**Temporal phosphorylation and network analysis of yeast signaling**

Dylan Cronin1,2,3, Anthony Gitter1,2

1University of Wisconsin-Madison, Madison, WI

2Morgridge Institute for Research, Madison, WI

3Bowling Green State University, Bowling Green, OH

**Abstract**

Current methods for identifying protein interaction networks due to a cellular stress response do not completely utilize all of the data that has now become available. The computational method, Temporal Pathway Synthesizer (TPS), uses temporal phosphoproteomic data to identify signaling pathways. Using this newfound temporal phosphorylation data, TPS can determine pathways by checking all possible pathways from the receptor node, and eliminate any potential paths that do not match the temporal data. TPS is used here to analyze the osmotic stress signal transduction pathways, and to determine whether or not the method is successful in identifying the potential proteins involved in the stress response compared to the literature already known. The results indicate that TPS is effective in its determination of the signaling pathway, but the methods also have room for improvement.

**Introduction**

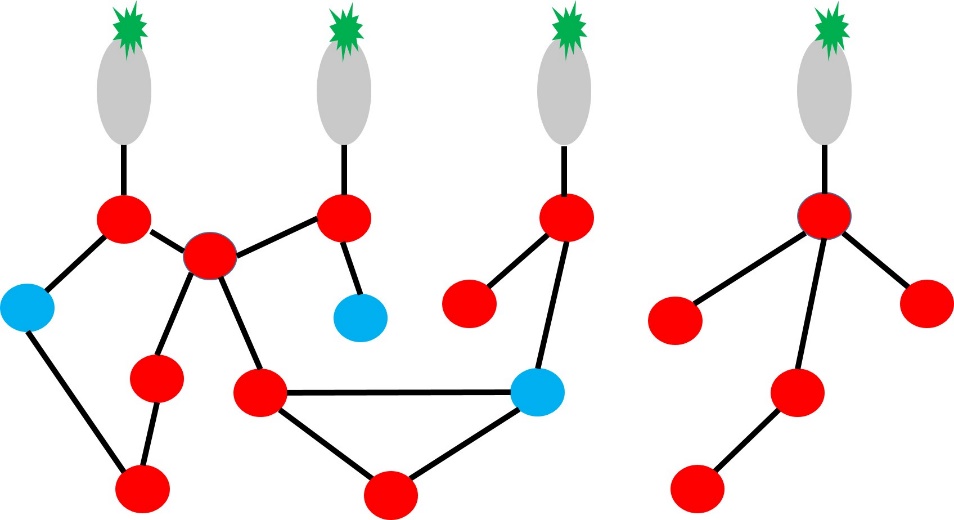
With increasingly more data becoming available, there is much more opportunity for the analyses of cellular processes in ways that were unimaginable before this wealth of information. One particular dataset, which is of great importance is protein activity tracked over time. Mass spectrometry enables one the ability to collect this phosphorylation data for analysis. In particular, mass spectrometry can provide a view of post-translational modifications in terms of many different types of modifications (Choudhary et al., 2010). Analyzing phosphoproteomic data can help identify the protein signaling networks that occur within the cell. When a stimulus response affects a cell, there is a cascade of protein interactions from the stimulus as the cell adapts to its new environment (Figure 1). This cascade can lead to a variety of effects from conformational changes to providing a regulatory role in downstream events. A ligand or chemical provides the initial stimulus to the cell, and then there are a series of protein-protein interactions to achieve some desired response (Kyriakis 2014). In signaling, these protein-protein interactions primarily occur through phosphorylation, and the end result is an effect on some specific transcription factors that bind to DNA and impact the expression of a particular gene. Network modeling serves to provide a method in which prior information in the pathway is not required, and use of this mass spectrometry data can provide insights that were not known otherwise. Not all proteins in an interaction network are phosphorylated, so it is quite possible that there is some other mechanism influencing some of the proteins, but the approach described will allow for these proteins to still exist in the network even without significant phosphorylation changes. Analyzing a protein network is necessary to understand more about the pathway and methods in which one could affect these signaling pathways.

Figure 1: This is a depiction of the cell response to a stress. The stress being some chemical (in green) binds to a receptor protein (gray ovals) cause a cascade of effects to proteins within the cell (red circles).

There are numerous methods to identify a subnetwork amongst the large set of interactions that phosphoproteomic data can provide, but refining these methods to make the models much more precise is necessary. Temporal phosphorylation data for specific proteins can allow for the elimination of possible pathways and further increase the confidence in a particular model. The software packages used in this analysis are Omics Integrator and Temporal Pathway Synthesizer. Omics Integrator uses protein phosphorylation data and an interactome of proteins to produce an unsigned, undirected network (Tuncbag et al., 2016). The Temporal Pathway Synthesizer utilizes a protein-protein interaction network along with temporal phosphorylation data to determine a signed, directed graph (Koskal et al.).

The microorganism of choice for identifying the effectiveness of the proposed model is Yeast because there is so much already known about this organism. Yeast have the ability to activate condition-specific transcript changes that provide specialized stress defenses, which allows for yeast to be a great microorganism for evaluating the proposed method (Hohmann & Mager, 2003).

**Methods**

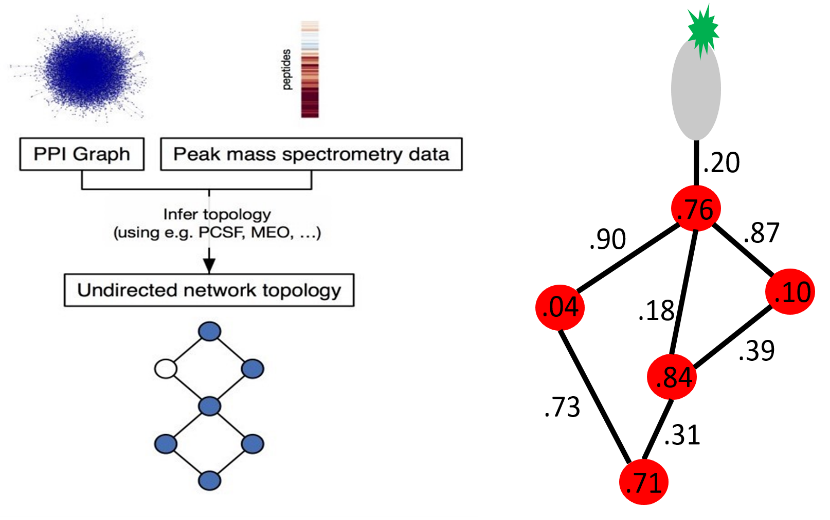
There are two existing modules to the Omics Integrator software package, the Forest Module and the Garnet module (Tuncbag et al., 2016). In this method, the Forest module is the sole module used from just the Omics Integrator software package because there is no gene expression data. The Forest module uses the Prize-Collecting Steiner Forest algorithm (PCSF) to predict a network structure based on phosphorylation data of each individual protein. PCSF uses the nodes within an interaction network as proteins, and assigns each protein a prize. Each prize in the interactome is determined based on the phosphorylation intensity for each protein after salt treatment relative to a control treatment. Then, the edges are assigned a weight based on a confidence score of the interaction between the proteins that connect to the edge. The higher the confidence that a particular interaction exists, the lower the weight on the edge. PCSF then uses these prizes and edge weights to maximize the prizes in the pathway and minimize the edge weights. Figure 2, for example, demonstrates high edge weights and low prize score for those nodes found on the left of the graph. Therefore, that particular connection of the graph would be removed from the set of possible, optimal protein-protein interaction networks. Maximizing the prizes and limiting the edge weights results in an undirected network with proteins that have been significantly phosphorylated with high confidence in the interactions between these proteins.

Figure 2: The workflow on the left demonstrates the process of the Forest module of Omics Integrator (Tuncbag et al., 2016). The image on the right is a depiction of the prizes (on each node) and edge weights (based on a confidence score) for the use of PCSF.

The Forest module uses a protein-protein interaction network and phosphoproteomic mass spectrometry data as input (Figure 2). This module also takes a set of up to six parameters, ω, β, D, μ, g, and garnetBeta. ω, β, and D are required parameters to run the Forest module. ω controls the number of trees in the network, β is responsible for the tradeoff between using more nodes and including less reliable edges, and D controls the depth of the network from the receptor protein down to the final node in the network. Then, μ, g, and garnetBeta are all optional parameters where μ restricts the abundance of hub nodes (nodes that have connections to many other proteins), g affects the convergence of the solution, and garnetBeta is used to scale Garnet output. Parameter sweeps were completed using ω, β, D, μ, and g. The sweeps of parameters were run as follows: beta was run from .25 to 10 in intervals of .25, μ was run from 0 to .1 in intervals of .005, ω was run from .5 to 10 in intervals of .5, D was set a constant of 10, and g was set at a constant of 1e-3. A parameter sweep was processed for each of these variables, and an optimal network was created from each run of these variables. These parameter sweeps produced 16800 different forests. This produced a set of many possible optimal solutions for the network, and these solutions were narrowed down by using only those networks that produced an optimal forest of greater than two nodes that were not phosphorylated significantly. Also, another selection of these possible optimal networks was made by selecting the networks with a maximized fraction of nodes that were significantly phosphorylated. Using the optimal parameters determined, the Forest module was run again with the determined set of parameters. The determined parameters were used to run the algorithm with noise added to the interaction costs, and the union of the forests obtained was used as input for TPS. This generates a protein interaction subnetwork that contains parallel paths between proteins instead of only the single best path.

TPS then uses this network produced by the Forest module as input. The software package also needs many other input files, including time series phosphorylation data, a significance test of every time point compared to the first time point, a significance test comparing every time point to its previous time point, a mapping file that declares which phosphorylation sites correspond to particular proteins, a source node for the receptor protein, and a threshold. The threshold was set at a constant of 0.01. TPS then uses this time series phosphorylation data to further refine the network found from the Forest module of Omics Integrator. First, TPS finds the direction of all of the edges in the interactome, whether they are directed or undirected, and explores all possible pathway models that align with this direction. Then, TPS checks from the source node to the last node if the particular pathway is in agreement with the time series data provided as input and the direction of all of the nodes in the network. TPS produces a final network that is in agreement with the temporal phosphoproteomic data.

**Results**

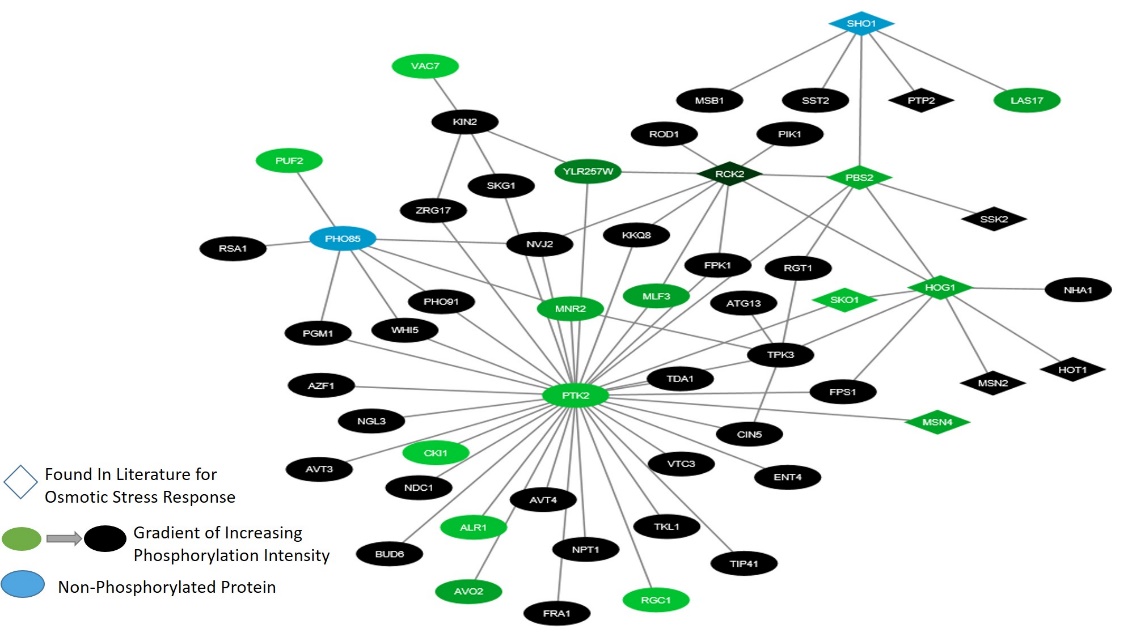
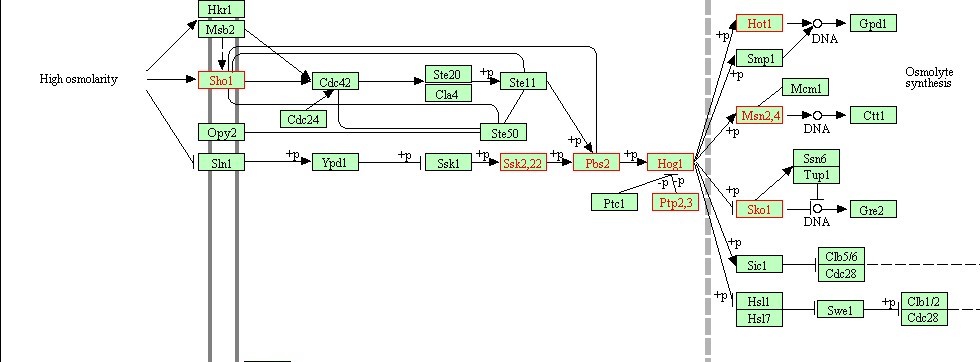
Taking the result from the TPS software package yields a final protein-protein interaction network (Figure 3). Figure 3 was visualized by inputting the network into Cytoscape v3.4.0 (Shannon 2003).The protein names in this network were mapped from systematic names to standard names using *Saccharomyces* Genome Database (Cherry 2012). Each node in the graph is representative of a protein in the network. The receptor node, used as a source for TPS input, was found in the network, and is located at the top of the network. Many nodes of the nodes deemed significant in this network were not found in the literature for the osmotic stress response in yeast. Then, the nodes found in Figure 4 were overlaid onto a KEGG pathway for osmotic stress response in yeast. The results indicate that there were eight nodes identified from the output of TPS in the KEGG pathway. Proteins found in the KEGG pathway that were in the network produced, include Sho1, Ssk,22, Pbs2, Hog1, Ptp,2,3, Hot1, Msn2,4, and Sko1 (Figure 4). The complete interaction network started with 3543 proteins. After using the Prize-Collection Steiner Forest algorithm within the Forest module of Omics integrator, this network was reduced down to 250 nodes. Finally, the network was further pruned to 55 total nodes in the final network (Figure 5).

Figure 4: This pathway is a pathway produced by KEGG, where each node in the pathway above represents a protein in the osmotic stress pathway for yeast (Kanehisa et al., 2012). The nodes with a red outline represent the nodes found in the output from TPS.

Figure 3: The image represents the protein-protein interaction network discovered by using TPS.

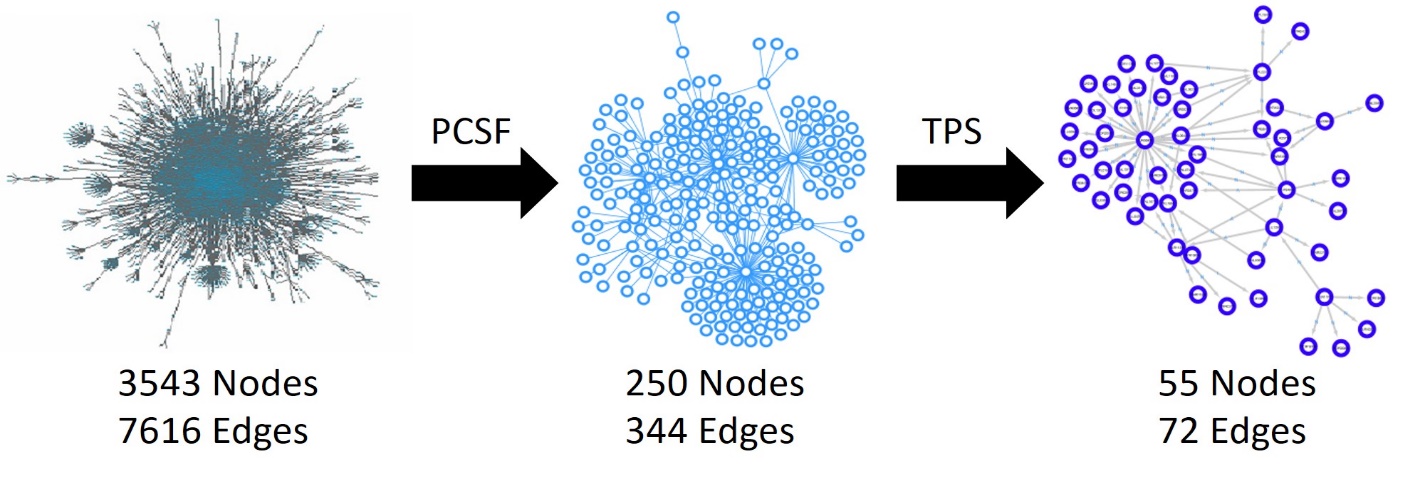
**Discussion**

Figure 5: The figure represents the first interaction network, with 3543 nodes, and after using PCSF, the graph was reduced to 250 nodes. Finally, after applying TPS to the data, 55 total nodes were obtained.

The results indicate that there is much more to learn about the yeast osmotic stress pathway, and that the method described was effective in producing a network that follows what is already known about the yeast pathway. Figure 3 reveals that in a network with 55 nodes, 8 of those nodes were found in literature as having a major impact in the osmotic stress pathway. 47 nodes that were found in the network produced were not found to be in the KEGG protein-protein interaction network (Kanehisa et al., 2012). Of the eight nodes found in the pathway that were present, Pbs2 and Ssk2,22, produced an interaction opposite of what is found in the KEGG pathway. In the network produce by TPS, Pbs2 phosphorylates Ssk2,22. This could indicate that there are alternative interactions possible amongst the nodes within the currently-understood model. Also, another protein of interest that seems to be greatly involved in the pathway, is the PTK2 protein, which is also not found in the KEGG osmotic stress pathway represented in Figure 4. PTK2 seems particularly interesting because it is a protein that has an interaction with many other proteins in the network, which indicates a crucial role in the pathway (Figure 3).

There are many aspects of this network that need to be analyzed further. Further examination of the interactions between some of these nodes is necessary because the interactions, in some cases, directly contradict those found in the literature. Also, it seems there are significant nodes that are highly connected that are also not a part of this well-understood pathway. Further additions to the network could also be made with the addition of undirected connections between proteins, and the alteration of the data by adding varying weights, instead of applying the same edge weight for all connections in the network. It is also necessary to investigate why there are proteins not shown in the final network produced by TPS that are present in the KEGG pathway. These proteins could be excluded due to not being present in the data, being removed due to the temporal phosphorylation data, or not having significant phosphorylation changes in the cell.

Overall, the results indicate that there is much more to be discovered about the osmotic stress pathway in yeast cells. Eight of the proteins in the network were identified in the literature for this pathway, but the majority of the proteins in the interaction network are novel to this pathway. As Figure 5 demonstrates, starting from 3543 proteins in the network, and reducing this large set of interactions down to just 55 nodes, with some of these nodes being present in the literature is a promising result.

References

Cherry J.M., Hong E.L., Amundsen C., Balakrishnan R., Binkley G., Chan E.T., Christie K.R., Costanzo M.C., Dwight S.S., Engel S.R., Fisk D.G., Hirschman J.E., Hitz B.C., Karra K., Krieger C.J., Miyasato S.R., Nash R.S., Park J., Skrzypek M.S., Simison M., Weng S., Wong E.D. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res. Jan;40(Database issue):D700-5 (2012).

C. Choudhary, M. Mann, Decoding signaling networks by mass spectrometry-based proteomics, Nat. Rev. Mol. Cell Biol. 11, 427–439 (2010).

Hohmann S, Mager P (eds) Yeast Stress Responses. pp 389. Springer-Verlag: Heidelberg (2003).

J. Kyriakis. In the Beginning, There Was Protein Phsophorylation. J Biol Chem. 289(14): 9460–9462 (2014).

Koskal, Ali Sinan et al. “Synthesizing Signaling Pathways from Temporal Phosphoproteomic Data.” Unpublished.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res: 40, D109-D114 (2012).

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.

Tuncbag, Nurcan, et al. "Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package." *PLoS Comput Biol* 12.4 (2016): e1004879.