Multi-omics analysis of metabolic genes perturbations in *Saccharomyces cerevisiae*

Yee Jher Chan, Yu-Pin Liang, Iddo Frieberg, Claus Kadelka, and Julie Dickerson

Iowa State University

Abstract

Understanding the effect of metabolic genes has been challenging due to the many confounding interactions between the metabolites and other biomolecules. In this study, we investigated the transcriptome and proteome of yeast auxotroph (single and combinatorial deletion of LEU2, MET15, HIS3, URA3) using co-expression network. The constructed network was able to identify clusters of proteins that are overexpressed in LEU2 and MET15 auxotrophs. The GO functionality analysis further details the molecular function and biological processes of those proteins, such as the increasing ribosomal activities in the LEU2 auxotroph. Using the sparse Partial Least Square method, features from both the transcriptomic and proteomic datasets were extracted and filtered to create a new network consisting of 90 genes and 13 proteins. This generated network was able to cluster genes and proteins between the auxotrophs. Therefore, the co-expression network showed promising results for identifying relationships between different groups with many features and can be used for understanding biological systems.

1. Introduction

Metabolic reactions within a cell are essential for growth and maintaining healthy cellular physiology. However, elucidating the genetic-metabolic interactions has been complicated by the highly connected network of the metabolites. Any perturbation in most of the metabolic genes will therefore result in system-wide consequences. One method to circumvent this confounding involvement of many other biomolecules is perturbing the metabolic pathways in the amino-acid and nucleobase biosynthesis, where the cells prefer to uptake metabolites over self-synthesis. The single-gene auxotrophs in those metabolic pathways have been used as markers for genetic studies. Alam et al. specifically investigated the single and combinatorial effects of HIS3, LEU2, URA3, and MET15 deletions on the transcriptome, proteome, and metabolome [1]. They found that up to 85% of gene expression was affected by metabolic background differences.

The investigation involving the transcriptome, proteome, and metabolome is only possible with the recent advancement of high-throughput technologies such as microarrays and RNA sequencing. However, it is still challenging to infer the large datasets and identify useful information. Dimensionality reduction methods can be used to quickly perceive the similarity between different groups. Principal components analysis (PCA) was frequently used to analyze multi-omics datasets and construct the clustering in each different condition using a linear dimensionality reduction technique [2]. PCA projects the variance from the original dataset to each principles component by eigenvectors and transforms the data to preserve the variance. T-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and

Projection (UMAP) are another two non-linear dimension reduction algorithm that worked well with biological datasets [3].

Another widely used method to infer large datasets is the co-expression network [4]. Co-expression networks have been predominantly employed to identify regulatory genes but can also be utilized for proteome and metabolome datasets. Canonical co-expression network first identifies relationships between genes based on correlation measures or mutual information, and then constructs a network where the nodes represent the genes, and the edges represent the similarity measures. Several network construction methods have been established, such as Genie3 [5], ARACNE [6], TIGRESS [7], etc. These methods are applicable to datasets with many features and low sample count. The constructed networks can then be used to identify regulators, hub genes, perform differential co-expression analyses, etc.

The response variables can also often be collected from multiple sources, such as RNA-seq and LC-MS [8]. For such datasets, PLS has recently obtained attention due to its ability to combine different data types. PLS method identifies linear combinations between the variables of different datasets to reduce dimensionality [9]. Unlike PCA, PLS maximizes the covariance between the latent variables, which has shown improved performance compared to PCA. Sparse Partial Least Squares (sPLS) is a variant of PLS, where it includes LASSO penalization to select only a few number of features from the datasets [10].

In this work, we attempted to use networks to identify key genes or protein clusters that characterize between different yeast strains of different metabolic backgrounds. The transcriptomic and proteomic datasets from [1] were employed; within the datasets,

expressions for one prototroph strain and fifteen other strains created from single or combinatorial deletions of HIS3 (H), LEU2 (L), URA3 (U), and MET15 (M) were included with three replicates each, resulting in a total of 48 samples. Dimensionality reduction techniques, specifically PCA, T-SNE, and UMAP, were first used to visualize the "closeness" between the 16 different strains. Genie3 was then used to construct the network, which is a feature selection method that combines with tree ensemble method. Next, differential expression analysis was performed on the datasets and visualized using the constructed network. The gene ontology (GO) terms were also analyzed for the upregulated and downregulated transcriptomes and proteomes to identify their functional relationships. Finally, sPLS method was used to integrate the two datasets and extract features among both transcriptome and proteome that best characterize different strains.

2. Methods

Transcriptomic and proteomic datasets

The transcriptomic and proteomic datasets are obtained from [1]. For the transcriptome dataset, any genes with expression value of 0 in all samples are removed, resulting in 5983 total genes.

Dimension reduction methods

All the dimensional reduction processes were performed in R. For PCA, *factoextra* and *FactoMineR* packages are used. For t-SNE, *Rtsne* package was used with a perplexity of 3. For UMAP, the default parameters were used under *M3C* library.

Network construction and visualization

GENIE3 was utilized from [11] and applied on each of the transcriptomic and proteomic datasets to generate a graph. For the transcriptomic dataset, 5983 genes and 48 samples (16 strains) were fed as input; For the proteomic dataset, 446 proteins and 48 samples were fed as input. All hyperparameters were kept at default except the number of attributes was set to "all" and number of threads to "1000". Interaction scores below 0.05 were removed. The generated graph was then imported into Cytoscape for visualization. Only the largest subnetwork from each graph was kept, leaving 1639 genes and 439 proteins.

Differential expression analysis

The *DESeq2* package was utilized in R for the differential expression analysis. This method normalizes the expression values based on the median of gene expression in a sample. The computed $\log_2$(fold change) was obtained and filtered for values above 0.2 (upregulated) and below -0.2 (downregulated).

GO functionality analysis

The upregulated or downregulated genes and proteins are imported as list to the STRING database online.

Datasets integration

For multiple datasets integration, the *mixOmics* package in R was used. "tune.spls" method was used to determine the number of variables to keep in each component. The results obtained was [45, 45] for "keepY" and [10, 3] for "keepX". Once the result was obtained from calling the

"spls" method, the "network" method was used to generate a correlation matrix (cutoff = 0.7) from the result. This correlation matrix was then imported into Cytoscape for visualization.

3. Results

3.1 Dimensionality Reduction

Principle Component Analysis (PCA), T-distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP) were being performed in our study in order to compare and investigate the similarity between different auxotrophs. PCA provided a generally well visualization for both datasets. PC1 provided 58.8% variance, and PC2 provided 10% variance in the transcriptomic dataset (Figure 1). PC1 provided 33.2% of the variance in the proteomic dataset and PC2 22% of the variance (Figure 2). The strain HU has the best clustering outcome, followed by HLUM and HLM, and L was relatively separated from U, M, and H in transcriptomic data (Figure 3). The strain H and HL had the best clustering outcome, followed by HLU, HLUM, and HLM, and L was relatively separated from U, M, and H in proteomics data (Figure 4). All clustering from proteomics data were comparably divided into three groups from the result of PCA. The outcome from the t-SNE in the transcriptomic dataset was significant. Most of the strains clustered in a superb manner (Figure 5). The t-SNE also yielded the same clustering result as PCA did in the proteomics data, with clustering comparably separated into three main groups (Figure 6). However, UMAP has the most minor clustering performance in both two transcriptomic and proteomic data, and it generally separated all the conditions into two main groups (Figure 7, 8). In short, L auxotroph was observed to be the most separated from prototroph compared to other single auxotrophs.

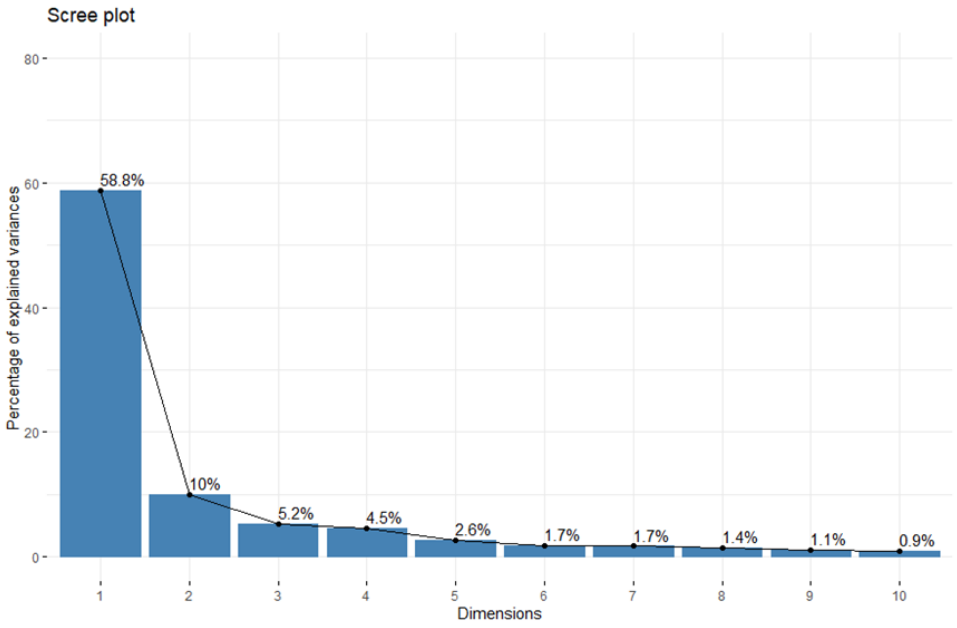

Figure 1. Scree plot from the result of the PCA in transcriptomic dataset.

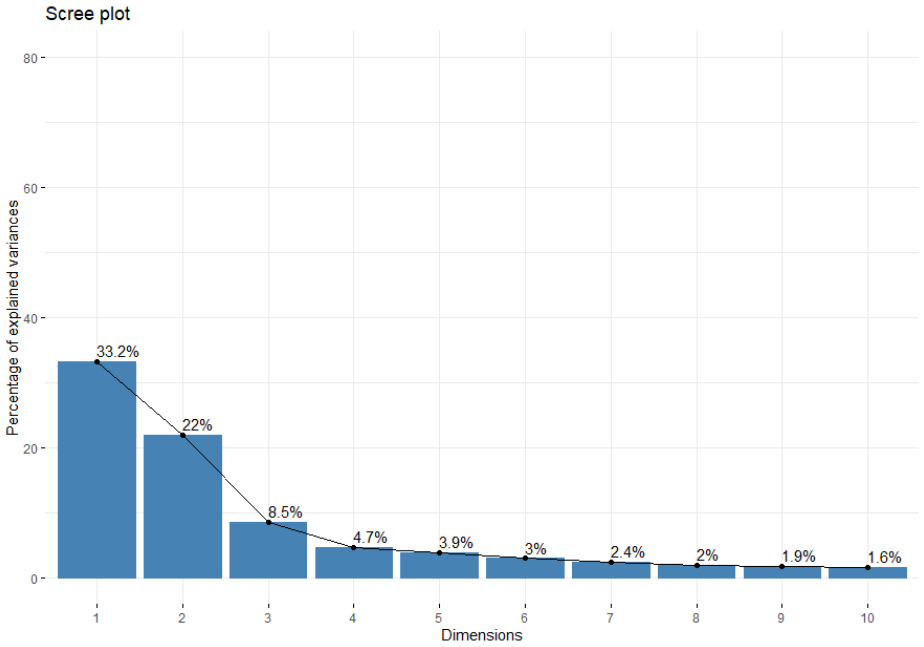

Figure 2. Scree plot from the result of PCA in Proteomics dataset.

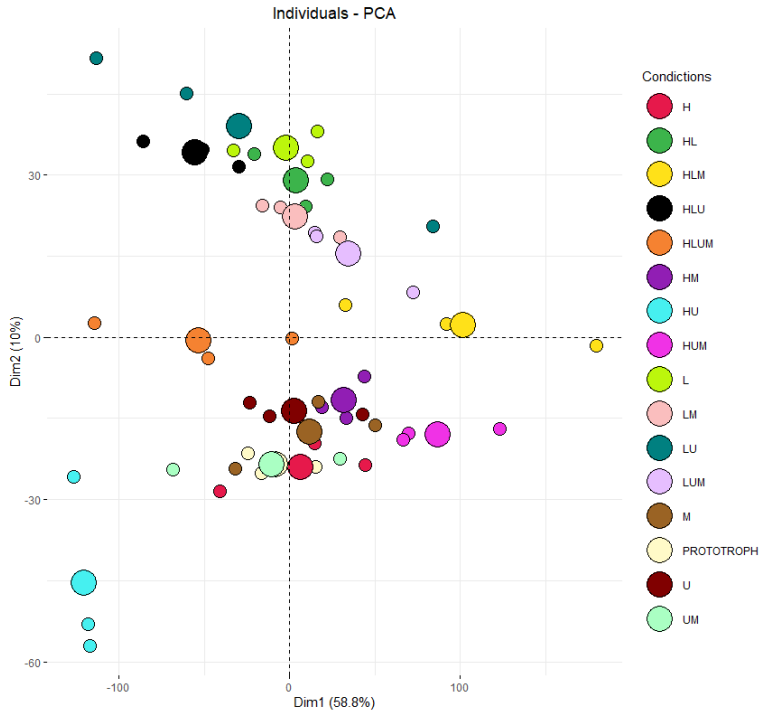

Figure 3. Two-dimensional PCA projection to PC1 and PC2 in transcriptomic dataset

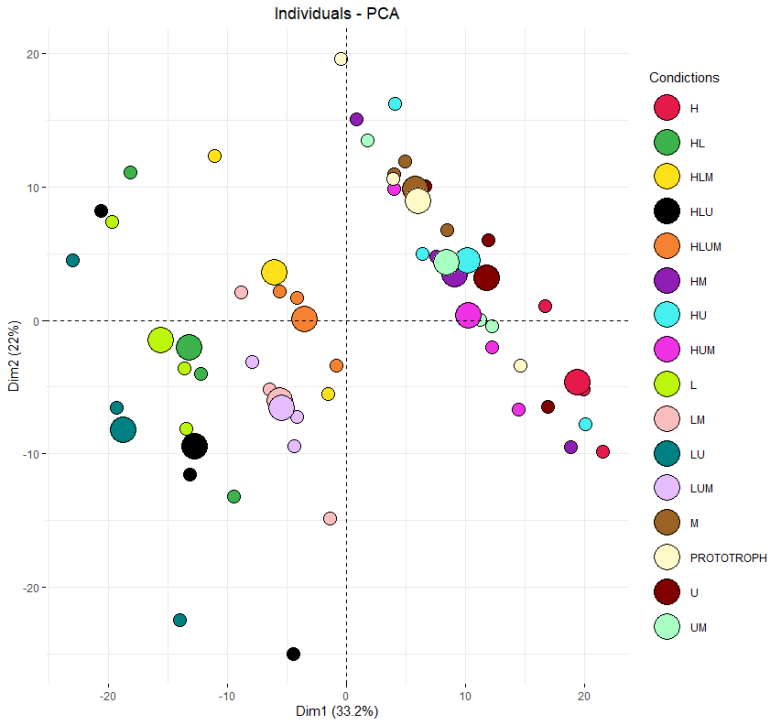

Figure 4. Two-dimensional PCA projection to PC1 and PC2 in proteomic dataset

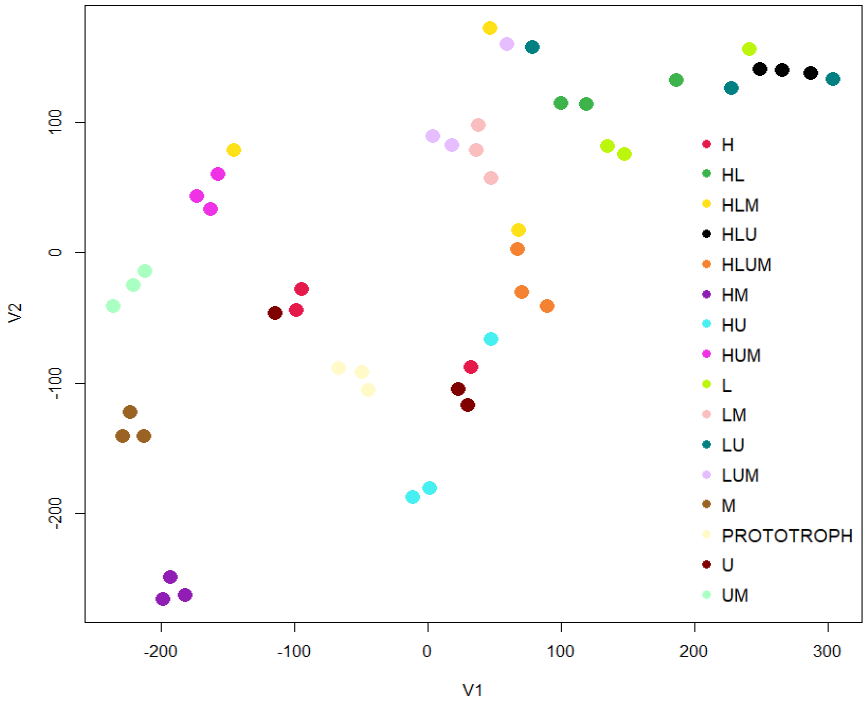

Figure 5. Clustering outcome from t-NSE in Transcriptomic dataset

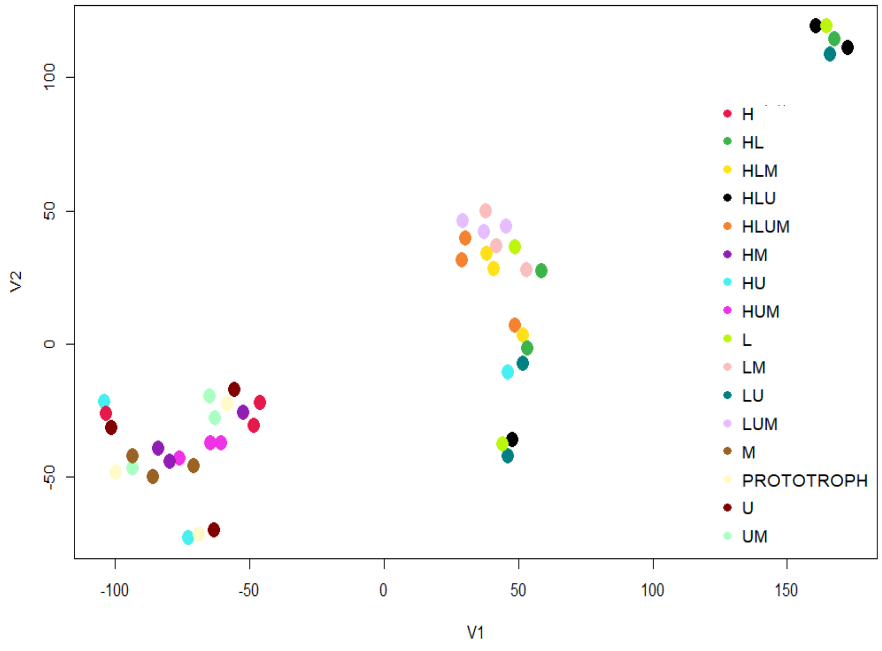

Figure 6. Clustering outcome from t-NSE in Proteomics dataset

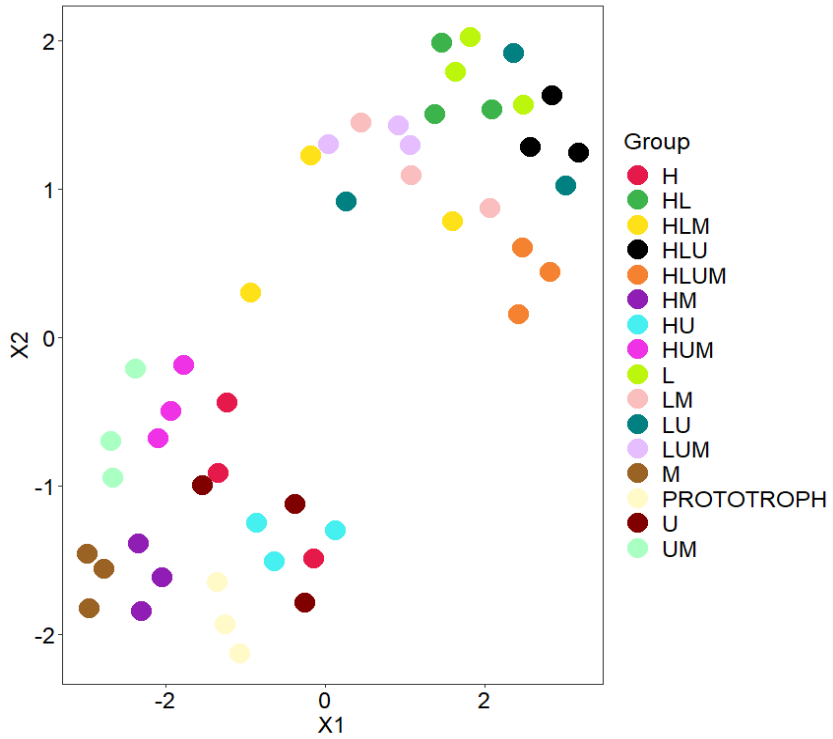

Figure 7. Clustering outcome from UMAP in transcriptomic dataset

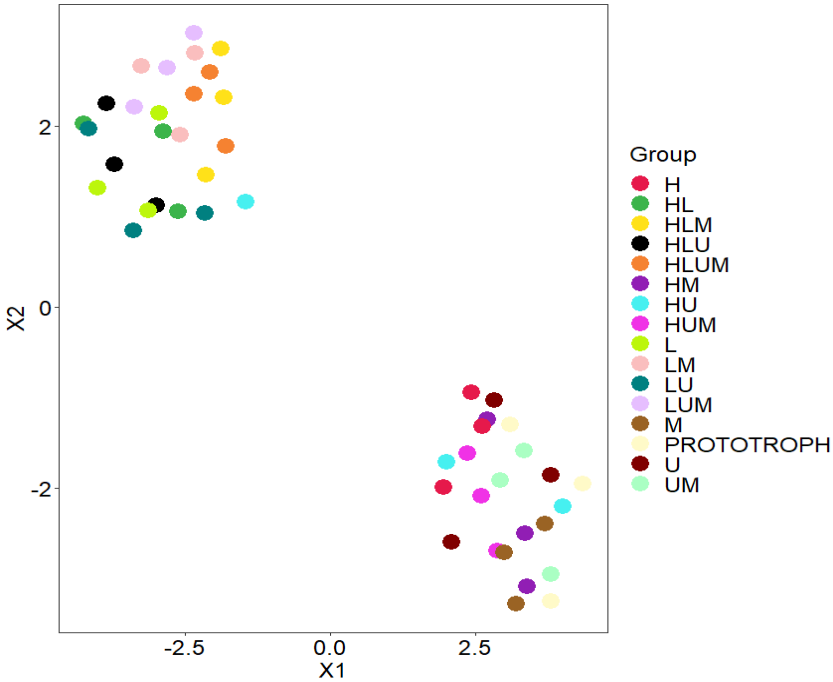

Figure 8. Clustering outcome from UMAP in proteomics dataset

3.2 Network construction

The transcriptome dataset that was input to the GENIE3 algorithm consists of 5983 transcriptomes and 48 samples that were generated from the 16 strains with three replicates each. The details regarding the algorithm, such as the hyperparameters used, can be found in the Methods section. The interaction table was imported into Cytoscape to visualize the network (Figure 9). Many small subnetworks and a large subnetwork were observed. Only the largest subnetwork was conserved due to the transcriptomes included in this largest subnetwork are more likely to have an important role in the regulatory processes among those strains. The largest subnetwork was then filtered again by the transcriptomes that exist in the STRING database, resulting in 1639 transcriptomes. A summary of the characteristics of the network can be found in Table 1.

A similar process was applied to the proteomic dataset with an initial number of 446 proteins and 48 samples. The generated network also showed a large subnetwork together with a few smaller subnetworks (Figure 10). When selecting only the largest subnetwork, 439 proteins remained. The network characteristics are summarized in Table 2. No overly clustered nodes were observed and the network resembles a common scale-free network.

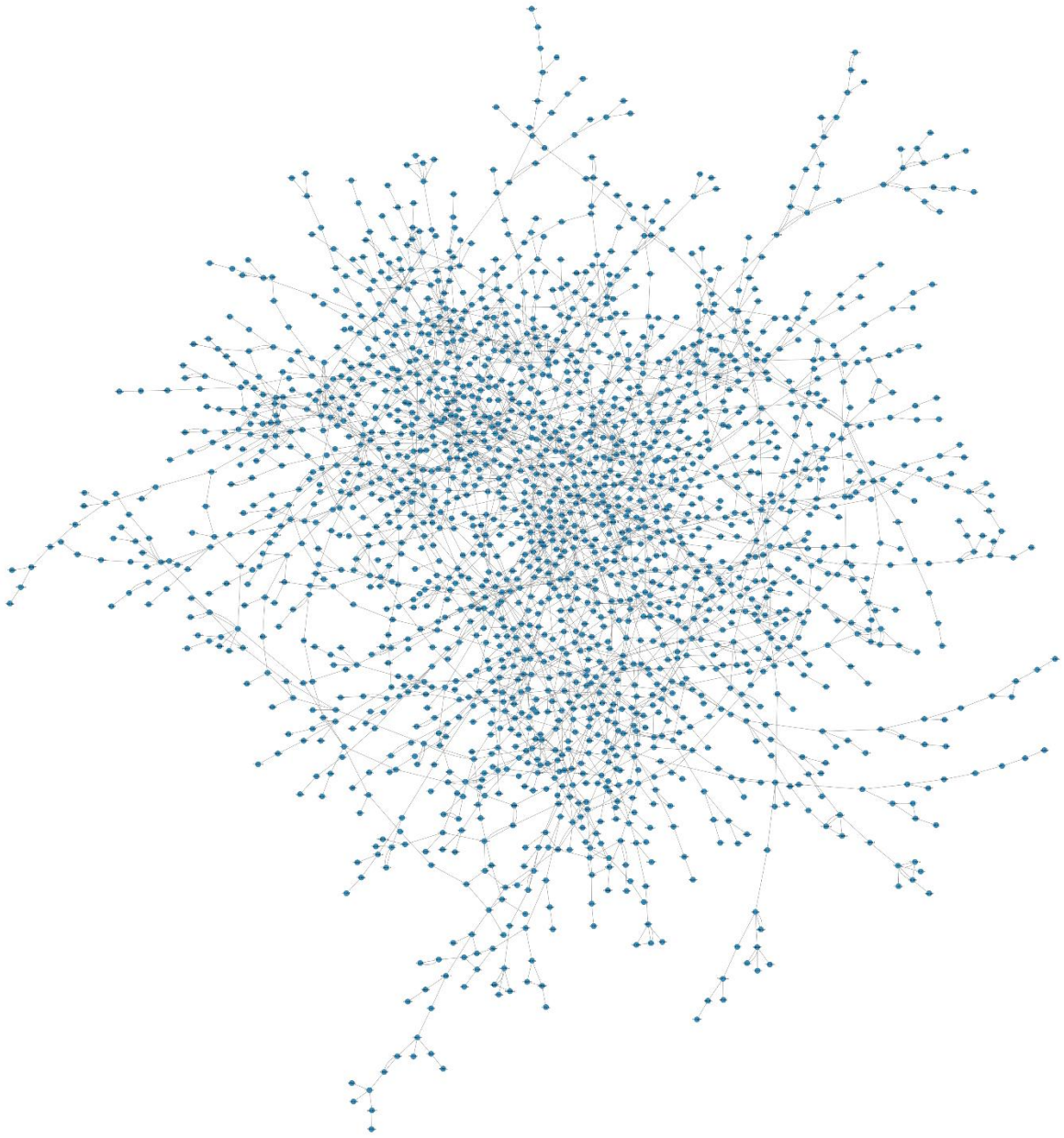

Figure 9. Transcriptome network generated from the transcriptomic dataset using GENIE3.

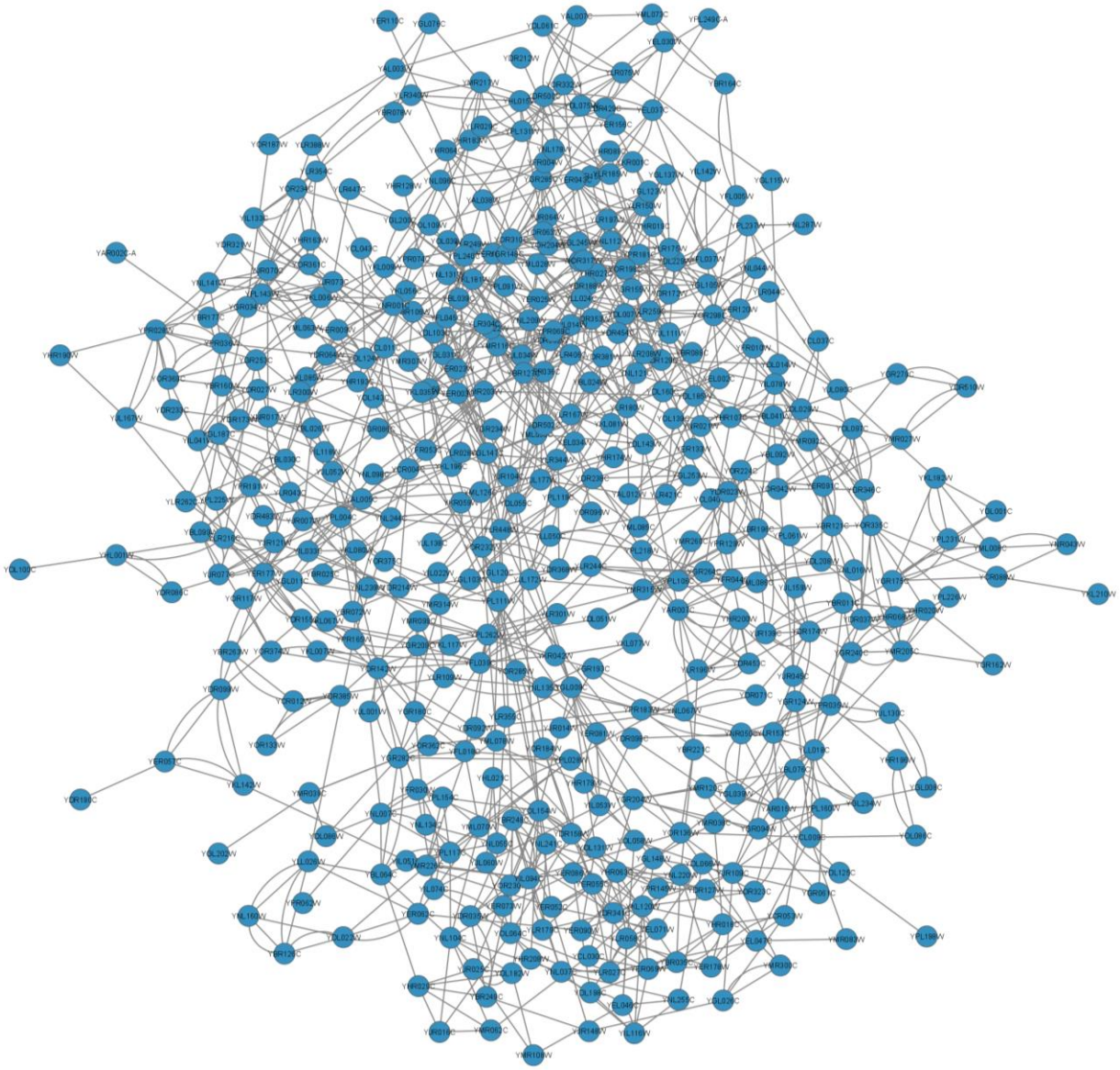

Figure 10. Proteome network generated from the proteomic dataset using GENIE3.

Table 1. Transcriptomic network characteristics.

Number of nodes	1639
Number of edges	2411
Average number of neighbors	2494
Network diameter	36
Network radius	20
Characteristic path length	14.497
Clustering coefficient	0.081

Table 2. Proteomic network characteristics.

Number of nodes	445
Number of edges	1276
Average number of neighbors	4.61
Network diameter	10
Network radius	6
Characteristic path length	5.068
Clustering coefficient	0.169

3.3 Differential expression analysis

The constructed networks can be used to help investigate and visualize the relationship between different strains in the regulatory networks. For this objective, three strains were selected, which are the single knockouts of leucine, methionine, and uracil gene strains. From the dimensionality reduction section above, it was shown that the leucine appears to be the most separated from the prototroph, followed by methionine and uracil. Uracil was observed to be very similar to prototroph in all dimensionality reduction algorithms. When performing the differential expression analysis, each of the three strains was compared to the prototroph for each transcriptomic and proteomic dataset. The obtained $\log_2$(fold change) values were then

incorporated into the network represented by the color of the nodes, where the red and blue nodes represent upregulation and downregulation of that transcriptome or proteome, respectively. The value differences of the $\log_2$(fold change) were also reflected by the color intensity with the maximum intensity at 1 and -1. For a consistent comparison, the same scale bar will be used for the following figures.

3.3.1 Leucine

For the single LEU2 knockout strain, most of the upregulated and downregulated genes are clustered around the center of the transcriptome network (Figure 11). Some upregulated (red) and downregulated (blue) clusters can be visualized. Out of the starting 5982 genes, 1311 and 1489 genes were upregulated ($\log_2$(fold change) > 0.2) and downregulated ($\log_2$(fold change) < -0.2), respectively. In the proteome network, clear clusters between the upregulated and downregulated proteins can be seen at the top and bottom of the network (Figure 12). Among the 446 proteins, 163 and 151 proteins are upregulated ($\log_2$(fold change) > 0.2) and downregulated ($\log_2$(fold change) < -0.2), respectively.

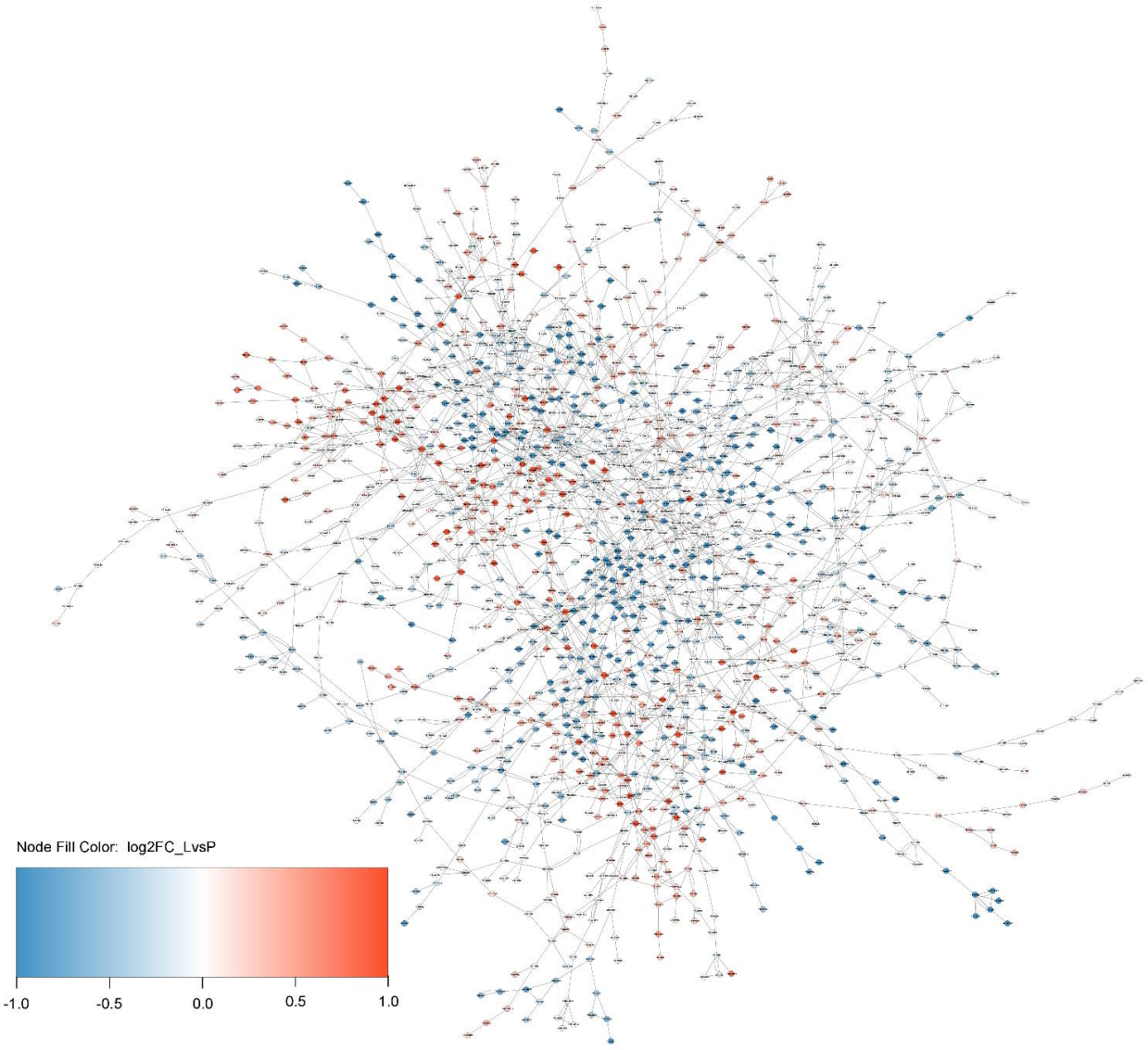

Figure 11. Transcriptomic network with nodes colored according to the log_2(fold change) values obtained from differential expression analysis between the LEU2 auxotroph and prototroph (red = upregulated, blue = downregulated).

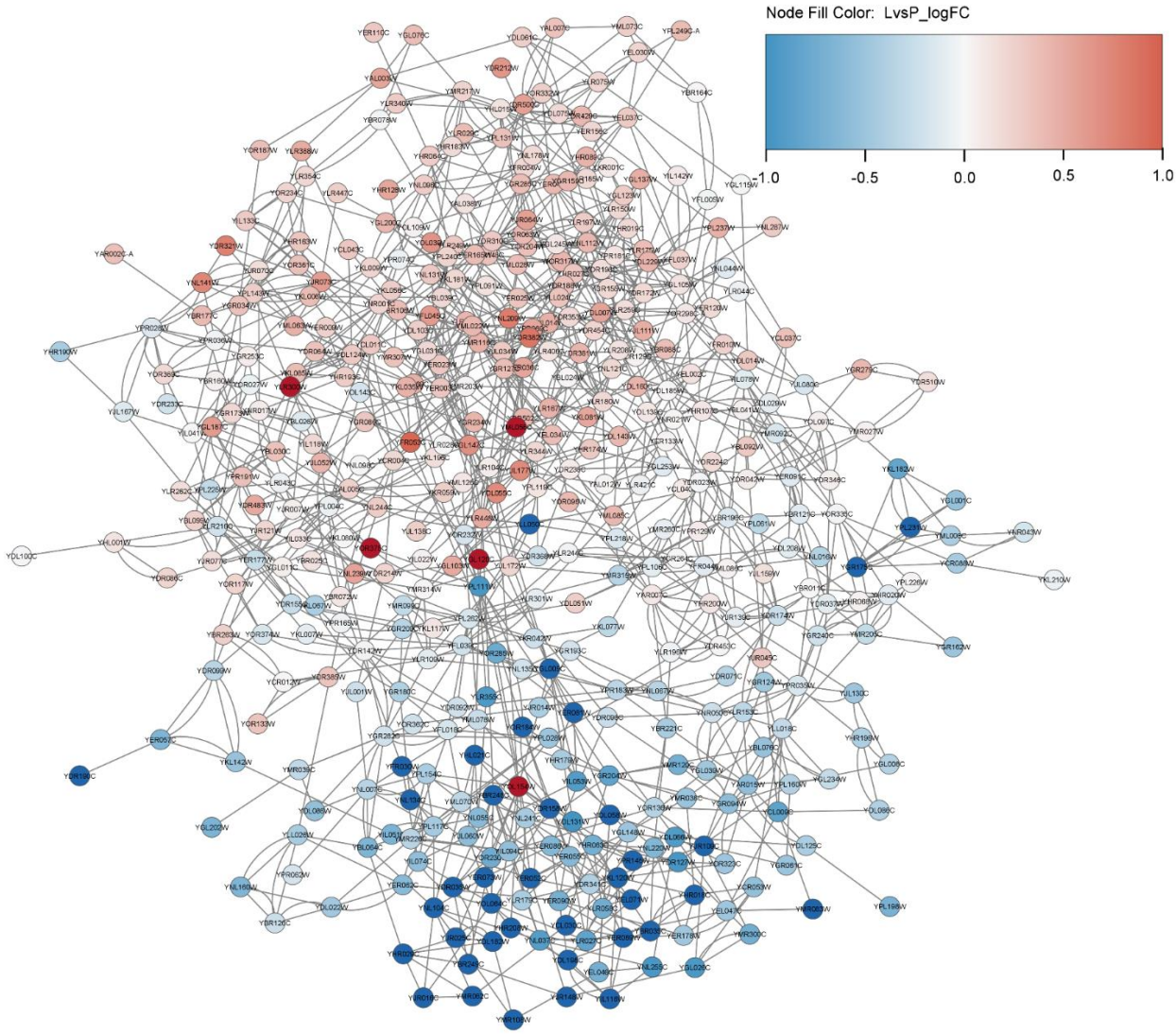

Figure 12. Proteomic network with nodes colored according to the log$_2$(fold change) values obtained from differential expression analysis between the LEU2 auxotroph and prototroph (red = upregulated, blue = downregulated).

3.3.2 Methionine

In the MET15 strain, there were more upregulated genes than downregulated genes, and they were clustered around the center of the transcriptome network (Figure 13). A downregulated cluster can be seen around the top half of the network, whereas some upregulated clusters can be found around the network. Out of the total 5982 genes, 882 and 713 genes were upregulated and downregulated, respectively. In the proteome network, the upregulated and downregulated proteins can be visually differentiated on the left and right of the network (Figure 14). Among the 446 proteins, 76 and 71 proteins are upregulated and downregulated, respectively.

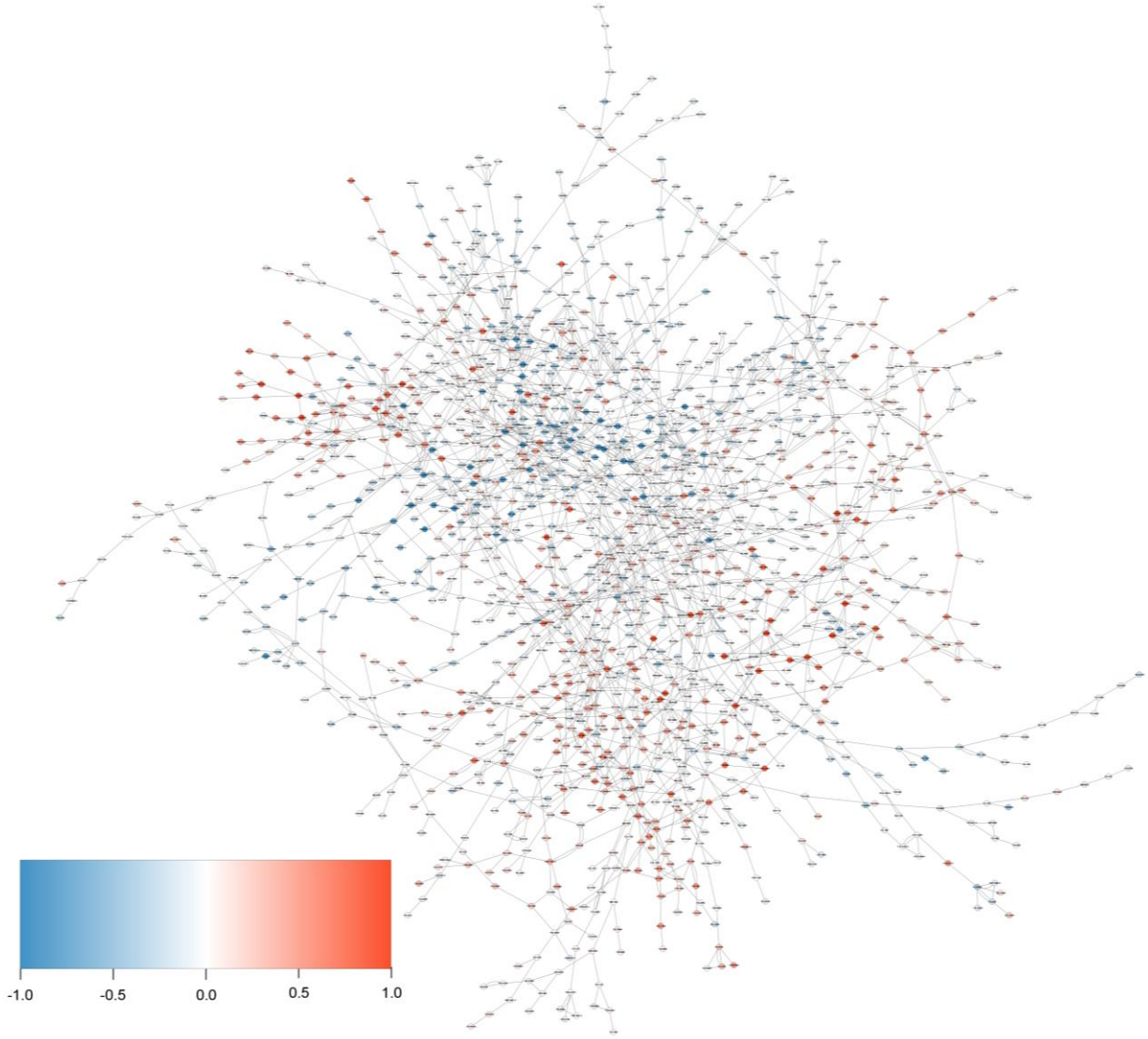

Figure 13. Transcriptomic network with nodes colored according to the log_2(fold change) values obtained from differential expression analysis between the MET15 auxotroph and prototroph (red = upregulated, blue = downregulated).

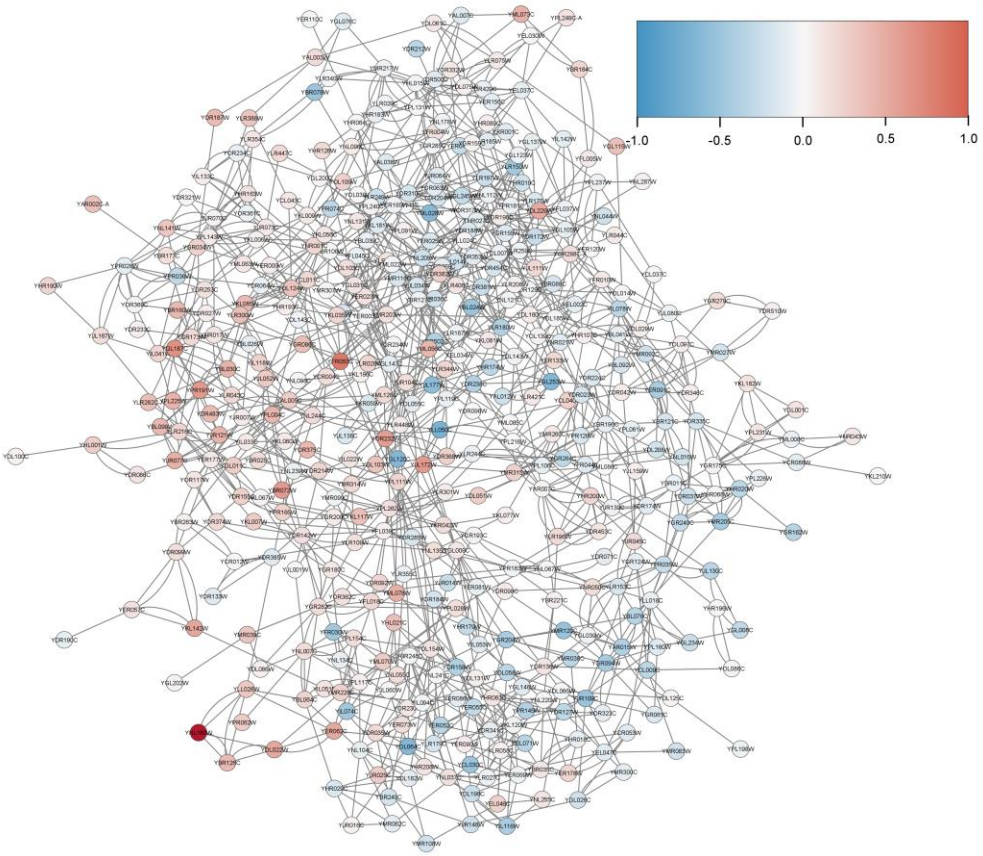

Figure 14. Proteomic network with nodes colored according to the log_2(fold change) values obtained from differential expression analysis between the MET15 auxotroph and prototroph (red = upregulated, blue = downregulated).

3.3.3 Uracil

In the URA3 strain, a cluster of upregulated genes can be found at the top left of the network, and not many genes are heavily downregulated (Figure 15). Out of the total 5982 genes, 758 and 474 genes were upregulated and downregulated, respectively. In the proteome network, the upregulated and downregulated proteins can be visually differentiated on the left and right of the network (Figure 16). Among the 446 proteins, 46 proteins were upregulated and 46 proteins were downregulated.

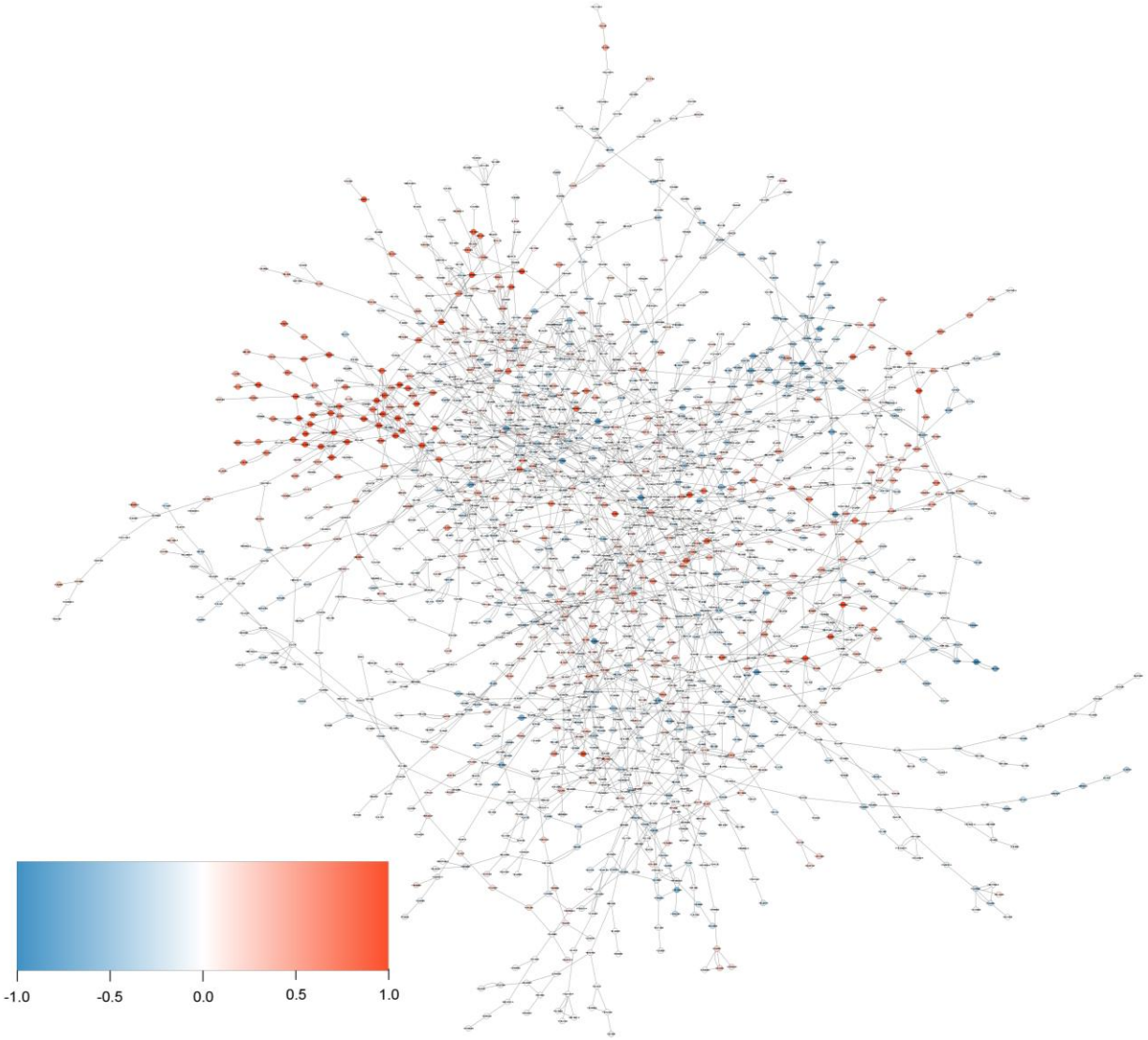

Figure 15. Transcriptomic network with nodes colored according to the log_2(fold change) values obtained from differential expression analysis between the URA3 auxotroph and prototroph (red = upregulated, blue = downregulated).

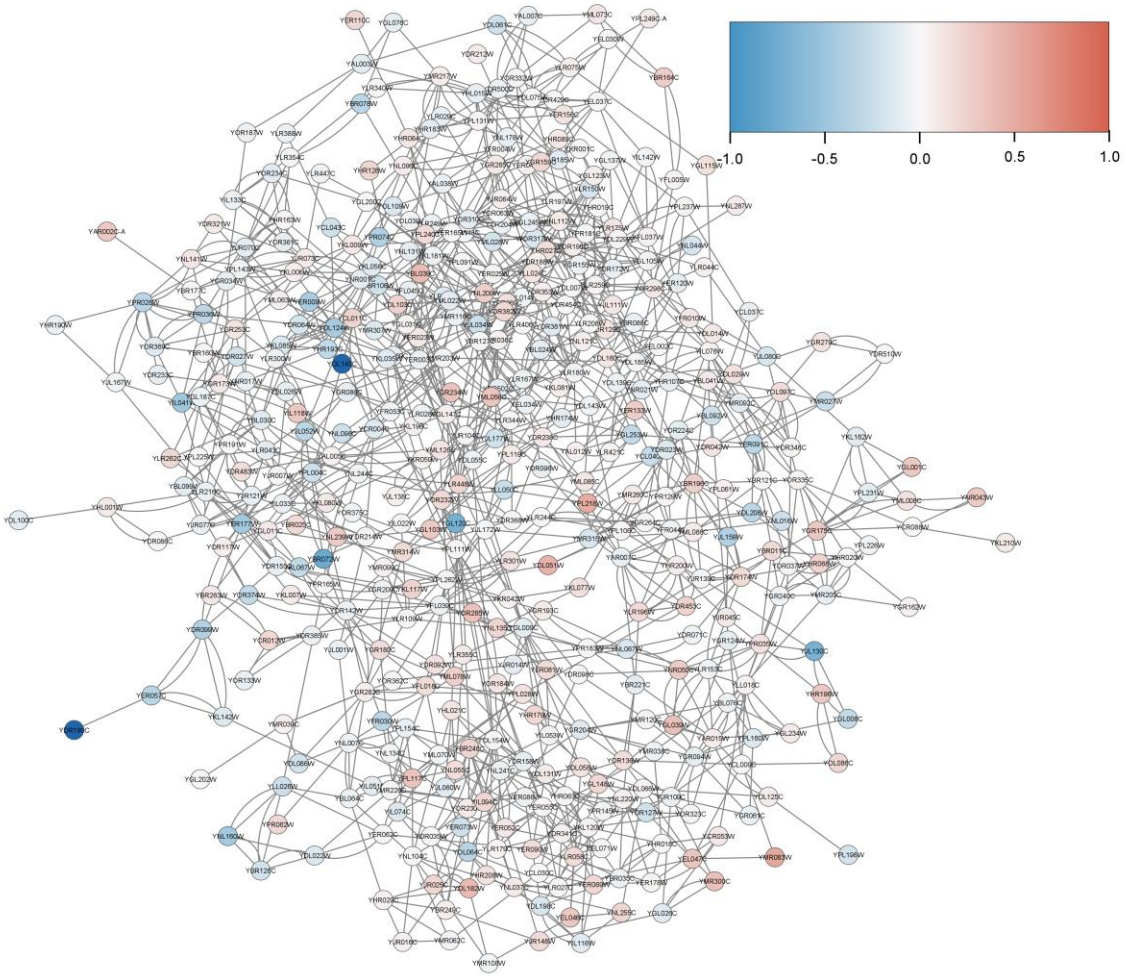

Figure 16. Proteomic network with nodes colored according to the log_2(fold change) values obtained from differential expression analysis between the URA3 auxotroph and prototroph (red = upregulated, blue = downregulated).

3.4 Gene Ontology (GO) terms analysis

Similar to the previous section, LEU2, MET15, and URA3 were specifically analyzed for their GO terms. Since the GO terms are similar when analyzing the transcriptomic and proteomic datasets of the same strain, only the result from the proteomic dataset is included. Additionally, only the first three terms of each of the GO aspects (cellular component, biological process, and molecular function) are included, ranked according to the false discovery rate.

3.4.1 Leucine

In the LEU2 strain, the downregulated proteins are mostly in the cytoplasm and are involved in metabolic processes. For molecular functions, those downregulated proteins are involved in oxidoreductase activity and DNA-binding transcription factor activity (Table 3). The upregulated proteins are predominantly in the ribosome and are related to translation processes. The molecular functions of those proteins are mainly for binding (Table 4).

Table 3. GO term analysis of 151 downregulated proteins in LEU2 strain.

#term ID	term description	observed gene count	background gene count	strength	false discovery rate
Cellular Component					
GO:0005737	Cytoplasm	145	4380	0.17	1.05E-16
GO:0005622	Intracellular	149	5255	0.1	8.81E-11
GO:0110165	Cellular anatomical entity	149	5509	0.08	4.15E-08
Biological Processes					
GO:0044281	Small molecule metabolic process	93	693	0.77	6.30E-47
GO:0019752	Carboxylic acid metabolic process	73	387	0.92	3.20E-43
GO:0044283	Small molecule biosynthetic process	67	324	0.96	1.70E-41
Molecular Function					
GO:0016491	Oxidoreductase activity	131	333	0.25	2.44E-05
GO:0000981	DNA-binding transcription factor activity, RNA polymerase II-specific	60	134	0.31	0.0051
GO:0003700	DNA-binding transcription factor activity	69	167	0.27	0.0061

Table 4. GO term analysis of 163 upregulated proteins in LEU2 strain.

#term ID	term description	observed gene count	background gene count	strength	false discovery rate
Cellular Component					
GO:0005840	Ribosome	43	260	0.83	8.14E-20
GO:0022626	Cytosolic ribosome	36	160	0.97	8.14E-20
GO:1990904	Ribonucleoprotein complex	56	519	0.65	6.69E-19
Biological Processes					
GO:0002181	Cytoplasmic translation	44	171	1.02	1.92E-26
GO:0006412	Translation	57	372	0.8	1.50E-25
GO:0043603	Cellular amide metabolic process	64	498	0.72	1.50E-25
Molecular Function					
GO:0005488	Binding	126	2877	0.25	1.72E-15
GO:1901363	Heterocyclic compound binding	101	1925	0.33	2.98E-15
GO:0097159	Organic cyclic compound binding	101	1936	0.33	3.03E-15

3.4.2 Methionine

In MET15 strain, the downregulated proteins are present in the cytoplasmic region and are involved in amino acid and carboxylic acid metabolic processes. Unlike LEU2, the molecular functions for these downregulated proteins are involved in biomolecules binding (Table 5). Interestingly, the upregulated proteins are also present in the cytoplasmic region and are involved in oxidation-reduction processes and nucleotide metabolic processes. The molecular functions of those upregulated proteins are catalytic and oxidoreductase activities (Table 6).

Table 5. GO term analysis of 71 downregulated proteins in MET15 strain.

#term ID	term description	observed gene count	background gene count	strength	false discovery rate
Cellular Component					
GO:0005622	Intracellular	71	5255	0.1	3.63E-05
GO:0005737	Cytoplasm	65	4380	0.15	0.00016
GO:0110165	Cellular anatomical entity	70	5509	0.08	0.0054
Biological Processes					
GO:0006520	Cellular amino acid metabolic process	33	246	1.1	4.78E-24
GO:0019752	Carboxylic acid metabolic process	38	387	0.97	4.78E-24
GO:0044281	Small molecule metabolic process	42	693	0.76	2.82E-20
Molecular Function					
GO:0036094	Small molecule binding	41	931	0.62	2.39E-14
GO:0000166	Nucleotide binding	39	868	0.63	6.87E-14
GO:0043168	Anion binding	38	946	0.58	4.23E-12

Table 6. GO term analysis of 76 upregulated proteins in MET15 strain.

#term ID	term description	observed gene count	background gene count	strength	false discovery rate
Cellular Component					
GO:0005737	Cytoplasm	72	4380	0.16	1.24E-06
GO:0005622	Intracellular	75	5255	0.1	0.00012
GO:0110165	Cellular anatomical entity	76	5509	0.08	0.00013
Biological Processes					
GO:0055114	Oxidation-reduction process	22	457	0.63	2.43E-05
GO:0009117	Nucleotide metabolic process	14	175	0.85	3.22E-05
GO:0044281	Small molecule metabolic process	26	693	0.52	3.22E-05
Molecular Function					
GO:0003824	Catalytic activity	49	2197	0.29	2.84E-05
GO:0016491	Oxidoreductase activity	16	333	0.63	0.00079

3.4.3 Uracil

In URA3 strain, no significant GO term in the cellular component and the biological process was observed for downregulated proteins. A slightly qualified false discovery rate in the molecular function was detected for oxidoreductase activity (Table 7). On the other hand, upregulated proteins are included in the cytoplasmic region and are involved in metabolic processes. The molecular functions of those proteins are catalytic and oxidoreductase activities (Table 8).

Table 7. GO term analysis of 46 downregulated proteins in URA3 strain.

#term ID	term description	observed gene count	background gene count	strength	false discovery rate
Molecular Function					
GO:0016903	Oxidoreductase activity, acting on the aldehyde or oxo group of donors	4	24	1.38	0.0486

Table 8. GO term analysis of 46 upregulated proteins in URA3 strain.

#term ID	term description	observed gene count	background gene count	strength	false discovery rate
Cellular Component					
GO:0005737	Cytoplasm	44	4380	0.16	0.0011
GO:0005622	Intracellular	46	5255	0.1	0.0072
GO:0110165	Cellular anatomical entity	46	5509	0.08	0.0415
Biological Processes					
GO:0044281	Small molecule metabolic process	21	693	0.64	5.33E-06
GO:0006520	Cellular amino acid metabolic process	13	246	0.89	1.91E-05
GO:0055114	Oxidation-reduction process	16	457	0.71	4.93E-05
Molecular Function					
GO:0003824	Catalytic activity	35	2197	0.36	3.74E-06
GO:0016491	Oxidoreductase activity	14	333	0.79	2.67E-05
GO:0016616	Oxidoreductase activity, acting on the ch-oh group of donors, nad or nadp as acceptor	6	76	1.06	0.008

3.5 Mix-omics integration

As high-throughput technologies continue to advance, datasets of different omics are more accessible, as is the case in [1]. Data integration method such as PLS allows for the combination of different datasets to visualize in a lower dimension. The following demonstrates the use of sPLS on transcriptomic and proteomic datasets to investigate the different metabolic auxotrophies. The sPLS methods selected 90 genes and 13 proteins from their corresponding datasets and calculated the correlation that maximizes the covariance between them. The resulting correlation matrix clearly separated the features into two clusters when plotted in the correlation circle plot (Figure 17). Few interpretations can be obtained from this plot: 1) the clusters are close to the outer circle, which indicates their importance in characterizing the different yeast strains; 2) the clusters are 90 degrees apart, which indicates that the features

between the two clusters are not correlated to each other, and each cluster is important on a component. The computed correlation can then be transformed into a network, and a similar differential expression analysis can be performed as in the earlier section. In this network, the nodes can be either a gene or a protein, annotated by a ".t" or ".p" at the end of the node label, respectively (Figure 18).

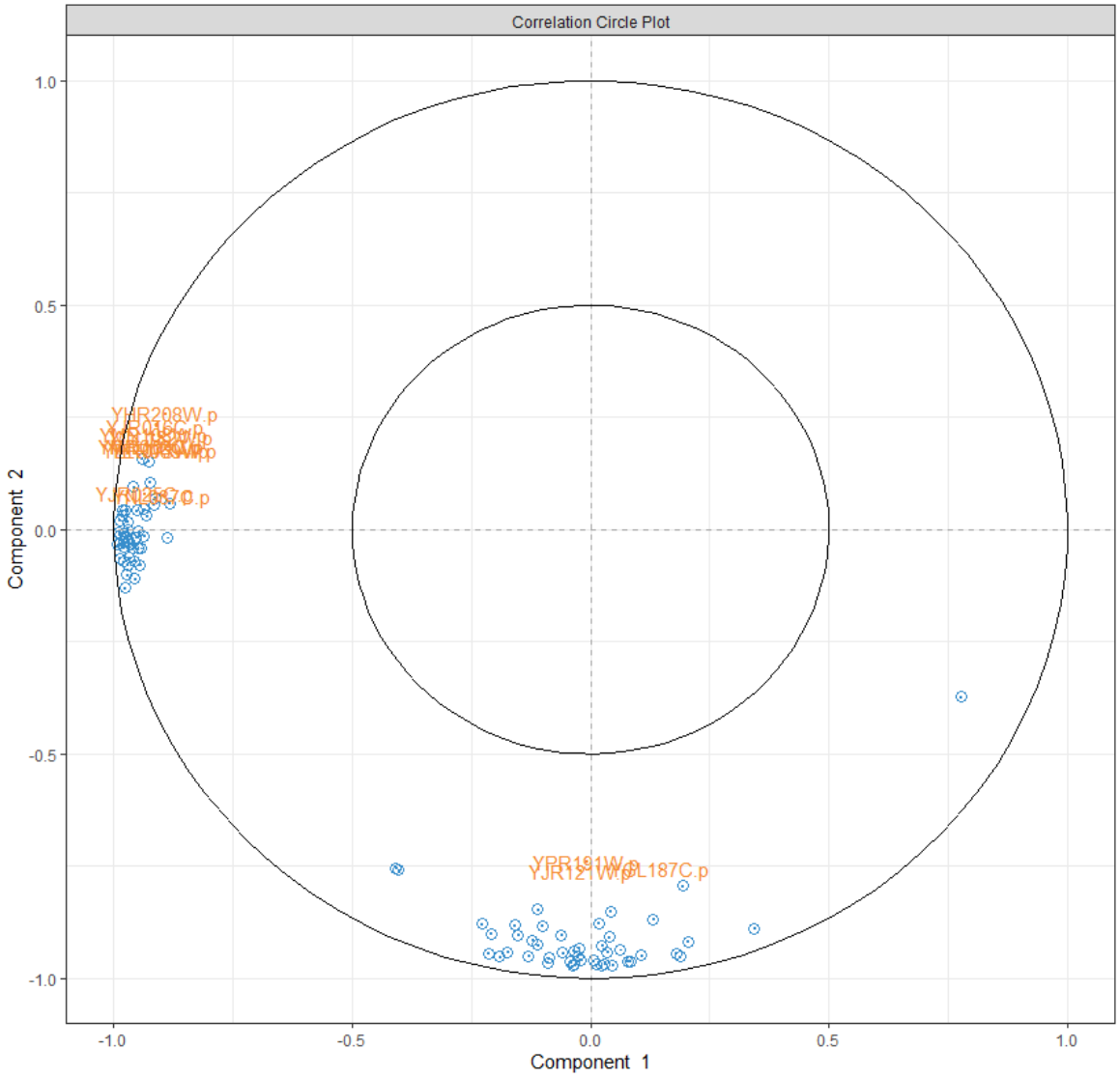

Figure 17. Correlation circle plot of the features selected using the sPLS method. Orange = proteins; blue = genes.

Similar to the earlier section, the differential expression of LEU2, MET15, and URA3 compared to the prototroph was performed and plotted on the network with the node's color representing $\log_2$(fold change) (Figure 18). Two subnetworks were formed resulting from the two clusters in the correlation circle plot above. For the LEU2 strain, the upregulated and downregulated genes and proteins are completely separated between the two subnetworks (Figure 18a). For the MET15 strain, the right subnetwork consists of all upregulated genes and proteins, whereas the left subnetwork contains a mix of upregulated and downregulated genes and proteins (Figure 18b). This indicates that the MET15 strain is more identical to the prototroph when compared to the LEU2 strain. For the URA3 strain, the right subnetwork instead consists of mostly downregulated genes and proteins, whereas the left subnetwork contains both upregulated and downregulated genes and proteins (Figure 18c). Note that the sPLS method was applied on all 48 samples (all 16 strains), indicating its good feature selection ability even when looking at the network clusters in just these three strains.

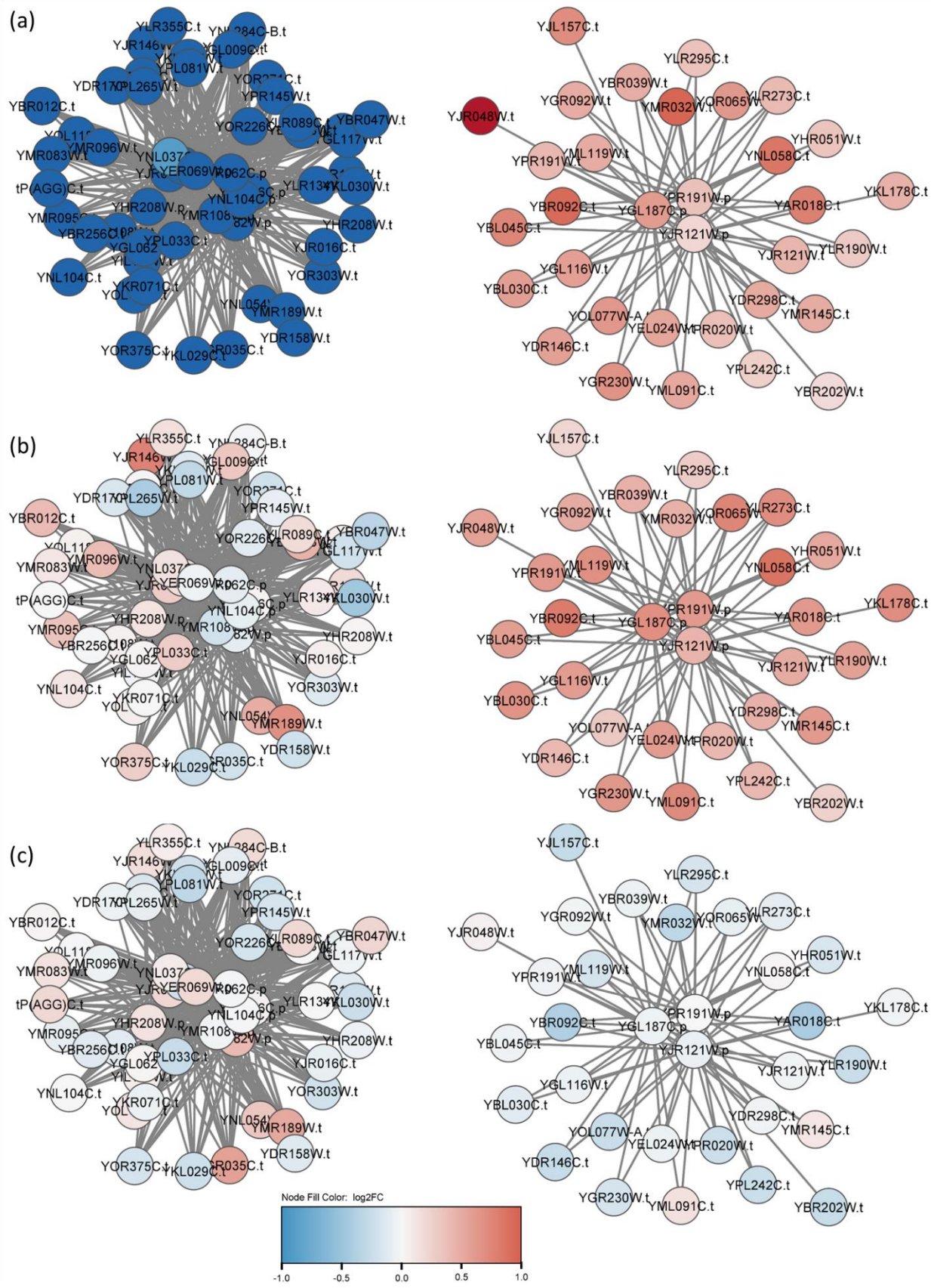
(a)
(b)
(c)
YJL157C.t
YLR295C.t
YBR039W.t
YGR092W.t
YMR032W.t
YJR048W.t
YLR273C.t
YHR051W.t
YPR191W.t
YNL058C.t
YKL178C.t
YBR092C.t
YAR018C.t
YBL045C.t
YJR121W.t
YLR190W.t
YGL116W.t
YBL030C.t
YDR298C.t
YMR145C.t
YOL077W-A.t
YEL024W.t
YDR146C.t
YPL242C.t
YGR230W.t
YML091C.t
YBR202W.t
YLR355C.t
YJR146W.t
YPL081W.t
YPL265W.t
YPR145W.t
YBR012C.t
YBR047W.t
YMR096W.t
YMR083W.t
YNL037C.t
YJR062C.p
YLR134W.t
YKL030W.t
YNL104C.p
YHR208W.p
YMR108W.p
YHR208W.t
YPL033C.t
YJR016C.t
YNL104C.t
YOR303W.t
YKR071C.t
YNL054W.t
YMR189W.t
YDR158W.t
YOR375C.t
YKL029C.t
YDR035C.t
Node Fill Color: log2FC
-1.0
-0.5
0.0
0.5
1.0

Figure 18. Transcriptome-proteome integrated network with differential expression analysis for (a) LEU2, (b) MET15, and (c) URA3 auxotrophs. The log_2(fold change) value is represented by the node color (red = upregulated, blue = downregulated).

4. Discussion

Metabolic-gene interactions are intricate, and recent high-throughput technologies may help to shine some light on this regard. In this work, the transcriptome and proteome of 16 *Saccharomyces cerevisiae* strains resulting from single or combinatorial deletions of LEU2, MET15, HIS3, and URA3 are employed from [1] for the investigation of metabolic-gene interactions. Using dimensionality reduction techniques such as PCA, T-SNE, and UMAP, a high-level distinction between the different strains can be identified. The LEU2 auxotroph itself was found to be very distinct in both the low dimension plot and in the experiment.

A co-expression network was constructed for further investigation and visualization of the features between different strains. For simplification, we selected LEU2, MET15, and URA3 single deletion strains, which were shown to be very different from prototroph (LEU2), slightly different from prototroph (MET15), and very similar (URA3) according to the dimension reduction plots. The network was able to illustrate some clusters between the upregulated and downregulated genes and proteins. In addition, the number of upregulated and downregulated genes and proteins also decreased from LEU2, MET15, and URA3 strains. This may indicate why LEU2 is the most distinguishable from the prototroph. The GO terms extracted from the upregulated proteins of LEU2 strain showed that the LEU2 strain had increased translational

processes and ribosomal activities. Practically, the strains with LEU2 deletion were indeed reported to have a consistent effect on growth rate [1]. Although other significant observations from other strains, the results obtained from GO functionality analysis could give a clue about the gene modification. Moreover, the network also showed clusters for some differentially expressed genes and proteins, which could indicate a special role or coordinated responses of those groups of proteins, despite possibly having different GO terms. However, one should also note that the expression data could be obtained at a steady-state or at any time point from random perturbation. This could result in incomplete capture of the actual system, as is also the limitation for other biological networks at the state of art. Integration of multiple datasets is one of the methods to possibly improve the understanding of the system, as is demonstrated in this work. Networks incorporating the dynamic expression of the omics can also be very vital for studying the system but comes with a higher computational cost. In general, both the experimental and computational end will require continuous development in the coming years to improve the accuracy and obtain a better understanding of the intricated biological system.

5. Conclusion

In this work, co-expression networks from transcriptomic and proteomic datasets were constructed to visualize yeast of different metabolic backgrounds. Using the dimensional reduction techniques (PCA, T-SNE, and UMAP), some strains appeared to be more distinguishable than the others when compared to the prototroph. Three strains were then selected for further analysis: LEU2, which is very distinguishable from the prototroph; MET15,

which is somewhat distinguishable; and URA3, which is hardly distinguishable. The upregulated and downregulated proteins (compared to prototroph) in LEU2 strain showed clear clusters in the constructed proteomic network. GO analysis further suggests that the upregulated proteins are significantly enriched in ribosomal activity, which justified the consistent effect on the growth rate for LEU2 perturbed strains. Clear clusters between upregulated and downregulated proteins were also seen for the MET15 strain, and hardly any cluster was observed for the URA3 strain. Using the sPLS method, the transcriptomic and proteomic datasets were combined, and a new network was constructed with 90 genes and 13 proteins (originally 5983 genes and 446 proteins). The new network consists of two subnetworks with key features that shows clusters of differentially expressed genes and proteins between different strains.

References

[1] M. T. Alam *et al.*, "The metabolic background is a global player in Saccharomyces gene expression epistasis," *Nat. Microbiol.*, vol. 1, no. 3, pp. 1–10, 2016.

[2] K. Pearson, "On lines and planes of closest fit to systems of points in space," *London, Edinburgh, Dublin Philos. Mag. J. Sci.*, 1901.

[3] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, 2008.

[4] S. van Dam, U. Võsa, A. van der Graaf, L. Franke, and J. P. de Magalhães, "Gene co-expression analysis for functional classification and gene-disease predictions," *Brief. Bioinform.*, vol. 19, no. 4, pp. 575–592, 2018.

[5] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, "Inferring regulatory networks from expression data using tree-based methods," *PLoS One*, vol. 5, no. 9, pp. 1–10, 2010.

[6] A. A. Margolin *et al.*, "ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinformatics*, vol. 7, no. SUPPL.1, pp. 1–15, 2006.

[7] A. C. Haury, F. Mordelet, P. Vera-Licona, and J. P. Vert, "TIGRESS: Trustful Inference of Gene REgulation using Stability Selection," *BMC Syst. Biol.*, vol. 6, 2012.

[8] O. A. Tomescu, D. Mattanovich, and G. G. Thallinger, "Integrative omics analysis. A study based on Plasmodium falciparum mRNA and protein data," *BMC Syst. Biol.*, vol. 8, no. 2, pp. 1–16, 2014.

[9] S. Wold, M. Sjöström, and L. Eriksson, “PLS-regression: A basic tool of chemometrics,” *Chemom. Intell. Lab. Syst.*, vol. 58, no. 2, pp. 109–130, 2001.

[10] K. A. Lê Cao, S. Boitard, and P. Besse, “Sparse PLS discriminant analysis: Biologically relevant feature selection and graphical displays for multiclass problems,” *BMC Bioinformatics*, vol. 12, no. June, 2011.

[11] “Genie3.” [Online]. Available: https://github.com/vahuynh/GENIE3.