BIOINFORMATICS

Vol. 22 no. 14 2006, pages e402–e407
doi:10.1093/bioinformatics/btl258

# Create and assess protein networks through molecular characteristics of individual proteins

Yanay Ofran[1,2,*,†], Guy Yachdav[1,2,3,†], Eyal Mozes[2], Ta-tsen Soong[2,4],
Rajesh Nair[1,2] and Burkhard Rost[1,2,3]

[1]Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street, New York, NY 10032, USA, [2]Columbia University Center for Computational Biology and Bioinformatics (C2B2), 1130 St. Nicholas Ave. Rm. 802, New York, NY 10032, USA, [3]NorthEast Structural Genomics Consortium (NESG), Columbia University, 1130 St. Nicholas Ave. Rm. 802, New York, NY 10032, USA and [4]Department of Biomedical Informatics, Columbia University, 630 West 168th Street, New York, NY 10032, USA

## ABSTRACT

**Motivation:** The study of biological systems, pathways and processes relies increasingly on analyses of networks. Most often, such analyses focus on network topology, thereby treating all proteins or genes as identical, featureless nodes. Integrating molecular data and insights about the qualities of individual proteins into the analysis may enhance our ability to decipher biological pathways and processes.

**Results:** Here, we introduce a novel platform for data integration that generates networks on the macro system-level, analyzes the molecular characteristics of each protein on the micro level, and then combines the two levels by using the molecular characteristics to assess networks. It also annotates the function and subcellular localization of each protein and displays the process on an image of a cell, rendering each protein in its respective cellular compartment. By thus visualizing the network in a cellular context we are able to analyze pathways and processes in a novel way. As an example, we use the system to analyze proteins implicated with Alzheimers disease and show how the integrated view corroborates previous observations and how it helps in the formulation of new hypotheses regarding the molecular underpinnings of the disease.

**Availability:** http://www.rostlab.org/services/pinat

**Contact:** pinat@rostlab.org; ofran@cubic.bioc.columbia.edu

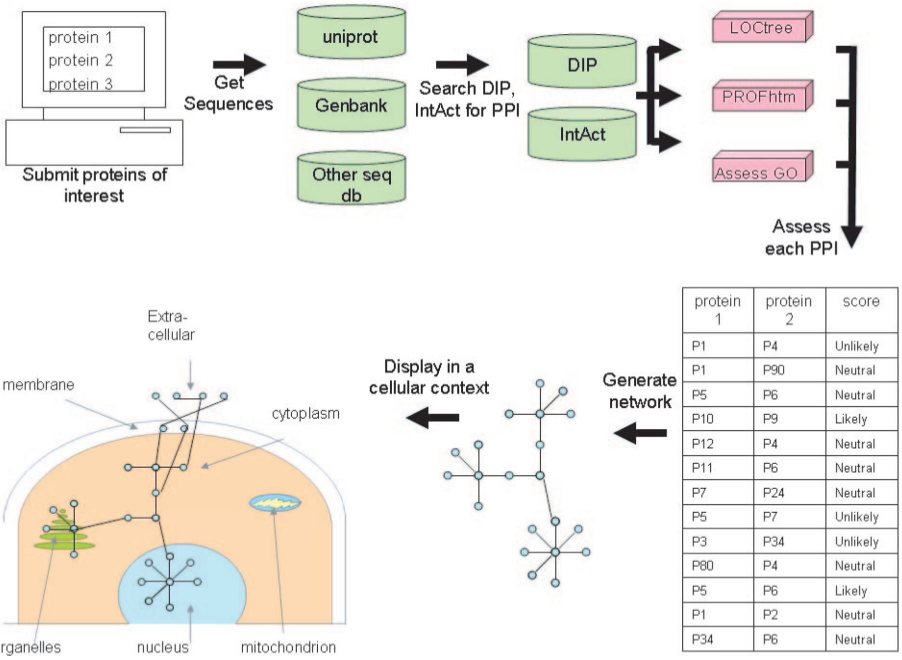
## 1 INTRODUCTION

Protein-protein interaction (PPI) networks are believed to constitute a valuable framework for the analysis of biological processes. Several studies attempt to characterize the topological properties of PPI networks as a whole (Barabasi and Oltvai, 2004), or of small, recurring elements within them (Wuchty, *et al*., 2003). The biological implications of such topological observations are still debated (Bork, *et al*., 2004). However, it has been suggested that the analysis of PPI networks can help identify biological ‘‘modules’’ namely networks of a limited number of proteins that interact to carry out a certain process or function (Ge, *et al*., 2003; Hartwell, *et al*., 1999). Parsing the topology of such networks could help decipher biological processes and assign function to un-annotated proteins that are implicated in these modules (Vazquez, *et al*., 2003).

The first step in the topological analysis of modules is the generation of PPI networks from pairwise protein-protein interactions. Numerous databases curate and sometime even predict protein-protein interactions based on various criteria (Bader, *et al*., 2001; Hermjakob, *et al*., 2004; Peri, *et al*., 2003; Rhodes, *et al*., 2005; von Mering, *et al*., 2005; Xenarios, *et al*., 2002; Zanzoni, *et al*., 2002). Although the vast majority of these data come from high-throughput experiments, they also include manually curated data from the literature. High-throughput PPI data are often rather noisy, and include a substantial amount of false positives (Cusick, *et al*., 2005). In particular, yeast two-hybrid experiments (Y2H) can yield false positive results of two kinds. (1) Experimental errors: two proteins observed to physically bind, may not interact in reality. (2) ‘‘*In vitro*’’ error: the conditions under which Y2H experiments are carried out may lead to interactions that do not occur *in vivo*. While the first type of errors can be reduced substantially by rather simple experimental adjustments, the second type of error is harder to control. The most effective approach thus far for identifying these false positives on a large-scale is through *in silico* analysis (Cusick, *et al*., 2005). Problems with the reliability or reproducibility of data are not confined to high throughput PPI dataset. A comparison of several datasets that were collected by experts from the literature revealed that the overlap between such sets is small (Ramani, *et al*., 2005), calling for caution when using them in an automatic manner. Thus, when using these data, it is imperative to assess the reliability of specific interactions.

Another problem with the analysis of PPI networks relates to data representation. Many higher-level studies of biological networks treat individual proteins as featureless nodes and focus their analysis on the topology of network graphs. Yet, the molecular details of the individual proteins are crucial for understanding and assessing networks. There is an essential connection between the structure of a PPI network and the molecular features of each protein. For example, most eukaryotic proteins are confined to particular subcellular compartments. Biological processes that span different compartments often consist of several modules. Each of these modules is typically localized to a different compartment, and a small number of proteins serve as connectors between compartments. The localization of a protein is instrumental for assessing PPI data, as proteins that reside in different compartments are less

[*]To whom correspondence should be addressed.
[†]Equal contribution.



| protein 1 | protein 2 | score |
|---|---|---|
| P1 | P4 | Unlikely |
| P1 | P90 | Neutral |
| P5 | P6 | Neutral |
| P10 | P9 | Likely |
| P12 | P4 | Neutral |
| P11 | P6 | Neutral |
| P7 | P24 | Neutral |
| P5 | P7 | Unlikely |
| P3 | P34 | Unlikely |
| P80 | P4 | Neutral |
| P5 | P6 | Likely |
| P1 | P2 | Neutral |
| P34 | P6 | Neutral |

**Fig. 1. Flow of the analysis and integration in PiNat.** PiNAT accepts protein names as input. Its output is the PPI network around these proteins, rendered on a figure of a cell representing the interplay between the different cellular compartments. It also returns a table with a score for the biological likelihood of each interaction. This is done by integrating molecular data regarding the individual proteins, PPI data and systems view of the process. In particular, using analysis of GO annotation and of predicted SCL PiNAT grades the interactions according to their biological likelihood and renders them at the appropriate SCL.

likely to interact than proteins that are located to the same compartment (Sprinzak, *et al*., 2003; von Mering, *et al*., 2002). Similarly, proteins with incompatible functional annotation are assumed not to be very likely to interact (Sprinzak, *et al*., 2003). It is increasingly acknowledged that there is a need for a framework that will integrate the micro level, namely the characteristics of each protein such as its localization and functional annotation, with the macro level, namely the topology of the network. A system of representation and analysis that will offer such integration may improve our ability to elicit reliable and useful insights from high-throughput PPI data.

Here we introduce PiNAT (Protein interaction Network Assessment Tool)—an automated system that generates PPI networks around proteins of interest. It automatically analyzes each sequence, assesses the reliability of the interactions based on molecular criteria, and displays the network within an image of a cell, in a way that represents the flow of the process between its compartments. The automatic assessment is based on the results of a large-scale meticulous analysis of a large, highly reliable dataset of PPI. Once a list of proteins is submitted to the system, the following sequence of events is initiated (Fig. 1):

(1) PiNAT automatically queries databases of PPIs and constructs a network of known PPIs.

(2) The sequences of all proteins are obtained.

(3) Based on these sequences the system predicts the subcellular localization for all proteins, including proteins that have no homologue with experimental annotations about localization.

(4) Each interaction is graded using a likelihood based on the predicted localization of the participating proteins.

(5) Where available, the GO annotation of each protein is obtained.

(6) Each interaction is graded based on the likelihood of interaction between proteins with these annotations.

(7) Finally, the network of interactions is displayed in a cellular context. It is readily visible from this display how the process flows between the different compartments of the cell.

We demonstrate the power of the system by generating, assessing and displaying the known fraction of the PPI network that underlies Alzheimer's disease (AD).

## 2 METHODS

### 2.1 Large-scale assessment of PPIs based on localization

We used the DIP core dataset (Deane, *et al*., 2002; Xenarios, *et al*., 2002) to generate the localization-based scores for PPIs. This is a large, reliable set of interactions each of which was observed by at least three different methods. We predicted the localization for each protein in this dataset and checked the probability of observing interactions between proteins from any combination of localizations.

Subcellular localization was predicted using two methods: (1) LOCtree, (Nair and Rost, 2005) that assigns the following major classes to eukaryotic proteins: extra-cellular space, cytoplasm, organelles, mitochondrion or nucleus. (2) PHDhtm, (Rost, *et al*., 1996) that predicts transmembrane helices. The sustained performance of both methods has been thoroughly established. LOCtree assigns each prediction a confidence level between 1 (low) and 10 (high). We considered LOCtree predictions with confidence scores <4 as ''low confidence'' and discarded them from the assessment. PHDhtm predicts whether or not a certain residue is embedded in a transmembrane helix (TMH) and assigns the prediction a confidence level between 0 (low) and 9 (high). We deemed a protein transmembrane if: (a) PHDhtm identified at least 20 transmembrane residues, and (b) the average confidence score of the 20 most reliable predictions was above 8.5. In the gold-standard analysis described below, we found that with these thresholds about 7% of the proteins were identified as transmembrane and approximately 60% of the nodes were predicted in a particular localization with high confidence. All other proteins were designated unknown.

For each pair of subcellular compartments, we assigned a likelihood grade of ‘‘likely’’, ‘‘unlikely’’ or ‘‘neutral’’, indicating the likelihood of interaction among proteins from these compartments. These grades were determined as follows: We ran LOCtree and PHDhtm for 4800 interactions from the DIP core set, involving a total of 2191 proteins. 1482 of the 2191 proteins were given a high-confidence prediction by either PHDhtm or LOCtree. Of the 4800 interactions, 2312 had high-confidence predictions for both proteins.

Since we had a total of 1482 proteins with high-confidence predictions, the total number of protein pairs—assuming symmetry—was 1,097,421; of these, 2312 ($\sim$1/475) were well-documented interactions. If we take as our null hypothesis that the knowledge of localization has no effect on the probability of interaction, the approximate expected number of well-documented interactions for each pair of compartments will be the total number of PPIs in this pair of compartments divided by 475. For each pair of compartments, we determined whether it was over- or under represented in the subset of well-documented interactions. We then used the binomial approximation to the cumulative hypergeometric probability distribution, to assign a p-value to this over- or under-representation. We used a p-value threshold of 0.01, i.e. we assigned each pair of categories a likelihood grade of ‘‘likely’’ if it was over-represented with a p-value <0.01; a grade of ‘‘low’’ if it was under-represented with a p-value <0.01; and a grade of ‘‘neutral’’ otherwise.

When analyzing a network, each edge is assigned a likelihood grade based on the predicted compartments of its two nodes. An assessment of likely or unlikely is assigned to an edge only if we have high-confidence predictions, from either PHDhtm or LOCtree, for both nodes. If one or both of the nodes have only low-confidence predictions, the edge is always assigned a neutral grade.

## 2.2 Automatic generation of networks

The first stage of the analysis in PiNAT is the automatic generation of a PPI network. This is done by taking the list of protein names submitted by the user and search both DIP (Xenarios, *et al*., 2002) and IntAct (Hermjakob, *et al*., 2004) for the interactions involving them. Users can specify what depth of the interaction tree around the proteins they are interested in. For example, a depth of 1 will retrieve all the proteins that interact with any of the query proteins; a depth of 2 will retrieve also the proteins that interact with the proteins at depth 1, and so forth. Finally, based on the protein names and accession numbers the sequences are retrieved from the relevant sequence database. It is also optional to submit to the PiNAT server a list of sequences or a complete interaction network.

## 2.3 Large-scale assessment of PPIs based on GO

Proteins in one biological process are more likely to interact than proteins in distinct processes. Therefore, we used the GO annotations of each protein in order to grade the likelihood of interaction between them. Since GO includes records inferred electronically (i.e. based on sequence or structure similarity), we only take the annotations that come from trusted experiments such as direct assays. We measured the distance between two GO terms as the information content of the minimum subsumer of the two terms (Lord, *et al*., 2003). Low information content reflects a highly specific concept shared by the two terms and indicates a close relationship between them. Since there is often more than one GO annotation available for a particular protein, for every annotation $c_k$ in protein $i$, we found its most similar term $c_{j\max}$ in protein $j$, and vice versa for each annotation in protein $j$. We then averaged these best similarities to obtain the GO score between the two proteins (Eqn. 1).

$$similarity\ (i,\ j) = \frac{\sum_{k=1}^{m} simGO\ (c_k, c_{j\max}) +\ \sum_{p=1}^{n} simGO\ (c_{i\max}, c_p)}{m+n} \qquad \text{Eqn.(1)}$$

where m and n are the respective numbers of annotations in $i$ and $j$, and $simGO(c_A, c_B)$ is the GO similarity between terms $c_A$ and $c_B$ according to the definition by Lord *et al*.

We used the proteins in the DIP core set and generated 100,000 random pairings of these proteins to derive a background distribution of GO similarity scores.

## 2.4 Display of networks in the cellular context

The predictions from LOCtree and PHDhtm are also used to visualize the location of the nodes in the network drawing in the following manner. Given the network and the predictions from LOCtree and PHDhtm, we generate a Graph Markup Language (GML) file for Cytoscape (Shannon, *et al*., 2003), placing each node in the drawing according to its predicted localization. Nodes in the drawing are divided among six groups: one for each of LOCtree’s five categories, and one for membranes. Note that we, incorrectly, assumed that all membrane proteins reside in the cytoplasmic membrane, due to the lack of accurate method that distinguishes *in silico* between proteins in different membranes. For purposes of placing a node in the drawing, we used the following intuition-based rules: a high-confidence prediction from PHDhtm overrides a high-confidence prediction from LOCtree; a high-confidence prediction from LOCtree overrides a low-confidence prediction from PHDhtm; and a low-confidence prediction from PHDhtm overrides a low-confidence prediction from LOCtree. There is also a seventh group for nodes for which LOCtree was unable to give even a low-confidence prediction; such cases, however, are relatively rare (<1%).

## 2.5 Alzheimer’s disease related pathway

A pathway of proteins implicated in Alzheimer was retrieved from the KEGG database (Goto, *et al*., 1997). The pathway includes 21 proteins and was manually gleaned from the literature. We used the 21 proteins as input to the PiNAT server, composed the network of interactions around them (depth=1), identified the most likely and the most unlikely interactions and rendered the network in a cellular context.

# 3 RESULTS AND DISCUSSION

## 3.1 Interactions across subcellular compartments

A first glance at the results of the large-scale assessment of PPIs based on subcellular localization (Table 1) confirmed the intuition: almost always, proteins from the same compartment had a higher chance of interacting with each other than do pairs of proteins from different compartments. The exceptions for the intra-compartment interactions were extracellular proteins that did not show a significant tendency to interact with each other. This makes biological sense, as extracellular proteins are often messengers that facilitate communication between cells. Hence, it is not surprising to find that they show only weak interaction preferences. In contrast, almost all low scores originated from distant compartments. Conversely, PPIs between nearby compartments were found to be likely. An exception to this was the interaction between transmembrane and cytoplasmic proteins, and between transmembrane and organellar proteins. While the first (transmembrane-cytoplasmic) was significantly lower than random, the latter (transmembrane-organellar) was significantly higher. This could be explained by the intricate trafficking system of

**Table 1.** scores for interactions between compartments

| | Extra cellular | Cytoplasmm | Orgnl | Mitochondrion | Nuclear | TM |
|---|---|---|---|---|---|---|
| Extra cellular | Neutral | | | | | |
| Cytoplasmic | Neutral | High | | | | |
| Orgnnellar | Neutral | Neutral | High | | | |
| Mitochondrial | Neutral | Low | Neutral | High | | |
| Nuclear | Low | High | Low | Low | High | |
| TM | Neutral | Low | High | Neutral | Low | High |

For each combination of subcellular compartments we computed whether the interaction between proteins from these compartments has a high probability, is neutral or has a low probability. The calculation is based on a large dataset of reliable PPI. Low, high significance refers to the expected probability under the null hypothesis ($H_0$: localization has no effect on the probability of interaction) with p-values <0.01 (for under- or over-representation). Combinations that did not differ significantly from the expectation were deemed neutral.

proteins to the membrane, which involves the organelles: globular proteins that interact with membrane proteins often get to their destination through the secretory pathway rather than through free diffusion in the cytoplasm.

### 3.2 Likely and unlikely interactions across GO

Fig. 2 shows the scores obtained from the analysis of GO annotations for positive interactions (i.e. interactions included in DIP core set) and negative interaction (randomly paired proteins from the core set). We found that over 52% of the positive interactions get a score higher than 3.25 while 95% of the negative ones get a score lower than 3.25. When using a lower threshold of 1.3, we could retain 81% of the positives but only reject half of the negatives. We therefore binned the GO similarity scores into three categories: ''likely'' for interactions with score greater than 3.25, ''unlikely'' for interactions with score smaller than 1.3, and ''neutral'' for any score within that range. Note, that since the ratio of interacting pairs to non- interacting pairs in a proteome is very small (Grigoriev, 2003), even at this cutoff false positive interactions will outnumber true positives. However, most of the negative interactions will be rejected while most of the positive ones will be accepted.

### 3.3 Alzheimer in the perspective of PiNAT

The main pathological manifestations in Alzheimer's are neuritic plaques and neurofibrillary tangles, both are abnormal protein aggregations. They differ in the proteins that are accumulated in them and in their localization. While the first occur in the extracellular space, the latter occur around the cytoskeleton in the cytoplasm (Chapman, *et al*., 2001). All the genes that were found to be linked to the disease are involved in the production or deposition of these aggregations (Selkoe, 2001). The mapping of compartments to the PPI network that is implicated in this process is, therefore, of particular interest. Fig. 3 shows the Alzheimer's related PPI, rendered by PiNAT into a cellular context. The main hub in this network is the Amyloid beta A4 protein (APP), a derivative of which constitutes the neuritic plaques. The function of APP is not entirely clear. It is generally accepted that it is a cell surface receptor (Selkoe, 2001).

However, it has been shown that APP is often cleaved and that part of it is secreted (Selkoe, 2001). It has also been reported that derivatives of APP were observed in the cytoplasm (Selkoe, 2001). APP can promote transcription activation, and it has been reported that some derivatives of it are located in the nucleus (Selkoe, 2001). The figure reflects this unclarity regarding APP's localization. It is displayed in the nucleus according to the LOCtree prediction (the prediction of LOCtree was given a confidence score of 4, which is our lower bound for accepting LOCtree predictions). However, PHDhtm identified a short transmembrane segment that due to its length fell just below the cutoff we set for considering proteins as transmembrane (Methods). Yet, the pattern of interactions around APP is in agreement with all the reported observations regarding its localization. APP interacts extensively with almost every compartment of the cell. For example, there are 17 proteins that are predicted to be nuclear in this network. Most of them are part of a connected component. However, if APP is removed from the network, the connected component immediately disintegrates leaving only two of the nuclear proteins connected to each other. Hence, PiNAT's display of the network corroborates the findings regarding APP's location.

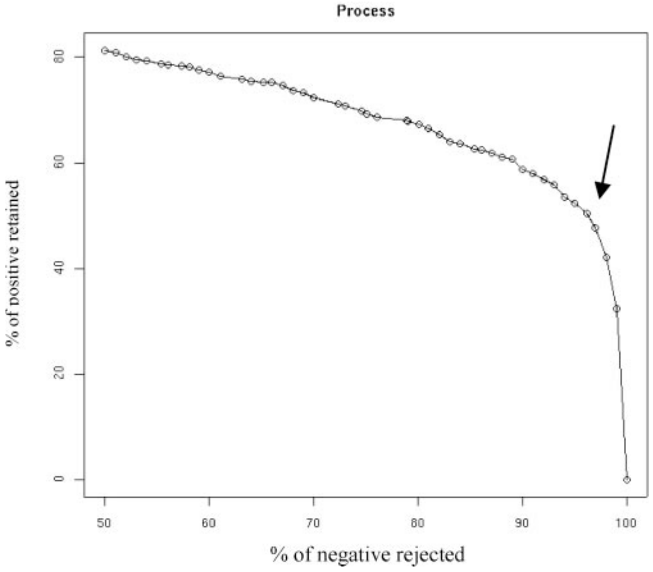


**Fig. 2.** Scores of interactions according to GO annotations on positive and negative data. Each point on the graph represents a certain score for interaction between proteins with different GO annotations. On the x axis is the percentage of negative (i.e. random) interaction that are below that score. On the y axis is the percentage of real PPI that get a score above that score. For a score 3.25 (arrow) 95% of the negative interactions will be rejected and the 52.4% of the positive ones will be retained.

Thirteen of the Alzheimer PPIs were deemed unlikely according to their localization (Table 2). Most of these interactions involved

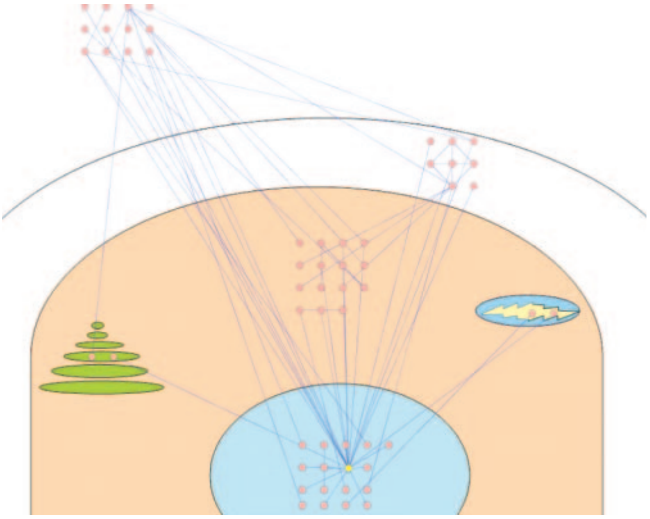


**Fig. 3. Output of PiNAT for Alzheimer's disease-related PPI network.** The proteins known to be involved in a pathway related to AD were used as an input to PiNat. Their PPIs were collected from DIP and IntAct to generate a network. Each protein in the network is displayed in its SCL according to the predictions of LOCtree and PHDhtm. The major hub of this network is the APP protein (colored yellow), whose metabolism is a major key for understanding the pathologies of AD. The myriad of interactions it has with many compartments of the cell is in agreement with suggestions regarding its functional diversity.

APP. This is understandable, as many interactions between APP, which was predicted to be nuclear, and proteins in non-nuclear compartments are considered unlikely. Still, knowing that APP and its derivatives can reside in different compartments calls for reevaluation of this classification. By and large, the two scoring schemes were in agreement about the likelihood of most interactions. The two scoring schemes radically disagreed for only one PPI, namely the interaction between NEDD8 Amyloid protein binding protein (ULA1_HUMAN) and APP (A4_HUMAN). The GO annotations of these proteins were very different. Hence their GO derived interaction score was low. However, since they were predicted to be in spatial proximity, the localization-based method scored the interaction between them as highly likely. This disagreement could also be ascribed to our poor understanding of APP and its function.

Interestingly, several of the proteins in this network have little or no functional annotation. Viewing them in the context of the cellular process can help to postulate hypotheses regarding the role they may have in Alzheimer's. It is widely accepted that inhibiting the cleavage of APP can slow down the advance of the disease and may even help prevent it. Thus, the exact localization in which the cleavage takes place is of a great interest. Identifying proteins in the network that may be involved in this cleavage may offer some important insights into this problem. Careful analysis of the localized network can suggest some additional insights into the molecular underpinnings of Alzheimer's disease and even help formulate new hypotheses. For example, it may be possible to determine which of the un-annotated proteins in the network may be involved in cleaving APP. Their SCL could serve to identify where the cleavage occurs. Clearly, such analysis is beyond the scope of this paper.

**Table 2.** Scores for Alzheimer related interactions

| protein 1 (UniProt) | protein 2 (UniProt) | SCL score | GO derived score |
|---|---|---|---|
| A4_HUMAN | A2MG_HUMAN | UNLIKELY | NEUT |
| A4_HUMAN | A4_HUMAN | LIKELY | NEUT |
| A4_HUMAN | ABB1_HUMAN | LIKELY | NEUT |
| A4_HUMAN | ABB2_HUMAN | LIKELY | NEUT |
| A4_HUMAN | ABB3_HUMAN | LIKELY | NEUT |
| A4_HUMAN | ACES_HUMAN | NEUT | UNLIKELY |
| A4_HUMAN | ACH7_HUMAN | UNLIKELY | NEUT |
| A4_HUMAN | APA1_HUMAN | UNLIKELY | NEUT |
| A4_HUMAN | APB1_HUMAN | NEUT | NEUT |
| A4_HUMAN | APE_HUMAN | UNLIKELY | NEUT |
| A4_HUMAN | ASP2_HUMAN | LIKELY | NEUT |
| A4_HUMAN | BACE1_HUMAN | UNLIKELY | UNLIKELY |
| A4_HUMAN | HCD2_HUMAN | UNLIKELY | UNLIKELY |
| A4_HUMAN | JIP1_HUMAN | LIKELY | NEUT |
| A4_HUMAN | LRP1_HUMAN | NEUT | UNLIKELY |
| A4_HUMAN | Q9UCX5 | LIKELY | NEUT |
| A4_HUMAN | SHC1_HUMAN | NEUT | UNLIKELY |
| A4_HUMAN | TGF1_HUMAN | UNLIKELY | NEUT |
| A4_HUMAN | TGFB2_HUMAN | UNLIKELY | NEUT |
| A4_HUMAN | TTHY_HUMAN | UNLIKELY | NEUT |
| A4_HUMAN | ULA1_HUMAN | LIKELY | UNLIKELY |
| ABB3_HUMAN | A4_HUMAN | LIKELY | NEUT |
| APH1A_HUMAN | PSN1_HUMAN | LIKELY | LIKELY |
| BIR2_HUMAN | CASP3_HUMAN | NEUT | NEUT |
| BIR2_HUMAN | CASP7_HUMAN | NEUT | NEUT |
| CASP3_HUMAN | BIR7_HUMAN | NEUT | NEUT |
| FLNA_HUMAN | PSN1_HUMAN | NEUT | NEUT |
| G3P2_HUMAN | A4_HUMAN | LIKELY | NEUT |
| GNB:3712673 | PSN1_HUMAN | UNLIKELY | NEUT |
| GSK3B_HUMA | REN3A_HUMAN | LIKELY | NEUT |
| IF38_HUMAN | NEP_HUMAN | NEUT | NEUT |
| LIPL_HUMAN | ACC2_HUMAN | NEUT | NEUT |
| LIPL_HUMAN | CSN6_HUMAN | NEUT | NEUT |
| LIPL_HUMAN | L7L2_HUMAN | NEUT | NEUT |
| LIPL_HUMAN | LRP1_HUMAN | NEUT | UNLIKELY |
| LIPL_HUMAN | PTN4_HUMAN | NEUT | UNLIKELY |
| LIPL_HUMAN | Q7L354 | NEUT | NEUT |
| LIPL_HUMAN | Q9P2H0 | NEUT | NEUT |
| LIPL_HUMAN | RL18A_HUMAN | NEUT | UNLIKELY |
| LIPL_HUMAN | ULA1_HUMAN | NEUT | UNLIKELY |
| LRP1_HUMAN | A2MG_HUMAN | NEUT | NEUT |
| NICA_HUMAN | APH1A_HUMAN | LIKELY | LIKELY |
| NICA_HUMAN | PSN1_HUMAN | LIKELY | LIKELY |
| O00193 | A2MG_HUMAN | UNLIKELY | NEUT |
| PEN2_HUMAN | APH1A_HUMAN | LIKELY | LIKELY |
| PEN2_HUMAN | NICA_HUMAN | LIKELY | LIKELY |
| PEN2_HUMAN | PSN1_HUMAN | LIKELY | LIKELY |
| Q7Z4Y5 | A2MG_HUMAN | NEUT | NEUT |
| Q9H5B5 | TAU_HUMAN | NEUT | NEUT |
| Q9UJZ5 | PSN1_HUMAN | NEUT | NEUT |
| R11A_HUMAN | PSN1_HUMAN | UNLIKELY | NEUT |
| RL10_HUMAN | PSN1_HUMAN | NEUT | NEUT |
| TAU_HUMAN | TBBX_HUMAN | NEUT | UNLIKELY |

List of PPI that were extracted from DIP and IntAct to generate the Alzheimer's related network. In the first and second columns are the UniProt protein names for each of the proteins in a given interaction. The third column is the SCL-based score for this interaction and the fourth column is the GO derived score

## 4 CONCLUSIONS

The integration of molecular knowledge and network structure can enhance our understanding of biological processes and of pathways. PiNAT, which is fully automated, offers a framework for combining the micro-level analysis of individual molecules and the macro-level of network topology. Users can submit a single protein, a list of proteins, or a whole network as an input. As an output they will receive a visual description of the predicted spatial flow of a pathway in the cell. In addition the user will get a list of scores for each interaction, based on different sequence-level analyses of the individual proteins. PiNAT is easily expandable. Thus, it will be possible to add many other molecular and network analyses to improve our insights into pathways and modules. Our example for the case of Alzheimer's illustrated just some aspects of the usefulness of PiNAT. The server is available at: www.rostlab.org/services/pinat

## 5 ACKNOWLEDGEMENTS

Thanks to Kazimierz O. Wrzeszczynski (Columbia) for helpful comments and discussion. This work was supported by the grants: RO1-GM64633-01 from the National Institutes of Health (NIH) and the grant RO1-LM07329-01 from the National Library of Medicine (NLM). Last, not least, thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases.

## 6 REFERENCES

Bader,G.D., Donaldson,I., Wolting,C., Ouellette,B.F., Pawson,T. and Hogue,C.W. (2001) BIND—The Biomolecular Interaction Network Database. *Nucleic Acids Res*, 29, 242–245.

Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5, 101–113.

Bork,P., Jensen,L.J., von Mering,C., Ramani,A.K., Lee,I. and Marcotte,E.M. (2004) Protein interaction networks from yeast to human. *Curr Opin Struct Biol*, 14, 292–299.

Chapman,P.F., Falinska,A.M., Knevett,S.G. and Ramsay,M.F. (2001) Genes, models and Alzheimer's disease. *Trends Genet*, 17, 254–261.

Cusick,M.E., Klitgord,N., Vidal,M. and Hill,D.E. (2005) Interactome: gateway into systems biology. *Hum Mol Genet*, 14 Spec No. 2, R171–181.

Deane,C.M., Salwinski,L., Xenarios,I. and Eisenberg,D. (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics*, 1, 349–356.

Ge,H., Walhout,A.J. and Vidal,M. (2003) Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet*, 19, 551–560.

Goto,S., Bono,H., Ogata,H., Fujibuchi,W., Nishioka,T., Sato,K. and Kanehisa,M. (1997) Organizing and computing metabolic pathway data in terms of binary relations. *Pac Symp Biocomput*, 175–186.

Grigoriev,A. (2003) On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Res*, 31, 4157–4161.

Hartwell,L.H., Hopfield,J.J., Leibler,S. and Murray,A.W. (1999) From molecular to modular cell biology. *Nature*, 402, C47–52.

Hermjakob,H., Montecchi-Palazzi,L., Lewington,C., Mudali,S., Kerrien,S., Orchard,S., Vingron,M., Roechert,B., Roepstorff,P., Valencia,A., Margalit,H., Armstrong,J., Bairoch,A., Cesareni,G., Sherman,D. and Apweiler,R. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res*, 32, D452–455.

Lord,P.W., Stevens,R.D., Brass,A. and Goble,C.A. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19, 1275–1283.

Nair,R. and Rost,B. (2005) Mimicking cellular sorting improves prediction of subcellular localization. *J Mol Biol*, 348, 85–100.

Peri,S., Navarro,J.D., Amanchy,R., Kristiansen,T.Z., Jonnalagadda,C.K., Surendranath,V., Niranjan,V., Muthusamy,B., Gandhi,T.K., Gronborg,M., Ibarrola,N., Deshpande,N., Shanker,K., Shivashankar,H.N., Rashmi,B.P., Ramya,M.A., Zhao,Z., Chandrika,K.N., Padma,N., Harsha,H.C., Yatish,A.J., Kavitha,M.P., Menezes,M., Choudhury,D.R., Suresh,S., Ghosh,N., Saravana,R., Chandran,S., Krishna,S., Joy,M., Anand,S.K., Madavan,V., Joseph,A., Wong,G.W., Schiemann,W.P., Constantinescu,S.N., Huang,L., Khosravi-Far,R., Steen,H., Tewari,M., Ghaffari,S., Blobe,G.C., Dang,C.V., Garcia,J.G., Pevsner,J., Jensen,O.N., Roepstorff,P., Deshpande,K.S., Chinnaiyan,A.M., Hamosh,A., Chakravarti,A. and Pandey,A. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, 13, 2363–2371.

Ramani,A.K., Bunescu,R.C., Mooney,R.J. and Marcotte,E.M. (2005) Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol*, 6, R40.

Rhodes,D.R., Tomlins,S.A., Varambally,S., Mahavisno,V., Barrette,T., Kalyana-Sundaram,S., Ghosh,D., Pandey,A. and Chinnaiyan,A.M. (2005) Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol*, 23, 951–959.

Rost,B., Fariselli,P. and Casadio,R. (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci*, 5, 1704–1718.

Selkoe,D.J. (2001) Alzheimer's disease: genes, proteins, and therapy. *Physiol Rev*, 81, 741–766.

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

Sprinzak,E., Sattath,S. and Margalit,H. (2003) How reliable are experimental protein-protein interaction data? *J Mol Biol*, 327, 919–923.

Vazquez,A., Flammini,A., Maritan,A. and Vespignani,A. (2003) Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol*, 21, 697–700.

von Mering,C., Jensen,L.J., Snel,B., Hooper,S.D., Krupp,M., Foglierini,M., Jouffre,N., Huynen,M.A. and Bork,P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*, 33, D433–437.

von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417, 399–403.

Wuchty,S., Oltvai,Z.N. and Barabasi,A.L. (2003) Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat Genet*, 35, 176–179.

Xenarios,I., Salwinski,L., Duan,X.J., Higney,P., Kim,S.M. and Eisenberg,D. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, 30, 303–305.

Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) MINT: a Molecular INTeraction database. *FEBS Lett*, 513, 135–140.