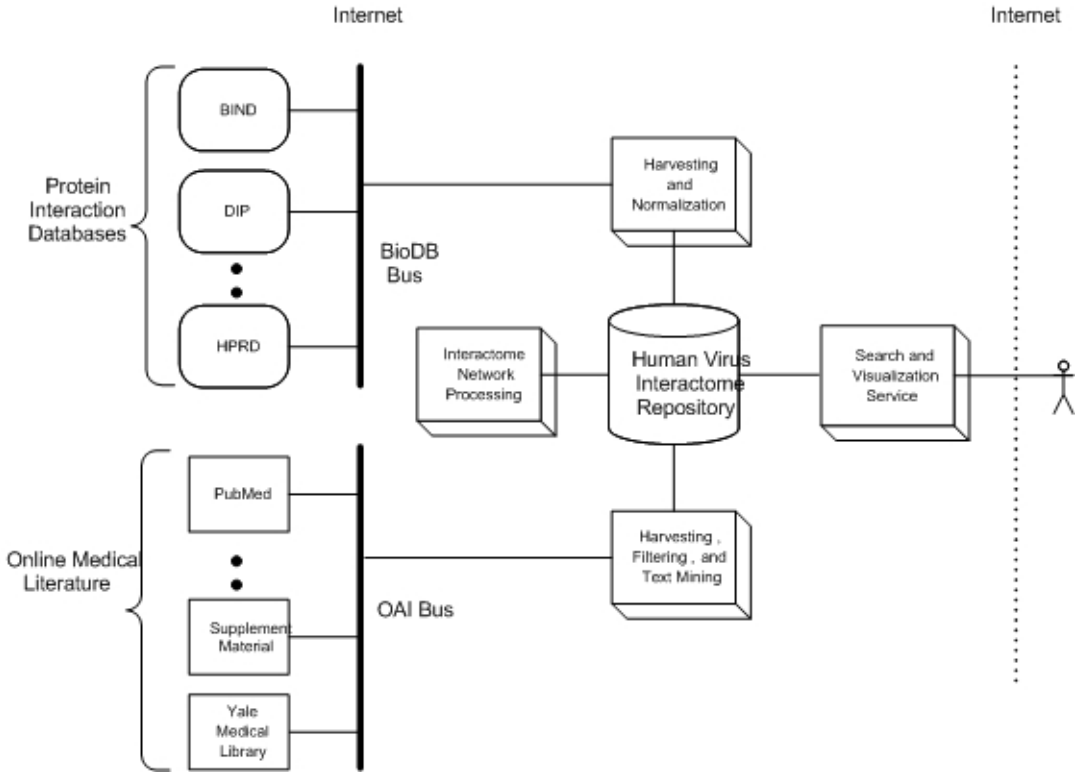

Figure 1: HVIR Architecture.

of dynamically updating proteomic information, is essential for scientists to quickly gather information on related viruses from the literature since it gets updated daily.

2 Approach

Our approach to build the HVIR is illustrated in Figure 1 below.

We are working with two types of information sources: Protein Interaction Databases and Online Medical Literature consisting of digital libraries and supplement material available on individual researchers' web sites. The major challenges in realizing our architecture are:

- Automation of periodical collection of information from protein interaction databases.
- Currently there are no protocol standards for sharing information electronically from these databases.
- Lack of standard representation of information in protein interaction databases makes it difficult to aggregate information from different sources.
- Harvesting supplementary material from published literature, available at the authors' web site. This information tends to be very unstructured.
- Adapting text mining algorithms for growing collections.
- Normalizing entity names in the context of new discoveries.
- User friendly search and visualization interfaces that will allow operations targeted to studying protein interactions such as visualizing subnetworks and performing transitive closures.
- Curating a dynamically changing repository.

3 Overview of HVIR Prototype

The objective of the current phase is to demonstrate the effectiveness of HVIR in a limited domain. We restrict HVIR for a specific virus, the human T-cell leukemia virus type 1 (HTLV-1), which is responsible for causing a type of leukemia and associated diseases in adults. A viral protein called Tax plays a central role in the development of these diseases through its influence on multiple cellular responses such as cell cycle control, signal transduction, transcription (making RNAs from DNA copies of the virus genome), and programmed cell suicide [1]. Tax is thus a "Swiss-army knife" protein because of its multiple cellular interactions.

For the current phase, we restrict the sources of information that we use for populating the HVIR repository. The architecture outlining the prototype is shown in Figure 2.

We use two protein interaction databases and selected literature from Pubmed as source of information for the HVIR repository. The data from database sources is harvested and normalized to a standardized format (for our prototype, we will use BIND XML). The information from Pubmed literature is extracted using text mining approaches. The

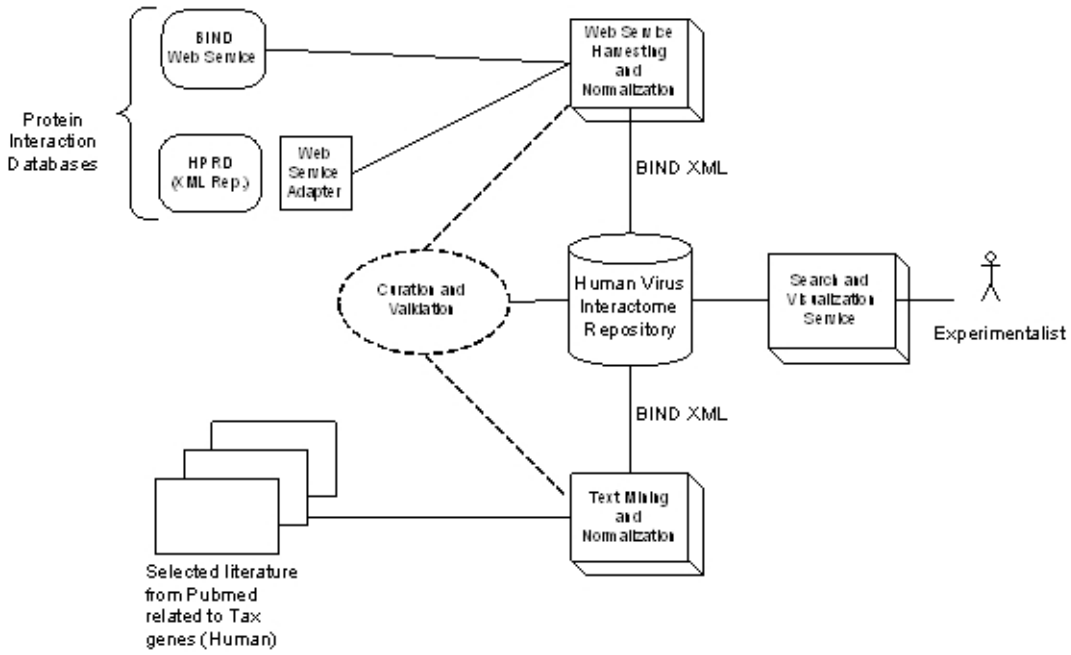

Figure 2: Prototype architecture.

information extracted is normalized and represented in BIND XML format and stored in HVIR repository. A human expert curates and validates the quality of data stored in the repository. This is an iterative process that has an impact on the text mining and normalization components. The search and visualization service provides an interface to experimentalist, which helps in predicting the interactions between viral and human proteins. We now discuss in details the major components of our prototype.

4 Processing Protein Interaction Databases

4.1 Literature Selection and Curation

We have manually searched and downloaded from the Pubmed database the most relevant papers listing interactions of the viral Tax protein with human proteins. Data from the manually created knowledgebase are used to create and fine tune syntax patterns for discovering protein interactions via mining. This knowledgebase is used to verify the correctness and effectiveness of the text-mining algorithms and the update strategies.

4.2 Text Mining and Normalization

Building the text-mining system is a challenging exercise due to the many keywords, synonyms, syntax patterns, and formats that describe protein interactions in the literature and databases [2]. E.g., there are 20 synonyms for the HTLV-1 virus in NIH's Virus genome database. We are developing algorithms to text-mine protein interactions and co-localizations from the literature.

4.3 Search and Visualization Service

We are using as the basis a network-drawing program for visualizing the interactome network, to present the interactome data to a user of HVIR. We have modified the standard program to allow for the representation of the interactions in a fashion that the levels of distances from Tax can be seen visually. That is all proteins interacting with tax that are a distance k away will all have the same x coordinates.

5 Status and Proposed Evaluation

Currently we have completed the prototype that includes some 40,000 protein interactions from the databases and also have text mined selected articles from PUBMED and integrated them into the HVIR database. Our visualization tools are functioning for allowing a user to select the number of levels to display and also will provide details for any interaction selected. The program will show all the paths from a selected protein back to Tax and finally the user can opt to show a network anchored around a selected protein with a diameter of k levels.

Specifically, the screenshot in Figure 3 shows the TAX protein in blue color and all its interactors. When we hover over a protein its gene name can be seen as a tooltip. Any protein one clicks on will change its color to red and its neighbors are highlighted with a darker border. If we would like to know more information about this protein we can click on the link on the right side of the window which will redirect the user to a different site.

Figure 4 shows all the proteins for up to level three. The path from any protein to tax can be highlighted by clicking on 'show path to tax' box at the top panel. The beginning and ending proteins are colored in blue and other proteins on the path are in red color.

Neighbors of a particular protein can be highlighted by clicking on 'highlight neighbors' box at the top panel. Neighbors can be easily identified as they are colored in yellow while the other are in

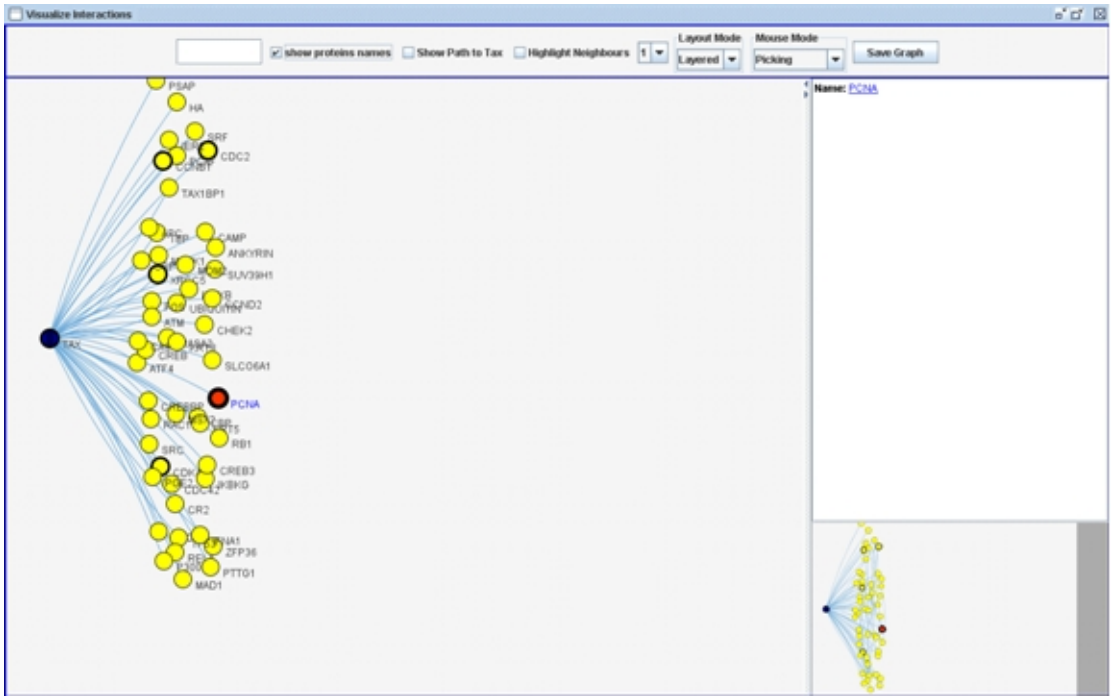

Figure 3: Tax and its first level interactors

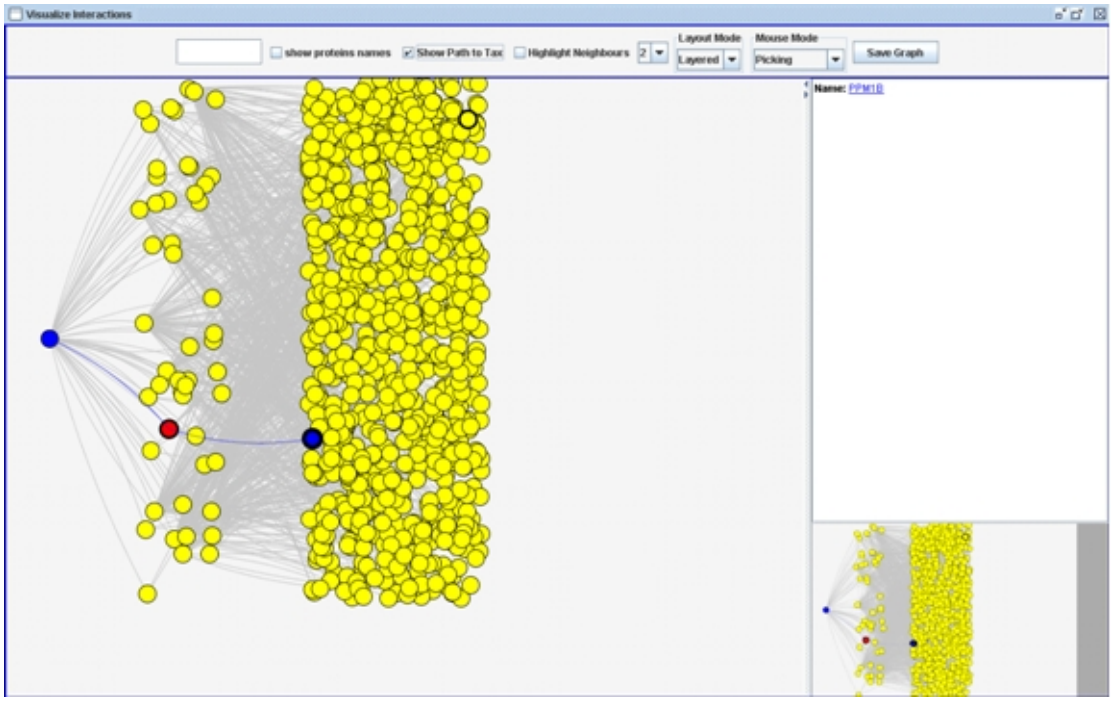

Figure 4: Shortest path from protein to TAX.

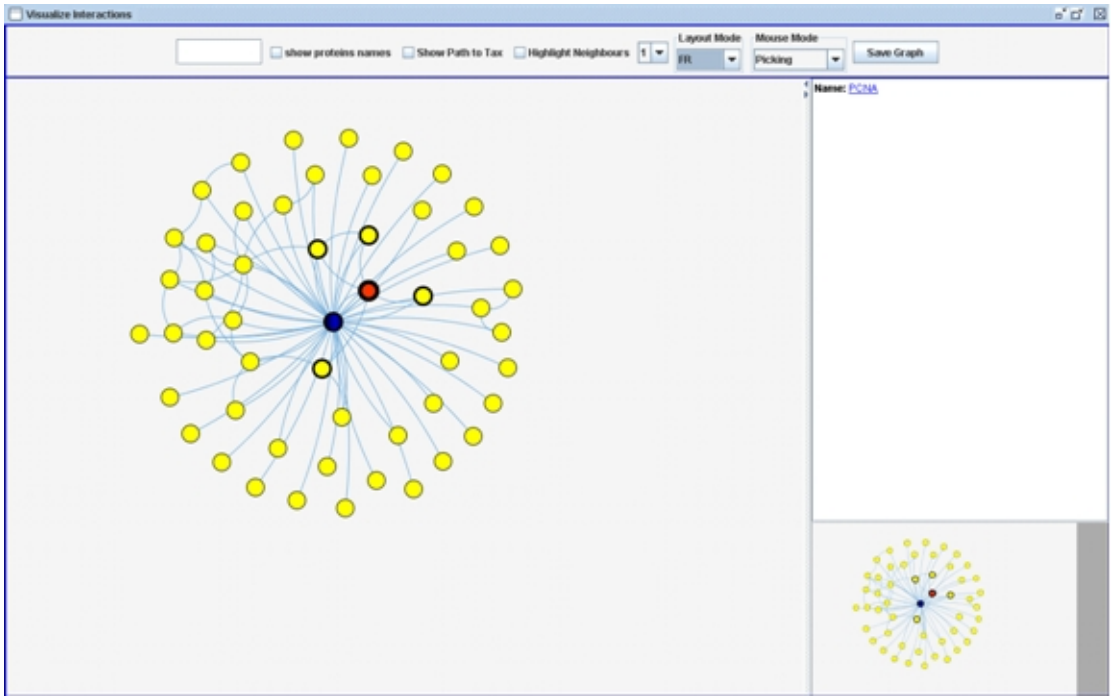

Figure 5: Neighborhood of a protein.

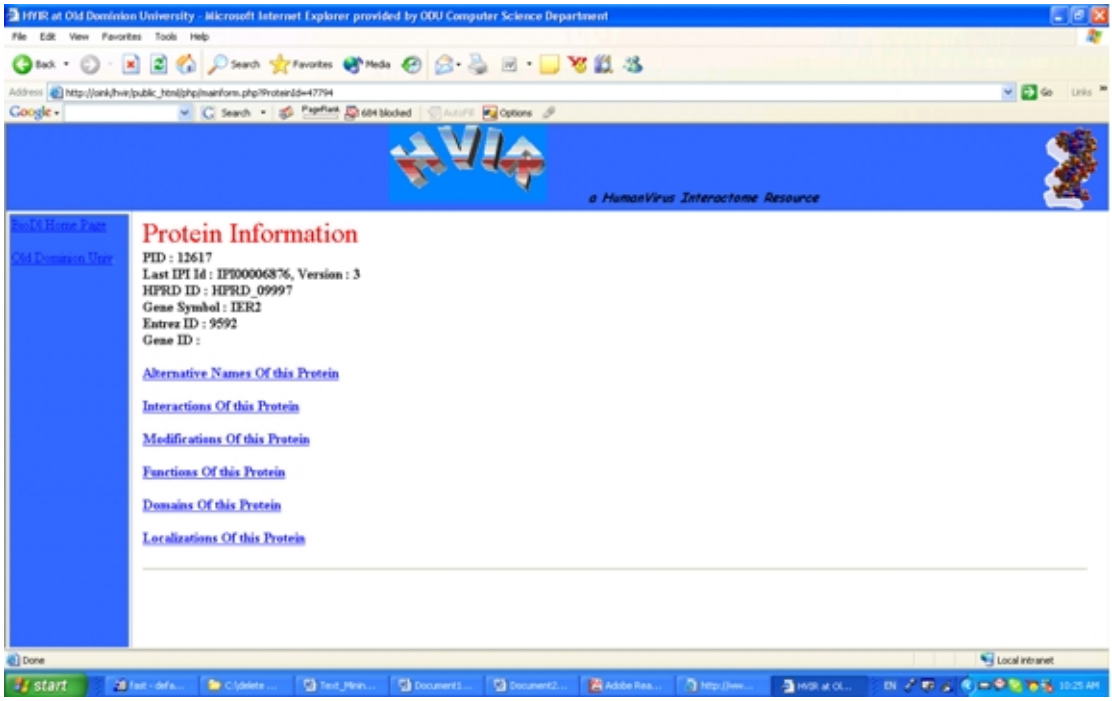

Figure 6: Protein information.

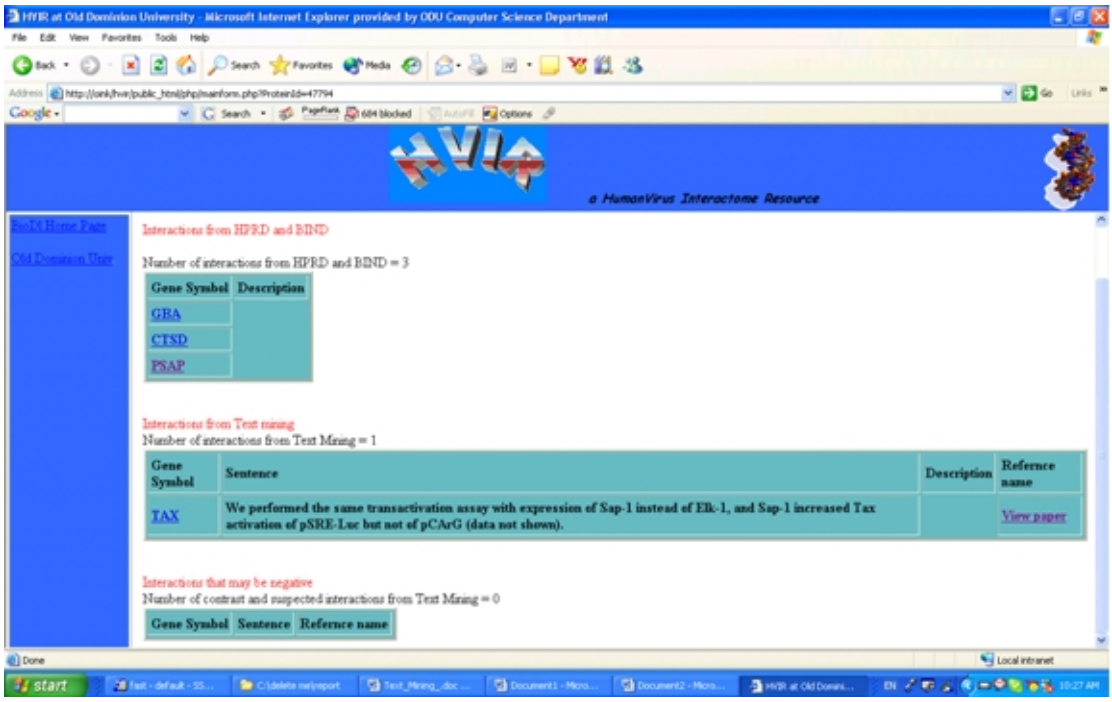

Figure 7: Protein interactions.

white. In Figure 5 we see a layered layout where the protein of central focus is at the left most corner and all it neighbors are in an higher level.

For any protein selected from the visualization, we provide more information about that protein like its HPRD id, IPI id, name, Entrez id as shown in Figure 6. We also provide some links to show the interactions, alternative names, modifications, functions and domains of that protein.

In Figure 7, for a selected protein, we show the interactions imported from other database as well as interactions extracted from our text mining tool whether it is a fact interaction or a contrast interaction. For the text mining results, the user can view the reference that contains this interaction along with the sentence at which this interaction occurred.

We will evaluate the design of HVIR for scalability, i.e., how effectively the algorithms employed in it deal with growing, heterogeneous, literature collections and databases. There are challenges involved in designing scalable text-mining algorithms, and in normalizing entity names in the context of new discoveries. As a simple example of the latter, consider a paper that concludes that two proteins reported by two different groups with different interaction partners are identical. This would cause the literature to be searched again for interactions to be discovered and the knowledgebase to be updated. Curating a dynamically changing repository by identifying and correcting inconsistent data is also difficult, due to the high error rates in experimental techniques such as yeast 2-hybrid and tagged affinity purification (TAP). This will be accomplished by developing probability scores for how likely two proteins are to interact using evidence from multiple experimental methods, data from homologous proteins in other organisms, and from the domain structure of proteins.

Heterogeneity of the sources of information on protein interactions is a major difficulty in designing a large-scale resource like the HVIR. In addition to yeast 2-hybrid and TAP experiments, microarray experiments, co-localization studies, novel proteomic technologies, and small-scale experiments provide interaction data. There are no information standards for representing experimental information in databases and for sharing them electronically, making it difficult to aggregate data from different sources.

Usability will be initially evaluated by the biologists on the team, and then by local virologists at EVMS and ODU.

References

[1] I. Azran, Y. Schavinsky-Khrapunsky, and M. Aboud. Role of Tax protein in HTLV1 leukemogenicity. *Retrovirology*, 1:20:24 pp, 2004.

[2] L. Hirschman, J. Park, J. Tsujii, L. Wong, and C Wu. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18: 15531561, 2002. doi:10.1093/bioinformatics/18.12.1553.

[3] C. Russell and R. Webster. The genesis of a pandemic influenza virus. *Cell*, pages 368–371, 2005.