Network-Based Contextualisation of LC-MS/MS Proteomics Data

by

Armin Guntram Geiger

Thesis presented in partial fulfilment of the requirements for the degree of Master of Science in Wine Biotechnology (Computational Biology) in the Faculty of AgriScience at Stellenbosch University

Institute for Wine Biotechnology,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Supervisor: Dr. D. Jacobson

December 2014

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: 28/09/2014

Abstract

Network-Based Contextualisation of LC-MS/MS Proteomics Data

A. Geiger
Institute for Wine Biotechnology,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Thesis: Master of Science in Wine Biotechnology (Computational Biology)
December 2014

This thesis explores the use of networks as a means to visualise, interpret and mine MS-based proteomics data.

A network-based approach was applied to a quantitative, cross-species LC-MS/MS dataset derived from two yeast species, namely *Saccharomyces cerevisiae* strain VIN13 and *Saccharomyces paradoxus* strain RO88.

In order to identify and quantify proteins from the mass spectra, a workflow consisting of both custom-built and existing programs was assembled. Networks which place the identified proteins in several biological contexts were then constructed. The contexts included sequence similarity to other proteins, ontological descriptions, proteins-protein interactions, metabolic pathways and cellular location.

The contextual, network-based representations of the proteins proved effective for identifying trends and patterns in the data that may otherwise have been obscured. Moreover, by bringing the experimentally derived data together with multiple, extant biological resources, the networks represented the data in a manner that better represents the interconnected biological system from which the samples were derived. Both existing and new hypotheses based on proteins relating to the yeast cell wall and proteins of putative oenological potential were investigated. These proteins were investigated in light of their differential expression between the two yeast species. Examples of proteins that were investigated included cell wall proteins such as GGP1 and

SCW4. Proteins with putative oenological potential included haze protection factor proteins such as HPF2. Furthermore, differences in capacity for malo-ethanolic fermentation between the two strains were also investigated in light of the protein data. The network-based representations also allowed new hypotheses to be formed around proteins that were identified in the dataset, but were of unknown function.

Uittreksel

Netwerk-Gebasseerde Kontextualisasie van LC-MS/MS Proteome Data

("Network-Based Contextualisation of LC-MS/MS Proteomics Data")

A. Geiger

Institute vir Wynbiotegnologie,
Universiteit van Stellenbosch,
Privaatsak X1, Matieland 7602, Suid Afrika.

Tesis: Magister in die Natuurwetenskappe in Wynbiotegnologie

Desember 2014

Hierdie studie verken die gebruik van netwerke om proteonomiese data te visualiseer, te interpreteer en te ontgin.

'n Netwerkgebaseerde benadering is gevolg ter ontleding van 'n kwantitatiewe LC-MS/MS datastel wat afkomstig was van twee gis-spesies nl, *Saccharomyces cerevisiae* ras VIN1 en *Saccharomyces paradoxus* ras RO88.

Die massaspektra is met bestaande en selfgeskrewe rekenaarprogramme verwerk om 'n werkvloei saam te stel ter identifisering en kwantifisering van die betrokke proteïene. Hierdie proteïene is dan aan bestaande biologiese databasisse gekoppel om die proteïene in biologiese konteks te plaas. Die gekontekstualiseerde is dan gebruik om biologiese netwerke van die data te bou. Die kontekste beskou onder meer lokalisering van selaktiwiteite, ontologiese beskrywings, ooreenkomste in aminosuur-volgordes en interaksies met bekende proteïene asook assosiasie en verbintenisse met metaboliese paaie.

Hierdie kontekstuele, netwerk-gebaseerde voorstelling van die betrokke proteïene het effektief duidelike data-tendense en patrone opgelewer wat andersins nie opmerkbaar sou wees nie. Daarby het die kombinering van eksperimentele data en bestaande biologiese bronne 'n beter perspektief aan die data-analise verleen. Beide bestaande en nuwe hipoteses tov gis-selwandproteïene en proteïene met moontlike wynkundige potensiaal is ondersoek in die lig van hul differensiële uitdrukking in die twee gis-spesies. Voorbeelde wat ondersoek is sluit in selwandproteïene soos GGP1 en SCW4 asook waasbeskermingsfaktorproteïen HPF2. Verskille tov kapasiteit mbt malo-etanoliese gisting is ook gevind. Die netwerk-gebaseerde voorstellings het ook aanleiding gegee tot die formulering van nuwe hipoteses mbt datastel-proteïene waarvan die funksies tans onbekend is.

Acknowledgements

I would like to acknowledge Carissa Blekker, Piet Jones and Debbie Weighill for assistance, advice and useful discussions, as well as Daniel Jacobson for supervision, input, advice and support. Special thanks to Debbie Weighill and Daniel Jacobson for editorial input and direction.

For funding I would like to acknowledge National Research Foundation, Institute for Wine Biotechnology, Swedish UD40 project, Winetech and THRIP

Dedications

To my wonderful wife Nelia, Loyal Family, Friends and Carl Terblanche (1989/03/31- 2013/01/31).

Contents

Declaration ii

Abstract iii

Uittreksel v

Acknowledgements vii

Dedications viii

Contents ix

List of Figures xi

List of Tables xvi

1 Introduction and Aims 1

1.1 Network-based Contextualisation of Omics Data 1
1.2 Aims 2
1.3 Summary 3

Bibliography 4

2 Literature review 1

2.1 Introduction 1
2.2 Proteomics 2
2.3 Liquid Chromatography (LC) 3
2.4 Mass Spectrometry 5
2.5 MS in Proteomics 12
2.6 Shotgun Proteomics with LC-MS/MS 13
2.7 MS Data Analysis 17
2.8 Contextualisation 20
2.9 Networks 27
2.10 Conclusion 28
2.11 Supplementary Materials 29

Bibliography **33**

3 Protein Identification, Quantification and Network-based Contextualisation **1**

3.1 Introduction 1

3.2 Methods 3

3.3 Results and Discussion 17

3.4 Summary and Conclusion 41

3.5 Supplementary Material 42

Bibliography **48**

4 Conclusion **1**

4.1 Concluding Remarks 1

4.2 Future Work 2

Bibliography **4**

List of Figures

2.1 **Diagram of an HPLC instrument.** (A) shows an overview of the instrument and its components. (B) is a simplified diagram of the output with z, x and y representing resolved analytes. Figure adapted from [20]. 4

2.2 **Diagram depicting ESI and the components involved.** (A) shows the high voltage power supply, metal capillary and flow of electrons between the components. (B) is an enlarged view of the liquid cone at the tip of the capillary. Figure adapted from [31]. . . . 6

2.3 **Diagram depicting MALDI coupled to a Time-of-Flight mass analyser.** The sample plate is a matrix where each entry is a sample spot. Sample molecules are ionised by gas-phase proton transfer from the matrix. The immediate area of laser excitation forms a plume, consisting of matrix and analyte ions, which directly enters the high-vacuum of the mass spectrometer. Figure taken from [34]. 8

2.4 **Diagram of a TOF analyser**. Figure adapted from [38] 8

2.5 **Diagram of a FT-ICR mass analyser**. Figure taken from [38] . 9

2.6 **Diagram of a Quadrupole mass analyser**. Figure taken from [40]. 10

2.7 **Schematic of a quadrupole ion trap**. Figure taken from [40]. . 10

2.8 **Cutaway diagram of orbitrap mass analyser.** The red arrow indicates the point where ions are injected into the orbitrap. Figure adapted from [43]. 11

2.9 **A diagram of the LTQ Orbitrap.** This is an example of a hybrid instrument and consists of three main parts. The first section after the source is a linear ion trap which is able a to detect MS and MS^n spectra. In the C-trap component ions are accrued and their energy dampened after which they are injected into the orbitrap. Figure taken from [44]. 12

2.10 **PMF and sequencing by tandem MS.** Two common methods for protein identification via MS. Figure taken from [49]. 13

2.11 **The workflow and layout of a generic MS proteomics experiment consisting of five steps.** Figure taken from [46]. . . . 14

2.12 **Experimental strategies for global quantitative MS-based proteomics.** Figure adapted from [60]. 16
2.13 **Generic data analysis workflow for LC-MS/MS data** 18
2.14 **Diagram showing an overview for protein identification from LC-MS/MS data.** The open circles A to D represent proteins in a mixed sample. The peptides are represented by the open squares. Figure taken from [81]. 19
2.15 **General steps in the OrthoMCL workflow.** 21
2.16 **Sensisitivity vs specifity plot for various orthology detection softwares.** The x-axis of the figure represents false positive rates and the y-axis represents false negative rates. Figure taken from [88]. 22

3.1 **Experimental design used to produce the VIN13 and R088 protein mass spectra.** Three biological replicates for VIN13 and RO88 respectively were produced. Each replicate sample was labelled with a unique Tandem Mass Tag (TMT) with Tag masses ranging from 126 Da to 131 Da. 4
3.2 **Overview of the TPP workflow utilised for the identification and quantification of proteins.** The ellipses represent programs or processes whilst the rectangles are inputs and or outputs. 5
3.3 **The protein - GO term network in unfiltered form.** Proteins are represented as circular nodes and GO terms are represented as hexagonal nodes. The protein node border reflects a continuous color mapping between black and white relating to the probability of the protein identification where black indicates a high quality identification and white indicates a low quality identification. The protein node color corresponds to a continuous color mapping reflecting the relative fold change attribute. Red shades indicate an increase in the quantity of the protein in VIN13 relative to the amount in RO88. Blue shades indicate a fold increase in the quantity of the protein in RO88 relative to the amount in VIN13. Darker shades indicate a larger fold change. A white protein node color indicates that no relative fold change could be calculated for this protein. An edge between a protein node and GO term indicates that the GO term is ascribed to the protein. 12
3.4 **Unfiltered protein family modules - GO term network.** The visual mapping style is the same as in Figure 3.3 except that the nodes here are protein family modules and not proteins. 13
3.5 **Protein metabolic network.** Proteins are represented as circular nodes and the metabolic pathway nodes are rounded rectangles. The visual mapping style for the protein nodes are the same as in Figure 3.3. An edge between a protein node and a pathway node means that the protein is involved in the pathway. 14

3.6 **Protein-protein interactome networks.** (A) and (B) are gene-based and physical-based protein-protein interaction networks respectively. The visual mapping style is the same as in Figure 3.3 except that there are no GO term nodes in this network. 16

3.7 **Protein-protein interactome module to GO term networks.** Figures (A) and (B) are gene-based and physical-based protein-protein interaction module networks respectively. The blue rectangular nodes are interaction modules and the green nodes are GO terms. 16

3.8 **The amount of proteins identified in both species and in VIN13 and RO88 only.** A total of 396 proteins are presented and contextualised as a result of the workflow. 169 of the 396 proteins were identified in both VIN13 and RO88, whilst 115 proteins could only be identified in VIN13 and 112 only identified in RO88. 20

3.9 **The amount of proteins for which fold changes could be calculated.** Of the 169 proteins identified in both species, fold changes could only be calculated for 111. However, of this 111, 81 proteins had only one quantitative peptide that met the criteria set in Section 3.2.5, whilst only 30 proteins had 2 or more peptides from which to calculate a relative fold change. Furthermore, only 77 of the fold changes were found as statistically significant by the method. 21

3.10 **Network of enriched cellular component GO terms.** This is a subnetwork of the protein - GO term network constrained to only enriched GO terms belonging to the category cellular component and the proteins that are associated with this subset of GO terms. The visual style is the same as in Figure 3.3. 22

3.11 **Network of enriched molecular function GO terms.** This is a subnetwork of the protein - GO term network constrained to only enriched GO terms belonging to the category molecular function and the proteins that are associated with this subset of GO terms. The visual mapping style is the same as in Figure 3.3. 23

3.12 **Network of enriched biological process GO terms.** This is a subnetwork of the protein - GO term network constrained to only enriched GO terms belonging to the category biological process and the proteins that are associated with this subset of GO terms. The visual mapping style is the same as in Figure 3.3. 24

3.13 **Protein-protein GO overlap networks.** (A) - (E) are protein-protein networks with Jaccard index thresholds of (A) 0.1 (B) 0.3 (C) 0.5 (D) 0.7 (E) 0.9. The networks consist only of protein nodes and the visual mapping style for the nodes is the same as in Figure 3.3. Edge sizes are scaled according to the Jaccard index between the two connected nodes, a thicker edge represent a higher Jaccard index. 25

3.14 **The protein-protein GO overlap network at a Jaccard index threshold of 0.2.** (A) is the entire protein-protein GO overlap network at this threshold. (B) and (C) are the GTPase group and Histone group subnetworks respectively. The visual mapping style and network type is the same as in Figure 3.13. 26

3.15 **Orthology module 81 with the GO terms that describe it.** The members of Ortholog module 81 are YNL031C, BUR5, YBL002W and SPT12 (histone group). Family modules are represented as circular nodes and GO terms are represented as hexagonal nodes. The size of the family module node corresponds to the number of proteins in the module. An edge between a protein family module node and a GO term indicates that the GO term is ascribed to the protein family module. The visual mapping style is the same as in Figure 3.3 except that the nodes here are protein family modules and not proteins. 27

3.16 **Subnetwork of the protein-GO term network filtered for all nodes relating to the cell wall.** The visual mapping style is the same as in Figure 3.3. The dashed arrows point to GO terms. 30

3.17 **A cell wall subnetwork of the protein-protein network with edges weighted by Jaccard index.** The visual mapping style is the same as in Figure 3.13. 31

3.18 **A simplified metabolic pathway diagram of the mechanisms for reoxidation of assimilatory NADH in *S. cerevisiae.*** The top half of the diagram shows the reoxidation of NADH via glycerol formation. The bottom half of the diagram shows MEF. The diagram contains some of the key enzymes, compounds and pathways involved. The rectangles with rounded edges are pathways, circular shapes are compounds and rectangles are enzymes. Dashed lines indicate an indirect pathway link and solid lines are direct biochemical reactions involving the enzymes. 36

3.19 **Protein metabolic network with TCA and glycerol metabolism subnetworks**. (A) is all of the identified proteins within a metabolic context. A selection of pathways are pointed out for reference. (B) is a subnetwork of (A) and shows the six TCA cycle proteins that had significant fold changes. (C) is also a subnetwork of (A) and shows two proteins that had significant fold changes and are involved in glycerol and glycerophospholipid metabolism. 37

3.20 **YHR138C in GO and interactome contexts respectively.** (A) is the GO context and (B) is the interactome context. The visual mapping style is the same as in Figure 3.3. 39

3.21 **Possible hypothesis for the function of YHR138C.** 41

S1 **Parameters used for X!Tandem.** 43

S2 **The output from Protein Prophet for VIN13 proteins.** The x-axis of the graph indicates the minimum probability and the y-axis represents both sensitivity and error. The red line indicates sensitivity, whilst the green line indicates error. The corresponding table provides the values from which the graph is plotted. 44

S3 **The output from Protein Prophet for RO88 proteins.** The x-axis of the graph indicates the minimum probability and the y-axis represents both sensitivity and error. The red line indicates sensitivity, whilst the green line indicates error. The corresponding table provides the values from which the graph is plotted. 44

S4 **The minimum spanning trees of the gene-based and physical-based protein-protein interaction networks respectively.** (A) is the minimum spanning tree of the gene-based protein-protein interaction network. (B) is the minimum spanning tree of the physical-based protein-protein interaction network. The visual mapping scheme is the same as in Figure 3.6. 45

S5 **Physical-based protein-protein interaction network filtered for one or more line of evidence.** The visual mapping scheme is the same as in Figure 3.6. 45

List of Tables

3.1 **Summary and figure reference for all networks constructed in Section 3.2.8.** . . . 17

3.2 **The number of edges that exist at different Jaccard index thresholds as determined for the protein-protein network.** . . 22

3.3 **A subselection of attributes belonging to the proteins of the GTPase group.** In the protein probability column the VIN13 probability is given first followed by the RO88 probability. If the probability of identification for the proteins was the same in both species, only one number is given. The number of GO terms and number of enriched GO terms columns give the amount of GO terms that describe the proteins and the number of enriched GO terms respectively. . . . 24

3.4 **A subselection of attributes belonging to the proteins of the cell wall subnetworks.** In the Probability of identification column the VIN13 probability is given first followed by the RO88 probability. If the probability of identification for the proteins was the same in both species, only one number is given. The number of quantitative peptides column indicates the number of peptides that were available to derive the fold change. The fold change column gives the relative fold change ratio for VIN13/RO88 as calculated in Section 3.2.5. The secretion signal column may contain either a Yes or No variable to indicate if the protein has a secretion signal defined within the databases. . . . 32

3.5 **A subselection of attributes belonging to the proteins shown in Figure 3.19.** In the protein probability column the VIN13 probability is given first followed by the RO88 probability. If the probability of identification for the proteins was the same in both species, only one number is given. The number of quantitative peptides column indicates the number of peptides that were available to derive the fold change. . . . 38

3.6 **A subselection of attributes for YHR138C and its interacting proteins.** The column descriptions are the same as in Table 3.5 . . . 38

S1 **The number of edges that exist at different Jaccard index thresholds as determined for the protein family modules GO overlap network.** 42

Chapter 1

Introduction and Aims

1.1 Network-based Contextualisation of Omics Data

The characteristics of biological systems arise from the interactions amongst the molecules of which they are comprised. However, the coherent behaviours and responses that are observed in living systems are not solely produced by the interactions between the individual molecules themselves. Instead, they are the result of large numbers of functionally diverse sets of components that interact selectively [1; 2; 3].

The advent of omic technologies have greatly advanced knowledge of living systems by offering system-wide snapshots of classes of molecules at a given point in time. However, given the interlinked, multi-layered systems from which these datasets are derived and the generally high volumes of data that these omic experiments produce, the tasks of data analysis and interpretation present many new challenges.

Interpreting such large datasets requires a multi-faceted approach, drawing on both predictive computational tools and extant knowledge contained within biological databases and literature. Furthermore, the systems-based nature of the problems necessitates that many layers of information be integrated. In order to facilitate such tasks, conceptual and physical frameworks for the integration of experimental data with relevant resources are needed.

Networks are ideal for the study and modelling of complex systems. A system consisting of interlinked components can be represented as a collection of nodes and edges and this form of representation offers not only an intuitive means for visualisation, but also a mathematical structure to which a variety of network analysis tools can be applied. Moreover, representing omics data as a network offers both a global and local perspective of the data, allowing a considerable amount of data to be visualised without obscuring patterns, trends and relationships [1; 2; 3]. Such network-based approaches have proven to be effective solutions to the problems of data contextualisation and interpretation

[4; 5; 6].

Historically, the fields of genomics, transcriptomics and metabolomics have received relatively more attention when compared to proteomics. This can mainly be attributed to the technical difficulties associated with large scale proteome analyses [7]. However, collective advances in chromatography, techniques for the ionisation of biomolecules, Mass Spectrometry (MS) and computational capacity have made large-scale proteomics comparable in power to other more established omics technologies [8].

1.2 Aims

This thesis focuses on the analysis and interpretation of MS-based proteomics data by making use of networks as a tool for representing proteins within various biological contexts. The multi-faceted nature and size of the datasets necessitates the use of both computational network analysis tools and visual network interpretation of the data. The aims of this work were as follows: 1) Identify and quantify proteins from a LC-MS/MS dataset derived from a cross-species yeast secretome sample set that was labelled with Tandem Mass Tags; 2) Construct networks placing the identified proteins in various biological contexts; 3) Use the networks to investigate and mine the data with the objectives of investigating existing hypotheses as well as formulating new hypotheses based on trends and patterns in the data. Each one of the aims are discussed in more detail below:

1. Various programs and complete software packages for the identification and quantification of proteins from LC-MS/MS data exist [9; 10; 11]. However, the identification and quantification of proteins from a quantitative, cross-species proteomics experiment requires a customised workflow in order for the data to be interpreted in a statistically defensible manner [12; 13]. Therefore, the first aim was to assemble a workflow that would allow for the identification and quantification of proteins from a quantitative, cross-species proteomics experiment.

2. Although proteins can be viewed as the effectors of the cell [14], they still form part of a larger intricate biological system [1; 2; 15]. Biological databases offer ways to represent this system and allow for the contextualisation of omics data. Proteins can be contextualised in a variety of manners such as their sequence similarity to other proteins, their ontological descriptions, with what other proteins they interact, metabolic pathways they are associated with and where in the cell they are active. Moreover, networks have been shown to be an effective means for the analysis and contextualisation of omic data [4]. However, the datasets have typically been viewed in isolation, using only a small portion of the the contextual resources at a time. Thus, the second aim was to

construct a variety of networks using many contexts, either building the information into the structure of the network itself or as attributes.

3. The final aim was to use the constructed networks to mine and interpret the data for both existing and new hypotheses relating to proteins that may be of oenological interest.

1.3 Summary

The quantity and quality of proteomics data is likely to increase in keeping with trends set by other omics technologies [8]. Given the nature and scale of the data produced by MS-based proteomics experiments, proteomics data mining is an area that is well suited for network-based approaches. This thesis involves the development of a workflow that is able to identify and quantify proteins using LC-MS/MS data derived from a quantitative, cross-species experimental design. Moreover, this study offers network-based contextualisations of the proteins which allows for intuitive visualisation and hypotheses generation.

Bibliography

[1] Kitano, H.: Computational systems biology. *Nature*, vol. 420, no. 6912, pp. 206–210, 2002.

[2] Hartwell, L.H., Hopfield, J.J., Leibler, S. and Murray, A.W.: From molecular to modular cell biology. *Nature*, vol. 402, pp. C47–C52, 1999.

[3] Barabasi, A.-L. and Oltvai, Z.N.: Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, 2004.

[4] Aittokallio, T. and Schwikowski, B.: Graph-based methods for analysing networks in cell biology. *Briefings in Bioinformatics*, vol. 7, no. 3, pp. 243–255, 2006.

[5] Palumbo, M.C., Colosimo, A., Giuliani, A. and Farina, L.: Functional essentiality from topology features in metabolic networks: a case study in yeast. *FEBS letters*, vol. 579, no. 21, pp. 4642–4646, 2005.

[6] Li, J., Zimmerman, L.J., Park, B.-H., Tabb, D.L., Liebler, D.C. and Zhang, B.: Network-assisted protein identification and data interpretation in shotgun proteomics. *Molecular Systems Biology*, vol. 5, no. 1, 2009.

[7] Gstaiger, M. and Aebersold, R.: Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nature Reviews Genetics*, vol. 10, no. 9, pp. 617–627, 2009.

[8] Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J. and Aebersold, R.: The quantitative proteome of a human cell line. *Molecular Systems Biology*, vol. 7, no. 1, 2011.

[9] Deutsch, E.W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B. and Prazen, B.: A guided tour of the Trans Proteomic Pipeline. *Proteomics*, vol. 10, no. 6, pp. 1150–1159, 2010.

[10] Kohlbacher, O., Reinert, K., Gröpl, C., Lange, E., Pfeifer, N., Schulz-Trieglaff, O. and Sturm, M.: TOPP - The OpenMS proteomics pipeline. *Bioinformatics*, vol. 23, no. 2, pp. e191–e197, 2007.

[11] Cox, J. and Mann, M.: MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, vol. 26, no. 12, pp. 1367–1372, 2008.

[12] Snijders, A.P.L., de Koning, B. and Wright, P.C.: Relative quantification of proteins across the species boundary through the use of shared peptides. *Journal of Proteome Research*, vol. 6, no. 1, pp. 97–104, 2007.

[13] Pandhal, J., Snijders, A.P.L., Wright, P.C. and Biggs, C.A.: A cross-species quantitative proteomic study of salt adaptation in a halotolerant environmental isolate using 15N metabolic labelling. *Proteomics*, vol. 8, no. 11, pp. 2266–2284, 2008.

[14] Baerenfaller, K., Grossmann, J., Grobei, M.A., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., Zimmermann, P., Grossniklaus, U., Gruissem, W. and Baginsky, S.: Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science*, vol. 320, no. 5878, pp. 938–941, 2008.

[15] Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.-M., Cruciat, C.-M. and Others: Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, vol. 415, no. 6868, pp. 141–147, 2002.

Chapter 2

Literature review

2.1 Introduction

The technological advances of the past two decades have facilitated the development of high throughput (HTP) omic technologies that have greatly advanced our knowledge of biological systems. A cornerstone in molecular biology is the flow of information from DNA to RNA to Protein. The genome of an organism provides information on its protein-coding capacity, whilst the transcriptome and proteome of an organism gives information about the genes that are being expressed at that point in time. Furthermore, it is clear that complexity in a living system does not simply arise from the net total of molecules of which it is comprised, but rather the contextual combination of these components [1; 2; 3].

In the past, the fields of genomics, transcriptomics and metabolomics have received relatively more attention when compared to proteomics. This can mainly be attributed to the technical difficulties associated with large scale proteome analyses [4]. Although great knowledge and biological insight has been gained from studies of genes and their expression, equal knowledge on the level of the proteome is required to complete the flow of information.

The field of proteomics involves a variety of various technical disciplines. However, for the characterisation of entire proteomes the use of Liquid chromatography (LC) coupled to Mass Spectrometry (MS) has been shown to be a very valuable tool allowing for both protein identification and quantification [5]. Specifically, the use of LC coupled to tandem mass spectrometry in shotgun proteomics approaches has illustrated the capacity to identify thousands of proteins from a single sample in one experiment [6; 7; 8; 9]. Thus, large-scale, high-accuracy proteomics is now comparable in power to other more established omics technologies [9].

Omic technologies generate high volumes of data and present the opportunity for system-wide understanding of living organisms. However, it is often difficult to extract information and observe trends within the data. Furthermore, turning such omic data into biological knowledge requires linking the experimental data to relevant external databases and literature-based resources.

Proteins can be contextualised in a variety of manners such as sequence similarity to other proteins, ontological descriptions, protein interactions, metabolic pathways which they are involved in and where in the cell they are active. Although all these contexts may be useful by themselves for the interpretation of proteomics data, in order to interpret a subset of identified proteins as part of the larger system within which they function, integration of the contexts is needed in order to gain the best possible systems view.

Networks are ideal for the study and modelling of complex systems and are thus well suited to act as the scaffold for the integration of multiple layers of biological data. Furthermore, in biology, networks have been successfully applied in a myriad of ways for the interpretation of data [10] and provide a means by which a complex system of interlinked components can be represented as a collection of nodes and edges. The network-based representation of proteins within a biological network enables a considerable amount of data to be visualised without obscuring patterns, trends and relationships that exists in the data.

In the sections below, proteomics is discussed with focus on LC-MS/MS and the subsequent data analysis that follows such an experiment. Furthermore, the challenges of quantitative cross-species proteomics are also discussed. Next, the interpretation of LC-MS/MS data is discussed with focus on a selection of contextual resources that are currently available.

2.2 Proteomics

The proteome of an organism can be defined as the protein content of any given cell including their isoforms, splice variants, post-translation modifications, interacting partners as well as higher order complexes [11]. Furthermore, cells do not have a single fixed proteome [12]. Instead the protein complement of a cell is dynamic and is determined by a combination of factors such as its genome, the environment the cell is in at that point in time and even the history of the cell and what circumstances it has previously encountered [13]. Also the abundance of a protein may vary greatly, having possible dynamic ranges that span five-fold [13].

There are several reasons why is it important to have systematic and quantitative information on proteins: 1) Proteins can be viewed as the effectors of biological function within a cell, thus making any information about them or their expression levels important [14]; 2) Any deviations from genome-based predicted protein models could possibly be evident in the proteome; 3) mRNA expression levels are often used to infer subsequent protein levels, however, the two do not always correlate well [15]. Therefore it is necessary to identify and quantify proteins in an organism precisely.

The field of proteomics involves a variety of technical disciplines, however, for the purpose of this work we will focus specifically on shotgun proteomics and the role of Liquid Chromatography (LC) coupled to Mass Spectrometry (MS) in this technique.

2.3 Liquid Chromatography (LC)

Chromatography allows for the separation, identification and purification of compounds from complex chemical mixtures [16]. The manner with which a compound or analyte distributes between two immiscible phases can be described by its distribution or partitioning coefficient. For two immiscible phases A and B at a given temperature, the partitioning coefficient can be calculated as $K_{\rm d} = \frac{[A]}{[B]}$ where [A] is the concentration of the analyte in phase A and [B] is the concentration of the analyte in phase B. Chromatography exploits the fact that analytes in a complex mixture have differing partitioning coefficients [17]. All chromatographic systems consist of an immobilised stationary phase and a mobile phase. The mobile phase may be liquid or gas (always liquid in LC) and is passed over or through the stationary phase after the sample that needs to be separated has been applied [18; 17]. During chromatographic separation, the analytes associate with the two phases to varying degrees such that the differences in the partitioning coefficients of the analytes result in their separation. Two concurrent interactions affect the behaviour of an analyte during chromatographic separation. The first involves the interaction of the analyte with the stationary phase and the second involves interactions caused by processes such as diffusion which oppose the desired separation [18; 17; 19].

Column chromatography is a form of chromatographic separation where the stationary phase is either applied to the wall of the column as a thin film or coated onto small discrete particles which are then packed into the column [18; 19]. The mixture of analytes to be separated is applied and subsequently eluated with the mobile phase. The mobile phase, also called the eluent, moves through the column either by gravity or a pump [18].

2.3.1 High-performance liquid chromatography (HPLC)

HPLC is the modern culmination of advances in liquid chromatography and is a form of column chromatography. In addition to a column, an HPLC instrument also consists of a delivery system for the mobile phase which can deliver flow into the column at a constant rate, a sample injector system to deliver samples in a reproducible manner and a detector. The latter gives a continuous record of the analytes in the eluate as they emerge from the column. Detection of an analyte is typically based on a physical parameter such as visible or ultra violet absorption or fluorescence [20]. Furthermore, an HPLC may be coupled to MS for further analysis of the analyte.

An overview of an HPLC instrument and its components is shown in Figure 2.1A [20]. Samples are placed on a tray from which they are automatically injected onto the column. Pumping of solvent through the column is continuous and compounds are sensed by a detector as they leave the column. The result is a plot of detector signal over time called a chromatograph. The chromatograph consists of a series of peaks that represent the elution of analytes as shown in Figure 2.1B.

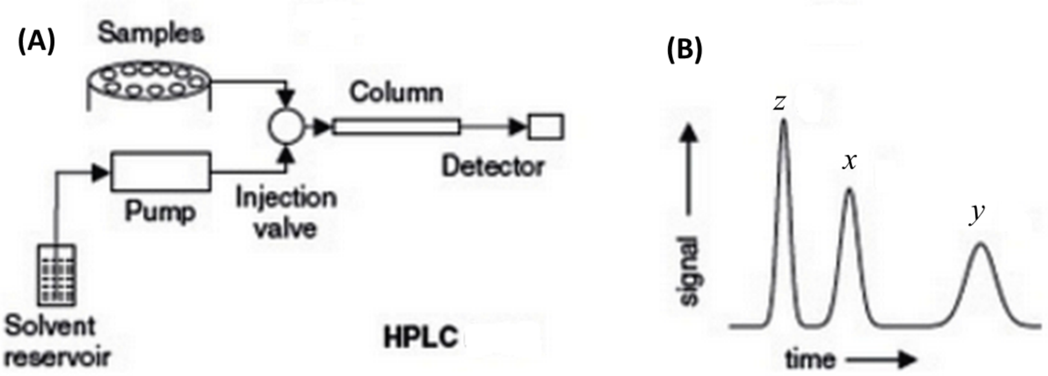


Figure 2.1: **Diagram of an HPLC instrument.** (A) shows an overview of the instrument and its components. (B) is a simplified diagram of the output with z, x and y representing resolved analytes. Figure adapted from [20].

Chromatographic separations depend upon the choice of stationary and mobile phases since these affect the partitioning coefficients of the analytes. Various combinations of mobile and stationary phases can be made and are defined by the type of equilibrium that forms between them. For the separation of peptides an adsorption equilibrium is often used.

Adsorption chromatography is based on the principle that a material has the ability to retain a molecule at its surface. This retention is due to non-ionic attractive forces such as hydrogen-bonding and van der Waals forces and these occur at specific adsorption sites. The adsorption sites have the ability to discriminate between types of molecules and may be occupied by either molecules of the analyte or the eluent. The proportion of analytes or eluant that occupies the adsorption sites is determined by their respective relative strength of interaction with the mobile phase. As eluent is constantly passed down the column, the separation of the analytes occurs as a result of differences in the binding strengths of the analytes [17].

2.3.1.1 Reverse phase high-performance liquid chromatography (RP-HPLC)

RP-HPLC [21] is a form of adsorption chromatography. In this experimental system the stationary phase is typically a non-polar n-alkylsilica based absorbent [22]. The mobile phase is relatively polar and consists of water, aqueous buffers, and organic solvents such as acetonitrile, methanol or mixtures thereof [17]. In RP-HPLC, the separation of analytes is predominantly determined by the composition of the mobile phase. The composition of the mobile phase may be altered as the chromatographic process proceeds and this technique is known as gradient elution. Gradient elution is typically applied for the separation of peptides and may involve alterations to the pH and/or gradual increases in the concentration of organic solvents [22]. In RP-HPLC, only non-polar interactions with analytes are possible since the stationary phase is basically inert [21; 22; 17].

2.3.1.2 Chromatographic Performance Parameters

The performance of a chromatographic separation can be measured by calculating measures such as the plate height and resolution. Chromatography columns can be viewed in terms of numerous adjacent zones in which there is sufficient space for an analyte to completely equilibrate between two phases. Each zone is termed a theoretical plate. The resolution of a chromatographic separation indicates the ability of the system to resolve one analyte peak from another [17].

2.3.1.3 Liquid Chromatography in Proteomics

The performance of a shotgun proteomics experiment is greatly affected by the ability to separate peptides as much as possible prior to MS. To this end, techniques such as HPLC are used due to their ability to separate tryptic peptides with high efficiency. Furthermore, approaches that make use of multiple LC steps coupled to MS have shown an increased ability to identify proteins [23; 24; 6; 7].

2.4 Mass Spectrometry

MS can be described as an analytical technique that can be used to identify the chemical composition of compounds on the basis of the mass-to-charge ratios of charged particles [25]. A generic mass spectrometer can be divided into three essential components, namely an ion source responsible for the ionisation of analytes in the sample, a mass analyser that can measure the mass-to-charge (m/z) ratio and a detector that can register the amount of ions at each m/z value [26]. Each one of the components are discussed in more detail below.

2.4.1 Ion Sources

The first step in MS analysis requires the ionisation of the sample. The development of electrospray ionisation (ESI) [27; 28] and matrix assisted laser desorption/ionisation (MALDI) [29; 30] represented a major breakthrough for MS and proteomics [5]. The development of these so-called 'soft' ionisation techniques solved the problem of generating ions from large non-volatile analytes such as proteins and peptides without inducing significant analyte fragmentation [5].

2.4.1.1 ESI

ESI involves the formation of charged droplets under the influence of an intense electric field and results in the production of gas phase ions from solutions containing dissolved ions [31]. The process can generally be divided into three steps, namely droplet formation, droplet shrinkage and desorption of gaseous ions.

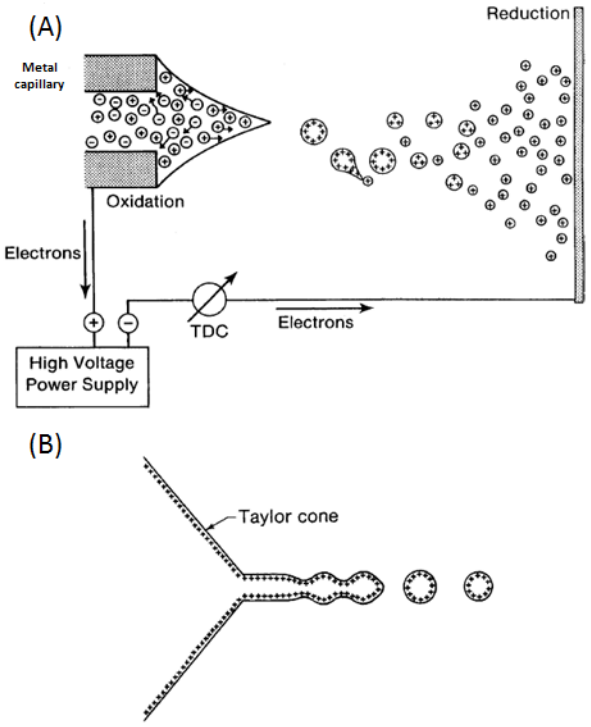


Figure 2.2: **Diagram depicting ESI and the components involved.** (A) shows the high voltage power supply, metal capillary and flow of electrons between the components. (B) is an enlarged view of the liquid cone at the tip of the capillary. Figure adapted from [31].

Figure 2.2A illustrates the ESI process and components of the device. Enrichment of positive electrolyte ions occurs at the meniscus of the solution due to the high electric field from the power supply and the needle-like dimension of the metal capillary. The cations will tend to migrate toward the counter electrode. As the net charge is pulled downfield and the liquid expands, the meniscus forms a cone from which a spray of small positive charged droplets is emitted. Droplet volume is reduced further by evaporation of the solvent. The continuous production of charged molecules is assisted by the electrochemical redox process. Figure 2.2B is an enlarged view of the liquid cone that forms at the tip of the capillary. This cone is called the Taylor cone. The least stable point of the cone is the tip, which subsequently extends into a filament where the charged droplets are formed [31; 27; 28; 32].

2.4.1.2 MALDI

Figure 2.3 gives an overview of the MALDI process. The first step in MALDI involves mixing the sample with excess matrix material such that the ratio between sample and matrix is in the range of 1:10000. The mixture is deposited on a sample plate and the solvent is evaporated which leaves sample-matrix crystals. An ultraviolet (UV) laser beam with a beam diameter of a few micrometers is directed at the sample for short pulses lasting only nanoseconds. This results in the simultaneous desorption and ionisation of both sample and matrix material and allows them to enter gas phase as intact ions. The matrix material serves as an absorbing medium for the UV light converting the incident light energy into molecular electronic energy and serves as a proton source for the ionisation of samples [29; 30; 33; 34].

2.4.2 Mass Analysers

For the purposes of proteomics work, there are several types of mass analysers in use including time-of-flight (TOF), quadrupole, Fourier transform ion cyclotron resonance (FT-ICR), ion trap and orbitrap. Each one of these instrument types makes use of different physical principles to obtain a mass-to-charge (m/z) ratio and are discussed in more detail below.

2.4.2.1 TOF Mass Analysers

TOF analysers use an electric field to accelerate ions which are then separated along a flight tube based on their different velocities. If the particles have the same charge, their velocity (and consequently the time they take to reach the detector) will depend only on the mass of the particle. Thus, the ions with the lowest mass will reach the detector first. Figure 2.4 shows the ions formed in the ion source. The ions are then separated in a field-free region of the flight tube before reaching the detector [35; 36; 37; 38].

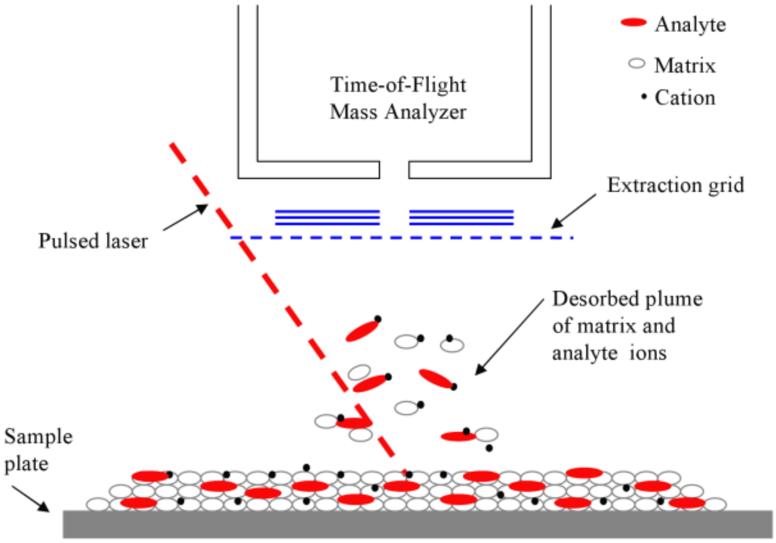


Figure 2.3: **Diagram depicting MALDI coupled to a Time-of-Flight mass analyser.** The sample plate is a matrix where each entry is a sample spot. Sample molecules are ionised by gas-phase proton transfer from the matrix. The immediate area of laser excitation forms a plume, consisting of matrix and analyte ions, which directly enters the high-vacuum of the mass spectrometer. Figure taken from [34].

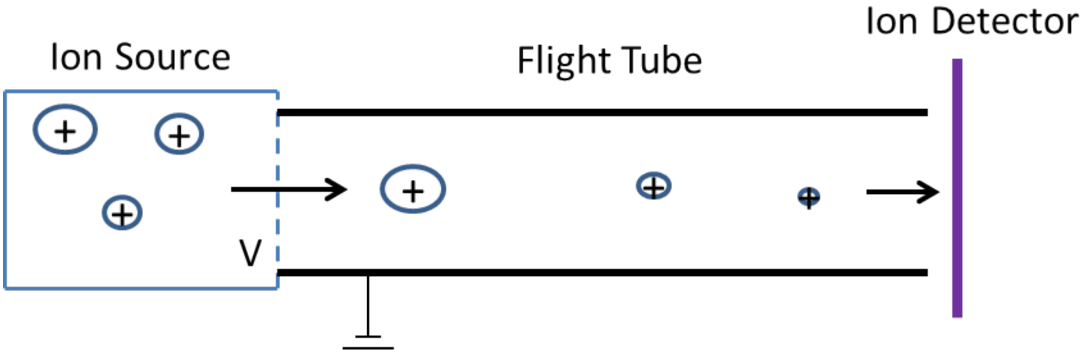


Figure 2.4: **Diagram of a TOF analyser**. Figure adapted from [38]

2.4.2.2 FT-ICR

A FT-ICR ion trap makes use of a magnetic field to trap ions inside an orbit as shown in Figure 2.5. When a moving charge enters a magnetic field it experiences a centripetal force which places the ion into orbit. The force on the ion due to the magnetic field is equal to the centripetal force on the ion. Detectors are placed at fixed positions in the mass analyser where they capture the electrical signal of ions which pass near or over them, thus producing a periodic signal. The m/z of the ion determines the frequency of its cycle which can be deconvoluted by performing a Fourier transform on the signal [38].

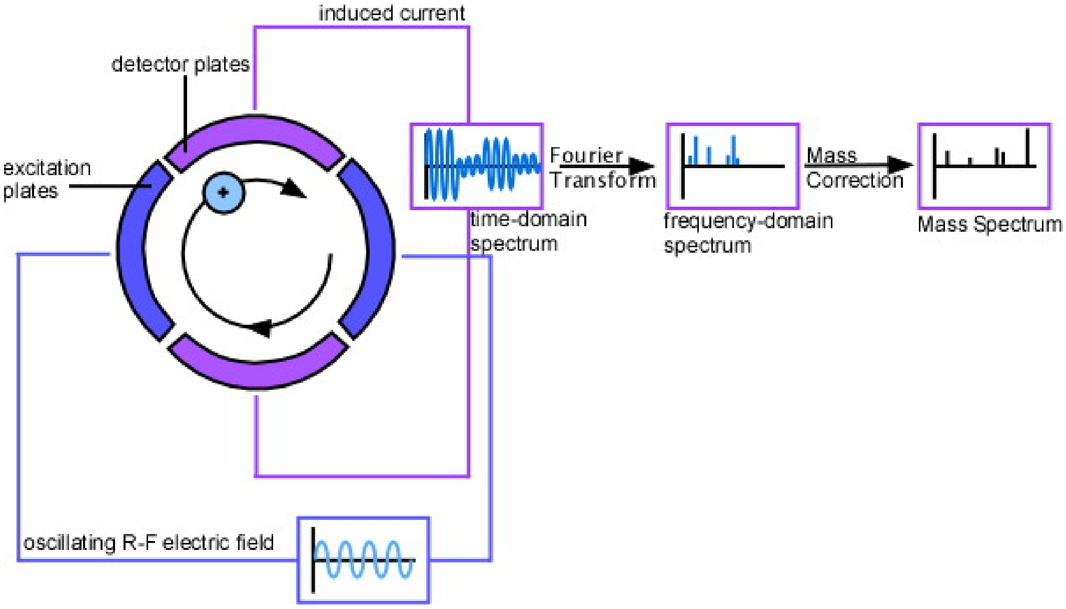


Figure 2.5: **Diagram of a FT-ICR mass analyser**. Figure taken from [38]

2.4.2.3 Linear Quadrupole Mass Analysers

Linear quadrupole mass analysers or filters consist of four parallel rods through which an electrical current is passed. As shown in Figure 2.6, a radio frequency (RF) quadrupole field is created between the four parallel rods and this electrical field is oscillated in a time-varying manner such that the paths of ions passing through the RF can be stabilised or destabilised. Thus, at a given time, ions of a desired m/z are allowed to pass through to the detector on a stable trajectory [39; 40; 41].

2.4.2.4 Quadrupole Ion Trap

The quadrupole ion trap is the three dimensional form of the linear quadrupole mass filter described in Section 2.4.2.3 and employs similar principles for operation. The quadrupole ion trap also uses an electric field for the separation of the ions by mass to charge ratios. However, in an ion trap, ions of desired m/z are held and then ejected. Furthermore, in the linear quadrupole mass analysers forces on the ion is in two dimensions, whereas in the ion trap the ion experiences forces in three dimensions.

As illustrated in Figure 2.7, the space in which the ions are trapped is defined by three electrodes, namely the central ring electrode and two adjacent

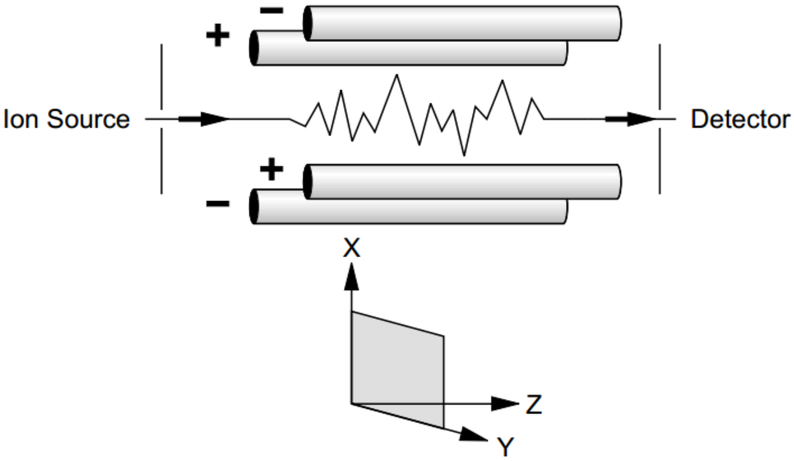


Figure 2.6: **Diagram of a Quadrupole mass analyser**. Figure taken from [40].

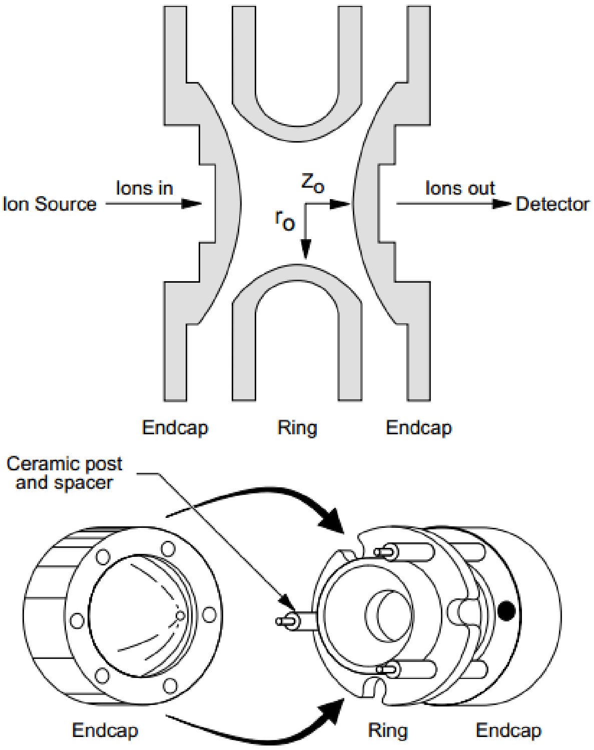


Figure 2.7: **Schematic of a quadrupole ion trap**. Figure taken from [40].

endcap electrodes. All three electrodes have hyperbolic surfaces. The apparatus is radially symmetrical with r_O and z_O representing the dimensions of the apparatus as shown in Figure 2.7. Potentials are applied to all the electrodes with the ring electrode having an alternating potential of constant RF but variable amplitude. The result is an electric field in the cavity of the analyser where ions of certain m/z values will orbit in the space. The orbits of ions with higher mass become more stable as the potential is increased and conversely the orbits of ions with lower mass become less stable and can then be ejected onto a detector [40; 41].

2.4.2.5 Orbitrap Mass Analyser

An orbitrap mass analyser differs from other types of mass analysers in the sense that it does not use magnets or RF to place ions in orbit. Instead ions are electrostatically trapped around a spindle shaped electrode (as shown along the z-axis in Figure 2.8) in the center of the chamber. The ions are attracted electrostatically toward the central electrode, however, a centrifugal force also arises from the initial tangential velocity of the ions. This centrifugal force compensates for the electrostatic force that an ion encounters. The result is that ions move in complex spiral patterns around the central electrode [42].

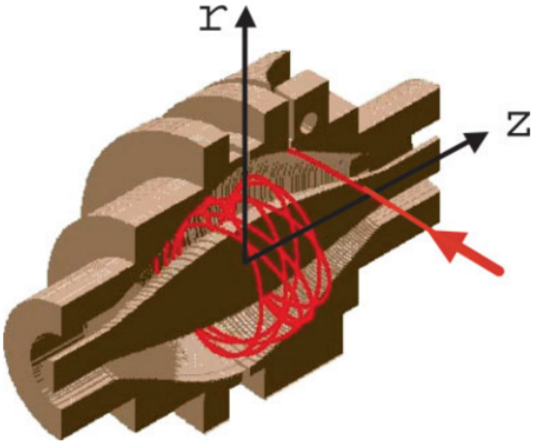


Figure 2.8: **Cutaway diagram of orbitrap mass analyser.** The red arrow indicates the point where ions are injected into the orbitrap. Figure adapted from [43].

Ion oscillation induces a signal voltage which can be picked up by outer electrode detector plates which are able to represent the oscillation of an ion in terms of an image current. The m/z of an ion is related to the image current that the detector will record. A Fourier transform is applied to the signals provided by the recorded image currents in order to form mass spectra [42].

2.4.2.6 Hybrid instruments

When several mass analysers of different types are used to construct a mass spectrometer, it is referred to as a hybrid instrument [38]. Figure 2.9 is a diagram of one such hybrid instrument that is commonly used for LC-MS/MS.

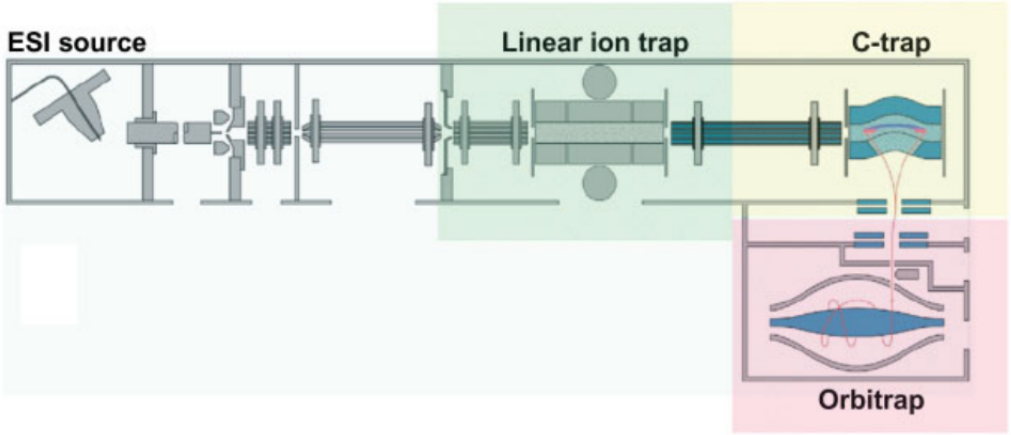


Figure 2.9: **A diagram of the LTQ Orbitrap.** This is an example of a hybrid instrument and consists of three main parts. The first section after the source is a linear ion trap which is able a to detect MS and MS^n spectra. In the C-trap component ions are accrued and their energy dampened after which they are injected into the orbitrap. Figure taken from [44].

2.4.3 Detectors

The third component of an MS device is the detector, which generates a record of the ions in the form of a mass spectrum. The mass spectrum is a frequency histogram indicating the intensity of an ion on the y-axis and the m/z value of the ion on the x-axis. The m/z is the relationship of the mass of a given ion, divided by the number of the charges that it has [45].

2.5 MS in Proteomics

MS is currently seen as the most valuable tool in proteomics [5]. MS, with regard to proteomics, can be used for protein identification, quantification, protein profiling and to study protein interactions and protein modifications [46].

Following electrophoretic or chromatographic separation of sample components, peptide mass fingerprinting (PMF) [47] and sequencing by tandem mass spectrometry [48] are two major methods used for the identification of proteins [49]. An overview of both methods is presented in Figure 2.10.

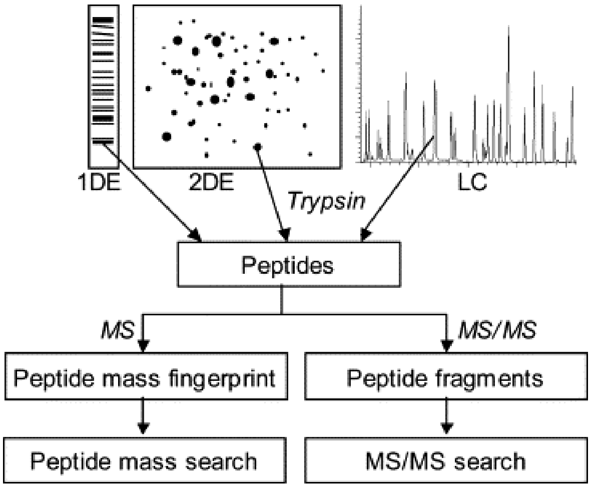


Figure 2.10: **PMF and sequencing by tandem MS.** Two common methods for protein identification via MS. Figure taken from [49].

As shown in Figure 2.10, proteins are separated by gel electrophoresis or liquid chromatography. A sequence-specific endoprotease such as trypsin is then used to cleave the proteins into peptides. Following the digestion, molecular masses are determined for the peptides. The obtained peptide masses are then compared to theoretical peptide masses of proteins.

2.6 Shotgun Proteomics with LC-MS/MS

Shotgun proteomics refers to a technique where a complex mixture of proteins are digested with an enzyme such as trypsin to give a mixture of peptides. This mixture is then separated using LC and the peptides are sequenced by tandem mass spectrometry (MS/MS). Automated database searching is done to identify proteins [50]. Identifying proteins using tandem mass spectrometry coupled with liquid chromatography is collectively known as LC-MS/MS.

2.6.1 LC-MS/MS

LC-MS/MS is a method that involves the use of more than one mass analyser in tandem with one another [48]. In a typical MS/MS experiment, a precursor ion is selected based on its mass by mass analyser 1 (MS1) and focused into a collision region that precedes a second mass analyser (MS2). The mass analysers can be arranged either in space as is the case with sector and triple quadrupole instruments, or in time as is the case with trapping instruments such as an ion trap. Key to LC-MS/MS is a process termed Collision Induced Dissociation (CID) [51]. This process involves the introduction of an inert gas into a collision zone. In the collision zone the inert gas molecules collide with the precursor ion. This process produces so called product ions from the precursor ion and the product ions can then be mass analysed by MS2 [48; 51].

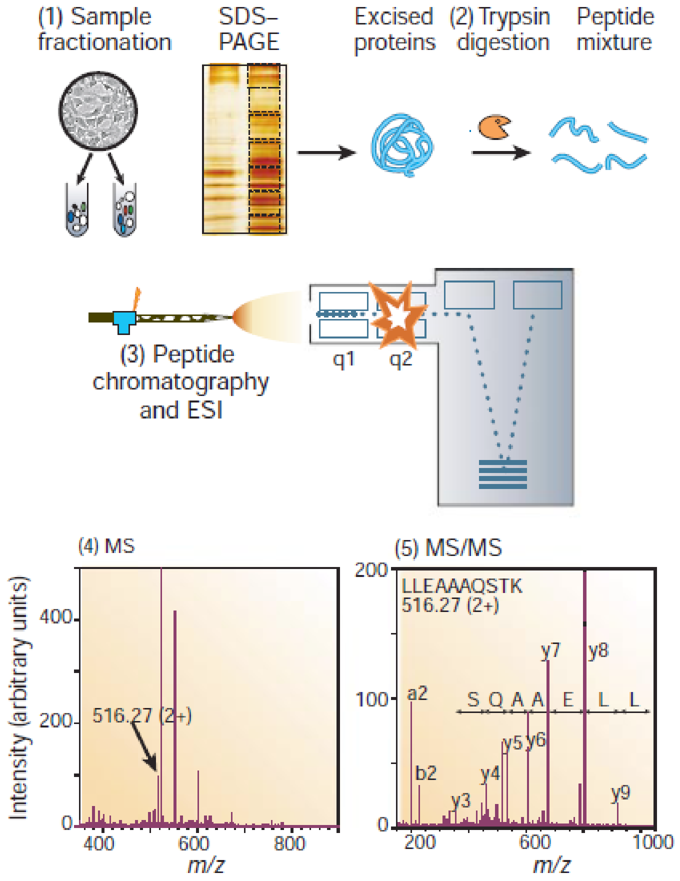


Figure 2.11: **The workflow and layout of a generic MS proteomics experiment consisting of five steps.** Figure taken from [46].

Figure 2.11 illustrates the five steps of a generic LC-MS/MS proteomics experiment: 1) Proteins are isolated from the biological sample; 2) Proteins are digested by enzymes such as trypsin; 3) Peptides are separated by HPLC which is followed by ionisation of the elute in the ion source of the mass spectrometer; 4) A mass spectrum of the peptides that eluted at this specific point in time is generated. This spectrum is known as the MS1 spectrum. A computer then generates list of these co-eluting peptides; 5) Each peptide ion (as identified by the first mass spectrum) is subject to energetic collision with an inert gas. This spectrum is known as the tandem or MS/MS spectrum.

Both the MS1 and MS/MS spectra for a single peptide ion are acquired within approximately one second. These spectra are then stored and used to match against a protein sequence data base. The desired outcome of the experiment is the identity of the proteins that constituted the original sample [52].

2.6.2 Sample Preparation

A number of variables influence the type and quantity of proteins that are extracted from a sample. These variables include the size of proteins to be extracted, their cellular location, the point of extraction relative to the cell's growth phase, the type of solvents that are used and extraction temperature [53]. Several methods for protein extraction from yeast have been described [54; 55; 56; 57]. Steps for selection of proteins occurring within certain cellular fractions may also be conducted. These fraction may be from organelles inside the cell or of proteins found in the growth media.

In order to prepare proteins for detection using MS, samples are often further separated into fractions based on their physical and chemical properties using techniques such as two dimensional sodium dodcecyl sulfate polyacrilamide gel electrophoresis (2D-SDS-PAGE) [58]. 2D-SDS-PAGE involves the separation of proteins based on their isoelectric point, followed by separation based on mass. Proteins may then be excised from the gel for digestion with trypsin and subsequent MS analysis.

2.6.3 Peptide Properties

The relatively large size and characteristics of intact proteins make direct protein sequencing with MS-based methods difficult [59]. Consequently, for proteome sequencing using LC-MS/MS, proteins are first cleaved into smaller peptides before analysis. Although peptides are relative more amenable to MS-based sequencing than proteins, they are highly variable as each peptide has its own set of physico-chemical properties that make it unique. These properties entail physical and chemical features such as molecular weight, isoelectric point and hydrophobicity of the peptide. Peptide properties have an impact on the performance of nearly every aspect of the LC-MS/MS method including the chromatographic and ionisation steps, fragmentation during CID and all subsequent events including data analysis and interpretation [60]. Further complexity is introduced by the occurrence of post-translational modification (PTM) of proteins. PTM is a form of processing used to control protein activity and involves chemical alteration to the structure of a protein such as phosphorolation, glycosylation and acetylation [61]. PTMs result in a difference in the mass of the protein relative to the molecular weight of the protein calculated from only the amino acid sequence and thus lead to further heterogeneity in the peptide population [62]. Moreover, there is also variability arising from sample preparation which includes factors such as differential degradation rates of proteins and peptides, the digestion efficiency of the cleavage enzyme and preparation-induced modifications such as methionine oxidation [63].

2.6.4 Quantification with LC-MS/MS

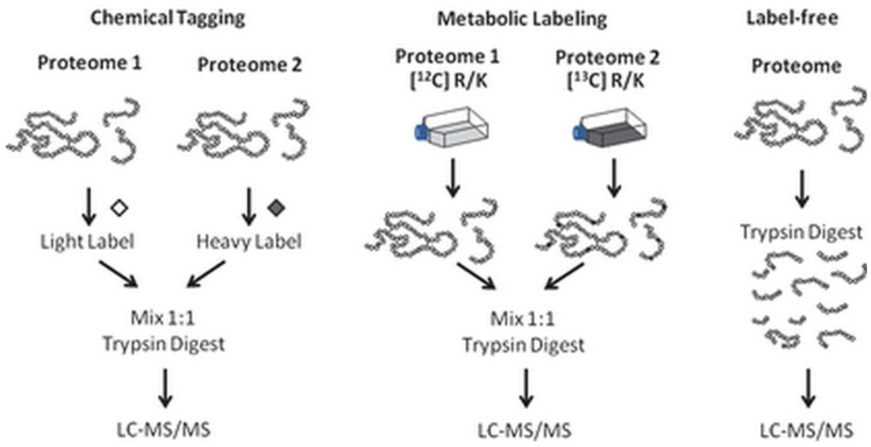


Figure 2.12: **Experimental strategies for global quantitative MS-based proteomics.** Figure adapted from [60].

Figure 2.12 shows experimental strategies that may be used for global quantitative proteomic strategies. These include chemical tagging strategies such as isobaric tagging, *in vitro* metabolic labelling as well as label-free techniques. Isobaric tagging may be performed with methods such as iCAT [64], iTRAQ [65] and Tandem Mass Tags (TMT) [66].

Tandem Mass Tagging involves the attachment of specific isobaric reagents to the primary amines of peptides from different samples. As shown in Figure 2.12, the samples are mixed after attachment of the isobaric labels. Labelled peptides of the same mass will eluate from the column at the same time and subsequently appear as a single peak in the MS spectrum.

Because the samples are mixed (Figure 2.12), the same peptide from each sample appears as a single peak in the MS spectrum. When a tagged peptide is subjected to CID, the peptides fragment to release reporter ions. The peak area of these reporter ions can then be captured and used to calculate the relative abundance of peptides between different samples [67].

2.6.5 Proteomics Across Species

Given the success of MS-based proteomic approaches on samples from individual species, it is only logical to want to extend the application to investigate the proteomes of different species under the same conditions. However, when cross-species proteomics experiments are attempted it adds complexity. A few groups have attempted these types of cross-species proteomics experiments either by using PMF, LC-Tandem MS or a combination of methods with varying degrees of success [68; 69; 70; 71; 72].

One of the requirements for a cross-species proteome comparison is that there must be some degree of sequence similarity between the species [69; 73]. However, even a high degree of overall sequence similarity does not ensure success seeing that even a single non-synonymous polymorphism between two strains of the same species may alter the properties of a peptide. For example, if a glycine residue is substituted for a tryptophan residue, the peptides will still appear to have a high degree of similarity, however, this single amino acid change leads to a mass difference of 129.06 Daltons [74]. Furthermore, these sequence changes may alter the physico-chemical properties of the peptide which consequently affects downstream processes as discussed in Section 2.6.3.

Several studies have tried to solve the problems encountered with cross-species proteomics [68; 71; 72; 75]. Approaches include combining MS-based protein identifications with other data such as amino acid composition, estimated intact protein mass and isoelectric point. These studies have shown that the number of identifications significantly improves using such combined approaches [68; 75]. Furthermore, it is possible to establish relative protein abundances using shared peptides [71; 72]. However, the availability of species or strain-specific sequenced genomes of the target species is important.

2.7 MS Data Analysis

After MS analysis, all mass spectra of the peptide fragments are written to a file or loaded into a database. The subsequent analysis may vary, although most approaches involve similar steps [76]. Figure 2.13 gives an overview of the steps involved in the analysis of spectra after the propriety peak detection and alignment steps. Each step is discussed in more detail below.

2.7.1 Conversion of MS Spectra

The first step entails the conversion of the spectra to a usable format. Most mass spectrometer vendors make use of proprietary data format for storage of the data recorded in the mass spectrometer, however, this creates difficulty for the use and development of software and downstream analysis of data. Hence, the first step in an MS data analysis workflow is to convert the data from its proprietary format to an open XML-based format [76].

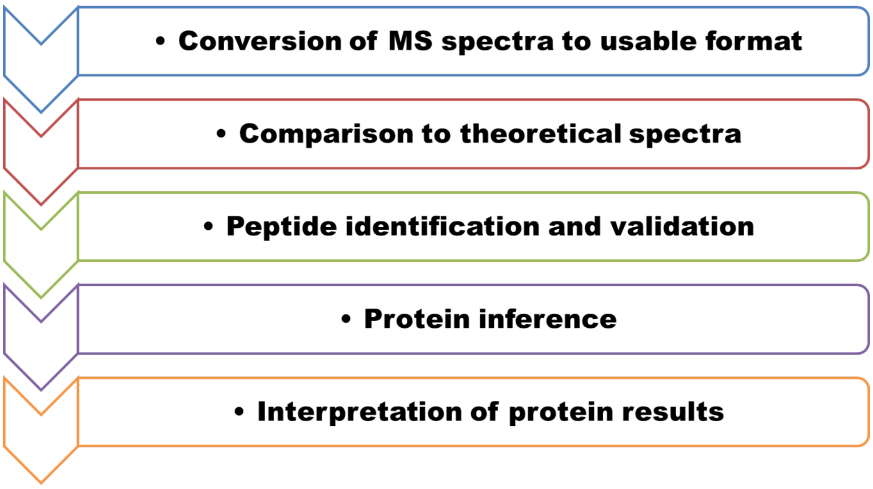

• Conversion of MS spectra to usable format
• Comparison to theoretical spectra
• Peptide identification and validation
• Protein inference
• Interpretation of protein results

Figure 2.13: **Generic data analysis workflow for LC-MS/MS data**

2.7.2 Spectral Matching

The second step involves matching the observed spectra to theoretical spectra. The theoretical peptide sequences are obtained by performing an *in silico* trypsin digest on a given list of protein sequences which is followed by matching the observed spectra to theoretical spectra using a spectrum interpretation algorithm. There are many programs that perform this function of which SEQUEST [77], Mascot [78] and X! Tandem [79] are well known examples. Most of these sequence searching programs make use of statistical scoring mechanisms which take into account the size of the sequence database being searched as well as the likelihood that the top spectrum match is a random event [47; 76; 79].

2.7.3 Statistical Validation of Peptide Assignments

The third step entails the statistical validation of the peptide assignments using the scores for the peptide assignments made by the sequence searching programs. The distributions of the scores are modelled as a mixture of two populations, namely correct and incorrect peptide assignments. Based on this mixed model, probabilities of correctness as related to all identifications can be calculated [76]. PeptideProphet [80] is a tool that implements such an approach.

2.7.4 Protein Inference

The fourth step deals with the identification of proteins based on the peptide evidence. When a peptide maps uniquely to a protein the identity of the protein can be inferred in a relatively straightforward manner. However, this is often not the case, especially for higher eukaryotic organisms, due to the frequent occurrence of multiple protein isoforms, protein families, peptides mapping to multiple proteins and database redundancies. ProteinProphet [81] is a program that uses probabilities of peptides associated with a given protein to calculate a probability of identification of that protein. It makes use of the Occam's razor principle to reduce the protein list to a minimal set of proteins that can explain all the observed peptides [81].

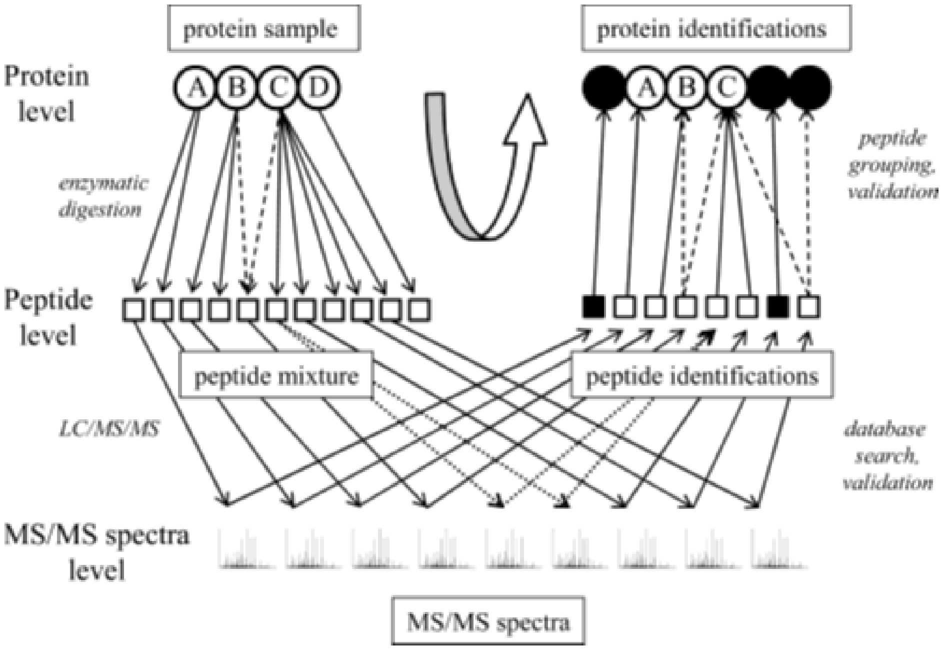


Figure 2.14: **Diagram showing an overview for protein identification from LC-MS/MS data.** The open circles A to D represent proteins in a mixed sample. The peptides are represented by the open squares. Figure taken from [81].

Figure 2.14 shows how protein identifications are made using peptide level information in a typical LC-MS/MS experiment of a complex protein mixture. The open circles A to D represent proteins in a mixed sample. Each sample protein is enzymatically cleaved into smaller peptides represented by the open squares. A peptide may be unique to a protein or it may be shared between sample proteins as indicated by the dashed arrows extending from sample proteins B and C. The peptides are then subjected to LC-MS/MS to produce spectra. Some peptides may be selected for fragmentation multiple times as indicated by the dotted arrows, whilst other peptides may not even be selected once. All the acquired MS/MS spectra are searched against the sequence database and assigned a best matching peptide. The best match may not be correct (as indicated by the black squares) and thus require validation.

Proteins are then inferred from the validated peptides. The open circles A, B and C represent proteins identified from the original sample and black circles represent incorrect protein identifications.

There are many tools available that can be pieced together to form a custom workflow to perform the tasks in Figure 2.13. However, there are only a few packages available that aim to provide a single environment that allows one to perform all of the steps required. One such workflow is the Trans Proteomic Pipeline (TPP) [82]. Other similar packages include the openMS proteomics pipeline (TOPP) [83] and Max Quant [84].

The TPP is a comprehensive suite of software tools that facilitates and standardises the analysis of LC-MS/MS data. It includes software tools for the representation and visualisation of MS data, peptide identification, validation, quantification and protein inference [82].

2.7.5 Protein Interpretation

The fifth step involves interpreting the proteins identified by the LC-MS/MS workflow. Even after protein assignments have been made, the output of a typical shotgun proteomics experiment may still be daunting. At this point, the output is usually in the form of a large spreadsheet consisting of hundreds of rows and several columns with each row representing a protein and columns representing identification attributes.

Although the proteins are identified as independent entities, they function as part of a greater connected system. Furthermore, most biological functions arise from interactions between proteins and other molecules [2]. Shotgun proteomics experiments are capable of offering quantitative snapshots of entire proteomes across species and experimental perturbations. However, to interpret the identified proteins as part of a complex system, context is needed.

2.8 Contextualisation

HTP-omic technologies generate high volumes of data and present the opportunity for system-wide understanding of living organisms. However, it is often difficult to extract information and observe trends within the data. Furthermore, turning HTP data into biological knowledge requires linking the experimental data to relevant external databases and literature-based resources.

Proteins can be contextualised in a variety of manners such as their sequence similarity to other proteins, their ontological descriptions, with what other proteins they interact, in what metabolic pathways they form part and where in the cell they are active. Each context is discussed in more detail below.

2.8.1 Orthology Detection

Determining sequence similarity between proteins may offer insight into their functions, evolution, cellular locations, post translational modifications and regulation. OrthoMCL was developed to meet the challenges of identifying orthologous groups across multiple taxa.

Using Markov clustering [85], this pipeline groups putative orthologs, co-orthologs and paralogs. Homologous proteins can be divided into two major types, namely orthologs and paralogs. Orthologs differ from paralogs in that they evolved from a common ancestor by speciation whilst the latter originate from duplication events [86; 87].

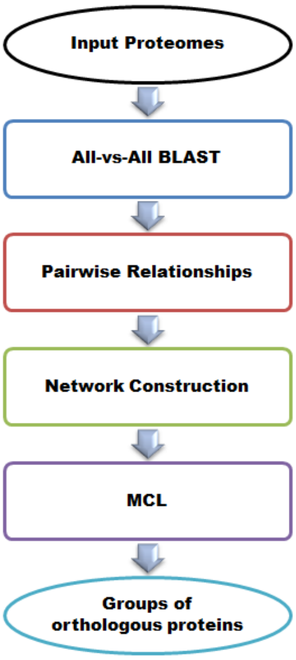
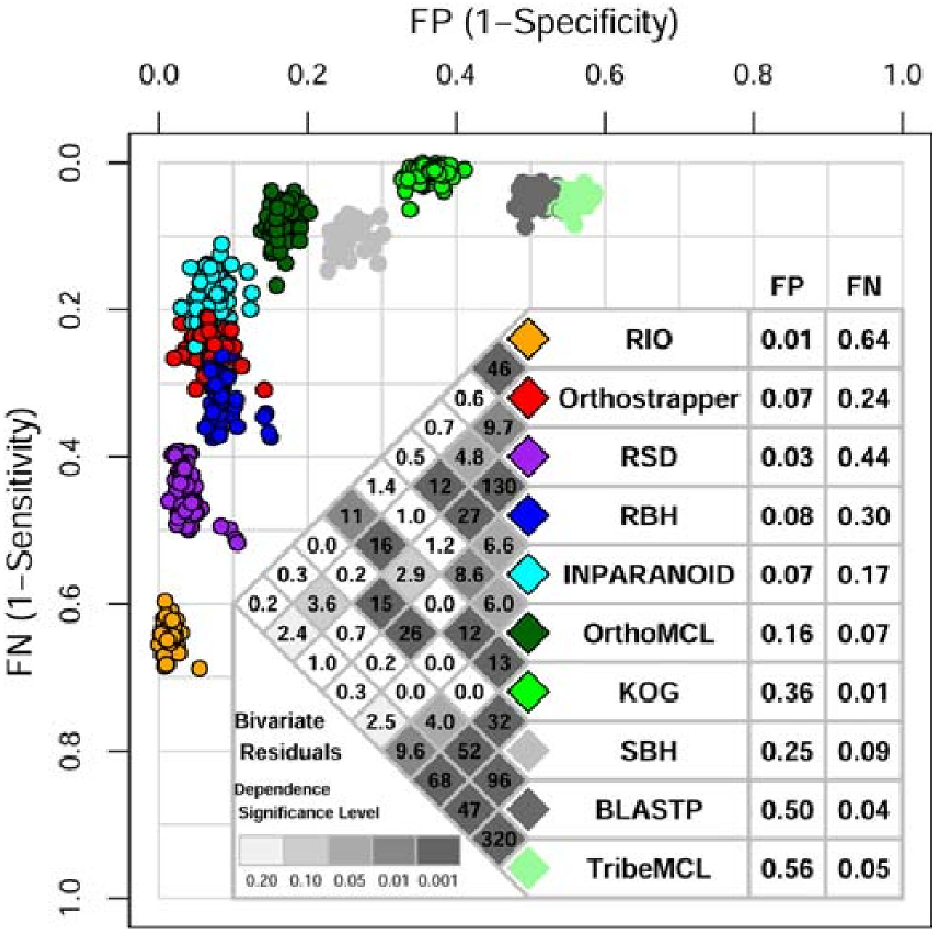


Figure 2.15: **General steps in the OrthoMCL workflow.**

Figure 2.15 shows the major steps in the OrthoMCL workflow: 1) OrthoMCL uses proteome sequence as input; 2) An all-vs-all BLAST is used to determine the amount of similarity between all the input proteins; 3) The best reciprocal BLAST hits are grouped into putative orthologs, co-orthologs and paralogs which are used to construct a network; 4) MCL is applied to the network in order to form groups of orthologous proteins. MCL includes a parameter called the inflation index which the user sets in order to control the granularity of the resulting clusters.

The performance of various orthology detection methods on a eukaryotic dataset was evaluated using a statistical technique called Latent Class Analysis

(LCA) [88]. When different orthology detection methods are compared, information regarding instances when methods are in agreement or dis-agreement can be obtained. LCA can utilise such comparative outputs to deduce information regarding sensitivities and specificities.



	FP	FN
RIO	0.01	0.64
Orthostrapper	0.07	0.24
RSD	0.03	0.44
RBH	0.08	0.30
INPARANOID	0.07	0.17
OrthoMCL	0.16	0.07
KOG	0.36	0.01
SBH	0.25	0.09
BLASTP	0.50	0.04
TribeMCL	0.56	0.05

Figure 2.16: **Sensisitivity vs specifity plot for various orthology detection softwares.** The x-axis of the figure represents false positive rates and the y-axis represents false negative rates. Figure taken from [88].

Figure 2.16 shows the results of a performance evaluation of various orthology detection methods [88]. Two algorithms, namely OrthoMCL and INPARANOID [89] showed the best overall balance of sensitivity and specificity with both these criteria greater than 80%. Another factor that sets OrthoMCL apart is its ability to cluster orthologs from multiple species, whilst INPARANOID is only able to identify orthologs across two species. OrthoMCL also outperforms the manually curated KOG database [90] and TribeMCL [91] when it comes to the within-group consistency of protein function and domain architecture.

2.8.2 The Gene Ontology

The Gene Ontology (GO) was created with the goal of producing a dynamic, controlled vocabulary that can be applied to all organisms even as new knowledge about the genes and gene products arise [92]. This controlled vocabulary may be used to describe the roles of genes and gene products in any organism.

GO is comprised of three independent ontologies, namely Biological Process (BP), Molecular Function (MF) and Cellular Component (CC). The ontologies themselves are trees or DAGs (directed acyclic graphs). The level of the node within the tree indicates how specific the GO term is. Each node in GO links to other kinds of external information such as gene and protein keyword databases that give more detailed information about the gene or its products.

Nodes within the BP ontology specify a biological process that a gene or its product contributes to. A biological process can be defined as ordered assemblies of molecular functions and often involves the chemical or physical transformation of molecules. Nodes within the MF ontology give information about the biochemical activity of a gene product. While BP and MF describe processes and functions, nodes within the CC ontology point to the places in the cell where a gene product is active.

The existence of a controlled vocabulary such as GO allows for the automated transference of biological annotations, via gene and protein sequence similarity, from model organisms to organisms that do not yet have the same level of information available.

2.8.2.1 Gene Ontology Enrichment

Identifying a subset of overrepresented or enriched GO terms from a larger set is one approach to narrow the focus of investigation. The Gene Ontology Enrichment Analysis Software Toolkit (GOEAST) is a web-based tool (available at http://omicslab.genetics.ac.cn/GOEAST/) that allows one to find significantly enriched GO terms among a given list of genes . GOEAST employs a hypergeometric test to determine which GO terms are significantly enriched in the database.

2.8.3 KEGG

The Kyoto Encyclopedia of Genes and Genomes (KEGG) project represents an effort to link genomic information with higher order functional assignments (available at http://www.genome.jp/kegg/). Making functional assignments is an ongoing process that requires the linking of a set of genes in the genome with a network of interacting molecules in the cell.

KEGG consists of three databases, namely 1) "Pathway" for representation of higher order functions in terms of a network of interacting molecules. 2) "Genes" which consist of a gene catalog for all the completely sequenced genomes and some partial genomes. 3) "Ligand" which is a collection of chemical compounds in the cell.

The KEGG metabolic pathway database is comprised of manually drawn pathway maps that include entries for carbohydrate, energy, lipid, nucleotide, and amino acid metabolism as well as glycan biosynthesis and metabolism of cofactors and vitamins. These pathways provide useful context for proteins by supplying information about their enzymatic function, which reactions they are involved in and what metabolites are affected.

2.8.4 BioGRID

The Biological General Repository for Interaction Datasets (BioGRID) is a database containing physical and genetic interactions (available at http://thebiogrid.org/). The current version of BioGRID is comprised of 749912 protein and genetic interactions from several major model organisms and represents a total 43149 publications [93].

The genetic interactions reveal functional relationships between and within regulatory modules [2], whilst the physical interactions provide information about the proteins' direct interactions with each other. Given that proteins can be viewed as the effectors of biological function within a cell [14], and nearly all cellular responses involve protein interactions [94], a resource such as the BioGRID is useful for the contextualisation and interpretation of proteomics data.

The BioGRID repository seeks to collate interaction data from a variety of platforms and experiments in a consistent and well annotated format. The first public release of BioGRID, originally termed GRID, consisted of interaction data generated from HTP two-hybrid assays and mass spectrometric platforms performed on samples from *S.cerevisiae* only. The BioGRID has since evolved into a resource for HTP interaction data from other species and now also contains numerous manually curated interactions sourced from focused studies or from the literature.

Protein-protein interactions contained within the BioGRID can be subdivided into gene-based interactions and physical interactions, of which each interaction is determined by various methods [95; 93]. The experimental techniques used to detect protein-protein interactions are listed and described in more detail in Section 2.11.1 of the supplementary materials. Experimental techniques such as the yeast two hybrid (2-H) system for detecting pairwise protein interactions [96; 97; 98] and the analysis of purified protein complexes via MS [99; 3] can both be viewed as HTP hypotheses generating tools. More recent platforms such as the synthetic genetic array (SGA) and molecular barcode (dSLAM) methods bring HTP capability to enable detection of synthetic lethal genetic interactions [100; 101]. The yeast 2-H system and the analysis of purified protein complexes via MS are discussed in more detail below.

2.8.4.1 The Yeast 2-H system

The yeast 2-H method is also known as an interaction trap. It is used to detect pairwise protein interactions and can be applied in a HTP manner. The system works by taking advantage of the attributes of the *Saccharomyces cerevisiae* GAL4 protein [102]. GAL4 is a transcriptional activator required for the expression of genes encoding enzymes for galactose utilization and, like other transcriptional activators, it is a modular protein that requires both DNA-binding (BD) and activation domains (AD). The 2-H assay operates by expressing two fusion proteins in yeast namely, the "hunter" and the "bait". The "hunter" protein is the possible binding partner fused to a yeast AD and the "bait" is the protein of interest which is fused to a yeast BD. A yeast strain is transformed with both constructs and the appropriate upstream activating sequence which is in close proximity to the reporter gene [102]. Thus, if the reporter gene is expressed it means that there was interaction between the "hunter" and the "bait" proteins [103].

2.8.4.2 MS-based Analysis of Purified Protein Complexes

Purified protein complexes may be identified by MS using a technique known as high throughput mass spectrometric protein complex identification (HMS-PCI) [99]. This technique allows protein-protein interactions to be observed directly using a tagged bait protein which is then followed by identification of the interacting partners via MS. The main steps of the assay include the following [63]: 1) cDNA of interest is cloned into a vector that equips it with an epitope tag; 2) The vector with the cDNA is then transformed into the cell of interest. The expressed protein here constitutes the "bait"; 3) Affinity purification using an antibody against the epitope is used to purify the lysate obtained from the lysed cells; 4) Competitive elution using a peptide that encodes the epitope is used to obtain the proteins that were bound specifically to the bait protein; 5) The proteins released via the competitive elution are then separated by gel electrophoresis followed by identification using MS.

Despite their hypotheses generating utility, HTP interactome datasets are often prone to high false positive and false negative rates [104; 105]. Each of the interaction detection methods have different biases and scoring systems and thus it is important to know what type of interaction is found, how it was derived and if there are multiple lines of evidence to support it.

2.8.5 Protein Targeting and the Fungal Secretome

Proteins possess inherent signals that dictate their transport and localisation within the cell [106]. Protein targeting is the process that governs the movement of proteins within the cell and this routing system relies on a variety of targeting signals [107].

Signal peptides (SPs) represent one type of targeting signal. This class of targeting signal is comprised of short transient peptides and is typically found at the amino terminus of the secreted protein [108]. SPs are typically comprised of 15-20 hydrophobic amino acid residues and are cleaved off during translocation of the protein across the membrane. Proteins containing an SP direct the ribosomes to the rough endoplasmic reticulum (ER) in order for polypeptide synthesis to be completed [109].

Knowing if a protein is targeted for transport or further processing can thus aid in the contextualisation and biological interpretation of proteomics data. The properties of SPs make them amenable to computational prediction and consequently a variety of tools exist that are able to identify proteins that are targeted for membrane translocation. A selection of the available tools include PrediSi [110], SPEPlip [111], Signal-CF [112], Signal-3L [113], Signal-BLAST [114] and SignalP-4.1 [115]. There are also several repositories for signal peptides of which SPdb [107] and FunSecKB are examples. Two of the repositories, namely SPdb [107] and FunSecKB [116] and one predictive tool, namely SignalP-4.1 [115] are respectively discussed in more detail below:

SPdb: SPdb is a repository of experimentally derived and computationally predicted signal peptides for archaea, prokaryotic and eukaryotic organisms [107]. In its current release (SPdb 5.1), there are 27433 entries, of which 2512 are experimentally verified signal sequences and 24921 are unverified signal sequences. SPdb gathers information from two sources, namely Uniprot protein sequence database [117] and the EMBL nucleotide sequence database [118].

FunSecKB: FunSecKB consists of secreted proteins derived from all available fungal protein data in the NCBI RefSeq database [116]. The knowledgebase is comprised of both manually curated entries as well as computationally assigned instances found with a workflow that includes SignalP [115], WolfPsort [119] and Phobius [120].

SignalP-4.1: Computational predictions of SPs is still very error prone. This is partially due to difficulty in algorithmically distinguishing SPs from N-terminal transmembrane helices [115]. SignalP-4.1 makes use of a neural-network-based method to predict SPs and is able to discriminate them from transmembrane helices [115].

All the contexts discussed in Section 2.8.1 to Section 2.8.5 may be useful by themselves for the interpretation of proteomics data. However, in order to interpret a subset of identified proteins as part of the larger system within which they function, integration of the contexts is needed in order to gain the best possible systems-view. In order to facilitate the integration of data with contextual resources a conceptual and practical framework is needed.

2.9 Networks

Networks are ideal for the study and modelling of complex systems and have been applied in many fields such as engineering, communications and computer science [121]. A network may be used to represent physical entities such as electrical circuits, roadways and molecules, as well as concepts such as ecosystems and sociological relationships [122; 123].

A network is also called a graph. A graph $G = (V,E)$ is a mathematical structure that is made up of two finite sets V and E. The elements of V are called vertices or nodes. The elements of E are called edges and each edge is defined as a set of two vertices [124; 121; 123].

A network can be visualised as a set of points (nodes) on a plane or in three dimensional space with the edges connecting these points [124; 123] and both the nodes and edges may take on any number of attributes. In addition to serving as an intuitive visualisation tool, the structure of the network and its topological features may also yield insight into the data. Furthermore, existing network analysis indices and tools such as shortest path [125] and clustering algorithms [85] may be applied to the network. Thus, networks are well suited as the scaffold for the contextualisation of biological data.

In modern molecular biology, networks have been successfully applied in a myriad of ways, including the representation of interactions between molecules [95; 126], modelling of neural networks [127] and predicting functional essentiality from topological features in metabolic networks [128]. For proteomics-specific applications, networks have been used to aid in protein identification [129] as well as interpretation of proteomics data from MS-based experiments [130]. A more in depth review of the role of networks in biology is provided by [10; 131].

When data is represented in tabular format it quickly becomes overwhelming and thus hampers the effective mining and utilisation of the data. Networks provide a means by which a complex system of interlinked components can be represented as a collection of nodes and edges. The network-based representation of proteins within a biological network enables a considerable amount of data to be visualised without obscuring patterns, trends and relationships that exist in the data. Several visualisation software packages such as Pajek [132], Cytoscape [133], Osprey [134] and Gephi [135] may be used to visualise and explore biological data as a network.

2.10 Conclusion

It is likely that MS-based proteomics will follow the trend of other omics platforms and that both data quality and quantity will increase as the cost of producing data will decrease [9]. The data is likely to come from hybrid MS instrumentation coupled to multiple rounds of LC. Various groups have demonstrated the capacity of such hybrid instruments and have reported datasets from which more than 10 000 proteins have been identified from a single experiment [6; 7; 8; 9]. Furthermore, it has been shown that, provided certain conditions are met, cross-species experiments can be performed with this technology [71; 72]. Additionally, both labelled and label-free strategies provide the capacity to add a quantitative dimension to these data sets [60].

The programs available for conducting the general steps of MS data analysis such as peptide spectral matching, peptide validation and protein inference are also likely to improve, however, the need to make customised workflows that are appropriate for the experimental design and biological aims is clear, especially for cross-species experiments.

Extant biological resources provide the information with which experimental data can be contextualised and these resources are constantly evolving and improving in both size and accuracy. Thus, as these biological resources move forward in parallel with MS-based proteomic technologies, the ability to mine these proteomic datasets will greatly improve. However, as the size of experimental data sets grow and the amount of contextual information increases, deriving biologically relevant insights and knowledge becomes more difficult.

It is understood that the characteristics of living organisms arises from the interaction of all the molecules of which they are comprised. However, it is not solely the interactions between the individual molecules themselves that gives rise to function, but rather large numbers of functionally diverse sets of components that interact selectively to produce the coherent behaviours and responses that are observed in the living system [1; 2; 136]. Thus, in order to interpret a dataset that offers a system-wide snapshot of the components, many contexts need to be brought together to facilitate a systems-based interpretation of the data. Networks provide a means by which a complex system of interlinked components can be represented as a collection of nodes and edges and are thus well suited to serve as a scaffold for the contextualisation of biological data. Network-based representation offers both a global and local perspective of the data [2; 1; 136] and have been shown to be effective solutions to the problems of data contextualisation and interpretation for various types of omic data [130; 10; 131].

2.11 Supplementary Materials

2.11.1 Biogrid Experimental Evidence Codes

Gene-based protein interactions contained within the BioGRID are determined by various methods including the following experimental systems [93]:

1. Dosage Growth Defect. Dosage refers to the over expression of a gene or the increased dosage of a gene. If over expression of the gene leads to a growth defect in a strain that is mutated or deleted for another gene, a genetic interaction between the two genes can be inferred.

2. Dosage Lethality. If over expression of the gene causes lethality in a strain that is mutated or deleted for another gene, a genetic interaction between the two genes can be inferred [137].

3. Dosage Rescue. If over expression of the gene rescues the lethality or growth defect in a strain that is mutated or deleted for another gene, a genetic interaction between the two genes can be inferred [138].

4. Phenotypic Enhancement. If the mutation or over expression of one gene results in the enhancement of any phenotype other than lethality or growth defect an interaction can be determined.

5. Phenotypic Suppression. If the mutation or over expression of one gene results in the suppression of any phenotype other than lethality or growth defect an interaction can be determined.

6. Negative Genetic Interactions. These are applicable to strains where the combination of mutations and/or deletions in separate genes, which by themselves cause no significant change in phenotype, results in a more severe fitness defect or lethality under a given condition. [139].

7. Positive Genetic Interactions. These are applicable to strains where the combination of mutations and/or deletions in separate genes, which by themselves cause no significant change in phenotype, results in a less severe fitness defect under a given condition. [139].

8. Synthetic Lethal Genetic Interaction. These can be identified when a specific mutant is screened for a second-site mutation that either suppresses or enhances the original phenotype [100].

9. Genetic Interaction Determined via Growth Defect. These can be identified when mutations in separate genes, of which each gene causes a no significant change in phenotype, result in a significant growth defect under a given condition when both mutations are combined in the same cell [140].

10. Genetic Interaction Determined via Synthetic Haplo-insufficiency. These are determined in strains where there are mutations or deletions in separate genes that cause no significant change in phenotype on its own and at least one of these genes must be hemizygous. When these mutated or deleted genes are combined in the same cell under certain conditions it should result in lethality.

11. Genetic Interaction Determined via Synthetic rescue. These can be identified when mutations or deletions of one gene rescues the lethality or growth defect of a strain mutated or deleted for another gene.

Physical protein interactions contained within the BioGRID are determined by various methods including the following experimental systems [93]:

1. Affinity Capture with Luminescence. A bait protein is tagged with luciferase and light is emitted when the prey protein immunoprecipitates with the bait. An epitope tag or polyclonal antibody may then be used to capture the prey protein from the cell extracts. An interaction can then be inferred between the bait and prey protein.

2. Affinity Capture with MS. The bait protein is affinity captured from cell extracts by either polyclonal antibody or epitope tag. MS-based methods can then be used to identify the prey protein.

3. Affinity Capture with RNA. An epitope tag or polyclonal antibody is used to capture the bait protein from the cell extracts. The RNA species associated with the bait protein can then be identified using Northern blot, real-time polymerase chain reaction (RT-PCR), affinity labelling, sequencing, or microarray analysis.

4. Affinity Capture using Western Blot. The bait protein is affinity captured from cell extracts by either a polyclonal antibody or an epitope tag. A second epitope tag or specific polyclonal antibody is used to capture the associated interaction partner using western blot analysis.

5. Interaction Determined via Biochemical Activity. The biochemical effect that one protein has on another is recorded as a type of modification. In this type of assay the substrate or "hit" protein is acted upon by the "bait" protein. For example, the bait protein may be a kinase that phoshorylates a substrate protein. Possible modifications that can be detected include Phosphorylation, Ubiquitination, Sumoylation, Dephosphorylation, Methylation, Prenylation, Acetylation, Deubiquitination, Proteolytic Processing, Glucosylation, Nedd(Rub1)ylation, Deacetylation and Demethylation.

6. Interaction Determined via Co-crystal Structure. X-ray crystallography, Nuclear Magnetic Resonance (NMR) or Electron Microscopy (EM) may be used to observe interactions on the atomic level.

7. Interaction Determined by Co-fractionation. When two or more protein subunits are present in a partially purified protein preparation an interaction can be inferred.

8. Interaction determined by Co-localization. An interaction can only be ascertained if two conditions are met. Firstly, co-localisation of two proteins in the cell must be established with indirect immunofluorescence. Secondly, if either one of the genes encoding the interacting proteins is deleted, the other protein should be incorrectly localised.

9. Interaction Determined by Co-purification. If two or more protein subunits are present in a purified protein complex an interaction can be inferred. In this method the observation must be made by classical biochemical fractionation or affinity purification and one or more additional fractionation steps.

10. Far Western Analysis. Protein mixtures are electrophoretically separated followed by the transfer of proteins to a membrane. The membrane is probed with one or more bait proteins. The location of the prey protein on the membrane is revealed if the bait and prey proteins form a complex together and thus a protein-protein interaction can be inferred.

11. Fluorescence Resonance Energy Transfer (FRET). Molecules are labelled with fluorophores and an interaction is inferred when close proximity of interaction partners is detected using fluorescence resonance energy transfer.

12. Protein-Fragment Complementation Assay (PCA). This assay relies on the joining of two complementary protein fragments to form a functional reporter protein, for instance the split-ubiquitin assay [141]. The "bait" and "prey" proteins are fused respectively to either the N- or C- terminal of a fragment reporter protein in such a manner that the "bait" and "prey" proteins are expressed as part of a reporter protein fragment. If there is an interaction between the two proteins, the complementary reporter peptide fragments will come together to form a functional reporter protein.

13. Interactions Determined Between Protein and Peptide. This type of interaction is inferred when a peptide, derived from an interaction partner, interacts with a protein. This category also includes phage display experiments [142].

14. Protein-RNA. Interactions of this type are derived from an *in vitro* assay where proteins and RNA interact.

15. Proximity Label-MS. The identification of the interacting protein in this assay is determined via similar methods as the second instance of this list, however, the bait protein in this instance is fused with a protein that selectively modifies a nearby protein. This proximal protein has a diffusible reactive product which can be detected and thus an interaction can be determined.

16. Reconstituted Complex. Interactions of this type are derived from *in vitro* assays using purified proteins.

17. The two hybrid assay operates by expressing two fusion proteins in yeast namely, the "hunter" and the "bait". The "hunter" protein is the possible binding partner fused to a yeast AD and the "bait" is the protein of interest which is fused to a yeast BD. A yeast strain is transformed with both constructs and the appropriate upstream activating sequence which is in close proximity to the reporter gene [102]. Thus, if the reporter gene is expressed it means that there was interaction between the "hunter" and the "bait" proteins [103].

Bibliography

[1] Kitano, H.: Computational systems biology. *Nature*, vol. 420, no. 6912, pp. 206–210, 2002.

[2] Hartwell, L.H., Hopfield, J.J., Leibler, S. and Murray, A.W.: From molecular to modular cell biology. *Nature*, vol. 402, pp. C47–C52, 1999.

[3] Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.-M., Cruciat, C.-M. and Others: Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, vol. 415, no. 6868, pp. 141–147, 2002.

[4] Gstaiger, M. and Aebersold, R.: Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nature Reviews Genetics*, vol. 10, no. 9, pp. 617–627, 2009.

[5] Aebersold, R. and Goodlett, D.R.: Mass spectrometry in proteomics. *Proteomics*, vol. 3, p. 5, 2001.

[6] Washburn, M.P., Wolters, D. and Yates, J.R.: Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, vol. 19, no. 3, pp. 242–247, 2001.

[7] Kislinger, T., Gramolini, A.O., MacLennan, D.H. and Emili, A.: Multidimensional Protein Identification Technology (MudPIT): Technical Overview of a Profiling Method Optimized for the Comprehensive Proteomic Investigation of Normal and Diseased Heart Tissue. *Journal of the American Society for Mass Spectrometry*, vol. 16, no. 8, pp. 1207–1220, 2005. ISSN 1044-0305.

[8] Michalski, A., Damoc, E., Hauschild, J.-P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M. and Horning, S.: Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Molecular & Cellular Proteomics*, vol. 10, no. 9, pp. M111–011015, 2011.

[9] Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J. and Aebersold, R.: The quantitative proteome of a human cell line. *Molecular Systems Biology*, vol. 7, no. 1, 2011.

[10] Aittokallio, T. and Schwikowski, B.: Graph-based methods for analysing networks in cell biology. *Briefings in Bioinformatics*, vol. 7, no. 3, pp. 243–255, 2006.

[11] Tyers, M. and Mann, M.: From genomics to proteomics. *Nature*, vol. 422, no. 6928, pp. 193–7, March 2003. ISSN 0028-0836.

[12] Evans, C.R. and Jorgenson, J.W.: The role of separation science in proteomics research. *Analytical and Bioanalytical Chemistry*, vol. 378, no. 8, pp. 1952–1961, 2004.

[13] Issaq, H.J.: The role of separation science in proteomics research. *Electrophoresis*, vol. 22, no. 17, p. 3629, 2001.

[14] Baerenfaller, K., Grossmann, J., Grobei, M.A., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., Zimmermann, P., Grossniklaus, U., Gruissem, W. and Baginsky, S.: Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science*, vol. 320, no. 5878, pp. 938–941, 2008.

[15] De Godoy, L.M.F., Olsen, J.V., Cox, J., Nielsen, M.L., Hubner, N.C., Fröhlich, F., Walther, T.C. and Mann, M.: Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*, vol. 455, no. 7217, pp. 1251–1254, 2008.

[16] Wilson, J.N.: A theory of chromatography. *Journal of the American Chemical Society*, vol. 62, no. 6, pp. 1583–1591, 1940.

[17] Wilson, K. and Walker, J.M.: *Principles and Techniques of Biochemistry and Molecular Biology*. Cambridge University Press, 2010.

[18] Giddings, J.C.: *Dynamics of Chromatography: Principles and Theory*. CRC Press, 2002.

[19] Poole, C.F. and Poole, S.K.: *Chromatography Today*. Elsevier, 2012.

[20] Snyder, L.R., Kirkland, J.J. and Dolan, J.W.: *Introduction to Modern Liquid Chromatography*. John Wiley & Sons, 2011.

[21] Regnier, F.E.: HPLC of proteins, peptides, and polynucleotides. *Analytical Chemistry*, vol. 55, no. 13, pp. 1298A–1306A, 1983.

[22] Aguilar, M.-I. and Hearn, M.T.W.: High-resolution reversed-phase high-performance liquid chromatography of peptides and proteins. *Methods in Enzymology*, vol. 270, pp. 3–26, 1996.

[23] Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., Garvik, B.M. and Yates, J.R.: Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology*, vol. 17, no. 7, pp. 676–682, 1999.

[24] McDonald, W.H., Ohi, R., Miyamoto, D.T., Mitchison, T.J. and Yates III, J.R.: Comparison of three directly coupled HPLC MS/MS strategies for identification of proteins from complex mixtures: single-dimension LC-MS/MS, 2-phase MudPIT, and 3-phase MudPIT. *International Journal of Mass Spectrometry*, vol. 219, no. 1, pp. 245–251, 2002.

[25] Chapman, J.R.: *Practical Organic Mass Spectrometry: a Guide for Chemical and Biochemical Analysis*. John Wiley & Sons, 1995.

[26] Yates III, J.R.: Mass spectrometry: from genomics to proteomics. *Trends in Genetics*, vol. 16, no. 1, pp. 5–8, 2000.

[27] Whitehouse, C.M., Dreyer, R.N., Yamashita, M. and Fenn, J.B.: Electrospray interface for liquid chromatographs and mass spectrometers. *Analytical Chemistry*, vol. 57, no. 3, pp. 675–679, 1985.

[28] Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F. and Whitehouse, C.M.: Electrospray ionization for mass spectrometry of large biomolecules. *Science*, vol. 246, no. 4926, pp. 64–71, 1989.

[29] Karas, M. and Hillenkamp, F.: Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical Chemistry*, vol. 60, no. 20, pp. 2299–2301, 1988.

[30] Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y., Yoshida, T. and Matsuo, T.: Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry*, vol. 2, no. 8, pp. 151–153, 1988.

[31] Kebarle, P.: A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry. *Journal of Mass Spectrometry*, vol. 35, no. 7, pp. 804–817, 2000.

[32] Wilm, M.S. and Mann, M.: Electrospray and Taylor-Cone theory, Dole's beam of macromolecules at last? *International Journal of Mass Spectrometry and Ion Processes*, vol. 136, no. 2, pp. 167–180, 1994.

[33] Karas, M., Bahr, U., Ingendoh, A., Nordhoff, E., Stahl, B., Strupat, K. and Hillenkamp, F.: Principles and applications of matrix-assisted UV-laser desorption/ionization mass spectrometry. *Analytica Chimica Acta*, vol. 241, no. 2, pp. 175–185, 1990.

[34] O'Brien, A.M.: *Environmental Proteomics and Mass Spectrometry: Characterization of Viable Microorganisms in Ambient Air*. ProQuest, 2007.

[35] Wiley, W.C. and McLaren, I.H.: Time-of-flight mass spectrometer with improved resolution. *Review of Scientific Instruments*, vol. 26, no. 12, pp. 1150–1157, 1955.

[36] Cotter, R.J.: Time-of-flight mass spectrometry for the structural analysis of biological molecules. *Analytical Chemistry*, vol. 64, no. 21, pp. 1027A–1039A, 1992.

[37] Guilhaus, M.: Special feature: Tutorial. Principles and instrumentation in time-of-flight mass spectrometry. Physical and instrumental concepts. *Journal of Mass Spectrometry*, vol. 30, no. 11, pp. 1519–1532, 1995.

[38] Downard, K.: *Mass Spectrometry: a Foundation Course.* Royal Society of Chemistry, 2004.

[39] Schwartz, J.C., Senko, M.W. and Syka, J.E.P.: A two-dimensional quadrupole ion trap mass spectrometer. *Journal of the American Society for Mass Spectrometry*, vol. 13, no. 6, pp. 659–669, 2002.

[40] Wong, P.S.H. and Graham Cooks, R.: Ion trap mass spectrometry. *Current Separations*, vol. 16, pp. 85–92, 1997.

[41] Cooks, R.G. and Kaiser Jr, R.E.: Quadrupole ion trap mass spectrometry. *Accounts of Chemical Research*, vol. 23, no. 7, pp. 213–219, 1990.

[42] Makarov, A.: Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Analytical Chemistry*, vol. 72, no. 6, pp. 1156–1162, 2000.

[43] Hu, Q., Noll, R.J., Li, H., Makarov, A., Hardman, M. and Graham Cooks, R.: The Orbitrap: a new mass spectrometer. *Journal of Mass Spectrometry*, vol. 40, no. 4, pp. 430–443, 2005.

[44] Scigelova, M. and Makarov, A.: Orbitrap mass analyzer - overview and applications in proteomics. *Proteomics*, vol. 6, no. S2, pp. 16–21, 2006.

[45] Steen, H. and Mann, M.: The ABC's (and XYZ's) of peptide sequencing. *Nature Reviews Molecular Cell Biology*, vol. 5, no. 9, pp. 699–711, 2004.

[46] Aebersold, R. and Mann, M.: Mass spectrometry-based proteomics. *Nature*, vol. 422, no. 6928, pp. 198–207, 2003.

[47] Pappin, D.J.C., Hojrup, P. and Bleasby, A.J.: Rapid identification of proteins by peptide-mass fingerprinting. *Current Biology*, vol. 3, no. 6, pp. 327–332, 1993.

[48] Hunt, D.F., Yates, J.R., Shabanowitz, J., Winston, S. and Hauer, C.R.: Protein sequencing by tandem mass spectrometry. *Proceedings of the National Academy of Sciences*, vol. 83, no. 17, pp. 6233–6237, 1986.

[49] Thiede, B., Höhenwarter, W., Krah, A., Mattow, J., Schmid, M., Schmidt, F. and Jungblut, P.R.: Peptide mass fingerprinting. *Methods*, vol. 35, no. 3, pp. 237–247, 2005.

[50] Nesvizhskii, A.I.: Protein identification by tandem mass spectrometry and sequence database searching. In: *Mass Spectrometry Data Analysis in Proteomics*, pp. 87–119. Springer, 2007.

[51] Biemann, K.: Sequencing of peptides by tandem mass spectrometry and high-energy collision-induced dissociation. *Methods in Enzymology*, vol. 193, pp. 455–479, 1990.

[52] Griffiths, W.J., Jonsson, A.P., Liu, S., Rai, D.K. and Wang, Y.: Electrospray and tandem mass spectrometry in biochemistry. *Biochemical Journal*, vol. 355, no. Pt 3, p. 545, 2001.

[53] Byers, M., Miflin, B.J. and Smith, S.J.: A quantitative comparison of the extraction of protein fractions from wheat grain by different solvents, and of the polypeptide and amino acid composition of the alcohol-soluble proteins. *Journal of the Science of Food and Agriculture*, vol. 34, no. 5, pp. 447–462, 1983.

[54] Conzelmann, A., Riezman, H., Desponds, C. and Bron, C.: A major 125-kd membrane glycoprotein of *Saccharomyces cerevisiae* is attached to the lipid bilayer through an inositol-containing phospholipid. *The EMBO Journal*, vol. 7, no. 7, p. 2233, 1988.

[55] Kolodziej, P.A. and Young, R.A.: [35] Epitope tagging and protein surveillance. *Methods in Enzymology*, vol. 194, pp. 508–519, 1991.

[56] Riezman, H., Hase, T., Van Loon, A.P., Grivell, L.A., Suda, K. and Schatz, G.: Import of proteins into mitochondria: a 70 kilodalton outer membrane protein with a large carboxy-terminal deletion is still transported to the outer membrane. *The EMBO Journal*, vol. 2, no. 12, p. 2161, 1983.

[57] Wright, A.P.H., Bruns, M. and Hartley, B.S.: Extraction and rapid inactivation of proteins from *Saccharomyces cerevisiae* by trichloroacetic acid precipitation. *Yeast*, vol. 5, no. 1, pp. 51–53, 1989.

[58] Rais, I., Karas, M. and Schägger, H.: Two-dimensional electrophoresis for the isolation of integral membrane proteins and mass spectrometric identification. *Proteomics*, vol. 4, no. 9, pp. 2567–2571, 2004.

[59] Nesvizhskii, A. and Keller, A.: A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry*, vol. 75, no. 17, pp. 4646–4658, 2003.

[60] Hawkridge, A.M.: Chapter 1 Practical Considerations and Current Limitations in Quantitative Mass Spectrometry-based Proteomics. In: *Quantitative Proteomics*, pp. 1–25. The Royal Society of Chemistry, 2014. ISBN 978-1-84973-808-8.

[61] Mann, M. and Jensen, O.N.: Proteomic analysis of post-translational modifications. *Nature Biotechnology*, vol. 21, no. 3, pp. 255–261, 2003.

[62] Nørregaard Jensen, O.: Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Current Opinion in Chemical Biology*, vol. 8, no. 1, pp. 33–41, 2004.

[63] Mann, M., Hendrickson, R.C. and Pandey, A.: Analysis of proteins and proteomes by mass spectrometry. *Annual Review of Biochemistry*, vol. 70, no. 1, pp. 437–473, 2001.

[64] Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H. and Aebersold, R.: Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, vol. 17, no. 10, pp. 994–999, 1999.

[65] Ross, P.L., Huang, Y.N., Marchese, J.N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S. and Others: Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Molecular & Cellular Proteomics*, vol. 3, no. 12, pp. 1154–1169, 2004.

[66] Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T. and Hamon, C.: Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. *Analytical Chemistry*, vol. 75, no. 8, pp. 1895–1904, 2003.

[67] Aggarwal, K., Choe, L.H. and Lee, K.H.: Shotgun proteomics using the iTRAQ isobaric tags. *Briefings in Functional Genomics & Proteomics*, vol. 5, no. 2, pp. 112–120, 2006.

[68] Cordwell, S.J., Wilkins, M.R., Cerpa-Poljak, A., Gooley, A.A., Duncan, M., Williams, K.L. and Humphery-Smith, I.: Cross-species identification of proteins separated by two-dimensional gel electrophoresis using matrix-assisted laser desorption ionisation/time-of-flight mass spectrometry and amino acid composition. *Electrophoresis*, vol. 16, no. 1, pp. 438–443, 1995.

[69] Wilkins, M.R. and Williams, K.L.: Cross-species protein identification using amino acid composition, peptide mass fingerprinting, isoelectric point and molecular mass: a theoretical evaluation. *Journal of Theoretical Biology*, vol. 186, no. 1, pp. 7–15, 1997.

[70] Verrills, N.M., Harry, J.H., Walsh, B.J., Hains, P.G. and Robinson, E.S.: Cross-matching marsupial proteins with eutherian mammal databases: Proteome analysis of cells from UV-induced skin tumours of an opossum (*monodelphis domestica*). *Electrophoresis*, vol. 21, no. 17, pp. 3810–3822, 2000.

[71] Snijders, A.P.L., de Koning, B. and Wright, P.C.: Relative quantification of proteins across the species boundary through the use of shared peptides. *Journal of Proteome Research*, vol. 6, no. 1, pp. 97–104, 2007.

[72] Pandhal, J., Snijders, A.P.L., Wright, P.C. and Biggs, C.A.: A cross-species quantitative proteomic study of salt adaptation in a halotolerant environmental

isolate using 15N metabolic labelling. *Proteomics*, vol. 8, no. 11, pp. 2266–2284, 2008.

[73] Lester, P.J. and Hubbard, S.J.: Comparative bioinformatic analysis of complete proteomes and protein parameters for cross-species identification in proteomics. *Proteomics*, vol. 2, no. 10, pp. 1392–1405, 2002.

[74] Wright, J.C., Beynon, R.J. and Hubbard, S.J.: Cross species proteomics. In: *Proteome Bioinformatics*, pp. 123–135. Springer, 2010.

[75] Cordwell, S.J., Basseal, D.J. and Humphery-Smith, I.: Proteome analysis of *Spiroplasma melliferum* (A56) and protein characterisation across species boundaries. *Electrophoresis*, vol. 18, no. 8, pp. 1335–1346, 1997.

[76] Deutsch, E.W., Lam, H. and Aebersold, R.: Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. *Physiological Genomics*, vol. 33, no. 1, pp. 18–25, 2008.

[77] Eng, J.K., McCormack, A.L. and Yates, J.R.: An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, vol. 5, no. 11, pp. 976–989, 1994.

[78] Cottrell, J.S. and London, U.: Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, vol. 20, no. 18, pp. 3551–3567, 1999.

[79] Craig, R. and Beavis, R.C.: TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, vol. 20, no. 9, pp. 1466–1467, 2004.

[80] Keller, A., Nesvizhskii, A.I., Kolker, E. and Aebersold, R.: Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry*, vol. 74, no. 20, pp. 5383–5392, 2002.

[81] Nesvizhskii, A.I., Keller, A., Kolker, E. and Aebersold, R.: A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry*, vol. 75, no. 17, pp. 4646–4658, 2003.

[82] Deutsch, E.W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B. and Prazen, B.: A guided tour of the Trans Proteomic Pipeline. *Proteomics*, vol. 10, no. 6, pp. 1150–1159, 2010.

[83] Kohlbacher, O., Reinert, K., Gröpl, C., Lange, E., Pfeifer, N., Schulz-Trieglaff, O. and Sturm, M.: TOPP - The OpenMS proteomics pipeline. *Bioinformatics*, vol. 23, no. 2, pp. e191–e197, 2007.

[84] Cox, J. and Mann, M.: MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, vol. 26, no. 12, pp. 1367–1372, 2008.

[85] van Donge, S.M.: *Graph Clustering by Flow Simulation*. Ph.D. thesis, University of Utrecht, 2000.

[86] Fitch, W.M.: Distinguishing homologous from analogous proteins. *Systematic Biology*, vol. 19, no. 2, pp. 99–113, 1970.

[87] Fitch, W.M.: Homology: a personal view on some of the problems. *Trends in Genetics*, vol. 16, no. 5, pp. 227–231, 2000.

[88] Chen, F., Mackey, A.J., Vermunt, J.K. and Roos, D.S.: Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PloS One*, vol. 2, no. 4, p. e383, 2007.

[89] O'Brien, K.P., Remm, M. and Sonnhammer, E.L.L.: Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Research*, vol. 33, no. suppl 1, pp. D476—-D480, 2005.

[90] Tatusov, R.L., Galperin, M.Y., Natale, D.A. and Koonin, E.V.: The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, vol. 28, no. 1, pp. 33–36, 2000.

[91] Enright, A.J., Van Dongen, S. and Ouzounis, C.A.: An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, vol. 30, no. 7, pp. 1575–1584, 2002.

[92] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, vol. 25, no. 1, pp. 25–9, May 2000. ISSN 1061-4036.

[93] Chatr-aryamontri, A., Breitkreutz, B.-J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., O Donnell, L. and Others: The BioGRID interaction database: 2013 update. *Nucleic Acids Research*, vol. 41, no. D1, pp. D816—-D823, 2013.

[94] Pawson, T. and Nash, P.: Assembly of cell regulatory systems through protein interaction domains. *Science*, vol. 300, no. 5618, pp. 445–452, 2003.

[95] Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A. and Tyers, M.: BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, vol. 34, no. suppl 1, pp. D535–D539, 2006.

[96] Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P. and Others: A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, vol. 403, no. 6770, pp. 623–627, 2000.

[97] Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S. and Sakaki, Y.: Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proceedings of the National Academy of Sciences*, vol. 97, no. 3, pp. 1143–1147, 2000.

[98] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y.: A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences*, vol. 98, no. 8, pp. 4569–4574, 2001.

[99] Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.-L., Millar, A., Taylor, P., Bennett, K., Boutilier, K. and Others: Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, vol. 415, no. 6868, pp. 180–183, 2002.

[100] Tong, A.H.Y., Evangelista, M., Parsons, A.B., Xu, H., Bader, G.D., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C.W.V., Bussey, H. and Others: Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, vol. 294, no. 5550, pp. 2364–2368, 2001.

[101] Pan, X., Yuan, D.S., Xiang, D., Wang, X., Sookhai-Mahadeo, S., Bader, J.S., Hieter, P., Spencer, F. and Boeke, J.D.: A robust toolkit for functional profiling of the yeast genome. *Molecular Cell*, vol. 16, no. 3, pp. 487–496, 2004.

[102] Fields, S. and Song, O.-k.: A novel genetic system to detect protein-protein interactions. *Nature*, vol. 340, no. 6230, pp. 245–6, July 1989. ISSN 0028-0836.

[103] Sobhanifar, S.: Yeast two hybrid assay: A fishing tale. *BioTeach Journal*, vol. 1, pp. 81–87, 2003.

[104] Bader, G.D. and Hogue, C.W.V.: Analyzing yeast protein–protein interaction data obtained from different sources. *Nature Biotechnology*, vol. 20, no. 10, pp. 991–997, 2002.

[105] Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. and Bork, P.: Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, vol. 417, no. 6887, pp. 399–403, 2002.

[106] Tatu, U.: Nobel Prize in Physiology or Medicine 1999. *Resonance*, vol. 5, no. 5, pp. 91–94, 2000.

[107] Choo, K.H., Tan, T.W. and Ranganathan, S.: SPdb-a signal peptide database. *BMC Bioinformatics*, vol. 6, p. 249, January 2005. ISSN 1471-2105.

[108] Heijne, G.: Signal Peptides. *eLS*, 1990.

[109] Blobel, G. and Dobberstein, B.: Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. *The Journal of Cell Biology*, vol. 67, no. 3, pp. 835–851, 1975.

[110] Hiller, K., Grote, A., Scheer, M., Münch, R. and Jahn, D.: PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Research*, vol. 32, no. suppl 2, pp. W375–W379, 2004.

[111] Fariselli, P., Finocchiaro, G. and Casadio, R.: SPEPlip: the detection of signal peptide and lipoprotein cleavage sites. *Bioinformatics*, vol. 19, no. 18, pp. 2498–2499, 2003.

[112] Chou, K.-C. and Shen, H.-B.: Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. *Biochemical and Biophysical Research Communications*, vol. 357, no. 3, pp. 633–640, 2007.

[113] Shen, H.-B. and Chou, K.-C.: Signal-3L: A 3-layer approach for predicting signal peptides. *Biochemical and Biophysical Research Communications*, vol. 363, no. 2, pp. 297–303, 2007.

[114] Frank, K. and Sippl, M.J.: High-performance signal peptide prediction based on sequence alignment techniques. *Bioinformatics*, vol. 24, no. 19, pp. 2172–2176, 2008.

[115] Petersen, T.N., Brunak, S.r., von Heijne, G. and Nielsen, H.: SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, vol. 8, no. 10, pp. 785–786, 2011.

[116] Lum, G. and Min, X.J.: FunSecKB: the Fungal Secretome KnowledgeBase. *Database : The Journal of Biological Databases and Curation*, vol. 2011, p. bar001, January 2011. ISSN 1758-0463.

[117] Consortium, U. and Others: The universal protein resource (UniProt). *Nucleic Acids Research*, vol. 36, no. suppl 1, pp. D190–D195, 2008.

[118] Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., Browne, P., van den Broek, A., Castro, M., Cochrane, G. and Others: The EMBL nucleotide sequence database. *Nucleic Acids Research*, vol. 33, no. suppl 1, pp. D29–D33, 2005.

[119] Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J. and Nakai, K.: WoLF PSORT: protein localization predictor. *Nucleic Acids Research*, vol. 35, no. suppl 2, pp. W585–W587, 2007.

[120] Käll, L., Krogh, A. and Sonnhammer, E.L.L.: Advantages of combined transmembrane topology and signal peptide prediction - the Phobius web server. *Nucleic Acids Research*, vol. 35, no. suppl 2, pp. W429–W432, 2007.

[121] Deo, N.: *Graph theory with applications to engineering and computer science.* PHI Learning Pvt. Ltd., 2004.

[122] Bollobás, B.: *Modern Graph Theory*, vol. 184. Springer, 1998.

[123] Gross, J.L. and Yellen, J.: *Graph theory and its applications second edition.* CRC press, 2006.

[124] Gibbons, A.: *Algorithmic Graph Theory.* Cambridge University Press, 1985.

[125] Kruskal, J.: On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. *Proceedings of the American Mathematical Society*, vol. 7, no. 1, pp. 48–50, 1956.

[126] Kanehisa, M. and Goto, S.: KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, January 2000. ISSN 0305-1048.

[127] Cartwright, H.M.: Artificial Neural Networks in Biology and Chemistry-The Evolution of a New Analytical Tool. In: *Artificial Neural Networks*, pp. 1–13. Springer, 2009.

[128] Palumbo, M.C., Colosimo, A., Giuliani, A. and Farina, L.: Functional essentiality from topology features in metabolic networks: a case study in yeast. *FEBS letters*, vol. 579, no. 21, pp. 4642–4646, 2005.

[129] Li, J., Zimmerman, L.J., Park, B.-H., Tabb, D.L., Liebler, D.C. and Zhang, B.: Network-assisted protein identification and data interpretation in shotgun proteomics. *Molecular Systems Biology*, vol. 5, no. 1, 2009.

[130] Deighton, R.F., Kerr, L.E., Short, D.M., Allerhand, M., Whittle, I.R. and McCulloch, J.: Network generation enhances interpretation of proteomic data from induced apoptosis. *Proteomics*, vol. 10, no. 6, pp. 1307–1315, 2010.

[131] Giuliani, A., Filippi, S. and Bertolaso, M.: Why network approach can promote a new way of thinking in biology. *Frontiers in Genetics*, vol. 5, 2014.

[132] Batagelj, V. and Mrvar, A.: Pajek-program for large network analysis. *Connections*, vol. 21, no. 2, pp. 47–57, 1998.

[133] Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, vol. 13, no. 11, pp. 2498–504, November 2003. ISSN 1088-9051.

[134] Breitkreutz, B.-J., Stark, C., Tyers, M. and Others: Osprey: a network visualization system. *Genome Biology*, vol. 4, no. 3, p. R22, 2003.

[135] Bastian, M., Heymann, S., Jacomy, M. and Others: Gephi: an open source software for exploring and manipulating networks. *ICWSM*, vol. 8, pp. 361–362, 2009.

[136] Barabasi, A.-L. and Oltvai, Z.N.: Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, 2004.

[137] Kroll, E.S., Hyland, K.M., Hieter, P. and Li, J.J.: Establishing genetic interactions by a synthetic dosage lethality phenotype. *Genetics*, vol. 143, no. 1, pp. 95–102, 1996.

[138] Díaz-Blanco, N.L. and Rodríguez-Medina, J.R.: Dosage rescue by UBC4 restores cell wall integrity in *Saccharomyces cerevisiae* lacking the myosin type II gene MYO1. *Yeast*, vol. 24, no. 4, pp. 343–355, 2007.

[139] Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L.Y., Toufighi, K., Mostafavi, S. and Others: The genetic landscape of a cell. *Science*, vol. 327, no. 5964, pp. 425–431, 2010.

[140] Collins, S.R., Miller, K.M., Maas, N.L., Roguev, A., Fillingham, J., Chu, C.S., Schuldiner, M., Gebbia, M., Recht, J., Shales, M. and Others: Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature*, vol. 446, no. 7137, pp. 806–810, 2007.

[141] Johnsson, N. and Varshavsky, A.: Split ubiquitin as a sensor of protein interactions in vivo. *Proceedings of the National Academy of Sciences*, vol. 91, no. 22, pp. 10340–10344, 1994.

[142] Clackson, T., Hoogenboom, H.R., Griffiths, A.D. and Winter, G.: Making antibody fragments using phage display libraries. *Nature*, vol. 352, no. 6336, pp. 624–8, August 1991. ISSN 0028-0836.

Chapter 3

Protein Identification, Quantification and Network-based Contextualisation

3.1 Introduction

The field of proteomics involves a variety of technical disciplines, however, for the characterisation of entire proteomes the use of Liquid chromatography (LC) coupled to Mass Spectrometry (MS) has shown to be a very valuable tool allowing for both protein identification and quantification [1]. Specifically, the use of approaches involving LC coupled to tandem mass spectrometry in shotgun proteomics have illustrated the capacity to identify thousands of proteins from a single sample in one experiment [2; 3; 4; 5].

LC-MS/MS involves the use of more than one mass analyser in tandem with one another [6]. In a typical LC-MS/MS experiment, a precursor ion is selected based on its mass by mass analyser one (MS1) and focused into a collision region that precedes a second mass analyser (MS2). The mass analysers can be arranged either in space as is the case with sector and triple quadrupole instruments, or in time as is the case with trapping instruments such as an ion trap. Key to LC-MS/MS is a process known as Collision Induced Dissociation (CID) [7]. This process involves the introduction of an inert gas into a collision zone. In the collision zone the inert gas molecules collide with the precursor ion. This process produces so called product ions from the precursor ion and the product ions can then be mass analysed by MS2.

The processing and analysis of proteomics data is a complex process that consists of multiple steps. These steps include: 1) Performing peak detection on the native spectra; 2) The processing of the raw data and distinguishing signal from noise; 3) Matching observed peptide spectra to theoretical spectra; 4) The *in silico* reassembly of peptides into proteins; 5) Validation of search results on the peptide and protein levels respectively; 6) Annotation and interpretation of the results. Despite the complexity of the process, the quality and quantity of MS-based analysis is likely to increase in keeping with the trend set by other high throughput technologies [8].

Typically, peak detection and processing of the raw LC-MS/MS data is performed online by proprietary software, however, many tools exist from various groups that can be pieced together to form a customised workflow that perform the rest of these steps. However, there are only a few packages available that aim to provide a single environment that allows one to perform all of the steps required. One such workflow is the Trans Proteomic Pipeline (TPP) [9]. Other similar packages include the openMS proteomics pipeline (TOPP) [10] and Max Quant [11]. In this work we will focus on the TPP and the various tools that are associated with it.

The TPP was selected for use because of its open-source nature and good description of its component programs and algorithms. Moreover, the TPP is comprehensive in that it provides tools for to perform all the required steps for a typical tandem MS experiment. Furthermore, the modular design of the pipeline means outputs of utilities can easily be linked to custom programs Deutsch2010.

Even after protein assignments have been made, the output of a typical shotgun proteomics experiment may still be daunting. At this point, the output is usually in the form of a large spreadsheet consisting of hundreds of rows and several columns with each row representing a protein and columns representing identification attributes. Additionally, quantitative strategies may provide information about protein abundances [12].

Although the proteins are identified as independent entities, they function as part of a greater, connected system. Furthermore, most biological functions arise from interactions between proteins and other molecules [13]. However, to interpret the identified proteins as part of a complex system, context is needed.

Proteins may be contextualised in a variety of manners such as their sequence similarity to other proteins, their ontological descriptions, with what other proteins they interact, the metabolic pathways in which they are active and where in the cell they are located. Although these contexts may be useful by themselves for the interpretation of proteomics data, in order to interpret a subset of identified proteins as part of the larger system within which they function, integration of the contexts is needed in order to gain the best possible systems view.

When data is represented in tabular format it quickly becomes overwhelming and thus hampers the effective mining and utilisation of the data. Networks provide a means by which a complex system of interlinked components can be represented as a collection of nodes and edges. A network-based representation of proteins within a biological network enables a considerable amount of data to be visualised without obscuring patterns, trends and relationships that exists in the data. Although much work has been done on the identification of proteins from LC-MS/MS data, tools for the contextualisation and interpretation of the proteins are still lacking.

This work describes a computational workflow that makes use of a combination of existing and custom made utilities in order to firstly identify and quantify proteins from cross-species LC-MS/MS data and secondly, place the protein identifications into a biological context using network-based methods. The method is demonstrated using a quantitative LC-MS/MS dataset derived from two yeast species grown under fermentative conditions.

The LC-MS/MS data was provided by Dr Thulile Ndlovu in the lab of Prof Florian Bauer and Dr Benoit Divol at the Institute for Wine Biotechnology, Stellenbosch. LC-MS/MS analysis was conducted at the Proteomics Core Facility at Sahlgrenska Academy, University of Gothenburg, Sweden. The original experimental aims behind the generation of the data were to investigate what proteins were present in the synthetic wine must after conducting fermentations to dryness using two yeasts, namely *S. cerevisiae* VIN13 and *S. paradoxus* RO88 respectively. The principle goals of the experiment were to determine the presence of haze protection factor proteins and possible ascertain differences in the relative abundances of these proteins between the two yeast species [14]. This work represents an effort to further mine the data using an alternative computational workflow for the identification, quantification and contextualisation of the proteins. Several examples are shown that illustrate the method's utility for both investigation of existing hypotheses and the formulation of new hypotheses.

3.2 Methods

3.2.1 Experimental Design and Sample Generation

Figure 3.1 illustrates the experimental design used. Three biological replicates for VIN13 and RO88 respectively were produced. The samples consisted of proteins present in the growth media after fermentations where run to dryness. Each replicate sample was digested with trypsin and the resulting peptides labelled with a unique Tandem Mass Tag (TMT). The samples were then pooled and separated using Strong Cation Exchange Chromatography (SCX) which resulted 13 fractions. These fractions were then analysed using a LTQ-Orbitrap-Velos (Thermo Fisher Scientific) mass spectrometer interfaced with an in-house constructed nano-LC column. Further details about sample generation and the LC-MS/MS workflow are described in the supplementary materials Section 3.5.2.

3.2.2 Data Analysis Objectives and Workflow

The objectives in terms of the data analysis strategy were as follows: 1) To identify what proteins were expressed in one or both of these yeast species as

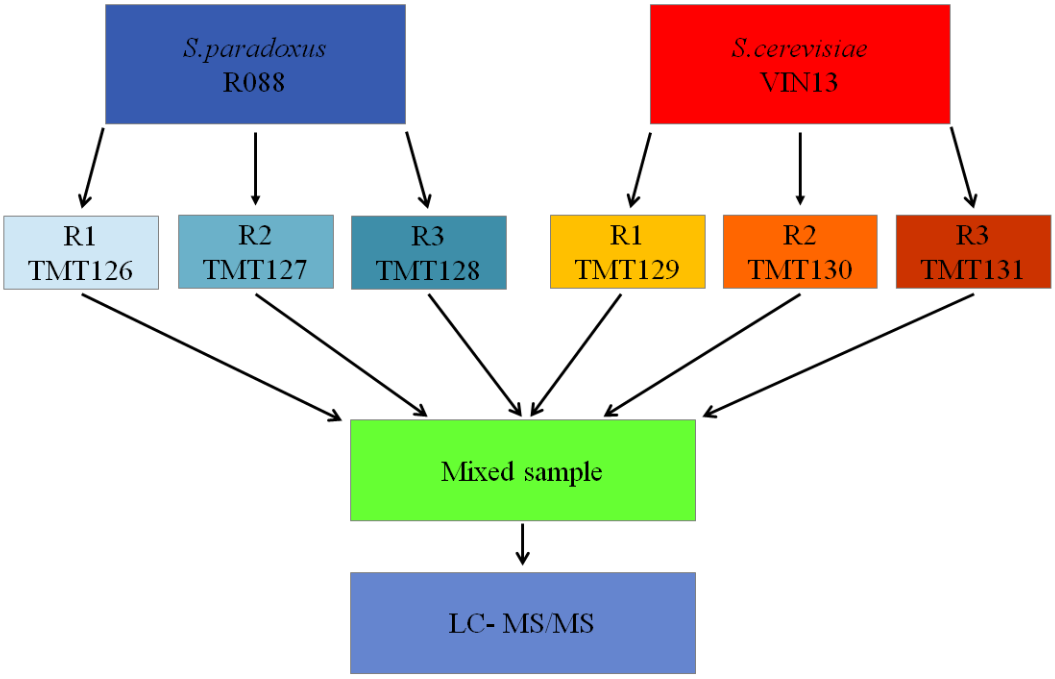


Figure 3.1: **Experimental design used to produce the VIN13 and R088 protein mass spectra.** Three biological replicates for VIN13 and RO88 respectively were produced. Each replicate sample was labelled with a unique Tandem Mass Tag (TMT) with Tag masses ranging from 126 Da to 131 Da.

well as to determine relative quantitative differences of the proteins produced by both species using the Tandem Mass Tag (TMT) signals; 2) To contextualise and visualise the data; 3) Mine the data to the full extent possible whilst remaining cognisant of bias and error that is inherent to these data types; 4) Report and present the data in a manner that is intuitive, easily interpretable and facilitates easy hypotheses generation and pattern recognition.

The workflow presented here consisted of the following main components: identification and quantification of proteins using the Trans Proteomic Pipeline [9], identification of orthologs using OrthoMCL [15], description of the data using the Gene ontology [16] supported by GO enrichment analysis using GOEAST [17], further contextualisation of identified proteins using KEGG biochemical pathways [18] and the BioGRID interactome database [19] and visualisation of networks using Cytoscape [20].

3.2.3 Identification and Quantification of Proteins from Spectra

Figure 3.2 outlines all the major steps and tools used during the identification and quantification of proteins using the TPP.

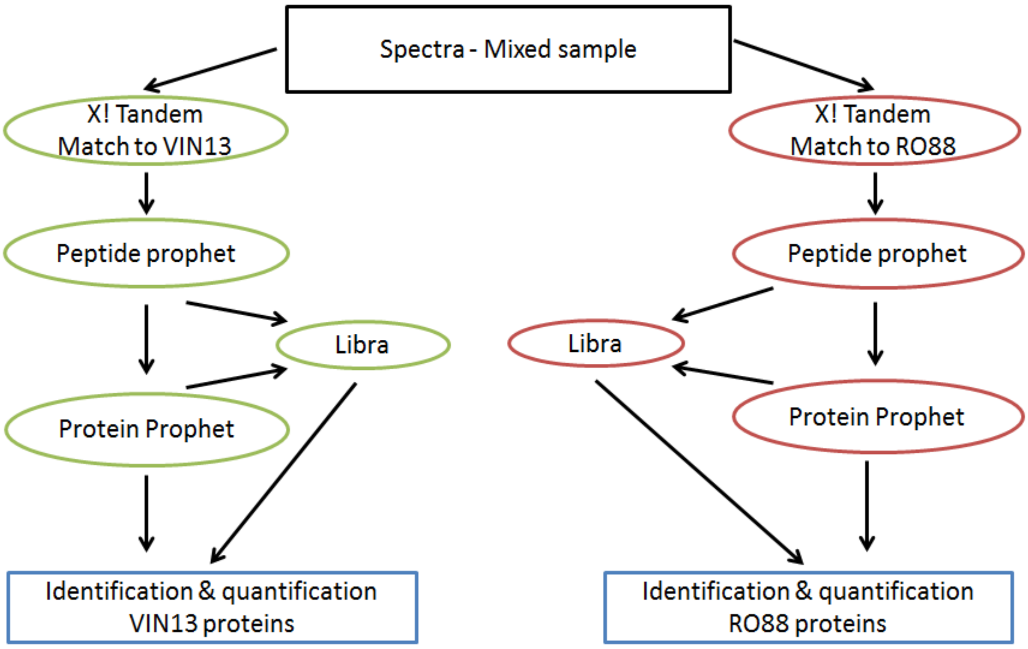


Figure 3.2: **Overview of the TPP workflow utilised for the identification and quantification of proteins.** The ellipses represent programs or processes whilst the rectangles are inputs and or outputs.

3.2.3.1 Database Search and TMT Quantification

MS raw data files from all 13 SCX fractions for the TMT 6-plex set were merged for relative quantification and identification using the Trans Proteomic Pipeline (TPP) version v4.6 OCCUPY rev 1, Build 201209261035 (MinGW) on Windows 7. The .raw format data files were converted to mzXML format using the msconvert tool in TPP. Database searching was performed with X!Tandem [21] as part of the TPP. Two workflows were run using separate species-specific proteome databases in order to ensure that the search space matched the samples as closely as possible.

The *S. cerevisiae* VIN13 proteome was obtained from http://www.uniprot.org/uniprot on 6 June 2013 and contained 3916 proteins. An *S. paradoxus* theoretical proteome was created from the corresponding *S. paradoxus* NRRL Y-17217 ORFs [22] downloaded from http://www.broadinstitute.org/annotation/fungi/comp_yeasts/downloads.html/S1b.ORFs/Spar_extended.fasta.gz on 4 Febraury 2013 and translated using the EMBOSS-6.5.7 [23] translate utility. The resulting *S. paradoxus* predicted proteome contained 4787 proteins.

3.2.3.2 X!Tandem Database Search Parameters

The parent monoisotopic mass error tolerance was set to -2 Da and +4 Da, cysteine carbamethylation and methionine oxidation were set as potential mass modifications. Appropriate mass modifications were allowed for the Tandem Mass Tags, namely static mass modifications at the N-terminal and Lysine (K), with potential mass modifications at the tyrosine residues. Semi-tryptic peptides and up to two missed cleavages were allowed. The tandemparameters.xml file used is provided in Figure S1 of the supplemental materials section.

3.2.3.3 Peptide and Protein Prophet

The peptide-spectral matches made by X!Tandem were then validated by Peptide Prophet [24]. The utility performed this validation by learning the distributions of search scores and peptide properties such as the number of termini compatible with enzymatic cleavage and the number of missed cleavages. For each result it computes the probability that the peptide assignment is correct or incorrect.

The output from Peptide Prophet is passed on to Protein Prophet [25] which in turn computes a probability that proteins are present on the basis of peptide assignments. Peptides that correspond to more than a single protein are apportioned among all corresponding proteins. Protein Prophet then derives a minimal list of proteins that is sufficient to account for the observed peptide assignments using an expectation maximisation algorithm. Both Peptide Prophet and Protein Prophet were used with their default settings.

3.2.3.4 Libra

To capture and process the quantitative TMT information, a modified version of a program called Libra [9] was used. Libra is is also incorporated in the TPP. Libra integrates the intensities of the TMT mass to charge ratios in an LC-MS/MS spectrum and stores the values at the peptide level. Protein Prophet then infers the simplest list of proteins consistent with the identified peptides and protein quantity is then derived from the group of peptides associated with the protein. Libra may be used to perform normalisation of the data and outlier removal, however, in this instance the cross-species experimental design necessitated the use of a modified algorithm. Thus, the unprocessed quantitation.tsv file containing the peptides and their raw tag values were retrieved and processed using custom-built Perl programs. The criteria for the modified algorithm and their intended purpose in the context of the cross-species experimental design are discussed more in Section 3.2.5.

From the TPP workflow, two species-specific outputs were obtained. Each output contained unfiltered data pertaining to protein groups, peptides, probabilities and other identification attributes whilst raw mass tag signals were obtained from the quantitation.tsv file. Perl programs were used in order to set thresholds and parse out the necessary information from these TPP outputs.

3.2.4 Orthology Detection

Two separate sequence comparisons were conducted using OrthoMCL v2.0.5 with default parameters in both cases. In the first comparison, orthologs were the detected between proteomes of *S. cerevisiae* VIN13 and *S. paradoxus* RO88. The outputs from OrthoMCL used in this instance were the lists of orthologs and co-orthologs between the species. These two files were then further processed using a custom written Perl program (score-based filter) in order to produce a list of unique protein pairs with one-to-one ortholog relationships.

For the second comparison, orthologs were detected between proteomes of *S. cerevisiae* VIN13, *S. paradoxus* and *S. cerevisiae* S288C. The S288C reference proteome was downloaded from the *Saccharomyces* Genome Database (SGD) from the following url: http://downloads.yeastgenome.org/sequence/S288C_reference/orf_protein/orf_trans_all.fasta.gz on 15 March 2013. The aim of this step was to associate protein IDs from our target organisms to the systemic IDs that exist for *S. cerevisiae* S288C via sequence similarity.

The establishment of this sequence-based relationship for a given protein enables one to annotate this protein with the extant knowledge contained within the Gene Ontology, biochemical pathways and interactome databases that have been used to annotate S288C. The list of orthologs, co-orthologs and family members (determined using a MCL inflation value of 10), were then concatenated into one reference file for further use. A small fraction of proteins failed to be grouped into families using OrthoMCL. For these proteins the best reciprocal BLAST match was used to assign ortholog relationships. If a protein was designated as ‘identified in both species’, it means that an ortholog systemic ID was identified in both VIN13 and RO88.

3.2.5 Inferring and Calculating Relative Protein Fold Change

Criteria for Quantitative Peptides: In order to calculate a relative fold change between two proteins the following criteria were designed to minimise false signals and derive the most accurate conclusion from the data: 1) The protein must be identified in both organisms; 2) there must be at least one peptide that is unique to the ortholog protein pair; 3) the peptide must have a complete set of TMT signals, in other words a TMT signal must exist from every replicate biological sample.

Normalisation of the TMT Signal: The following normalisation and outlier removal was conducted on a subset of peptides successfully labelled with TMTs: 1) All peptides where one or more of the TMT signals were absent were excluded; 2) Peptides with a peptide probability less than 0.5 were not used; 3) Each peptide channel was normalised by the sum of that peptide's channels; 4) The average and standard deviation (σ) of all the signals for a given peptide were calculated; 5) If a given signal value deviated by more than two σ the signal was not used.

Relative Fold Change Ratio: A quantitative signal for the protein was calculated by taking the average of all of its quantitative peptide constituents that passed the stipulated peptide filtering requirement. What remained for each protein was a value associated with each TMT. An average of the three biological replicates for RO88 (TMT 126, 127 and 128) and VIN13 (TMT 129, 130 and 131) was calculated respectively. A ratio of relative protein abundance for a given protein-pair was taken as the ratio of VIN13/RO88. When the the ratio VIN13/RO88 was less than 1, the negative inverse of the ratio was reported. To determine if the fold change for a given protein is statistically significant between the two organisms a simple two tailed t-test for two independent samples was conducted using the package Statistics::TTest, Version 1.1.0 by Yun-Fang Juan (http://search.cpan.org/~yunfang/Statistics-TTest-1.1.0/TTest.pm). To correct for multiple hypotheses testing Benjamini-Hochberg p-value adjustment [26] was used and implemented in R [27]. The adjusted p-value threshold for significance was set at 0.1.

3.2.6 Unresolved Identifications and Multi-ortholog Proteins

The protein sequence comparisons performed in Section 3.2.4 revealed numerous cases in the data where the ortholog relationship between the three species was ambiguous. In other words, clear ortholog relationships could not be established by OrthoMCL or alternatively by taking the best reciprocal BLAST hit relationship. These proteins were attributed as having multiple orthologs. Some cases where the same set of peptide spectra were matched equally well to more than one protein by the TPP were also attributed as ambiguous. In instances where no such ambiguity was detected, the protein was attributed as single-ortholog or unambiguous in order to distinguish it from the multi-ortholog cases.

3.2.7 Resources for Network Contextualizations

Several databases and resources were used to place the protein identifications into relevant biological contexts. The resources are described in more detail below:

3.2.7.1 Kyoto Encyclopaedia for Genes and Genomes (KEGG)

In order to contextualise the proteins within a biochemical context, metabolic pathways as defined by KEGG were used. The yeast metabolic network containing the relationships between proteins, reactions, compounds and their respective pathways was obtained in .xml format and parsed using custom Perl programs.

3.2.7.2 The Gene Ontology

The Gene Ontology (GO) is a controlled vocabulary of terms for describing gene product characteristics [16]. The GO annotation file for *S. cerevisiae* S288C was downloaded from the following url: ftp://ftp.geneontology.org/pub/go/gene-associations/gene_association.sgd.gz on 25 March 2013.

A Perl program was written in order to connect the identified proteins to the Gene Ontology. As inputs, this program uses the information acquired during the second sequence comparison as described in Section 3.2.4, the proteins identified via the TPP along with their identification and quantitative attributes and a gene association file for *S. cerevisiae* S288C.

The contents of the gene association file connects the systematic ID and other gene name aliases for *S. cerevisiae* S288C to corresponding GO-IDs. This step effectively transfers Gene Ontology terms to the identified proteins via the determined ortholog relationship with *S. cerevisiae* S288C. The GO-IDs were then used to retrieve the GO terms that corresponds to each GO-ID.

3.2.7.3 Gene Ontology Enrichment Analysis Software Toolkit (GOEAST)

GO enrichment analysis was performed using the customised analysis tool of the GOEAST web-utility [17]. As background for the GO-EAST analysis, a file containing all the systemic IDs of S288C and their associated GO terms was created. As a target, an input list containing all the systemic IDs of the proteins identified was used. The output from the utility is a tab delimited file that consisted of the following information: For each enriched GO term, a corresponding log-odds ratio and a p-value exists. The log-odds ratio is the measure for enrichment and the p-value may be used to determine if the enrichment is statistically significant.

3.2.7.4 BioGRID

A database containing both protein and genetic interactions for *S. cerevisiae* was obtained from http://thebiogrid.org/download.php on 8 August 2013 [19].

The database was split into genetic interactions and protein interactions respectively. Custom written Perl programs were used to parse out the appropriate information from the database.

3.2.7.5 Proteins with a Secretion Signal

In order to check if the identified proteins had any secretion signals or were known to be secreted, two different databases were used as reference, namely the Signal Peptide database (SPdb) [28] and the Fungal Secretome KnowledgeBase (FunSecKB) [29]. Additionally, a predictive tool called SignalP-4.1 [30] was also used. These three sources were used to construct a custom reference file containing proteins with a secretion signal. The systemic-and/or common name IDs of the identified proteins were used to match against the custom reference file. If a match occurred the relevant information was incorporated as a protein attribute and was displayed as such in the Cytoscape sessions.

3.2.8 Contextual Network Construction

3.2.8.1 Overview

Using the outputs obtained from the Sections 3.2.3 to 3.2.7 a variety of networks were constructed in order to provide different biological contexts for the identified proteins, such as functional context through GO terms, biochemical context in KEGG metabolic pathways and interaction context defined by BioGRID. Each of these networks places either a protein or a module of proteins in any one of the contexts.

When visualised, all of the networks have a visual style that serves to represent the proteins and their attributes. Node shapes relate to the type of nodes in the network. Proteins and protein family modules are ellipses, GO terms are hexagons and metabolic pathway and protein interaction module nodes appear as squares with rounded edges in their respective networks. The protein node border reflects a continuous color mapping relating to the probability of the protein identification. Black indicates a high quality identification (highest probability of 1), whilst lower quality identifications take lighter shades of grey.

The protein node color corresponds to a continuous color mapping reflecting the relative fold change attribute. Red shades indicate a large fold increase in the quantity of the protein in VIN13 relative to the amount in RO88 with darker red shades indicating a larger fold change. Blue shades indicate a large fold increase in the quantity of the protein in RO88 relative to the amount in VIN13 with darker blue shades indicating a larger fold change. If the protein node color is white, it means no relative fold change could be calculated for this protein. Only proteins with statistically significant fold changes have shaded node colors. The protein node color style also applies to protein family module nodes. Relevant information about the protein is built into the visualisation or the structure of the network and available in tabular format as node attributes. A selection of attributes that accompany protein nodes include a protein ID and aliases, a description of the protein, probability of protein identification, the orthology group type, presence of a secretion signal and associated GO terms. Quantitated proteins have additional attributes, namely the number of peptides contributing to the quantitative signal for the protein, the fold change and the adjusted p-value for fold change significance.

The construction of each network is described below. All networks were created using custom Perl programs. The outputs were generated in .sif format and visualised in Cytoscape version 2.8.3.

3.2.8.2 Protein - GO Term network

The protein to GO term network consists of two types of nodes, namely protein and GO term nodes. A protein node is connected to a GO term node via an edge if the gene is associated with the GO term as defined by the gene association file. This network can be seen in Figure 3.3.

3.2.8.3 Protein-Protein GO Overlap Network

The Jaccard index is a set overlap similarity metric and was used here as a measure of GO term overlap between proteins. It is calculated by dividing the size of the intersection by the size of the union of two sets [31; 32]. A set in this instance refers to the GO terms that are associated with a given protein. The Jaccard index was calculated for every possible pair of identified proteins resulting in a matrix from which networks can be constructed. The Jaccard index varies between 0 and 1, with a value of 1 indicating complete set overlap (the proteins have the same set of associated GO terms), whilst zero indicates no set overlap. A network was constructed in which each node represents a protein and each edge represents the overlap between the GO terms associated with the two proteins as quantified by the Jaccard index. This network can be filtered based on a Jaccard index cutoff value. By doing this, one is able to create network views based on the desired level of connectivity in the network. Edge attributes for this network include the Jaccard index and a list of the shared GO terms for each pair of proteins. A selection of the resulting networks is shown in Figure 3.13 of Section 3.3.5.

3.2.8.4 Protein Family Modules - GO term network

As discussed in Section 3.2.6, a large fraction of the identified proteins in the dataset matched equally well to multiple orthologs. This ambiguity makes it difficult to assign attributes and context to such proteins when treating them as individuals. In order to contextualise and represent these ambiguous proteins, protein family modules were created. To define the members of a

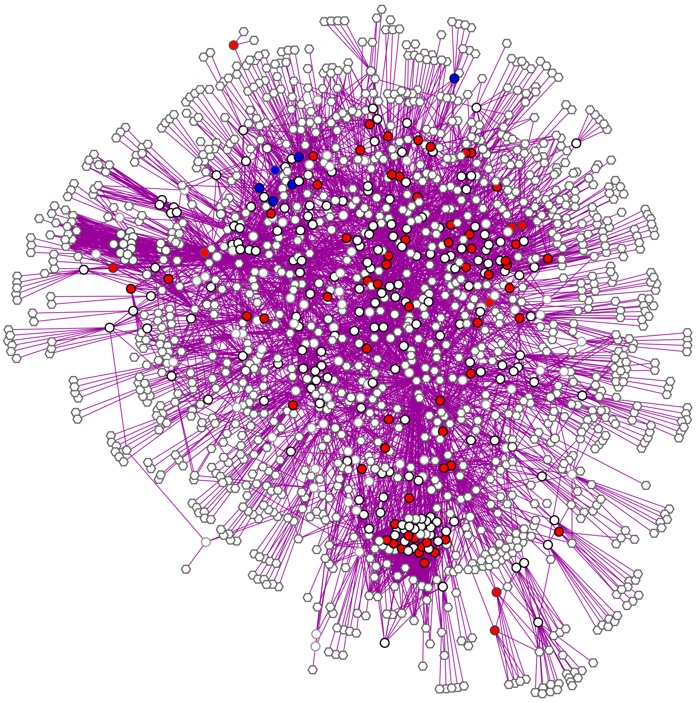


Figure 3.3: **The protein - GO term network in unfiltered form.** Proteins are represented as circular nodes and GO terms are represented as hexagonal nodes. The protein node border reflects a continuous color mapping between black and white relating to the probability of the protein identification where black indicates a high quality identification and white indicates a low quality identification. The protein node color corresponds to a continuous color mapping reflecting the relative fold change attribute. Red shades indicate an increase in the quantity of the protein in VIN13 relative to the amount in RO88. Blue shades indicate a fold increase in the quantity of the protein in RO88 relative to the amount in VIN13. Darker shades indicate a larger fold change. A white protein node color indicates that no relative fold change could be calculated for this protein. An edge between a protein node and GO term indicates that the GO term is ascribed to the protein.

---

module, the protein families derived from OrthoMCL analysis of the three species performed in Section 3.2.4 were used. Thus, a protein family module may consist of one or more S288C derived systemic IDs and one or more VIN13 and/or RO88 IDs. For the proteins which had to be grouped according to best reciprocal BLAST matches, modules were created from these BLAST-defined families. In the instances where there was a shared member between an OrthoMCL family module and a BLAST module, a new combined module with a unique ID was formed.

For the purpose of this workflow, family modules are characterised as one of two types, namely modules consisting of non ambiguous members and modules consisting of ambiguous members. Modules consisting of non ambiguous members have only one systemic ID, tied to one ortholog in VIN13 and/or RO88 respectively. Modules consisting of ambiguous members contain one or more systemic ID, tied to one or more orthologs in VIN13 and/or RO88. Proteins that matched equally well to more than one spectrum, and thus were unresolved by the TPP may also form part of these modules.

For the modules that contain multiple VIN13 and RO88 proteins of which a subset contain proteins with quantitative mass tag information, an agglomerative relative fold change signal was calculated for the module using the average of the VIN13/RO88 ratios for the proteins in that module.If proteins within the same module were found with opposing quantitative patterns, an attribute indicating this was created.

A network was constructed linking each protein family module to GO terms associated with the proteins within that module. This network (Figure 3.4) consisted of two types of nodes, namely protein family module nodes and GO term nodes. A protein family module node is connected to a GO term node via an edge if a protein within the module is associated with the GO term. Each protein family module has various attributes displayed in the attribute columns of the Cytoscape session, including module members, number of members, module type, agglomerative relative fold change (when applicable), which member proteins carry the quantitative signal, what member proteins have a secretion signal and if so in what database. A list of GO terms that describe all of the module members is also provided. Protein family modules and their attributes were constructed using custom written Perl programs.

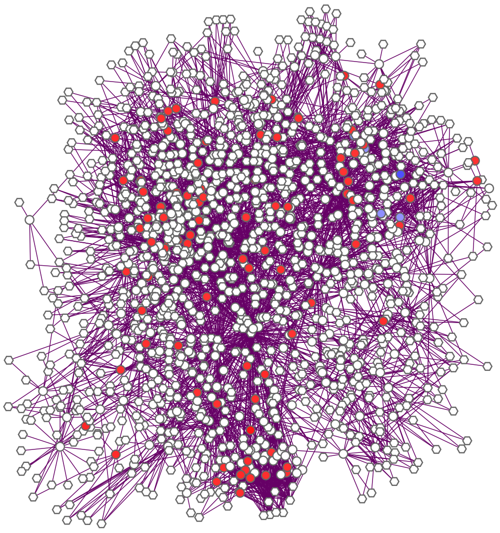


Figure 3.4: **Unfiltered protein family modules - GO term network.** The visual mapping style is the same as in Figure 3.3 except that the nodes here are protein family modules and not proteins.

3.2.8.5 Protein Modules GO Overlap Network

This network is constructed in a similar fashion to the protein-protein GO network described in Section 3.2.8.3, however, the nodes here are the protein family modules described in Section 3.2.8.4.

3.2.8.6 Protein - Metabolic Pathway network

The KEGG biochemical pathway data was used to construct this network, shown in Figure 3.5. The identified proteins are linked to the pathway nodes via their systemic ID, thus an identified protein will be linked to its corresponding pathway(s) with an edge. Pathways are linked together with an edge if they share a compound.

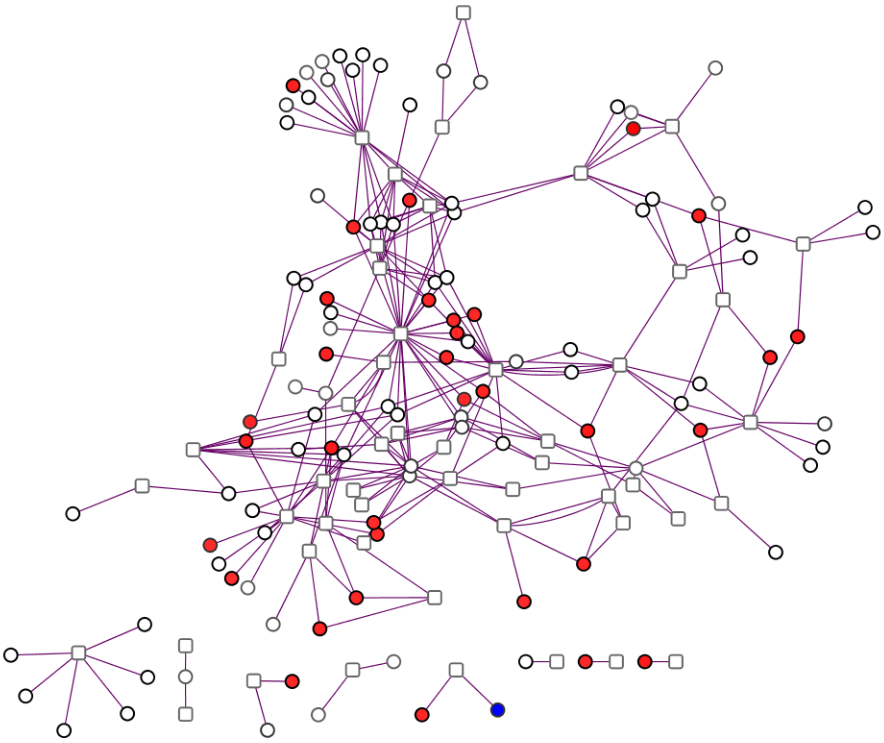


Figure 3.5: **Protein metabolic network.** Proteins are represented as circular nodes and the metabolic pathway nodes are rounded rectangles. The visual mapping style for the protein nodes are the same as in Figure 3.3. An edge between a protein node and a pathway node means that the protein is involved in the pathway.

3.2.8.7 Protein-Protein Interactome Networks

The BioGRID interactome database was used to construct the protein-protein interactome networks. If a known interaction existed between any of the identified proteins, an edge was formed between these two proteins. Each edge between interacting proteins was attributed with the nature of the interaction which entails the following attributes: the type of interaction which may be physical or genetic, the method used to detect the interaction as well as a publication identifier.

All genetic interactions are determined by genetic interference experiments and assays. The three types of genetic interference methods applicable here are additive genetic interactions defined by inequality, suppressive genetic interaction defined by inequality and synthetic genetic interaction defined by inequality.

The physical protein-protein interactions refer to direct physical interaction of two proteins or co-existence in a stable complex. The interaction detection methods relevant for this dataset include affinity chromatography, two hybrid assays, fluorescent resonance energy transfer, protein complementation assays, pull downs and enzymatic studies.

When the protein-protein interactome network is visualised in Cytoscape, its edge dense nature makes it difficult to visually interpret, thus two further methods were used to generate different views of the interactome network. Firstly, the network was split into genetic-based interactions and physical protein-protein interactions. These networks can be seen in Figure 3.6. The protein interactions were further split into a subset of protein interactions with more than one line of evidence as shown in Figure S5.

For both the gene-based and physical-based protein interaction networks, Minimum Spanning Tree (MST) views of the original networks were created (Figure S4). MSTs are sparser subnetworks connecting all the nodes of the original network and thus reflect and retain much of information that was in the original network whilst reducing complexity.

The edges of the original unfiltered interactome networks were undirected and all have equal weights. Kruskal's minimum spanning tree algorithm [33] was then applied to these unfiltered interactome networks which resulted in MSTs of the original networks respectively.

3.2.8.8 Protein Interactome Module Networks

In order to create modules of interacting proteins, the protein-protein interaction networks in Section 3.2.8.7 were clustered using the Markov Clustering Algorithm [34].

For the genetic interactome network an inflation value of 1.7 was used, resulting in the creation of 50 clusters. For the physical-based protein interactome network an inflation value of 4.9 was used which resulted in the creation

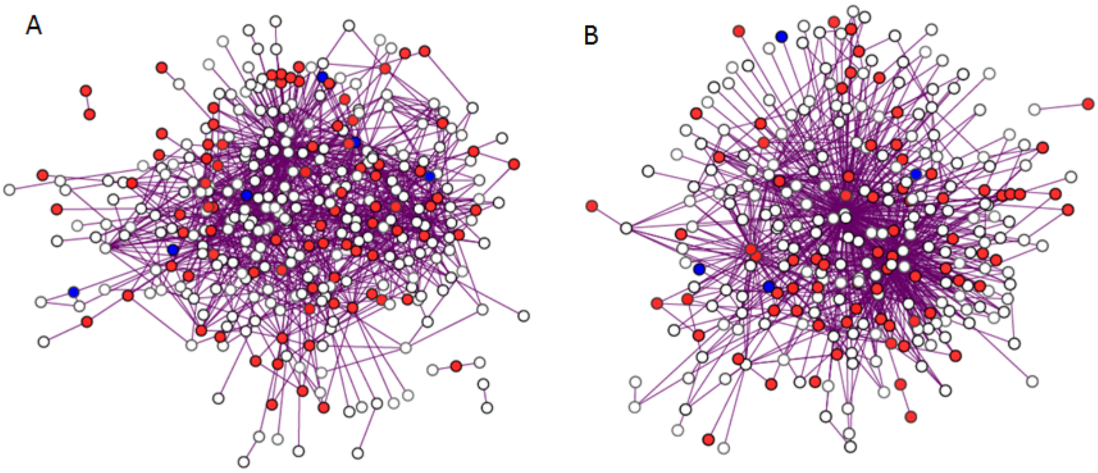


Figure 3.6: **Protein-protein interactome networks.** (A) and (B) are gene-based and physical-based protein-protein interaction networks respectively. The visual mapping style is the same as in Figure 3.3 except that there are no GO term nodes in this network.

of 52 clusters. Networks where the interactome modules connect to the GO terms of the proteins within the modules were created and can be seen in Figure 3.7.

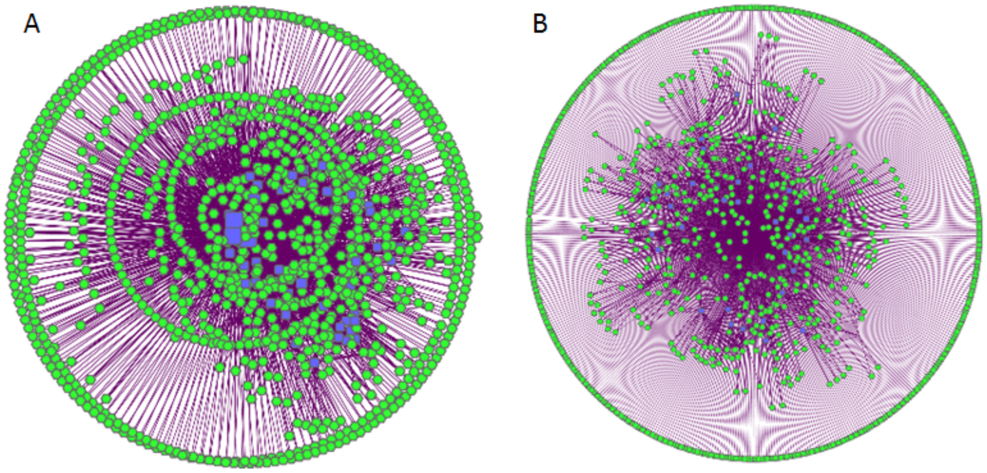


Figure 3.7: **Protein-protein interactome module to GO term networks.** Figures (A) and (B) are gene-based and physical-based protein-protein interaction module networks respectively. The blue rectangular nodes are interaction modules and the green nodes are GO terms.

Table 3.1: **Summary and figure reference for all networks constructed in Section 3.2.8.**

Network type	Figure reference
Protein - GO term	3.3
Protein family modules - GO term	3.4
Metabolic pathway	3.5
Protein interactome	3.6
Protein interaction modules - GO	3.7
Protein - enriched GO term for Cellular Component	3.10
Protein - enriched GO term for Molecular Function	3.11
Protein - enriched GO term for Biological Process	3.12
Protein - GO overlap	3.13
Protein - GO overlap 0.2	3.14
Protein interactome - MSTs	S4
Physical-based protein-protein interaction multiple lines evidence	S5

3.3 Results and Discussion

3.3.1 Overview

Presented and discussed in this section are the results of the network constructions described in Section 3.2.8. Any of the attributes provided can be used to filter and create subnetworks from the main networks. Such filters can be used in numerous combinations and thus effectively creates a relatively easy and quick method by which one can investigate the networks for existing hypotheses or formulate new hypotheses based on what is observed.

The examples of network interpretation presented in this section serve to illustrate the utility of the method and showcase instances where it is most valuable. However, it is not meant as an exhaustive interpretation of the data.

Given the size of the dataset and the multiple combinations of filtering criteria that may be used, a myriad of subnetworks can be generated. The relevance of these outputs are largely determined by the aims and objectives. One of the main goals of this workflow was to provide a set of contextualised networks from which one can explore the results of an experiment.

Table 3.1 is a summary of all the networks constructed (Section 3.2.8) and provides a figure reference.

3.3.2 Filtering Criteria

Any attribute assigned to a node or an edge in any of the networks may be used as a filtering criteria from which one can constrain the data within a given context. Furthermore, these criteria may be used in combination with one another and may span multiple network types. Some of the criteria are described in sections 3.3.2.1 through 3.3.2.3 below.

3.3.2.1 Protein Probability

The protein probability represents the probability that a protein has been correctly identified. Factored into the calculation of the protein probability are variables from both the peptide level and protein level, including the probabilities of the peptides that group with the protein, percentage of protein coverage and number of unique peptides [25]. The protein probability is thus a good metric of data quality and can be thresholded in order to create contextualised outputs at the desired level of error versus sensitivity. The ability to filter the data at the contextualised level is advantageous because it is easy to see what data is removed by a given filtering criteria and why the data is excluded. This provides one with more information from which to make decisions about thresholds in a data dependant manner as opposed to applying arbitrary cutoffs. In order to interpret the networks, it is advisable that the identified proteins first be filtered based on their probability. A probability threshold may be chosen based on the information obtained from protein prophet. Figures S2 and S3 illustrate this information for the identifications made in VIN13 and RO88 respectively.

3.3.2.2 Differential Identifications

Differential identification refers to the ability to infer the presence of a protein expressed by one organism and not by the other in a mixed sample context. However, caution must be taken when interpreting the differential identification results. What can be inferred from these results is that a given protein could only be detected in one organism and not in the other using the mass spectrometry and data analysis workflow described. It does not necessarily mean that the protein was not present in the sample. There can be multiple reasons why a protein was identified in one organism and not in the other. The following scenarios are all plausible for this dataset. The peptides of a protein may be present in very low quantities in one species relative to the other, leading to a detection bias against that protein. The thresholds which are used to determine ortholog relationships may, if set too stringently, exclude certain legitimate ortholog relationships. Conversely, if the ortholog criteria is too lax, illegitimate ortholog assignments will lead to an over-represented count of proteins identified in both species.

3.3.2.3 Relative Fold Change

When interpreting the relative protein fold changes it is crucial to take heed of the number of peptides that contribute to the quantitative signal. Multiple quantitative peptides are required in order to calculate a standard deviation for the quantitative signal. Special care must be taken when interpreting the relative protein fold change in cases where there was only one quantitative peptide available from which to calculate the relative protein fold change.

3.3.2.4 GO Terms

The Gene Ontology provides a useful controlled vocabulary that can be used to contextualise the data. In addition, the Gene Ontology consists of three distinct hierarchies or categories represented as independent directed acyclic graphs (DAGs) or directed acyclic networks [16]. In this work, the three hierarchies, namely molecular function, biological process and cellular component offer a valuable means for contextual constraint of the data since the GO category attribute can be used to create subnetworks.

3.3.2.5 GO Enrichment

When GO terms are statistically significantly enriched in this context, it means that the GO terms connected to the proteins are over-represented in the dataset. Thus, constraining network outputs to only enriched GO terms and their proteins may provide biological insight.

3.3.2.6 Proteins with a Secretion Signal

Since this experiment involved analysis of the secretome, one of the desired outcomes was to see if any proteins with known secretion signals could be observed in the data. In total, 15 proteins with known secretion signals were detected. Of these 15 instances, 10 were detected in SignalP-4.1 only, two instances were detected in the Fungal Secretome database only, another two instances were detected in SPdb only and only one instance had a secretion signal defined by both the Fungal Secretome database and SignalP-4.1 database.

3.3.3 Overview of Results

A total of 396 proteins are presented and contextualised as a result of the workflow. 169 of the 396 proteins were identified in both VIN13 and RO88, whilst 115 proteins could only be identified in VIN13 and 112 only identified in RO88. Figure 3.8 shows these results as a Venn diagram.

Of the 169 proteins identified in both species, a relative fold change could only be calculated for 111 of these proteins. Of the 111 proteins with a calculated fold change, 81 instances had only 1 peptide that met the criteria set by

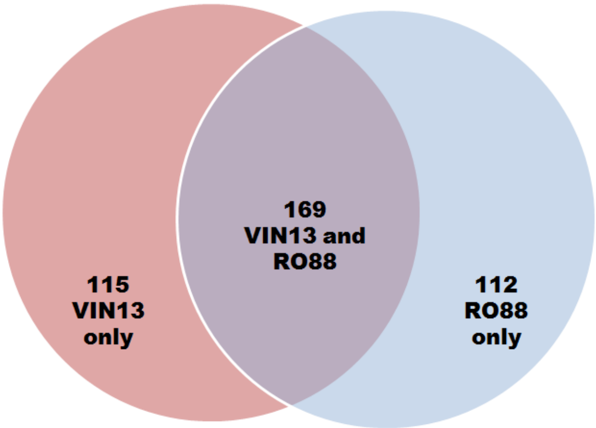


Figure 3.8: **The amount of proteins identified in both species and in VIN13 and RO88 only.** A total of 396 proteins are presented and contextualised as a result of the workflow. 169 of the 396 proteins were identified in both VIN13 and RO88, whilst 115 proteins could only be identified in VIN13 and 112 only identified in RO88.

the method for calculation of the relative fold change. The other 30 instances had 2 or more peptides available for the calculation of the relative fold change with the highest amount of peptides reaching 5. Figure 3.9 shows the result as a Venn diagram. Furthermore, of the 111 calculated fold changes, only 77 were statistically significant. 70 of these 77 indicated a higher relative protein abundance in VIN13 with only 7 instances indicating higher relative protein abundance in RO88.

3.3.4 The Protein - GO Term Networks

The unfiltered Protein - GO term network (Figure 3.3) consists of 1501 nodes and 4498 edges. 396 of the nodes are protein nodes and 1104 are GO term nodes. 561 of the GO term nodes belonged to domain biological process, 189 to cellular component and 354 to molecular functions. A GO term may connect to multiple proteins, whilst a protein may connect to multiple GO terms. As can be seen in Figure 3.3, this network is very dense. In order to sparsify this network, it was filtered to retain only enriched GO terms. 95 GO terms were statistically significantly enriched in the dataset, 41 of these belonged to the category biological process, 21 to cellular component and 34 to molecular function. Protein - GO term subnetworks for each of the categories were created using only the enriched GO terms and are displayed in Figures 3.10, 3.11, and 3.12.

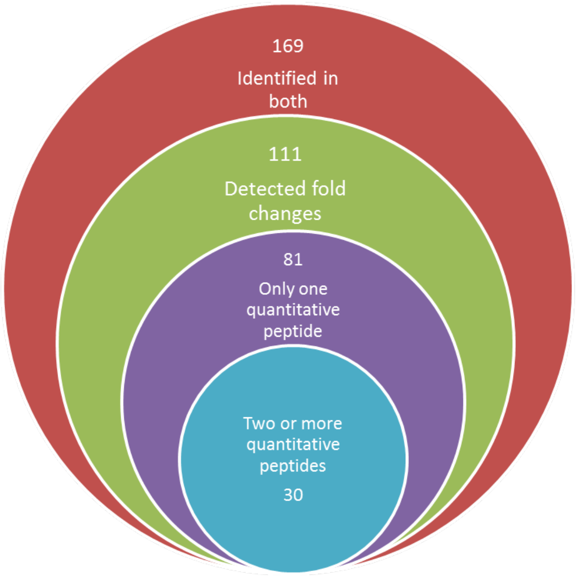

169
Identified in both
111
Detected fold changes
81
Only one quantitative peptide
Two or more quantitative peptides
30

Figure 3.9: **The amount of proteins for which fold changes could be calculated.** Of the 169 proteins identified in both species, fold changes could only be calculated for 111. However, of this 111, 81 proteins had only one quantitative peptide that met the criteria set in Section 3.2.5, whilst only 30 proteins had 2 or more peptides from which to calculate a relative fold change. Furthermore, only 77 of the fold changes were found as statistically significant by the method.

3.3.5 Protein - Protein GO Overlap Networks

The protein-protein GO overlap network illustrates the similarity between proteins based on their functional annotation in terms of associated GO terms, quantified using the Jaccard index. Table 3.2 shows the decrease in the total number of edges as the Jaccard index threshold is increased. At a Jaccard index threshold of 0.0 every protein is connected to at least 1 other protein, since every protein, if it is defined within the gene ontology, must have at least a high level GO term ascribed to it. At a Jaccard index of 0.1, proteins with very little connection to other proteins are revealed, whilst at a Jaccard index of 0.9 only proteins with high similarity in terms of the GO terms ascribed to them remain connected. It was also observed that systemic IDs that cluster in the same protein families were also described by many of the same or exactly the same GO terms.

Figure 3.13 illustrates the information in Table 3.2 and demonstrates that using a set overlap measure such as the Jaccard index may be a useful filtering criteria. Furthermore, this approach can be used to identify groups of functionally similar proteins at predetermined threshold values and is also useful to find protein instances where the annotation appears non-obvious. Thus, a network-based view of GO term overlap provides one with a different perspec-

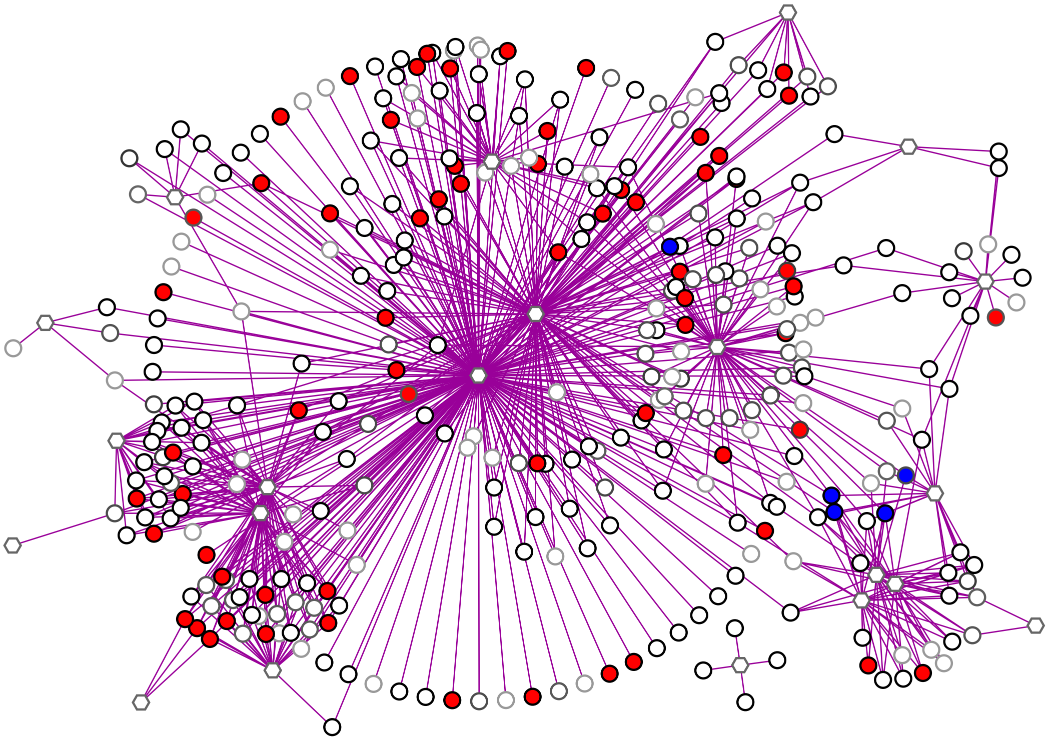


Figure 3.10: **Network of enriched cellular component GO terms.** This is a subnetwork of the protein - GO term network constrained to only enriched GO terms belonging to the category cellular component and the proteins that are associated with this subset of GO terms. The visual style is the same as in Figure 3.3.

Table 3.2: **The number of edges that exist at different Jaccard index thresholds as determined for the protein-protein network.**

Jaccard Index Threshold	Number of Edges
0.0	42383
0.1	10495
0.2	3006
0.3	1762
0.4	1205
0.5	782
0.6	509
0.7	311
0.8	177
0.9	90

tive on the dataset from which additional and possibly obscured information about the proteins can be gleaned. More specific results from this approach

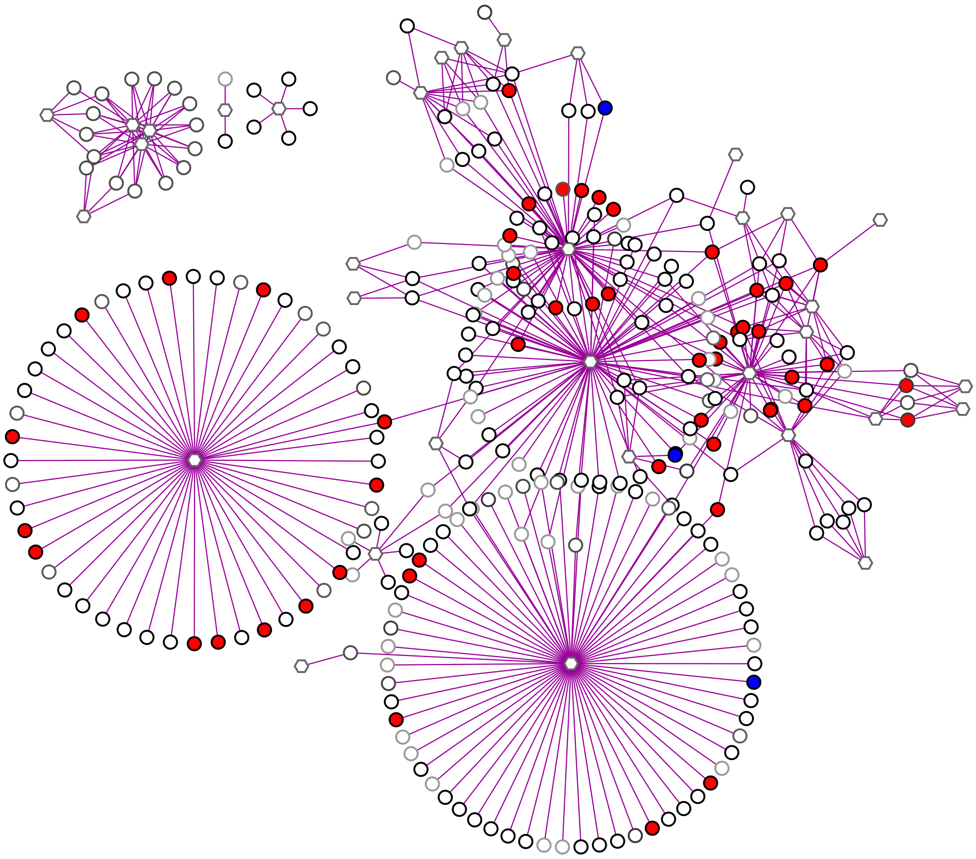


Figure 3.11: **Network of enriched molecular function GO terms.** This is a subnetwork of the protein - GO term network constrained to only enriched GO terms belonging to the category molecular function and the proteins that are associated with this subset of GO terms. The visual mapping style is the same as in Figure 3.3.

---

are described in Section 3.3.5.1.

3.3.5.1 Protein - Protein GO Overlap Network with a Jaccard Index Threshold of 0.2

A subnetwork of the protein-protein GO overlap network was created where an edge can only exist if the Jaccard index between two nodes is greater than or equal to 0.2. This network is displayed in Figure 3.14A. Apparent in this network were 22 individual proteins and four groupings that became disconnected, indicating that these proteins share only a few GO terms with the other proteins in the dataset.

Table 3.3 contains a summary of some of the attributes for the proteins in the GTPase group. Now that the view has been constrained to a group of interest one may choose to decide on other filtering criteria, such as probability

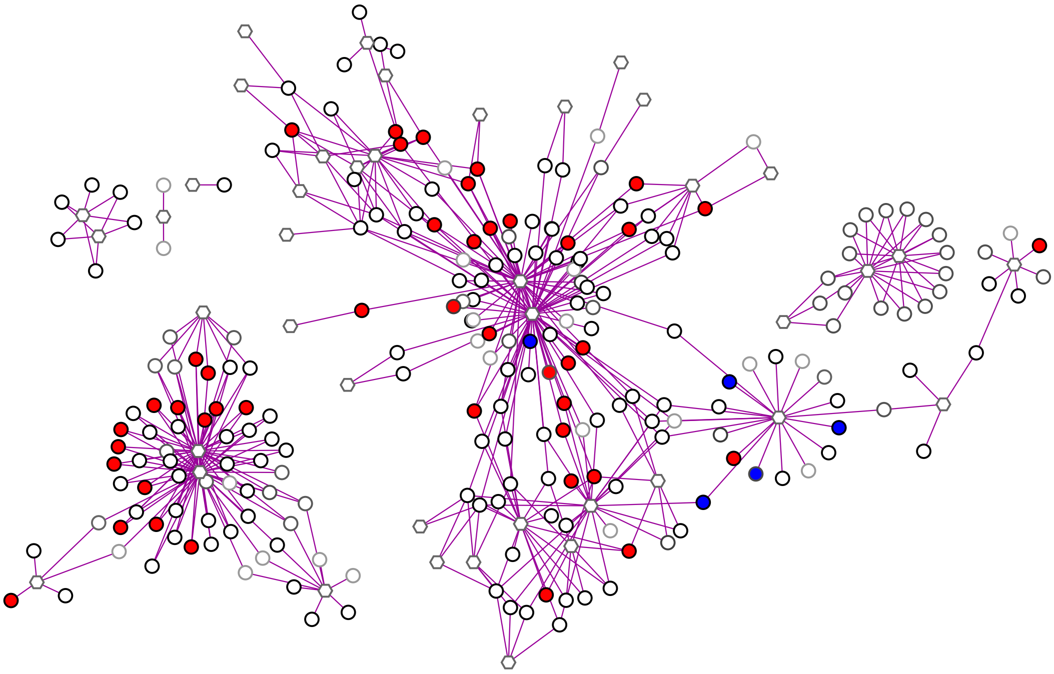


Figure 3.12: **Network of enriched biological process GO terms.** This is a subnetwork of the protein - GO term network constrained to only enriched GO terms belonging to the category biological process and the proteins that are associated with this subset of GO terms. The visual mapping style is the same as in Figure 3.3.

Table 3.3: **A subselection of attributes belonging to the proteins of the GTPase group.** In the protein probability column the VIN13 probability is given first followed by the RO88 probability. If the probability of identification for the proteins was the same in both species, only one number is given. The number of GO terms and number of enriched GO terms columns give the amount of GO terms that describe the proteins and the number of enriched GO terms respectively.

Protein ID	Identified in species	Probability of identification	Number of GO Terms	Number enriched GO terms
YOR101W	RO88-only	0.9628	13	3
YLR289W	VIN13-only	0.2829	13	3
YLL001W	Both	0.3594 ; 0.3032	18	2
KRH2	RO88-only	0.2818	14	2
SRA1	RO88-only	0.4033	11	3
GLC5	Both	0.9812 ; 1	17	3

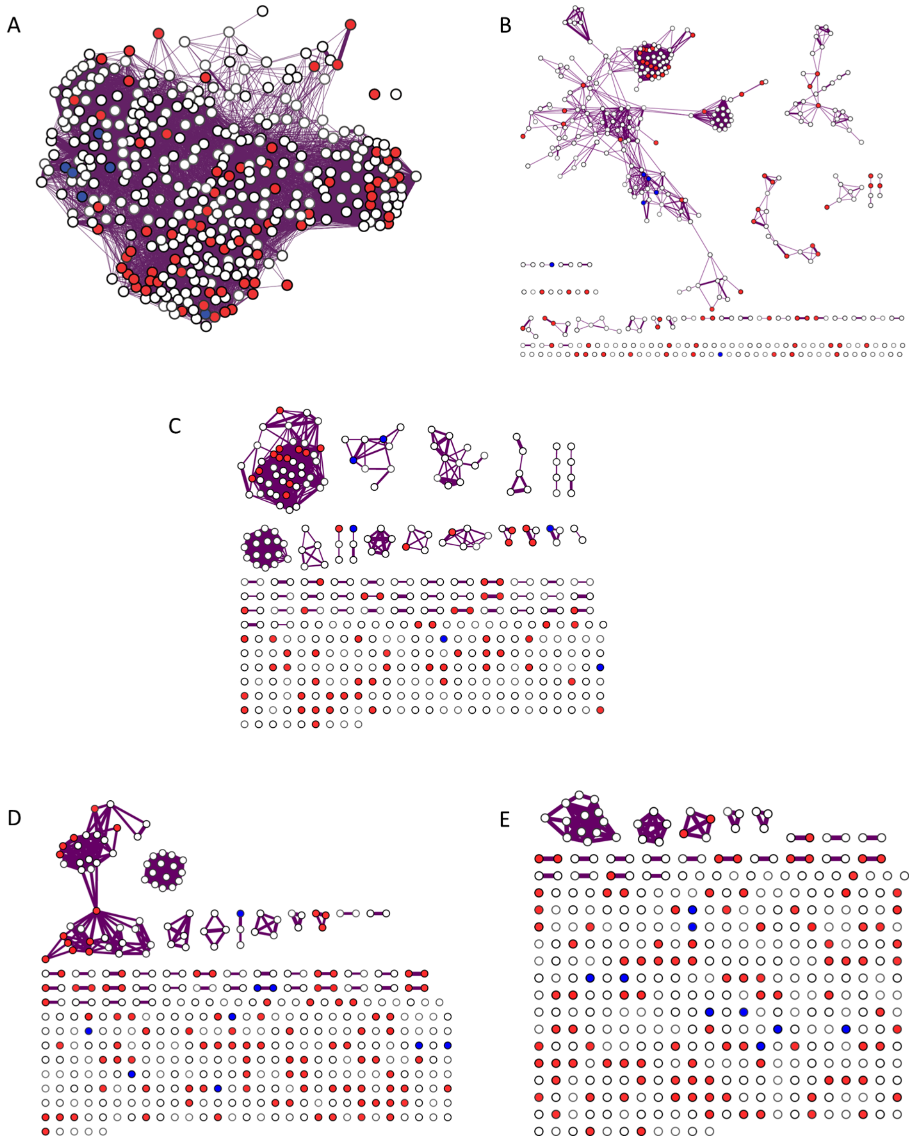


Figure 3.13: **Protein-protein GO overlap networks.** (A) - (E) are protein-protein networks with Jaccard index thresholds of (A) 0.1 (B) 0.3 (C) 0.5 (D) 0.7 (E) 0.9. The networks consist only of protein nodes and the visual mapping style for the nodes is the same as in Figure 3.3. Edge sizes are scaled according to the Jaccard index between the two connected nodes, a thicker edge represent a higher Jaccard index.

of identification, to further filter the output. In this case, a justifiable choice may be made to single out proteins such as KRH2 because it was identified in only one of the target species at a relatively low probability. What this example illustrates is the ability and relative ease with which the method allows a grouping of proteins that shares biologically relevant connections to be identified and interpreted prior to the application of a data quality threshold.

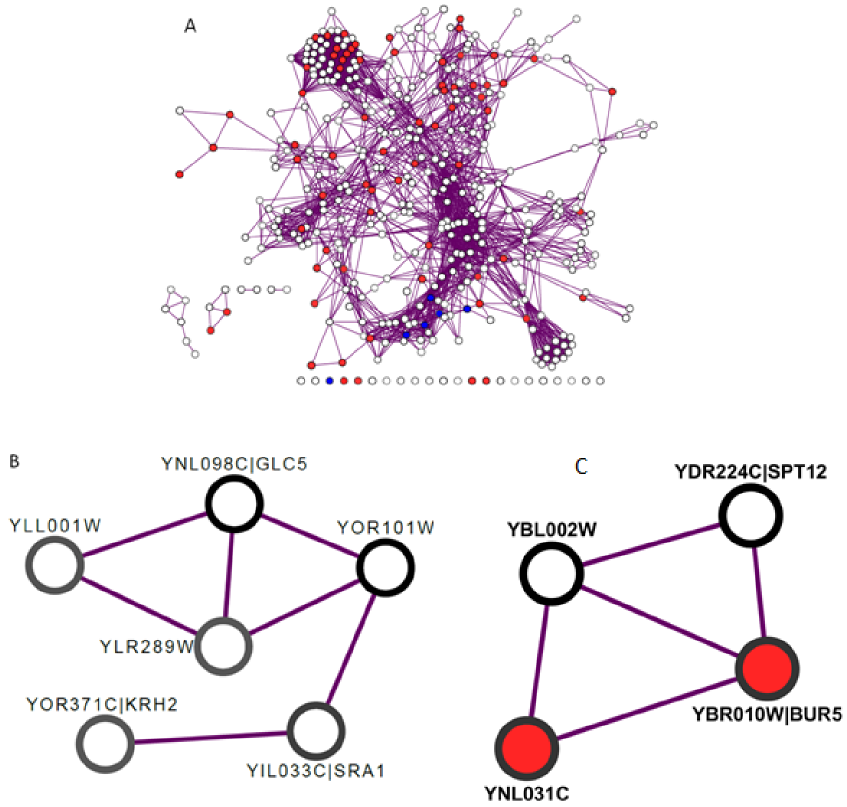


Figure 3.14: **The protein-protein GO overlap network at a Jaccard index threshold of 0.2.** (A) is the entire protein-protein GO overlap network at this threshold. (B) and (C) are the GTPase group and Histone group subnetworks respectively. The visual mapping style and network type is the same as in Figure 3.13.

This is in contrast to approaches where relatively arbitrary determined data quality thresholds are applied first, and what remains of the data thereafter is then interpreted.

Another one of the four groupings that become disconnected at the 0.2 threshold is illustrated in 3.14C (hereafter referred to as the Histone group) and is comprised of YNL031C, BUR5, YBL002W and SPT12. This example was chosen to illustrate how the workflow described is capable of dealing with ambiguities and redundancies encountered with data of this kind.

YBL002W is Histone H2B and is identified only in VIN13 with a protein probability of 1.0 and has 9 GO terms associated with it, none of which are enriched. SPT12 is also described as Histone H2B and is identified only in RO88 with a protein probability of 0.998 and has 11 GO terms ascribed to it, none of which are enriched. 9 GO terms are shared between SPT12 and YBL002W, however, SPT12 has 2 GO terms not ascribed to YBL002W.

Unlike YBL002W and SPT12, YNL031C and BUR5 are both described as Histone protein H3 and correspond to the same set of identified mass spectra. They are identified in both VIN13 and RO88 with protein probabilities of 0.6328 and 0.6523 respectively and these proteins were found to be 6.14 fold more abundant in VIN13. YNL031C has 7 GO terms ascribed to it, none of which are enriched, whilst BUR5 has 9 GO terms ascribed to it, none of which are enriched. YNL031C and BUR5 are flagged as ambiguous because they belong to a multiple ortholog group and they correspond to two proteins in RO88 to which the spectra match equally well.

Consequently, YNL031C and BUR5 are best viewed in the context of the protein family modules networks discussed in Section 3.2.8.4. By doing this, it becomes clear that both YNL031C and BUR5 are grouped in the same module, namely Ortholog module 81, which has 5 members (Figure 3.15). We can now also easily obtain a non-redundant view of this module and the GO terms ascribed to the member proteins of this module.

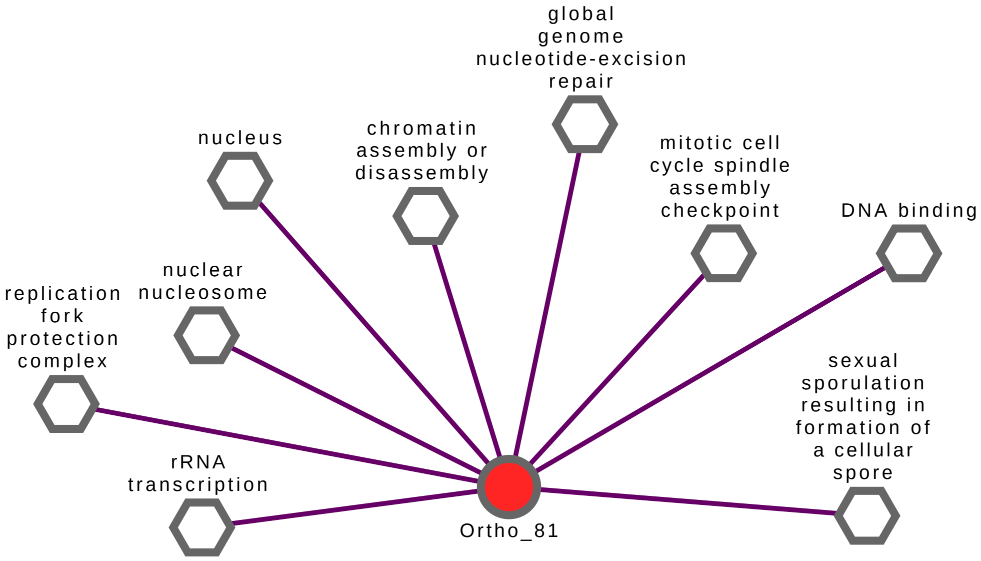

global genome nucleotide-excision repair
nucleus
chromatin assembly or disassembly
mitotic cell cycle spindle assembly checkpoint
DNA binding
nuclear nucleosome
replication fork protection complex
sexual sporulation resulting in formation of a cellular spore
rRNA transcription
Ortho_81

Figure 3.15: **Orthology module 81 with the GO terms that describe it.** The members of Ortholog module 81 are YNL031C, BUR5, YBL002W and SPT12 (histone group). Family modules are represented as circular nodes and GO terms are represented as hexagonal nodes. The size of the family module node corresponds to the number of proteins in the module. An edge between a protein family module node and a GO term indicates that the GO term is ascribed to the protein family module. The visual mapping style is the same as in Figure 3.3 except that the nodes here are protein family modules and not proteins.

The above example illustrates how one may use multiple network-based contextualisations of the data to interpret a given grouping of proteins. Furthermore, the use of the protein family modules network illustrates how redundancy created by annotation may be obviated in an automated fashion.

3.3.6 Proteins Relating to the Yeast Cell Wall

The fungal cell wall is a highly dynamic cellular organelle with four major functions, namely the stabilisation of internal osmotic conditions, protection against environmental and physical stresses, maintaining the shape of the cell and acting as a scaffold to which proteins can attach [35].

The yeast cell wall has two layers and consists mainly of polysaccharides with three sugars, namely mannose, glucose and N-acetylglucosamine as the predominant building blocks [36]. The inner layer is composed of β-1,3 glucan and chitin, whilst the outer layer consists of mainly β-1,6 glucan and heavily glycosylated mannoproteins [35]. Given the vital functions of the cell wall, it is of great interest from both a fundamental biological perspective and an applied point of view. In this section we focus especially on parietal proteins and/or yeast mannoproteins identified in the dataset with positive oenological properties and the role that they fulfill in the cell wall, as well as those proteins that are actively secreted into the growth media.

3.3.6.1 Oenological Functions of Parietal Yeast Proteins

Several oenological functions of parietal yeast mannoproteins have been described: 1) Yeast mannoproteins can combine with anthocyanins and tannins in wine leading to increased colour stability [37] and decreased astringency resulting in a wine with more body, better mouthfeel and with an increased resistance to oxidation [38]; 2) The growth of malolactic bacteria in wine is stimulated by the presence of parietal mannoproteins [39]; 3) Crystallization of tartrate salt can be prevented with the use of mannoproteins and can thus aid in achieving tartrate stability in wine [40]; 4) Mannoproteins and aromatic compounds may interact during the winemaking process and this interaction occurs especially during ageing of the wine on the lees. Interactions between yeast proteins and aromatic metabolites in wines can lead to modifications of volatility and aromatic intensity of wines as well as contribute to overall aroma stability in wines [38]; 5) Heat stability can be conferred to wines due to the presence of certain mannoproteins [41]; 6) Mannoproteins play a considerable role in the adsorption of Ochratoxin A [42; 43; 44], a dangerous fungal secondary metabolite often found in grapes, grape juices and wines [45]; 7) The passive release of molecules due to yeast autolysis while wine ages on the lees increases the mannoprotein level and the amount of yeast-derived amino acids in the wine [46]. It is believed that this winemaking practice may protect the wine from oxidation and add to the complexity of aroma and flavour to the wine; 8) Mannoproteins are of importance for wines where the *flor* technique is applied. Film-forming yeasts or flor yeasts spontaneously develop on the surface of the wine, forming a thick mat of cells called the velum. Velum yeast posess a 49-kDa hydrophobic cell wall mannoprotein which correlates with velum formation and surface hydrophobicity [47]; 9) Mannoproteins are also of particular interest in the manufacture of several sparkling wines for the role that they play in the flocculation of yeast strains [48; 49]; 10) A common problem during the production of white wines is the formation of haze, a phenomina that occurs predominantly due to the relatively slow rate at which grape proteins denature and precipitate. Several glycoproteins have been observed to reduce visible haziness by decreasing the particle size of the haze [50; 51; 52; 53].

The amount and type of proteins released by yeast during the wine making process and ageing on the lees is very much dependant on the specific yeast strain used and the nutritional conditions of the must [38]. Furthermore, it has been shown that the strain of yeast used determines the influence that the mannoproteins have and that the proteins released during the fermentation process itself are more reactive than those released during yeast autolysis [37].

Due to their positive oenological properties [38] and wine haze reduction potential [50], mannoproteins and the genetic determinants involved with their release were a group of high interest for this study. Proteins of interest included targets such as the haze protection factor proteins (HPFs) YOL155C (also known as HPF1) and YDR055W (also known as HPF2 or PST1) [54], β-1,6 exoglucanases such as EXG1 (YLR300W) [51] and proteins involved with chitin metabolism such as Chitin synthase III (YBR023C) [52] and Chitin transglycosylase (YGR189C) [53].

A current hypothesis suggests that many of the proteins listed above are secreted into the growth media and are produced in higher abundance in RO88 relative to VIN13 [14]. In order to find evidence in support of or against this hypothesis, known relevant target proteins were investigated in the network context. Also, all of the constructed networks were queried for the string "cell wall" starting with the protein to GO term network. By executing this query, all proteins that match this string in their description or in the GO terms ascribed to them were returned and a cell-wall-themed subnetwork was created. This network allowed the investigation of all the identified proteins under the cell wall theme with the added GO context.

The cell wall subnetwork in Figure 3.16 consists of 45 nodes of which 36 are proteins and nine are GO terms. Statistically significant fold changes were found for seven of the proteins, four of which were more abundant in RO88, whilst three were found to be more abundant in VIN13. Seven of the proteins also have known secretion signals. Of the nine GO terms in the network, "cell wall", "fungal-type cell wall" and "fungal-type cell wall organization" are of the highest degree. These GO terms are relatively high level GO terms and thus give somewhat unspecific knowledge about the proteins that they describe. The other six GO terms, such as "cell wall mannoprotein biosynthetic process", are more specific.

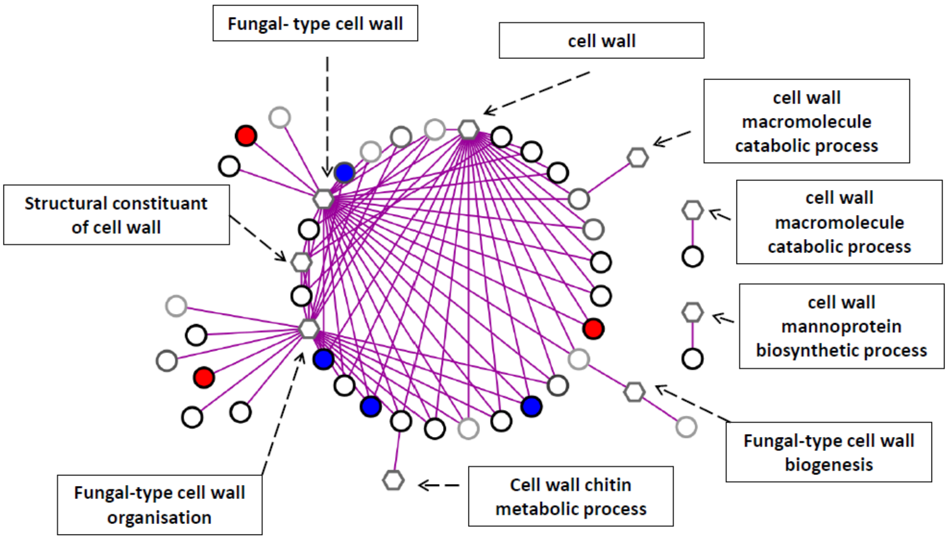

Fungal- type cell wall
cell wall
cell wall macromolecule catabolic process
Structural constituent of cell wall
cell wall macromolecule catabolic process
cell wall mannoprotein biosynthetic process
Fungal-type cell wall biogenesis
Cell wall chitin metabolic process
Fungal-type cell wall organisation

Figure 3.16: **Subnetwork of the protein-GO term network filtered for all nodes relating to the cell wall.** The visual mapping style is the same as in Figure 3.3. The dashed arrows point to GO terms.

The cell wall subnetwork in Figure 3.16 can be further filtered by selecting for only proteins with significant fold changes between the species. This subselection of proteins can then be visualised in the GO term overlap network and is illustrated in Figure 3.17.

The protein pair with the highest amount of GO overlap is HPF2 and SSR1 with a Jaccard index of 0.6. A selection of low level enriched GO terms for this subnetwork include terms such as "membrane", "fungal-type cell wall", "fungal-type cell wall organization", "extracellular region" and "anchored to membrane". Table 3.4 provides a selection of attributes that describe the seven proteins in this network.

All seven proteins shown in Figure 3.17 are of interest for fundamental cell wall biology and their possible positive oenological traits which are described in Section 3.3.6.1. These proteins are discussed in more detail below.

3.3.6.2 HPF2

YDR055W, also known as HPF2, is a cell wall mannoprotein that is capable of reducing the particle size of aggregate proteins, however, the mechanism by which it confers this haze protective ability is not yet fully understood [54]. It has been found that HPFs do not prevent wine proteins from forming aggregates, instead it is the manner in which the wine proteins aggregate that is altered [55]. Furthermore, it has been suggested that HPFs act by competing with wine proteins for other wine components and by this mode of action prevent the formation of protein aggregations which are large enough

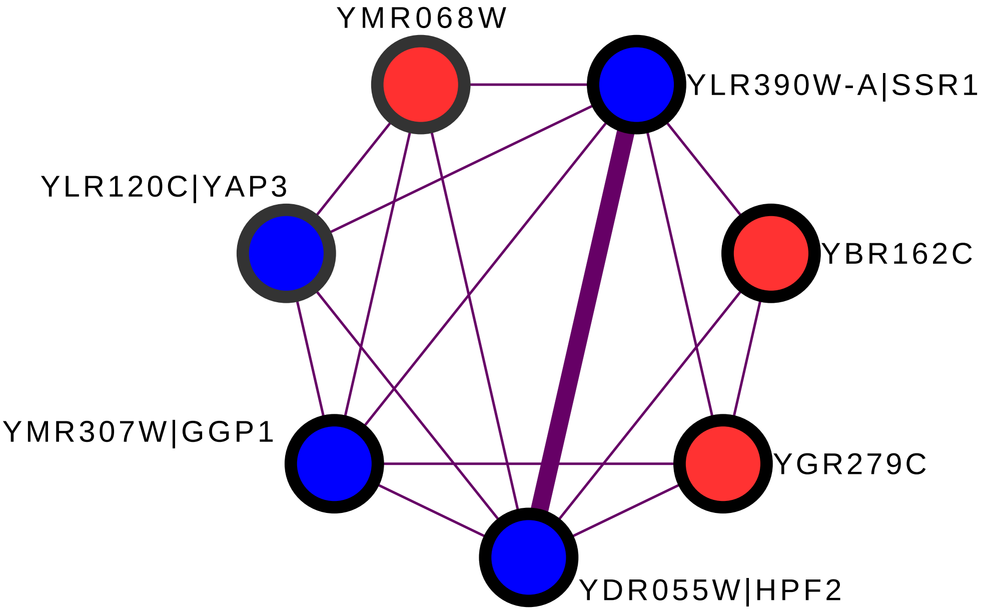


Figure 3.17: **A cell wall subnetwork of the protein-protein network with edges weighted by Jaccard index.** The visual mapping style is the same as in Figure 3.13.

to be detected as haze [54].

As shown in Table 3.4, HPF2 has a secretion signal and has been shown to be secreted by regenerating protoplasts [56]. The data confirms the presence of HPF2 in the secretome at higher relative abundance in RO88 when compared to VIN13. However, it has been suggested that the glycan structure and possible strain-specific manner of post translational glycan modification is of high importance for the role that this protein plays in white wine haze protection in addition to the quantity at which this protein is present [57].

3.3.6.3 Structural Cell Wall Proteins

Much of the current research pertaining to cell wall rigidity has focused on the polysaccharide components. However, it has been suggested that cell wall proteins should also be considered as important for cell wall rigidity [58]. Many cell wall proteins are modified by the addition of short O-linked sugar chains. It is thought that these sugar chains affect the secretory process of cell wall proteins and may directly contribute to cell wall rigidity [59]. Several of the proteins that affect the structure of the cell wall in either a direct or indirect manner identified in this dataset are illustrated in Figure 3.17 and summarised in Table 3.4. These proteins include SSR1, YBR162C, YGR279C and YMR068W, each of which is discussed in more detail below.

The main layer of the cell wall is believed to be a mesh-like structure consisting of proteins, 1,6-β-glucan and chitin that are cross-linked with the

Table 3.4: **A subselection of attributes belonging to the proteins of the cell wall subnetworks.** In the Probability of identification column the VIN13 probability is given first followed by the RO88 probability. If the probability of identification for the proteins was the same in both species, only one number is given. The number of quantitative peptides column indicates the number of peptides that were available to derive the fold change. The fold change column gives the relative fold change ratio for VIN13/RO88 as calculated in Section 3.2.5. The secretion signal column may contain either a Yes or No variable to indicate if the protein has a secretion signal defined within the databases.

Protein-ID	Probability of identification	Fold change	Number of quantitative peptides	Secretion signal
HPF2	1	-1.22	3	Yes
GGP1	1	-1.51	2	No
SSR1	0.9999	-1.33	1	No
YAP3	0.59 ; 0.6554	-2.64	1	No
YBR162C	1	1.45	1	Yes
YGR279C	1	1.49	3	No
YMR068W	1	2.33	1	No

side chains of 1,3-β-glucan. However, before cross-linking with the 1,3-β-glucan side chains can occur these side chains need to be modified. YMR307W, also known as GGP1, is a β-1,3-glucanosyltransferase and is believed to provide the enzymatic activity for the modification of the side chains [60], hence performing a vital role in the formation of the fungal cell wall. Another protein that modifies components of the cell wall is YGR279C (also known as SCW4) which is similar to glucanases. The paralog for YGR279C, namely YMR305C (also known as SCW10), was also identified in both species, however, no fold change could be calculated. It was previously found that SCW4 and SCW10 may play a direct or indirect role in the anchoring of proteins to β-1,6-glucan [61; 62; 63]. Furthermore, it has been suggested that, in addition to glucanase function, SCW4 and SCW10 also act as transglucosylases that provide the the necessary glucan polymers required for stabilisation of the fusion stage during yeast mating [63].

Whereas both GGP1 and SCW4 provide direct enzymatic modification of cell wall components, a protein, namely SSR1, that can be described as a core structural cell wall component, was also observed in this dataset. SSR1 is a glycoprotein that is located in the inner layer of the cell wall and associates with glucan [64]. YMR068W, a protein with regulatory function, was also identified. YMR068W (also known as AVO2) is a component of a protein complex containing the Tor2p kinase and other proteins referred to as the TOR complex 2 (TORC2). TORC2 has two main known functions: Firstly, it is required for progression in the G1 phase of the cell cycle and also signals initiation of translation. It shares these functions with its homolog TORC1. Secondly, TORC2 is involved in the polarised distribution of actin in the cytoskeleton and this function is unique to TORC2. It has been proposed that given these two functions, TORC2 possibly integrates temporal and spatial control of cell growth [65].

Another protein that is covalently bound to the cell wall is YBR162C, also known as TOS1. This protein is currently of unknown function, however, mutants with this gene knocked out are highly resistant to treatment with *beta*-1,3-glucanase [66]. Furthermore, a transcription factor study using DNA microarrays revealed that *YBR162C* is a target for the SBF transcription factor which is under the control of cell cycle regulation [67]. Genes activated by SBF are predominantly involved in yeast cell budding, and in membrane and cell-wall biosynthesis [67]. TOS1 also has a predicted secretion signal, is upregulated in VIN13 and seeing that its function is still unknown, it may be a protein of potential oenological interest.

TOS1, SCW4 and AVO2 were all present in VIN13 in higher abundance when compared to RO88, whilst CCW14 and GGP1 are more abundant in RO88. Given the roles and differential expression of these proteins by VIN13 and RO88, it is possible that these two yeast differ quite significantly in terms of their respective cell wall composition and regulation even when grown in isolation under the same conditions.

3.3.6.4 Yapsins

YLR120C (also known as YAP3 or YPS1) is an aspartic protease and belongs to a family of five glycosyl phosphatidyl inositol-linked aspartyl proteases also known as yapsins. The paralog for YPS1, namely YPS2, was also identified in the dataset, but was identified in RO88 only with a relatively low protein probability of 0.4728. The yapsin family of proteases are believed to process cell wall proteins involved in the maintenance of cell wall integrity [68]. YPS1 is active on the cell surface [69] where it is able to cleave at clusters of basic amino acids (C terminal to basic residues) within peptides and proteins [70]. Expression of *YPS1* is induced during periods of cell wall stress and remodelling as shown by genome-wide expression experiments [71; 72; 73; 74]. Further evidence supporting the induction of *YPS1* can be found in the results of quantitative immunoblotting experiments [75]. When the cells were shifted from 24°C to 37°C , YPS1 levels increased 12-fold. Also, YPS1 could not be detected with immunofluorescence at 24°C but was detected at 37°C showing fluorescence localised at the plasma membrane.

In summary, YPS1 is expressed in a temperature dependant manner [75] and according to the data it is upregulated RO88 when compared to VIN13. Furthermore, there is a difference in the optimum growth temperature for VIN13 and RO88. Thus, given these facts, YPS1 and indeed the yapsin protein family, are interesting targets for their possible oenological role, especially in RO88.

3.3.7 Proteins Relating to Malo-Ethanolic Fermentation (MEF)

In order to make sense of the data, parameters such as the growth conditions and time point of sample extraction must be kept in mind. The yeasts were grown separately in MS300 media (Section 3.5.2.1) and fermentations were allowed to run to dryness. In other words, just prior to the point of protein extraction glucose was limiting in the media.

For the greater portion of time during the growth of the yeast used in this study, metabolism is fermentative for both species and the Crabtree regulatory system [76] is in effect. Under these conditions mitochondrial activities are restricted [77] and carbon flow is steered away from biosynthesis towards ethanol production [78]. However, even under fermentative conditions, some biosynthetic activity is still required and is essential for the survival of the organism [79]. Biosynthetic processes produce NADH and consume NADPH resulting in a redox imbalance with NADH needing to be reoxidised. Alcoholic fermentation is a redox neutral process and can thus not account for the reoxidation of assimilatory NADH. In *S. cerevisiae* and other yeast, the formation of glycerol is a well understood mechanism by which the redox balance is restored [80]. An auxillary pathway for the regeneration of NADH involving malic acid and the malo-ethanolic pathway in yeast has also been proposed [81].

A key difference of oenological interest between *S. paradoxus* strain RO88 and other members of the *Saccharomyces* genus is the ability of RO88 to reduce the amount of L-malic acid in the must via malo-ethanolic (ME) fermentation, whilst still being able to produce a wine of good quality [82].

Notable variations in the degradation of L-Malic acid within the *Saccharomyces sensu stricto* group have been observed. The degradation of L-malic acid also appears to correlate with the optimal growth temperature of the individual strains. L-Malic acid synthesis was observed with cryotolerant species of *S. cerevisiae* such as *S. bayanus*, *S. pastorianusS. uvarum*, whilst thermotolerant strains of *S. cerevisiae* and *S. paradoxus* (such as strain RO88) were able to degrade between 40 and 48% of L-malic acid [83; 84; 85].

Furthermore, it was previously found that strain RO88 of *S. paradoxus* was able to degrade 38% of malic acid in Chardonnay must [82]. Increased expression of the malic enzyme gene in strain RO88 of *S. paradoxus* was also observed towards the end of fermentation when glucose was depleted [82]. It is noteworthy that VIN13 is characterised as having an optimum growth temperature of between 12-16°C and can thus be classified as a cryotolerant strain of *S. cerevisiae* [86].

S. cerevisiae can only use L-Malic acid in the presence of one or more fermentable carbon sources. Deletion of the gene encoding the malic acid enzyme (EC:1.1.1.38) revealed that it was non essential for the survival of the organism. There appears to be no active transport system for L-Malic acid in *S. cerevisiae*, however, mitochondrial L-Malic acid transporters do exist [82].

A study investigating the underlying mechanisms that control the ability of a yeast to degrade extracellular L-Malic acid during alcoholic fermentation was conducted using three different species of *Saccharomyces* and the results showed that all three had varying abilities to degrade L-Malic acid [82]. *S. bayanus* EC1118 and *S. cerevisiae* 71B were only able to degrade 8 and 17% of L-Malic acid respectively, however, *S. paradoxus* RO88 was able to degrade 28-38% of L-Malic acid [82]. Concomitant gene expression analysis revealed that increased expression of the malic acid enzyme led to increased degradation of L-Malic acid [82]. Different promoter sequences also exist between *S. paradoxus* RO88, *S.bayanus* EC1118 and *S. cerevisiae* 71B. Hence, it was proposed that different transcriptional regulatory mechanisms in these strains may explain the ability of *S. paradoxus* RO88 to degrade L-Malic acid to a greater extent [79].

Figure 3.18 shows two ways by which *S. cerevisiae* and other yeast are able to regenerate NADH. Under these experimental conditions, the glycerol formation pathway is likely active and up regulated for VIN13. Given the enhanced malic acid degradation ability of *S. paradoxus* RO88 [82], the pathway involving malic acid is likely active in RO88 under the experimental conditions defined in Section 3.5.2.1.

Although MAE1 was not detected in the protein dataset at hand, many protein constituents playing a role in the TCA cycle were identified. Six were found to be relatively more abundant in VIN13 when compared to RO88. These proteins are illustrated in Figure 3.19 B and include GLU1, ACO2, ACN17, IDH2, PDA1 and IDP1. Table 3.5 summarises a selection of attributes ascribed to these six proteins. GLU1, also known as ACO1, is the aconitase enzyme and ACO2 is a putative mitochondrial aconitase isozyme with high sequence similarity to ACO1. Both the identification and quantification of GLU1 and ACO2 are based on the same set of peptide spectra. IDP1 is a mitochondrial NADP-specific isocitrate dehydrogenase, whilst IDH2 is a subunit of mitochondrial NAD(+)-dependent isocitrate dehydrogenase. PDA1 is the E1 alpha subunit of the pyruvate dehydrogenase (PDH) complex and ACN17 is an iron-sulfur protein subunit of succinate dehydrogenase. Thus, all six of these proteins have direct enzymatic function within the mitochondria or form part of larger complexes that are involved in the TCA cycle.

The data possibly suggest that under the same stipulated growth conditions, the two yeast are making use of two different pathways to maintain the redox balance for core biosynthetic reactions in lieu of the predominant redox-neutral ethanol production. The suppositions being that *S. paradoxus* RO88 uses the auxillary pathway for the regenaration of NADH involving malic acid

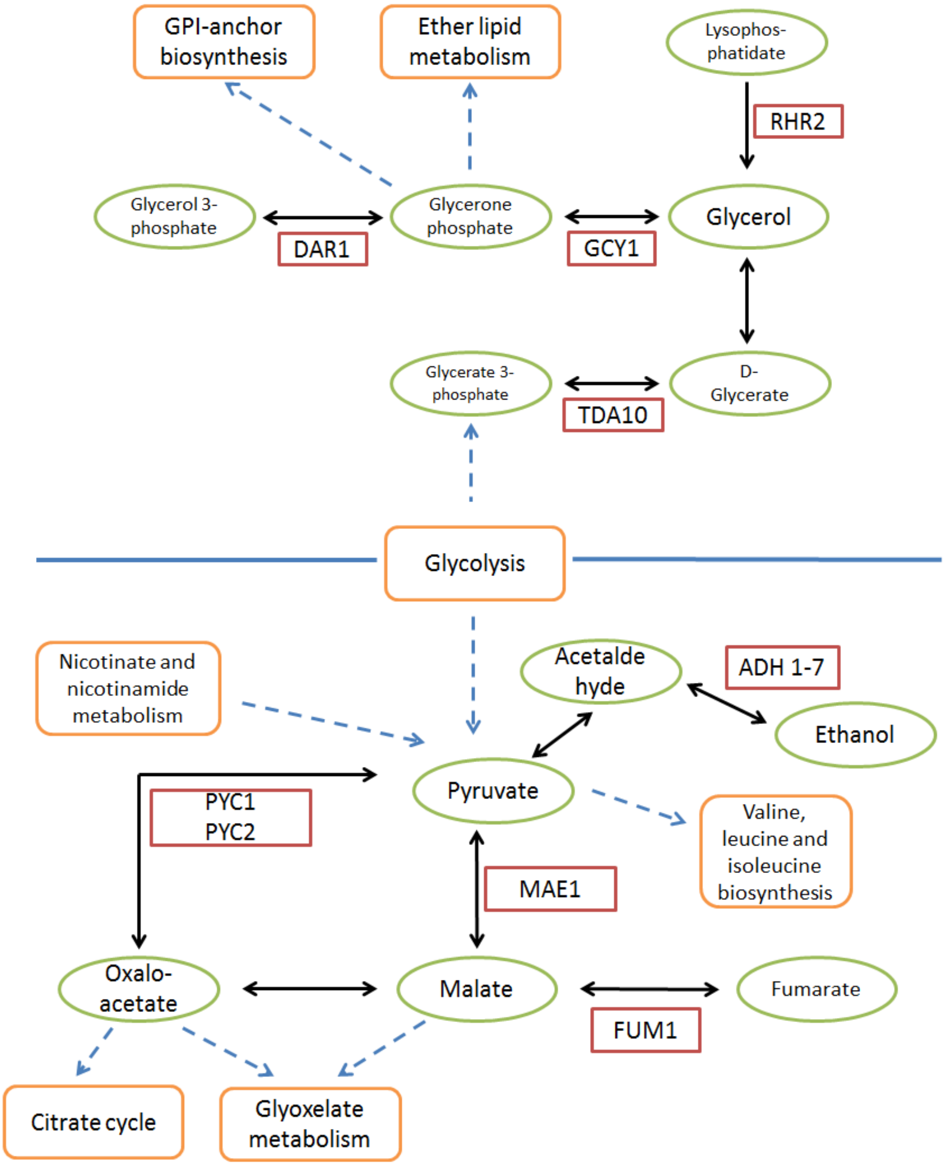


Figure 3.18: **A simplified metabolic pathway diagram of the mechanisms for reoxidation of assimilatory NADH in *S. cerevisiae*.** The top half of the diagram shows the reoxidation of NADH via glycerol formation. The bottom half of the diagram shows MEF. The diagram contains some of the key enzymes, compounds and pathways involved. The rectangles with rounded edges are pathways, circular shapes are compounds and rectangles are enzymes. Dashed lines indicate an indirect pathway link and solid lines are direct biochemical reactions involving the enzymes.

---

and the malo-ethanolic pathway [81] and that *S. cerevisiae* VIN13 is maintaining redox balance via the formation of glycerol [80]. Further support for this

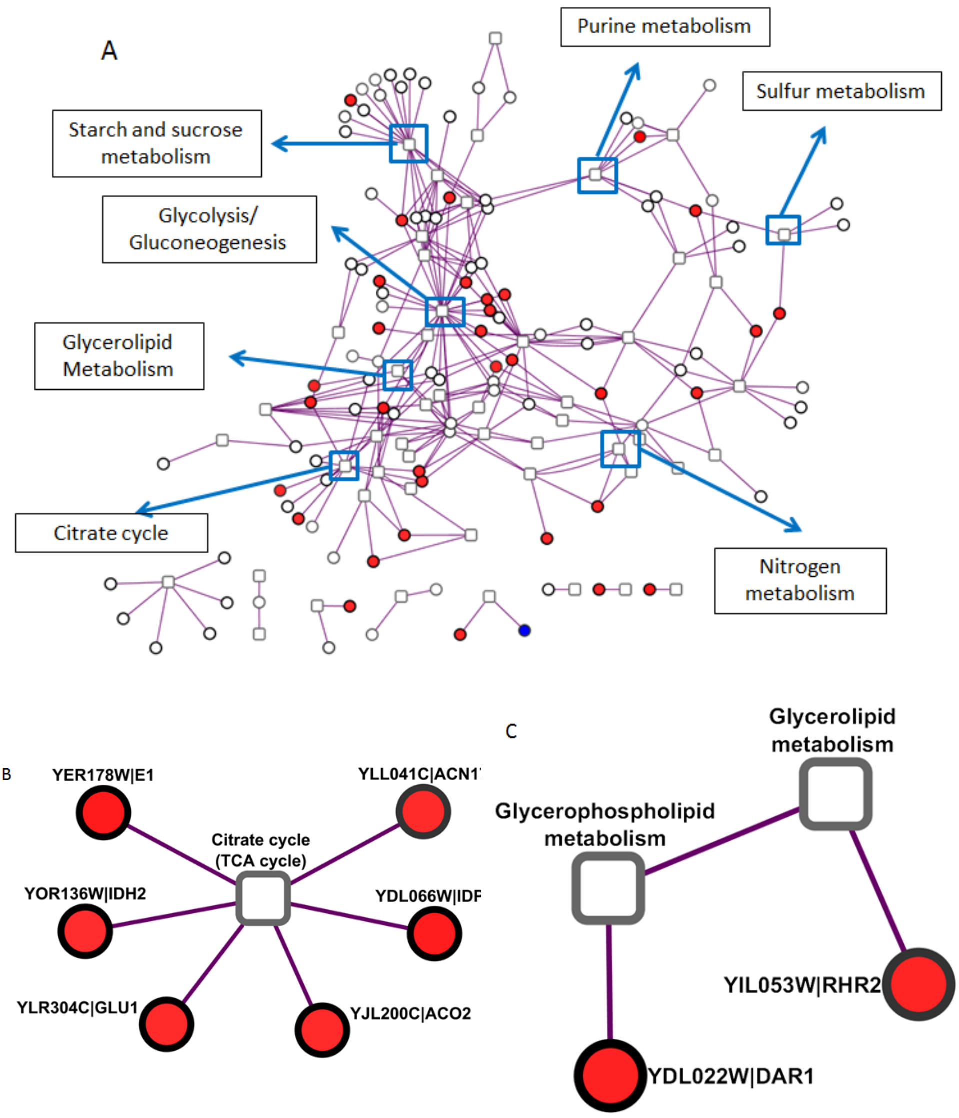


Figure 3.19: **Protein metabolic network with TCA and glycerol metabolism subnetworks**. (A) is all of the identified proteins within a metabolic context. A selection of pathways are pointed out for reference. (B) is a subnetwork of (A) and shows the six TCA cycle proteins that had significant fold changes. (C) is also a subnetwork of (A) and shows two proteins that had significant fold changes and are involved in glycerol and glycerophospholipid metabolism.

hypothesis is evidenced by higher abundance levels of the glycerol producing enzymes, DAR1 and RHR2 in VIN13 as shown in Figure 3.19C. Additional attributes for DAR1 and RHR2 are shown in Table 3.5. RHR2, also known as GGP1, is a constitutively expressed glycerol-1-phosphatase and is the enzymatic step prior to the formation of glycerol as shown in Figure 3.18. RHR2 catalyses the formation of glycerol from lysophosphatidate. This enzyme is unidirectional and thus an increase in it's abundance leads to increased glycerol production. DAR1, also known as GPD1, is a NAD-dependent glycerol-3-phosphate dehydrogenase and catalysis the formation of glycerone phosphate from glycerol 3-phosphate as shown in Figure 3.18. GCY1 then uses glycerol 3-phosphate as a substrate for the production of glycerol.

Table 3.5: **A subselection of attributes belonging to the proteins shown in Figure 3.19.** In the protein probability column the VIN13 probability is given first followed by the RO88 probability. If the probability of identification for the proteins was the same in both species, only one number is given. The number of quantitative peptides column indicates the number of peptides that were available to derive the fold change.

Protein-ID	Probability of identification	Fold change	Number of quantitative peptides
IDP1	1	8.74	1
IDH2	0.9656 ; 0.9867	3.4	1
YER178W	0.9669 ; 0.9441	9.57	1
ACN17	0.7402 ; 1	3.79	1
ACO2	0.9928 ; 1	4.15	1
GLU1	0.9928 ; 1	4.15	1
RHR2	0.5276 ; 0.7162	5.96	1
DAR1	0.9916 ; 0.9956	7.0	1

3.3.8 YHR138C - a Protein of Unknown Function

YHR138C is described as a protein of unknown function but has three GO terms, namely "cellular component", "endopeptidase inhibitor activity" and "vacuole fusion, non-autophagic" ascribed to it as shown in Figure 3.20 A.

Table 3.6: **A subselection of attributes for YHR138C and its interacting proteins.** The column descriptions are the same as in Table 3.5

.

Protein-ID	Probability of identification	Fold change	Number of quantitative peptides
YHR138C	0.5405 ; 0.9886	1.5	1
AIM3	0.6786 (RO88 only)	na	na
ILV3	1	2.9	2
LYS11	0.989 (VIN13 only)	na	na

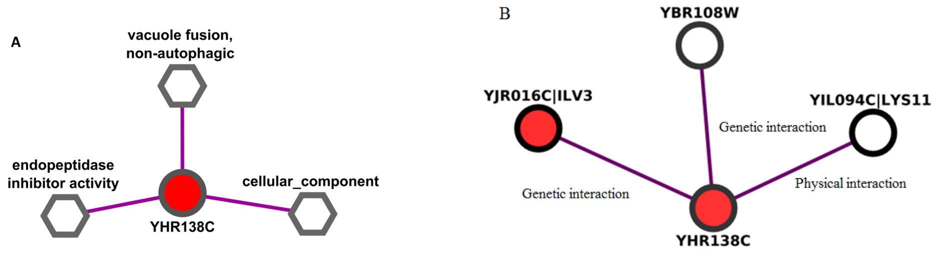


Figure 3.20: **YHR138C in GO and interactome contexts respectively.** (A) is the GO context and (B) is the interactome context. The visual mapping style is the same as in Figure 3.3.

3.3.8.1 YHR138C Interactions

Relatively little is known about YHR138C, however, it is similar to PBI2, a protein required for efficient vacuole inheritance [87]. It has been demonstrated that when both *YHR138C* and *PBI2* are knocked out in *S. cerevisiae*, highly fragmented vacuoles indicative of faulty vacuole fusion are observed [88]. However, in the respective individual gene mutant strains, the vacuole formation was unaffected [88].

According to the BioGRID database, YHR138C has three protein-protein interactions with other proteins in the dataset and these are illustrated in Figure 3.20 B. Two of these interactions were detected via genetic interference, namely YBR108W (also known as AIM3) [89] and ILV3 [90]. AIM3 is a protein that inhibits barbed-end actin filament elongation [91] and ILV3 catalyzes the third step in the common pathway leading to the biosynthesis of branched-chain amino acids [92].

YHR138C also has another genetic interaction [93] with ARP1 (not identified in this data). ARP1 is an actin-related protein which forms part of the dynactin complex that is required for nuclear migration and spindle orientation [94]. In addition to the genetic interactions, YHR138C also has a known physical association with LYS11 [95]. LYS11 is a NAD-linked homo-isocitrate dehydrogenase located in the mitochondria and is responsible for catalysing the fourth step of lysine biosynthesis [96].

3.3.8.2 The Role of Vacuoles in Yeast

Whilst actively growing, the cells of *S. cerevisiae* and other yeast have several prominent vacuoles that are functionally similar to plant vacuoles and mammalian lysosomes. Historically, vacuoles have been viewed only as "end-points" or terminal compartments in the biosynthetic and endocytic pathways [97]. Although acting as a terminal compartment is a vital function, this relatively simplistic view describes the vacuole to be no more than a compartment where unwanted materials and obsolete components from either the cytoplasm or extracellular space are sent to be degraded and recycled. A recent review suggest that vacuoles are more than just "end-points" but also act as "crossroads" that are able to dynamically respond to changes in the extracellular environment [97].

Vacuoles in yeast, like many other subcellular organelles, are not synthesised anew during cell division. Instead they are inherited from the mother cell [98]. During the early stages of the S-phase in *S. cerevisiae*, a tubular-vesicular "segregation" structure is projected by the vacuole into the newly formed bud. It is via this "segregation" structure that the daughter cell can receive maternal vacuolar vesicles, which can then fuse to establish the daughter vacuoles [99; 100; 101].

The redistribution of vacuoles from mother to daughter cell is an actin-dependent process that requires the temporal and spatial control of physical vacuole movement as well as adjustment of vacuolar function during this redistribution process [102]. Thus, the literature provides logical possible explanations for the interactions between YHR138C, PBI2 and ARP1 respectively. However, the link between YHR138C, ILV3 and LYS11 is less clear.

3.3.8.3 YHR138C Interaction With ILV3 and LYS11

A possible explanation for the interactions of YHR138C with ILV3 and LYS11 may be found by looking at vesicle formation and its role during endocytotic processes and how it relates to other factors such as vacuoles in the cell, actin, changes in the cell wall and changes in nutritional availability.

Actin cortical patches are known endocytic sites and endocytic proteins have also been found to colocalize with these actin patches [103; 104; 105; 106]. Furthermore, it is known that cell surface proteins may enter the cells by endocytosis [107] of which the first step involves transiting to early endosomes. This is followed by intersection of the endosomes with the carboxypeptidase Y pathway as multivesicular bodies and then transport to the vacuole. One such cell surface protein known to be edocytosed is GAP1, a general amino acid permease. When cytosolic amino acids are limited, GAP1 is targeted for degradation via endocytosis [108].

Nutrients become limiting toward the end of fermentation and this is the point at which proteins for this dataset were extracted. Thus, under these nutrient-limiting conditions it is possible that GAP1 will be targeted for ubiquitination via endocyctosis and amino acid biosynthetic capacity will increase. From the data and literature presented it is known that the protein of unknown function, YHR138C, is involved with endocytosis and vacuoles in possibly more than one way. Additionally, vacuoles are cellular bodies known to be dynamic and responsive to nutrient changes [97]. Furthermore, YHR138C interacts with two enzymes at the core of amino acid metabolism which links it to nitrogen metabolism. Thus, from this summation one can formulate the hypothesis that YHR138C is somehow involved or connected to a nitrogen-source driven regulatory circuit. Figure 3.21 illustrates the hypothesis diagrammatically.

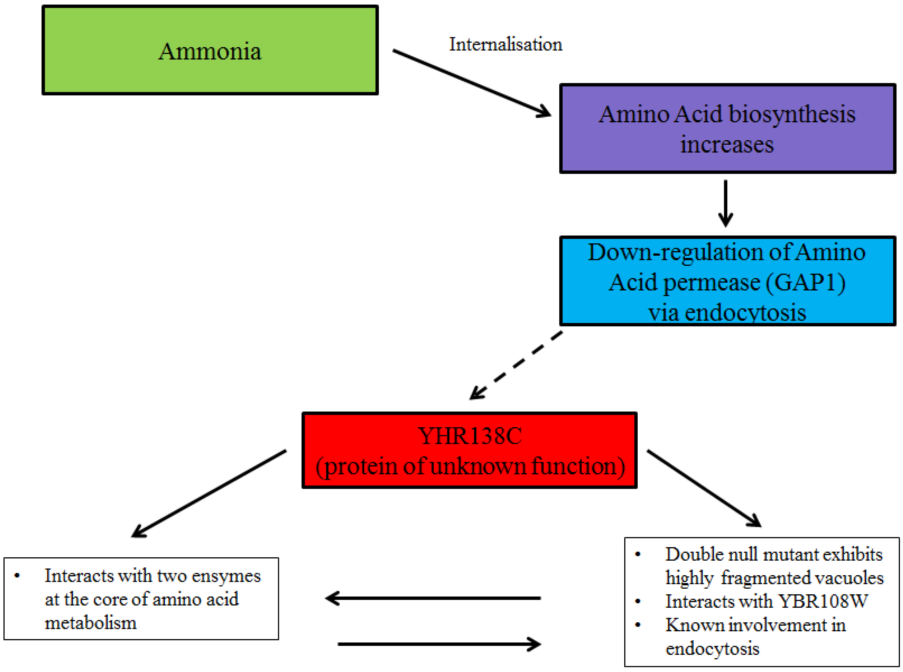

Ammonia
Internalisation
Amino Acid biosynthesis increases
Down-regulation of Amino Acid permease (GAP1) via endocytosis
YHR138C (protein of unknown function)
Interacts with two ensymes at the core of amino acid metabolism
Double null mutant exhibits highly fragmented vacuoles
Interacts with YBR108W
Known involvement in endocytosis

Figure 3.21: **Possible hypothesis for the function of YHR138C.**

3.4 Summary and Conclusion

The examples presented here illustrate only a few of many possible cases of how the respective network contextualisations of the data can be combined to investigate existing hypotheses, or formulate new hypotheses from the dataset. Also, the workflow described allows one much more control over data quality at the point of interpretation without being overwhelmed by the shear volume of information that is available. Furthermore, the multiple filtering criteria available and visual nature of the network representations facilitate easy pattern recognition as well as reporting of the results. Thus, the method is successful in bringing together captured LC-MS/MS data, combining it with database and literature resources in a manner that is statistically defensible, and allows for the maximal extraction of biologically relevant knowledge from the experiment.

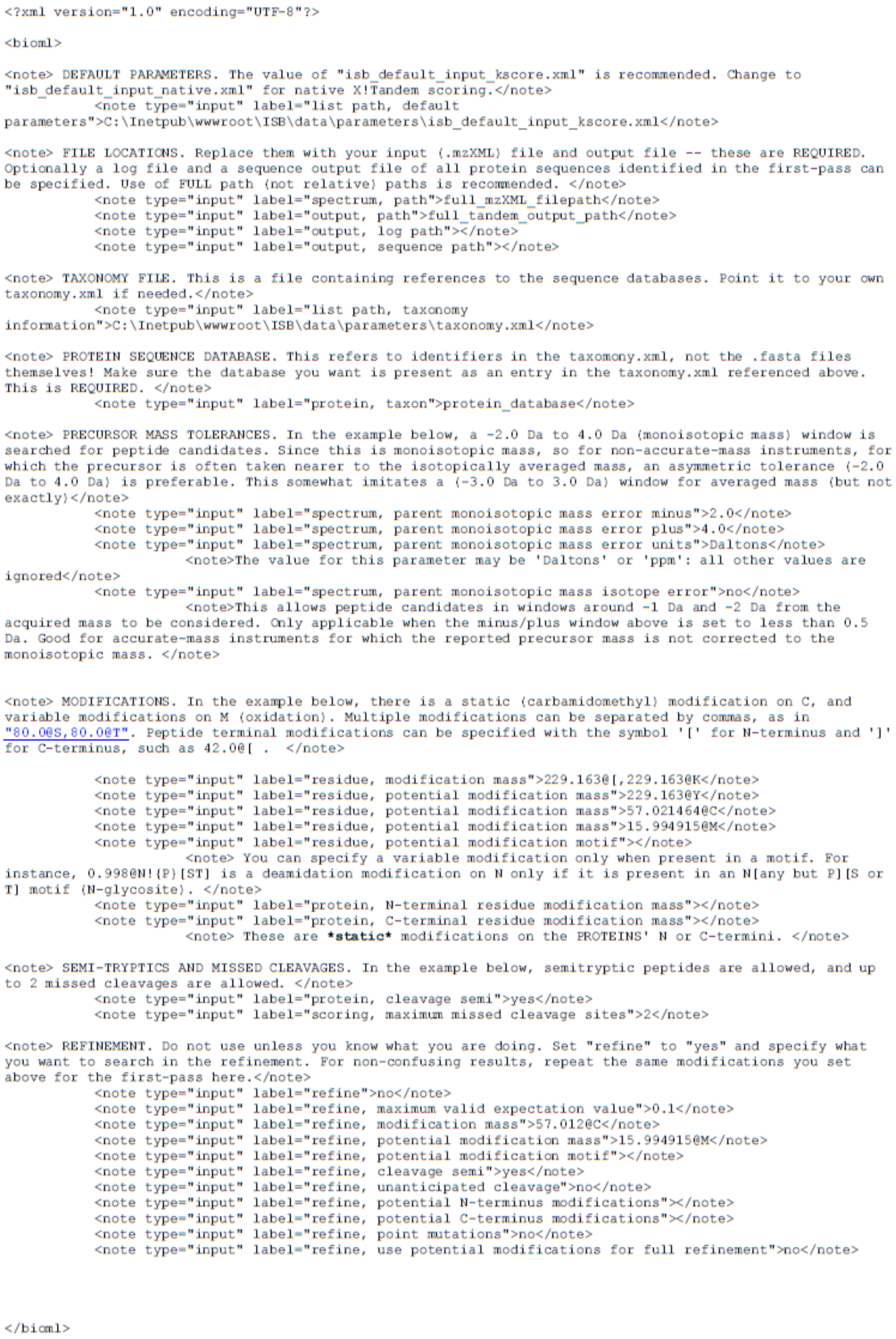
3.5 Supplementary Material

3.5.1 Supplementary Figures and Tables

Table S1 shows the decrease in the total number of edges as the threshold for edge creation between protein family module nodes is increased. This network is similar to the network described in Section 3.3.5, the difference is in the node type.

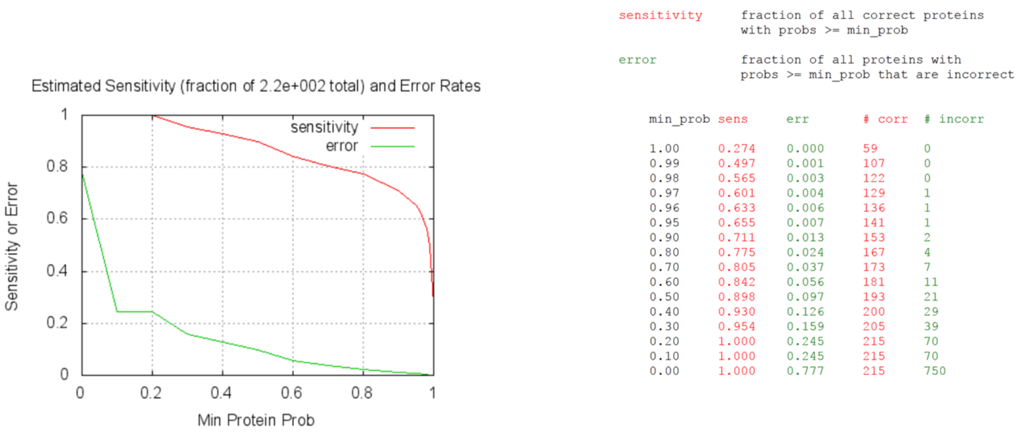
Table S1: **The number of edges that exist at different Jaccard index thresholds as determined for the protein family modules GO overlap network.**

Jaccard index threshold	number of edges
0.0	31526
0.1	7572
0.2	2035
0.3	1075
0.4	676
0.5	407
0.6	250
0.7	120
0.8	50
0.9	18



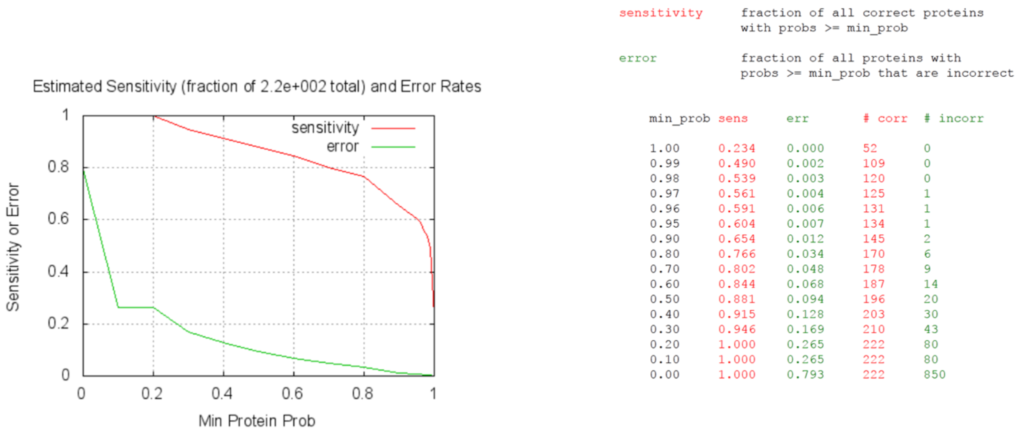
```
<?xml version="1.0" encoding="UTF-8"?>

<bioml>

<note> DEFAULT PARAMETERS. The value of "isb_default_input_kscore.xml" is recommended. Change to
"isb_default_input_native.xml" for native X!Tandem scoring.</note>
          <note type="input" label="list path, default
parameters">C:\Inetpub\wwwroot\ISB\data\parameters\isb_default_input_kscore.xml</note>

<note> FILE LOCATIONS. Replace them with your input (.mzXML) file and output file -- these are REQUIRED.
Optionally a log file and a sequence output file of all protein sequences identified in the first-pass can
be specified. Use of FULL path (not relative) paths is recommended. </note>
          <note type="input" label="spectrum, path">full_mzXML_filepath</note>
          <note type="input" label="output, path">full_tandem_output_path</note>
          <note type="input" label="output, log path"></note>
          <note type="input" label="output, sequence path"></note>

<note> TAXONOMY FILE. This is a file containing references to the sequence databases. Point it to your own
taxonomy.xml if needed.</note>
          <note type="input" label="list path, taxonomy
information">C:\Inetpub\wwwroot\ISB\data\parameters\taxonomy.xml</note>

<note> PROTEIN SEQUENCE DATABASE. This refers to identifiers in the taxomony.xml, not the .fasta files
themselves! Make sure the database you want is present as an entry in the taxonomy.xml referenced above.
This is REQUIRED. </note>
          <note type="input" label="protein, taxon">protein_database</note>

<note> PRECURSOR MASS TOLERANCES. In the example below, a -2.0 Da to 4.0 Da (monoisotopic mass) window is
searched for peptide candidates. Since this is monoisotopic mass, so for non-accurate-mass instruments, for
which the precursor is often taken nearer to the isotopically averaged mass, an asymmetric tolerance (-2.0
Da to 4.0 Da) is preferable. This somewhat imitates a (-3.0 Da to 3.0 Da) window for averaged mass (but not
exactly)</note>
          <note type="input" label="spectrum, parent monoisotopic mass error minus">2.0</note>
          <note type="input" label="spectrum, parent monoisotopic mass error plus">4.0</note>
          <note type="input" label="spectrum, parent monoisotopic mass error units">Daltons</note>
                    <note>The value for this parameter may be 'Daltons' or 'ppm': all other values are
ignored</note>
          <note type="input" label="spectrum, parent monoisotopic mass isotope error">no</note>
                    <note>This allows peptide candidates in windows around -1 Da and -2 Da from the
acquired mass to be considered. Only applicable when the minus/plus window above is set to less than 0.5
Da. Good for accurate-mass instruments for which the reported precursor mass is not corrected to the
monoisotopic mass. </note>


<note> MODIFICATIONS. In the example below, there is a static (carbamidomethyl) modification on C, and
variable modifications on M (oxidation). Multiple modifications can be separated by commas, as in
"80.0@S,80.0@T". Peptide terminal modifications can be specified with the symbol '[' for N-terminus and ']'
for C-terminus, such as 42.0@[ .  </note>

          <note type="input" label="residue, modification mass">229.163@[,229.163@K</note>
          <note type="input" label="residue, potential modification mass">229.163@Y</note>
          <note type="input" label="residue, potential modification mass">57.021464@C</note>
          <note type="input" label="residue, potential modification mass">15.994915@M</note>
          <note type="input" label="residue, potential modification motif"></note>
                    <note> You can specify a variable modification only when present in a motif. For
instance, 0.998@N!{P}[ST] is a deamidation modification on N only if it is present in an N[any but P][S or
T] motif (N-glycosite). </note>
          <note type="input" label="protein, N-terminal residue modification mass"></note>
          <note type="input" label="protein, C-terminal residue modification mass"></note>
                    <note> These are *static* modifications on the PROTEINS' N or C-termini. </note>

<note> SEMI-TRYPTICS AND MISSED CLEAVAGES. In the example below, semitryptic peptides are allowed, and up
to 2 missed cleavages are allowed. </note>
          <note type="input" label="protein, cleavage semi">yes</note>
          <note type="input" label="scoring, maximum missed cleavage sites">2</note>

<note> REFINEMENT. Do not use unless you know what you are doing. Set "refine" to "yes" and specify what
you want to search in the refinement. For non-confusing results, repeat the same modifications you set
above for the first-pass here.</note>
          <note type="input" label="refine">no</note>
          <note type="input" label="refine, maximum valid expectation value">0.1</note>
          <note type="input" label="refine, modification mass">57.012@C</note>
          <note type="input" label="refine, potential modification mass">15.994915@M</note>
          <note type="input" label="refine, potential modification motif"></note>
          <note type="input" label="refine, cleavage semi">yes</note>
          <note type="input" label="refine, unanticipated cleavage">no</note>
          <note type="input" label="refine, potential N-terminus modifications"></note>
          <note type="input" label="refine, potential C-terminus modifications"></note>
          <note type="input" label="refine, point mutations">no</note>
          <note type="input" label="refine, use potential modifications for full refinement">no</note>



</bioml>
```

Figure S1: **Parameters used for X!Tandem.**



```
sensitivity     fraction of all correct proteins
                with probs >= min_prob

error           fraction of all proteins with
                probs >= min_prob that are incorrect


    min_prob sens      err        # corr  # incorr

    1.00     0.274     0.000      59      0
    0.99     0.497     0.001      107     0
    0.98     0.565     0.003      122     0
    0.97     0.601     0.004      129     1
    0.96     0.633     0.006      136     1
    0.95     0.655     0.007      141     1
    0.90     0.711     0.013      153     2
    0.80     0.775     0.024      167     4
    0.70     0.805     0.037      173     7
    0.60     0.842     0.056      181     11
    0.50     0.898     0.097      193     21
    0.40     0.930     0.126      200     29
    0.30     0.954     0.159      205     39
    0.20     1.000     0.245      215     70
    0.10     1.000     0.245      215     70
    0.00     1.000     0.777      215     750
```

Figure S2: **The output from Protein Prophet for VIN13 proteins.** The x-axis of the graph indicates the minimum probability and the y-axis represents both sensitivity and error. The red line indicates sensitivity, whilst the green line indicates error. The corresponding table provides the values from which the graph is plotted.



```
sensitivity     fraction of all correct proteins
                with probs >= min_prob

error           fraction of all proteins with
                probs >= min_prob that are incorrect


    min_prob sens      err        # corr  # incorr

    1.00     0.234     0.000      52      0
    0.99     0.490     0.002      109     0
    0.98     0.539     0.003      120     0
    0.97     0.561     0.004      125     1
    0.96     0.591     0.006      131     1
    0.95     0.604     0.007      134     1
    0.90     0.654     0.012      145     2
    0.80     0.766     0.034      170     6
    0.70     0.802     0.048      178     9
    0.60     0.844     0.068      187     14
    0.50     0.881     0.094      196     20
    0.40     0.915     0.128      203     30
    0.30     0.946     0.169      210     43
    0.20     1.000     0.265      222     80
    0.10     1.000     0.265      222     80
    0.00     1.000     0.793      222     850
```

Figure S3: **The output from Protein Prophet for RO88 proteins.** The x-axis of the graph indicates the minimum probability and the y-axis represents both sensitivity and error. The red line indicates sensitivity, whilst the green line indicates error. The corresponding table provides the values from which the graph is plotted.

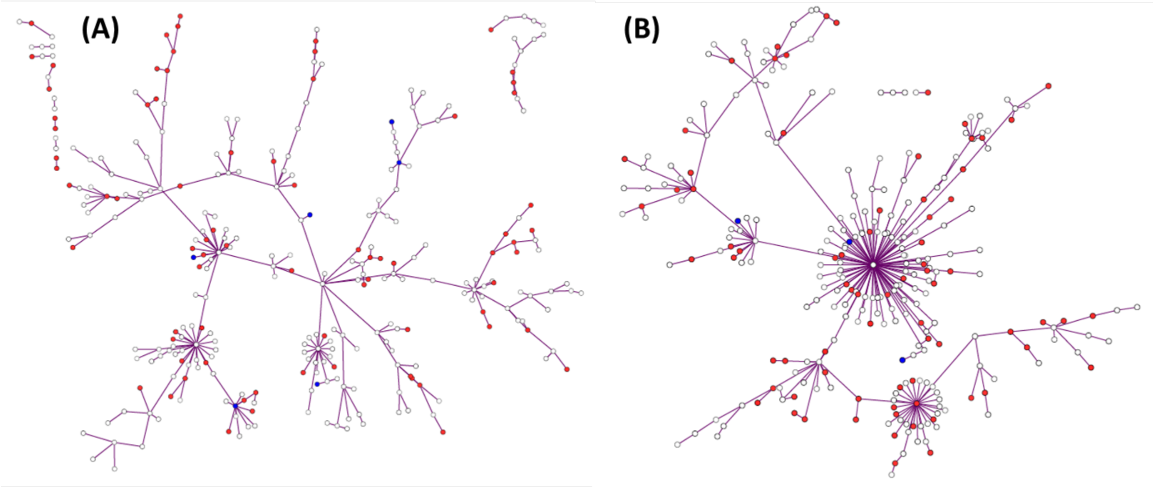


Figure S4: **The minimum spanning trees of the gene-based and physical-based protein-protein interaction networks respectively.** (A) is the minimum spanning tree of the gene-based protein-protein interaction network. (B) is the minimum spanning tree of the physical-based protein-protein interaction network. The visual mapping scheme is the same as in Figure 3.6.

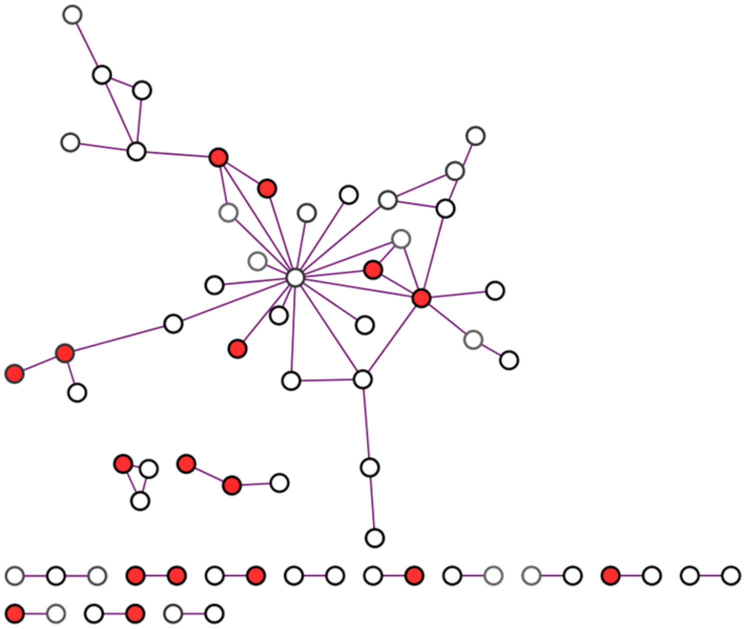


Figure S5: **Physical-based protein-protein interaction network filtered for one or more line of evidence.** The visual mapping scheme is the same as in Figure 3.6.

3.5.2 Supplementary Materials and Methods

3.5.2.1 Fermentation media and conditions

This study made use of *S. cerevisiae* strain VIN13 and *S. paradoxus* strain RO88 [82]. Both yeast were used to ferment to dryness chemically defined MS300 media [109] containing 200 g/l glucose and fructose. Fermentations were carried out in triplicate with yeast pre-cultures grown in YPD broth (BD Becton, Dickinson and Company, catelog number 242820). Synthetic must were inoculated to obtain a final concentration of 10^6 cells/ml. All fermentations were carried out in Erlenmeyer flasks using a 100 ml working volume. Vessels were closed with fermentation caps and no agitation was used. The fermentations were conducted in a room maintained at 25°C. Residual glucose and fructose concentrations were less than 5 g/l as measured using a D-glucose/fructose kit (Amersham).

3.5.2.2 Protein Purification for TMT Analysis

Protein purification was carried out following the protocol described [110]. Fermented MS300 was centrifuged at 5000 rpm for 5 minutes to remove cells and concentrated with Millipore Membrane Centrifugal Filter devices with a molecular weight cut-off of 10kDa. An ice-cold ethanol solution containing 15% (w/v) trichloro-acetic acid (TCA) was used to dilute concentrates at 4°C and the pellet was washed with ice-cold ethanol and centrifuged. The vacuum dried protein pellet was solubilized in 100 μl of 6 M urea, 2 M thiourea and 10 mM DTT. The proteins were then alkylated with 50 mM iodoacetamide for 40 minutes at room temperature in the dark.

3.5.2.3 TMTs for Relative Quantification

Total protein concentration was determined using Pierce BCA Protein Assay (Thermo Scientific). After pooling tubes in each group (VIN13 and RO88, respectively) there was 85 μg in each sample. Each sample was diluted with 0.5 M TEAB (triethyl ammonium bicarbonate) and then diluted with milli-Q water to a 4-fold dilution to a pH $>$8. To each sample, SDS (sodium dodecyl sulphate) solution to a final concentration of 0.1% and trypsin (dissolved in milli-Q water) with a ratio of 1:10 was added. Digestion was done overnight at 37°C .

3.5.2.4 Labelling with TMT reagents

TMT reagents 126, 127 and 128 for RO88 and 129, 130 and 131 for VIN13 were dissolved in ethanol and added to the respective sample according to the manufacturer's protocol. After labelling, the samples were combined and concentrated. TMT-labelled peptides were separated with Strong Cation Exchange Chromatography (SCX). The concentrated peptides were acidified by

10% formic acid and diluted with SCX solvent A (25 mM ammonium formate, pH 2.8, 20% acetonitrile (ACN)) and injected onto a PolySULFOETHYL A ™SCX column (2.1 mm i.d. x 10 cm length, 5 μm particle size, 300 Å pore size). SCX chromatography and fractionation was carried out on an ÄKTA purifier system (GE healthcare) at 0.25 mL/min flow rate using the following gradient: 0% B (500 mM ammonium formate, pH 2.8, 20% ACN) for 5 min; 0-40% B for 20 min; 40-100% B for 10 min and 100% B held for 10 min. UV absorbance at 254 and 280 nm was monitored while fractions were collected at 0.5 mL intervals. The peptide containing fractions were desalted on Pep-Clean™C18 spin columns according to manufacturer's instructions (Thermo Fisher Scientific) and dried down in a Speed Vac.

3.5.2.5 LC-MS/MS Analysis on LTQ-Orbitrap-Velos

The desalted and dried fractions were reconstituted into 0.1% formic acid and analysed on a LTQ-Orbitrap-Velos (Thermo Fisher Scientific) interfaced with an in-house constructed nano-LC column. Two-micro liter sample injections were made with an Easy-nLC autosampler (Thermo Fisher Scientific, Inc., Waltham, MA, USA), running at 200 nl/min. The peptides were trapped on a pre-column (45 x 0.075 mm i.d.) and separated on a reversed phase column, 200 x 0.075 mm, packed in-house with 3 μm Reprosil-Pur C18-AQ particles. The gradient was as followed; 0-90 min 5-37% acetonitrile (ACN), 0.1% formic acid, 90-93 min 37-90% ACN, 0.1% formic acid and the last 5 min at 90% ACN, 0.1% formic acid.

3.5.2.6 LTQ-Orbitrap Velos Settings

LTQ-Orbitrap Velos settings were as follows: spray voltage 1.4 kV; 1 microscan for MS1 scans at 60 000 resolutions (m/z 400) and full MS mass range m/z 400-2000. The LTQ-Orbitrap Velos was operated in a data-dependent mode with one MS1 FTMS scan precursor ions followed by CID (collision induced dissociation) and HCD (high energy collision dissociation), MS2 scans of the five most abundant protonated ions in each FTMS scan. The settings for the MS2 were as follows: 1 microscans for HCD-MS2 at 7500 resolution (at m/z 400); mass range m/z 100-2000 with a collision energy of 50%; 1 microscans for CID-MS2 with a collision energy of 30%.

Bibliography

[1] Aebersold, R. and Goodlett, D.R.: Mass spectrometry in proteomics. *Proteomics*, vol. 3, p. 5, 2001.

[2] Washburn, M.P., Wolters, D. and Yates, J.R.: Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, vol. 19, no. 3, pp. 242–247, 2001.

[3] Kislinger, T., Gramolini, A.O., MacLennan, D.H. and Emili, A.: Multidimensional Protein Identification Technology (MudPIT): Technical Overview of a Profiling Method Optimized for the Comprehensive Proteomic Investigation of Normal and Diseased Heart Tissue. *Journal of the American Society for Mass Spectrometry*, vol. 16, no. 8, pp. 1207–1220, 2005. ISSN 1044-0305.

[4] Michalski, A., Damoc, E., Hauschild, J.-P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M. and Horning, S.: Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Molecular & Cellular Proteomics*, vol. 10, no. 9, pp. M111–011015, 2011.

[5] Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J. and Aebersold, R.: The quantitative proteome of a human cell line. *Molecular Systems Biology*, vol. 7, no. 1, 2011.

[6] Hunt, D.F., Yates, J.R., Shabanowitz, J., Winston, S. and Hauer, C.R.: Protein sequencing by tandem mass spectrometry. *Proceedings of the National Academy of Sciences*, vol. 83, no. 17, pp. 6233–6237, 1986.

[7] Biemann, K.: Sequencing of peptides by tandem mass spectrometry and high-energy collision-induced dissociation. *Methods in Enzymology*, vol. 193, pp. 455–479, 1990.

[8] Domon, B. and Aebersold, R.: Challenges and opportunities in proteomics data analysis. *Molecular & Cellular Proteomics*, vol. 5, no. 10, pp. 1921–1926, 2006.

[9] Deutsch, E.W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B. and Prazen, B.: A guided tour of the Trans Proteomic Pipeline. *Proteomics*, vol. 10, no. 6, pp. 1150–1159, 2010.

[10] Kohlbacher, O., Reinert, K., Gröpl, C., Lange, E., Pfeifer, N., Schulz-Trieglaff, O. and Sturm, M.: TOPP - The OpenMS proteomics pipeline. *Bioinformatics*, vol. 23, no. 2, pp. e191–e197, 2007.

[11] Cox, J. and Mann, M.: MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, vol. 26, no. 12, pp. 1367–1372, 2008.

[12] Hawkridge, A.M.: Chapter 1 Practical Considerations and Current Limitations in Quantitative Mass Spectrometry-based Proteomics. In: *Quantitative Proteomics*, pp. 1–25. The Royal Society of Chemistry, 2014. ISBN 978-1-84973-808-8.

[13] Hartwell, L.H., Hopfield, J.J., Leibler, S. and Murray, A.W.: From molecular to modular cell biology. *Nature*, vol. 402, pp. C47–C52, 1999.

[14] Ndlovu, T.: *Mannoprotein production and wine haze reduction by wine yeast strains*. Ph.D. thesis, Stellenbosch: Stellenbosch University, 2012.

[15] Li, L., Stoeckert, C.J. and Roos, D.S.: OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, vol. 13, no. 9, pp. 2178–89, September 2003. ISSN 1088-9051.

[16] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, vol. 25, no. 1, pp. 25–9, May 2000. ISSN 1061-4036.

[17] Zheng, Q. and Wang, X.-J.: GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Research*, vol. 36, no. Web Server issue, pp. W358–63, July 2008. ISSN 1362-4962.

[18] Kanehisa, M. and Goto, S.: KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, January 2000. ISSN 0305-1048.

[19] Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A. and Tyers, M.: BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, vol. 34, no. suppl 1, pp. D535–D539, 2006.

[20] Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, vol. 13, no. 11, pp. 2498–504, November 2003. ISSN 1088-9051.

[21] Craig, R. and Beavis, R.C.: TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, vol. 20, no. 9, pp. 1466–1467, 2004.

[22] Kellis, M., Patterson, N., Endrizzi, M., Birren, B. and Lander, E.S.: Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, vol. 423, no. 6937, pp. 241–254, 2003.

[23] Rice, P., Longden, I. and Bleasby, A.: EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics : TIG*, vol. 16, no. 6, pp. 276–7, June 2000. ISSN 0168-9525.

[24] Keller, A. and Nesvizhskii, A.: Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry*, vol. 74, no. 20, pp. 5383–5392, 2002.

[25] Nesvizhskii, A. and Keller, A.: A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry*, vol. 75, no. 17, pp. 4646–4658, 2003.

[26] Benjamini, Y. and Hochberg, Y.: Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[27] R Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.
Available at: http://www.r-project.org

[28] Choo, K.H., Tan, T.W. and Ranganathan, S.: SPdb-a signal peptide database. *BMC Bioinformatics*, vol. 6, p. 249, January 2005. ISSN 1471-2105.

[29] Lum, G. and Min, X.J.: FunSecKB: the Fungal Secretome KnowledgeBase. *Database : The Journal of Biological Databases and Curation*, vol. 2011, p. bar001, January 2011. ISSN 1758-0463.

[30] Petersen, T.N., Brunak, S.r., von Heijne, G. and Nielsen, H.: SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, vol. 8, no. 10, pp. 785–786, 2011.

[31] Jaccard, P.: Étude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, vol. 37, no. 142, pp. 547–579, 1901.

[32] Hamers, L., Hemeryck, Y., Herweyers, G., Janssen, M., Keters, H., Rousseau, R. and Vanhoutte, A.: Similarity measures in scientometric research: the jaccard index versus Salton's cosine formula. *Information Processing & Management*, vol. 25, no. 3, pp. 315–318, 1989.

[33] Kruskal, J.: On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. *Proceedings of the American Mathematical Society*, vol. 7, no. 1, pp. 48–50, 1956.

[34] van Donge, S.M.: *Graph Clustering by Flow Simulation*. Ph.D. thesis, University of Utrecht, 2000.

[35] Klis, F.M., Boorsma, A. and De Groot, P.W.J.: Cell wall construction in *Saccharomyces cerevisiae*. *Yeast (Chichester, England)*, vol. 23, no. 3, pp. 185–202, February 2006. ISSN 0749-503X.

[36] Cabib, E., Roh, D.H., Schmidt, M., Crotti, L.B. and Varma, A.: The yeast cell wall and septum as paradigms of cell growth and morphogenesis. *The Journal of Biological Chemistry*, vol. 276, no. 23, pp. 19679–82, June 2001. ISSN 0021-9258.

[37] Escot, S., Feuillat, M., Dulau, L. and Charpentier, C.: Release of polysaccharides by yeasts and the influence of released polysaccharides on colour stability and wine astringency. *Australian Journal of Grape and Wine Research*, vol. 7, no. 3, pp. 153–159, 2001.

[38] Caridi, A.: Enological functions of parietal yeast mannoproteins. *Antonie van Leeuwenhoek*, vol. 89, no. 3-4, pp. 417–22, 2006. ISSN 0003-6072.

[39] Guilloux-Benatier, M. and Chassagne, D.: Comparison of components released by fermented or active dried yeasts after aging on lees in a model wine. *Journal of Agricultural and Food Chemistry*, vol. 51, no. 3, pp. 746–751, 2003.

[40] Gerbaud, V., Gabas, N., Laguerie, C., Blouin, J., Vidal, S., Moutounet, M. and Pellerin, P.: Effect of wine polysaccharides on the nucleation of potassium hydrogen tartrate in model solutions. *Chemical Engineering Research & Design*, vol. 74, no. 7, pp. 782–790, 1996.

[41] Batista, L., Monteiro, S., Loureiro, V.B., Teixeira, A.R. and Ferreira, R.B.: The complexity of protein haze formation in wines. *Food Chemistry*, vol. 112, no. 1, pp. 169–177, January 2009. ISSN 03088146.

[42] Dvegowda, G., Raju, M. and Swamy, H.: Mycotoxins: novel solutions for their counteraction. *Feedstuffs (USA)*, 1998.

[43] Piva, A. and Galvano, F.: Nutritional approaches to reduce the impact of mycotoxins. *Biotechnology in the Feed Industry*, pp. 381–399, 1999.

[44] Baptista, A.S., Horii, J., Calori-Domingues, M.A., da Gloria, E.M., Salgado, J.M. and Vizioli, M.R.: The capacity of manno-oligosaccharides, thermolysed yeast and active yeast to attenuate aflatoxicosis. *World Journal of Microbiology and Biotechnology*, vol. 20, no. 5, pp. 475–481, 2004.

[45] Zimmerli, B. and Dick, R.: Ochratoxin A in table wine and grape-juice: Occurrence and risk assessment. *Food Additives & Contaminants*, vol. 13, no. 6, pp. 655–668, 1996.

[46] Chassagne, D., Charpentier, C., Guilloux-Benatier, M., Alexandre, H. and Feuillat, M.: Influence de l'autolyse des levures apres fermentation sur le d{é}veloppement de Brettanomyces-Dekkera dans le vin. *International Journal of Vine and Wine Sciences*, vol. 35, no. 3, pp. 157–164, 2001.

[47] Alexandre, H., Blanchet, S. and Charpentier, C.: Identification of a 49-kDa hydrophobic cell wall mannoprotein present in velum yeast which may be implicated in velum formation. *FEMS Microbiology Letters*, vol. 185, no. 2, pp. 147–150, 2000.

[48] Suzzi, G., Romano, P. and Zambonelli, C.: Flocculation of wine yeasts: frequency, differences, and stability of the character. *Canadian Journal of Microbiology*, vol. 30, no. 1, pp. 36–39, 1984.

[49] Klis, F.M., Mol, P., Hellingwerf, K. and Brul, S.: Dynamics of cell wall structure in *Saccharomyces cerevisiae*. *FEMS Microbiology Reviews*, vol. 26, no. 3, pp. 239–256, 2002.

[50] Dupin, I. and Stockdale, V.: *Saccharomyces cerevisiae* Mannoproteins That Protect Wine from Protein Haze: Evaluation of Extraction Methods and Immunolocalization. *Journal of Agricultural and Food Chemistry*, pp. 1086–1095, 2000.

[51] Nebreda, A., Villa, T., Villanueva, J. and del Rey, F.: Cloning of genes related to exo-β-glucanase production in Saccharomyces cerevislae: characterization of an exo-β-glucanase structural gene. *Gene*, vol. 41, pp. 245–259, 1986.

[52] Shaw, J., Mol, P. and Bowers, B.: The function of chitin synthases 2 and 3 in the *Saccharomyces cerevisiae* cell cycle. *The Journal of cell biology*, vol. 114, no. 1, 1991.

[53] Rodríguez-Peña, J. and Cid, V.: A novel family of cell wall-related proteins regulated differently during the yeast life cycle. *Molecular and Cellular Biology*, 2000.

[54] Brown, S.L., Stockdale, V.J., Pettolino, F., Pocock, K.F., Barros Lopes, M., Williams, P.J., Bacic, A., Fincher, G.B., Hø j, P.B. and Waters, E.J.: Reducing haziness in white wine by overexpression of *Saccharomyces cerevisiae* genes YOL155c and YDR055w. *Applied Microbiology and Biotechnology*, vol. 73, no. 6, pp. 1363–1376, October 2006. ISSN 0175-7598.

[55] Waters, E.J., Wallace, W., Tate, M.E. and Williams, P.J.: Isolation and partial characterization of a natural haze protective factor from wine. *Journal of Agricultural and Food Chemistry*, vol. 41, no. 5, pp. 724–730, 1993.

[56] Pardo, M., Monteoliva, L., Pla, J., Sánchez, M., Gil, C. and Nombela, C.: Two-dimensional analysis of proteins secreted by *Saccharomyces cerevisiae* regenerating protoplasts: a novel approach to study the cell wall. *Yeast (Chichester, England)*, vol. 15, no. 6, pp. 459–72, April 1999. ISSN 0749-503X.

[57] Schmidt, S.A., Tan, E.L., Brown, S., Nasution, U.J., Pettolino, F., Macintyre, O.J., de Barros Lopes, M., Waters, E.J. and Anderson, P.A.: Hpf2 glycan structure is critical for protection against protein haze formation in white wine. *Journal of Agricultural and Food Chemistry*, vol. 57, no. 8, pp. 3308–3315, 2009.

[58] Gentzsch, M. and Tanner, W.: The PMT gene family: protein O-glycosylation in *Saccharomyces cerevisiae* is vital. *The EMBO Journal*, vol. 15, no. 21, p. 5752, 1996.

[59] Hernandez, L.M., Alvarado, E., Ballou, C.E., Armero, J., Abad, A., Sentandreu, R. and Zueco, J.: Deletion of New Covalently Linked Cell Wall Glycoproteins Alters the Electrophoretic Mobility of Phosphorylated Wall Components of *Saccharomyces cerevisiae*. *Journal of Bacteriology*, vol. 181, no. 10, pp. 3076–3086, 1999.

[60] Ragni, E., Fontaine, T., Gissi, C., Latgè, J.P. and Popolo, L.: The Gas family of proteins of *Saccharomyces cerevisiae*: characterization and evolutionary analysis. *Yeast (Chichester, England)*, vol. 24, no. 4, pp. 297–308, April 2007. ISSN 0749-503X.

[61] Sites, J.B.C.A.: Architecture of the Yeast Cell Wall. *Journal of Biological Chemistry*, 1997.

[62] Montijn, R. and Rinsum, J.V.: Glucomannoproteins in the cell wall of *Saccharomyces cerevisiae* contain a novel type of carbohydrate side chain. *Journal of Biological Chemistry*, vol. 269, no. 30, pp. 19338–19342, 1994.

[63] Cappellaro, C., Mrsa, V. and Tanner, W.: New Potential Cell Wall Glucanases of *Saccharomyces cerevisiae* and Their Involvement in Mating. *Journal of Bacteriology*, vol. 180, no. 19, pp. 5030–5037, 1998.

[64] Moukadiri, I., Armero, J., Abad, A., Sentandreu, R. and Zueco, J.: Identification of a mannoprotein present in the inner layer of the cell wall of *Saccharomyces cerevisiae*. *Journal of Bacteriology*, vol. 179, no. 7, 1997.

[65] Schmidt, a., Kunz, J. and Hall, M.N.: TOR2 is required for organization of the actin cytoskeleton in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 24, pp. 13780–5, November 1996. ISSN 0027-8424.

[66] Yin, Q.Y., de Groot, P.W.J., Dekker, H.L., de Jong, L., Klis, F.M. and de Koster, C.G.: Comprehensive proteomic analysis of *Saccharomyces cerevisiae* cell walls: identification of proteins covalently attached via glycosylphosphatidylinositol remnants or mild alkali-sensitive linkages. *The Journal of Biological Chemistry*, vol. 280, no. 21, pp. 20894–901, May 2005. ISSN 0021-9258.

[67] Peyrot, G., Sands, D., Soni, P. and Travanty, E.: Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, vol. 409, no. January, 2001.

[68] Krysan, D., Ting, E. and Abeijon, C.: Yapsins are a family of aspartyl proteases required for cell wall integrity in *Saccharomyces cerevisiae*. *Eukaryotic Cell*, 2005.

[69] Komano, H., Seeger, M. and Gandy, S.: Involvement of cell surface glycosylphosphatidylinositol-linked aspartyl proteases in α-secretase-type cleavage and ectodomain solubilization of human Alzheimer β-amyloid precursor protein in yeast. *Journal of Biological Chemistry*, 1998.

[70] Komano, H., Rockwell, N. and Wang, G.: Purification and characterization of the yeast glycosylphosphatidylinositol-anchored, monobasic-specific aspartyl protease yapsin 2 (Mkc7p). *Journal of Biological Chemistry*, 1999.

[71] Causton, H.C., Ren, B., Koh, S.S., Harbison, C.T., Kanin, E., Jennings, E.G., Lee, T.I., True, H.L., Lander, E.S. and Young, R.A.: Remodeling of yeast genome expression in response to environmental changes. *Molecular Biology of the Cell*, vol. 12, no. 2, pp. 323–337, 2001.

[72] García, R., Bermejo, C., Grau, C., Pérez, R., Rodríguez-Peña, J.M., Francois, J., Nombela, C. and Arroyo, J.: The global transcriptional response to transient cell wall damage in *Saccharomyces cerevisiae* and its regulation by the cell integrity signaling pathway. *Journal of Biological Chemistry*, vol. 279, no. 15, pp. 15183–15195, 2004.

[73] Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D. and Brown, P.O.: Genomic expression programs in the response of yeast cells to environmental changes. *Molecular biology of the cell*, vol. 11, no. 12, pp. 4241–4257, 2000.

[74] Jung, U.S. and Levin, D.E.: Genome-wide analysis of gene expression regulated by the yeast cell wall integrity signalling pathway. *Molecular Microbiology*, vol. 34, no. 5, pp. 1049–1057, 1999. ISSN 1365-2958.

[75] Ash, J., Dominguez, M. and Bergeron, J.: The yeast proprotein convertase encoded by YAP3 is a glycophosphatidylinositol-anchored protein that localizes to the plasma membrane. *Journal of Biological Chemistry*, 1995.

[76] Deken, R.D.: The Crabtree effect: a regulatory system in yeast. *Journal of General Microbiology*, 1966.

[77] Perlman, P. and Mahler, H.: Derepression of mitochondria and their enzymes in yeast: regulatory aspects. *Archives of Biochemistry and Biophysics*, 1974.

[78] Pronk, J., Steensma, H. and Dijken, J.V.: Pyruvate metabolism in *Saccharomyces cerevisiae*. *Yeast*, vol. 12, pp. 1607–1633, 1996.

[79] Volschenk, H., van Vuuren, H.J.J. and Viljoen-Bloom, M.: Malo-ethanolic fermentation in Saccharomyces and *Schizosaccharomyces*. *Current Genetics*, vol. 43, no. 6, pp. 379–91, September 2003. ISSN 0172-8083.

[80] Dijken, J.P. and Scheffers, W.: Redox balances in the metabolism of sugars by yeasts. *FEMS Microbiology Letters*, vol. 32, no. 3-4, pp. 199–224, April 1986. ISSN 03781097.

[81] Boles, E., de Jong-Gubbels, P. and Pronk, J.T.: Identification and characterization of MAE1, the *Saccharomyces cerevisiae* structural gene encoding mitochondrial malic enzyme. *Journal of Bacteriology*, vol. 180, no. 11, pp. 2875–82, June 1998. ISSN 0021-9193.

[82] Redzepovic, S., Orlic, S., Majdak, A., Kozina, B., Volschenk, H. and Viljoen-Bloom, M.: Differential malic acid degradation by selected strains of Saccharomyces during alcoholic fermentation. *International Journal of Food Microbiology*, vol. 83, no. 1, pp. 49–61, May 2003. ISSN 01681605.

[83] Castellari, L., Ferruzzi, M. and Magrini, A.: Unbalanced wine fermentation by cryotolerant vs. non-cryotolerant Saccharomyces strains. *Vitis*, vol. 52, pp. 49–52, 1994.

[84] Rainieri, S., Zambonelli, C., Giudici, P. and Castellari, L.: Characterisation of thermotolerant *Saccharomyces cerevisiae* hybrids. *Biotechnology Letters*, vol. 20, no. 6, pp. 543–547, 1998.

[85] Rainieri, S. and Zambonelli, C.: The enological traits of thermotolerant Saccharomyces strains. *American Journal of Enology and Viticulture*, vol. 49, no. 3, 1998.

[86] Engel, S.R. and Cherry, J.M.: The new modern era of yeast genomics: community sequencing and the resulting annotation of multiple *Saccharomyces cerevisiae* strains at the Saccharomyces Genome Database. *Database : the Journal of Biological Databases and Curation*, vol. 2013, p. bat012, January 2013. ISSN 1758-0463.

[87] Schu, P., Suarez Rendueles, P. and Wolf, D.H.: The proteinase yscB inhibitor (PBI2) gene of yeast and studies on the function of its protein product. *European Journal of Biochemistry*, vol. 197, no. 1, pp. 1–7, April 1991. ISSN 0014-2956.

[88] Xu, Z., Sato, K. and Wickner, W.: LMA1 binds to vacuoles at Sec18p (NSF), transfers upon ATP hydrolysis to a t-SNARE (Vam3p) complex, and is released during fusion. *Cell*, vol. 93, no. 7, pp. 1125–34, June 1998. ISSN 0092-8674.

[89] Michelot, A., Costanzo, M., Sarkeshik, A., Boone, C., Yates, J.R. and Drubin, D.G.: Reconstitution and protein composition analysis of endocytic actin patches. *Current biology : CB*, vol. 20, no. 21, pp. 1890–9, November 2010. ISSN 1879-0445.

[90] Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L.Y., Toufighi, K., Mostafavi, S. and Others: The genetic landscape of a cell. *Science*, vol. 327, no. 5964, pp. 425–431, 2010.

[91] Michelot, A., Grassart, A., Okreglak, V., Costanzo, M., Boone, C. and Drubin, D.G.: Actin filament elongation in Arp2/3-derived networks is controlled by three distinct mechanisms. *Developmental Cell*, vol. 24, no. 2, pp. 182–195, 2013.

[92] Velasco, J.A., Cansado, J., Peña, M., Kawakami, T., Laborda, J. and Notario, V.: Cloning of the dihydroxyacid dehydratase-encoding gene (ILV3) from *Saccharomyces cerevisiae*. *Gene*, vol. 137, no. 2, pp. 179–185, 1993. ISSN 0378-1119.

[93] Schöner, D., Kalisch, M., Leisner, C., Meier, L., Sohrmann, M., Faty, M., Barral, Y., Peter, M., Gruissem, W. and Bühlmann, P.: Annotating novel genes by integrating synthetic lethals and genomic information. *BMC Systems Biology*, vol. 2, no. 1, p. 3, 2008.

[94] Muhua, L., Karpova, T.S. and Cooper, J.A.: A yeast actin-related protein homologous to that in vertebrate dynactin complex is important for spindle orientation and nuclear migration. *Cell*, vol. 78, no. 4, pp. 669–679, 1994. ISSN 0092-8674.

[95] Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P. and Others: Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, vol. 440, no. 7084, pp. 637–643, 2006.

[96] Mark Zabriskie, T. and D. Jackson, M.: Lysine biosynthesis and metabolism in fungi. *Natural Product Reports*, vol. 17, no. 1, pp. 85–97, 2000.

[97] Li, S. and Kane, P.: The yeast lysosome-like vacuole: endpoint and crossroads. *Biochimica et Biophysica Acta (BBA)-Molecular*, vol. 1793, no. 4, pp. 650–663, 2009.

[98] Warren, G. and Wickner, W.: Organelle inheritance. *Cell*, vol. 84, no. 3, pp. 395–400, 1996.

[99] Weisman, L.S. and Wickner, W.: Intervacuole exchange in the yeast zygote: a new pathway in organelle communication. *Science*, vol. 241, no. 4865, pp. 589–591, 1988.

[100] De Mesquita, D.S.G., ten Hoopen, R. and Woldringh, C.L.: Vacuolar segregation to the bud of *Saccharomyces cerevisiae*: an analysis of morphology and timing in the cell cycle. *Journal of General Microbiology*, vol. 137, no. 10, pp. 2447–2454, 1991.

[101] Raymond, C.K., Roberts, C.J., Moore, K.E., Howald, I. and Stevens, T.H.: Biogenesis of the vacuole in *Saccharomyces cerevisiae*. *Int. Rev. Cytol*, vol. 139, pp. 59–120, 1992.

[102] Weisman, L.S.: Organelles on the move: insights from yeast vacuole inheritance. *Nature Reviews Molecular Cell Biology*, vol. 7, no. 4, pp. 243–252, 2006.

[103] Engqvist-Goldstein, A.s.E.Y. and Drubin, D.G.: Actin assembly and endocytosis: from yeast to mammals. *Annual Review of Cell and Developmental Biology*, vol. 19, no. 1, pp. 287–332, 2003.

[104] Huckaba, T.M., Gay, A.C., Pantalena, L.F., Yang, H.-C. and Pon, L.A.: Live cell imaging of the assembly, disassembly, and actin cable–dependent movement of endosomes and actin patches in the budding yeast, *Saccharomyces cerevisiae*. *The Journal of cell Biology*, vol. 167, no. 3, pp. 519–530, 2004.

[105] Kaksonen, M., Sun, Y. and Drubin, D.G.: A pathway for association of receptors, adaptors, and actin during endocytic internalization. *Cell*, vol. 115, no. 4, pp. 475–487, 2003.

[106] Sekiya-Kawasaki, M., Groen, A.C., Cope, M.J.T.V., Kaksonen, M., Watson, H.A., Zhang, C., Shokat, K.M., Wendland, B., McDonald, K.L., McCaffery, J.M. and Others: Dynamic phosphoregulation of the cortical actin cytoskeleton and endocytic machinery revealed by real-time chemical genetic analysis. *The Journal of Cell Biology*, vol. 162, no. 5, pp. 765–772, 2003.

[107] Piper, R.C., Cooper, A.A., Yang, H. and Stevens, T.H.: VPS27 controls vacuolar and endocytic traffic through a prevacuolar compartment in *Saccharomyces cerevisiae*. *The Journal of Cell Biology*, vol. 131, no. 3, pp. 603–617, 1995.

[108] Helliwell, S.B., Losko, S. and Kaiser, C.A.: Components of a ubiquitin ligase complex specify polyubiquitination and intracellular trafficking of the general amino acid permease. *The Journal of Cell Biology*, vol. 153, no. 4, pp. 649–662, 2001.

[109] Bely, M., Sablayrolles, J.-M. and Barre, P.: Automatic detection of assimilable nitrogen deficiencies during alcoholic fermentation in oenological conditions. *Journal of Fermentation and Bioengineering*, vol. 70, no. 4, pp. 246–252, 1990.

[110] Palmisano, G., Antonacci, D. and Larsen, M.R.: Glycoproteomic profile in wine: a 'sweet'molecular renaissance. *Journal of Proteome Research*, vol. 9, no. 12, pp. 6148–6159, 2010.

Chapter 4

Conclusion

4.1 Concluding Remarks

Large-scale, MS-based proteomics is comparable in power to other more established omics technologies [1], however, it also comes with the challenges of interpreting high throughput datasets. Networks are well suited for the contextualisation, interpretation and mining of such data. In this work, networks were used as a conceptual framework within which hundreds of proteins, identified via LC-MS/MS, could be placed into biological contexts representative of the system from which the sample set was derived. The network-based contextualisation of the dataset allowed for the observation of trends, patterns and non-obvious biological connections. Moreover, existing hypotheses were investigated and new hypotheses were formulated.

4.1.1 Aim 1

Sections 2.6.3 through 2.6.5 of Chapter 2 stress the importance of peptide physico-chemical properties, especially for LC-MS/MS datasets derived from quantitative cross-species proteomics experiments. In order to appropriately analyse a dataset of this nature, a workflow was assembled to facilitate the identification and quantification of proteins. The workflow, which consisted of both existing and custom-written programs is presented in Sections 3.2.3 through 3.2.6 of Chapter 3. The workflow was designed with the goal of minimising false positive and false negative protein identifications. This was done by matching the observed spectra to theoretical spectra derived from organism-specific databases, thus ensuring that the search space matched the samples as closely as possible. This resulted in species-specific protein identifications, with each protein having a specific probability of identification.

The interpretation of quantitative information within this cross-species experiment required the application of customised criteria for isobarically-labelled peptide selection and reporting of relative fold change which were discussed in more detail in Section 3.2.5 of Chapter 3. Application of the peptide criteria resulted in a reduced risk of false quantitative signal for proteins. Thus, the first project aim of identifying and quantifying proteins from a LC-MS/MS dataset derived from an isobarically labelled, cross-species secretome sample was achieved.

4.1.2 Aim 2

Networks have been shown as an effective means for the analysis and contextualisation of omics data [2]. The second aim was to construct networks placing the identified proteins in various biological contexts. This aim was achieved with the networks described in Section 3.2.8 of Chapter 3. The biological resources used to provide context are described in more detail in Section 2.8 of Chapter 2.

4.1.3 Aim 3

The utility of the method as a tool to explore the data with the objectives of investigating existing and formulating new hypotheses was illustrated through a selection of examples of potential oenological relevance. In particular, proteins related to the cell wall, malo-ethanolic fermentation as well as proteins of unknown function were investigated within the biological contexts provided by the networks. These integrated biological contexts allowed for deeper understanding and interpretation of the identified proteins and the relationships amongst them, thus addressing the third aim of this thesis.

4.2 Future Work

The analysis of data derived from shotgun proteomics experiments is an area of research that is well suited for the application of network-based methods, not only as a visualisation and contextual scaffold, but also for problems such as protein identification and inference [3]. The work conducted in this Master's thesis may be extended in various ways. The areas of focus may be split into two parts: 1) Problems dealing with the identification and quantification of proteins; 2) The interpretation of the large numbers of proteins derived from whole proteome experiments.

4.2.1 Protein Identification and Quantification

Confidence in protein identifications may be increased by taking a consensus view from various peptide spectral matching programs and peptide and protein validation programs. The protein inference problem in particular is one area that has potential for increasing the amount of proteins that can be extracted from a LC-MS/MS dataset [3]. More sophisticated methods for the integration of quantitative signals from isobarically-labelled peptides may also be applied [4]. Moreover, improved experimental design, sample preparation and peptide separation may improve the amount of proteins characterised as well as the confidence in their identifications and quantitative signals.

4.2.2 Protein Interpretation

The construction of contextual networks presented in this work draws on the existence of extant biological knowledge from a variety of resources. These resources are continuously developing with both the volume and accuracy of the knowledge contained within them increasing. Thus, as the size of proteomics datasets and the amount of information with which to describe them increases, the size of the resulting networks constructed will also increase. Although initial visual interpretation may be hampered, networks are well suited for the mining of large datasets and a myriad of network analysis tools exists. Moreover, utilisation of such network analysis tools and investigation of the network topology itself may further increase the ability to observe system-wide patterns and trends.

In summary, a quantitative, cross-species LC-MS/MS dataset was successfully analysed, contextualised, interpreted and mined using a customised workflow facilitating protein identification, quantification and network-based contextualisation.

Bibliography

[1] Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J. and Aebersold, R.: The quantitative proteome of a human cell line. *Molecular Systems Biology*, vol. 7, no. 1, 2011.

[2] Aittokallio, T. and Schwikowski, B.: Graph-based methods for analysing networks in cell biology. *Briefings in Bioinformatics*, vol. 7, no. 3, pp. 243–255, 2006.

[3] Li, J., Zimmerman, L.J., Park, B.-H., Tabb, D.L., Liebler, D.C. and Zhang, B.: Network-assisted protein identification and data interpretation in shotgun proteomics. *Molecular Systems Biology*, vol. 5, no. 1, 2009.

[4] Breitwieser, F.P., Muller, A., Dayon, L., Kocher, T., Hainard, A., Pichler, P., Schmidt-Erfurth, U., Superti-Furga, G., Sanchez, J.-C., Mechtler, K. and Others: General statistical modeling of data from protein relative expression isobaric tags. *Journal of Proteome Research*, vol. 10, no. 6, pp. 2758–2766, 2011.