

University of Nevada, Reno

**Developing bioinformatic tools for phosphorylation site co-regulation
and correlation: The plant Cellulose Synthase Complex as a case study**

A thesis submitted in partial fulfillment
of the requirements for the degree of

Bachelor of Science in Biochemistry and Molecular Biology and the Honors Program

by

Joseph J, Thomas

Dr. Ian S. Wallace, Thesis Advisor

May 2017

**UNIVERSITY
OF NEVADA
RENO**

THE HONORS PROGRAM

We recommend that the thesis
prepared under our supervision by

Joseph J. Thomas

entitled

**Developing bioinformatic tools for phosphorylation site co-regulation
and correlation: The plant Cellulose Synthase Complex as a case study**

be accepted in partial fulfillment of the
requirements for the degree of

Bachelor of Science in Biochemistry and Molecular Biology

Dr. Ian S. Wallace, Ph.D., Thesis Advisor

Tamara Valentine, Ph.D., Director, Honors Program

May 2017

Abstract

Cellulose is the most abundant biopolymer on the planet, and this paracrystalline polysaccharide has a variety of industrial applications ranging from textiles to biofuel production. Regulation and polymerization of cellulose in plant cell walls are complex pathways consisting of many proteins with phosphorylated residues discovered through phosphoproteomic surveys. As a ubiquitous form of post-translational modification, phosphorylation by protein kinases is a crucial form of both discrete protein and whole metabolic or signaling pathway regulation. Despite the importance of protein phosphorylation, the relationships between these modifications are poorly understood at the proteomic scale. Through this work, a variety of phosphorylation site network visualization formats were developed to directly visualize the number, frequency, and sequence conservation of phosphorylation sites within a given protein or network of proteins. Further comparison of the primary sequence flanking the phosphorylated residue will help identify motifs conserved across proteins within the network, indicative of protein kinases acting on multiple targets within the pathway and playing an important regulatory role in quickly modulating different steps in the pathway in response to extra- or intra-cellular signals. Overall, this work may lead to streamlined workflows to connect tens of thousands of experimentally supported phosphorylation sites to thousands of protein kinases.

Table of Contents

Abstract.....	i
Table of Contents.....	ii
List of Tables.....	iii
List of Figures.....	iv
Introduction.....	1
Methods.....	4
Results.....	8
Discussion.....	19
References.....	24

List of Tables

Table 1. Protein kinase prediction for yeast glycolysis.....	18
--	----

List of Figures

Fig. 1. General workflow of network construction.....	11
Fig. 2. Visualization of occurrence count and residue conservation.....	11
Fig. 3. Primary cell wall biosynthesis network with residue conservation incorporated.....	12
Fig. 4. Plot of occurrence count vs. residue conservation for primary cell wall CSC and associated proteins.....	13
Fig. 5. CesA3 internal peptide comparison.....	14
Fig. 6. Primary cell wall cellulose biosynthesis interprotein sequence comparisons....	15
Fig. 7. Overall yeast glycolysis network.....	16
Fig. 8. Plot of occurrence count vs. residue conservation for yeast glycolytic enzymes.....	16
Fig. 9. Yeast glycolysis interprotein sequence comparison.....	17

Introduction

Cellulose is the most abundant component of plant cell walls, and this paracrystalline polysaccharide acts as the chief load bearing component opposing the turgor-driven cell growth that underlies plant cell expansion. Cellulose in the cell wall also plays a crucial role in influencing overall cell shape (Jones et al., 2016). Cellulose is composed of multiple (18-24) β -1,4-linked glucan chains that are organized into a paracrystalline microfibril ((Newman et al., 2013; Thomas et al., 2013). Cellulose polymerization is catalyzed by a large plasma membrane-localized transmembrane protein complexes known as cellulose synthase complexes (CSCs) (Kimura et al., 1999). Imaging of the CSC via transmission electron microscopy has shown it to be a hexamer of proposed trimers arranged in a six-fold symmetrical structure known as a “rosette”. (Nixon et al., 2016). Additionally, live-cell imaging of fluorescently labeled CSCs indicate that these complexes functionally associate with microtubules *in vivo* and that CSCs move at the plasma membrane with a velocity of approximately 250 nm/ min (Paredes et al., 2006). At least three isoforms of the 10-member Cellulose Synthase A family (CESAs) comprise functional CSCs in Arabidopsis (Taylor et al., 2003; Persson et al., 2007; Desprez et al., 2007). The specific isoforms present in the CSC vary depending on whether the complex produces cellulose in the primary or secondary cell wall. Primary cell wall biosynthesis requires CesA1, CesA3, and at least one of the partially redundant CesA2., CesA5, CesA6, or CesA9 isoforms (Desprez et al., 2007; Persson et al. 2007). Conversely, cellulose for the secondary cell wall is produced by CesA4, CesA7, and CesA8 (Turner and Somerville, 1997). Primary cell walls are present in almost all plant cells, whereas secondary cell walls provide extra rigidity to mature, non-dividing cells. In

addition to these core catalytic components, a host of other proteins have been experimentally demonstrated to play roles in the regulation of biosynthesis. These proteins include the β -1,4-endoglucanase KORRIGAN1 (KOR1), the Cellulose Synthase Interactive 1 protein family (CSI1) shown to associate the CSC with underlying cortical microtubules, the Companion of Cellulose synthase proteins (CC) aiding with microtubule stability under salt-stress, and the dynamin-related protein 1A (DRP1A) implicated with cell plate formation (Mansoori et al., 2014; Li et al., 2012; Endler et al., 2015; Collings et al., 2007). While the precise function of some of these proteins in the overall biosynthetic machinery are still being fully elucidated, proteomic data indicates that many of these proteins are phosphorylated, suggesting that post-translational phosphorylation may be a critical regulatory strategy in cellulose biosynthesis (Nühse et al., 2004; Taylor et al., 2007; Facette et al., 2013; Nakagami et al., 2010).

Protein phosphorylation on serine, threonine, and tyrosine by protein kinases is one of the most widely prevalent post-translational modifications for regulating protein activity in eukaryotes (Manning et al., 2002; Huber, 2007). Addition of phosphate groups to these residues can affect conformation, activity, subcellular localization, or alter protein-protein interactions with non-substrate proteins (Newman et al., 2013; Ross et al., 2013). In the *Arabidopsis thaliana* genome, approximately 1000 protein kinases have been identified, a handful of which have been implicated in cellulose biosynthesis through genetic analyses (Arabidopsis Genome Institute, 2000; Hématy et al., 2007; Hématy and Höfte, 2008; Xu et al., 2008)). Protein kinases recognize short, specific stretches of amino acids flanking the residue to be phosphorylated in the primary sequence; therefore, identification of substrate specificity motifs prevalent throughout a

metabolic pathway hints at kinases with sweeping regulatory effects within that process (Kemp et al., 1975; Ren et al., 2008). Investigation of the phosphorylated residues within a peptide chain is crucial in developing a complete model of a protein's overall behavior, regulation, and impact in cellular processes. In terms of metabolic pathway regulation, the rate at which phosphorylation and other post-translational modifications occur compared to transcriptional changes offers an explanation to how a cell can rapidly and adeptly respond to changes in their internal and external environments (Tripodi et al., 2015). A comprehensive understanding of the kinases acting on proteins within a metabolic or signaling pathway is necessary to fully grasp how it is modulated in accordance to various stimuli.

Numerous phosphoproteomic surveys have been performed in a wide range of organisms, and more recent applications of novel mass spectrometry techniques have yielded thousands of *in vivo* experimentally supported phosphorylation sites (Mann et al., 2003). Visualization and global analysis methods for phosphorylation networks is lacking for higher eukaryotic organisms. Conducted studies focus primarily on kinase interactions, and network-based analysis of residue characteristics remains an unexplored bioinformatics territory (Fiedler et al., 2009; Arodz et al., 2015). The construction of networks summarizing information from previous studies can be useful in elucidating phosphorylated proteins or phosphorylation sites of interest that can then be further studied in focused experiments. In this thesis, methods for the construction of metabolic networks detailing various phosphoproteomic data are explored as well as the development of new analytical techniques for the identification of wide-acting protein kinases that can be explored in future experiments.

Methods

Phosphorylation site data collection:

The proteins and reactions comprising metabolic pathways of interest were identified using AraCyc (<http://pmn.plantcyc.org/organism-summary?object=ARA>) on the Plant Metabolic Network database. Using AraCyc, enzymes for each step in a pathway (both experimentally and computationally determined) were identified along with their Arabidopsis Information Resource (TAIR) accession numbers. AraCyc gave only *Arabidopsis thaliana* proteins. The phosphoproteomic database PhosPhAt (<http://phosphat.uni-hohenheim.de/>) was then used to identify all experimentally supported phosphorylation sites within each target protein. Individual phosphorylation residues were identified using PhosPhAt, and the number of mass spectra indicating each experimentally supported residue was observed was counted as an occurrence. Different peptides containing the same phosphorylation site were not counted as a separate residue. For the *Saccharomyces cerevisiae* glycolysis network, Jessica Hernandez gathered all the sequence and proteomic data from the Saccharomyces Genome Database (<http://www.yeastgenome.org/>).

Residue conservation:

Using the accession numbers provided by AraCyc, each primary amino acid sequence for each protein was used as a BLASTp query. The top ten results from different plant species were exported as a .txt file, which was then aligned using ClustalX program (<http://www.clustal.org/clustal2/>). The species used include *Arabidopsis thaliana*,

Camelina sativa, *Medicago truncatula*, *Vitis vinifera*, *Oryza sativa*, *Solanum tuberosum*, *Solanum lycopersicum*, *Zea mays*, *Lotus japonicus*, *Triticum aestivum*, and *Cucumis sativus*. Conservation scores were calculated as a percentage of the number of times the residue in *A. thaliana* was conserved in the other nine sequences.

Glycolytic enzymes in yeast were put through a similar BLAST screen. Since yeast is not a plant, the first nine non-*S. cerevisiae* were selected to calculate the conservation score. In this case, conservation scores were calculated based on the residue in *S. cerevisiae* compared to the residues in the same position among the other species after the sequence alignment.

Sequence analysis:

Using the program BioEdit (<http://www.mbio.ncsu.edu/bioedit/bioedit.html>), alignment files created in ClustalX (<http://www.clustal.org/clustal2/>) were cut down to a short sequence consisting of a phosphorylated residue and 14 flanking amino acids (7 on each side of the residue). The online resource WebLogo (<http://weblogo.berkeley.edu/>) took these shortened, aligned sequence files and created an image depicting which amino acids were present at each position in the sequence. The more times a specific amino acid was present at a given position, the larger its graphical representation on the logo.

Construction of network:

Proteomic data (protein ID, phosphorylation site IDs, and frequency of phosphorylation) was coded into a Microsoft Excel spreadsheet in a three-column format. The .xlsx file was used as the framework for constructing the network in the program Cytoscape

(<http://www.cytoscape.org/>). Each column was assigned either as a source node, target node, or edge variable. Duplicate edges leading to the same node (i.e. same phosphorylated residue among different protein) were removed and replaced with separate residue nodes for each protein. Occurrences were entered as a long integer variable, which could then be stylistically modified by creating a continuous mapping for all edges that made a color gradient based off the minimum and maximum values. Conservation was then added to the network's table as another long integer variable that was mapped as residue node background color.

Sequence Comparison

Using the previously produced logo images, an average sequence was established for each phosphorylation residue and its flanking primary sequence. The most heavily represented amino acid at a given position was used to make this average sequence. A spreadsheet formula was developed to compare each of these 15-amino acid-long peptides with the other peptides from the same protein on a position-by-position basis. Two metrics were used to create individual positions scores. First, a conservation score of either zero (0) or one (1) was assigned to each residue position. If one of the peptides had no amino acid present in that position while the other did, the conservation score was set to 0. If the positions matched in the presence, or lack thereof, of any amino acid, the conservation score was set to 1. Next, the properties of the amino acid were compared with a property score. If the same amino acid was conserved, the property score was set to 1. If there were two amino acids at the position and they both had similar chemical properties (nonpolar, polar, acidic, or basic), a score of 0.5 was given. Lastly, completely

dissimilar amino acids scored 0. The position score was calculated by multiplying the conservation score by the property score. The total score for a single comparison was calculated by the summing the individual position scores. This is represented in Equation 1 where s is the total similarity score, c_i is the conservation score at the i^{th} position, r_i is the residue score at the i^{th} position, and n is the length of the peptides being compared.

Equation 1

$$s = \sum_{i=1}^n c_i \times r_i$$

Kinase Search

Kinases most likely to target a given pair of peptides were found using the community resource atlas known as NetPhorest developed by Miller et al. (2008) and Horn et al. (2004) (<http://netphorest.info/index.shtml>). The comparison data had outlier scores that were much higher than the median, typically between isoformic residues with near-identical flanking amino acids. Determination of “significant comparisons” was established to be those scoring greater than 6 at the least, though it was modified to 7 for the yeast network considering the noise produced from the multitude of 6-scoring comparisons. Only non-outlier, significant comparisons were paired with a kinase. For the yeast network, this meant looking at comparisons with a score between 7 and 8. After the algorithm output the most likely kinases to phosphorylate the residues in either protein of the comparison, the lists were parsed for kinases that appeared on both. From there, kinase(s) with the most similar total score were selected (kinases with similar probabilities of phosphorylating each residue).

To better understand the work flow for this project, Figure 1 below was developed, showing the website or program each piece of information was collected from and is funneled into.

Results

The Arabidopsis CSC represents an attractive model system to develop protein kinase correlation networks because this multiprotein complex consists of known interacting proteins and because numerous phosphorylation sites within the CSC have been identified. To begin to develop these techniques, we first developed visualization methods to display networks with multiple phosphorylation sites. First, a Cytoscape network was made for the proteins involved in the synthesis of cellulose for the Arabidopsis primary cell wall cellulose synthesis machinery and incorporated the phosphorylation site's number of occurrences in proteomic studies as well as conservation of these phosphorylation sites among *Arabidopsis* and nine other plant species as discussed in Materials and Methods (Figure 2 and Figure 3). Figure 2 provides the basic schema used in the rest of the study for representing this data for an individual protein: central nodes (boxes) represent proteins with their phosphorylated residues encircling them, connected by edges (lines) with some data mapped to it (in this case, occurrence count). Figure 4 compares the conservation and occurrence of each residue in Figure 3, showing a weak positive relationship between occurrence count and conservation score. Although higher conservation is not direct indicator of high occurrence count, residues with higher occurrence counts can be predicted to be better conserved.

Intraprotein sequence comparisons of phosphorylation sites were carried out first chiefly because of the complexity of interprotein comparisons; the lower number of combinations for intraprotein comparisons made it more conducive to developing more robust phosphorylation site conservation algorithms. Figure 5a shows the final product with CesA3 as an example. Although the key for the CesA3 comparisons goes from 1 to 5 only, the maximum possible score possible is 15 for two identical sequences, indicating that different kinases are acting on each residue and that CesA3's activity is modulated by many protein kinases. Figure 5b shows the sequence logos for S176 and S226, the residues with the highest comparison in Figure 5a. The comparison between the two scored a 5 which is fairly low, and the tenuous similarity between the two is seen in the logos.

When all the interprotein comparisons for the primary cell wall network were included, the quantity of low-similarity scores greatly impeded the visualization of significant sequence comparisons. Two of the highest scoring edges were between protein isoforms and were excluded by making them transparent and white. The edge between CesA6 S11 and KOR1 S37 scored a 12, but the upper limit for the color gradient was set to the next highest value of 7.5 to prevent skewing of the data set. The lower limit to be opaque was set to 6, reducing the noise from 5 and 5.5 scored edges that represented ~30% comparable sequences. Contrast among the remaining edges was increased by limiting the range of the green-yellow-red to the same as that of the transparency mapping (6 to 7.5). Figures 6b, c, and d display the sequence logos generated for the three most similar pairs of residues in the network. As expected from their comparison score, these logos have much more crossover with each other than those

in Figure 5b. Although the sequence is not the exact same, amino acids with similar chemical properties appear in the same position. Interestingly, of the 18 colored edges in the resulting image (Figure 6), only three connect to residues with 10% conservation. More tightly conserved residues are more likely to have sequence similarity with other residues in the network, suggesting certain sequence motifs have been preserved over evolutionary time with important implications for the overall metabolic pathway.

To test the potential predictive properties these visualization techniques, networks for glycolysis in *Saccharomyces cerevisiae* were constructed. This network is comprised of: hexokinase (HXK1/2), phosphoglucose isomerase (PGI1), phosphofructokinase (PFK1/2), aldolase (FBA1), triose phosphate isomerase (TPI1), GAP dehydrogenase (GAPDH1/2/3), phosphoglycerate kinase (PGK1), phosphoglycerate mutase (GPM1), enolase (ENO1/2), and pyruvate kinase (PYK1/2). The base network (Figure 7) contained residue occurrence and conservation information in a similar manner to Figure 3a. The plot of occurrence count vs. conservation score (Figure 8) showed the same trends as seen in Figure 4. Once interprotein comparison information was added, removal of edges between residues of isoforms and using the same minimum value as in the cellulose biosynthesis network, 6, produced the visual seen in Figure 9. The highest scoring residue pairs from Figure 9 along with a random sampling of the lower scoring pairs were used in the NetPhorest kinase prediction. A summary of the prediction is seen in Table 1, which has been sorted and colored to group residue pairs with the same predicted kinase.

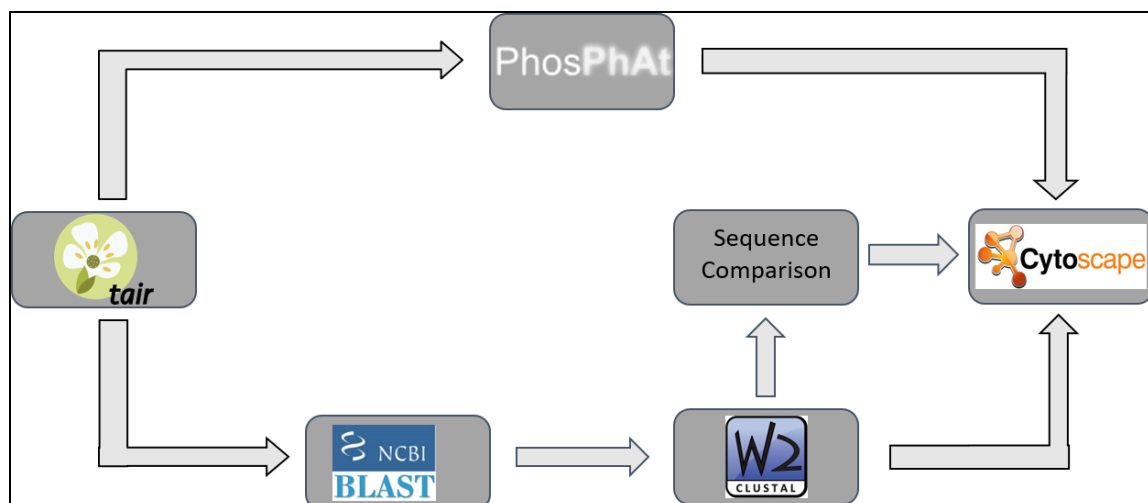


Figure 1. General workflow of network construction. All sequence information is obtained through TAIR and funneled through either PhosPhAt or BLAST and ClustalX to obtain the final data that is input into the Cytoscape network.

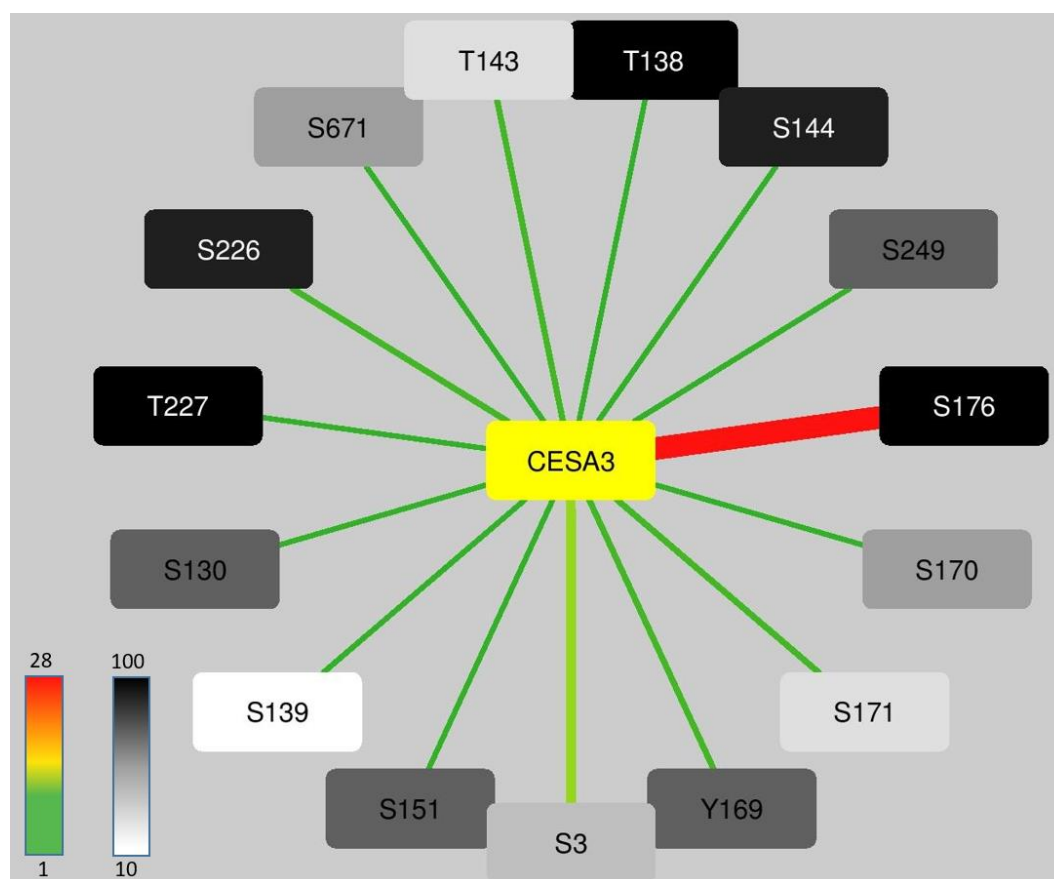


Figure 2. Visualization of occurrence count and residue conservation. Central nodes are proteins with known phosphorylated residues radiating outward. Occurrence count was mapped to edge color per the key, and edge thickness was increased proportional to

the occurrence count. Conservation was mapped to node color on a white-black gradient as seen in the key.

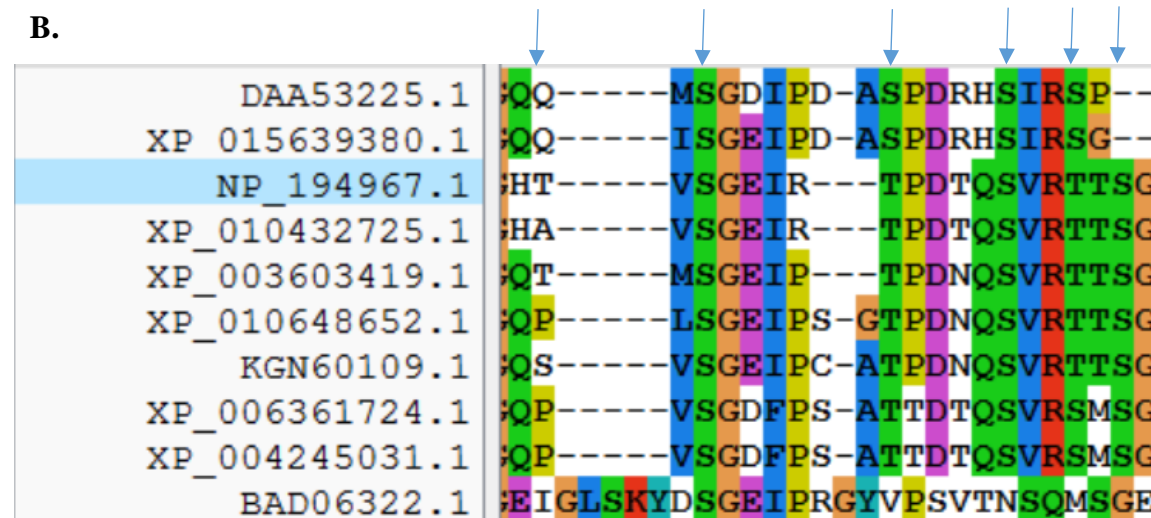
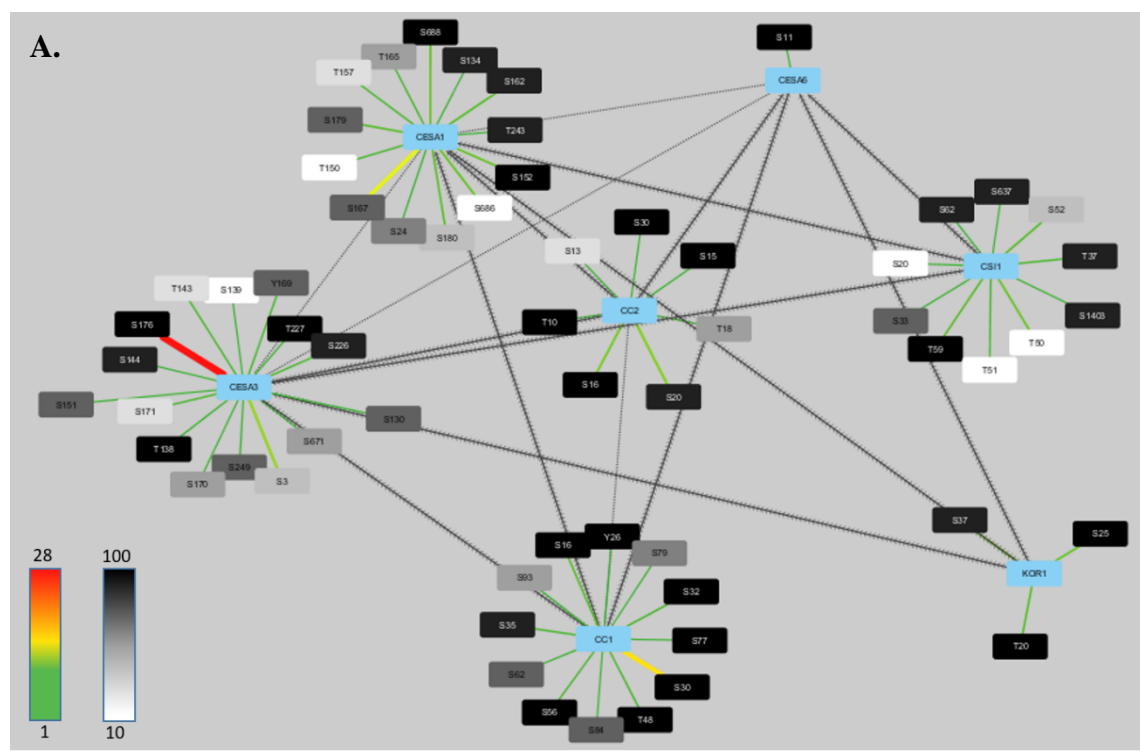


Figure 3. Primary cell wall biosynthesis network with residue conservation incorporated. (a) Like Figure 2, central nodes are proteins with known phosphorylated extending outward and occurrence count and conservation mapped to edge color and node color, respectively, per the keys. Edge thickness was also increased as the occurrence count increased. Black edges indicate protein interactions (serrated lines) or protein isoforms. (b) Sample sequence alignment for Cesa1 with the *Arabidopsis*

sequence highlighted in light blue. Arrows designate the phosphorylation residues in the segment conservation was calculated for (T150, S152, T157, S162, T165, and S167).

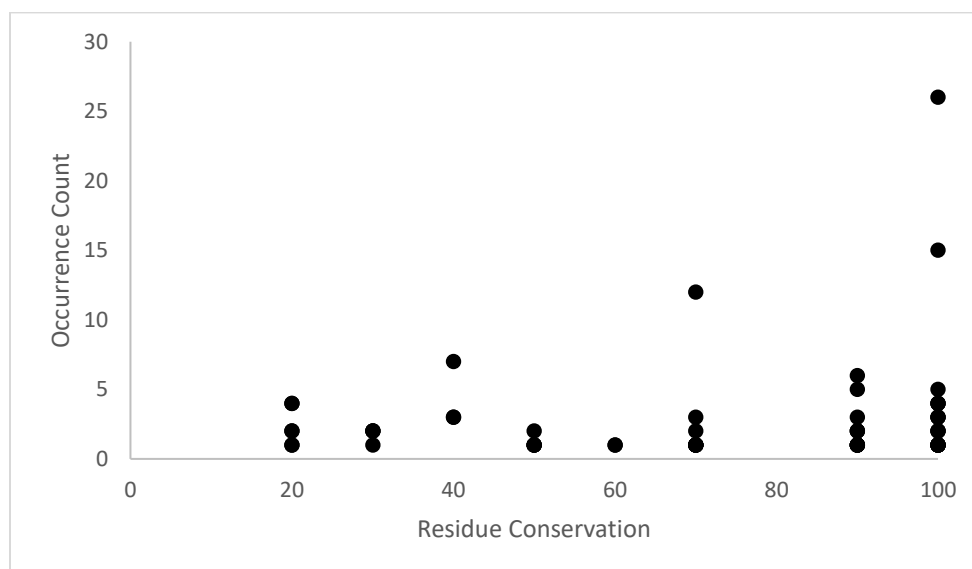
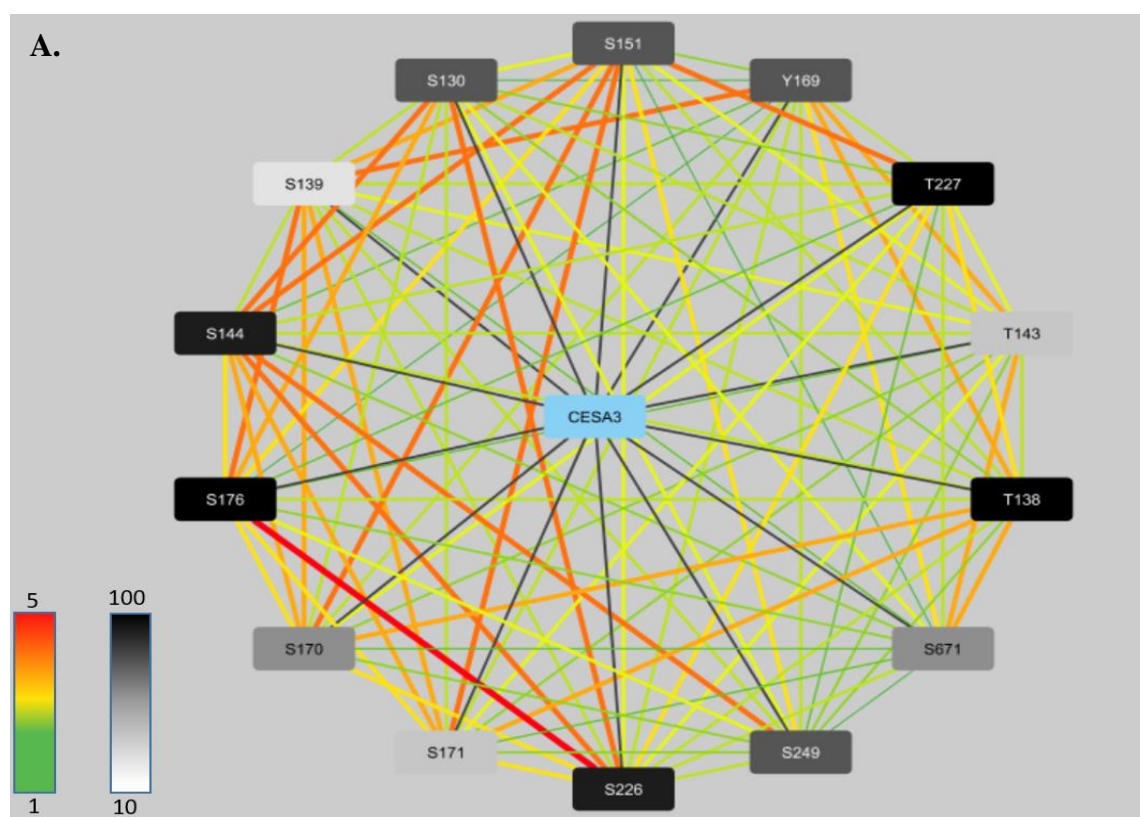


Figure 4. Plot of occurrence count vs. residue conservation for primary cell wall CSC and associated proteins. Values were taken from the network seen in Figure 3a.



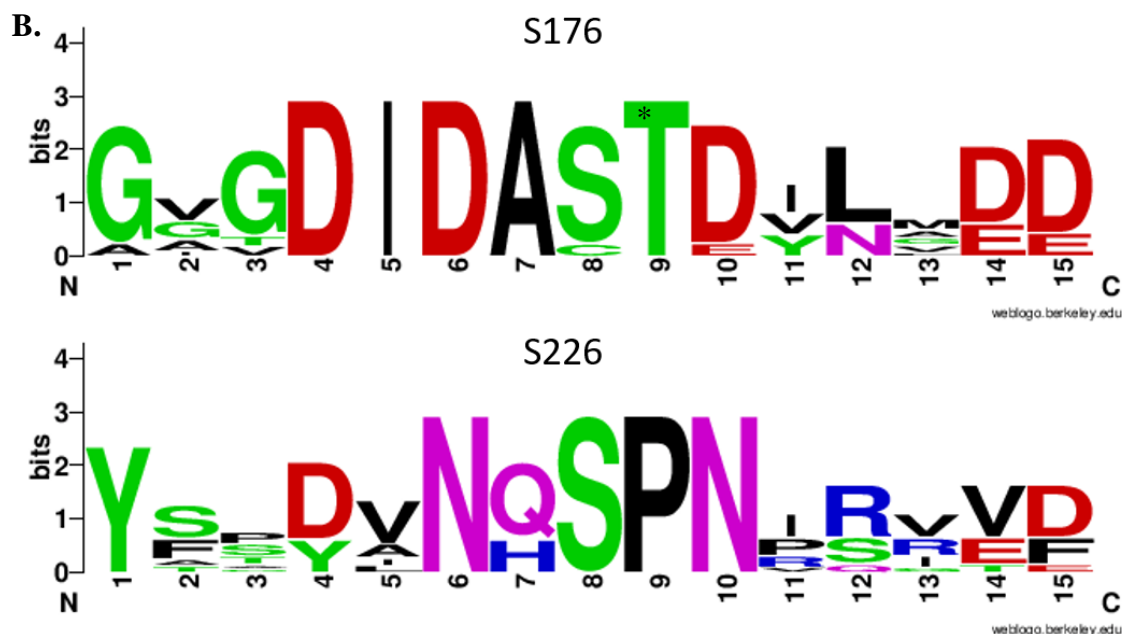


Figure 5. CesA3 internal peptide comparison. (a) Comparison totals were mapped to edge color according per the green-yellow-red gradated key. Thickness was mapped as well with higher similarity correlating with thicker edge lines. Previously calculated conservation scores for each phosphorylation residue was also includes as white-black gradated node color per the key. (b) Sequence logos for S176 and S226. The size of the letter(s) at each position is proportional to the residue's presence in the aligned sequences. Asterisks denote the phosphorylated residue.

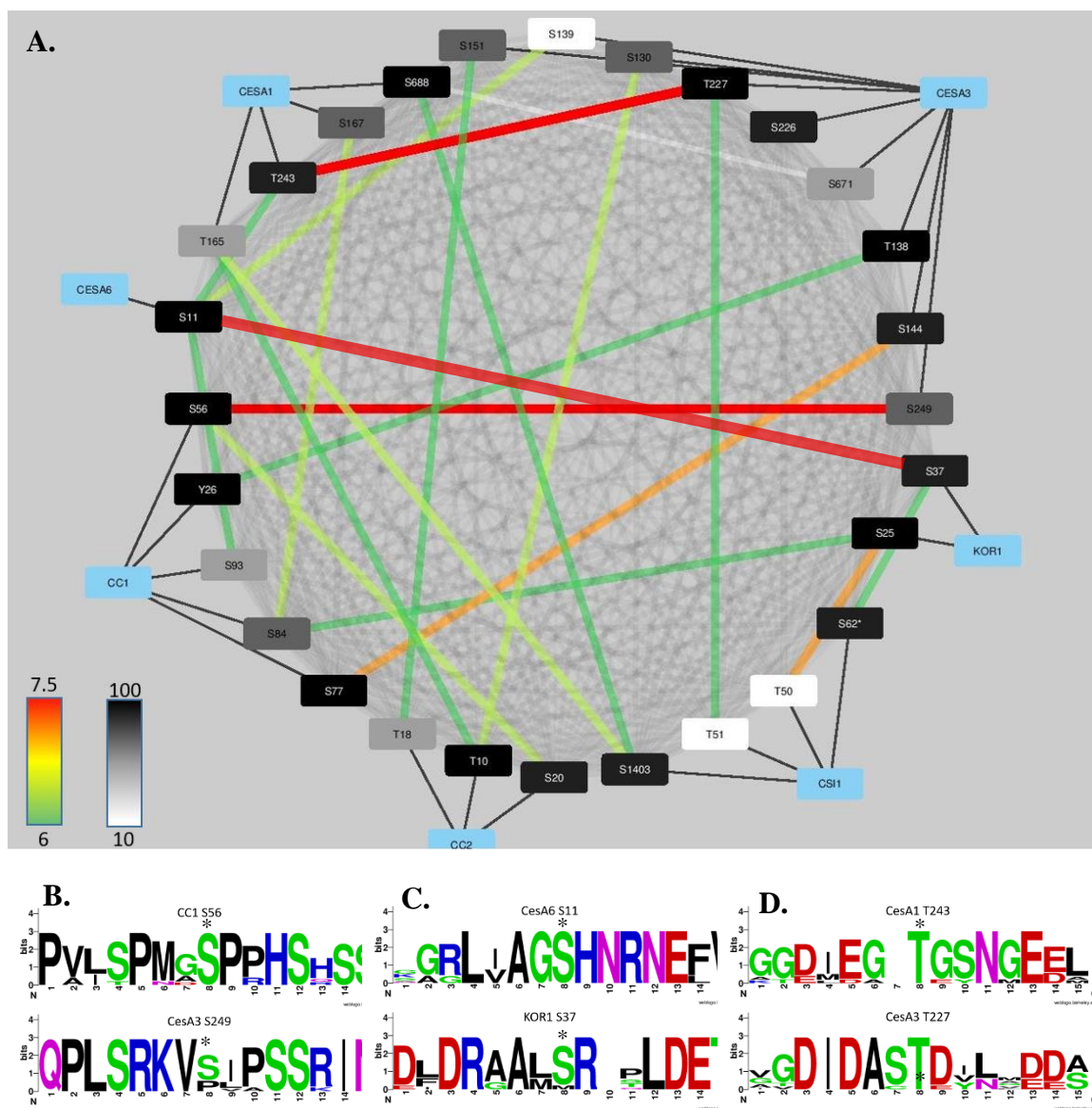


Figure 6. Primary cell wall cellulose biosynthesis interprotein sequence comparisons. (a) Comparison totals were mapped to edge color per the green-yellow-red gradated key as well as edge thickness. Previously calculated conservation scores for each phosphorylation residue were also included as white-black gradated node color per the key. Faint white lines represent high sequence similarity between residues on isoforms. (b) (c) and (d) Show sequence logos for the three highest scoring comparisons with the size of the letter at each position proportional to the residue's presence among the aligned sequences. Asterisks denote the phosphorylated residue.

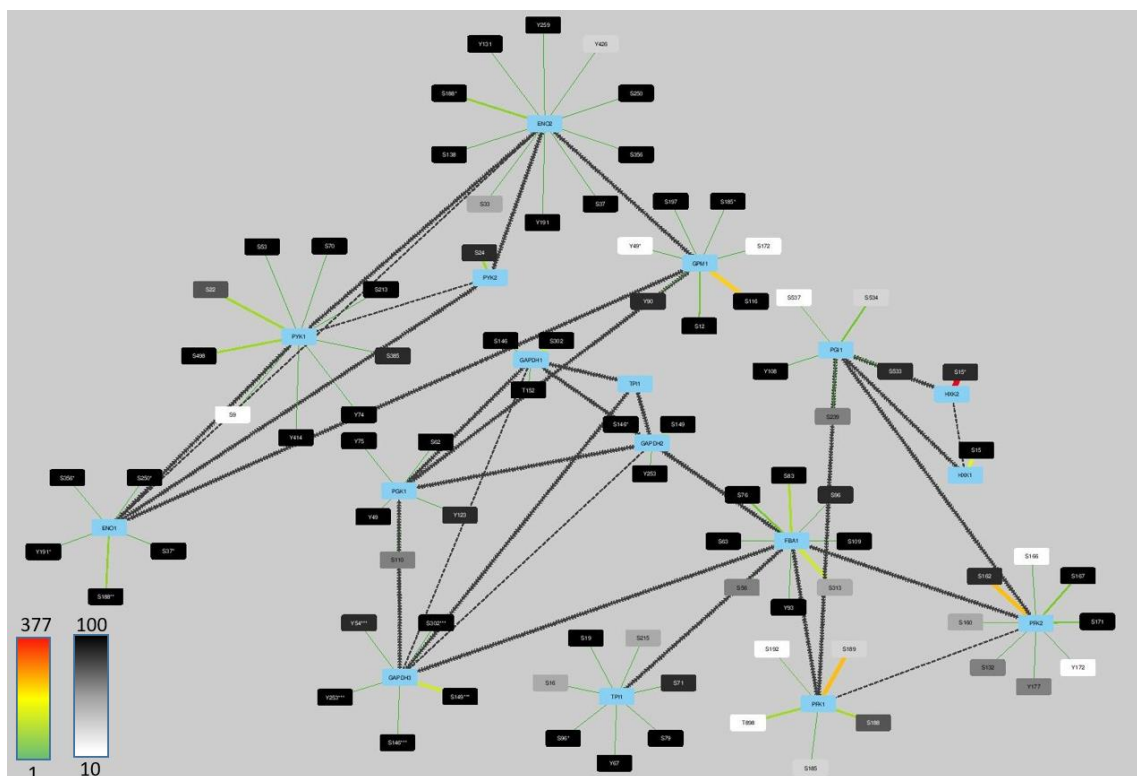


Figure 7. Overall yeast glycolysis network. Central nodes are proteins with known phosphorylated residues radiating outward. Occurrence count was mapped to edge color per the key, and edge thickness was positively correlated with occurrence count. Black edges indicate protein interactions (serrated lines) or protein isoforms. Conservation was mapped to node color on a white-black gradient as seen in the key.

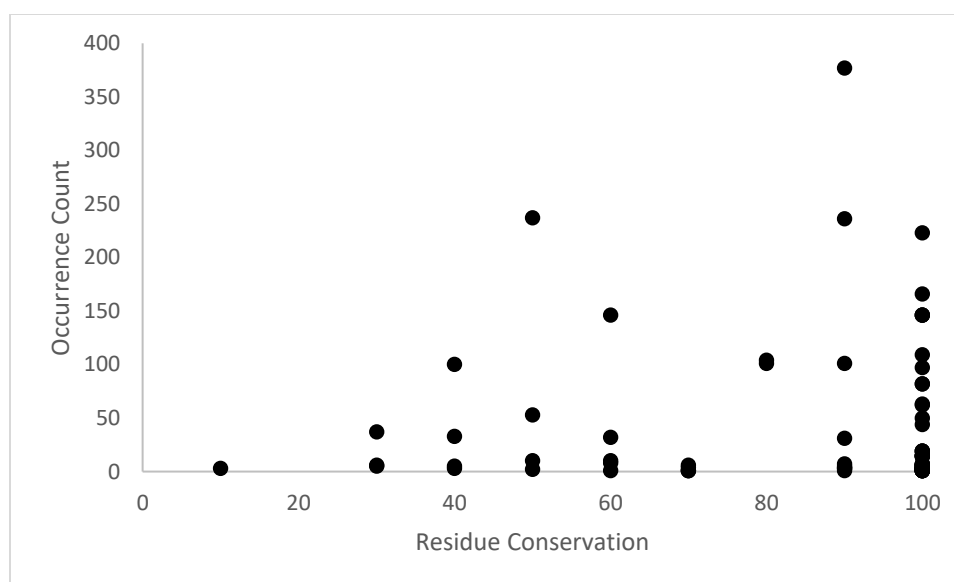


Figure 8. Plot of occurrence count vs. residue conservation for yeast glycolytic enzymes. Values were taken from the network seen in Figure 7.

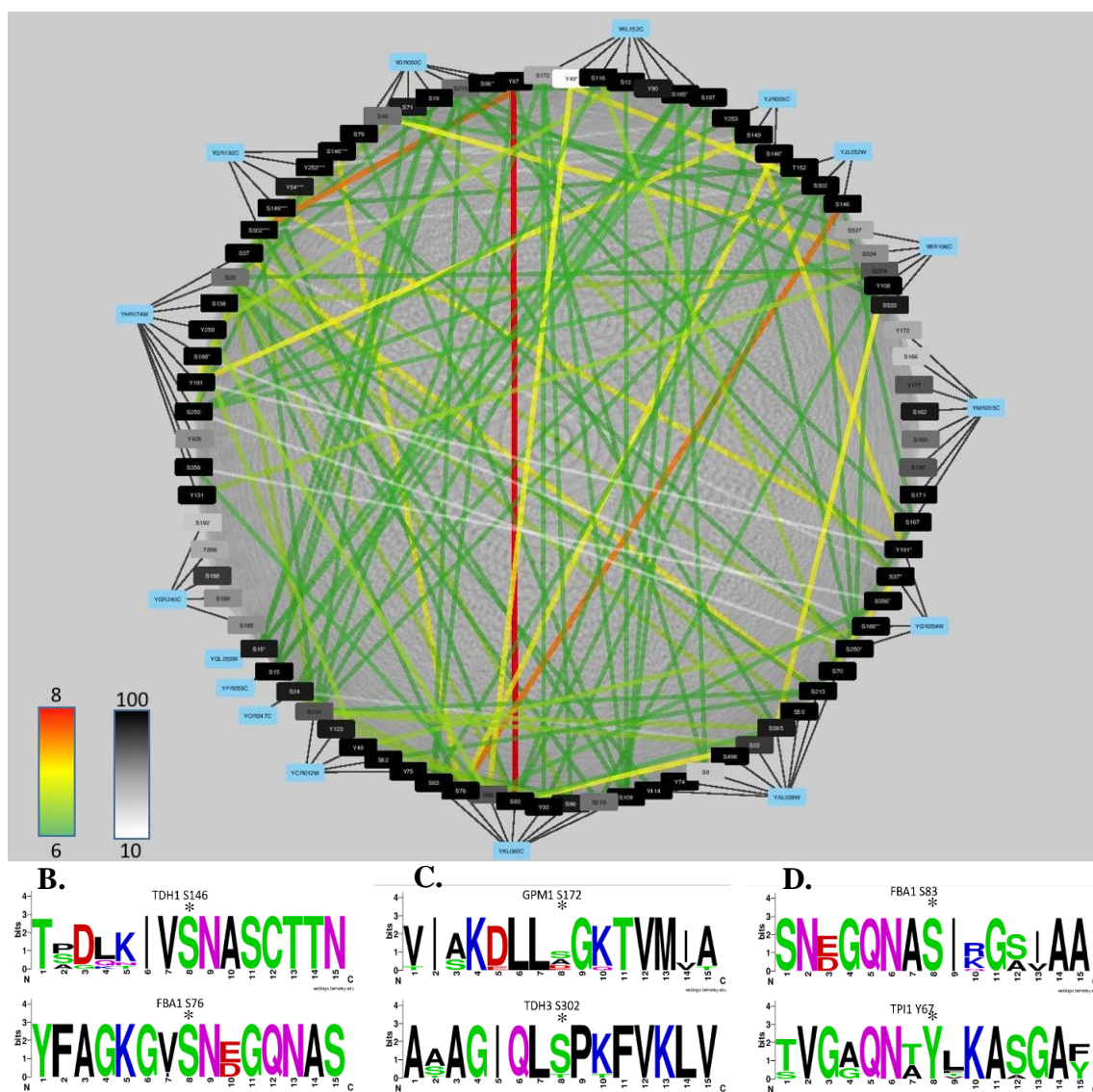


Figure 9. Yeast glycolysis interprotein sequence comparison. (a) Comparison totals were mapped to edge color according to the green-yellow-red gradated key as well as edge thickness. Previously calculated conservation scores for each phosphorylation residue was also includes as white-black gradated node color per the key. Faint white lines represent high sequence similarity between residues on isoforms. (b) (c) and (d) Show sequence logos for the three highest scoring comparisons with the size of the letter at each position proportional to the residue's presence among the aligned sequences.

Table 1. Protein kinase prediction for yeast glycolysis. Sample of residue pairs connected by visible edges from Figure 9. The score column represents the probability NetPhorest assigned to the kinase potential involvement in phosphorylation based on sequence alone. “No match” indicates NetPhorest had no predications for the indicated residue or for both. “No similarity” indicates that there were no kinases on that appeared on both residue’s prediction list.

Residue #1	Residue #2	Kinase Group	Score (#1, #2)	Comparison
FBA1 S76	TDH2 S146	YMR216C	0.04, 0.04	7
TDH1 S146	FBA1 S76	YMR216C	0.05, 0.04	6.5
TDH1 S146	PFK2 S167	YMR216C	0.05, 0.04	6
TDH1 T152	TDH2 S149	YMR216C	0.05, 0.09	6
TDH1 T152	PYK2 S24	YMR216C	0.05, 0.14	6
TDH1 S302	PGK1 S110	YMR216C	0.04, 0.05	6
TDH1 S302	ENO1 S250	YMR216C	0.04, 0.06	6
PGI1 S534	TPI1 S16	KIN1,2 group	0.12, 0.17	7
PGI1 S534	ENO2 S250	KIN1,2 group	0.12, 0.6	6
PGI1 S534	ENO1 S250	KIN1,2 group	0.12, 0.6	6
GAPDH3 S146	FBA1 S76	PTK1,2 group	0.04, 0.03	7
PGI1 S537	TPI1 S19	PTK1,2 group	0.09, 0.10	6
TDH1 S302	FBA1 S83	PTK1,2 group	0.03, 0.04	6
GAPDH3 S302	GPM1 S173	SCH9 PKC1 group	0.04, 0.04	7.5
TDH1 S302	GPM1 S173	SCH9 PKC1 group	0.04, 0.04	7
TDH1 S302	ENO2 S250	SCH9 PKC1 group	0.04, 0.05	6
PGI1 S239	ENO2 S138	YDR283C	0.15, 0.11	6.5
PGI1 S239	FBA1 S313	YDR283C	0.15, 0.08	6
TDH1 T152	GAPDH3 S149	YDR283C	0.05, 0.07	6
ENO1 S37	PYK1 S213	STE20 CLA4 SKM1 group	0.03, 0.06	7
ENO2 S37	PYK1 S213	YCL024W	0.04, 0.04	7
FBA1 S83	PYK1 S22	YDL028C	0.09, 0.15	7
TDH1 T152	FBA1 S63	YFL033C	0.14, 0.10	6
PGI1 S239	ENO2 S33	YAR018C	0.10, 0.08	6
PGI1 S533	TPI1 S215	KIN82 FPK1 group	0.13, 0.05	6
TDH1 S302	TPI1 S16	PHO85 CDC28 group	0.08, 0.06	6
FBA1 S83	TPI1 Y67	(no similarity)	n/a	8
TPI1 Y67	GAPDH3 S302	(no match)	n/a	7.5
ENO1 Y191	TDH2 S149	(no match for #1)	n/a	7

ENO1 Y191	GAPDH3 S149	(no match for #1)	n/a	7
GAPDH3 S149	ENO1 Y191	(no match for #2)	n/a	7
FBA1 S56	GPM1 Y49	(no match)	n/a	7
ENO2 Y191	TDH2 S149	(no match)	n/a	7
PGI1 Y108	PYK1 S385	(no similarity)	n/a	7
PGI1 S239	ENO2 Y131	(no similarity)	n/a	6.5

Discussion

At the start of the project, residue conservation was expected to have some level of relation with other variables encoded in the network. As seen in Figures 4 and 8, there is a slight positive correlation between a residue's occurrence count and conservation score. Although the occurrence count cannot be predicted solely based on the conservation score, the opposite is true to an extent. The higher the occurrence count for a given amino acid residue, the more likely it is to have a high conservation score. Residues scoring highly in both parameters may be important to the protein's overall behavior and should be explored in future experiments. There is another relationship between sequence and residue conservation that is seen well in Figure 6a. For the most part, high scoring edges connect nodes with the same or similar conservation scores. This may be indicative of either a sequence motif that has only more recently evolved or one that is widely seen through the phylogeny of the selected plant species. The former could be a non-crucial regulatory or functional phosphorylation site, while the latter would be a site necessary for the protein to function properly.

For the interprotein comparisons seen in Figure 9, only 18 of the total 378 comparison edges fell in the range of 6 – 7.5 that excluded isoform and insignificant

comparisons. There does not appear to be any residue that has high sequence similarity with multiple other residues of different proteins that in turn connect to each other. Conserved sequence motifs present throughout a whole network of proteins are expected to behave in a transitive fashion: if sequence A is like sequence B, and B is like sequence C, then C should be like A, even if the “likeness” between each varies. Given the visualization parameters used to create Figures 6a and 9a, this relationship would create something that looks like a subnetwork, a small group of interconnected nodes within the overall network. However, this does not mean the pairs of highly similar residues picked out from the network are not important. Similarity between residues in two proteins could still represent a protein kinase acting on two separate proteins in the network. Protein kinases phosphorylating two separate phosphorylation sites in different proteins within a pathway would still play an important role in modulating the pathway in response to different stimuli. Interestingly, there are chains of sequence similarity within the network. For example, there is one chain between Cesa3 S249, CC1 S56, and CC2 S20. The scores of each individual edge composing the chain varies. These could hint at the evolution of the primary structure around the residue; as more mutations accumulate, the oldest and newest residues would gradually become more and more dissimilar, leading to the linear chain.

The protein kinase search for the yeast glycolysis residues yielded results for the majority of the random samples selected, as can be seen in Table 1. For every residue input, the NetPhorest algorithm produces a list of the protein kinases it predicts to catalyze the phosphorylation. The results are each paired with a probability factor determined by sequence comparison of the query and known substrates of the kinase

(Miller et al., 2008). Selection of the most likely kinase to act on two given residues typically bypassed the most probable for either as determined by the algorithm. For the sample of residue pairs queried, a number of them returned the same hit as highlighted in Table 1. The KIN1/KIN2 group are a pair of homologs implicated with exocytosis while the PTK1/PTK2 group are a pair involved in ion transport (Erez et al., 2002; Elbert et al., 2005). SCH9 and PKC1 play roles in controlling cell growth (Levin et al., 1990; Fabrizio et al., 2001). YMR216C and YDR283C have been shown to be involved in mRNA metabolism and translation (Siebel et al., 1999; Kubota et al., 2001). While none of these protein have been demonstrated to directly interact with the proteins in the glycolytic pathway, three of them have been tied to other cell functions related to glycolysis. Fabrizio et al. (2001) demonstrated that SCH9 acts in pathways mediating glucose-dependent signaling, glycolysis, and growth. Goossens et al. (2000) implicated PTK2 with the control of an H⁺-ATPase in response to glucose metabolism. Lastly, YMR216C has been connected with oxidative stress in yeast, a factor influenced by glycolysis (Leiro et al., 2012). Interestingly, higher comparison scores as seen in the Cytoscape network did not serve as an indication that a kinase could be identified for a residue pair, as seen by the results for FBA1 S83 and TPI1 Y67 with a comparison score of 8 yet no predicted protein kinase.

For the future directions of this project, the first step that needs undertaking is the automation of the process of gathering the proteomic information, collecting and trimming the sequences, and calculating sequence comparison scores. The first two points are easy to do manually, but automation of the steps would be a boon for the process, particularly for protein networks with either many constituent proteins or a

plethora of phosphorylation sites on each. Coding a program to gather proteomic information, collect sequences, and trim would also be a simple task to carry out. The sequence comparison is by and large the most time-consuming part of the whole process; for n phosphorylation sites, there are $\frac{n!}{2 \times (n-2)!}$ combinations that need to be carried out.

Over the course of this project, the steps have been automated as much as possible to find comparisons, but the code for each step needs to be worked on so that running the comparisons is not as taxing in terms of computer processing as it is currently.

The other, and more important, step as the project proceeds is to run kinase activity assays between the proteins within the yeast glycolysis network and its predicted protein kinases. For these assays, the protein needs to be in its native conformation; tertiary and quaternary structure can play pivotal roles in the interaction between the protein and the kinase that are lost if the assay is done with small peptides. If the activity assay reveals the kinase acts on the protein of interest, the identity of the phosphorylated residue needs to be confirmed through mass spectrometry analysis of the substrate protein. Hypothetically, the predicted kinases in Table 1 should show activity in the initial assay and be inactive in the second. Following these assays, knockout mutants of the protein kinase can be obtained to determine the effects their absence has on the metabolic pathway through phenotypic changes and metabolite concentration comparisons.

This method of data visualization and phosphoresidue analysis is excellent for the condensation of a variety of data, but its ability to demonstrate co-regulation and correlation among phosphorylated residues needs to be tested further and improved. As it is, the algorithm bluntly compares amino acids in a position. Similarity between amino

acids is determined solely based on chemical properties of the R-group (e.g. acidic, basic, nonpolar, and polar). A side chain's physical features, which may not necessarily overlap with the chemical property-based groupings, are not considered. For example, β -branched amino acids such as Val, Ile, and Thr or long-chain amino acids like Met and Lys receive r scores of 0 or 0.5 despite their physical similarities. The algorithm needs to be refined to take this into account and go beyond a trinary scoring system that lumps certain residues together to the exclusion of other measures of similarity. Despite the lack of patterns consistent throughout the networks, they do suggest that there are protein kinases that potentially work on at least two proteins. Following experiments testing the efficacy of the predicted kinases in phosphorylating the glycolytic proteins they have been implicated with will determine if the technique has predictive capacity.

References

- Arabidopsis Genome Institute. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, *408*(6814), 796-815. doi:10.1038/35048692
- Arodz, T., & Bonchev, D. (2015). Identifying influential nodes in a wound healing-related network of biological processes using mean first-passage time. *New Journal of Physics*, *17*, 10. doi:10.1088/1367-2630/17/2/025002
- Brown, R. M. (1987). The biosynthesis of cellulose. *Food Hydrocolloids*, *1*(5), 345-351. doi:10.1016/S0268-005X(87)80024-3
- Cheng, S.-H., Willmann, M. R., Chen, H.-C., & Sheen, J. (2002). Calcium Signaling through Protein Kinases. The Arabidopsis Calcium-Dependent Protein Kinase Gene Family. *Plant Physiology*, *129*(2), 469-485. doi:10.1104/pp.005645
- Collings, D. A., Gebbie, L. K., Howles, P. A., Hurley, U. A., Birch, R. J., Cork, A. H., Hocart, C. H., Arioli, T., & Williamson, R. E. (2007). Arabidopsis dynamin-like protein DRP1A: a null mutant with widespread defects in endocytosis, cellulose synthesis, cytokinesis, and cell expansion. *Journal of Experimental Botany*, *59*(2), 361-376. doi:10.1093/jxb/erm324
- Desprez, T., Juraniec, M., Crowell, E. F., Jouy, H., Pochylova, Z., Parcy, F., Hofte, H., Gonneau, M., & Vernhettes, S. (2007). Organization of cellulose synthase complexes involved in primary cell wall synthesis in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(39), 15572-15577. doi:10.1073/pnas.0706569104
- Elbert, M., Rossi, G., & Brennwald, P. (2005). The yeast Par-1 homologs Kin1 and Kin2 show genetic and physical interactions with components of the exocytic machinery. *Molecular Biology of the Cell*, *16*(2), 532-549. doi:10.1091/mbc.E04-07-0549
- Endler, A., Kesten, C., Schneider, R., Zhang, Y., Ivakov, A., Froehlich, A., Funke, N., & Persson, S. (2015). A Mechanism for Sustained Cellulose Synthesis during Salt Stress. *Cell*, *162*(6), 1353-1364. doi:10.1016/j.cell.2015.08.028
- Erez, O., & Kahana, C. (2002). Deletions of SKY1 or PTK2 in the *Saccharomyces cerevisiae* *trk1Δtrk2Δ* mutant cells exert dual effect on ion homeostasis. *Biochemical and Biophysical Research Communications*, *295*(5), 1142-1149. doi:10.1016/S0006-291X(02)00823-9

- Fabrizio, P., Pozza, F., Pletcher, S. D., Gendron, C. M., & Longo, V. D. (2001). Regulation of Longevity and Stress Resistance by Sch9 in Yeast. *Science*, 292(5515), 288-290. doi:10.1126/science.1059497
- Facette, M. R., Shen, Z., Björnsdóttir, F. R., Briggs, S. P., & Smith, L. G. (2013). Parallel Proteomic and Phosphoproteomic Analyses of Successive Stages of Maize Leaf Development. *The Plant Cell*, 25(8), 2798-2812. doi:10.1105/tpc.113.112227
- Fiedler, D., Braberg, H., Mehta, M., Chechik, G., Cagney, G., Mukherjee, P., Silva, A. C., Shales, M., Collins, S. R., van Wageningen, S., Kemmeren, P., Holstege, F. C. P., Weissman, J. S., Keogh, M., Koller, D., Shokat, K. M., & Krogan, N. J. (2009). Functional Organization of the *S-cerevisiae* Phosphorylation Network. *Cell*, 136(5), 952-963. doi:10.1016/j.cell.2008.12.039
- Goossens, A., Natalia de la, F., Forment, J., Serrano, R., & Portillo, F. (2000). Regulation of Yeast H⁺-ATPase by Protein Kinases Belonging to a Family Dedicated to Activation of Plasma Membrane Transporters. *Molecular and Cellular Biology*, 20(20), 7654-7661. doi:10.1128/MCB.20.20.7654-7661.2000
- Horn, H., Schoof, E. M., Kim, J., Robin, X., Miller, M. L., Diella, F., Palma, A., Cesareni, G., Jensen, L. J., & Linding, R. (2014). KinomeXplorer: an integrated platform for kinome biology studies. *Nature Methods*, 11(6), 603-604. doi:10.1038/nmeth.2968
- Huber, S. C. (2007). Exploring the role of protein phosphorylation in plants: from signalling to metabolism. *Biochemical Society Transactions*, 35(Pt 1), 28-32.
- Hématy, K., Sado, P.-E., Van Tuinen, A., Rochange, S., Desnos, T., Balzergue, S., Pelletier, S., Renou, J.P., & Höfte, H. (2007). A Receptor-like Kinase Mediates the Response of Arabidopsis Cells to the Inhibition of Cellulose Synthesis. *Current Biology*, 17(11), 922-931. doi:10.1016/j.cub.2007.05.018
- Hématy, K., & Höfte, H. (2008). Novel receptor kinases involved in growth regulation. *Current Opinion in Plant Biology*, 11(3), 321-328. doi:10.1016/j.pbi.2008.02.008
- Jones, D. M., Murray, C. M., Ketelaar, K. J., Thomas, J. J., Villalobos, J. A., & Wallace, I. S. (2016). The Emerging Role of Protein Phosphorylation as a Critical Regulatory Mechanism Controlling Cellulose Biosynthesis. *Frontiers in Plant Science*, 7, 684. doi:10.3389/fpls.2016.00684
- Kemp, B. E., Bylund, D. B., Huang, T.-S., & Krebs, E. G. (1975). Substrate Specificity of the Cyclic AMP-Dependent Protein Kinase. *Proceedings of the National Academy of Sciences of the United States of America*, 72(9), 3448-3452. doi:10.1073/pnas.72.9.3448

- Kimura, S., Laosinchai, W., Itoh, T., Cui, X., Linder, C. R., & Brown, R. M., Jr. (1999). Immunogold Labeling of Rosette Terminal Cellulose-Synthesizing Complexes in the Vascular Plant *Vigna angularis*. *The Plant Cell*, *11*(11), 2075-2086. doi:10.1105/tpc.11.11.2075
- Kubota, H., Ota, K., Sakaki, Y., & Ito, T. (2001). Budding yeast GCN1 binds the GI domain to activate the eIF2 α kinase GCN2. *The Journal of Biological Chemistry*, *276*(20), 17591-17596.
- Leiro, A. G., Lombardero, S. R., Vazquez, A. V., Siso, M. I. G., and Cerdan, M.E. 2012. The yeasts genes ROX1, IXR1, SKY1 and their effect upon enzymatic activities related to oxidative stress. *Oxidative Stress*. Edited by V.I. Lushchak and H. Semchyshyn. InTech, Rijeka, Croatia. In press. ISBN 979-953-307-574-6.
- Levin, D. E., Fields, F. O., Kunisawa, R., Bishop, J. M., & Thorner, J. (1990). A candidate protein kinase C gene, PKC1, is required for the *S. cerevisiae* cell cycle. *Cell*, *62*(2), 213-224. doi:10.1016/0092-8674(90)90360-Q
- Li, S., Lei, L., Somerville, C. R., & Gu, Y. (2012). Cellulose synthase interactive protein 1 (CS11) links microtubules and cellulose synthase complexes. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(1), 185-190. doi:10.1073/pnas.1118560109
- Linding, R., Jensen, L. J., Ostheimer, G. J., van Vugt, M. A. T. M., Jorgensen, C., Miron, I. M., Diella, F., Colwill, K., Taylor, L., Elder, K., Metalnifov, P., Nguyen, V., Pasculescu, A., Jin, J., Park, J. G., Samson, L. D., Woodgett, J. R., Russell, R. B., Bork, P., Yaffe, M. B., & Pawson, T. (2007). Systematic discovery of in vivo phosphorylation networks. *Cell*, *129*(7), 1415-1426. doi:10.1016/j.cell.2007.05.052
- Mann, M., & Aebersold, R. (2003). Mass spectrometry-based proteomics. *Nature*, *422*(6928), 198-207. doi:10.1038/nature01511
- Manning, G., Plowman, G. D., Hunter, T., & Sudarsanam, S. (2002). Evolution of protein kinase signaling from yeast to man. In (Vol. 27, pp. 514-520). LONDON: Elsevier Ltd.
- Mansoori, N., Timmers, J., Desprez, T., Claire, L. A. K., Dianka, C. T. D., Vincken, J.-P., Richard G. F. V., Höfte, H., Vernhettes, S., & Trindade, L. M. (2014). KORRIGAN1 Interacts Specifically with Integral Components of the Cellulose Synthase Machinery: e112387. *PLoS One*, *9*(11). doi:10.1371/journal.pone.0112387
- Miller, M. L., Jensen, L. J., Diella, F., Jørgensen, C., Tinti, M., Li, L., Hsiung, H., Parker, S. A., Bordeaux, J., Sicheritz-Ponten, T., Olhovsky, M., Pasculescu, A.,

- Alexander, J., Knapp, S., Blom, N., Bork, P., Li, S., Cesareni, G., Pawson, T., Turk, B. E., Yaffe, M. B., Brunak, S., & Linding, R. (2008). Linear motif atlas for phosphorylation-dependent signaling. *Science Signaling*, 1(35), ra2-ra2. doi:10.1126/scisignal.1159433
- Mueller, S. C., & Brown, R. M. (1980). Evidence for an Intramembrane Component Associated with a Cellulose Microfibril-Synthesizing Complex in Higher Plants. *The Journal of Cell Biology*, 84(2), 315-326. doi:10.1083/jcb.84.2.315
- Newman, R. H., Hu, J., Rho, H. S., Xie, Z., Woodard, C., Neiswinger, J., Cooper, C., Shirley, M., Clark, H. M., Hu, S., Hwang, W., Jeong, J. S., Wu, G., Lin, J., Gao, X., Ni, Q., Goel, R., Xia, S., Ji, H., Dalby, K. N., Birnbaum, M. J., Cole, P. A., Knapp, S., Ryazanov, A. G., Jack, D. J., Blackshaw, S., Pawson, T., Gingras, A., Desiderio, S., Pandey, A., Turk, D. E., Zhang, J., Zhu, H., & Qian, J. (2013). Construction of human activity-based phosphorylation networks. *Molecular Systems Biology*, 9(1), 655-n/a. doi:10.1038/msb.2013.12
- Nixon, B. T., Mansouri, K., Singh, A., Du, J., Davis, J. K., Lee, J. G., Slabaugh, E., Vandavasi, V. G., O'Neill, H., Roberts, E. M., Roberts, A. W., Yingling Y. G., & Haigler, C. H. (2016). Comparative Structural and Computational Analysis Supports Eighteen Cellulose Synthases in the Plant Cellulose Synthesis Complex. *Scientific Reports*, 6, 28696. doi:10.1038/srep28696
- Nühse, T. S., Stensballe, A., Jensen, O. N., & Peck, S. C. (2004). Phosphoproteomics of the Arabidopsis Plasma Membrane and a New Phosphorylation Site Database. *The Plant Cell*, 16(9), 2394-2405. doi:10.1105/tpc.104.023150
- Paredez, A. R., Somerville, C. R., & Ehrhardt, D. W. (2006). Visualization of Cellulose Synthase Demonstrates Functional Association with Microtubules. *Science*, 312(5779), 1491-1495. doi:10.1126/science.1126551
- Persson, S., Paredez, A., Carroll, A., Palsdottir, H., Doblin, M., Poindexter, P., Khitrov, N., Auer, M., & Somerville, C. R. (2007). Genetic evidence for three unique components in primary cell-wall cellulose synthase complexes in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, 104(39), 15566-15571. doi:10.1073/pnas.0706592104
- Ren, S., Uversky, V. N., Chen, Z., Dunker, A. K., & Obradovic, Z. (2008). Short Linear Motifs recognized by SH2, SH3 and Ser/Thr Kinase domains are conserved in disordered protein regions. *BMC Genomics*, 9 Suppl 2(Suppl 2), S26-S26. doi:10.1186/1471-2164-9-S2-S26
- Ross, K. E., Arighi, C. N., Ren, J., Huang, H., & Wu, C. H. (2013). Construction of protein phosphorylation networks by data mining, text mining and ontology

integration: analysis of the spindle checkpoint. *Database: the journal of biological databases and curation Journal Article*, 2013, bat038.

- Siebel, C. W., Feng, L., Guthrie, C., & Fu, X.-D. (1999). Conservation in Budding Yeast of a Kinase Specific for SR Splicing Factors. *Proceedings of the National Academy of Sciences of the United States of America*, 96(10), 5440-5445. doi:10.1073/pnas.96.10.5440
- Taylor, N. G., Howells, R. M., Huttly, A. K., Vickers, K., & Turner, S. R. (2003). Interactions among Three Distinct Cesa Proteins Essential for Cellulose Synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 100(3), 1450-1455. doi:10.1073/pnas.0337628100
- Taylor, N. G. (2007). Identification of cellulose synthase AtCesA7 (IRX3) in vivo phosphorylation sites—a potential role in regulating protein degradation. *Plant Molecular Biology*, 64(1), 161-171. doi:10.1007/s11103-007-9142-2
- Thomas, L. H., Forsyth, V. T., Šturcová, A., Kennedy, C. J., May, R. P., Altaner, C. M., Apperley, D. C., Wess, T. J., & Jarvis, M. C. (2013). Structure of Cellulose Microfibrils in Primary Cell Walls from Collenchyma. *Plant Physiology*, 161(1), 465-476. doi:10.1104/pp.112.206359
- Tripodi, F., Nicastro, R., Reghellin, V., & Coccetti, P. (2015). Post-translational modifications on yeast carbon metabolism: Regulatory mechanisms beyond transcriptional control. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1850(4), 620-627. doi:10.1016/j.bbagen.2014.12.010
- Turner, S. R., & Somerville, C. R. (1997). Collapsed Xylem Phenotype of Arabidopsis Identifies Mutants Deficient in Cellulose Deposition in the Secondary Cell Wall. *The Plant Cell*, 9(5), 689-701. doi:10.1105/tpc.9.5.689
- Xu, S.-L., Rahman, A., Baskin, T. I., & Kieber, J. J. (2008). Two Leucine-Rich Repeat Receptor Kinases Mediate Signaling, Linking Cell Wall Biosynthesis and ACC Synthase in Arabidopsis. *The Plant Cell*, 20(11), 3065-3079. doi:10.1105/tpc.108.063354