Viewing proteomic experiments in context with known protein networks

Radu Jianu* Richard Park+ Arthur Salomon+ David H. Laidlaw*

Abstract— We present a system that allows proteomic researchers to view data derived from phosphorylation experiments in context with known protein interaction networks. We hypothesize that such a visual integration between new, experimental data and known interactions, done in an automated and streamlined fashion, enhances the proteomic research process.

We start from hand-drawn pathway representations, augment them with relevant protein interactions extracted from public databases, produce a novel, explorable visualization, and overlay results from phosphorylation experiments; these experiments indicate how proteins respond to stimuli and are one of the main methods of signaling pathway research. The system has been developed iteratively, in response to feedback obtained from collaboration with researchers using phosphorylation to study mast-cell signaling pathways. It was also recently deployed and integrated into their research lab. The main consequences of our work are: we enable researchers to put their experimental data in context with available knowledge about protein interactions, a task not supported by previous software; we introduced a novel network visualization method that works well for proteomic pathways but may have a wider application area; we identified tasks and work patterns in proteomics research, useful information in designing future applications.

Index Terms— pathway visualization, signaling networks, phosphorylation experiments, proteomic workflows, explorable networks.

1 Introduction

One protein can produce changes in the state of another protein through various biochemical mechanisms; such a relation between two proteins is commonly referred to as a *protein interaction*.

Protein interactions are the way activity is triggered and coordinated at the cell level: a stimulus interacts with a receptor protein located on the cell membrane; the receptor protein interacts with several proteins inside the cell which in turn can pass the message even further [Fig. 1]. A configuration of proteins in different states, or the synthesis of new DNA that can happen if the signal reaches the nucleus of the cell, determines the cellular outcome.

Such cascades of protein interactions that occur in a certain cell often in response to an external stimulus are conceptually grouped into a *signaling pathway*.

Understanding the details of how a pathway works, i.e. how information "flows" through protein interactions in response to a stimulus will lead to a molecular understanding of disease and provide hints about possible drug targets. Let us consider the example of the mast-cell that is responsible for the effects associated with allergy. The presence of allergen will trigger a protein signaling cascade within the mast cell causing it to release granules of histamine into the body and produce the known allergy and asthma symptoms [Fig2]. Current drugs to treat allergy block the perception of histamine by the target cells and do not inhibit the initial release of histamine from mast cells. However, understanding the intrinsic details of the mast-cell pathway could allow us to intervene at the cell level and block the production of histamine altogether without interfering with the other functions that the cell performs.

It seems natural that a way to study protein pathways is to artificially stimulate cells and measure how the proteins respond. One of the most important and common modification to proteins is phosphorylation. This modification of certain amino acids within a protein consists of adding a phosphate group to a protein molecule, significantly altering its function. It is also important to note that a protein can be phosphorylated at multiple locations within a single protein at distinct phosphorylation sites.

Current methods in proteomics are capable of the collection of information about phosphorylation on hundreds or thousands of proteins at a time. Since the number of proteins in general is on the order of tens of thousands, many of the proteins detected in the phosphorylation experiment will be unfamiliar to the researcher investigating the data. Also, previously discovered protein interactions can explain why and how a protein was triggered and have to be taken into consideration. Unfortunately, this type of information is also too large to be memorized by researchers.

As a result, the first step in using the phosphorilation data set is exploring it, discovering how it relates to already existent knowledge and making predictions about the data's significance. This process is currently highly manual. A researcher must read hundreds of papers, query online databases one protein at a time, track highly complex protein networks through memory and constantly take notes. Proteomic researchers can spend months exploring the data manually.

Here, we present a visualization system that can be fully integrated into the proteomic research pipeline, allowing the researcher to explore how results from phosphorylation experiments relate to already existent information in the field. We concentrate mainly on protein-protein interactions networks available from public sources, which we visualize in ways intuitive to proteomic

- *Radu Jianu is with Brown University. E-Mail: jr@cs.brown.edu*
- *Richard Park is with Brown University. E-Mail: richer_park@brown.edu*
- *Arthur Salomon is with Brown University. E-Mail: as@brown.edu*
- *David H. Laidlaw is with Brown University. E-Mail: dhls@cs.brown.edu*

researchers. Exploration of these known data elements is guided by the new experimentally derived quantitative proteomic data overlaid on top of the protein-protein interaction network.

We hypothesize that this approach will allow the researcher to become familiar with his experimental results much faster than using previously available tools, will help determine the significance of various parts of his data and ultimately build new, testable hypothesis about how the cell functions.

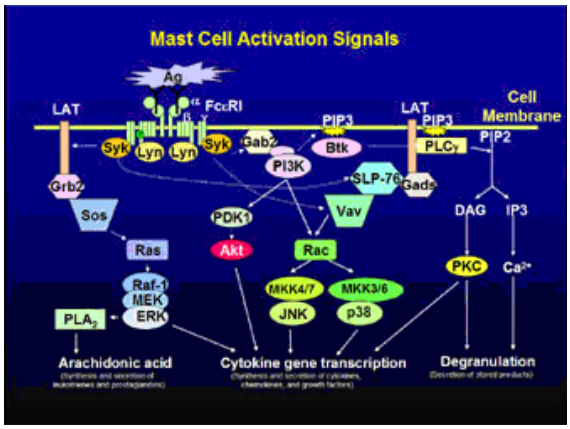

Fig. 1 Mast Cell Signaling Pathway

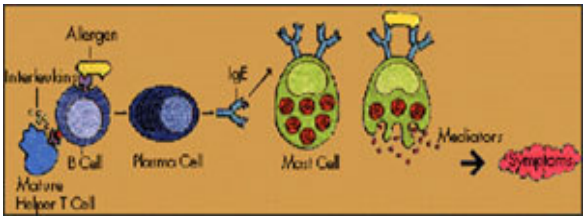

Fig. 2 Histamine production as a result of allergen stimulation of the mast cell

2 Related Work

The majority of work related to our methods is in the area of protein interaction network visualization. Visual integration of results from high-throughput proteomic experiments with known protein data is less common in the visualization or bioinformatics literature.

Discussions on drawing metabolic pathways started long ago with papers proposing conventions on representing interactions of biochemical entities [1][2][3]. These papers generally refer to the popular method of hand-edited drawings of pathways [Fig. 1]. These are overview, static depictions of how the pathway is known to function. The main advantages of this method are that the information contained in the pathways has a high degree of confidence and that they are aesthetically appealing. However, there are many disadvantages: they are available at cost or only for a limited range of protein pathways due to the necessity of human intervention, they provide little detail, and they cannot be extended with additional information, queried or altered in any way. A more technologically advanced variation of this kind of representations are the pathway databases. Kegg [4] and Biocarta [5] are repositories of hand assembled pathway illustrations also available in XML format and providing some basic querying capabilities.

With the proliferation of databases containing protein interactions dynamical solutions that build interactive representations guided by user input have emerged. Many of the protein interaction databases now include a visual exploration component that accepted proteins as input and produce various types of node-edge representation of their neighborhood PPID[6], STRING[7], MINT[8].

However, these visualizations where mainly intended to aid the user in browsing the database content rather than performing complex analytical tasks. So, although some include helpful features such as associating and displaying confidence scores for interactions, they often cannot deal with very large networks and do not have sufficient expressive power for extensive research activities.

More recently, the drawbacks of these lightweight visualizations have been recognized and a couple of more advanced stand-alone visualization systems appeared. Most notably Cytoscape [9], VisANT [10] or NAViGaTOR [11] are some of the leading protein pathway visualization systems. These applications usually offer multiple representation methods based on graph drawing algorithms such as [12], saving session capabilities and a plethora of features intended to make pathway analysis easier.

However, several problems can be identified if trying to use these tools for advanced protein interaction analysis. A common problem is that viewing a space larger than a few dozen proteins becomes difficult and results in a cluttered and illegible image. Heavyweight systems such as Cytoscape or NAViGaTOR usually allow the user to choose from several representation methods; unfortunately these are all instances of general graph-drawing methods that fail to exploit particular features of protein interaction networks.

Another issue that arises with the use of general graph-drawing algorithms (e.g. force-directed methods, simulated annealing etc) instead of tailored protein network algorithms is that they don't produce layouts that are consistent with the mental model of the protein researchers. Proteins need to be located in the final image according to rules that researchers intuitively take for granted: receptors at the top/left, information flowing top-down or left-to-right etc. Inconsistencies with these general conventions tend to make it harder for the researchers to find and reason about the information in the visualization.

A major shortcoming of current applications that our work addresses is the lack of flexibility to tailor the application to the needs of particular researchers or labs. Current visualization systems are hardwired to specific data-sources. Researchers cannot, for example, overlay their experimental results on top of existing knowledge, bring in new information that is not part of the database and use this information to highlight or filter out data. We are aware of very few systems that allow researchers to overlay some sort of experimental results onto protein interaction networks and of no method that collates protein phosphorilation experimental data with protein interaction networks.

We base our approach on the hypothesis that it is beneficial to allow researchers to have the research process guided by new experimental data rather than already discovered knowledge.

3 Methods

3.1 Overview

We start from hand-drawn overview representations of pathways that the users replicate into the system using a simple interface. We use the pathway overview as a seed around which we build a network with interactions from the STRING database. We create a network layout using an approach based on the initial overview of the pathway. We than create glyph-like representations of phosphorilation data which we attach to proteins. Finally, we use a focus + context method of exploring the network.

The methods section is structured as follows: we present the data we use; we discuss the protein layout generation and we illustrate the focus + context approach and exploration metaphors; we discuss a couple of details of implementation; finally we talk about the evaluation process.

3.2 Data Sources

3.2.1 Protein Interactions

We choose the STRING database as our main protein interaction source. Here, interactions are automatically derived by considering multiple sources of evidence. A major advantage of STRING is that a confidence score is associated to each interaction by considering

the reliabilities of the sources that derived the interaction. It is important to mention that one such source is "previous knowledge", a term that encapsulates interactions found in other available databases. A convenient consequence of this is that information from multiple databases not only makes its way into String but is also automatically curated and scored based on confidence.

The data we extract from STRING and use to drive our network generation is basically represented by a list of protein identifier pairs and an associated interaction confidence score.

We choose to use String as our primary protein interaction source mainly due to the opportunity we saw in using confidence scores to allow the user to explore the data at different levels of certainty. In addition, many other databases we might have considered should indirectly be present in String as "previous knowledge".

3.2.2 Experimental Data

The phosphorylation experimental data indicates both the timing and magnitude of change in phosphorylation in response to a stimulus over the course of cellular signaling. Thus, an example of possible result set is presented in [Fig 3]. Each line in the table represents a protein phosphorylation site. A line consists of the name of the protein that contains the site, the name of the site and the relative abundance of that phosphorylation throughout the time-course of the signaling cascade.

Protein Name	Phos. Site	0 Sec.	10 Sec	30 Sec	60 Sec
CDC2	Y15	18393108	24914670	14781200	20607424
CDC2	Y19	46185919	24914670	14781200	20261662
CDC2	T14Y15	9947569	39173197	28549987	26190809
CDC2	T14Y15	79070	133640	105107	153486
Cofilin 1	T88	355326	472624	327896	531363
Cofilin 1	Y89	228514	457787	335049	531363
cofilin - mouse	Y68		77714	93815	197222
Cofilin 1	Y140	23635	45114	40576	58584
Grb4	Y110		35092	126947	162665
Enolase 2	Y44	193534	230583	186331	199597
EphA3	Y31	82517	188907	175389	236533
EphA3	Y937	94181	106234	153040	242064
CHERP	Y682		24818	27002	18486
Ezrin	Y270	149590	201873	190872	196484

Fig. 3

3.3 Protein Layout

We initially experimented with using a general purpose graph-drawing algorithm (simulated annealing) to generate protein network representation. However, feedback from proteomic researchers was negative mainly due to aspects discussed in the related work section: mental model inconsistencies and clutter.

Our solution was to involve the researcher, at a small degree, in the layout generation. Since the most aesthetically pleasing pathway representations are considered to be the hand-assembled ones [Fig. 1], we use these representations as the starting point of our layout strategy. The user replicates the hand-assembled pathway into our visualization system using a simple interface: proteins are placed on a 2D canvas with simple mouse clicks, labeled and dragged to the desired locations. Protein connections are added by pair-wise selection of already placed proteins.

This first step results in what we will call a "pathway skeleton" which is an overview image of the pathway, very similar to hand-drawn signaling pathways. It has the advantage of being familiar to the researcher and containing only established proteins and interactions. We will refer to proteins belonging to the "pathway skeleton" as principal proteins.

Around the pathway skeleton we build a protein network using STRING interactions. The pathway skeleton is used as seeding and a graph is grown in breadth like fashion for a number of levels specified by the user. A minimum confidence threshold under which interaction are not considered has to be set. Proteins added to the network in this step and not present in the pathway skeleton will be referred to as secondary proteins.

The layout method is based on the fact that principal proteins already have a user assigned position. A secondary protein is placed by interpolating between the known positions of principal proteins, depending on their distance from the secondary protein being processed. Distance between two proteins, is represented by the number of interactions (network edges) between the two proteins.

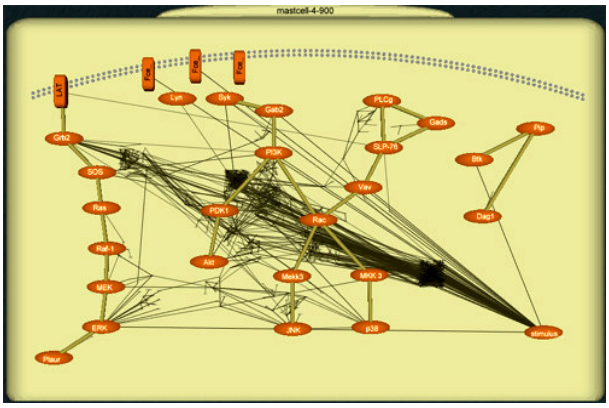

Fig. 4 Mast cell network extracted from STRING for a confidence value of 900 (out of 1000) and 4 levels out. The network contains approximately 1200 nodes and 3500 edges

The final layout is obtained by applying a few steps of a general force directed algorithm to locally adjust the layout and minimize overlap. An example of final output is presented in [Fig. 4].

3.4 Views

3.4.1 General Considerations

We have adopted a focus + context approach. Network structures and patterns can be observed in a global view of the pathway while exploration is performed in a detail rich one level at a time, radial view. The global and the exploration view coexist as two parallel 3D planes [Fig. 5].

Exploration is started by selecting a protein in the global view. A semi-transparent plane containing the selected protein in the center and its interactors around it is created above the global-view. Useful information about proteins is attached as glyphs to the protein nodes. The center of the exploration plane is located straight above the position of the center protein in the global view.

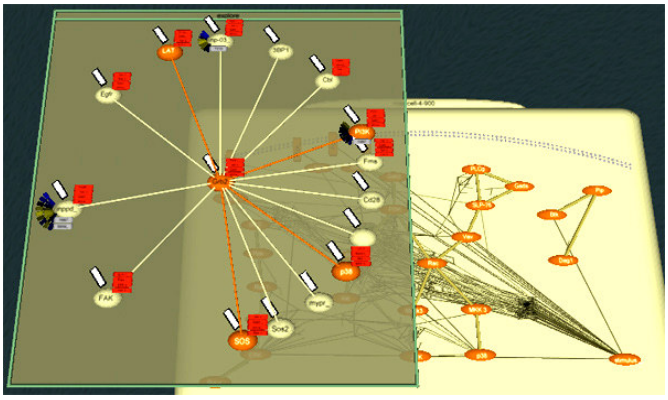

Fig. 5 Global and local views coexisting in exploration mode

The user can continue exploring by clicking on one of the interactors. With a smooth animation the exploration plane will glide over the global view and center itself above the global position of the new center protein. Simultaneously, the interactors of the old center are removed and the ones corresponding to the new center added. The process is illustrated in [Fig 6].

The focus + context approach yields the following advantages: the single level view provide space that enables us to display information that would clutter the global view; the global view on the other hand orients the user, ensures mental map preservation during exploration sessions and provides the ability of performing global tasks such as identifying highly connected proteins or finding paths.

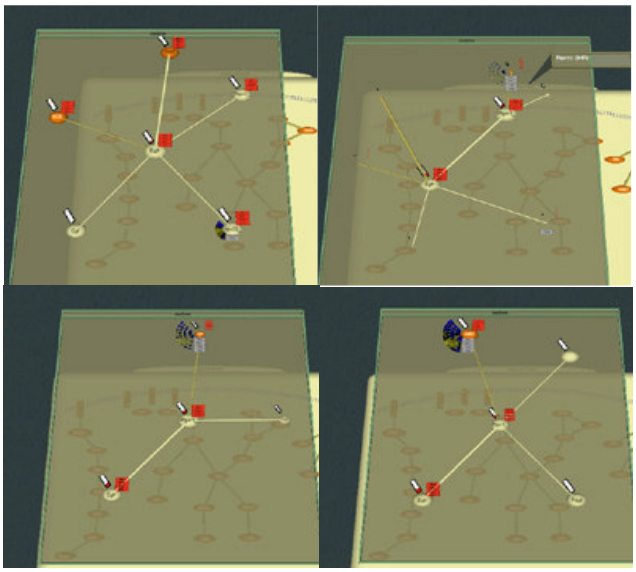

Fig. 6 Transitioning from one protein to another (order is: upper left, upper right, lower left, lower right)

3.4.2 Exploration view

The exploration view is represented by a semi-transparent 3D plane that moves over the global view [Fig 5]. It uses a radial layout to arrange neighboring proteins around a protein of interest located in the center of the plane. The layout is generated by placing the interactors at equal distance from the center node and the same direction relative to the center as they are in the global layout. Then, we apply a repulsive force on the nodes while keeping the edges rigid. This will move the nodes away from each other and minimize overlap.

Primary proteins are colored with the same color both in the global and in the exploration view; also, their movement is slower during the repulsion phase than that of secondary proteins. Because this last measure attempts to keep principal proteins on the direction copied from the global view, the layout in the explore view is likely to resemble the one in the global view.

Finally, we use several metaphors to aid the exploration process. First, the signpost glyph [Fig 7], informs the user what principal proteins can be reached if following a particular interaction. It also provides information about the number of jumps required to get to a principal protein.

A second metaphor is that of exploration bars [Fig 7], a glyph indicating how much of the neighborhood of a node has been explored. As edges leaving a node are explored the exploration bar is filled; as edges leaving its neighbors are covered the exploration bar is filled again, this time with smaller increments. The process can continue for an arbitrary number of degrees of separation.

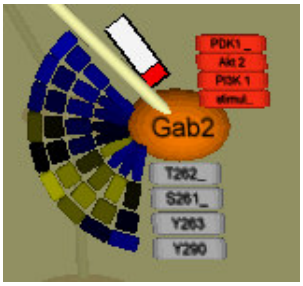

Fig. 7 Protein Glyphs. radial heatmaps for 4 phosphorylation sites; signpost with 4 destinations; exploration bar.

3.5 Phosphorylation Data

The experimental data used as input by our visualization is presented in [Fig.3]. This data undergoes normalization by the values of the highest change in protein expression observed in the experiment and is then transformed in a color heatmap representation. The colors range from blue (low expression) to yellow (high expression). They are then attached as a radial glyph [Fig. 8] to proteins in the explore view. As specified in the introduction section, a protein can change in several places; a protein node can thus have several heatmaps.

3.6 The phosphorilated proteins list

A consequence of the work patterns identified together with our collaborators is the addition of a list containing the experimentally revealed proteins. The researcher can use this list to systematically investigate his data set: phosphorylated proteins and their neighborhoods are explored one by one and they can be marked if considered significant. The explored bar glyph is displayed in the list as well to help the researcher keep track of explored versus unexplored proteins.

3.7 The stimulus node

Using the pathway skeleton as the only seed for network generation will cause the experimental data that is not connected through known interactions to the pathway to be lost. We therefore introduce a fake node that has all the experimental proteins connected to it. We call this node the “stimulus node” since conceptually all phosphorylation in the experiment were triggered by the artificially induced stimulus. The stimulus node allows the network to be grown both from the known pathway and from the experimental data. Investigating how these two networks connect is one of the interesting tasks that researchers might perform.

3.8 Evaluation

We have built the system iteratively constantly receiving feedback from our collaborators and redesigning aspects of the software. The initial layout, using simulated annealing, had to be changed due to its failure to comply with proteomic pathway drawing conventions. The focus + context approach was the result of a necessity of both detail and global views of the pathway. Finally, discussions have led to the identification and clarifications of the tasks that could not be performed without this software as well as the work patterns of proteomic researchers. This led to the introduction of the stimulus node and the phosphorylated protein list that would allow the researcher to analyze the data methodically.

4 Results

Probably the main result of our work is that proteomics researchers are now able to harness the advantages of high throughput phosphorylation experiments in ways impossible before.

As stated by our collaborators working in the biomedical department a researcher required several months to digest the result set of a phosphorylation experiment using previously available tools. A typical workflow involved analyzing and documenting the experimentally revealed proteins one at a time by querying web-databases and reading queries. Interactions were rarely considered because of lack of proper software. Our collaborators feel that by using our system and combine the known mast cell signaling network, the protein-protein interaction network STRING and quantitative data about phosphorylation, they will be able to generate more reliable hypotheses faster and build a more complete model of the structure of the signaling pathway.

Another result is identifying the problem of using general graph drawing techniques on protein pathways and providing an alternative user guided layout method. According to our collaborators the visual representations look more familiar then usual spider web diagrams, and proteins are easier to spot. We also hypothesize that this layout method can be useful in many other network analysis areas where a subset of nodes is deemed more important.

We also propose a novel method of exploiting the software by integrating it in the research pipeline of a proteomics lab. The phosphorylation experiment is run; the data is automatically processed and stored in a database; the visualization system loads the newly acquired data onto pre-constructed pathways and is ready for use. This mode of operation removes the tedious aspects of data management and allows the user to directly proceed to the more interesting aspects of discovering new things about protein interactions.

Finally, a notable result was identifying what the previous work process was and how it would be improved by visualization software such as ours. Manually querying each individual protein from the phosphorilation experiment will be replaced by systematically going through the a protein list; interactions, that were previously not considered because of the difficulties associated with mentally tracking networks can now be involved in the research process; connecting the network around the pathway with the network of the experimental proteins can now attempted.

5 Discussion

5.1 Layout

The user guided pathway layout was a consequence of two factors: information about "correct" placement of proteins in a visual representation is currently not available and is subject to individual researchers; and the representations deemed most aesthetically pleasing are the hand-drawn pathways.

However, we think that the method we proposed addresses a more general need for network visualizations where certain nodes have to be placed at specific positions or spatial relation to each other. Although we have identified this need in protein pathway visualizations, we hypothesize that this approach could be useful in other cases as well. There are many networks where a small subset of nodes is deemed to be more important or at least distinguishable from the others. In social networks for instance these could be individuals in key positions or with greater influence in the community. Having rules that specify where these "special" nodes are located in the drawings, or giving the user the possibility to place those nodes at desired locations and organizing the rest of the network around these nodes, can yield many benefits:

- it enables the user to become familiar with the network faster by providing some well established anchor points
- keeping the same set of important nodes but regenerating the rest of the network according to some different criteria will keep the layout of the important nodes in place which has several advantages:
    - it can reflect important differences between the networks;
    - it allows the user to adapt to the new network more quickly
- the placement of nodes can have more meaning instead of being driven by aesthetical criteria only.

5.2 Focus + Context choice

We believe that opting for a focus+context approach was in our case beneficial due to the nature of the tasks that we assume will be performed: exploring the experimental dataset in detail at a protein and interaction level; finding structure in the network as a whole is secondary.

5.3 Data issues

Working with protein data proved to be a much challenging task than we first anticipated due mainly to naming issues within protein databases in general and STRING in particular. At the biological level proteins are identified by a unique amino acid sequence. However, a certain proteins name and sequence goes through various changes as each protein is studied in more detail. Thus querying for the same protein across databases is non trivial. In addition to that, because of errors, sequences that are not proteins in their own right but are just part of larger proteins might be present in database with their own identifier and name. Furthermore, almost identical sequences that biologically represent the same protein can show up as two different proteins.

Phosphorylation experiments also produce sequences that need to be translated into STRING identifiers. To be able to integrate them with STRING data we need to obtain corresponding STRING identifiers for these sequences. Some simple features such as querying the STRING database for fragments of the proteins sequences such as those revealed in a proteomics experiment in addition to protein names or identifiers needed to be in our system to clarify the "true" protein in STRING that corresponds to a protein discovered in the experiment.

Another problem that we were confronted with was that many protein identifiers in STRING lack meaningful names or abbreviations because the protein sequence was not assigned a name. Since a protein identifier by itself does not provide much information to the researcher we had to use complementary databases, such as ENSEMBLE, to provide some naming for the proteins. Also, our software allows users to manually name and provide annotation to proteins.

An option that we considered was switching to another interaction database such as the Human Protein Reference Database (HPRD). The disadvantage over STRING is that it does not providence a confidence score; however a great advantage would is that HPRD is manually curated and thus is completely non-redundant.

6 Conclusion

We have introduced and tested a new approach to visual analysis of proteomic experimental data by putting it in context with what is already known in the field. We allow phosphorilation experimental data to be overlaid on protein interaction networks generated from data extracted from public databases. We have developed a new protein network layout technique, coupled with a focus + context approach. Collaborators evaluated the system as useful, considering that it allows them to digest experimental data in ways impossible with previous tools.

We identified proteomic works patterns, problems related to the layout of protein signaling pathways and proposed developing visualization applications that are integrated in the research pipeline of proteomic labs.

References

[1] Michal G. Biochemical Pathways [Poster]. Boehringer Mannheim GmbH, 1993

[2] Michal G. On Representation of Metabolic Pathways. *Biosystems 47: 1-7*, 1998

[3] Kohn K.W. Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol. Bio. Cell 10, 2703-2734,1999*

[4] Kanehisa M., Goto S., Hattori M., Aoki-Kinoshita K.F., Itoh M., Kawashima S., Katayama T., Araki M., and Hirakawa M. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res. 34, D354-357,* 2006.

[5] www.biocarta.com

[6] Husi H, Grant SG. Construction of a protein-protein interaction database (PPID) for synaptic biology. In Neuroscience Databases: a Practical Guide, 51-62, 2002.

[7] von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, Bork P. *Nucleic Acids Res. 35(Database issue):D358-62.*,2007

[8] Chatr-aryamontri A., Ceol A.,Montecchi Palazzi L., Nardelli G., Schneider M.V., Castagnoli L., Cesareni G. MINT: the Molecular INTeraction database. *Nucleic Acids Res.,* 2006.

[9] Shannon P., Markiel A., Ozier O., Baliga N.S., Wang J.T., Ramage D., Amin N., Schwikowski B., Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res., 13, 2498–2504,* 2003

[10] Hu Z., Mellor J., Wu J.,DeLisi C. VisANT: an online visualization and analysis tool for biological interaction data. *BMC Bioinformatics 5:17 doi:10.1186/1471-2105-5-17*, 2004

[11] http://ophid.utoronto.ca/navigator/

[12] Fruchtermann T.M.J., Reingold E.M. Graph drawing by force-directed placement. *Software, Practice and Experience 21 1129-1164,* 1991