Networks and Multivariate Statistics as applied to Biological Datasets and Wine-related Omics

by

Daniel A. Jacobson

Dissertation presented for the degree of
Doctor of Philosophy (AgriSciences)
at
Stellenbosch University
Institute for Wine Biotechnology, Faculty of AgriSciences

Supervisor: Prof M Vivier

December 2013

Declaration

By submitting this dissertation 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: 2013/03/01 .....

Abstract

Networks and Multivariate Statistics as applied to Biological Datasets and Wine-related Omics

D. Jacobson

Dissertation: PhD (Wine Biotechnology)

December 2013

Introduction: Wine production is a complex biotechnological process aiming at productively coordinating the interactions and outputs of several biological systems, including grapevine and many microorganisms such as wine yeast and wine bacteria. High-throughput data generating tools in the fields of genomics, transcriptomics, proteomics, metabolomics and microbiomics are being applied both locally and globally in order to better understand complex biological systems. As such, the datasets available for analysis and mining include *de novo* datasets created by collaborators as well as publicly available datasets which one can use to get further insight into the systems under study. In order to model the complexity inherent in and across these datasets it is necessary to develop methods and approaches based on network theory and multivariate data analysis as well as to explore the intersections between these two approaches to data modelling, mining and interpretation.

Networks: The traditional reductionist paradigm of analysing single components of a biological system has not provided tools with which to adequately analyse data sets that are attempting to capture systems-level information. Network theory has recently emerged as a new discipline with which to model and analyse complex systems and has arisen from the study of real and often quite large networks derived empirically from the large volumes of data that have collected from communications, internet, financial and biological systems. This is in stark contrast to previous theoretical approaches to understanding complex systems such as complexity theory, synergetics, chaos theory, self-organised criticality, and fractals which were all sweeping theoretical constructs based on small toy models which proved unable to address the complexity of real world systems.

Multivariate Data Analysis: Principle components analysis (PCA) and Partial Least Squares (PLS) regression are commonly used to reduce the di-

mensionality of a matrix (and amongst matrices in the case of PLS) in which there are a considerable number of potentially related variables. PCA and PLS are variance focused approaches where components are ranked by the amount of variance they each explain. Components are, by definition, orthogonal to one another and as such, uncorrelated.

Aims: This thesis explores the development of Computational Biology tools that are essential to fully exploit the large data sets that are being generated by systems-based approaches in order to gain a better understanding of wine-related organisms such as grapevine (and tobacco as a laboratory-based plant model), plant pathogens, microbes and their interactions. The broad aim of this thesis is therefore to develop computational methods that can be used in an integrated systems-based approach to model and describe different aspects of the wine making process from a biological perspective. To achieve this aim, computational methods have been developed and applied in the areas of transcriptomics, phylogenomics, chemiomics and microbiomics.

Summary: The primary approaches taken in this thesis have been the use of networks and multivariate data analysis methods to analyse highly dimensional data sets. Furthermore, several of the approaches have started to explore the intersection between networks and multivariate data analysis. This would seem to be a logical progression as both networks and multivariate data analysis are focused on matrix-based data modelling and therefore have many of their roots in linear algebra.

Uittreksel

Netwerke en Meerveranderlike statistiek toegepas op Biologiese Datastelle en wyn-verwante Omika

("Networks and Multivariate Statistics as applied to Biological Datasets and Wine-related Omics")

D. Jacobson

Tesis: PhD

Desember 2013

Inleiding: Wynproduksie is 'n komplekse biotegnologiese proses wat mik op die produktiewe koördinering van verskeie interaksies en uitsette van verskeie biologiese sisteme. Hierdie sisteme sluit in die wingerd, wat van besondere belang is, asook die wyn gis en wyn bakterieë. Hoë-deurset data generasie word huidiglik beide globaal en plaaslik toegepas in die velde van genomika, transkriptomika, proteomika, metabolomika en mikrobiomika. As sulks is hierdie tipe datastelle beskikbaar vir ontleding, bemyning en verkening. Die datastelle kan de novo gegenereer word, met behulp van medewerkers, of dit kan vanuit die publieke databasisse gewerf word waar sulke datastelle dikwels beskikbaar gemaak word sodat verdere insig verkry kan word met betrekking tot die sisteem onder studie. Die hoë-deurset datastelle onder bespreking bevat 'n hoë mate van inherente kompleksiteit, beide ten opsigte van ditself asook tussen verskeie datastelle. Om ten einde hierdie datastelle en hul inherente kompleksiteit te modelleer is dit nodig om metodes en benaderings te ontwikkel wat gesetel is in netwerk teorie en meerveranderlike statistiek. Verdermeer is dit ook nodig om die kruisings tussen netwerk teorie en meerveranderlike statistiek te verken om sodoende die modellering, bemyning, verkening en interpretasie van data te verbeter.

Netwerke: Die tradisionele reduksionistiese paradigma, waarby enkele komponente van 'n biologiese sisteem geontleed word, het tot dusver nie voldoende metodes en gereedskap gelewer waarmee datastelle, wat streef om sisteemvlak informasie te bekom, geontleed kan word nie. Netwerk teorie het na vore gekom as 'n nuwe dissipline wat toegepas kan word vir die model-skepping en ontleding van komplekse sisteme. Dit stem uit die studie van egte, dikwels groot netwerke wat empiries afgelei word uit die groot volumes data wat tans

na vore kom vanuit kommunikasie-, internet-, finansiële- en biologiese sisteme. Dit is in skrille kontras met vorige teoretiese benaderings wat gestreef het om komplekse sisteme te verstaan met konsepte soos kompleksiteits teorie, "synergetics", chaos teorie, self-georganiseerde kritikaliteit en fraktale. Al die bogeneomde is breë teoretiese konstrukte, gebasseer op relatief kleinskaal modelle, wat nie instaat was om oplossings vir die kompleksiteit van egte-wêreld sisteme te bied nie.

Meerveranderlike Data-analise: Hoofkomponente-ontleding (PCA) en "Partial Least Squares (PLS)" regressie word dikwels gebruik om die dimensionaliteit van 'n matriks (en tussen matrikse in die geval van PLS) te verminder. Hierdie matrikse bevat dikwels 'n aansienlike groot hoeveelheid moontlikverwante veranderlikes. PCA en PLS is variansie gedrewe metodes en behels dat komponente gerang word deur die hoeveelheid variansie wat elke component verduidelik. Komponente is by definisie ortogonaal ten opsigte van mekaar en as sulks ongekorreleerd.

Doelwitte: Hierdie tesis verken die ontwikkeling van verskeie "Computational Biology" metodes wat noodsaaklik is om ten volle die groot skaal datastelle te benut wat tans deur sisteem-gebasseerde benaderings gegenereer word. Die doel is om beter begrip en kennis van wyn verwante organismes te kry, hierdie organismes sluit in die wingerd (met tabak as laboratorium-gebasseerde plant model), plant patogene en microbes sowel as hulle interaksies.

Die breë mikpunt van hierdie tesis is dus om gerekenaardiseerde metodes te ontwikkel wat gebruik kan word in 'n geintergreerde sisteem-gebaseerde benadering tot die modellering en beskrywing van verskillende aspekte van die wynmaak proses vanuit 'n biologiese standpunt. Om die mikpunt te bereik is gerekenaardiseerde metodes ontwikkel en toegepas in die velde van transkriptomika, filogenomika, chemiomika en mikrobiomika.

Opsomming: Die primêre benadering geneem in hierdie tesis is die gebruik van netwerke en meerveranderlike data-ontleding metodes om hoë-dimensie datastelle te ontleed. Verdermeer, verskeie van die metodes begin om die gemeenskaplike grond tussen netwerke en meerveranderlike data-ontleding te verken. Dit blyk om 'n logiese progressie te wees, aangesien beide netwerke en meerveranderlike data-ontleding gefokus is op matriks-gebaseerde data modellering en dus gewortel is in liniêre algebra.

Acknowledgements

I would like to express my sincere gratitude to the following people:

Melané Vivier, Erik Alexandersson, Antonio Ferreira, Evodia Setati, Florian Bauer, Piet Jones, Debbie Weighill, Armin Geiger, Guy Emerton and the staff and students of the IWBT.

I would like to thank the National Research Foundation, Winetech and the Swedish UD40 programme for funding.

Dedications

Bruce, Phyllis and Danielle Jacobson (father, mother and daughter, respectively).

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Aims

1.1 Introduction

The making of wine is a complex process that is affected by the plants and microbes present in the vineyard, the actions of yeast and other microbes in the cellar and the complex chemistry of aging that occurs in the bottle. The conditions under which grapevines are grown in a vineyard will affect their gene expression and metabolism and therefore the chemical content of the resulting grape berries. The chemical content of the grape berries is the starting point for further metabolism by the microorganisms responsible for the fermentation of crushed grapes (grapemust) into wine. The microorganisms present in a fermentation include those that are resident on the grape berries in the vineyard as well as those introduced (intentionally or unintentionally) in the cellar. Once grape berries have been fermented the resulting wine is either stored in barrels for further maturation (and later bottled) or directly bottled. Once bottled, the flavour and aroma profiles of wine change over time due to a complex set of chemical reactions. Thus, wine can be viewed as a complex system that is dependent upon the conditions present at each step of the grape growing and wine making process.

Past research, challenged by the complexity of this process, has been largely descriptive, and has frequently not been able to establish causalities between the various environmental or external inputs, the ecological dynamics of the process and the outputs generated by the various biological systems. However, future improvements of viticultural and oenological methods and practices and of wine-associated organisms will require a fundamental understanding of these causal relationships. In order to achieve this, high-throughput data generating tools in the fields of genomics (determining the entire sequence of an organism and identifying the location and structure of of sequence features such as genes), transcriptomics (the genome-wide profiling of gene expression patterns under different conditions), proteomics (the identification and quantification of as many of an organisms proteins as possible), metabolomics (the study of

as many-small-metabolites-as-possible in a system [7]) and microbiomics (the study of microbial populations in an ecosystem) are being applied both locally and globally. As such, the datasets available for analysis and mining include *de novo* datasets created by collaborators as well as publicly available datasets which one can use to get further insight into the systems that are being studied.

1.1.0.1 Grapevine

The availability of the grapevine genome has propelled grapevine research into the genomics era [17]. The whole genome sequence with the associated ability to analyse the transcriptome, and to a lesser extent the proteome and metabolome gives grapevine researchers the ability to implement systems biology approaches and study the grapevine's molecular responses to the environment (biotic and abiotic) or to viticultural treatments/actions. The current challenge is to predict and visualise the changes of the transcriptome, proteome and metabolome within molecular networks (for example, metabolic or signalling pathways), during a given experiment.

Research areas of considerable local interest for which datasets have been and are being generated include gene expression and metabolite profiling of grapevines in vineyard settings as well as the study of a plant's response to a pathogen such as *Botrytis cinerea*. In this thesis the focus has been on the development of statistical methods that are appropriate for the analysis of data resulting from field studies as well as methods to analyse datasets derived from transgenic lines with *B. cinera* resistance phenotypes caused by upregulation of a specific *Vitis vinifera* defence gene expressed in tobacco. Tobacco was used as it has been previously established as a *B. cinera* pathosystem [12] and it is relatively easy to work with in the laboratory compared to grapevines.

1.1.0.2 Vineyard Microbiomics

Grapes harbour a wide variety of microorganisms that play different roles in wine-making. The microbial diversity associated with grapes evolves throughout grape ripening stages [4, 15] and may be affected by various factors such as farming practices, grape berry health and geographical location [5, 6]. Yeasts play a pivotal role in alcoholic fermentation, and in the past decade a substantial body of research has shown that non-*Saccharomyces* yeasts contribute to wine flavour and aroma although they may not always persist in high ethanol content. However, most of the knowledge we have acquired to date is based on cultivation-based approaches and does not provide a holistic overview of the wine microbiome and the metabolic contributions of all the organisms. The microbiomics data focused on in this thesis was from a study using molecular methods in addition to cultivation-based approaches to investigate the heterogeneity of yeast populations within vineyards as well as the affect of different farming practices on the yeast populations across contiguous vineyards.

1.1.0.3 Yeast

Yeast of the species *Saccharomyces cerevisiae* are primarily responsible for alcoholic fermentation, the conversion of grapemust to wine. Several omic and cross-omic datasets have been created locally. These datasets often present difficult statistical challenges due to the variation in fermentation replicates and interpretive challenges as such large datasets are best viewed with as much extant biological context as possible. This thesis has focused on the development of methods which can aid in the analysis and interpretation of metabolomic and transcriptomic datasets so as to exploit the accumulated scientific knowledge that is available regarding the molecular biology of the model system *S. cerevisiae*.

1.1.0.4 Bottle Aging

Wine is a complex mixture of tens of thousands of different compounds, many of which contribute to flavour, aroma and mouth-feel characteristics and others that have important health benefits (e.g. resveratrol). As mentioned above, these compounds have their origins in the original grapes as well as the enzymatic conversions that are brought about by microorganisms during fermentation and by slow chemical reactions that happen in the bottle during aging.

At present, only several hundred compounds have been identified and most chemical analysis of wine to date has focused on a subset of this limited number of known compounds. Unfortunately, this "targeted" type of analysis is often inadequate to understand the underlying biological and chemical networks responsible for generating the compounds present in wine as it represents such a small fraction of the compounds that are present and interacting. In this regard, this thesis has focused on the development of methods to do "untargeted" chemical analysis with the goal of better understanding how the chemical composition of wine changes during aging and the underlying mechanisms responsible for these changes. This work is done with the hope that when armed with a better understanding of the entire system one can better identify key compounds of interest that need identification and further study.

1.1.0.5 Modelling Complexity

In order to model the complexity inherent in and across these datasets it is necessary to develop methods and approaches based on network theory and multivariate data analysis as well as to explore the intersections between these two approaches to data modelling, mining and interpretation. Thus, this thesis will touch on each of these aspects of this complex system with a focus on the use of networks for the development of new computational methods for the analysis and interpretation of omics datasets generated either from a wine related organism (or model species thereof) or wine itself.

1.1.1 Networks

The traditional reductionist paradigm of analysing single components of a biological system has not provided tools with which to adequately analyse datasets that are attempting to capture systems-level information.

Network theory has emerged as a new discipline with which to model and analyse complex systems. Network theory is not only useful in biology but has applications in many other disciplines including communications [13], economics [9], computer networks [8], physics, linguistics and ecology [2] to name but a few. Network theory has arisen from the study of real and often quite large networks derived empirically from the large volumes of data that has been emerging from communications, internet, financial and biological systems. This is in stark contrast to previous theoretical approaches to understanding complex systems such as complexity theory, synergetics, chaos theory, self-organised criticality, and fractals which were all sweeping theoretical constructs based on small toy models which proved unable to address the complexity of real world systems [3].

1.1.2 Multivariate Data Analysis

Principle components analysis (PCA) is a dimension reduction approach in which the dimensionality of a matrix representing many different variables is reduced to a smaller set of orthogonal components. The data is transformed into this reduced component space and plotted in two dimensions for visual interpretation. Although PCA in its earliest form was proposed by Pearson in 1901 [14] and Hotelling in 1933 [10] it did not become widely used before the computers became commonly available. It has since been used in nearly every discipline that produces numeric data in need of analysis [11, 18, 16].

Partial Least Square (PLS) combines PCA and multiple linear regression in an effort to predict dependent variables from a set of independent variables. PLS analysis involves the use of two matrices, $X_{i\times j}$ and $Y_{i\times k}$. The X matrix contains objects (i) and independent variables (j). The Y matrix contains the same objects (i) and dependent variables, (k). PLS regression endeavours to find principal components from X that are the best at predicting the variables in Y with the simultaneous decomposition of X and Y trying to explain as much of the covariance between the two matrices as possible [19, 20, 1].

1.1.3 Aims

This thesis explores the development of Computational Biology tools that are essential to fully exploit the large datasets that are being generated by systems-based approaches in order to gain a better understanding of wine-related organisms, including grapevine and tobacco (as a laboratory-based plant model),

vineyard-based microbial populations, and yeast, as well as the complex sets of chemical changes that occur during aging.

The broad aim of this thesis is therefore to develop computational methods that can be used in an integrated systems-based approach to model and describe different aspects of the wine making process from a biological perspective. To achieve this aim, methods have been developed in the areas of transcriptomics, phylogenomics, chemiomics and microbiomics. Each these omics datasets come with their own set of challenges that need to be addressed. The aims of this thesis involve the development of methods to address the following challenges:

1. In order to better understand the complex changes that occur in the volatile components during the aging of wine an untargeted GC-FID dataset of wines aged from between 2 and 60 years was generated by a collaborator in Portugal. The untargeted analysis of chromatographic datasets requires a significant amount of pre-processing in order to account for the differences between repeated chromatographic measurements. **The first aim of this thesis is to develop pre-processing methods using wavelets with which to create a refined alignment of chromatograms.**

2. In un-targeted chemiomics most of the compounds detected are of unknown structure and function. As such, methods are needed with which to do internal contextualisation via the modelling of correlative relationships amongst samples and compounds in order to better understand the complex chemical mechanisms at play. **The second aim of this thesis is to develop multivariate and network-based methods with which to model untargeted chemiomics datasets to show the correlative relationships amongst the compounds in order to further study the underlying chemical mechanisms occurring during wine aging.**

3. Transcriptomic datasets usually contain thousands to tens of thousands of variables which can present considerable statistical challenges in the data analysis. Most statistical methods currently used in the analysis of microarray-based gene expression datasets can not handle data generated with considerable variance between replicates due to unknown orthogonal variance commonly present in field studies. Gene set analysis has been shown to be useful in this regard, however very little research has been done on gene set generation. **The third aim of this thesis is to explore new ways to generate gene sets to be used in gene set analyses.**

4. Gene set analysis, as currently implemented, also has high false positive and false negative rates. **The fourth aim of this thesis is to develop**

methods to reduce the false positive and false negative rates of gene set analysis.

5. Transcriptomic datasets also present significant challenges involved in their biological interpretation. The most common output of most microarray analysis methods, a list of differentially expressed genes, is suboptimal for the generation of biological knowledge from the dataset. To create biological insights from such a dataset the experimental results need to be placed in the context of extant knowledge for further interpretation. **Thus, the fifth aim of this thesis is to develop methods to integrate as much annotation (extant knowledge) as possible about the genes and their biological context into network models which can be used for data visualisation and interpretation.**

6. In order to investigate *Botrytis-cinera*-resistant transgenic lines of tobacco the global gene expression of two such transgenic lines (*VvPGIP1* lines 37 and 45) over-expressing a plant defence gene were compared to that of wild type (SR1) with the use of the TIGR 10K potato microarray by the plant biotechnology group at the IWBT. The analysis of data from this spotted potato cDNA microrarray cross-hybridised with tobacco transcripts presented a number of challenges: the gene to probe relationships are ambiguous, the original annotation was sparse and tobacco is not a model organism so there was relatively little information with which to contextualise the results. **The sixth aim of this thesis is to use phylogenomic information to determine orthologous relationships between tobacco, potato and nine sequenced plant genomes and, combined with a cross-hybridisation model, to develop a network-based framework for the analysis, annotation and interpretation of cross-species microarray data.**

7. Cultivation-based and molecular datasets were generated in order to investigate the spatial heterogeneity of microbial communities within vineyards and to determine if farming practices have an impact on microbial population structure across vineyards. The analysis of this microbiomic dataset needed the development of new methods with which to sample, delineate and interpret microbial population structures, their geospatial distribution and their responses to environmental conditions or perturbations. **The seventh aim of this thesis is to develop novel mixed-model networks, which combine sample correlations and microbial community distribution probabilities to use for the analysis of vineyard microbiomes.**

1.1.4 Summary

The primary approaches taken in this thesis have been the development and use of networks and multivariate data analysis methods to analyse these highly dimensional datasets derived form various areas of importance to wine making. Furthermore, several of the approaches have started to explore the intersection between networks and multivariate data analysis. This would seem to be a logical progression as both networks and multivariate data analysis are focused on matrix-based data modelling and therefore have many of their roots in linear algebra.

References

[1] Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 1, pp. 97–106.

[2] Albert, R. and Barabási, A. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, vol. 74, no. 1, p. 47.

[3] Barabási, A.-L. (2011 December). The network takeover. *Nature Physics*, vol. 8, no. 1, pp. 14–16. ISSN 1745-2473.

[4] Barata, A., Malfeito-Ferreira, M. and Loureiro, V. (2012 March). The microbial ecology of wine grape berries. *International journal of food microbiology*, vol. 153, no. 3, pp. 243–59. ISSN 1879-3460.

[5] Barata, A., Santos, S.C., Malfeito-Ferreira, M. and Loureiro, V. (2012 August). New insights into the ecological interaction between grape berry microorganisms and Drosophila flies during the development of sour rot. *Microbial ecology*, vol. 64, no. 2, pp. 416–30. ISSN 1432-184X.

[6] Barata, A., Seborro, F., Belloch, C., Malfeito-Ferreira, M. and Loureiro, V. (2008 April). Ascomycetous yeast species recovered from grapes damaged by honeydew and sour rot. *Journal of applied microbiology*, vol. 104, no. 4, pp. 1182–91. ISSN 1365-2672.

[7] Cevallos-Cevallos, J.M., Reyes-De-Corcuera, J.I., Etxeberria, E., Danyluk, M.D. and Rodrick, G.E. (2009). Metabolomic analysis in food science: a review. *Trends in Food Science & Technology*, vol. 20, no. 11-12, pp. 557–566. ISSN 0924-2244.

[8] Eckmann, J.-P. and Moses, E. (2002 April). Curvature of co-links uncovers hidden thematic layers in the World Wide Web. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 9, pp. 5825–9. ISSN 0027-8424.

[9] Hidalgo, C.A., Klinger, B., Barabási, A.-L. and Hausmann, R. (2007 July). The product space conditions the development of nations. *Science (New York, N.Y.)*, vol. 317, no. 5837, pp. 482–7. ISSN 1095-9203.

[10] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, vol. 24, no. 6, p. 417.

[11] Jolliffe, I.T. (2002). *Principal Component Analysis, Second Edition*. Springer. ISBN 0-387-95442-2.

[12] Joubert, D.A., Slaughter, A.R., Kemp, G., Becker, J.V.W., Krooshof, G.H., Bergmann, C., Benen, J., Pretorius, I.S. and Vivier, M.A. (2006 December). The grapevine polygalacturonase-inhibiting protein (VvPGIP1) reduces Botrytis cinerea susceptibility in transgenic tobacco and differentially inhibits fungal polygalacturonases. *Transgenic research*, vol. 15, no. 6, pp. 687–702. ISSN 0962-8819.

[13] Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J. and a L Barabási (2007 May). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 18, pp. 7332–6. ISSN 0027-8424.

[14] Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572.

[15] Prakitchaiwattana, C.J., Fleet, G.H. and Heard, G.M. (2004 September). Application and evaluation of denaturing gradient gel electrophoresis to analyse the yeast ecology of wine grapes. *FEMS yeast research*, vol. 4, no. 8, pp. 865–77. ISSN 1567-1356.

[16] Sanguansat, P. (ed.) (2012 February). *Principal Component Analysis - Multidisciplinary Applications*. InTech. ISBN 978-953-51-0129-1.

[17] Troggio, M., Vezzulli, S., Pindo, M., Malacarne, G., Fontana, P., Moreira, F.M., Costantini, L., Grando, M.S., Viola, R. and Velasco, R. (2008). Beyond the Genome , Opportunities for a Modern Viticulture : A Research Overview. *American journal of enology and viticulture*, vol. 59, no. 2, pp. 117–127.

[18] Wold, S., Esbensen, K. and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52.

[19] Wold, S., Ruhe, A., Wold, H. and Dunn III, W.J. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, vol. 5, no. 3, pp. 735–743.

[20] Wold, S., Sjöström, M. and Eriksson, L. (2002). Partial least squares projections to latent structures (PLS) in chemistry. *Encyclopedia of computational chemistry*.

Chapter 2

Literature Review

2.1 Introduction

As this thesis is focused on the use and creation of computational methods for the analysis of biological and chemical data sets, this literature review will focus on the computational methods employed rather than on the various biological or chemical domains in which these methods have been applied. This will include a review of networks, principal components analysis (PCA), gene set analysis (GSA), chromatographic alignment and orthology detection. This is a rather broad range of topics, however, as was mentioned in the Chapter 1, the golden thread through this thesis is the use of networks and multivariate statistics for data analysis and visualisation. Each of these approaches are used in the research chapters of this thesis as either pre-processing methods on data that will be used for multivariate data analysis or network creation, or used in conjunction with networks to facilitate data analysis.

2.2 Networks

Over the past 15 years there has been an unprecedented explosion of -omics data. The dramatic improvements in the speed and cost of nucleotide sequencing technology has made it possible to sequence the genomes of thousands of organisms as well as many different members of the same species. In addition, it has become not only possible but indeed common to follow the expression of all of the genes in a a set of samples. Concomitant with this has been the improvement in shotgun proteomics and the emergence of large scale targeted and untargeted metabolomics. The result of this flood of data is the ability to monitor many of the components of biological systems in ways that were unimaginable two decades ago. However, this new -omics era also presents unprecedented challenges in data analysis. The traditional reductionist paradigm of analysing single components of a biological system in relative understanding did not provide tools with which to adequately analyse data sets that are

attempting to capture systems-level information as captured by the various -omics data types described in Chapter 1.

Fortunately, also in the past 15 years, network theory has emerged as a new discipline with which to model and analyse complex systems. Network theory is not only useful in biology but has applications in many other disciplines including communications [126], economics [73], computer networks [42], physics [2], linguistics [2] and ecology [2] to name but a few. Network theory has arisen from the study of real and often quite large networks derived empirically from the large volumes of data that has been emerging from communications, internet, financial and biological systems. This focus on real world systems, as opposed to several previous theoretical constructs which were based on small synthetic models, may explain why network theory has succeeded where other theories have failed to model the truly complex systems found in a biological context.

In order to understand networks one must first understand the basics of graph theory. Viewed from a mathematical perspective a network is described as a graph. A graph is defined as two sets,

$$G = \{V, E\} \tag{2.2.1}$$

where V is a set of nodes (also known as vertices) V_1, V_2, ... V_i and E is a set of edges that connect two nodes (elements of V). Visually, graphs are often shown as a set of circles (or other geometric shapes), each representing a node; nodes that are connected by an edge are denoted by having a line between them as shown in Figure 2.1 [2].

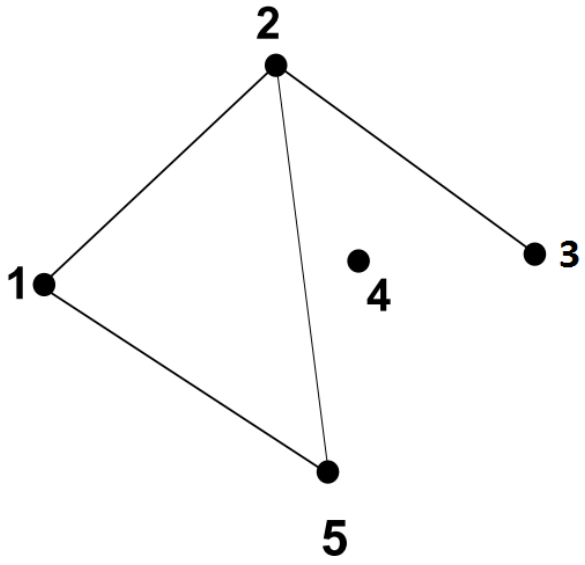


Figure 2.1: A visual representation of a network with 5 nodes and 4 edges. $V = \{1, 2, 3, 4, 5\}$ and $E = \{\{1, 2\}, \{1, 5\}, \{2, 3\}, \{2, 5\}\}$ (modified from [2]).

Networks can be considerably more complex than that shown in Figure 2.1. For example the yeast metabolic network shown in Figure 2.2 is more typical of the networks used to represent complex biological systems.

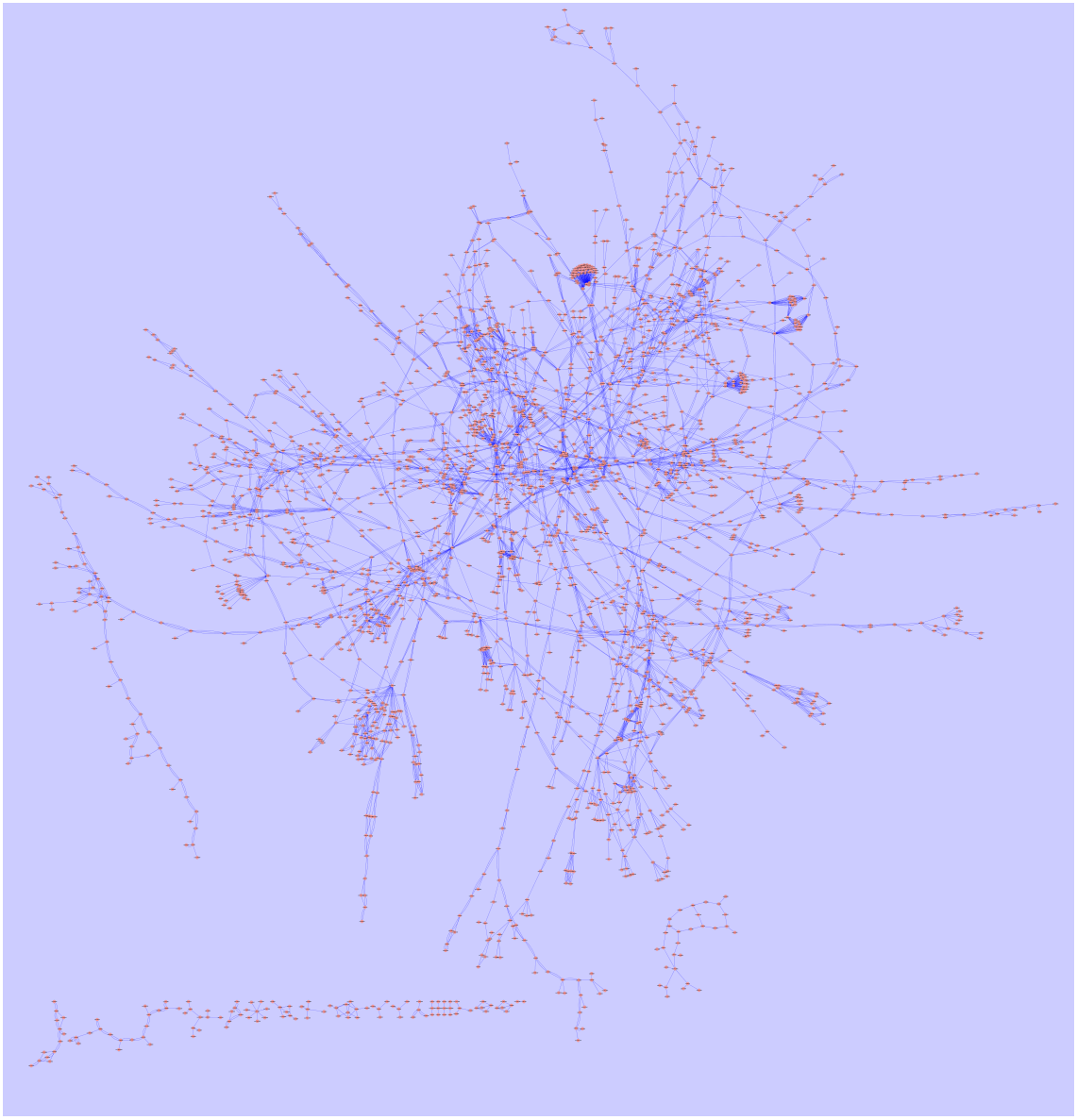


Figure 2.2: Visual representation of the Yeast metabolic network derived from the KEGG database. It is clear from the bottom of this figure that the there are several isolated subgraphs, but the majority of genes and metabolites are connected in the core network shown at the top. Furthermore this is a relatively sparse network that does appear to have complex but visible topological features.

Networks can also be represented by matrices. The simplest matrix repre-

sentation of a network is known as the adjacency matrix $A(G)$. An adjacency matrix is a square matrix with each node represented in a row and a column. An edge between two nodes is denoted by a 1 and the lack of an edge between two nodes is denoted as a 0 [2]. Thus the adjacency matrix of the network shown in Figure 2.1 can be seen in Table 2.1.

Table 2.1: Adjacency matrix for the network seen in Figure 2.1.

	node1	node2	node3	node4	node5
node1	0	1	0	0	1
node2	1	0	1	0	1
node3	0	1	0	0	0
node4	0	0	0	0	0
node5	1	1	0	0	0

As this was an unweighted network the edges are all defined to be of equal weight 1. It is possible to define a weighted network where each edge is assigned a scalar value. In this case the adjacency matrix is simply defined with the value of the edge weight, w_e, present rather than 1. Thus if the $w_{\{node1,node2\}} = 5$, $w_{\{node1,node5\}} = 1$, $w_{\{node2,node3\}} = 2$, $w_{\{node2,node5\}} = 10$ then the weighted adjacency matrix would be as shown in Table 2.2.

Table 2.2: Weighted adjacency matrix for the network seen in Figure 2.1 but with edge weights as defined in the text.

	node1	node2	node3	node4	node5
node1	0	5	0	0	1
node2	5	0	2	0	10
node3	0	2	0	0	0
node4	0	0	0	0	0
node5	1	10	0	0	0

This ability to assign weights to edges is an important feature of network annotation as it allows one to assign quantitative data to the edges between nodes which can in turn be used by algorithms to treat edges differently based in these edge weights. The distance between nodes in a network is defined as the sum of the edge weights traversed in going from one node to the other [11]. As such, in this weighted network the shortest path between node3 to node5 would traverse node2 and node1 as the sum of the edge weights, $2+5+1 = 8$, which is actually shorter than path length traversing from node3 to node2 to node5, $2 + 10 = 12$. On the other hand, in the unweighted network, in which all of the edges have the same weight, the shortest path would indeed be the path that traversed from node3 to node2 to node5. In more complex networks determining the shortest paths between nodes is a difficult problem

and several algorithms that solve it have been published, most notably those by Dijkstra [33], Bellman-Ford [11, 12], Floyd [54] and Johnson [78]. A minimum spanning tree for a connected, undirected network, is a tree that includes all of the nodes in the network with a minimum total edge cost [156]. Calculating the minimum spanning tree is quite related to determining the shortest path between two nodes and will be discussed in greater detail later in this chapter.

An adjacency matrix can also be transformed into a Laplacian matrix as an alternative matrix representation of a network. The Laplacian matrix

$$l_{i,j} := \begin{cases} deg(v(i)) & \text{if } i = j, \\ -1 & \text{if } i \neq j \text{ and } v(i) \text{ is adjacent to } v(j), \\ 0 & \text{otherwise} \end{cases} \tag{2.2.2}$$

where $deg(v(i))$ denotes the degree of $v(i)$, i.e. the number of edges incident to $v(i)$. Thus, the Laplacian matrix is the difference between the diagonal Degree Matrix (D) and the Adjacency Matrix (A) [117].

$$L = D - A \tag{2.2.3}$$

The Lapalcian matrix for the unweighted network described above can be seen in Table 2.3.

Table 2.3: Laplacian matrix for the unweighted network seen in Figure 2.1.

	node1	node2	node3	node4	node5
node1	2	-1	0	0	-1
node2	-1	3	-1	0	-1
node3	0	-1	1	0	0
node4	0	0	0	0	0
node5	-1	-1	0	0	2

A number of graph theoretic properties of a graph can be derived from its Laplacian matrix and the eigenvector and eigenvalues thereof, including the number of connected components in the graph; its algebraic connectivity (Fielder value); etc. [117]. In fact, the PCA of a graph matrix and spectral graph clustering have been linked previously by Saerens *et al.* [140].

The following subsections will include a review of the use of networks as both models of extant knowledge that may have built up over many decades of research (e.g. the many years of research in biochemistry that has yielded our current understanding of metabolism) as well as structures for exploratory data analysis of data sets such as correlation and probability networks as well as minimum and maximum spanning trees.

2.2.1 Biological Networks

Networks are useful in representing complex sets of interactions and therefore in modelling different types of extant biological knowledge and/or novel experimental data sets.

Networks have been used to describe the complex array of genetic interactions arising from gene deletion studies in which pairs of genes are deleted and the double deletion mutants compared to wild type and the single deletion mutants. Any detectable phenotype can be screened this way, the most common one being a fitness or growth phenotype. The data is then used to create a network that reflects the affects of deleting gene pairs and the network analysed in order to find functional modules of interest [76, 145, 155]. A study of particular note was performed by Costanzo *et al.* [29] in which they examined 5.4 million gene-gene deletion pairs in *Saccharomyces cerevisiae* for synthetic genetic interactions. A network based on the comparison of genetic interaction vectors showed that genes involved in similar processes were collocated in specific topological regions of the network. The network showed connectivity across all biological processes in the cell and as such was proposed to be a model for cellular pleiotropy. One of the networks derived from this work can be seen in Figure 2.3.

The relationships between human genetic diseases and the genes associated with them has been analysed as a network which allowed one to see the genetic relationships between diseases and genes. This effort provided a roadmap of disease associations that will be useful for future studies [65, 9].

Transcription factor networks have been used for a wide variety of purposes including the elucidation of key metabolic control mechanisms that may account for many of the metabolic difference between yeast strains [139], the regulation of stress responses [60] and yeast spore germination [57] to name but a few.

Protein-protein interaction (PPI) networks are derived from affinity purification/mass spectrometry and yeast-two-hybrid datasets. PPI networks have been used to infer information about proteins' functions and to identify network motifs that give a better understanding of the systems associated with different diseases such as cancer [148] or for identifying protein complexes [16, 174]

An interesting example of the creation of a network from extant knowledge is the flavour network. Anh *et al.* [1] used Fenaroli's handbook of Flavour Ingredients in order to create a bipartite network of the flavour compounds found in different food ingredients. As can be seen in Figure 2.4 the network contains an edge between each ingredient with it's constituent flavour compounds. Such a network allows one to see quickly which ingredients share flavour compounds. They then created a flavour network of ingredients by creating edges between ingredients that shared flavour compounds. The edges were weighted based on the number of ingredients that were shared between

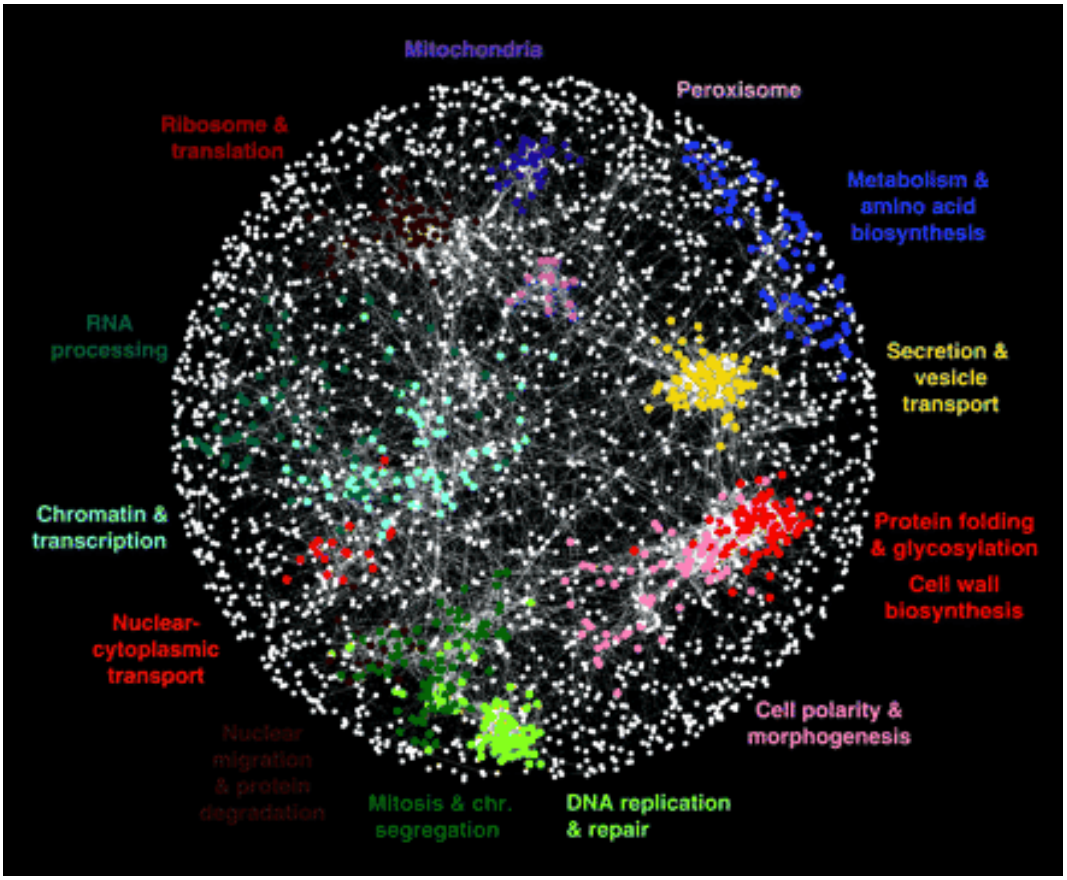


Figure 2.3: Pearson Correlation network of gene interaction vectors. Node colouring indicates GO enrichment of the terms listed in the figure (from [29]).

them.

Figure 2.4 shows this process for two recipes. In order to explore the impact of ingredient choices and the underlying flavour compounds on cuisines around the world they built a large network consisting of 381 ingredients and 1,021 flavour compounds found in these ingredients. The resulting network was quite dense, so for visualisation purposes they developed a backbone extraction method in order to select the most significant edges. The resulting network can be seen in Figure 2.5. Note that different categories of ingredients segregate to reasonably distinct topological areas in the network. The flavour network was used to evaluate food pairing as a function of network topology in order to determine if ingredient pairs that are highly connected in the flavour network are favoured or avoided in recipes from around the world. In order to test this the authors used 56,498 recipes representing North American, Western European, Southern European, Latin American, and East Asian cuisines. Among other things they found that North American and Western European cuisines tend to have ingredients that share flavour compounds. However, East Asian and Southern European cuisines tend to have ingredients that do not

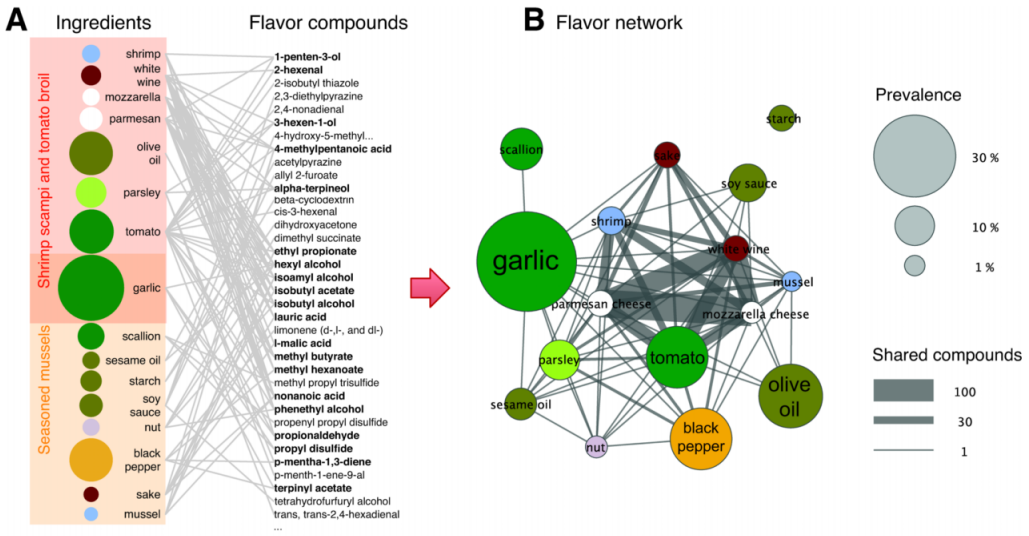


Figure 2.4: Flavor network. (A) Ingredients and flavour compounds are considered as nodes and an edge is created between ingredients and the flavour compounds they contain yielding a bipartite network. (B) Edges are created between ingredients if they share one or more flavor compounds. Edges are weighted (and visually scaled) with the number of flavour compounds shared between ingredients. Nodes are scaled to indicate the prevalence of the ingredients in recipes (modified from [1]).

share flavour compounds. This paper starts to give a chemical understanding of the clear empirical difference between cuisines from different regions and is a striking example of the power of network modelling of extant data that was not originally collected with networks in mind at all. One can imagine that a similar study using networks to model the chemical composition of wines with the use of untargeted metabolomics, when combined with sensory and or wine-food pairing data (and taking into account the flavour compounds present in the food) could also lead to fascinating insights.

2.2.1.1 Metabolic networks

Metabolic networks are the results of decades of research in biochemistry. Biochemists often would study a single metabolic enzyme at a time, carefully characterizing the enzyme and the reaction that it catalyses. The product of one metabolic reaction is usually the substrate of another reaction. Thus, as the number of reactions studied grew it was possible to link reactions together into pathways. As such, our knowledge of metabolism was created incrementally from the ground up with small sets of binary relationships between substrates, products and the enzymes that catalysed their conversion. These reactions were extracted out of the literature over time into textbooks and eventually into the Boehringer wall chart [113]. However, paper-based representations of

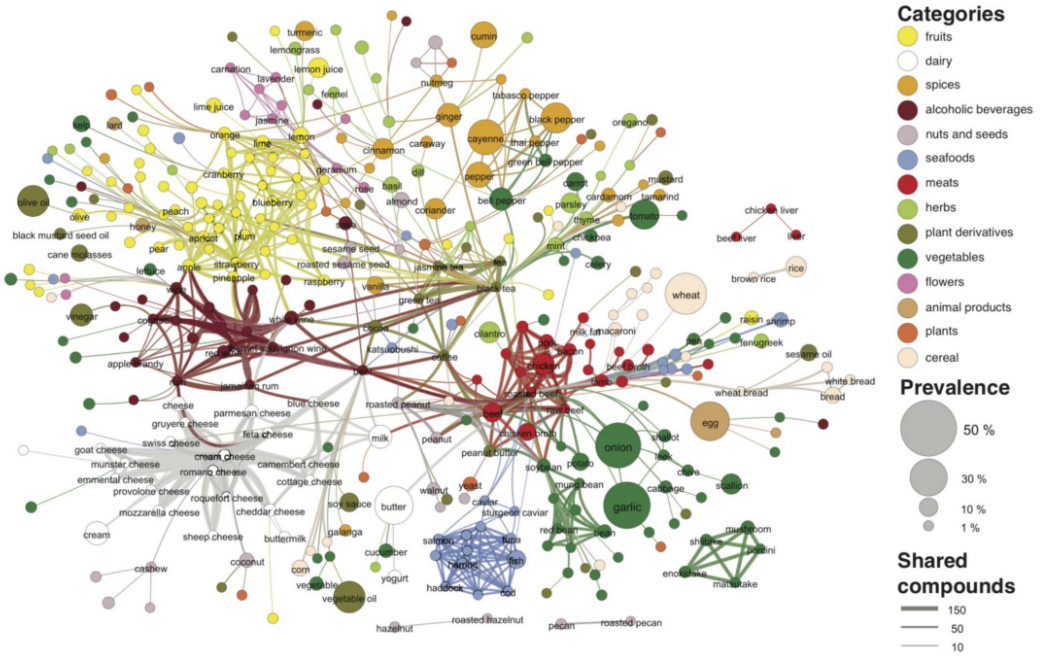


Figure 2.5: Backbone of the complete flavor network. Ingredients are nodes; nodes are coloured by categories; nodes are scaled to indicate ingredient prevalence in recipes. Edges denote that a significant number of flavour compounds are shared by two ingredients; edges are scaled to indicate the number of flavour compounds shared (from [1]).

metabolism were very limiting as they did not allow for computational analysis of the information or its structure.

As a result, a number of different efforts have been made to computerise information about metabolism. Many of the currently available metabolic databases can be grouped into a few families as shown in Table 2.4.

Each member of the metabolic database families shown in Table 2.4 shares a common schema and code base with which to create and query each of the different species databases within each family. MetaCyc, KEGG, Model SEED and Reactome each use their own reference pathway database in order to predict the pathways present in a newly sequenced organism. In contrast the BiGG family of databases do not have a reference database, rather each database was created manually by Palsson's group at the University of California San Diego [144]. In a hybrid approach, the databases in the Model SEED family [72] were created with the use of a pipeline involving both automated processing and manual curation. On the other hand the databases in the KEGG family [82] as well as those in the MetaCyc family [18] were made by a variety of different groups making use of their respective reference databases and software. Due to this automated approach the KEGG and MetaCyc families each cover well over a thousand species from many different kingdoms and

Table 2.4: Summary metabolism database families (from [86]).

Database family	MetaCyc	KEGG	Model SEED	Reactome	BiGG
web address	biocyc.org	www.genome.jp/kegg/	www.theseed.org/models	www.reactome.org	bigg.ucsd.edu
curation	+	−	−	+	+
number of organisms	>1000	>1000	>200	21	6
genome	+	+	+	−	−
proteome	+	+	+	+	−
reactions	+	+	+	+	+
metabolites	+	+	+	+	+
pathways	+	+	+	+	−
registration required	_a	−	_a	−	+

[a](registration is required for building models, but not for viewing existing models)

phyla [86].

The KEGG and Metacyc families have their origins in databases that were started in the early to mid 1990s. The Metacyc/Biocyc family of databases and software tools [19, 88, 18] started in Peter Karp's group as Ecocyc [85, 89, 87, 90], a database focused on the metablolism of a single species, *Escherichia coli*. This effort was generalised to apply to other species in a system known as Metacyc [91] and the resulting software has been used by other domain specialists to create over 1000 species-specific pathway databases [86].

KEGG was the database used for the metabolic network created in Chapter 4 of this thesis and, as such, will be the primary focus of the rest of this section. KEGG was initially created by computerising the binary relationships between metabolic compounds and enzymes as found in paper form in the Boehringer wall chart, and to a lessor extent from a compilation by the Japanese Biochemical Society, some text books and a few online resources [82]. KEGG's primary model was simply collections of the binary metabolic relationships (edges) into subgraphs of metabolism (pathways). The primary interface for KEGG is pathway focused with individual pathways represented as metabolic wiring diagrams as seen in KEGG's representation of the *Saccharomyces cerevisiae* TCA cycle in Figure 2.6. KEGG's visualisation approach has been to display all of the known components of a particular pathway across all species and then to denote with green boxes those enzyme functions corresponding to EC Enzyme numbers that are thought to be present in the organism of interest.

By using this approach of a master template for metabolism, KEGG has been able to add metabolic reconstructions of many species as their genomes have been sequenced. The use of orthologous inference allows them to predict what enzyme functions are present in a newly sequenced organism and then generate diagrams with those functions highlighted with green boxes [125,

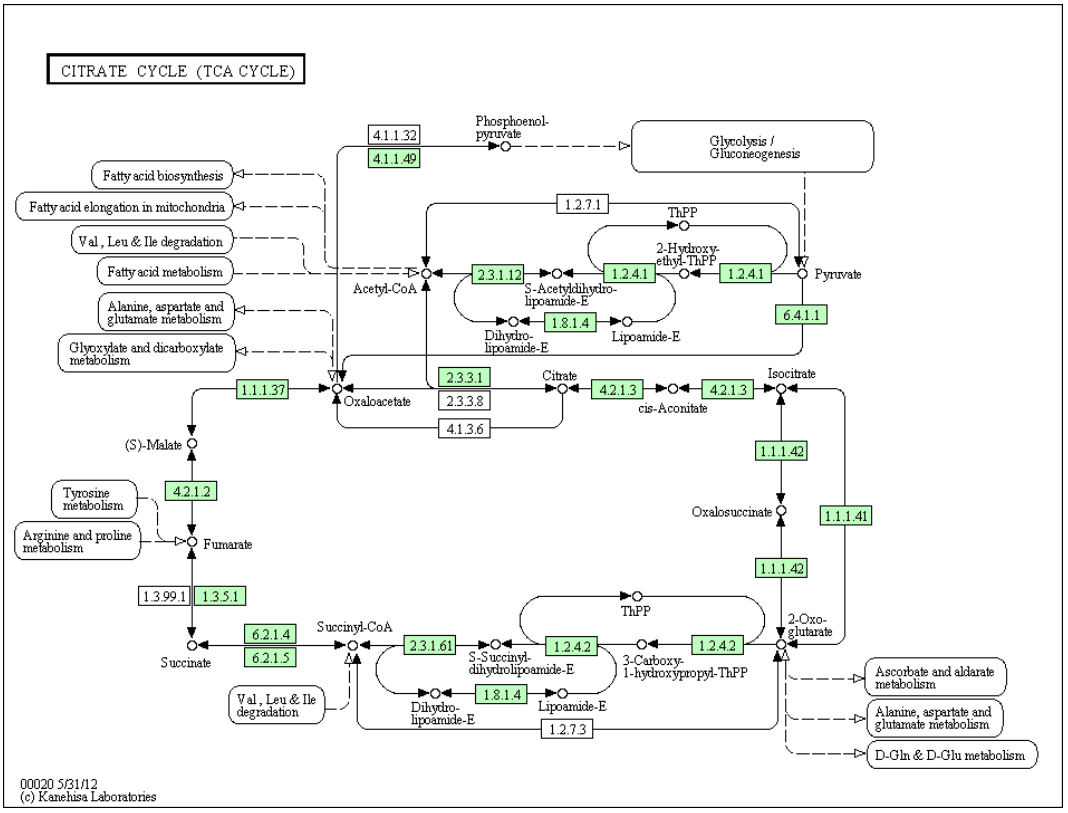


Figure 2.6: KEGG's representation of the TCA cycle of *Saccharomyces cerevisiae*.

83]. More recent work on KEGG has focused on its ortholgous relationships between genomes, its functional hierarchies and changes to its visual interface to allow the display of differential expression values of the enzymes represented in pathways as shown in Figure 2.7, to make it a useful tool for the localised interpretation of transcriptome or proteome data sets [84].

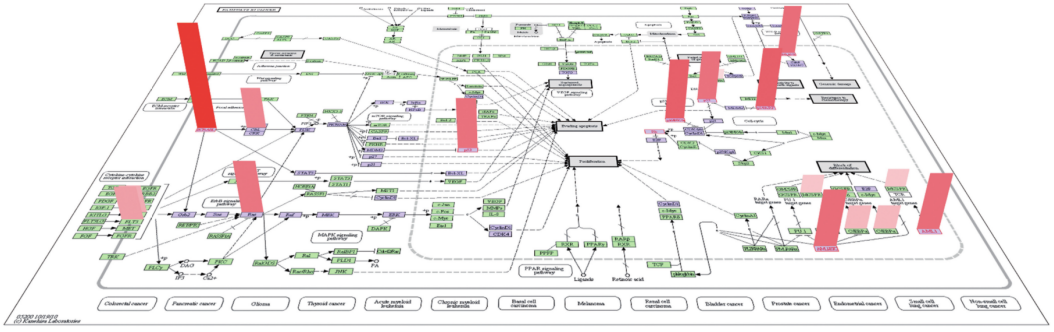


Figure 2.7: Visualisation of differential expression of enzymes in KEGG Pathways (from [84]).

KEGG's primary focus to this day is on pathways. The database is not described or distributed as a network. Rather, it is distributed as a series of XML files, each describing a different pathway, and within the pathway describing the relationships (edges) between metabolic compounds, reactions and enzymes. However, by parsing the nodes (compounds, enzymes and reactions) and relationships amongst them (edges) one can reconstruct the entire known metabolic network of an organism as seen in Figure 2.2.

KEGG provides inferred pathway information for over 2000 species. However, their orthology model (best BLAST [4] hit) is overly simplistic and their biggest focus is on primary metabolism. As such, although it is one of the most well known pathway databases there are others which tend to focus on specific sets of organisms and apply more domain expertise to those organisms and provide more information about secondary metabolism. An example of this is the plant-centric collection of pathway databases that work under or are affiliated with the Plant Metabolic Network (PMN), namely PlantCyc and the associated single-species/taxon databases: AraCyc (Arabidopsis), CassavaCyc (cassava, yucca), ChlamyCyc (a green alga), CornCyc (corn, maize), GrapeCyc (wine grape), MossCyc (moss), PapayaCyc (papaya), PoplarCyc (poplar), SelaginellaCyc (a lycophyte) and SoyCyc (soybean) [21, 178]. These databases are all based on the Metacyc software.

2.2.2 Data derived networks

Our current knowledge of the interconnections in biological, ecological and chemical systems is woefully incomplete. One of the goals of systems biology and systems chemistry (the study of as many components as possible of complex chemical mixtures and the chemical mechanisms inherent therein) in the -omics era is to infer and model the underlying systems with the use of high throughput data sets. As such, one of the major goals in systems biology and systems chemistry is network reconstruction in which one is trying to use data sets to infer what the underlying process or community might look like and form hypothesis as to how it may be regulated. In contrast to using networks to model extant knowledge, this section will focus on methods with which to build networks that are generated directly from experimental data sets with the use of correlation metrics or probabilities with which to weight the edges.

2.2.2.1 Correlation networks

Correlation networks have been used in a wide variety of fields including geophysics [69], geology [149, 13], finance and economics [107, 150], psychology [7], climatology [36], ecology [179] and molecular biology [49, 102, 177, 100, 6, 141]. Most of this section will focus on the use of correlation networks in transcriptomics but the concepts can be applied to nearly any multivariate data set.

Most biological objects, be they genes, proteins, metabolites, cells or organisms do not function in isolation but rather are members of groups or networks which work collectively to carry out different functions within a cell, organism or population. At the level of gene transcription the coordinated behaviour of gene expression is, in part, attributable to transcriptional regulation. One example of this can be seen in the peroxisome proliferator-activated receptor (PPAR) transcription factors. PPARs transcription factors control the expression of hundreds of genes across a wide variety of molecular and biological functions. Such coordinated transcriptional activity enables an organism to react quickly to changes in its environment. From many examples like this it is often possible to infer that genes that are coexpressed over a range of different experimental or natural perturbations are somehow involved in similar biological functions [49]. This phenomena is known as the "guilt-by-association principle" and an inference that is commonly made is that if one observes that a gene of unknown function is co-expressed with genes of known function, then the unknown gene may be hypothesised to be somehow involved in the function of the known genes. This has been shown to be true for many genes across a broad range of species including plants, [6], humans [106, 171], yeast [172], mice [171] and rats [171]. Coregulatory relationships amongst genes are often evolutionarily conserved among yeast, flies, worms and humans and have been used to find common functional modules in the expression correlation networks for these species [153] as well as between maize and rice [53] amongst others [141, 115].

In order to create a gene expression correlation network each microarray probeset is represented as a vector with each column being the normalised fluorescence value for that probe across different experimental conditions (perturbations). An all-against-all comparison is done so that a correlation coefficient is calculated for every possible probeset pair. As discussed later it is possible to use many different comparison metrics with which to compare probeset vectors. However, by far the most frequently used comparison metric is the Pearson product-moment correlation coefficient [129]:

$$r = \frac{\Sigma_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\Sigma_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{2.2.4}$$

where $\bar{X}$ and $\bar{Y}$ are the means of the X and Y vectors respectively and n is the number of elements of each vector.

A weighted correlation network includes the r value as the weight of each edge between probesets. An unweighted correlation network simply includes all edges whose $|r|$ value is above some predetermined threshold. A thresholded, weighted network includes all edges whose $|r|$ value is above a predetermined threshold and assigns r to be the weight of each edge [177]. Genes with a high degree, i.e. many edges connecting it to other genes, indicating that its

expression is correlated with many other genes, are typically known as hubs in the network [49].

Two approaches are often taken for the analysis of gene expression correlation networks, namely targeted and untargeted. The targeted approach uses a preselected set of genes of interest as bait with which to fish for correlation neighbours in the global network. The genes of interest can be selected in any manner of ways from extant knowledge gathered from the literature and/or from previous experiments in the laboratory that indicate that they may be involved in related biological processes. Breadth-first searches are conducted with a preselected radius using the target genes as initiation points. A breadth first search traverses a network by indexing all of the neighbours of an initial node and then then visiting all of their neighbours in turn, avoiding nodes that have already been visited (to prevent infinite loops). This process can be visualised in Figure 2.8.

This approach can be used to find network modules amongst the preselected genes of interest and/or for finding new genes which may be involved in the biological process being studied (as defined by the original genes of interest).

A good example of the use of targeted correlation network mining can be seen in a study of flavonoid and phenylpropanoid-related genes. In this study fifty-four known genes encoding proteins associated with flavonoid and phenylpropanoid metabolism were included in a large correlation network. Initial modules were formed by simply looking at the correlation relationship amongst these 54 genes, yielding four interconnected network modules. Subsequently, each of these modules was used as bait with which to extend the modules to include genes not previously known to be associated with flavonoid or phenylpropanoid metabolism. These unknown genes were proposed to be good candidates for further characterisation to determine if and how they are involved in flavonoid or phenylpropanoid metabolism [141]. The process for the first step can be visualised in Figure 2.9.

The untargetted approach on the other hand uses the entire network of all genes or a very large group of genes that result in an atlas of gene coexpression. Often these larger networks are examined with a range of network metrics in order to learn more about the properties inherently embedded in the network or for *de novo* module mining [141]. A recent example of this was an *Arabidopsis* gene network created with the use of a graphical Gaussian model [100]. A Gaussian network is "an undirected probabilistic graphical model estimating the conditional dependence between variables" [100]. A Gaussian network is derived from Pearson correlation coefficients which have been conditioned against the correlation with all other objects. The authors claim that the sparser networks that they generated with Gaussian networks were more representative of the underlying metabolic pathways that they were modelling. This interesting approach certainly warrants further exploration, perhaps in combination with other types of correlation coefficients to be conditioned.

As mentioned above, the Pearson correlation coefficient is the most fre-

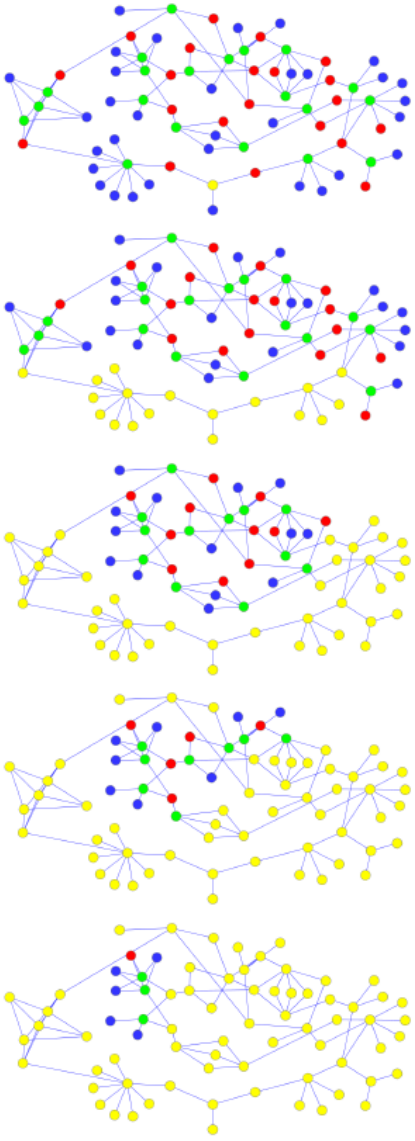


Figure 2.8: Visualisation of a breadth-first search, each step in this example has a radius of 2. Yellow indicates a visited node.

quently used metric to construct correlation networks. However, there are a number of other vector comparison metrics which could be used for edge weighting. One such example was given above where the Pearson coefficient was conditioned with graphical Gaussian model which had significant effects on the properties of the resulting network. In a recent paper Kumari *et al.* [102] compared eight different vector comparison metrics, namely, Spearman rank correlation [151], Weighted Rank Correlation [135], Kendall [93], Hoeffding's D measure [74], Theil-Sen [161], [146], Rank Theil-Sen [102], Distance Covariance [97], and Pearson [129] in an effort to see the effects of each metric on

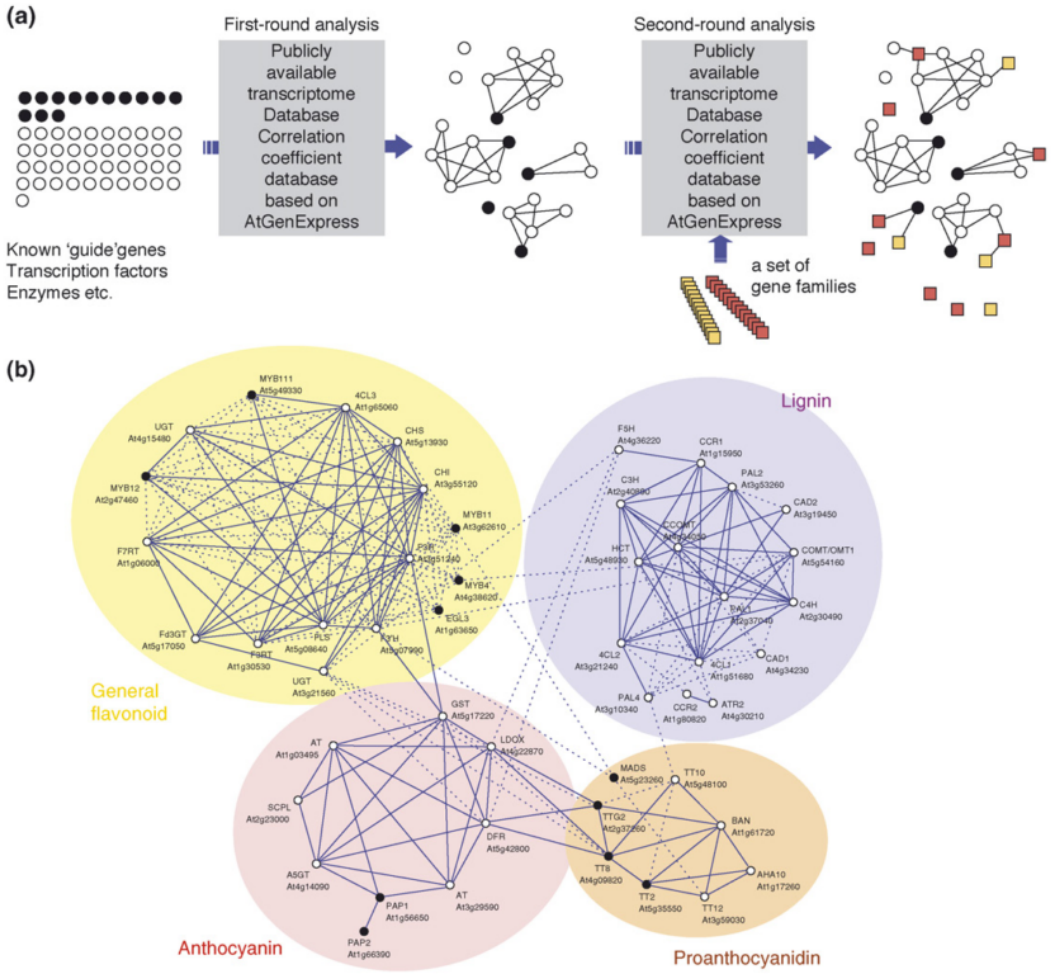


Figure 2.9: Mining correlation networks for modules and unknown genes involved in flavonoid and phenylpropanoid metabolism. (a) Targeted network mining. Fifty-four genes known to be involved in flavonoid and phenylpropanoid metabolism are used to find modules in the correlation network. First, correlation networks ($r > 0.5$) are built amongst the 54 genes to form 4 interconnected modules which can be seen in (b) (modified from [141]).

their tasks of interest, pathway-gene association and finding coordinated transcription factors in this case. They found that the different metric did differ in performance with regard to different tasks. More specifically, they found that the Spearman, Hoeffding and Kendall metrics performed well in the gene-to-pathway association task and the Theil-Sen, Rank Theil-Sen, Spearman, and Weighted Rank methods were better able to identify coordinated transcription factors involved in regulating the same biological processes. Furthermore, they found that the Pearson and Distance Covariance methods had distinctly different behaviours than the other methods [102]. To some degree this is not surprising as Pearson is designed to find linear correlation whilst the other methods are designed to look at different types of non-linear correlations. As

such, some metrics will likely perform better with some data sets and desired patterns than others. Another potential issue in this study is the use of co-existence in pathways as its judgement criteria. As our knowledge of pathways is incomplete and the vast majority of pairs found are not in pathways together this could be measuring a secondary noise affect rather than a primary phenomena. Rather than declare some metric to be better or worse than others perhaps it is better to simply acknowledge that each metric will likely give a different perspective on the data. This is exactly the approach taken in a paper that compared Pearson correlation to a mutual information metric for climatology data. They concluded that there are differences in the results using the two metrics and that mutual information provides another interesting perspective with which to interpret climatology data sets [36]. It would seem that there are opportunities to further explore the effects of different metrics on a range of network properties.

There are a number of things to keep in mind when using correlation networks. First and foremost is that correlation is not the same as causation. It is possible for two variables to be coincidentally correlated even though the underlying causes of their covariation are completely unrelated. As such, correlation network analysis should be thought of as high throughput hypothesis generation with further investigation needed to confirm or refute causation. Furthermore, only transcriptionally coregulated genes will be found with this method. As such, genes that are part of the same biological function but not transcriptionally coregulated will be missed by this approach. In addition there can be other layers of regulation in a biological system (translation, transport, degradation, kinetics, etc.) that can confound this analysis and lead to false positives or false negatives. In addition, the strength of correlations between genes will depend on the number and types of conditions examined. It is also possible that a set of genes are strongly coregulated under some conditions and not under others. Thus it is possible for global correlation values to be quite low while local correlation values for those same genes may be quite high [6, 141].

2.2.2.2 Probability networks

Probability matrices, also known as Markov matrices, stochastic matrices and substitution matrices are the common foundation for several mathematical models relevant to molecular biology and networks. Probability network models in turn are the combination of graph theory and probability theory which attempt to address both uncertainty and complexity. Probability theory can be viewed as annotating the edges between nodes and the nodes can be viewed as states (samples in some cases). Many multivariate probabilistic approaches are simply special cases of probability network formalism, including mixture models, factor analysis, Markov chains, hidden Markov models (HMMs), Kalman filters, Markov networks, Bayesian Networks and Markov random fields [80].

These approaches can be categorised as those that use undirected graphs and those that use directed graphs. Three of the most commonly used probabilistic networks in biology are Bayesian networks (directed graph), hidden Markov models (directed graph) and Markov networks (undirected graph). All of the networks used in this thesis are undirected graphs.

Hidden Markov models are the simplest kind of a dynamic Bayesian network. HMMs and Bayesian networks are outside the scope of this thesis but for the sake of completeness HMMs will be discussed here as an example of dynamic Bayesian networks that are perhaps the most well known applications of probability matrices in molecular biology as they used for sequence database searching or gene finding. To define an HMM of a gene family one first examines a multiple alignment of that family. A probability matrix is generated for each position in the sequence as follows: Create an $m \times n$ matrix where m is the number of sequences that have been aligned and n the number of nucleotides in the consensus sequence. Consider the alignment shown in Figure 2.10:

```
A C A - - - A T G
T C A A C T A T C
A C A C - - A G C
A G A - - - A T C
A C C G - - A T C
```

Figure 2.10: Alignment of DNA sequences, $m = 5$, $n = 9$ (from [98]).

The probability of finding a specific nucleotide (A,C,G,T) at any position is simply the number of those nucleotides present divided by m. After the third position in the alignment, 3 out of 5 sequences have "insertions". Therefore the probability of making an insertion is 3/5 and thus 2/5 for not making one. This process is called column normalisation and results in a probability matrix at each position. These probability matrices are then used in an HMM to calculate an overall transition probability for any sequence of interest as shown in Figure 2.11. Each box in Figure 2.11 represents a state, which contains the probability matrix for the nucleotides at that position (as denoted by the histograms). The transitions between states are indicated by arrows, the thickness of which has been scale to indicate their probability.

In order to calculate the probability of any given sequence one simply multiplies the probability of the observed nucleotide by the probability of the transition and so on, in this case the probability of the consensus sequence would be [98]:

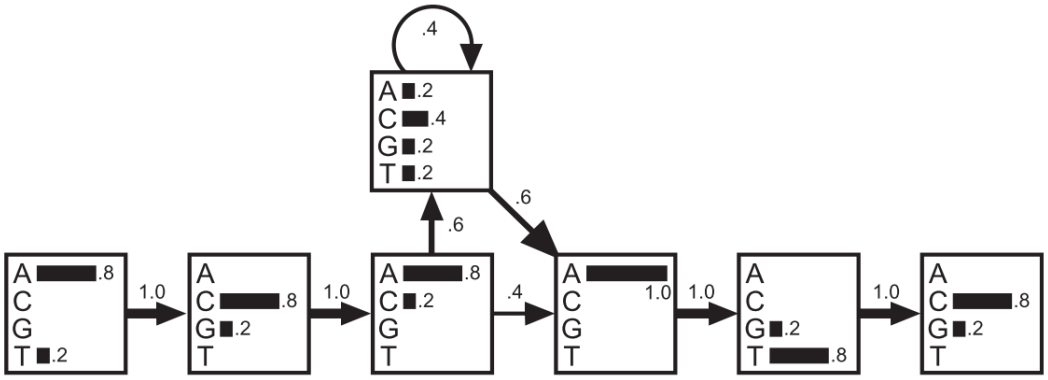


Figure 2.11: Hidden Markov Model (from [98]).

$$\begin{aligned} P(\text{ACACATC}) =& (0.8 \times 1) \times (0.8 \times 1) \times (0.8 \times 0.6) \\ & \times (0.4 \times 0.6) \times (1 \times 1) \times (0.8 \times 1) \times 0.8 \\ =& \sim 0.047 \end{aligned} \tag{2.2.5}$$

The probabilities of each of the sequences in the alignment can be seen in Table 2.5.

Table 2.5: Probabilities of the aligned sequences as well as the Consensus and "Exception" sequences (from [98]).

	Sequence	Probability ×100	Log odds
Consensus	A C A C - - A T C	4.7	6.7
Original	A C A - - - A T G	3.3	4.9
sequences	T C A A C T A T C	0.0075	3.0
	A C A C - - A G C	1.2	5.3
	A G A - - - A T C	3.3	4.9
	A C C G - - A T C	0.59	4.6
Exceptional	T G C T - - A G G	0.0023	-0.97

HMMs have been used to study the compositional heterogeneity in DNA [27] for differing evolutionary rates in phylogeny construction [50] as well as for examining genome composition [28]. The use of HMMs as probabilistic profiles for protein families was first proposed in 1993 [71]. The use of HMMs to search protein databases for the identification of protein family members was introduced by Krogh *et al.* [99] and revisited by Barrett *et al.* [10].

Markov networks, which are undirected graphs, are used in Chapter 6 and are used as the basis for Markov clustering in Chapter 5 of this thesis. Both of these approaches are really modelling a random walk on a graph which can be viewed as a special case of a Markov chain. Markov chains are defined as a system that undergoes state transitions between a finite number of states.

Markov chains are considered to be without memory, meaning that the next state depends only on the current state irrespective of anything that preceded it [118]. Markov chains are usually based on a directed graph so that the transition between states may not have the same probability depending on the direction of the transition. However, the Markov chain of a random walk is said to have the property of reversibility, which means that the probability of traversing an edge in one direction or the other are effectively equal. In order to represent the transition probabilities of a random walk on a network, the weighted adjacency matrix of the network is first converted into a stochastic matrix. This is achieved by dividing each element of each column by the sum of that column. The result is a Markov matrix in which the probability of a random walk from a node going to another node is represented by each element of the originating node's column. Thus the ij element of the matrix is the probability of travelling from node j to node i [167, 46]. This stochastic adjacency matrix is the first step in the Markov clustering of a graph with MCL which is discussed in more detail in the orthology detection subsection of this chapter.

2.2.3 Minimum and maximum spanning trees

As previously mentioned, a minimum spanning tree (MiST) for a connected, undirected network, is a tree that includes all of the nodes in the network with a minimum total edge cost [156]. An MiST can thus be used to extract the core backbone of a network representing the shortest distance between all nodes. A maximum weight spanning tree (MaST) can be achieved with the use of a minimum spanning tree algorithm simply by taking the inverse of the edge weights. For a correlation network, with the correlation metric, r, assigned as edge weights, this can be achieved by substituting in the correlation distance measure, d, as the edge weight. Correlation distance is simply defined as follows:

$$d = 1 - |r| \tag{2.2.6}$$

The most well known algorithms for calculating an MiST are those by Dijkstra [33], Bellman-Ford [11, 12], Krustall [101], Floyd [54] and Johnson [78]. Johnson actually uses the Bellman-Ford algorithm to transform a network in order to remove negative weights which allows Dijkstra's algorithm to be used on the transformed network.

Dijkstra's algorithm can be described as follows [33]:
Start at an initial node. The distance of a node X is the distance (sum of the edge weights) from the initial node to X. The algorithm assigns initial distances to all nodes and uses a step by step approach to try and improve upon them.

1. Set the distance to the initial node to zero and the distance to all other nodes to infinity.
2. Declare all nodes to be unvisited.
3. Set the state of the initial node to be current.
4. Create the unvisited set which contains all of the nodes except the initial node.
5. Examine each of the unvisited neighbours of the current node and calculate their preliminary cumulative distances.
6. If the preliminary cumulative distance is less than the previously recorded tentative distance of a node, store this distance as the new distance for the node. Neighbouring nodes are not marked as "visited" at this time, and remain in the unvisited set.
7. After examining all neighbour nodes, change the state of the current node to visited and delete it from the unvisited set.
8. If the smallest tentative distance among the nodes in the unvisited set is infinity, then stop.
9. Choose the unvisited node with the smallest preliminary distance, change its state to current, return to step three.

MSTs (MiSTs and/or MaSTs) have been used in a wide range of areas of biology including phylogeny [112, 23], multiple sequence alignment [70], gene expression data analysis [173, 134], immunology [59], proteomics [68], genotype clustering [8], panbiogeography [41], haplotype analysis [160] and genome assembly [22] to name a few.

It appears that there has not been a previous report of using an MST on a combined correlation and probability network as was done in Chapter 6 of this thesis.

2.3 Principal components analysis

Principle components analysis (PCA) is commonly used to reduce the dimensionality of matrix in which there are a considerable number of potentially related variables. PCA is a variance focused approach where components are ranked by the amount of variance they each explain. Components are, by definition, orthogonal to one another and, as such, uncorrelated. The sample and variable vectors are projected onto the new principal-component-based coordinate system and can be plotted to view the relationships amongst them in this new coordinate space [79].

The origins of PCA can be traced back to Pearson's geometric approach in 1901 [130] and Hotelling's algebraic approach in 1933 [75]. Since computers have become commonly available the use of PCA has become widespread with literally thousands of papers published that have used PCA. PCA has been used on multivariate data from nearly every discipline imaginable including (but not limited to) taxonomy, biology, pharmacy, finance, agriculture, ecology, health, physics, architecture, psychology, chemistry, climatology, demography, economics, food research, genetics, geology, meteorology and oceanography [79, 170, 143] . Most relevant to this thesis is its use in the analysis of networks and chromatographic profiles. The use of PCA on the matrices associated with networks has been very sparse, with the prime example being the connection between the PCA of a graph matrix and spectral graph clustering [140]. The use of PCA in the evaluation of chromatographic profiles has been widespread and was extensively reviewed by Cserhati [30]. Rather than focus on the broad range of uses of PCA this review will focus on the geometric interpretation and mathematical description thereof.

2.3.1 Geometric interpretation

This process can be described geometrically in terms of a Euclidean coordinate system as follows: The first principal component (P_1) can be viewed as a line that passes through the mean of the data set and is the best fit that encompasses the minimum of sum of the squares of the data points from said line as can be seen in Figure 2.12.

To create the second principal component P_1 is first subtracted from the data points and the same line fitting procedure is followed, creating P_2 which is orthogonal to P_1 and so on for each principal component [170].

The influence of each variable on each component can be seen as the cosine of the angle between the vector of the original variable axis and the principal component vector. The closer the principal component is to the variable axis, the smaller the angle and therefore the higher the loading coefficient, p. Thus, variables with similar loading values are having similar affects on the principal component. The loading values of variables are an approximation of their covariance and variables with similar loading values are often interpreted to be correlated with one another [79].

2.3.2 Mathematical description

2.3.2.1 Matrix-based description of a PCA model

If X is a data matrix composed of objects (rows) and variables (columns) then it can be described in terms of principal components as follows:

$$X = M + E \tag{2.3.1}$$

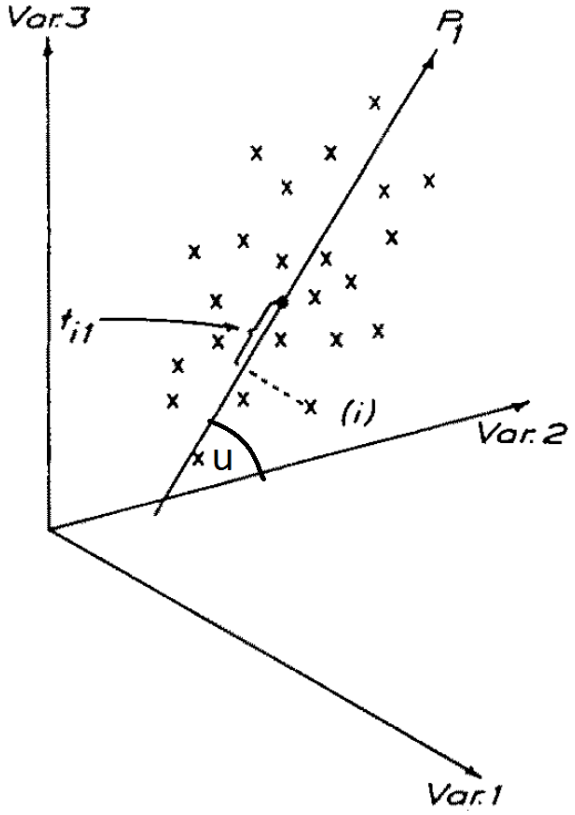


Figure 2.12: A data set plotted in three dimensions with the first principal component, P_1, shown. The principal component score of data point i is shown as t_{i1} and is the distance from the mean centre of the data set to the point of the orthoganal projection of i onto P_1. The cosine of angle u determines the influence (loading) of variable 2 and P_1 (modified from [170]).

where M is a principal components model and E is noise. The principal components model is composed of scores and loadings matrices so this can be further defined as:

$$X = TP^T + E \tag{2.3.2}$$

where T is a scores matrix, P is a loadings matrix and E is a matrix of the residuals (noise) not described by T and P. As each component contains its own scores, t, and loadings, p, vectors thus can be further described as:

$$X = t_1p_1^T + t_2p_2^T + ... + t_Ap_A^T + E \tag{2.3.3}$$

where A is the the total number of principal components calculated. This decomposition can be visualised in Figure 2.13 [170].

Thus a simple overview of how principal components are calculated is as follows:

- t_1 and p_1 are determined from X.

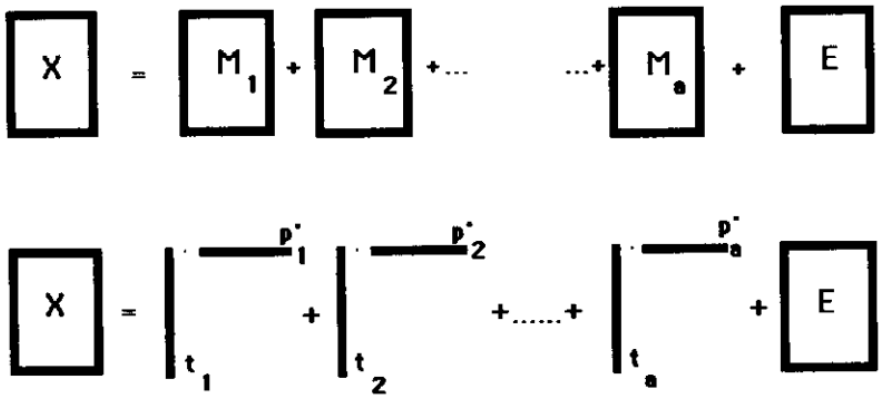


Figure 2.13: The matrix X can represented as a sum of matrices M_i and a matrix of residuals, E. M_i can be further decomposed into the outer products of the score and loading vectors, t_i, and $p_i^{'}$ respectively (from [170]).

- the first principal component is subtracted from X:

$$E_1 = X - t_1 p_1^T \tag{2.3.4}$$

- calculate t_2 and p_2 from E_1
- continue this process iteratively for all desired components

There are several different ways to calculate principal components, this review will focus on the singular value decomposition method and the approach taken by the NIPALS algorithm.

2.3.2.2 Singular value decomposition

The calculation of principal components can be transformed into an eigenvalue-eigenvector problem for a positive-semidefinite symmetric matrix [79].

For a matrix, A, and a vector x, if:

$$Ax = \lambda x \tag{2.3.5}$$

where λ is a real number, then x is an eigenvector of A. The scaling factor λ is known as the eigenvalue.

First, mean centring is done by subtracting the sample mean (mean of each row) from each row element of X. The singular value decomposition of X can be described as:

$$X = W\Sigma V^T \tag{2.3.6}$$

W is the matrix of *eigenvectors* of the covariance matrix XX^T.
Σ is a rectangular diagonal matrix with the *eigenvalues* of XX^T in the diagonal.

V is the matrix of *eigenvectors* of X^TX.

The PCA model that retains the original dimensionality of X is then given by:

$$\mathbf{Y}^T = \mathbf{X}^T\mathbf{W} \tag{2.3.7}$$
$$= \mathbf{V}\mathbf{\Sigma}^T\mathbf{W}^T\mathbf{W} \tag{2.3.8}$$
$$= \mathbf{V}\mathbf{\Sigma}^T \tag{2.3.9}$$

Each column of Y^T contains the scores of successive principal components. In order to reduce the dimensionality one can project the data onto a reduced component space using the L singular vectors of W_L.

$$Y = W_L^T X = \Sigma_L V^T \tag{2.3.10}$$

where

$$\Sigma_L = I_{L\times m}\Sigma \tag{2.3.11}$$

with $I_{L\times m}$ being a rectangular identity matrix.

2.3.2.3 NIPALS algorithm

The NIPALS algorithm for calculating principal component models can be described as follows [170].

1. Select the column in X with the largest variance as a proxy for t.
2. Create a loading vector, p^T
$$p^T = \frac{t^TX}{t^Tt} \tag{2.3.12}$$
3. Normalise p by multiplying by c where.
$$c = \frac{1}{\sqrt{p^Tp}} \tag{2.3.13}$$
4. Create a new score vector
$$t = \frac{Xp}{p^Tp} \tag{2.3.14}$$
5. Test for convergence between consecutive score vectors. If the sum of the squared differences of all elements of the score vectors is below a preset threshold then move on to step 6, else return to step 2. If convergence is not reached within a set number of components, end loop.
6. Create matrix of residuals, E, by subtracting the model for the principal component just determined from X.
$$E = X - tp^T \tag{2.3.15}$$
Use E as X in the next iteration.

2.3.2.4 NIPALS vs singular value decomposition

The NIPALS algorithm is computationally faster than the singular value decomposition method and is specifically designed to calculate the first set of principal components in cases where they will describe the vast majority of the variance in the data. For applications where the variance is spread out across the components and one does not necessarily want or need dimension reduction the singular value decomposition approach is more appropriate.

2.4 Gene Set Analysis

Gene Set Analysis (GSA) has proven to be a useful approach to microarray analysis. The underlying principle of GSA is that aggregate scores are assigned to each Gene Set based on all the individual gene scores within that set. There have been several different methods proposed to assign scores to gene sets [128, 127, 119, 63, 64, 94, 154, 162]. Of the approaches published to date, Gene Set Enrichment Analysis (GSEA) [119, 154] seems to have become the most commonly used. Of issue though is the fact that GSEA is based on a modified Kolmogorov-Smirnov test. This test can exhibit a lack of sensitivity; is difficult to employ in practical use, and requires at least 1000 permutations to be run. However, it has recently been found [77] that a one-sample Z-test can be very effective with gene sets for detecting shifts from the mean (sets that collectively show up or down regulation of their constituent genes). Unfortunately, this will not identify gene sets that have a balance of both up and down regulated genes as there will not be the requisite shift from the mean, but in statistical terms is rather a change in scale. However, a chi-squared test can be used to good effect to detect such changes in scale and thus find gene sets that exhibit a mixture of up and down regulation [77]. Furthermore, Irizarry *et al.* [77] have shown that the use of a combination of the computationally simple and rapid Z-test and chi-squared methods outperform GSEA. Dinu *et al.* [34] have extended the Significance Analysis of Microarrays to Gene Set Analysis (SAM-GS). Of further interest is the method described by Efron and Tibshirani [43] which uses a maxmean statistic to target gene sets with only a fraction of the genes differentially expressed and the approach of Falcon and Gentleman [48] which takes into account the fact that overlap exists between different gene sets. A good review of the various statistical approaches has been written by Goeman and Bühlmann [62].

Goeman and Bühlmann point out that most of the null hypotheses used by the statistical tests employed in GSA methods can be categorised as competitive or self contained. A competitive test compares the differential expression found in a gene set to a standard based on all of the other genes on the microarray. This is akin to the classic enrichment approach that many people will be familiar with in GO Enrichment (often using a Fisher Exact test). A

self-contained test, on the other hand compares the gene set in isolation (irrespective of the genes outside the gene set) to a fixed standard. Goeman and Bühlmann argue quite strongly that self-contained tests that use sample (vs. gene) permutations for p value calculation are the appropriate methods to use. [62].

Nam and Kim [123] take this argument a step further with the introduction of a third class of approaches which they term *mixed*. As the name implies two different approaches are applied, typically starting with sample permutations at the gene set level and followed by a data set wide sample permutation test of all of the gene sets. Nam and Kim state that their preference is for mixed approaches as they avoid the shortcomings of the other methods [123]. Of all of the methods evaluated, only those by Mootha/Subramanian [119, 154] *et al.* and Efron and Tibshirani [43] fall into this category. A table summarizing Nam and Kim's categorisation of different GSA methods can be seen in Table 2.6

Table 2.6: Summary of GSA Methods (modified from [123]).

Authors	Year	Name	Statistical test	Self-contained versus competitive	Gene versus ample randomization
Virtaneva *et al.*	2001		sample randomization	self-contained	sample
Pavlidis *et al.*	2002		gene randomization	competitive	gene
Mootha *et al.*	2003	GSEA	sample randomization	mixed	sample
Breslin *et al.*	2004	Catmap	gene randomization	competitive	gene
Goeman *et al.*	2004	globaltest	sample randomization	self-contained	sample
Smid *et al.*	2004	GO-Mapper	z-test	competitive	gene
Volinia *et al.*	2004	GOAL	gene randomization	competitive	gene
Barry *et al.*	2005	SAFE	sample randomization	competitive	sample
Beh-Shaul *et al.*	2005		Kolmogorov–Smirnov test	competitive	gene
Boorsma e al.	2005	T-profiler	*t*-test	competitive	gene
Kim *et al.*	2005	PAGE	z-test	competitive	gene
Lee *et al.*	2005	ErmineJ	sample randomization	competitive	gene
Subramanian *et al.*	2005	GSEA	sample randomization	mixed	gene
Tian *et al.*	2005	Q1, Q2	gene or sample randomization	competitive or self-contained	gene or sample
Tomfohr *et al.*	2005	PLAGE	sample randomization	self-contained	sample
Edelman *et al.*	2006	ASSESS	sample randomization	competitive	sample
Kong *et al.*	2006		Hotelling's T squared	self-contained	sample
Nam *et al.*	2006	ADGO	z-test	competitive	gene
Saxena *et al.*	2006	AE	sample randomization	competitive	sample
Scheer *et al.*	2006	JProGO	Fisher's exact test, Kolmogorov–Smirnov test, *t*-test, unpaired Wilcoxon's test	competitive	gene
Al-Shahrour *et al.*	2007	Fatiscan	Fisher's exact test, hypergeometric test	competitive	gene
Backes *et al.*	2007	GeneTrail	Fisher's exact test, hypergeometric test, sample randomization	competitive	gene or sample
Cavalieri *et al.*	2007	Eu.Gene Analyzer	Fisher's exact test, sample randomization	competitive	gene or sample
Dinu *et al.*	2007	SAM-GS	sample randomization	self-contained	sample
Efron *et al.*	2007	GSA	sample randomization	mixed	sample
Newton *et al.*	2007	Random set	z-test	competitive	gene

The method of Efron and Tibshirani is an extension of GSEA [119, 154] in which they explored different test statistics for the GSEA algorithm. They argue quite strongly that for p value generation any gene-set-based approach should not only compare set scores to those from permutations of the sample labels, but should also compare them to scores from sets generated by random gene selection. Using this approach they then tested five different test statistics, namely mean, mean.abs, maxmean, GSEA and GSEA.abs, in order to determine which would be the most appropriate one to test five different (simulated) scenarios. They proposed that the maxmean statistic is the only one which generated low p values in all five scenarios [43].

2.4.1 Maxmean Method

The maxmean method algorithm can be described briefly as follows [43]:

1. Calculate a summary statistic, z_i, for each gene, by calculating a two sample t-statistic for two-class data and then converting it into a z value.

$$z_i = \Phi^{-1}(F_{n-2}(t_i)) \tag{2.4.1}$$

where Φ is the standard normal cumulative distribution function (cdf) while F_{n-2} is the cdf for a t distribution having n-2 degrees of freedom.

2. For each gene-set S, choose a summary statistic S = s(z) using the maxmean statistic as follows:

First define a two-dimensional scoring function $s(z_i)$:

$$s(z_i) = (s^{(+)}(z_i), s^{(-)}(z_i)), \begin{cases} s^{(+)}(z_i) & = max(z_i, 0) \\ s^{(-)}(z_i) & = -min(z_i, 0) \end{cases} \tag{2.4.2}$$

The maxmean test statistic is then defined as:

$$S_{max} = max\{\bar{s}_S^{(+)}, \bar{s}_S^{(-)}\} \tag{2.4.3}$$

3. Standardize S by its randomisation mean and standard deviation

$$S^{'} = (S - mean_s)/stdev_s \tag{2.4.4}$$

For maxmean, this can be computed from the genewise means and standard deviations, without having to draw random sets of genes.

$$S^{**}_{max} = max\{\bar{s}_S^{(+)**}, \bar{s}_S^{(-)**}\} \tag{2.4.5}$$

4. Compute permutations of the class labels and recompute $S^{'}$ on each permuted dataset, yielding permutation values $S^{'1}, S^{'1}, ...S^{'B}$.. These permutation values are used to estimate p values for each gene-set score $S^{'}$, and false discovery rates applied to these p values for the collection of gene-set scores.

2.5 Orthology Detection

Microarray annotation is crucial to the analysis of the data captured from a gene expression experiment. The more annotation available for each probe on a microarray the more biological interpretation of the experimental results will be possible. Chapter 5 required the use of genome scale annotation of two EST data sets, namely tobacco and potato, in order to properly annotate the microarray that was used. The strategy taken for genome scale annotation was one of inference by orthology. Genes in different organisms carry differing levels of functional annotation due to the different experiments present in the literature that have yielded extant knowledge about gene function in different organisms. The assumption commonly made in genome annotation is that the basic function of a gene in one organism will be similar to its ortholog in a different organism. By defining orthologous families one can then infer annotation from multiple experimental histories and thus maximise the annotation captured for the organism being annotated. This approach is not without it's pitfalls as orthology is an evolutionary relationship rather than guaranteed to be a functional relationship. That said, the community consensus has been that lots of inference with some inaccuracy is better than no inference (and thus unannotated genomes). This section will serve as a brief review of the orthology detection algorithms available at present. Only methods that can compare multiple genomes simultaneously will be considered.

A few definitions discussed by Kuzniar *et al.* [103] will be helpful for the following section. **Homology** is the hypothesis that genes found in different species and which have a reasonable degree of sequence similarity have descended from a single gene in an ancestor of both species. **Paralogs** are genes that have come from a duplication event of a single sequence. **In-paralogs** are paralogs that are derived from a duplication event within a specific species. **Out-paralogs** are genes that are derived from a duplication event that occurred prior to speciation. **Orthologs** are genes that have resulted from a homologous sequence derived by a speciation event from a single sequence in a latest common ancestor.

The network-based methods that can handle multiple genomes simultaneously include COGS [158], MultiParanoid [3] and OrthoMCL [109]. Phylogeny-based methods for ortholog detection include COCCO-CL [81], HOPS, LOFT [164], RAP [39], RIO [183], PhIGS [32], PhOG [116], PhyOP [67] and TreeFam [108]. Phylogeny-based methods, with their requirements for multiple sequence alignment and tree construction, don't scale well and as such are not suited for complete genomes [103]. Thus, taking into consideration that the central theme of this thesis is the use of networks in data analysis, only network-based ortholog detection algorithms will be discussed here.

2.5.1 Network-Based Ortholog Detection Methods

2.5.1.1 Clusters of Orthologous Groups (COGs) of proteins

This algorithm uses best BLAST hits (BeTs) across proteomes and requires there to be congruent network triangles of BeTs in order to define protein families. Briefly, the algorithm finds all triangles formed by BeTs, combining those that share a side of a triangle until no new ones can be added [158]. Figure 2.14 show examples of simple to more complex COGs found by this method. COGs were manually curated and functionally annotated and, as such, were commonly used ten years ago. Unfortunately, the COGs database has not been updated since 2003 making them of limited utility now.

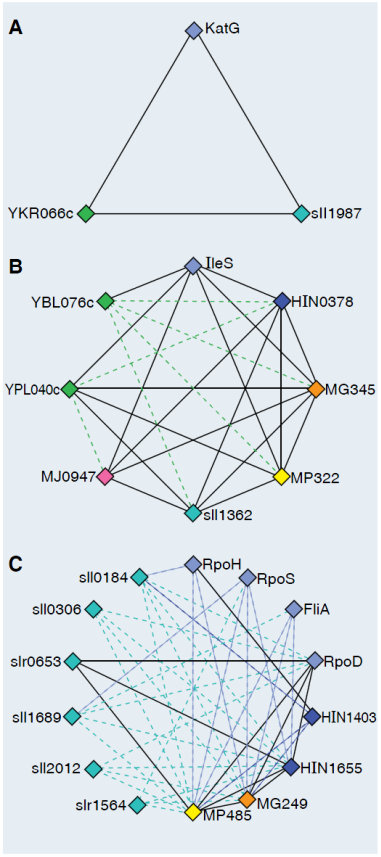


Figure 2.14: Examples of COGs. "Solid lines show symmetrical BeTs. Broken lines show asymmetrical BeTs, with color corresponding to the species for which the BeT is observed" (from [158]).

2.5.1.2 MultiParanoid

This approach [3] uses the InParanoid algorithm [24] to compare species pairwise and then constructs orthologous groups with a single linkage approach.

A major drawback of this method is that it can only be used for a few species which are roughly evolutionarily equidistant from a common ancestor. Briefly, the pariwise algorithm, InParanoid, uses a series of rules to merge or seperate orthologous groups. The best reciprocal BLAST hits are considered to be orthologs and used as starting points for the detection of in-paralogs in both species.

2.5.1.3 OrthoMCL

The OrthoMCL [109] algorithm integrates BLASTP, the normalisation of the relationships between inparalogs, co-orthologs and orthologs, the creation of a similarity network and it's partitioning based on Markov Clustering. This process is illustrated in Figure 2.15 and each step will be discussed in greater detail below.

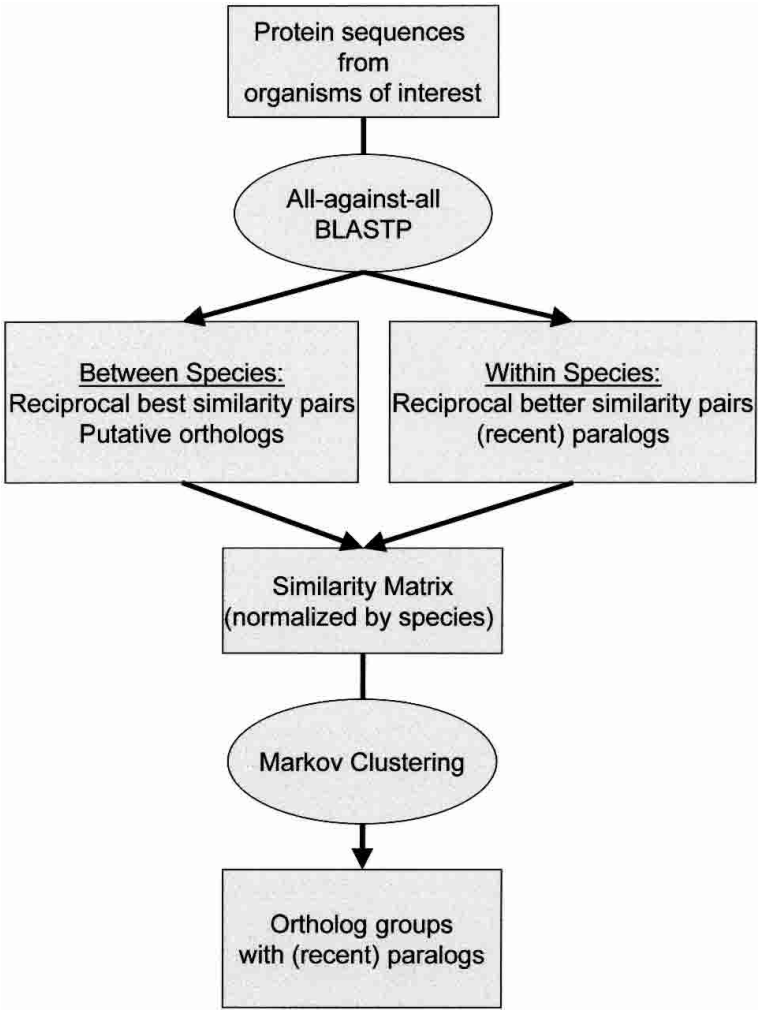


Figure 2.15: Flow chart for the OrthoMCL algorithm(from [109]).

All-against-all BLASTP

All protein sequences from all the organisms of interest are compared to one another with an all-against-all BLASTP query. This is a computationally intensive step that can result in billions of results. Care must be taken to make sure that E-value thresholds set for the BLASTP query are not larger than $1x10^{-5}$, and that the number of results per query is set very high (>10,000) in order not to inadvertently exclude proteins pairs that should be considered by the OrthoMCL algorithm. This step requires the use of a compute cluster and some mechanism with which to parallelize the BLASTP queries. The percent matches for each each pair are calculated and only pairs with E-value less than $1x10^{-5}$ and a percent match greater than 40% (a preset threshold which can be changed) are retained.

Paritioning into ortholog, co-ortholog and inparalog pairs

The next step splits the resulting BLASTP pairs into different groups, namely ortholog, co-ortholog and inparalog pairs. Orthologs are algorithmically defined as matches of proteins across species that have the best reciprocal E-values with each other than with proteins from other species. Inparalogs on the other hand are defined as matches of proteins within the same species that have the best reciprocal E-values with each other than either sequence has to any other protein from another species. Finally, co-orthologs are defined as matches of proteins from the same species which are linked by orthology and inparalogy [109]. These relationships can be visualised in Figure 2.16.

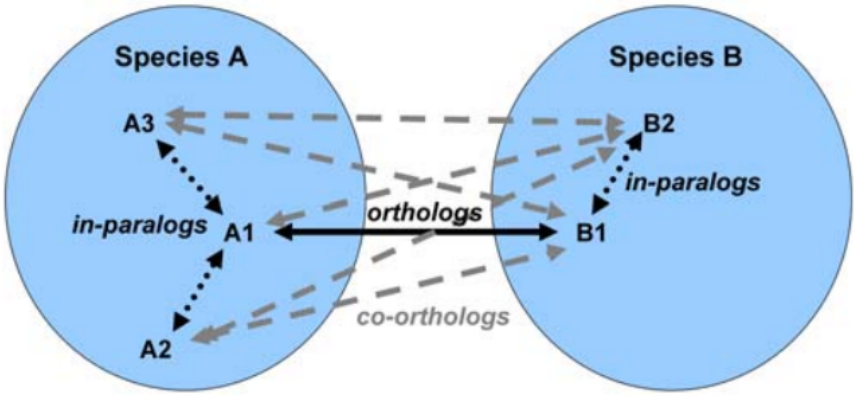


Figure 2.16: Relationships of orthologs, inparalogs and co-orthologs(from [24]).

Normalisation

The resulting pairs of orthologs, co-orthologs and inparalogs are each given a weight defined by the following equation [109]:

$$Weight = -\log_{10}(\text{E-value}_{av}) \tag{2.5.1}$$

E-value_{av} is the average of the reciprocal E-values. These weights are then normalised based on the category that their protein pair has been assigned to.

Orthologs are normalised by species pairs as follows: first the average weight of all ortholog pairs between two species is calculated.

$$\text{Average weight} = \frac{1}{N}\sum_{o=1}^{N} w_o \tag{2.5.2}$$

where N represents the number of ortholog pairs between those two species and w_o is the weight of each pair. The *normalised* weight n_o assigned to each ortholog pair o is then created by dividing the ortholog pair weight by the Average weight [109]:

$$n_o = \frac{w_o}{\text{Average weight}} \tag{2.5.3}$$

n_c, the normalised weight of co-orthologs is determined in the same way as orthologs but N in this case is the number of co-ortholog pairs rather than the number of ortholog pairs between two species.

The normalisation of inparalogs is treated quite differently and is dependent on a set of weights. A *normalising set of weights* can be found via the following procedure. If some of the inparalog pairs in a species contain a member which is also part of an ortholog pair, *the normalising set of weights* are defined as the weights of inparalog pairs that do contain at least one member of an ortholog pair. If none of the inparalog pairs contain any members that are part of an orthlog pair then the *normalising set of weights* is simply those of the set of all inparalog pairs in that species. This can be described mathematically as follows:

$$n_i = \frac{w_i}{average(N_s)} \tag{2.5.4}$$

where N_s is a *normalising set of weights* for a given species [109].

Normalisation is necessary because inparalogs have been derived from a recent duplication event in that species. As such, they are very likely to be more similar to each other than orthologs and co-orthologs are. If no normalisation is performed the clustering will be skewed towards the highly similar inparalogs and orthologs and co-orthologs will tend to be squeezed out of the clusters [109].

Once all of the normalised weights for all of the ortholog, co-ortholog and inparalog pairs have been assigned, each pair can simply be considered to be an edge in a undirected similarity network with proteins acting as nodes and the normalised weights assigned as edge weights. To define groups of orthologous proteins the similarity graph can be partitioned by Markov Clustering with the MCL algorithm.

Markov Clustering

The Markov Clustering (MCL) algorithm partitions a network in such a way that dense regions are kept together and the sparse connections between them are pruned away. This is achieved as follows: The weighted adjacency matrix of a network is first converted into a stochastic matrix. This is achieved by dividing each element of each column by the sum of that column. The result is a Markov matrix in which the probability of a random walk from a node going to another node is represented by each element of the originating node's column. Thus the ij element of the matrix is the probability of travelling from node j to node i. Adjacency matrices are square by definition so the resulting Markov matrix is square and stochastic.

Clustering is performed by alternating expansion and inflation of the matrix. Expansion is simply achieved by taking the normal matrix product of the matrix itself:

$$E(A) = A \times A \tag{2.5.5}$$

where E is the expansion operator, A is the Markov matrix and "$\times$" is the normal matrix product [166, 165]. Inflation is achieved by taking the rth Hadamard (Γ_r) power of the matrix and subsequently renormalising each of the new columns so that each element can be described as:

$$\Gamma_r(A_{ij}) = \frac{(A_{ij})^r}{\sum_{r,j}(A)} \tag{2.5.6}$$

where r is the inflation index [167, 166].

For r $>$ 1, inflation changes the probabilities of random walks that start from a node such that more probable walks (higher element values) are favoured over less probable walks (lower element values).

Expansion simulates longer random walks. It changes the probabilities associated with all node pairs. Longer paths are more common within clusters than between clusters, so the probability values assigned to edges of node pairs that are in the same cluster will likely be comparatively high as, due to the higher density of paths within a cluster, there are several paths going from one node to another node within that cluster. Inflation will increase the probabilities of walks within clusters and will decrease the likelihood of walks between clusters.

Iterating between expansion and inflation will partition the network into different components that lack paths between each other but are highly connected within themselves. Clustering is complete when the further expansion and inflation has no affect on the elements in the matrix. i.e. it is *doubly idempotent* [167, 166].

This process can be visualised in Figure 2.17.

Increasing the inflation parameter, r, will increase the granularity of the clusters. If desired a type of hierarchical clustering is possible by running the algorithm repeatedly with increasing inflation values each time.

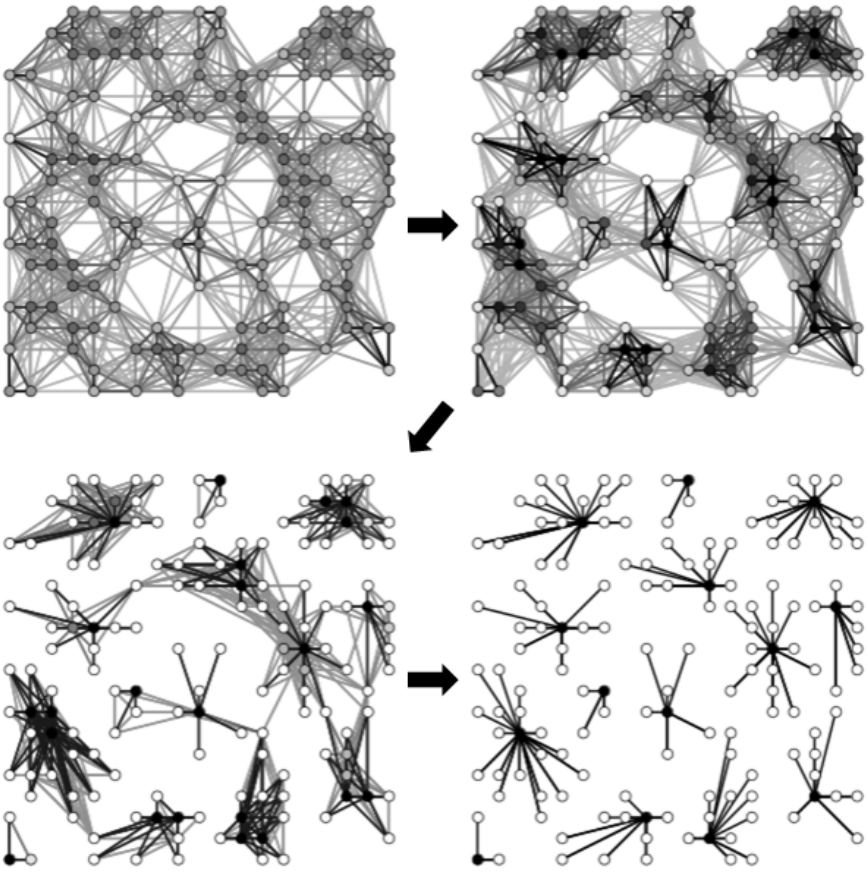


Figure 2.17: Iterative rounds of expansion and inflation generate clusters from a complex network(from [167, 166]).

The use of MCL on the normalised similarity matrix results in clusters of orthologs, co-orthologs and inparalogs for protein families in and across species.

2.5.2 Comparison of Ortholog Detection Methods

In order to obtain an objective assessment of the performance of network-based and phylogeny-based ortholog detection algorithms Chen *et al.* [24] constructed a set of 27,562 protein sequences that could be used by all methods and were common to COG/KOG as it had been manually curated. These sequences were from six eukaryotic genomes (*Arabidopsis thaliana, Caenorhabditis elegans, Drosophila melanogaster, Homo sapiens, Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe*). It should be noted that this is a relatively small scale data set as the number of proteins encoded in many plant genomes are in excess of 30,000 per genome.

They used latent class analysis (LCA) as a multivariate technique to assess algorithm performance. With the use of LCA they were able to calculate the

sensitivity (false negatives)and selectivity (false positives) for a number of ortholog detection algorithms including RIO (Resampled Inference of Orthology) [183], Orthostrapper/HOPS (Hierarchical grouping of Orthologous and Paralogous Sequences) [152], RSD (Reciprocal Smallest Distance), Reciprocal Best Hit (RBH) [24], Single-way Best Hit (SBH) [24], COG (Cluster of Orthologous Groups)/KOG (euKaryotic Orthologous Groups) [157], InParanoid [138], OrthoMCL [109], TribeMCL [45] and BLASTP. The methods compared include representatives of both phylogeny-based and network-based approaches.

As expected there is considerable variation between the selectivity and sensitivity of ortholog detection algorithms. The phylogeny-based methods tend to have a low level of false positives and a higher level of false negatives. The extreme example is RIO which has a 1% false positive rate but a 64% false negative rate. In contrast the network-based methods tend to have much lower false negative rates and higher false positive rates, with the extreme example being TribeMCL with a 56% false positive rate and a 5% false negative rate. The most desirable status would be low false negative and low false positive rates. InParanoid and OrthoMCL appear to have the best trade-offs in this regard as can be seen in Figure 2.18. However, InParanoid can only analyse two species at a time and its multi-species implementation, MultiParanoid, suffers from "tree conflicts" with more divergent species [3]. As such, OrthoMCL does seem to be the best approach for ortholog detection across a large, divergent set of species.

It is also notable that OrthoMCL, which is a totally automated algorithm outperforms COG/KOG, even though the latter has been manually curated. A specific example of this can be seen in Figure 2.19 which demonstrates how OrthoMCL splits a KOG group into two groups which appear to be more appropriate groupings based on EC annotation and domain architecture [24].

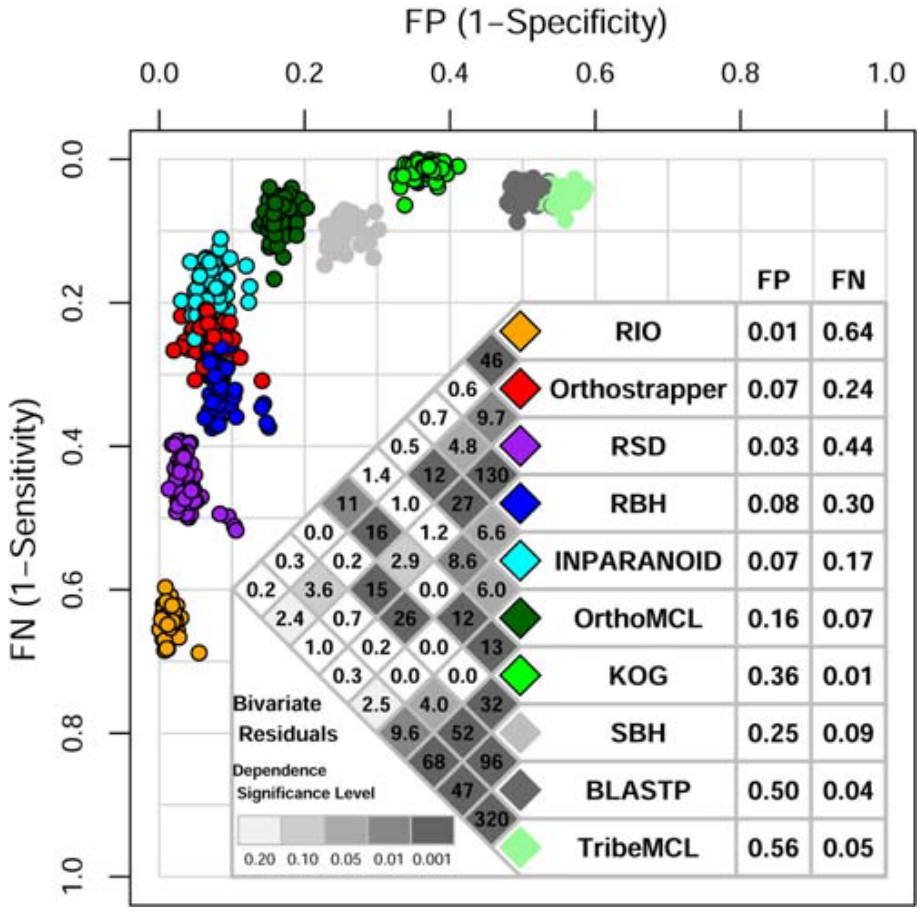


Figure 2.18: Comparision of sensitivity (false negatives, FN) and selectivity (false positives, FP) for a number of ortholog detection algorithms (from [24]).

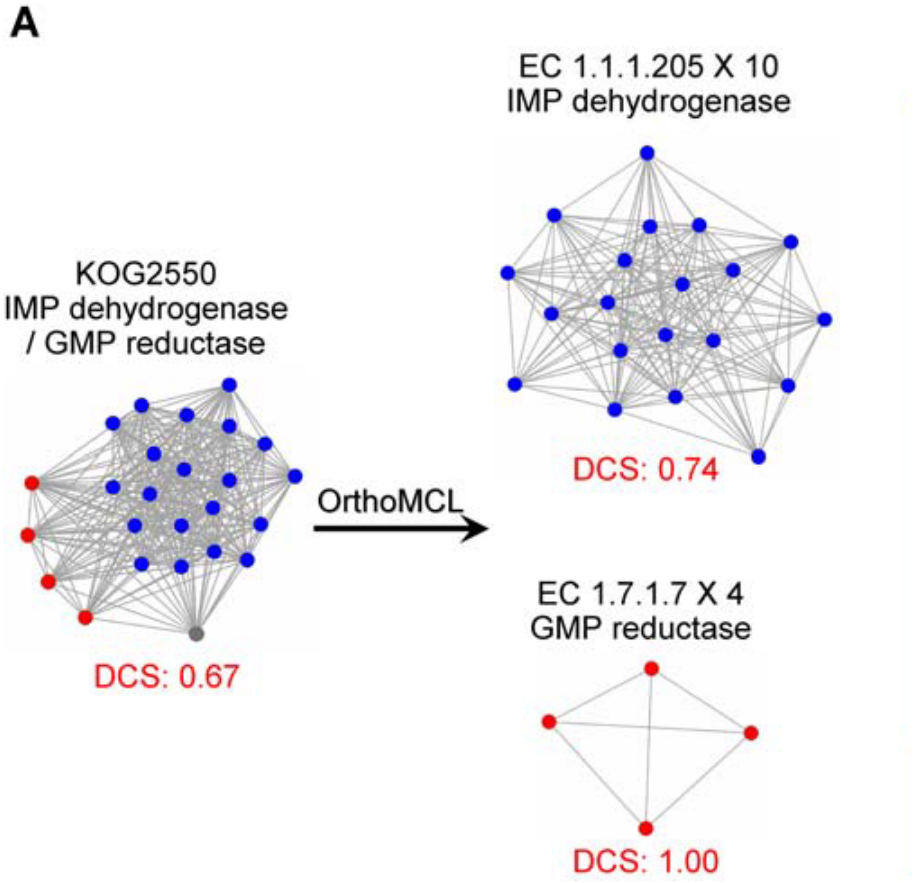


Figure 2.19: OrthoMCL splits a KOG group into two groups which appear to be more appropriate groupings based on EC annotation and domain architecture. (modified from [24]).

2.6 Chromatographic Preprocessing: Alignment, Refinement and Peak Selection

2.6.1 Context

Metabolomics is defined as the study of as many-small-metabolites-as-possible in a system [20]. Chapter 3 describes an example of chemiomics which is defined to be the study of the relationships between as many chemical compounds as possible in a complex chemical (non-enzymatic) system with the use of chromatography.

In order to accomplish this with chromatographic profiling two strategies can be employed: (i) targeted analysis, using *a priori* knowledge of which compounds to analyse, which requires their identification and manual quantification or (ii) untargeted analysis where one tries to detect as many compounds as possible in order to acquire sample fingerprints which will be submitted to multivariate analyses and network analysis for further classification. Compound identification and quantification is then performed on the variables that are found to be responsible for the classification of the original samples. Chapter 3 addresses the potential use of a univariate signal for a untargeted approach in order to (i) classify samples according to a given process or perturbation, (ii) evaluate the feasibility of developing a screening procedure to select candidates related to the process and (iii) provide insight into the chemical mechanisms which are affected by the perturbation. In order to achieve this it was necessary to use and develop methods for data pre-processing and visualisation tools to assist an analytical chemist to view and interpret complex multidimensional data sets.

Chromatography has long been used for the separation of molecules enabling both the quantitative and qualitative analysis of constituents in complex mixtures. Targeted chromatography has been used extensively in the analysis of the chemical composition of wine, particularly with regard to finding compounds that contribute to flavour and aroma properties [133, 114, 40, 105, 51, 52, 47, 120].

The separation of the different components of a complex mixture generates a fingerprint representing the chemical composition of the sample. Chromatographic fingerprints taken from samples under different experimental conditions can then be used to explain the changes caused by a perturbation [66]. Due to the separation performed by the column, it is possible to identify structures based on the elution time in a given chromatographic profile.

Chromatographic data has a huge number of variables and PCA and Partial Least Squares (PLS) are multivariate analysis (MVA) visualisation techniques that allow for the interpretation of multidimensional data sets. When multivariate analysis involves large datasets, variable selection processes play an important role because they eliminate the less significant or non-informative

variables. The overall aim of any variable selection technique is to capture variables from the original dataset that are most specifically related to the problem of interest and to exclude those variables that are affected by other sources of variation. As described in section 2.3, PCA is a non-supervised technique that decomposes the original variables of a data set into two matrices: the score and the loading matrices. The scores matrix contains information about the samples, which are described in terms of their projection onto the principal components. The loading matrix contains information about the variables which are also described in terms of their projection onto the principal components. Consequently, the use of GC fingerprints with MVA should make it possible to extract considerable amounts of information from complex mixtures. The tandem of GC-MVA can be viewed as a middle ground between a rich detector, which provides structural information (NMR), and detectors like FTIR and UV-Vis. Furthermore, as discussed in Chapter 3, network analysis can be used to try to understand the underlying kinetic systems at play that leads to the consumption and formation of compounds in a complex mixture.

However, in order to perform either MVA or network analysis, the signals have to be of the same length and the corresponding variables must occur in the same column of the matrix to be analysed. Unfortunately, chromatographic data does not meet these criteria [131]. The conditions between each chromatographic analysis vary subtly and can include slight differences in temperature, pressure, mobile phase composition, sample composition (the so called "matrix affect"), and column condition to name a few. All of these differences can affect the migration of compounds through a column and therefore their elution time. Thus, chromatography is inherently irreproducible as no two samples are run under exactly the same conditions. As such, chromatographic data requires the use of preprocessing algorithms with which to prepare the raw data for down stream analysis. The two most important pre-processing steps used in Chapter 3 involve chromatographic alignment with time domain warping and peak selection and peak alignment refinement with the use of wavelets. As such, this portion of the literature review will focus on Warping-based alignment methods and Wavelets.

2.6.2 Alignment: Warping Methods

Given that there is time domain variability between chromatograms and given that the variability is not a constant throughout the chromatogram there is the need to find local solutions to account for the variability and to then subsequently reconstruct the global alignment. As such, most of the chromatographic alignment methods use some form of a dynamic programming [11] approach to solve the problem as it is particularly well suited for local to global solutions. The primary methods used in the literature for chromatogram alignment include dynamic time warping (DTW), correlation opti-

mised warping (COW), parametric time warping (PTW) and semi-parametric time warping (STW).

2.6.2.1 Dynamic Time Warping

DTW was previously used in a variety of contexts, including speech processing [31, 121, 122, 136, 137], bird song analysis [5, 61, 95, 104], process monitoring for an industrial emulsion polymerisation process [92] and neurophysiology [26]. The first proposal for the use of DTW as applied to chromatograms came in 1987 [169].

The application of DTW to the alignment of two chromatograms (R and T) can be described as follows: First consider each chromatogram as a time indexed collection of intensity values. The lengths (number of time points) of R and T are denoted as L_R and L_T respectively. R_j is the intensity value of chromatogram R at time index j and T_i is the intensity value of chromatogram T at time index i. A distance matrix is created which contains the Euclidean distance between all time points (j = 1,...,L_R) in chromatogram R versus all of the time points (i = 1,...,L_T) in chromatogram T. The Euclidean distance in this instance is simply defined as the square root of the square of the difference between R_j and T_i [131]:

$$d(i,j) = \sqrt{(T_i - R_j)^2} \tag{2.6.1}$$

The goal now is to define the minimal cumulative distance through this matrix, whilst keeping reasonably close to the upwards diagonal as can be seen in Figure 2.20.

In order to limit the search space and to avoid aligning vastly different areas of the chromatograms (in most cases it would not make sense to align the beginning of one chromatogram to the end of the other) three major constraints are used, namely allowable predecessors, diagonal zone and a slope.

Allowable Predecessors Constraint

In order to limit the directionality and locality of the search space, the local path that can be followed when calculating the minimal cumulative difference is constrained by only allowing three neighbouring predecessors to be considered for every (i, j) pair in the matrix, namely (i-1, j), (i-1, j-1) and (i, j-1). This can be visualised as a directed graph as shown in Figure 2.21.

M - Diagonal Zone Constraint

In order to limit the overall locality of the search space the global path that can be followed when calculating the minimal cumulative difference is constrained by limiting the matrix to those cells that are within some defined constant M of the upward diagonal of the matrix as illustrated in Figure 2.22.

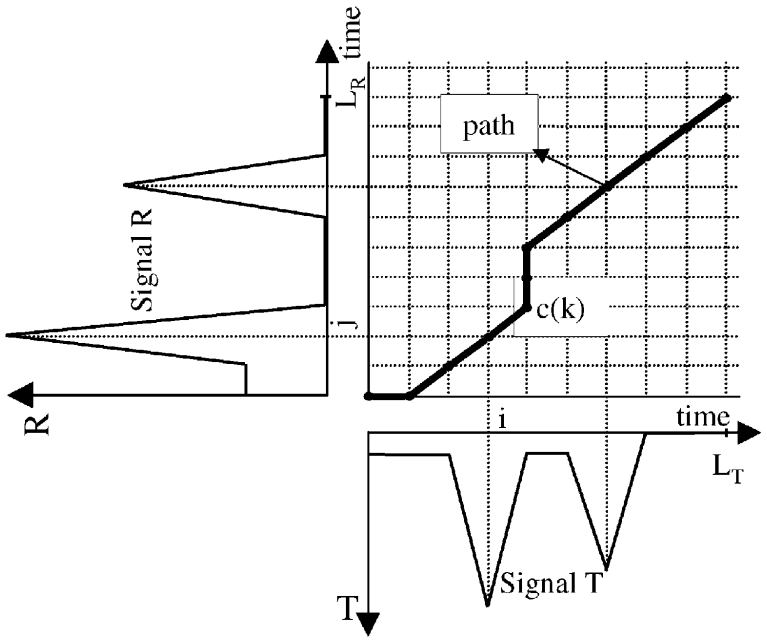


Figure 2.20: Conceptual Overview of Dynamic Time Warping (from [131]).

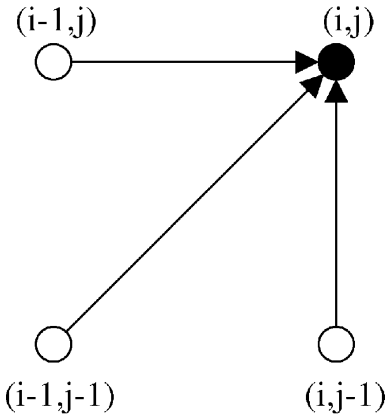


Figure 2.21: Allowable predecessors in Dynamic Time Warping (from [131]).

Non-negative Slope Constraint

The path through the matrix shown in Figure 2.20 can be defined as a sequence F of K points defined as [131]:

F = {c(1),c(2),...,c(k),...,c(K)}

where

c(k) = [i(k),j(k)]

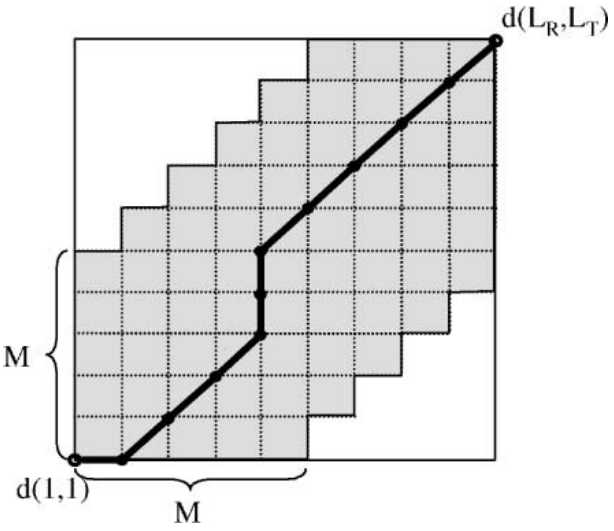


Figure 2.22: Diaganol zone used to constrain the global search space (from [131]).

As a further constraint to enforce locality in the search space the path can be forced to be monotonous and to contain only non-negative slopes by the use of the following criteria [131]:

$$i(k+1) \geq i(k) \tag{2.6.2}$$

and

$$j(k+1) \geq j(k) \tag{2.6.3}$$

Pragmatically this means that the path depicted in Figure 2.20 can only travel horizontally, vertically or in an upward diagonal from left to right, never from right to left or downwards. This prevents the discovery of local minima in the search path that are too far apart (and therefore would introduce discontinuity) in the two chromatograms to be attributable to the same compounds.

Minimal Cumulative Distance Path Construction

With the use of the distance matrix and these three constraints, the minimal cumulative distance path can be constructed as follows. First a new matrix is constructed by calculating the three local cumulative distances as derived from the three allowable predecessors for each point (i, j). The lowest value of these three cumulative distances for (i, j) is then placed in the new matrix. The step can be represented mathematically as follows :

$$D(i,j) = min \begin{cases} D(i-1,j) + d(i,j) \\ D(i-1,j-1) + d(i,j) \\ D(i,j-1) + d(i,j) \end{cases} \tag{2.6.4}$$

where $D(1,1) = d(1,1)$, $D(i,j)$ represents the localised cumulative distance point at that position in the new matrix and $d(i,j)$ is the local distance from the original distance matrix [131]. The diagonal zone constraint is used to only calculate the cumulative distance points in the allowable search zone. The starting point for path construction is at the end points of the two chromatograms (L_R, L_T). As shown in Figure 2.23, in a process following the allowable predecessors and non-negative slope constraints, the three allowable neighbours are examined and the one with the lowest value selected as the next point. This process is continued iteratively until the starting point $(1,1)$ is reached.

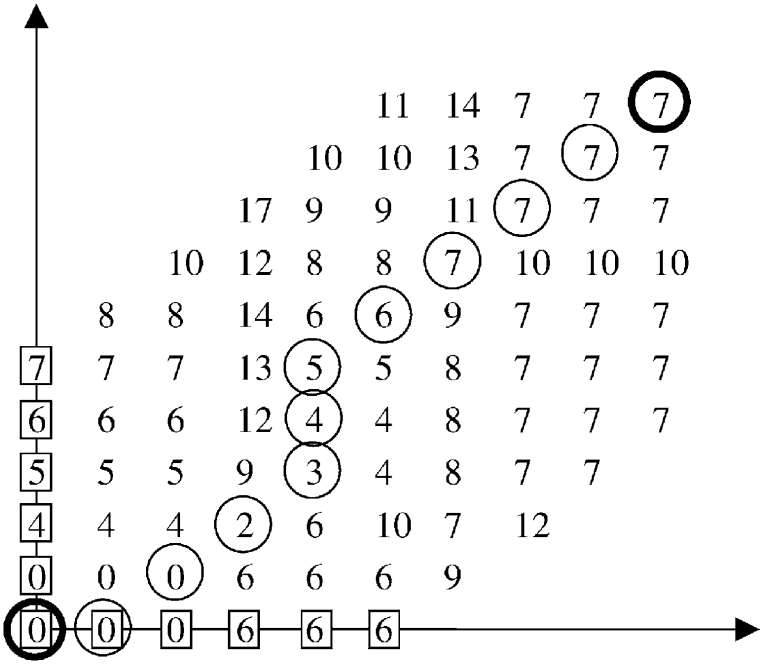


Figure 2.23: Minimal Path Construction (from [131]).

Synchronisation

Once the minimal cumulative path has been determined, synchronisation is the final step required to align the chromatograms. Alignment can be done symmetrically or asymmetrically. A symmetric alignment will result in the warping of both chromatograms toward each other while an asymmetric alignment will take one of the chromatograms as the reference and the other will be warped to it. For a symmetric alignment all of the points in the common path are used to obtain the warped chromatograms. If the minimal cumulative path contains a horizontal transition then the R_j value is taken twice.

For an asymmetric alignment, if T is declared to be the reference, the vertical transitions are treated differently. A vertical transition occurs when

more than one of the points in R are aligned with the same point in T, in this case, the average of those points is determined and aligned with the corresponding point in T. At the end of this procedure the length of the warped chromatogram is the same as the length of the reference chromatogram [131].

2.6.2.2 Correlation Optimised Warping

Correlation Optimised Warping (COW) is a specific instance of the more general DTW approach described above [163] that was first proposed as a method for chromatographic alignment by Nielson *et al.* [124]. In COW all operations are assumed to be asymmetric, that is that there is a target chromatogram, T, to which the other chromatogram, R, is being aligned via warping. However, in contrast to DTW a distance metric is not used, rather a correlation coefficient derived from segment-based warping is used as follows.

Both chromatograms are divided into N sub-segments of length m. If the chromatograms are of unequal length there is an adjusted factor Δ to account for the different initial segment lengths m of the two chromatograms.

$$\Delta = (L_T/N) - m \tag{2.6.5}$$

A slack parameter t is declared which defines how many warpings will be considered for each segment. The sequences of warpings, denoted as u, to be considered are contained in the set defined by $(\Delta - t; \Delta + t)$. If the range of warpings to be considered is defined by (-3; 3), then there are seven possible end points for the segment that is being warped, namely (-3, -2, -1, 0, 1, 2, 3). The relationship between Δ, m and t can be seen in Figure 2.24.

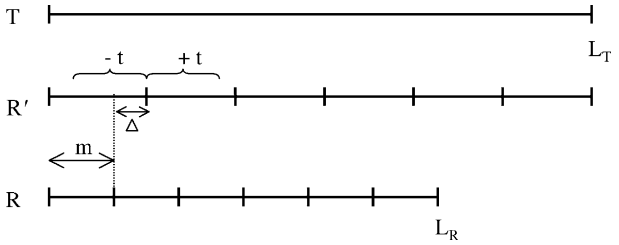


Figure 2.24: Relationship between Δ, m and t (from [131]).

A matrix F is then generated with size (N+1, L_T+1) and all initial elements of the matrix set to equal $-\infty$. Element f(N+1, L_T+1) is then set to zero indicating that the last point of the two are fixed together as aligned.

A series of new vectors are created from each segment of R as follows. For each warping end point the time index of the end point is adjusted and the vector values recalculated by linear interpolation. This yields a vector with the same number of elements as the target vector but which has been compressed (if the warping index, u, was < 0) or stretched (if the warping index, u, was > 0), or unchanged (if the warping index, u, was $= 0$). Each of these warped

vectors is then compared to the corresponding target vector and the degree of similarity captured by a correlation coefficient, cc:

$$cc = \frac{(w_T - \bar{w}_T)^T (w_{R'} - \bar{w}_{R'})}{std(w_T) std(w_{R'})} \tag{2.6.6}$$

where std(w) is the standard deviation of the vector and $\bar{w}$ is the mean of the vector w [131].

The correlation coefficient is added to the benefit function from the previous ending point. Analogous to the procedure discussed for DTW, only the highest value of this cumulative correlation coefficient is actually kept in $\boldsymbol{F}$. Now a different matrix $\boldsymbol{U}$ is created with the same size of $\boldsymbol{F}$. The warping index, u that corresponds to highest cumulative correlation coefficient as deposited in $\boldsymbol{F}$ is placed in the same position in $\boldsymbol{U}$. The algorithm starts in the next to last row of $\boldsymbol{F}$ and continues iteratively up to the first row of $\boldsymbol{F}$. The final construction of the warped signal is done by simply following the path defined in $\boldsymbol{U}$ and using the warping indices to select the best interpolated vector. This process can be visualised in Figure 2.25.

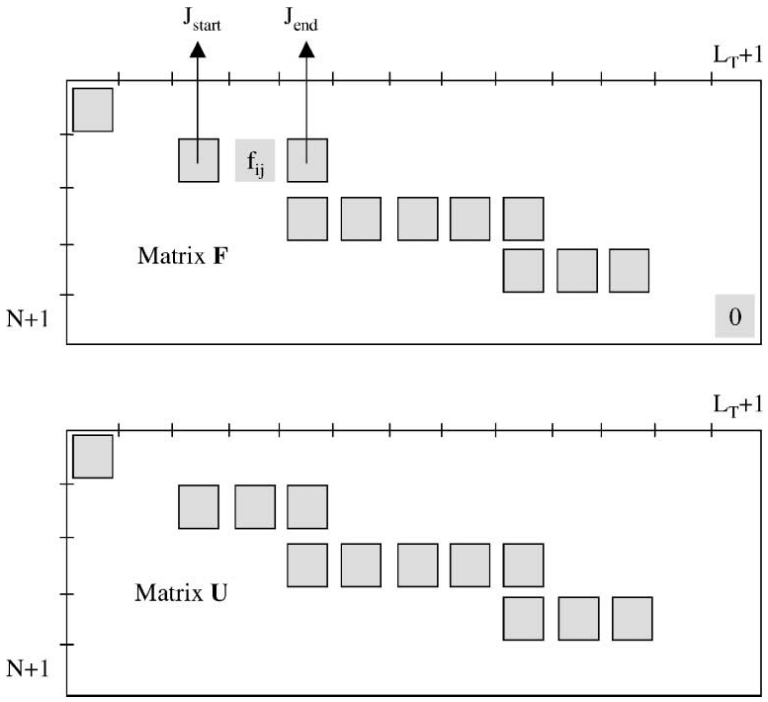


Figure 2.25: $\boldsymbol{F}$ and $\boldsymbol{U}$ matrices, $m = 3$, $L_T = 12$ and $t = 1$ (from [131]).

It is notable that both DTW and COW are quite akin to the shortest path problem in networks where one wants to find the shortest route through a complex network. One could view the distance and/or minimal cumulative distance matrix discussed in the DTW section above as a weighted adjacency matrix of a network with the values representing edge weights. With a bit

of matrix pruning with M and one additional constraint one could use Dijkstra's algorithm [33] (which requires global information of the network) or the Bellman-Ford algorithm [12, 55], which only requires local information (neighbouring nodes), to find the shortest path.

2.6.2.3 Parametric and Semi-parametric Time Warping

Parametric Time Warping (PTW) attempts a global solution to the alignment of two chromatograms by the use of a warping function that utilizes a second degree polynomial. Automated coefficient optimisation is performed in order to yield the minimum of the squared residuals between the two chromatograms. PTW is strongly affected by baseline variation and noise so smoothing and baseline correction are required pre-processing steps [44]. Although fast, PTW has been shown to be very poor at properly aligning moderately complex chromatograms [168]. This is likely due to the fact that PTW is attempting a global solution but the shifts happening between chromatograms are not consistent across the signal but are rather local in nature with shifts in different directions happening in different regions of the chromatogram.

Semi-parametric Time Warping (STW) is a variant of PTW where a series of B-splines constructed from polynomials is used as the warping function [168]. As such, STW is better able to cope with the localised variation encountered across chromatograms.

2.6.2.4 Comparison of Dynamic Time Warping, Correlation Optimised Warping, Parametric Time Warping and Semi-parametric Time Warping

DTW has been shown to be very sensitive to peak size [131] due to the fact that it is based on a distance metric. If the concentration of the compounds is quite different between two chromatograms, which in an experimental setting is often a desirable trait, then the distance metric of peaks that are already aligned will still be large as the peak heights will be quite different.

COW addresses this by using a correlation coefficient to decide on the appropriate warpings between subsegments of the two chromatograms. Segment size depends on N, the number of segments into which the chromatograms were divided. However, if the peak shifts are significant and the segments aren't large enough then peak alignment won't occur as the peaks will be on different segment indices across the two chromatograms. Furthermore, the slack parameter, t, defines how much stretching and compression will occur. The number of segments and the amount of warping have direct impacts on computational load as the higher value for each of them leads to more computations that must occur. N and t, also affect each other as the higher N, the shorter the segments and therefore the higher proportional affect t will have on interpolation and therefore correlation. Clearly the labour and computationally intensive

part of COW is the optimisation of N and t. It has been shown that such optimisation will have a large affect on the ultimate quality of the alignment [168]. A further potential drawback of COW is that these two parameters are set globally whilst they are dealing with local variability. Future work could focus on a method to iteratively do local optimisation for these parameters in an automated fashion.

PTW is a fast global approach that doesn't take local variability into account and it's alignment accuracy suffers accordingly as it was shown to be the worst performer of these algorithms to an unacceptable level [168].

STW uses multiple splines (default 40) to address the local variability and in that sense is similar to the segmentation strategy of COW. STW is considerably faster than COW and on complex chromatograms has been shown to have similar to better performance than COW [168]. However, it is prone to overfitting and the number of splines to use needs to be optimised as well, which at present can only be done manually.

From an accuracy point of view STW $\geq$ COW $>$ DTW $>$ PTW. STW is considerably faster than COW, so would seem to be an attractive option. However, outside of the paper comparing it to other methods in 2006 [168], there seems to be no use of it in the literature. Perhaps this is due to the fact that there does not appear to be a software implementation of STW available. COW is commonly found in use in the literature, perhaps because there are several software implementations thereof.

2.6.3 Wavelets

None of the alignment algorithms described above yield perfectly aligned chromatograms. Chapter 3 describes a procedure using wavelets with which one can further refine the alignment of a large number of chromatograms and to reduce the size of the data set dramatically by use of a peak indexing and quantification approach.

A brief introduction to wavelets

As shown in Figure 2.26, a wavelet transform consists of taking a mother wavelet (simply a predetermined wave form) and translating it (shifting a time point at a time) across the signal of interest and calculating how similar the wavelet is to the signal present at each iteration. Once this has been done for each time point the wavelet is scaled (compressed or stretched) and the new wave form translated across each position of the signal and the similarity to the signal calculated. In doing so one generates a list of correlation coefficients which, when indexed to the wave form scale and time represent both frequency and time domain information about the signal. This information can then be used to analyse the data to remove noise (smoothing) or to identify signal features present throughout the time domain.

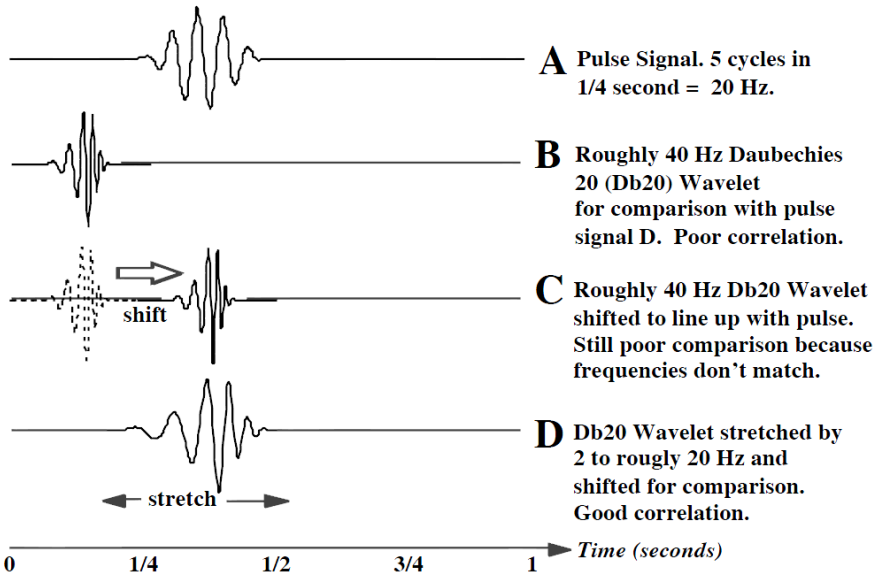


Figure 2.26: Illustration of a wavelet transform (from [56]).

Mathematically, a wavelet transfomation can be described as follows:

$$CWT_x^{\psi}(\tau, s) = \Psi_x^{\psi}(\tau, s) = \frac{1}{\sqrt{|s|}} \int x(t)\psi^* \left(\frac{t-\tau}{s}\right) dt \tag{2.6.7}$$

where τ and s are the translation and scale parameters, ψ is the mother wavelet (see below) and x(t) is the signal to be transformed. The result of such a transformation is a two-dimensional matrix of wavelet coefficients ($CWT_x^{\psi}(\tau, s)$) corresponding to each pair of τ and s used.

As shown in Figure 2.27, different mother wavelets can be used and are usually selected due to their similarity to the signal of interest. For chromatographic data the Mexican Hat wavelet is often used due to its similarity to a chromatographic peak, although other waveforms can be used if the goal is to select for the noise component of a signal.

2.6.3.1 Application of wavelets on analytical signals

Wavelets have been used on a wide variety of signal types including process monitoring [175], flow injection analysis [14], mass spectrometry [37, 111, 96], IR spectroscopy [15], capillary electrophoresis [110], image analysis [176] and chromatography [180, 110, 17, 181, 25, 182, 159, 132, 176, 35, 142, 66].

Wavelets used in chromatography

The intensity value across a GC-FID chromatogram is affected by the presence or absence of a peak caused by the ionisation of a molecule in the sample or the presence or absence of chemical or random noise. As such, the observed signal

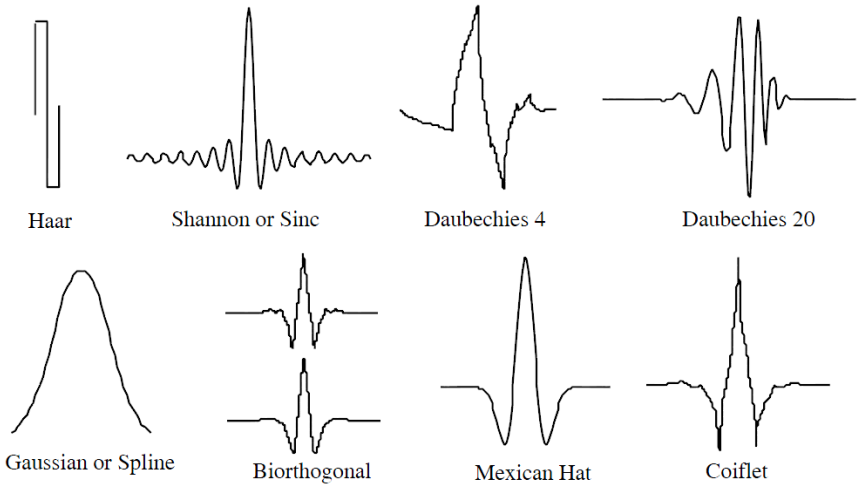


Figure 2.27: Different type of mother wavelets (from [56]).

(which varies as a function of time) is an arithmetic combination of the signal seen from a true ionised molecule, the baseline, stochastic (white) noise and a function to describe and account for the heteroscedasticity of the stochastic noise. The most common use of wavelets on chromatographic data sets has been for signal denoising and baseline correction [180, 110, 181, 182, 159, 176, 35, 142, 132, 17, 35, 66].

Wavelets used for Peak Detection

More recently wavelets have been used for peak detection in mass spectrometry data [38] and there are two recent instances of wavelets being used for peak detection in chromatographic data [25, 38]. However, Chen *et al.* [25] simply state that they use there own Matlab code to do so. They provide no description of what they did nor do they give any reference that describes their approach so this is somewhat difficult to evaluate.

Du *et al.* [38] developed a wavelet-based application for peak detection in very noisy peptide mass spectrometry data. They used a Mexican Hat mother wavelet and calculated the wavelet coefficients for 33 different scale values (1 to 66 in increments of two) and translated it over each position in the spectra. The resultant two-dimensional matrix of wavelet coefficients was then plotted as seen in Figure 2.28 where the ridges in the coefficient space can be clearly seen to correspond to mass spec peaks.

In order to make peak calls they identified the ridges in the coefficient space. Ridge identification was done by linking the local maxima as follows. The local maxima at each scale value were determined with a method similar to the one used in the PROcess R package by Gentlemen [58] employed with a sliding window. Those maxima were then linked as vertically connected lines that pass a minimum gap threshold for continuity. An example of the ridge

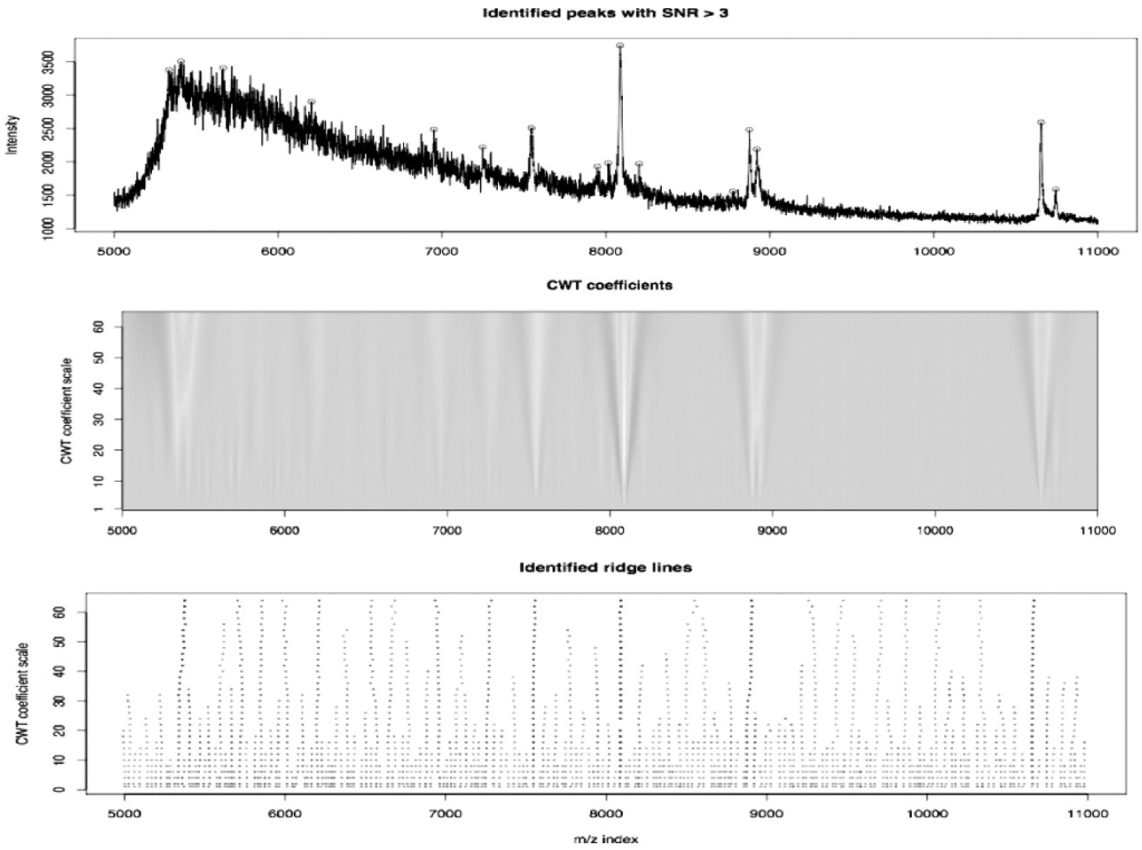


Figure 2.28: Wavlet-based peak detection. SNR = signal to noise ratio; CWT = continuous wavelet transform (from [38]).

lines found can be see in Figure 2.28. Peaks are then selected based on 1) the scale of the highest intensity coefficient on the ridge line being within a defined range, 2) the ridge line lengths being larger than a preset threshold and 3) a signal to noise ratio threshold. A putative peak's signal strength is taken to be the highest coefficient on the ridge line within its scale range. Noise was taken to "be the 95-percentage quantile of the absolute CWT coefficient values (a=1) within a local window surrounding the peak" [38].

Wavelets used for chromatographic peak detection and subsequent alignment

Zhang *et al.* used the wavelet algorithm proposed for mass spectrometry data by Du *et al.* [38] on chromatographic data for peak detection and then used a differential evolution algorithm to perform chromatographic alignments [180]. However, this work was done mostly on very limited synthetic data sets and only two real chromatograms. As such, it is difficult to tell how robust it will be with many different real chromatograms. They report that it performs very similarly to COW for alignment and is less likely to change peak heights

slightly as COW is prone to do. The code from this paper is available but is undocumented and replete with software bugs, we have been unable to get it to work on real chromatograms.

Finally, building on their previous work Zhang *et al.* [181] have developed multiscale peak alignment (MSPA) which used wavelets for peak detection and then used other methods for iterative alignment. Briefly, their method works as follows. The chromatograms are transformed into the 2D wavelet coefficient matrix and peaks detected as was done by Du *et al.* [38] and Zhang *et al.* [180]. However, for the mother wavelet they use a Haar wavelet instead of the Mexican Hat wavelet. They then segment the chromatogram (or its subsegments) into smaller segments with the use of a Shannon information content metric [147] in an iterative fashion. They define the Shannon information metric for peaks to be as follows. First the signal is row normalised to a total value of one, thus creating a probability distribution. The information content of each peak h_i is then defined by the following equation [181]:

$$h_i = -log_2 \frac{p_i}{\sum p_i} \tag{2.6.8}$$

where p_i is the area of peak i. When the h_i becomes smaller, a small uncertainty and a large peak is indicated. This was used to rank order the peaks by size and the larger peaks aligned first. This process is illustrated in Figure 2.29.

They then identify putative shifts with the use of fast Fourier transform cross correlation. They then combine adjacent segments and determine if the correlation coefficients increase to find the optimal regional shift in order to avoid local optima that could yield an incorrect alignment.

Unlike their previous work they tested MSPA on over 100 real chromatograms and the results do look encouraging (see Figure 2.30). However, by looking at the individual peak alignments it does appear that there is room for further refinement.

Correlation optimised warping has dominated the chromatogram alignment field to date. It is superior to dynamic time warping and available as software implementations in several different languages. However it's major draw back is the need for parameter optimisation and its computational complexity, which makes it slow. Semi-parametric Time Warping is an attractive option as it is fast and needs minimal parameter optimisation. However, its lack of availability as implemented software and its lack of use by different groups on real chromatograms are significant drawbacks. Wavelets have been extensively used for the denoising and baseline correction of chromatographic data sets. However, more recently both in this thesis (Chapter 3) and in work by Zhang *et al.* the wavelet-base peak detection algorithm originally developed for mass spectrometry data is being used as part of chromatographic alignment approaches. MSPA, The most recent work by Zhang *et al.* does look very encouraging but still needs to be compared to other methods to determine its comparative accuracy and speed. Its lack of need for parameter optimisation is very attractive.

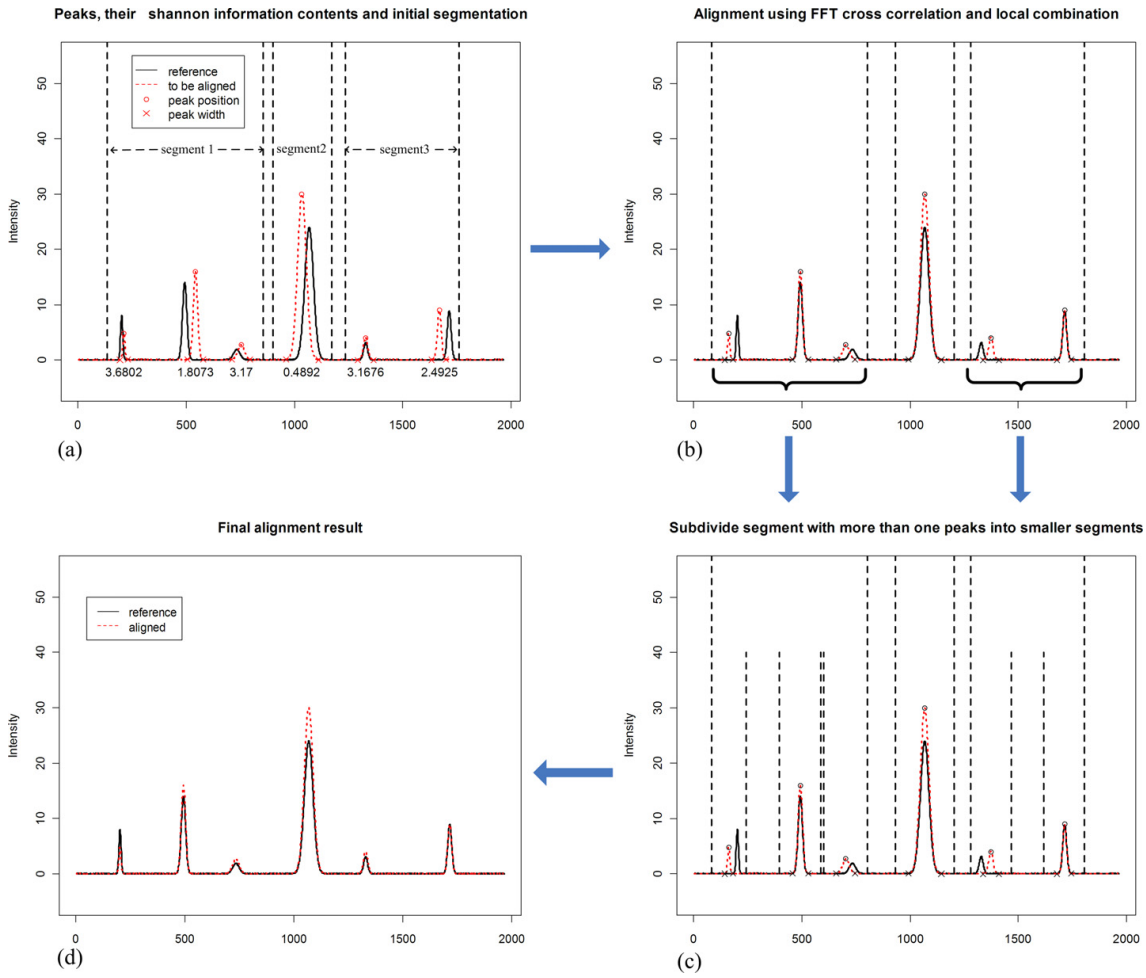


Figure 2.29: Iterative segmentation and alignment (modified from [181]).

Visual inspection indicates that its final results would still need refining by the method described in Chapter 3 were it to be used as a preprocessing step in network reconstruction. Future work could include an approach whereby each of the segments identified by MSPA is used as a mother wavelet. This mother wavelet could then be used in a continuous wavelet transformation of that region of the reference sequence. The warping is simply the translated and scaled wavelet resulting from the maximum wavelet coefficient.

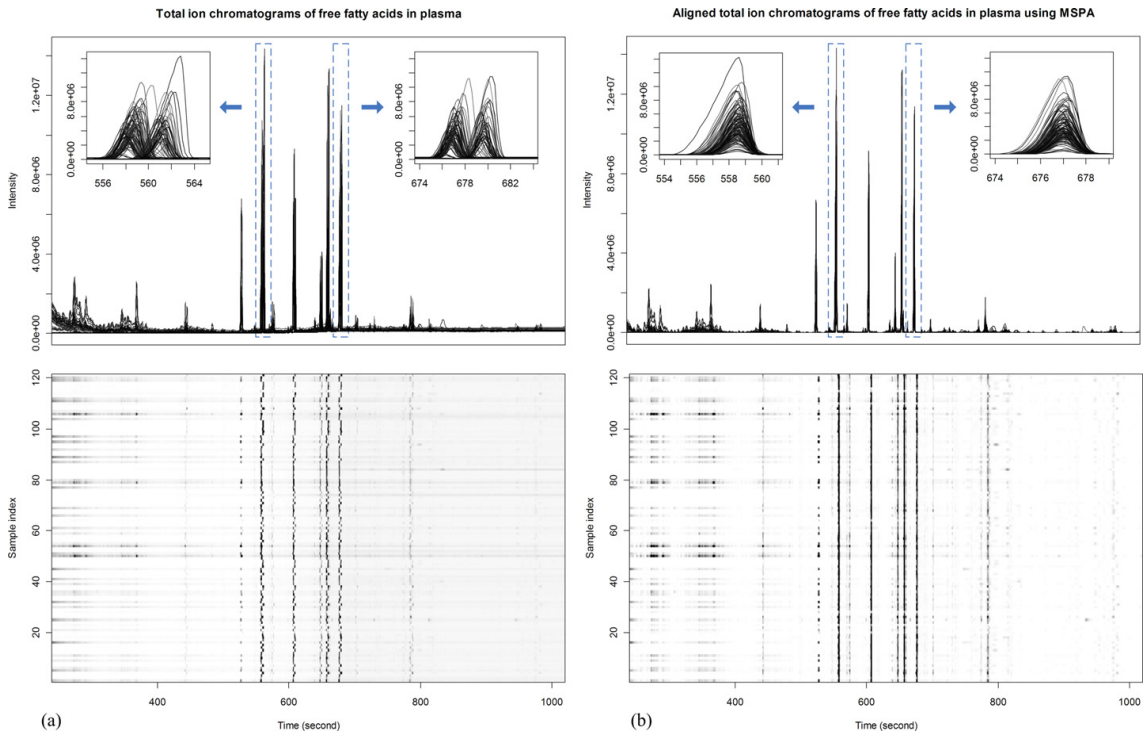


Figure 2.30: MSPA Alignment Results (from [181]).

2.7 Conclusions

Network theory has emerged as a new discipline with which to model and analyse complex systems and has arisen from the study of real networks derived from large volumes of data from a wide variety of disciplines. Networks have been used to represent extant knowledge and to represent correlative relationships derived from novel data sets. It has been noted that the the metrics used to create correlation networks can have a significant impact on the resulting network topology. The properties of networks and their associated matrices and graph-based algorithms can be leveraged to gain further insight into datasets. From a wine-based perspective, networks show considerable promise as a model which can handle much of the complexity inherent in -omics-based analysis of the wine making process.

A wide variety of approaches and statistical tests have been applied to gene set analysis. GSEA and the maxmean methods would seem to be the most robust methods as they used a mixed model and avoid the major pitfalls of the competitive and gene permutation methods. Maxmean has been shown to outperform GSEA across a number of different scenarios. The vast majority of the GSA literature is devoted to the development and comparison of gene set statistics with an almost complete absence of a discussion about the generation of gene sets to be used with GSA. As such, there seems to be significant scope for a research into gene set generation. The very challenging gene expression

datasets that result from from vineyard and cellar studies will be better served by gene set analysis than with traditional statistical methods that are unable to handle the orthogonal variation commonly found in these types of datasets.

Neither grapevine nor tobacco (used at the IWBT as a laboratory-based pathosystem) are model organisms and, as such, there is a limited amount of knowledge and data available about them at the -omics level. Thus, it is important to be able to infer information from other plant genomes in order to take advantage of the vast amount of extant knowledge available about other species, both as annotation as well as the information resident in expression datasets. In this regard, the use of ortholog detection algorithms is very important to wine science as it allows one to leverage knowledge derived by the broader plant research community and put the results from wine-related research into a broader biological context.

Correlation optimised warping has dominated the chromatogram alignment field to date. It is superior to dynamic time warping and available as software implementations in several different languages. However it's major draw back is the need for parameter optimisation and its computational complexity, which makes it slow. Semi-parametric Time Warping is an attractive option as it is fast and needs minimal parameter optimisation. However, its lack of availability as implemented software and its lack of use by different groups on real chromatograms are significant drawbacks. Wavelets have been extensively used for the denoising and baseline correction of chromatographic data sets. However, more recently both in this thesis (Chapter 3) and in work by Zhang *et al.* the wavelet-base peak detection algorithm originally developed for mass spectrometry data is being used as part of chromatographic alignment approaches. MSPA, The most recent work by Zhang *et al.* does look very encouraging but still needs to be compared to other methods to determine its comparative accuracy and speed but its lack of need for parameter optimisation is very attractive. Visual inspection indicates that its final results would still need refining by the method described in Chapter 3 were it to be used as a preprocessing step in network reconstruction. Future work could include an approach whereby each of the segments identified by MSPA is used as a mother wavelet. This mother wavelet could then be used in a continuous wavelet transformation of that region of the reference sequence. The warping is simply the translated and scaled wavelet resulting from the maximum wavelet coefficient. One of the significant difficulties in gaining further understanding of the chemistry of wine from a systems perspective is the relative sparsity of known compounds for targeted analysis. Systemic insights will only come when a larger portion of the compounds in wine can be analysed simultaneously in order to identify which compounds are important to identify and study further. As such, the use and improvement of chromatographic alignments hold tremendous promise in the untargeted analysis of grapevine and microbial metabolites throughout the wine making process as well as the chemical compounds that reactions result from bottle aging.

References

[1] Ahn, Y.-Y., Ahnert, S.E., Bagrow, J.P. and Barabási, A.-L. (2011 January). Flavor network and the principles of food pairing. *Scientific reports*, vol. 1, p. 196. ISSN 2045-2322.

[2] Albert, R. and Barabási, A. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, vol. 74, no. 1, p. 47.

[3] Alexeyenko, A., Tamas, I., Liu, G. and Sonnhammer, E.L.L. (2006). Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*, vol. 22, no. 14, pp. e9—-e15.

[4] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990 October). Basic local alignment search tool. *Journal of molecular biology*, vol. 215, no. 3, pp. 403–10. ISSN 0022-2836.

[5] Anderson, S.E., Dave, A.S. and Margoliash, D. (1996 August). Template-based automatic recognition of birdsong syllables from continuous recordings. *The Journal of the Acoustical Society of America*, vol. 100, no. 2 Pt 1, pp. 1209–19. ISSN 0001-4966.

[6] Aoki, K., Ogata, Y. and Shibata, D. (2007 March). Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant & cell physiology*, vol. 48, no. 3, pp. 381–90. ISSN 0032-0781.

[7] Arlin, M. (1976 March). Causal priority of social desirability over self-concept: a cross-lagged correlation analysis. *Journal of personality and social psychology*, vol. 33, no. 3, pp. 267–72. ISSN 0022-3514.

[8] Astobiza, I., Tilburg, J.J., Piñero, A., Hurtado, A., García-Pérez, A.L., Nabuurs-Franssen, M.H. and Klaassen, C.H. (2012 January). Genotyping of Coxiella burnetii from domestic ruminants in northern Spain. *BMC veterinary research*, vol. 8, p. 241. ISSN 1746-6148.

[9] Barabási, A.-L. (2007 July). Network medicine–from obesity to the "diseasome". *The New England journal of medicine*, vol. 357, no. 4, pp. 404–7. ISSN 1533-4406.

[10] Barrett, C., Hughey, R. and Karplus, K. (1997 April). Scoring hidden Markov models. *Computer applications in the biosciences : CABIOS*, vol. 13, no. 2, pp. 191–9. ISSN 0266-7061.

[11] Bellman, R. (1954). Some Problems in the Theory of Dynamic Programming. *Econometrica*, vol. 22, no. 1, pp. pp. 37–48. ISSN 00129682.

[12] Bellman, R. (1958). On a routing problem. *Quarterly of Applied Mathematics*, vol. 16, pp. 87–90.

[13] Berggren, W.A. and Aubry, M.P. (1996). A late Paleocene-early Eocene NW European and North Sea magnetobiochronological correlation network. *Geological Society, London, Special Publications*, vol. 101, no. 1, pp. 309–352.

[14] Bos, M. and Hoogendam, E. (1992 September). Wavelet transform for the evaluation of peak intensities in flow-injection analysis. *Analytica Chimica Acta*, vol. 267, no. 1, pp. 73–80. ISSN 00032670.

[15] Bos, M. and Vrielink, J. (1994 April). The wavelet transform for preprocessing IR spectra in the identification of mono- and di-substituted benzenes. *Chemometrics and Intelligent Laboratory Systems*, vol. 23, no. 1, pp. 115–122. ISSN 01697439.

[16] Cai, B., Wang, H., Zheng, H. and Wang, H. (2012 January). Detection of protein complexes from affinity purification/mass spectrometry data. *BMC systems biology*, vol. 6 Suppl 3, p. S4. ISSN 1752-0509.

[17] Cappadona, S., Levander, F., Jansson, M., James, P., Cerutti, S. and Pattini, L. (2008 July). Wavelet-based method for noise characterization and rejection in high-performance liquid chromatography coupled to mass spectrometry. *Analytical chemistry*, vol. 80, no. 13, pp. 4960–8. ISSN 1520-6882.

[18] Caspi, R., Altman, T., Dreher, K., Fulcher, C.A., Subhraveti, P., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Pujar, A., Shearer, A.G., Travers, M., Weerasinghe, D., Zhang, P. and Karp, P.D. (2012 January). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic acids research*, vol. 40, no. Database issue, pp. D742–53. ISSN 1362-4962.

[19] Caspi, R., Foerster, H., Fulcher, C.A., Hopkinson, R., Ingraham, J., Kaipa, P., Krummenacker, M., Paley, S., Pick, J., Rhee, S.Y., Tissier, C., Zhang, P. and Karp, P.D. (2006 January). MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic acids research*, vol. 34, no. Database issue, pp. D511–6. ISSN 1362-4962.

[20] Cevallos-Cevallos, J.M., Reyes-De-Corcuera, J.I., Etxeberria, E., Danyluk, M.D. and Rodrick, G.E. (2009). Metabolomic analysis in food science: a review. *Trends in Food Science & Technology*, vol. 20, no. 11-12, pp. 557–566. ISSN 0924-2244.

[21] Chae, L., Lee, I., Shin, J. and Rhee, S.Y. (2012 April). Towards understanding how molecular networks evolve in plants. *Current opinion in plant biology*, vol. 15, no. 2, pp. 177–84. ISSN 1879-0356.

[22] Chaisson, M.J. and Pevzner, P.a. (2008 March). Short read fragment assembly of bacterial genomes. *Genome research*, vol. 18, no. 2, pp. 324–30. ISSN 1088-9051.

[23] Chatterjee, A. and Mistry, N. (2012 December). MIRU-VNTR profiles of three major Mycobacterium tuberculosis spoligotypes found in western India. *Tuberculosis (Edinburgh, Scotland)*. ISSN 1873-281X.

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

[25] Chen, F., Xue, J., Zhou, L., Wu, S. and Chen, Z. (2011 October). Identification of serum biomarkers of hepatocarcinoma through liquid chromatography/mass spectrometry-based metabonomic method. *Analytical and bioanalytical chemistry*, vol. 401, no. 6, pp. 1899–904. ISSN 1618-2650.

[26] Chi, Z., Wu, W., Haga, Z., Hatsopoulos, N.G. and Margoliash, D. (2007 February). Template-based spike pattern identification with linear convolution and dynamic time warping. *Journal of neurophysiology*, vol. 97, no. 2, pp. 1221–35. ISSN 0022-3077.

[27] Churchill, G.A. (1989). Stochastic models for heterogeneous DNA sequences. *Bulletin of mathematical biology*, vol. 51, no. 1, pp. 79–94.

[28] Churchill, G.A. (1992). Hidden Markov chains and the analysis of genome structure. *Computers & chemistry*, vol. 16, no. 2, pp. 107–115.

[29] Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L.Y., Toufighi, K., Mostafavi, S., Prinz, J., St Onge, R.P., VanderSluis, B., Makhnevych, T., Vizeacoumar, F.J., Alizadeh, S., Bahr, S., Brost, R.L., Chen, Y., Cokol, M., Deshpande, R., Li, Z., Lin, Z.-Y., Liang, W., Marback, M., Paw, J., San Luis, B.-J., Shuteriqi, E., Tong, A.H.Y., van Dyk, N., Wallace, I.M., Whitney, J.a., Weirauch, M.T., Zhong, G., Zhu, H., Houry, W.a., Brudno, M., Ragibizadeh, S., Papp, B., Pál, C., Roth, F.P., Giaever, G., Nislow, C., Troyanskaya, O.G., Bussey, H., Bader, G.D., Gingras, A.-C., Morris, Q.D., Kim, P.M., Kaiser, C.a., Myers, C.L., Andrews, B.J. and Boone, C. (2010 January). The genetic landscape of a cell. *Science (New York, N.Y.)*, vol. 327, no. 5964, pp. 425–31. ISSN 1095-9203.

[30] Cserháti, T. (2010 January). Data evaluation in chromatography by principal component analysis. *Biomedical chromatography : BMC*, vol. 24, no. 1, pp. 20–8. ISSN 1099-0801.

[31] Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366.

[32] Dehal, P.S. and Boore, J.L. (2006 January). A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. *BMC bioinformatics*, vol. 7, p. 201. ISSN 1471-2105.

[33] Dijkstra, E.W. (1959 December). A note on two problems in connexion with graphs. *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271. ISSN 0029-599X.

[34] Dinu, I., Potter, J.D., Mueller, T., Liu, Q., Adewale, A.J., Jhangri, G.S., Einecke, G., Famulski, K.S., Halloran, P. and Yasui, Y. (2007 January). Improving gene set analysis of microarray data by SAM-GS. *BMC bioinformatics*, vol. 8, p. 242. ISSN 1471-2105.

[35] Dixon, S.J., Brereton, R.G., Soini, H.A., Novotny, M.V. and Penn, D.J. (2007). An automated method for peak detection and matching in large gas chromatography-mass spectrometry data sets. *Journal of Chemometrics*, , no. January, pp. 325–340.

[36] Donges, J.F., Zou, Y., Marwan, N. and Kurths, J. (2009 July). Complex networks in climate dynamics. *The European Physical Journal Special Topics*, vol. 174, no. 1, pp. 157–179. ISSN 1951-6355.

[37] Du, P., Kibbe, W.a. and Lin, S.M. (2006 September). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics (Oxford, England)*, vol. 22, no. 17, pp. 2059–65. ISSN 1367-4811.

[38] Du, P., Kibbe, W.A. and Lin, S.M. (2006 September). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics (Oxford, England)*, vol. 22, no. 17, pp. 2059–65. ISSN 1367-4811.

[39] Dufayard, J.-F., Duret, L., Penel, S., Gouy, M., Rechenmann, F. and Perrière, G. (2005 June). Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics (Oxford, England)*, vol. 21, no. 11, pp. 2596–603. ISSN 1367-4803.

[40] e Silva, H., de Pinho, P., Machado, B.P., Hogg, T., Marques, J.C., Câmara, J.S., Albuquerque, F. and Silva Ferreira, A.C. (2008). Impact of forced-aging process on Madeira wine flavor. *Journal of agricultural and food chemistry*, vol. 56, no. 24, pp. 11989–11996.

[41] Echeverría-Londoño, S. and Miranda-Esquivel, D.R. (2011 January). Marti-Tracks: a geometrical approach for identifying geographical patterns of distribution. *PloS one*, vol. 6, no. 4, p. e18460. ISSN 1932-6203.

[42] Eckmann, J.-P. and Moses, E. (2002 April). Curvature of co-links uncovers hidden thematic layers in the World Wide Web. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 9, pp. 5825–9. ISSN 0027-8424.

[43] Efron, B. and Tibshirani, R. (2007 June). On testing the significance of sets of genes. *Annals of Applied Statistics*, vol. 1, no. 1, pp. 107–129. ISSN 1932-6157.

[44] Eilers, P.H.C. (2004 January). Parametric time warping. *Analytical chemistry*, vol. 76, no. 2, pp. 404–11. ISSN 0003-2700.

[45] Enright, A.J., Kunin, V. and Ouzounis, C.A. (2003 August). Protein families and TRIBES in genome sequence space. *Nucleic acids research*, vol. 31, no. 15, pp. 4632–8. ISSN 1362-4962.

[46] Enright, A.J., Van Dongen, S. and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, vol. 30, no. 7, pp. 1575–1578.

[47] Escudero, A., Cacho, J. and Ferreira, V. (2000). Isolation and identification of odorants generated in wine during its oxidation: A gas chromatography–olfactometric study. *European Food Research and Technology*, vol. 211, no. 2, pp. 105–110.

[48] Falcon, S. and Gentleman, R. (2007 January). Using GOstats to test gene lists for GO term association. *Bioinformatics (Oxford, England)*, vol. 23, no. 2, pp. 257–8. ISSN 1367-4811.

[49] Farber, C.R. and Lusis, A.J. (2008 January). Integrating global gene expression analysis and genetics. *Advances in genetics*, vol. 60, no. 07, pp. 571–601. ISSN 0065-2660.

[50] Felsenstein, J. and Churchill, G.A. (1996 January). A Hidden Markov Model approach to variation among sites in rate of evolution. *Molecular biology and evolution*, vol. 13, no. 1, pp. 93–104. ISSN 0737-4038.

[51] Ferreira, A.C.S., Barbe, J.C. and Bertrand, A. (2003). 3-Hydroxy-4, 5-dimethyl-2 (5 H)-furanone: a key odorant of the typical aroma of oxidative aged port wine. *Journal of agricultural and food chemistry*, vol. 51, no. 15, pp. 4356–4363.

[52] Ferreira, A.C.S., Hogg, T. and de Pinho, P.G. (2003). Identification of key odorants related to the typical aroma of oxidation-spoiled white wines. *Journal of agricultural and food chemistry*, vol. 51, no. 5, pp. 1377–1381.

[53] Ficklin, S.P. and Feltus, F.A. (2011 July). Gene coexpression network alignment and conservation of gene modules between two grass species: maize and rice. *Plant physiology*, vol. 156, no. 3, pp. 1244–56. ISSN 1532-2548.

[54] Floyd, R.W. (1962). Algorithm 97: shortest path. *Communications of the ACM*, vol. 5, no. 6, p. 345.

[55] Ford, L.R. and Fulkerson, D.R. (1962). *Flows in Networks*, vol. 16 (5) The. Princeton University Press.

[56] Fugal, D. (2009). *Conceptual Waveletes in Digital Signal Processing*. 1st edn. Space & Signals Technologies, LLC, San Diego. ISBN 0982199457.

[57] Geijer, C., Pirkov, I., Vongsangnak, W., Ericsson, A., Nielsen, J., Krantz, M. and Hohmann, S. (2012 October). Time course gene expression profiling of yeast spore germination reveals a network of transcription factors orchestrating the global response. *BMC genomics*, vol. 13, no. 1, p. 554. ISSN 1471-2164.

[58] Gentleman R, Carey V, Huber W, Irizarry R, D.S. (ed.) (2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer.

[59] Gironi, M., Saresella, M., Rovaris, M., Vaghi, M., Nemni, R., Clerici, M. and Grossi, E. (2013 January). A novel data mining system points out hidden relationships between immunological markers in multiple sclerosis. *Immunity & ageing : I & A*, vol. 10, no. 1, p. 1. ISSN 1742-4933.

[60] Gitter, A., Carmi, M., Barkai, N. and Bar-Joseph, Z. (2012 October). Linking the signaling cascades and dynamic regulatory networks controlling stress responses. *Genome research*. ISSN 1549-5469.

[61] Glaze, C.M. and Troyer, T.W. (2006 January). Temporal structure in zebra finch song: implications for motor coding. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 26, no. 3, pp. 991–1005. ISSN 1529-2401.

[62] Goeman, J.J. and Bühlmann, P. (2007 April). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics (Oxford, England)*, vol. 23, no. 8, pp. 980–7. ISSN 1367-4811.

[63] Goeman, J.J., Oosting, J., Cleton-Jansen, A.-M., Anninga, J.K. and van Houwelingen, H.C. (2005 May). Testing association of a pathway with survival using gene expression data. *Bioinformatics (Oxford, England)*, vol. 21, no. 9, pp. 1950–7. ISSN 1367-4803.

[64] Goeman, J.J., van de Geer, S.A., de Kort, F. and van Houwelingen, H.C. (2004 January). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics (Oxford, England)*, vol. 20, no. 1, pp. 93–9. ISSN 1367-4803.

[65] Goh, K.-i., Cusick, M.E., Valle, D., Childs, B., Vidal, M. and Barabási, A.-L. (2007 May). The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 21, pp. 8685–90. ISSN 0027-8424.

[66] Gong, Q., Li, P., Ma, S., Indu Rupassara, S. and Bohnert, H.J. (2005). Salinity stress adaptation competence in the extremophile Thellungiella halophila in comparison with its relative Arabidopsis thaliana. *The Plant Journal*, vol. 44, no. 5, pp. 826–839.

[67] Goodstadt, L. and Ponting, C.P. (2006 September). Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS computational biology*, vol. 2, no. 9, p. e133. ISSN 1553-7358.

[68] Grimes, M.L., Lee, W.-J., van der Maaten, L. and Shannon, P. (2013 January). Wrangling phosphoproteomic data to elucidate cancer signaling pathways. *PloS one*, vol. 8, no. 1, p. e52884. ISSN 1932-6203.

[69] Gu, C. and Davidsen, J. (2010 December). Statistical properties of aftershocks. *AGU Fall Meeting Abstracts*, p. B1412.

[70] Gusfield, D. (1993 January). Efficient methods for multiple sequence alignment with guaranteed error bounds. *Bulletin of mathematical biology*, vol. 55, no. 1, pp. 141–54. ISSN 0092-8240.

[71] Haussler, D., Krogh, a., Mian, I. and Sjolander, K. (1993). Protein modeling using hidden Markov models: analysis of globins. *[1993] Proceedings of the Twenty-sixth Hawaii International Conference on System Sciences*, pp. 792–802.

[72] Henry, C.S., DeJongh, M., Best, A.A., Frybarger, P.M., Linsay, B. and Stevens, R.L. (2010 September). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology*, vol. 28, no. 9, pp. 977–82. ISSN 1546-1696.

[73] Hidalgo, C.A., Klinger, B., Barabási, A.-L. and Hausmann, R. (2007 July). The product space conditions the development of nations. *Science (New York, N.Y.)*, vol. 317, no. 5837, pp. 482–7. ISSN 1095-9203.

[74] Hoeffding, W. (1948). A non-parametric test of independence. *The Annals of Mathematical Statistics*, vol. 19, pp. 546–557.

[75] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, vol. 24, no. 6, p. 417.

[76] Huang, T., Zhang, J., Xu, Z.-P., Hu, L.-L., Chen, L., Shao, J.-L., Zhang, L., Kong, X.-Y., Cai, Y.-D. and Chou, K.-C. (2012 May). Deciphering the effects of gene deletion on yeast longevity using network and machine learning approaches. *Biochimie*, vol. 94, no. 4, pp. 1017–25. ISSN 1638-6183.

[77] Irizarry, R. and Wang, C. (2009). Gene set enrichment analysis made simple. *Statistical methods in medical research*, vol. 18, no. 6, pp. 565–575.

[78] Johnson, D.B. (1977). Efficient algorithms for shortest paths in sparse networks. *Journal of the ACM (JACM)*, vol. 24, no. 1, pp. 1–13.

[79] Jolliffe, I.T. (2002). *Principal Component Analysis, Second Edition*. Springer. ISBN 0-387-95442-2.

[80] Jordan, M.I. (2004). Learning in graphical models. vol. 19, no. 1, pp. 140–155.

[81] Jothi, R., Zotenko, E., Tasneem, A. and Przytycka, T.M. (2006). COCO-CL: hierarchical clustering of homology relations based on evolutionary correlations. *Bioinformatics*, vol. 22, no. 7, pp. 779–788.

[82] Kanehisa, M. (1996). Toward pathway engineering: a new database of genetic and molecular pathways. *Science & Technology Japan*, vol. 59, pp. 34–38.

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

[84] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. and Tanabe, M. (2012 January). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, vol. 40, no. Database issue, pp. D109–14. ISSN 1362-4962.

[85] Karp, P.D. (1992 August). A knowledge base of the chemical compounds of intermediary metabolism. *Computer applications in the biosciences : CABIOS*, vol. 8, no. 4, pp. 347–57. ISSN 0266-7061.

[86] Karp, P.D. and Caspi, R. (2011 September). A survey of metabolic databases emphasizing the MetaCyc family. *Archives of toxicology*, vol. 85, no. 9, pp. 1015–33. ISSN 1432-0738.

[87] Karp, P.D. and Paley, S.M. (1994 January). Representations of metabolic knowledge: pathways. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology ; ISMB.*, vol. 2, pp. 203–11. ISSN 1553-0833.

[88] Karp, P.D., Paley, S.M., Krummenacker, M., Latendresse, M., Dale, J.M., Lee, T.J., Kaipa, P., Gilham, F., Spaulding, A., Popescu, L., Altman, T., Paulsen, I., Keseler, I.M. and Caspi, R. (2010 January). Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Briefings in bioinformatics*, vol. 11, no. 1, pp. 40–79. ISSN 1477-4054.

[89] Karp, P.D. and Riley, M. (1993 January). Representations of metabolic knowledge. *Proceedings of the nternational Conference on Intelligent Systems for Molecular Biology ; ISMB.*, vol. 1, pp. 207–15. ISSN 1553-0833.

[90] Karp, P.D., Riley, M., Paley, S.M. and Pelligrini-Toole, A. (1996 January). EcoCyc: an encyclopedia of Escherichia coli genes and metabolism. *Nucleic acids research*, vol. 24, no. 1, pp. 32–9. ISSN 0305-1048.

[91] Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Paley, S.M. and Pellegrini-Toole, A. (2000 January). The EcoCyc and MetaCyc databases. *Nucleic acids research*, vol. 28, no. 1, pp. 56–9. ISSN 0305-1048.

[92] Kassidas, A., MacGregor, J.F. and Taylor, P.a. (1998 April). Synchronization of batch trajectories using dynamic time warping. *AIChE Journal*, vol. 44, no. 4, pp. 864–875. ISSN 00011541.

[93] Kendall, M.G. (1948). Rank correlation methods.

[94] Kim, S.-Y. and Volsky, D.J. (2005 January). PAGE: parametric analysis of gene set enrichment. *BMC bioinformatics*, vol. 6, p. 144. ISSN 1471-2105.

[95] Kogan, J.A. and Margoliash, D. (1998 April). Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: a comparative study. *The Journal of the Acoustical Society of America*, vol. 103, no. 4, pp. 2185–96. ISSN 0001-4966.

[96] Koo, I., Zhang, X. and Kim, S. (2011 July). Wavelet- and Fourier-transform-based spectrum similarity approaches to compound identification in gas chromatography/mass spectrometry. *Analytical chemistry*, vol. 83, no. 14, pp. 5631–8. ISSN 1520-6882.

[97] Kosorok, M.R. (2009 January). On Brownian Distance Covariance and High Dimensional Data. *The annals of applied statistics*, vol. 3, no. 4, pp. 1266–1269. ISSN 1932-6157. `arXiv:1010.0297v2`.

[98] Krogh, A. (1998). An Introduction to Hidden Markov Models for Biological Sequences. In: Salzberg, SL, Searls, DB, Kasif, S. (ed.), *Computational Methods in Molecular Biology*, pp. 45–63. Elsevier.

[99] Krogh, A., Brown, M., Mian, I., Sjölander, K. and Haussler, D. (1994 February). Hidden Markov models in computational biology. Applications to protein modeling. *Journal of molecular biology*, vol. 235, no. 5, pp. 1501–31. ISSN 0022-2836.

[100] Krumsiek, J., Suhre, K., Illig, T., Adamski, J. and Theis, F.J. (2011 January). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC systems biology*, vol. 5, no. 1, p. 21. ISSN 1752-0509.

[101] Kruskal, J.B. (1956). 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.

[102] Kumari, S., Nie, J., Chen, H.-S., Ma, H., Stewart, R., Li, X., Lu, M.-Z., Taylor, W.M. and Wei, H. (2012 January). Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. *PloS one*, vol. 7, no. 11, p. e50411. ISSN 1932-6203.

[103] Kuzniar, A., van Ham, R., Pongor, S. and Leunissen, J.A.M. (2008). The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet*, vol. 24, no. 11, pp. 539–551.

[104] Lachlan, R.F., Verhagen, L., Peters, S. and Cate, C.T. (2010 February). Are there species-universal categories in bird song phonology and syntax? A comparative study of chaffinches (Fringilla coelebs), zebra finches (Taenopygia

guttata), and swamp sparrows (Melospiza georgiana). *Journal of comparative psychology (Washington, D.C. : 1983)*, vol. 124, no. 1, pp. 92–108. ISSN 1939-2087.

[105] Lavigne, V., Pons, A., Darriet, P. and Dubourdieu, D. (2008). Changes in the sotolon content of dry white wines during barrel and bottle aging. *Journal of agricultural and food chemistry*, vol. 56, no. 8, pp. 2688–2693.

[106] Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J. and Pavlidis, P. (2004 June). Coexpression analysis of human genes across many microarray data sets. *Genome research*, vol. 14, no. 6, pp. 1085–94. ISSN 1088-9051.

[107] Leibon, G., Pauls, S., Rockmore, D. and Savell, R. (2008 December). Topological structures in the equities market network. *Proceedings of the National Academy of Sciences*, vol. 105, no. 52, pp. 20589–20594. ISSN 0027-8424.

[108] Li, H., Coghlan, A., Ruan, J., Coin, L.J., Hériché, J.-K., Osmotherly, L., Li, R., Liu, T., Zhang, Z., Bolund, L., Wong, G.K.-S., Zheng, W., Dehal, P., Wang, J. and Durbin, R. (2006 January). TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic acids research*, vol. 34, no. Database issue, pp. D572–80. ISSN 1362-4962.

[109] Li, L., Jr, C.J.S. and Roos, D.S. (2003). OrthoMCL : Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*, pp. 2178–2189.

[110] Liu, B.-F., Sera, Y., Matsubara, N., Otsuka, K. and Terabe, S. (2003 September). Signal denoising and baseline correction by discrete wavelet transform for microchip capillary electrophoresis. *Electrophoresis*, vol. 24, no. 18, pp. 3260–5. ISSN 0173-0835.

[111] Liu, Y. (2009 September). Feature extraction and dimensionality reduction for mass spectrometry data. *Computers in biology and medicine*, vol. 39, no. 9, pp. 818–23. ISSN 1879-0534.

[112] Lu, C.L., Tang, C.Y. and Lee, R.C.-T. (2003 September). The full Steiner tree problem. *Theoretical Computer Science*, vol. 306, no. 1-3, pp. 55–67. ISSN 03043975.

[113] M, G. (1992). Biological Pathways, Ed 3. Boehringer Mannheim, Mannheim, Germany.

[114] Marchand, S., de Revel, G. and Bertrand, A. (2000). Approaches to wine aroma: release of aroma compounds from reactions between cysteine and carbonyl compounds in wine. *Journal of agricultural and food chemistry*, vol. 48, no. 10, pp. 4890–4895.

[115] Markowetz, F. and Troyanskaya, O.G. (2007 July). Computational identification of cellular networks and pathways. *Molecular bioSystems*, vol. 3, no. 7, pp. 478–82. ISSN 1742-206X.

[116] Merkeev, I.V., Novichkov, P.S. and Mironov, A.A. (2006 January). PHOG: a database of supergenomes built from proteome complements. *BMC evolutionary biology*, vol. 6, p. 52. ISSN 1471-2148.

[117] Merris, R. (1994 January). Laplacian matrices of graphs: a survey. *Linear Algebra and its Applications*, vol. 197-198, pp. 143–176. ISSN 00243795.

[118] Meyn, S. and Tweedie, R. (1993). Heuristics. In: *Markov chains and stochastic stability*, pp. 3–23. Springer Verlag.

[119] Mootha, V.K., Lindgren, C.M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrå le, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D. and Groop, L.C. (2003 July). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, vol. 34, no. 3, pp. 267–73. ISSN 1061-4036.

[120] Moutounet, M., Rabier, P., Puech, J.L., Verette, E. and Barillere, J.M. (1989). Analysis by HPLC of extractable substances in oak wood. Application to a Chardonnay wine. *Sci. Aliments*, vol. 9, no. 1, pp. 35–51.

[121] Myers, C., Rabiner, L. and Rosenberg, A. (1980). Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 6, pp. 623–635.

[122] Myers, C.S. and HABINER, L.F. (1981). A Comparative Study of Several Dynamic Time-Warping Algorithms for Connected-Word. *Bell System Technical Journal*, vol. 60, no. 7, pp. 1389–1409.

[123] Nam, D. and Kim, S.-y. (2008). Gene-set approach for expression pattern analysis. *Access*, vol. 9, no. 3.

[124] Nielsen, N.-P.V., Carstensen, J.M. and Smedsgaard, J.r. (1998 May). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A*, vol. 805, no. 1-2, pp. 17–35. ISSN 00219673.

[125] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999 January). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research*, vol. 27, no. 1, pp. 29–34. ISSN 0305-1048.

[126] Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J. and a L Barabási (2007 May). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 18, pp. 7332–6. ISSN 0027-8424.

[127] Pavlidis, P., Qin, J., Arango, V., Mann, J.J. and Sibille, E. (2004 June). Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochemical research*, vol. 29, no. 6, pp. 1213–22. ISSN 0364-3190.

[128] Pavlidis P, Lewis DP, N.W. (2002). Exploring gene expression data with class scores. In: *Pac Symp Biocomput*, pp. 474–485.

[129] Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Philos. Trans. Royal Soc. London Ser. A*, vol. 187, pp. 253–318.

[130] Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572.

[131] Pravdova, V., Walczak, B. and Massart, D.L. (2002). A comparison of two algorithms for warping of analytical signals. vol. 456, no. September 2001, pp. 77–92.

[132] Prikler, S., Pick, D. and Einax, J.W. (2012 May). Comparing different means of signal treatment for improving the detection power in HPLC-ICP-MS. *Analytical and bioanalytical chemistry*, vol. 403, no. 4, pp. 1109–16. ISSN 1618-2650.

[133] Pripis-Nicolau, L., De Revel, G., Bertrand, A. and Maujean, A. (2000). Formation of flavor components by the reaction of amino acid and carbonyl compounds in mild conditions. *Journal of agricultural and food chemistry*, vol. 48, no. 9, pp. 3761–3766.

[134] Prom-on, S., Chanthaphan, A., Chan, J.H. and Meechai, A. (2011 February). Enhancing Biological Relevance of a Weighted Gene Co-Expression Network for Functional Module Identification. *Journal of Bioinformatics and Computational Biology*, vol. 09, no. 01, pp. 111–129. ISSN 0219-7200.

[135] Quade D, S.I. (1992). A survey of weighted rank correlation. In: P.K. Salama, PK, S.I. (ed.), *Order statistics and nonparametrics: theory and applications*, pp. 213–225. Elsevier Science Publishers B.V., Amsterdam.

[136] Rabiner, L., Rosenberg, A. and Levinson, S. (1978). Considerations in dynamic time warping algorithms for discrete word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, no. 6, pp. 575–582.

[137] Rabiner, L. and Schmidt, C. (1980). Application of dynamic time warping to connected digit recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 377–388.

[138] Remm, M., Storm, C.E.V. and Sonnhammer, E.L.L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons1. *J Mol Biol*, vol. 314, no. 5, pp. 1041–1052.

[139] Rossouw, D., Jacobson, D. and Bauer, F.F. (2012 January). Transcriptional regulation and the diversification of metabolism in wine yeast strains. *Genetics*, vol. 190, no. 1, pp. 251–61. ISSN 1943-2631.

[140] Saerens, M., Fouss, F., Yen, L. and Dupont, P. (2004). The principal components analysis of a graph, and its relationships to spectral clustering. *Machine Learning: ECML 2004*, pp. 371–383.

[141] Saito, K., Hirai, M.Y. and Yonekura-Sakakibara, K. (2008 January). Decoding genes with coexpression networks and metabolomics - 'majority report by precogs'. *Trends in plant science*, vol. 13, no. 1, pp. 36–43. ISSN 1360-1385.

[142] Sanchez-Ponce, R. and Guengerich, F.P. (2007 May). Untargeted analysis of mass spectrometry data for elucidation of metabolites and function of enzymes. *Analytical chemistry*, vol. 79, no. 9, pp. 3355–62. ISSN 0003-2700.

[143] Sanguansat, P. (ed.) (2012 February). *Principal Component Analysis - Multidisciplinary Applications*. InTech. ISBN 978-953-51-0129-1.

[144] Schellenberger, J., Park, J.O., Conrad, T.M. and Palsson, B.O. (2010 January). BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC bioinformatics*, vol. 11, p. 213. ISSN 1471-2105.

[145] Schlitt, T. and Brazma, A. (2002 January). Learning about gene regulatory networks from gene deletion experiments. *Comparative and functional genomics*, vol. 3, no. 6, pp. 499–503. ISSN 1531-6912.

[146] Sen, P.K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, vol. 63, no. 324, pp. 1379–1389.

[147] Shannon, C.E., Weaver, W., Blahut, R.E. and Hajek, B. (1949). *The mathematical theory of communication*, vol. 117. University of Illinois press Urbana.

[148] Shen, R., Goonesekere, N.C. and Guda, C. (2012 January). Mining functional subgraphs from cancer protein-protein interaction networks. *BMC systems biology*, vol. 6 Suppl 3, p. S2. ISSN 1752-0509.

[149] Smith, D.G. and Fewtrell, M.D. (1979 February). A use of network diagrams in depicting stratigraphic time-correlation. *Journal of the Geological Society*, vol. 136, no. 1, pp. 21–28. ISSN 0016-7649.

[150] Smith, R.D. (2009 June). The Spread of the Credit Crisis: View from a Stock Correlation Network. *Journal of the Korean Physical Society*, vol. 54, no. 6, p. 2460. ISSN 0374-4884.

[151] Spearman, C. (1987). The proof and measurement of association between two things. *The American journal of psychology*, vol. 100, no. 3/4, pp. 441–471.

[152] Storm, C.E.V. and Sonnhammer, E.L.L. (2002). Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics*, vol. 18, no. 1, pp. 92–99.

[153] Stuart, J.M., Segal, E., Koller, D. and Kim, S.K. (2003 October). A gene-coexpression network for global discovery of conserved genetic modules. *Science (New York, N.Y.)*, vol. 302, no. 5643, pp. 249–55. ISSN 1095-9203.

[154] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P. (2005 October). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–50. ISSN 0027-8424.

[155] Szappanos, B., Kovács, K., Szamecz, B., Honti, F., Costanzo, M., Baryshnikova, A., Gelius-Dietrich, G., Lercher, M.J., Jelasity, M., Myers, C.L., Andrews, B.J., Boone, C., Oliver, S.G., Pál, C. and Papp, B. (2011 July). An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nature genetics*, vol. 43, no. 7, pp. 656–62. ISSN 1546-1718.

[156] Tarjan, R.E. (1982). Sensitivity Analysis of Minimum Spanning Trees and Shortest Path Trees. *INFO. PROC. LETT.*, vol. 14, no. 1, pp. 30–33.

[157] Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N. and Others (2003). The COG database: an updated version includes eukaryotes. *BMC bioinformatics*, vol. 4, no. 1, p. 41.

[158] Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997). A genomic perspective on protein families. *Science*, vol. 278, no. 5338, pp. 631–637.

[159] Tautenhahn, R., Böttcher, C. and Neumann, S. (2008 January). Highly sensitive feature detection for high resolution LC/MS. *BMC bioinformatics*, vol. 9, p. 504. ISSN 1471-2105.

[160] Teacher, A.G.F. and Griffiths, D.J. (2011 January). HapStar: automated haplotype network layout and visualization. *Molecular ecology resources*, vol. 11, no. 1, pp. 151–3. ISSN 1755-0998.

[161] Theil, H. (1992). A rank-invariant method of linear and polynomial regression analysis. *Henri TheilâĂŹs Contributions to Economics and Econometrics*, pp. 345—-381.

[162] Tian, L., Greenberg, S.A., Kong, S.W., Altschuler, J., Kohane, I.S. and Park, P.J. (2005 September). Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 38, pp. 13544–9. ISSN 0027-8424.

[163] Tomasi, G., van den Berg, F. and Andersson, C. (2004 May). Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics*, vol. 18, no. 5, pp. 231–241. ISSN 0886-9383.

[164] van der Heijden, R., Snel, B., Van Noort, V. and Huynen, M. (2007). Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC bioinformatics*, vol. 8, no. 1, p. 83.

[165] van Dongen, S. (2000). MCL - a cluster algorithm for graphs. \url{http://micans.org/mcl/}.

[166] Van Dongen, S. (2008). Graph clustering via a discrete uncoupling process. *SIAM J MATRIX ANAL A*, vol. 30, no. 1, pp. 121–141.

[167] van Dongen, S.M. (2000). *Graph clustering by flow simulation*. Ph.D. thesis, Centre for Mathematics and Computer Science.

[168] van Nederkassel, a.M., Daszykowski, M., Eilers, P.H.C. and Heyden, Y.V. (2006 June). A comparison of three algorithms for chromatograms alignment. *Journal of chromatography. A*, vol. 1118, no. 2, pp. 199–210. ISSN 0021-9673.

[169] Wang, C.P. and Isenhour, T.L. (1987). Time-warping algorithm applied to chromatographic peak matching gas chromatography/Fourier transform infrared/mass spectrometry. *Analytical Chemistry*, vol. 59, no. 4, pp. 649–654.

[170] Wold, S., Esbensen, K. and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52.

[171] Wolfe, C.J., Kohane, I.S. and Butte, A.J. (2005 January). Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC bioinformatics*, vol. 6, p. 227. ISSN 1471-2105.

[172] Wu, L.F., Hughes, T.R., Davierwala, A.P., Robinson, M.D., Stoughton, R. and Altschuler, S.J. (2002 July). Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters. *Nature genetics*, vol. 31, no. 3, pp. 255–65. ISSN 1061-4036.

[173] Xu, Y., Olman, V. and Xu, D. (2002 April). Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees. *Bioinformatics (Oxford, England)*, vol. 18, no. 4, pp. 536–45. ISSN 1367-4803.

[174] Yong, C.H., Liu, G., Chua, H.N. and Wong, L. (2012 January). Supervised maximum-likelihood weighting of composite protein networks for complex prediction. *BMC systems biology*, vol. 6 Suppl 2, p. S13. ISSN 1752-0509.

[175] Zainuddin, Z., Wan Daud, W.R., Pauline, O. and Shafie, A. (2011 December). Wavelet neural networks applied to pulping of oil palm fronds. *Bioresource technology*, vol. 102, no. 23, pp. 10978–86. ISSN 1873-2976.

[176] Zhai, H.L., Hu, F.D., Huang, X.Y. and Chen, J.H. (2010 January). The application of digital image recognition to the analysis of two-dimensional fingerprints. *Analytica chimica acta*, vol. 657, no. 2, pp. 131–5. ISSN 1873-4324.

[177] Zhang, B. and Horvath, S. (2005 January). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, vol. 4, no. 1, pp. 1–43. ISSN 1544-6115.

[178] Zhang, P., Dreher, K., Karthikeyan, A., Chi, A., Pujar, A., Caspi, R., Karp, P., Kirkup, V., Latendresse, M., Lee, C., Mueller, L.A., Muller, R. and Rhee, S.Y. (2010 August). Creation of a genome-wide metabolic pathway database for Populus trichocarpa using a new approach for reconstruction and curation of metabolic pathways for plants. *Plant physiology*, vol. 153, no. 4, pp. 1479–91. ISSN 1532-2548.

[179] Zhang, W., Aldrich, P., Tacutu, R. and Budovsky, A. (2011). Constructing ecological interaction networks by correlation analysis: hints from community sampling. *Network Biology*, vol. 1, no. 2, pp. 81–98.

[180] Zhang, Z.-M., Chen, S. and Liang, Y.-Z. (2011 January). Peak alignment using wavelet pattern matching and differential evolution. *Talanta*, vol. 83, no. 4, pp. 1108–17. ISSN 1873-3573.

[181] Zhang, Z.-M., Liang, Y.-Z., Lu, H.-M., Tan, B.-B., Xu, X.-N. and Ferro, M. (2012 February). Multiscale peak alignment for chromatographic datasets. *Journal of chromatography. A*, vol. 1223, pp. 93–106. ISSN 1873-3778.

[182] Zhao, W., Sankaran, S., Ibáñez, A.M., Dandekar, A.M. and Davis, C.E. (2009 August). Two-dimensional wavelet analysis based classification of gas chromatogram differential mobility spectrometry signals. *Analytica chimica acta*, vol. 647, no. 1, pp. 46–53. ISSN 1873-4324.

[183] Zmasek, C.M. and Eddy, S.R. (2002). RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC bioinformatics*, vol. 3, no. 1, p. 14.

Chapter 3

Chemiomics

Accepted for publication as:

Jacobson D., Monforte, A.R., Silva Ferreira A.C. (2013). Untangling the chemistry of Port wine aging with the use of GC-FID, multivariate statistics and network reconstruction. Journal of Agricultural and Food Chemistry.

3.1 Abstract

Chromatography separates the different components of complex mixtures and generates a fingerprint representing the chemical composition of the sample. The resulting data structure will depend on the characteristics of the detector used, univariate for devices such as a flame ionization detector (FID) or multivariate for mass spectroscopy (MS). This study addresses the potential use of a univariate signal for a non-targeted approach in order to (i) classify samples according to a given process or perturbation, (ii) evaluate the feasibility of developing a screening procedure to select candidates related to the process and (iii) provide insight into the chemical mechanisms which are affected by the perturbation. In order to achieve this it was necessary to use and develop methods for data pre-processing and visualization tools to assist an analytical chemist to view and interpret complex multidimensional data sets.

Dichloromethane Port wine extracts were collected using GC-FID; the chromatograms were then aligned with correlation optimized warping (COW) and subsequently analyzed with multivariate statistics (MVA) by principal component analysis (PCA) and partial least squares regression (PLS-R). Furthermore, wavelets were used for peak calling and alignment refinement and the resulting matrix used to perform kinetic network reconstruction via correlation networks and maximum spanning trees. Network-target correlation projections were used to screen for potential chromatographic regions/peaks related to aging mechanisms. Results from PLS between aligned chromatograms and target molecules showed a high X to Y correlation of 0.91, 092 and 0.89; with 5-hydroxymethylfurfural (HMF) (Maillard), acetaldehyde (oxidation) and 4,5-

dimethyl-(5H)-3-hydroxy-2-furanone, respectively. The context of the correlation (and therefore likely kinetic) relationships amongst compounds detected by GC-FID and the relationships between target compounds within different regions of the network can be clearly seen.

Keywords: Univariate signal, Aging, Mechanisms, Preprocessing, GC-FID, PCA, PLS, Network Theory, Kinetic Network Reconstruction

3.2 Introduction

Port wine is a fortified wine produced in the Douro region of Portugal. After the vinification process wines are exclusively barrel aged (Tawnys) or matured for two years in a cask and then bottled (Vintage).

The aromatic profile of port wine changes during aging as the result of several underlying mechanisms. Therefore, if one wants to understand or modulate the sensory attributes of Port it is important to understand these mechanisms and the inter-connections amongst them. Several of the mechanisms are to a large extent already described as Maillard [26, 20, 21, 22, 32, 28, 12] or oxidation [40, 6, 43, 11, 23], nevertheless the overlaps between these two mechanisms is not well known.

In Port wines sotolon was recognized as a key molecule in the "perceived age" of barrel storage Port wine and consequently in the aroma quality of the final product. Its concentration can range from a few dozen ug/L in a young wine to 1 mg/L in wines older than 50 years. The odor threshold has been estimated to be 19 ug/L [15, 14].

The Maillard reaction has been suggested by several authors to be responsible for the formation of sotolon as a product of a reaction involving hexoses and pentoses in the presence of cysteine [22] and from the aldol condensation of butane-2,3-dione and hydroxyacetaldehyde [37]. On the other hand, several papers have linked sotolon formation with oxidation [15, 42, 4, 16, 13]. Both oxygen and temperature influence sotolon concentration, which suggests that its origin involves a connection between oxidation and Maillard mechanisms [29].

Therefore, wine aging is a complex system which requires more information to be analysed in order to better understand the mechanisms at play. Given this, techniques that are able to capture information about a broader range of compounds participating in the aging process are necessary in order to achieve a better understanding thereof.

Metabolomics is defined as the study of "as many-small-metabolites-as-possible" in a system [1]. In this paper we attempt to describe an example of chemiomics which we define to be the study of the relationships between as many chemical compounds as possible in a complex chemical (non-enzymatic) system. In order to accomplish this by chemical profiling, two strategies can be employed: (i) "targeted analysis", using a priori knowledge of which com-

pounds to analyse, which requires their identification and manual quantification or (ii) "untargeted analysis" where one tries to detect as many compounds as possible in order to acquire sample fingerprints which will be submitted to multivariate analyses and network analysis for further contextualisation. The identification and quantitation is then performed on the variables that are found to be associated with the principal components/correlation vectors as determined by multivariate and network analysis.

Spectroscopic detectors, such as those based on UV-Vis (ultraviolet-visible), FTIR (Fourier transform infrared) or NMR (nuclear magnetic resonance) spectroscopy, are largely employed to obtain chemical fingerprints which can be used for sample classification as well for chemical quantitation [35, 27, 44, 39, 9, 19, 24, 3, 2, 41]. The extremely convoluted resulting spectra can be further processed with MVA techniques that compensate, to a certain extent, for the absence of structural information in complex chemical mixtures. In spite of the versatility of these detectors the absence of structural information, due to the extremely convoluted signal, constitutes a major drawback in obtaining molecular identifications if there is a need to study complex systems that require kinetic contextualization.

Chromatography has long been used for the separation of molecules enabling both the quantitative and qualitative analysis of constituents in complex mixtures. The separation of the different components of a complex mixture generates a fingerprint representing the chemical composition of the sample. Chromatographic fingerprints taken from samples under different experimental conditions can then be used to explain the changes caused by a perturbation [18]. Due to the separation performed by the column it is possible to identify structures based on the elution time in a given chromatographic profile.

Chromatographic data has a huge number of variables and PCA and PLS are MVA visualization techniques that allow for the interpretation of multidimensional data sets. When multivariate analysis involves large datasets, variable selection processes play an important role as they eliminate the less significant or non-informative variables. The overall aim of any variable selection technique is to capture variables from the original dataset that are most specifically related to the problem of interest and to exclude those variables that are affected by other sources of variation.

PCA is a non-supervised technique that decomposes the original variables of a data set into two matrices: the scores and the loadings matrices. The scores matrix contains information about the samples, which are described in terms of their projection onto the principal components. The loadings matrix contains information about the variables which are also described in terms of their projection onto the principal components. The loadings can also be interpreted as the contribution of the variables for the observed scores distribution.

Consequently, the use of GC fingerprints with MVA should make it possible to extract considerable amounts of information from complex mixtures. The

tandem of GC-MVA is, in our perspective, a middle ground between a rich detector, which provides structural information (NMR), and detectors like FTIR and UV-Vis.

Therefore the aim of this study is to validate the feasibility of using univariate chromatographic data, in particular gas chromatography with a flame ionization detector (GC-FID), as a screening procedure to classify complex chemical mixtures, such as wine samples, to identify which compounds are responsible for differences and to perform network reconstructions that may indicate the underlying kinetic relationships and mechanisms.

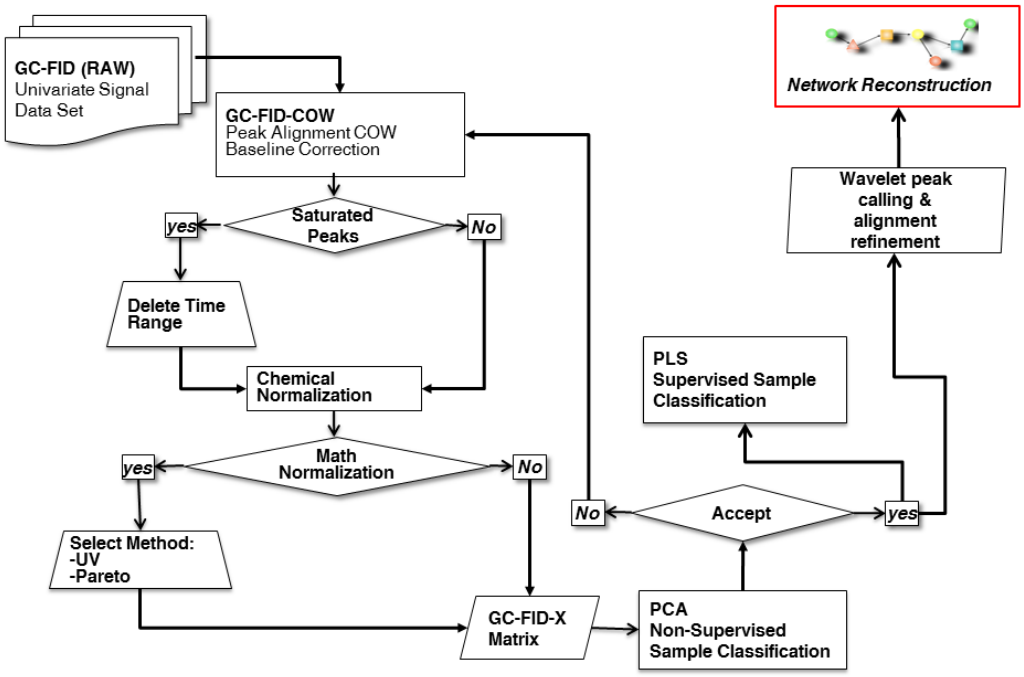


Figure 3.1: Proposed workflow for univariate (chromatographic) signal processing.

3.3 Materials and Methods

3.3.1 Reagents

The chemicals, 3-octanol (97%) 3-hydroxy-4,5-dimethyl-2(5H)-furanone (=99.5%), 5-methylfurfural (=99.5%), 5-hydroxymethylfurfural (=99.5%), acetaldehyde (=99.5%), ethyl lactate (98%), 5-(ethoxymethyl)furfural (=99.5%), acetic acid (=99.5%), 2,3-butanediol (98%), diethyl succinate (=99.5%), 2-phenylethanol

(=99.5%), diethyl malate (>97%), succinic monoethyl ester (=99.5%), benzaldehyde (=99.5%), octanoic acid (=99.5%), hexanoic acid (=99.5%) aspartame (>99%), glutamine (>99%), cysteine (>99%), serine (>99%), glycine (>99%), arginine (>99%), gamma-aminobutyric acid (>99%) alanine (>99%), tyrosine (>99%), valine (>99%), phenylalanine (>99%), leucine (>99%), ornitin (>99%), lysine (>99%), homoserine (>98%), norvaline (>98%), homocysteine (>99%), 2-sulfanylethanol (98%), tetraphenylborate (>99.5%), iodoacetic acid (>99%), o-phtaldialdehyde (>99%), and n-alkanes (C11-C22) were obtained from Sigma-Aldrich, USA. The cis-5-hydroxy-2-methyl-1,3-dioxane (cis-dioxane), cis-4-hydroxymethyl-2-methyl-1,3-dioxolane (cis-dioxolane), trans-4-hydroxymethyl-2-methyl-1,3-dioxolane (trans-dioxolane), trans-5-hydroxy-2-methyl-1,3-dioxane (trans-dioxane) were synthetized according to Maillard [25].

Dicloromethane (HPLC grade) was purchased from LabScan, Sowinskiego, Gliwice. Anhydrous sodium sulfate and methanol (HPLC grade) were obtained from Merck, Darmstadt, Germany.

3.3.2 Port Wine samples

The 37 samples used in this study were between 2 and 60 years of age: one sample for 2, 14, 19, 23, 35, 40, 42, 48, 54, 57 and 60 years of age, two samples for 7 and 20 years of age, three samples for 4 and 5 years of age and ten samples of 10 years of age. All wines were matured in oak barrels. These samples were supplied by the "Instituto do Vinho do Porto e do Douro". The wines were made following standard traditional winemaking procedures for Port wine and have been certified.

3.3.3 Analytical procedure

3.3.3.1 Volatiles Extraction

A liquid-liquid extraction was performed to extract the volatile fraction from each sample. The procedure used was as follows: 5 g of anhydrous sodium sulphate and 50 μL of internal standard (3-octanol) were added to 50 mL of sample and were extracted twice with 5 mL of dicloromethane using a magnetic stir bar for 5 minutes per extraction and 2 mL of the resulting organic phase were concentrated under a nitrogen stream 4 times. The extract was then analysed by GC (Agilent 5980, USA) with FID detection. Two microliters of the extract were injected. Chromatographic conditions were the following; column BP-21 (50 m $\times$ 0,25 mm $\times$ 0,25 μm) fused silica (SGE, Portugal); hydrogen (5.0, Air-liquide, Portugal); 1.2 mL/min flow rate; injector temperature, 220 °C; oven temperature, 40 °C for 1 min programmed at a rate of 2 °C/min to 220 °C, maintained during 30 min; splitless time, 0.5 min; split flow, 30 mL/min.

In order to facilitate identification, the Kovat's index for each peak was calculated as described by Van den Dool and Kratz [7]. This determination was performed on polar phase columns, BP21 (50 m × 0.25 mm × 0.25 μm).

3.3.3.2 Amino Acids Analysis

Twenty one amino acids were analysed in the Port wine samples: aspartic acid, glutamic acid, cysteine, asparagine, histidine, serine, glycine, arginine, threonine, alanine, g-aminobutyric acid, tyrosine, ethanolamine, valine, methionine, tryptophan, phenylalanine, isoleucine, leucine, ornithine and lysine. The methodology used was that described by Pipris-Nicolau et al. [33].

3.3.3.3 Acetaldehyde, furanic compounds and sotolon analysis

These analyses were done as described by Silva Ferreira et al. [15].

3.3.3.4 Data pre-processing

The ASCII file of the chromatographic data obtained from each sample was extracted and a matrix created containing all of the chromatograms. The intensities of each elution point in each chromatogram were normalized by dividing each value by the intensity of the internal standard (3-octanol) found in that chromatogram. The raw dataset (GC-FID) was then imported into The Unscrambler™X 10.1 (Camo, Sweden), where the first stage of the alignment of chromatograms was performed using Correlation Optimized Warping (GC-FID-COW). This algorithm aligns chromatograms by means of sectional linear stretching and compression, which shifts the peaks of one chromatogram to correlate with those of the other chromatograms in the dataset [17]. The saturated peaks were then removed and the baseline corrected (GC-FID-COW-saturated removed and baseline correction). The resulting matrix (GC-FID-X) was then used for multivariate data analysis as described in the statistical analysis section.

3.3.3.5 Statistical analysis

The data was analysed with PCA and PLS-R using either Qlucore (Lund, Sweden) or SIMCA-P+ 12.0.1 (Umetrics, Norway). PCA shows similarities between samples projected on a plane and makes it possible to determine which variables determine these similarities and in what way. PLS is used to extract factors related to one or more response values. PLS validation was performed by the cross-validation method.

3.3.3.6 Kinetic Network Reconstruction

In order to attempt to reconstruct the underlying kinetic network, the fingerprint needed to be further compressed to a single value for each putative molecule detected by GC-FID. Thus, each chromatographic peak needed to be replaced with a single value for the intensity and retention-time, at the apex of each peak and therefore a more refined alignment procedure was required. This was achieved as follows: An "average chromatogram" was created by taking the mean of the values at each elution point in the GC-FID-X matrix. The average chromatogram and the sample chromatograms were smoothed with the Savitzky-Golay method (settings: left=15, right=15, polynomial degree=0) [36] as implemented in The Unscrambler™X 10.1 (Camo, Sweden). The wavelet method of Du et al. [10], which was originally developed for peak calling in peptide mass-spec data, was adapted for finding peak centres in chromatographic data and a Mexican hat wavelet used to determine the location of all of the peaks in the average and sample chromatograms. A custom built Perl program was integrated with the R-based wavelet method to achieve this. The distances between the locations of all of the peaks in each sample chromatogram and the locations of the peaks in the average chromatogram were calculated. It was observed that there was a correlation between peak height and the amount of peak centre shift that occurred across chromatograms and we thus devised a two-step process for aligning sample peaks to those of the average chromatogram. If the (internal-standard-normalized) height of the average peak was greater than 2 and the distance to the nearest sample peak was less than 0.3 minutes then the intensity value of the sample peak was assigned as the sample value at the average peak location. However, if the (internal-standard-normalized) height of the average peak was less than 2 and the distance to the nearest sample peak was less than 0.15 minutes then the intensity value of the sample peak was assigned as the sample value at the average peak retention time. This algorithm was implemented in Perl.

As a result of this process a new matrix was created that contained the retention time of all peaks in the average chromatogram and the intensity values of all of the peak centres from each sample aligned to these average retention times. Thus a vector was created for each peak (presumably representing a compound) across all samples. An all-against-all comparison was done calculating the Pearson correlation between each and every peak vector. As such, one is able to track the increase or decrease of compounds (peaks) during the aging process and determine the correlative relationships amongst them. We applied a Pearson correlation threshold of 0.8 and represented the remaining relationships as a mathematical graph in order to form a correlation network with the nodes representing peaks and the edges weighted with the Pearson correlations between the peak vectors. In order to reconstruct the most likely kinetic network underlying the set of chemical reactions involved in the aging process, a maximum spanning tree was created by transforming the edge

weights into inverse correlations (by taking the difference between the number 1 and the absolute correlation values) and the subsequent use of a minimum spanning tree (mst) algorithm [8] on the this inverse correlation network. A minimum spanning tree represents the shortest possible path through a graph and, as such, selects for the smallest inverse correlation (i.e highest correlation) pairs between all nodes in the network. The resulting networks were visualized in Cytoscape [38].

3.4 Results / Discussion

3.4.1 Principal Component Analysis

Our initial goal for the use of PCA was to examine the intrinsic variation in the data set prior to alignment in order to determine if the volatile fraction of the samples followed a trend related to age. However, when using the GC-FID matrix described above some samples did not follow the latent age variable described by PC1, namely the 4 and 60 year old samples (score plot not shown). The analyses global workflow is described in Figure 3.1.

The loading plot in Figure 3.2 shows higher levels of acetic acid, 2,3-butanediol, diethyl succinate, diethyl malate, phenylethanol and succinic mono ethyl ester present in the older samples. The esterification process appears to be the most prevalent reaction amongst the compounds apparent in the loading plot. The organic acids naturally present in grape must, such as malic acid and those present as the result of fermentation, such as lactic, succinic and acetic acids all react with ethanol to yield the esters seen in the loading plot [34]. However, these molecules are out of the detector's linear response range, so they needed to be eliminated. Furthermore, the chromatograms must be aligned because an unavoidable characteristic of all chromatographic data is that the retention times for the peaks in the chromatograms shift slightly from one analysis to another. To address this problem correlation optimized warping (COW) was used to align all of the chromatograms.

A new fingerprint was created (GC-FID-COW), by the removal of saturated peaks and baseline correction to yield the GC-FID-X matrix, which was subsequently analysed by PCA. The new score plot shows the same latent age variable described by PC1, but the explained variance of the first two components is 74% (Figure 3.3) and the samples that did not previously follow the age vector now do so after alignment.

The score-plots in Figure 3.3A and Figure 3.3B show a clear trend related to wine age, suggesting that the chemical mechanisms are correlated with time, across the first principal component with the first two components explaining 74% of the variance. It appears that the latent age vector remains intact whether the data is mathematically normalised or not.

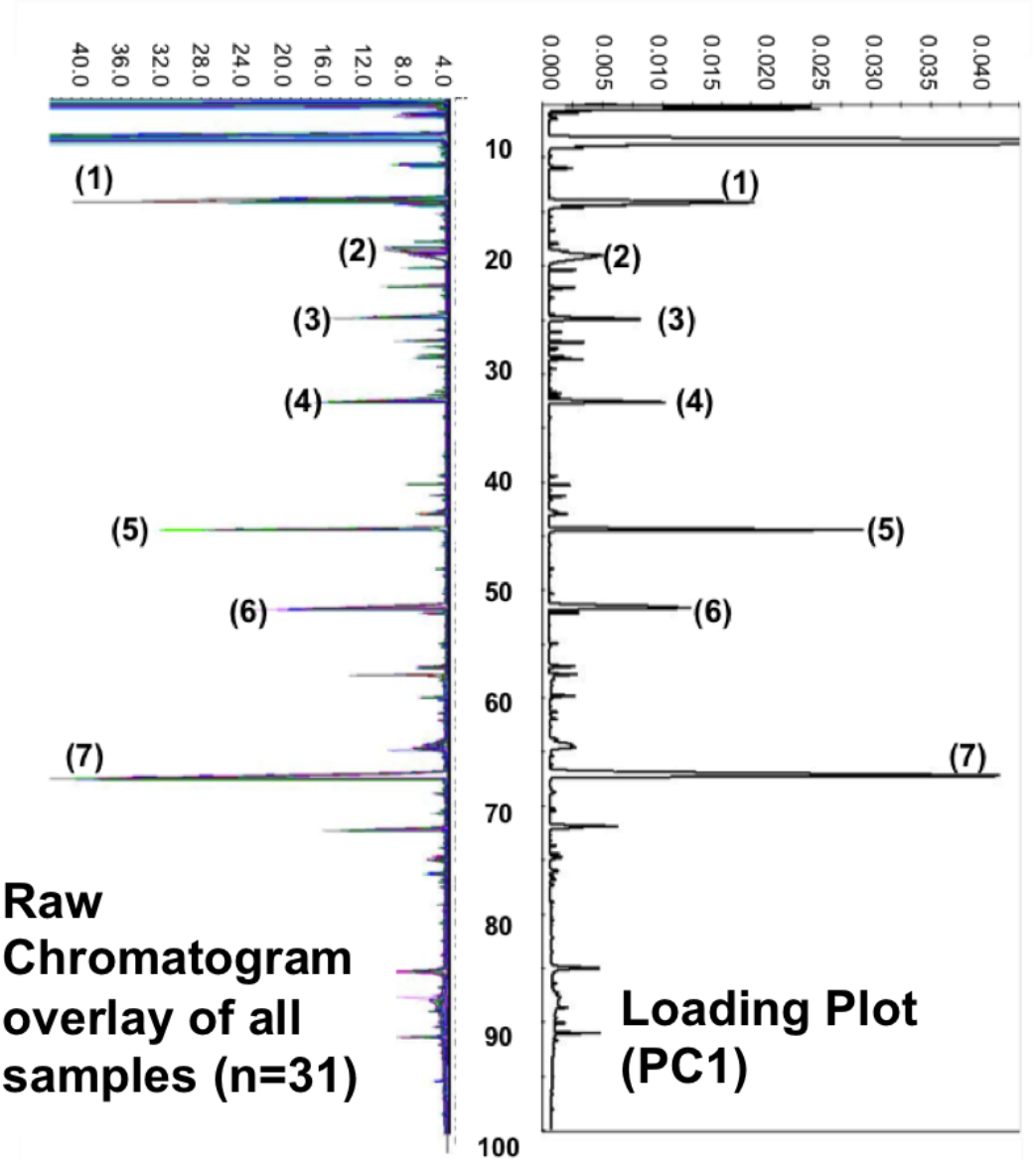


Figure 3.2: Raw chromatogram overlay of all samples (n=31) and Loading plot (PC1) representing the average chromatogram GC-FID chromatogram. (1) ethyl lactate, (2) acetic acid, (3) 2,3-butanediol, (4) diethyl succinate, (5) phenylethanol, (6) diethyl malate and (7) succinic monoethyl ester.

It is important to note that the data for both the PCA and PLS analyses was not mathematically centered or normalized as is commonly done to give all variables equal impact on the model. Centering is usually used as a matter of convenience for display and mathematically has no impact on the multivariate model. When we view the loadings we choose not to center the data in order to not have negative scores along PC1 and thus to have no negative peaks on the loading plot in order to keep the loading plot looking as

much like a normal chromatogram as possible. We are aware that our choice not to normalize means that peaks with higher intensities will have a larger impact on the model and be more apparent in the loading plots shown in Figures 3.2 and 3.3. However, this allows the patterns visible in the loading plots to be recognizable to an analytical chemist and therefore is easily read and interpreted as a normal chromatogram. Mathematical normalization unfortunately makes the standard chromatographic patterns unrecognizable as it rescales every peak to the same amount of variance. In addition, as can be seen from Figure 3.3A and Figure 3.3B, normalization does not change the fact that PC1 comprises the age vector with the biggest affect of normalization changing the sample distribution along PC2, which is likely to represent vintage and vinification technology effects. This suggests that the aging of wine largely overwhelms the differences between wines that are present due to the season they were made in or variations amongst the approaches used to make them (different yeast strains, temperatures, crushing mechanisms, fermentation tanks, barrels used for maturation, etc.). The primary purpose of PCA and PLS in our pipeline is as a graphical user interface for analytical chemists to use as a screening step for univariate data sets. As such, we strove to present the analytical chemist with a multivariate interpretative environment with which they would be as familiar as possible, namely chromatographic fingerprints with which they have large experience. Thus, we feel that, for this part of the analysis, the visual representation of the chromatographic loading plots outweighs the assignment of equal weights to every variable in the model. Furthermore, we address this variable normalization issue with the use of network reconstruction via Pearson correlation networks and maximum spanning trees. In the network analysis, every peak is analyzed and has an equal opportunity to form a part of the network and Pearson correlation includes vector normalization.

During aging there are likely to be several different mechanisms involved, including oxidation and Maillard reactions. In PC1 the samples correlate with the age of the wine, which points out that the overall kinetic system overrides any one specific mechanism. As such, the connections between the mechanisms at play are more relevant to sample classification than the contributions of any individual mechanism.

After alignment, compounds which appear to correlate with port wine aging as found in the loading plots were: cis-dioxane, cis-dioxolane, trans-dioxolane trans-dioxane, octanoic acid and HMF as shown in Figure 3.3C.

The cis and trans dioxane and dioxolane are formed by the condensation reaction between glycerol and acetaldehyde. These molecules were identified in Port wine by Silva Ferreira et al. [5] who noted that they increased with age, and, as such, could be used as age markers for port wine kept under oxidative conditions. Furanic compounds, HMF, 5MF and furfural are thought to be products from the Maillard reaction, formed by the fragmentation or cyclization of 3-deoxyosone, a highly reactive intermediary of the reaction [30].

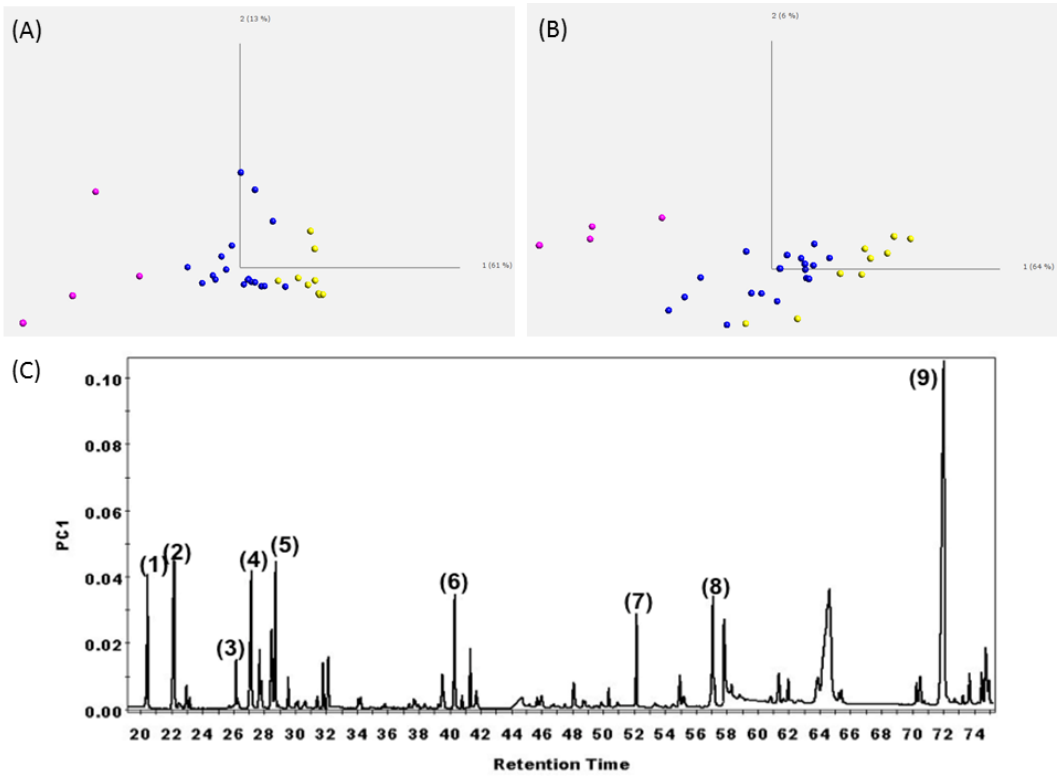


Figure 3.3: PCA score plots of cleaned and COW-aligned chromatograms: (A) un-normalized (B) normalized. Colours denote wines of age 2 to 7 years (yellow), 10 to 42 years (blue) and 48 to 60 years (pink). (C) Loading plot of PC1 with 9 of the peaks identified as (1) furfural, (2) cis dioxane, (3) benzaldehyde, (4) 5MF, (5) cis dioxolane, (6) trans dioxolane, (7) octanoic acid, (8) unknown and (9) HMF.

Barrel oak can also be a source of HMF and furfural [31].

3.4.2 Partial Least Squares Analysis

Partial Least Squares (PLS) analysis was used on the GC-FID-X matrix in an effort to associate specific peaks/fingerprint regions with mechanisms known to be involved in aging. It is worth noting that PLS was not used in its traditional role as a method with which to build calibrated, predictive models (that would therefore be built with training sets and validated with independent test sets). Rather, the goal of our use of PLS-1 was simply as a method with which to perform principal-component-based regressions in an effort to identify sets of peaks that were associated with known mechanistic markers or potential precursors for volatile compounds. Molecules that are thought to be associated with different mechanisms were selected and quantified from each sample and used as markers to try and find other compounds in GC-FID-X

that may be related with the same mechanism. Acetaldehyde and HMF were used as markers for oxidation and the Maillard reaction respectively. Sotolon was also used in an effort to gather more information about its origin. The concentration of each of the marker molecules was determined for each sample and the resulting vectors used as a second data block in PLS.

The resulting PLS coefficient plots show the variables which correlate with each mechanism-marker. Some variables have a positive value, which means that these have kinetic vectors which correlate with that of the mechanism-marker, and some have negative values, which indicate that they have an inverse correlation with the kinetic vector of the mechanism-marker.

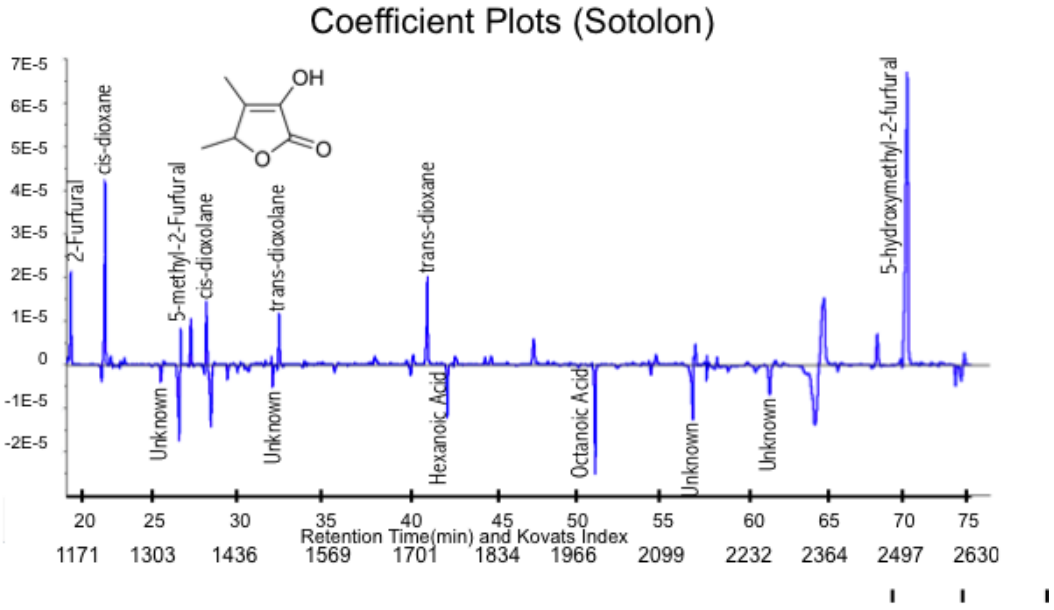


Figure 3.4: PLS *b* coefficients for Sotolon as the Y vector with 7 latent variables.

For sotolon the correlation is 0.89 (over 7 components) for molecules such as furfural, 5-MF, cis dioxane, cis-dioxolane, trans-dioxane, trans-dioxolane and HMF. We also found some organic acids with negative correlations, which indicates that they were being consumed as sotolon was being formed (Figure 3.4). The model had a correlation of 0.91 for HMF (a Maillard reaction marker) and 0.92 to acetaldehyde (an oxidation marker) for 7 latent variables. The PLS loading plot for sotolon was very similar to those seen for HMF and acetaldehyde which means that the mechanisms are correlated, and during aging contribute in the same way to the dynamics of the overall process.

Some amino acids, namely, valine, alanine, arginine, glutamine and aspartate, had relatively high (0.69-0.83) inverse correlations with a number of

peaks in the volatile profile which were themselves correlated to Maillard reaction markers such as such as HMF and furfural. Thus it seems likely that these amino acids are major Maillard aroma precursors.

3.4.3 Network Reconstruction

Figure 3.5 shows the maximum spanning tree derived from the correlation network between all peaks. Each node represents the center of a peak (Kovats index) and each edge represents the best correlation between the peaks. Fold changes between 2 year old and 60 year old wines were calculated for each peak and the nodes colored accordingly with shades of red representing increasing concentration and blue representing decreasing concentration. The thickness of each edge has been scaled to represent the level of correlation (thicker lines mean higher correlation values). Furthermore, the size of each node has been scaled to represent the number of correlations it had with other peaks above a threshold of 0.8.

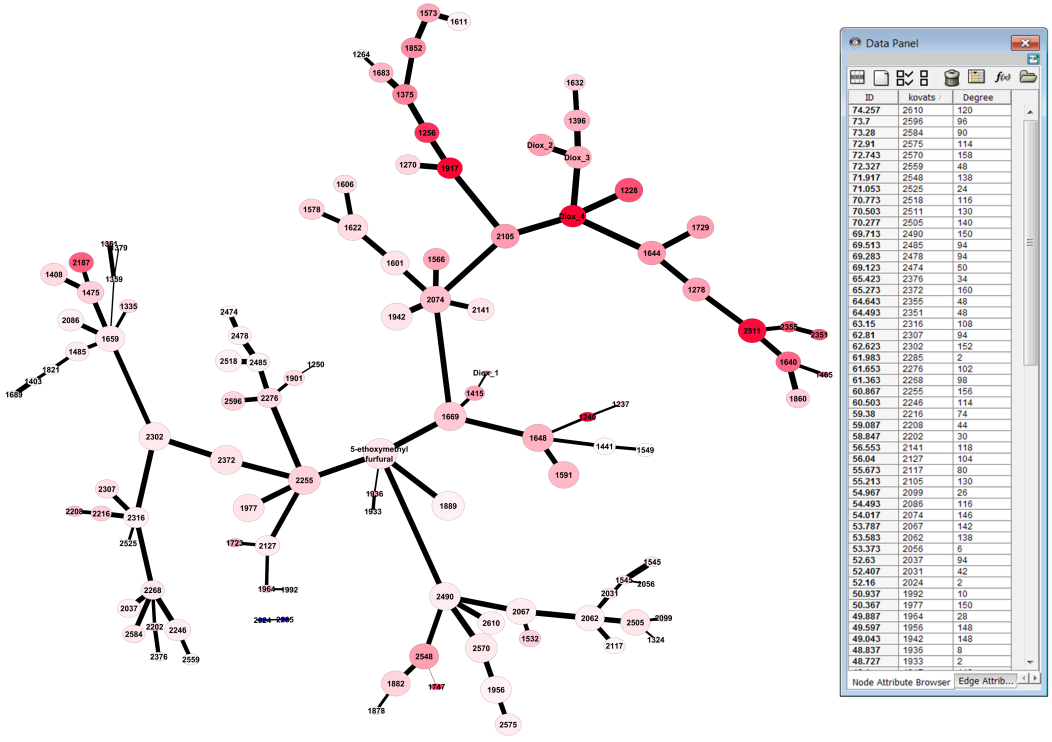


Figure 3.5: Putative Kinetic Network. Nodes are coloured in shades of red based on the fold change from 2 to 60 years. Node sizes are scaled by the number of other nodes (peaks) that are correlated to them above a Pearson threshold of 0.8. Edge thickness is scale by the degree of correlation between its two nodes. (Dioxanes in the network are labelled as follows: cisdioxane: Diox1; cisdioxolane : Diox2-; transdioxolane: Diox3 and transdioxane: Diox4).

Using pure standards as markers for Maillard and oxidation together with the Kovats index it is possible to explore and extract more information from the proposed network. In fact the cyclic acetals of glycerol and ethanal (oxidation products) cluster together on the upper part of Figure 3.5. In addition, 5-ethoxymethylfurfural, an ester of a major Amadori product (HMF), links two major branches of the network. As such, the network illustrates the aging process with the continuous formation of substances absent in young wines, which explains the aging character of wine. The volatile compounds that relate to 5-ethoxymethylfurfural are co-expressed during aging, thus presenting similar kinetics. Future work will focus on their identification using the respective Kovats index and rich information detectors like MS. The network representation captures some of the dynamics of the aging process based on the underlying kinetics. In fact those compounds that are highly correlated to one another (>0.9) are likely to have the same kinetic order and the network thus enables one to screen molecules according to their kinetic parameters.

We propose that this network is an approximate representation of the underlying chemical reaction network during aging. The higher the level of correlation (and therefore the nearer any two peaks are to each other in the network) the higher the probability is that they participate in the same or neighboring reactions. The correlation between compounds drops as you move further away in a chemical reaction network, as the intervening kinetics of each reaction will cause differences at each step. There are no doubt intermediate compounds for some reactions that were not detected by FID. The network is robust to this missing data as the intervening steps will simply be represented by a lower correlation value of an edge between compounds that were detected. This correlative approach of course is not proof of causation but rather serves as a useful tool for hypothesis generation in order to prioritise the identification of unknown compounds represented by the peaks.

By using correlation values to targeted compounds (or other variables such as age) we found that we could highlight the regions of the network that are closely associated with them and therefore likely involved in their formation or consumption. In order to explore regions of the network that may be related to age and particular mechanisms, Pearson correlation values between the peaks and each of the target vectors (sample age, sotolon, acetaldehyde, HMF, glutamate and alanine) were loaded into Cytoscape as node annotations. Alanine and glutamate were selected as target vectors because they were the amino acids best correlated with the GC-FID-X matrix. By sorting the nodes by correlation values and selecting the nodes corresponding to a correlation value with a target above some threshold, portions of the network which correlated with each target vector could be visualized in aqua as shown in Figure 3.6.

Figure 3.6A shows the nodes with a 0.86 Pearson correlation to the age of the wines. There is a clearly defined subnetwork that corresponds to age and represents compounds involved in the aging process. It was clear in the PCA diagrams that there are a group of compounds that correspond to aging

which are responsible for the first principal component. It is likely that the compounds responsible for the second principle component are due to differences between the vintages of the starting wines. We can see the same pattern in this network where there are a number of compounds that do not correlate well with age and are likely reflecting vintage differences amongst the wines.

Figures 3.6 B, C and D show the regions of the network (colored as aqua) that correlate (Pearson threshold 0.86) with sotolon, HMF and acetaldehyde, respectively. It is clear that there is considerable overlap between the subnetworks correlated with these three compounds and, as such, it is possible that there is a mixture of oxidation and maillard reactions at play in forming these compounds. The subnetwork that correlates with age at a Pearson threshold of 0.86 in Figure 3.6A clearly overlaps to a very large degree with the subnetworks defined by the correlations to acetaldehyde, HMF and sotolon. The arrow in Figure 3.6A shows the node that negatively correlates to both alanine and glutamate (-0.8 Pearson threshold) and as such likely represents the entry point to the volatile network. The fact that the anti-correlation is relatively low (-0.8 to -0.83) probably indicates that there are one to several intermediates between the amino acids and their products' entry into the volatile network.

We believe that we have demonstrated that the approach described here using GC-FID univariate data, when used as sample fingerprints, can be used to classify the age of Port wine and to predict potential molecules involved in this process by the deconvolution of time and the kinetics of different aging mechanisms. We began this analysis with the hope that multivariate analysis and network reconstruction would be useful tools with which to study mechanisms related to a perturbation. The PCA score plots allow for sample classification and the visualization of larger peaks that correspond with aging. The PLS loading plots provide the analytical chemist with a familiar set of patterns, namely virtual chromatograms which point out the larger peaks that appear to be associated with marker compounds for oxidation and Maillard mechanisms. The network reconstruction is very useful in visualizing the relationships between all the compounds detected via GC-FID and their changes in concentration over time. This view of the data should provide considerably more information in an effort to understand the probable kinetic contexts of the molecules represented by peaks in each chromatogram. Furthermore, it is possible to identify regions of the network that appear to be involved in the formation or consumption of target compounds. As such, the approach described here should indeed be a very powerful tool for the further study of mechanisms and kinetic networks in complex mixtures.

In conclusion, univariate chromatographic signals are less expensive compared to NMR or MS and therefore constitute a valuable tool in a bio-analytical pipeline. However, the use this type of data in an untargeted approach required the development and use of new data processing methods and graphical user interfaces in order to extract the maximum amount of information from the data. The approach reported here enables us to: 1) minimize the cost of analy-

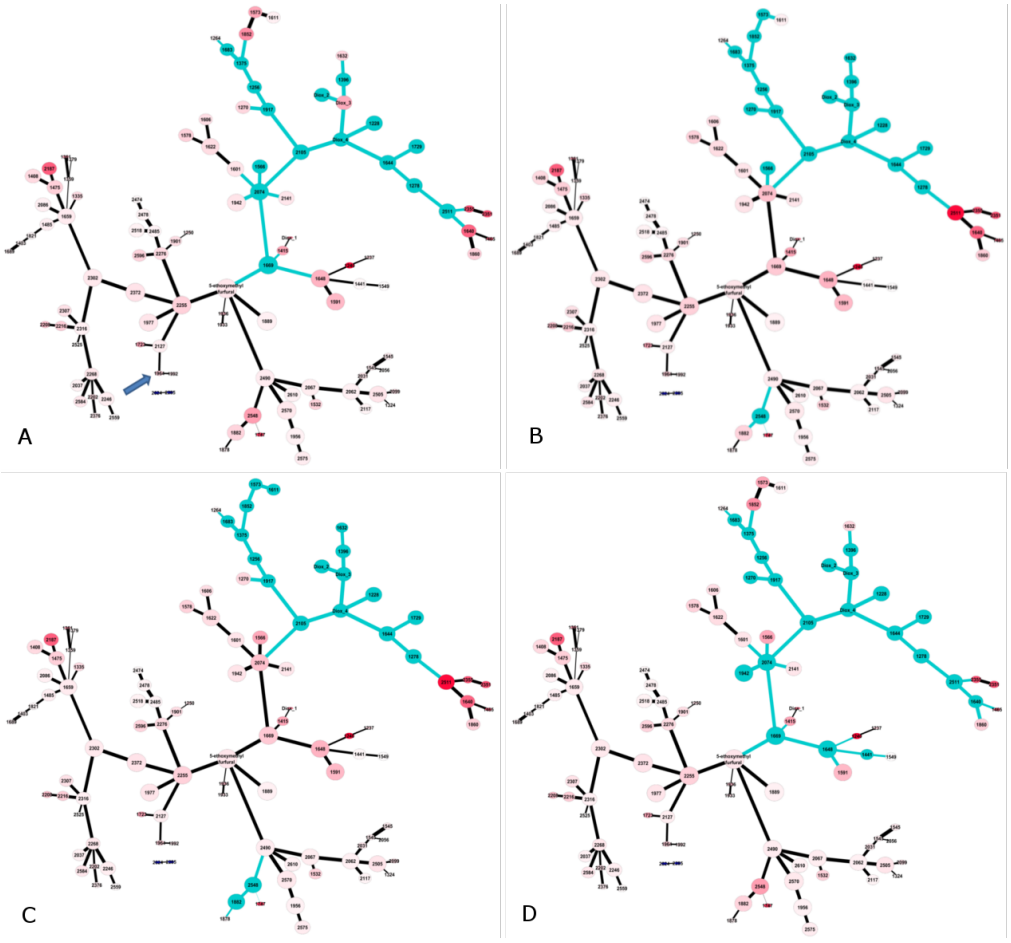


Figure 3.6: Subnetworks correlating to A) Age, B) Sotolon, C) HMF and D) Acetaldehyde. Nodes (compounds) with strong Pearson correlations to these target vectors are colored with aqua.

sis; 2) perform sample classification and contextualization; 3) perform process monitoring of aging or other time series; 4) consider the prospect of building databases from the large amounts of univariate data already available; 5) screen for correlations with known mechanism markers; 6) select biomarkers for identification and further study and 7) explore the putative kinetic network for a greater understanding of the process being studied.

3.4.4 Acknowledgments

The authors would like to thank Piet Jones, Guy Emerton and Debbie Weighill for useful discussions about the networks presented in this paper.

This research was funded by the project "Wine Metrics: Revealing the Volatile Molecular Feature Responsible for the Wine Like Aroma a Critical Task Toward the Wine Quality Definition." (PTDC/AGR-ALI/121062/2010), partially supported by ESB/UCP plurianual funds through the POS-Conhecimento

Program that includes FEDER funds through the program COMPETE (Programa Operacional Factores de Competitividade) by Portuguese national funds through FCT (Fundacao para a Ciencia e a Tecnologia), Winetech, the Technology and Human Resources Programme and the South African National Science Foundation.

3.5 Author Contributions

Performed the experiments: ARM, ACSF. Analyzed the data: DJ, ARM, ACSF. *Multivariate Analysis:* DJ, ACSF. *Wavelet-based alignment method development:* DJ. *All Network-based analysis:* DJ, *Wrote the paper:* DJ, ARM, ACSF.

References

[1] Cevallos-Cevallos, J.M., Reyes-De-Corcuera, J.I., Etxeberria, E., Danyluk, M.D. and Rodrick, G.E. (2009). Metabolomic analysis in food science: a review. *Trends in Food Science & Technology*, vol. 20, no. 11-12, pp. 557–566. ISSN 0924-2244.

[2] Consonni, R., Cagliani, L.R., Guantieri, V. and Simonato, B. (2011). Identification of metabolic content of selected Amarone wine. *Food Chemistry*, vol. 129, no. 2, pp. 693–699.

[3] Cuadros-Inostroza, A., Giavalisco, P., Hummel, J., Eckardt, A., Willmitzer, L. and PenÌČa-CorteÌĄs, H. (2010). Discrimination of wine attributes by metabolome analysis. *Analytical chemistry*, vol. 82, no. 9, pp. 3573–3580.

[4] Cutzach, I., Chatonnet, P. and Dubourdieu, D. (1998). Role du sotolon dans l'arome des vins doux naturels. Influence des conditions d'elevage et de vieillissement. *J. Int. Sci. Vigne Vin*, vol. 32, no. 4, p. 223.

[5] da Silva Ferreira, A.C., Barbe, J.C. and Bertrand, A. (2002). Heterocyclic acetals from glycerol and acetaldehyde in port wines: evolution with aging. *Journal of agricultural and food chemistry*, vol. 50, no. 9, pp. 2560–2564.

[6] Danilewicz, J.C. (2003). Review of reaction mechanisms of oxygen and proposed intermediate reduction products in wine: Central role of iron and copper. *American journal of enology and viticulture*, vol. 54, no. 2, pp. 73–85.

[7] den Dool, H. and Dec Kratz, P. (1963). A generalization of the retention index system including linear temperature programmed gas-liquid partition chromatography. *Journal of Chromatography A*, vol. 11, pp. 463–471.

[8] Dijkstra, E.W. (1959 December). A note on two problems in connexion with graphs. *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271. ISSN 0029-599X.

[9] Ding, Y., Wu, E.Q., Liang, C., Chen, J., Tran, M.N., Hong, C.H., Jang, Y., Park, K.L., Bae, K.H., Kim, Y.H. and Others (2011). Discrimination of cinnamon bark and cinnamon twig samples sourced from various countries using HPLC-based fingerprint analysis. *Food Chemistry*, vol. 127, no. 2, pp. 755–760.

[10] Du, P., Kibbe, W.A. and Lin, S.M. (2006). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, vol. 22, no. 17, pp. 2059–2065.

[11] Du Toit, W.J., Marais, J., Pretorius, I.S. and Du Toit, M. (2006). Oxygen in must and wine: A review. *S. Afr. J. Enol. Vitic*, vol. 27, no. 1, p. 76.

[12] e Silva, H., de Pinho, P., Machado, B.P., Hogg, T., Marques, J.C., Câmara, J.S., Albuquerque, F. and Silva Ferreira, A.C. (2008). Impact of forced-aging process on Madeira wine flavor. *Journal of agricultural and food chemistry*, vol. 56, no. 24, pp. 11989–11996.

[13] Escudero, A., Cacho, J. and Ferreira, V. (2000). Isolation and identification of odorants generated in wine during its oxidation: A gas chromatography–olfactometric study. *European Food Research and Technology*, vol. 211, no. 2, pp. 105–110.

[14] Ferreira, A., Avila, I. and Guedes de Pinho, P. (2005). Sensorial Impact of Sotolon as the" Perceived Age" of Aged Port Wine. In: *ACS symposium series*, vol. 908, pp. 141–159. ACS Publications.

[15] Ferreira, A.C.S., Barbe, J.C. and Bertrand, A. (2003). 3-Hydroxy-4, 5-dimethyl-2 (5 H)-furanone: a key odorant of the typical aroma of oxidative aged port wine. *Journal of agricultural and food chemistry*, vol. 51, no. 15, pp. 4356–4363.

[16] Ferreira, A.C.S., Hogg, T. and de Pinho, P.G. (2003). Identification of key odorants related to the typical aroma of oxidation-spoiled white wines. *Journal of agricultural and food chemistry*, vol. 51, no. 5, pp. 1377–1381.

[17] Gong, F., Liang, Y.Z., Fung, Y.S. and Chau, F.T. (2004). Correction of retention time shifts for chromatographic fingerprints of herbal medicines. *Journal of Chromatography A*, vol. 1029, no. 1, pp. 173–183.

[18] Gong, F., Wang, B.T., Chau, F.T. and Liang, Y.Z. (2005). Data preprocessing for chromatographic fingerprint of herbal medicine with chemometric approaches. *Analytical letters*, vol. 38, no. 14, pp. 2475–2492.

[19] Gu, H. (2009). *NMR and MS-based metabolomics/ Development and applications*. Ph.D. thesis, Purdue University.

[20] Hodge, J. (1953). Dehydrated foods, chemistry of browning reactions in model systems. *Journal of Agricultural and Food Chemistry*, vol. 1, no. 15, pp. 928—-943.

[21] Hofmann, T. and Schieberle, P. (1995). Evaluation of the key odorants in a thermally treated solution of ribose and cysteine by aroma extract dilution techniques. *Journal of agricultural and food chemistry*, vol. 43, no. 8, pp. 2187–2194.

[22] Hofmann, T. and Schieberle, P. (1997). Identification of potent aroma compounds in thermally treated mixtures of glucose/cysteine and rhamnose/cysteine using aroma extract dilution techniques. *Journal of agricultural and food chemistry*, vol. 45, no. 3, pp. 898–906.

[23] Kilmartin, P.A. (2009). The oxidation of red and white wines and its impact on wine aroma. *Chemistry in New Zealand*, vol. 73, no. 2, pp. 79–83.

[24] Lee, J.E., Hwang, G.S., Van Den Berg, F., Lee, C.H. and Hong, Y.S. (2009). Evidence of vintage effects on grape wines using NMR-based metabolomic study. *Analytica chimica acta*, vol. 648, no. 1, pp. 71–76.

[25] Maillard, B. (1971). *Additions radicalaires de diols et de leurs derives diesters et acetals cycliques*. Ph.D. thesis, Universite de Bordeaux.

[26] Maillard, L.C. (1912). Action des acides amines sur les sucres formation des melanoidines par voie methodique. *Council of Royal Academy Science*, vol. 2, no. 154, pp. 66–68.

[27] Manuel, A., Gonçalves, F., Barros, A.S. and Delgadillo, I. (2002). Fourier transform infrared spectroscopy and chemometric analysis of white wine polysaccharide extracts. *Journal of agricultural and food chemistry*, vol. 50, no. 12, pp. 3405–3411.

[28] Marchand, S., de Revel, G. and Bertrand, A. (2000). Approaches to wine aroma: release of aroma compounds from reactions between cysteine and carbonyl compounds in wine. *Journal of agricultural and food chemistry*, vol. 48, no. 10, pp. 4890–4895.

[29] Martins, R., Lopes, V. and Ferreira, A.C. (2009). Port Wine Oxidation Management: A ChemoInformatics Approach.

[30] Martins, S., Jongen, W.M.F. and Van Boekel, M. (2000). A review of Maillard reaction in food and implications to kinetic modelling. *Trends in food science & technology*, vol. 11, no. 9, pp. 364–373.

[31] Moutounet, M., Rabier, P., Puech, J.L., Verette, E. and Barillere, J.M. (1989). Analysis by HPLC of extractable substances in oak wood. Application to a Chardonnay wine. *Sci. Aliments*, vol. 9, no. 1, pp. 35–51.

[32] Pripis-Nicolau, L., De Revel, G., Bertrand, A. and Maujean, A. (2000). Formation of flavor components by the reaction of amino acid and carbonyl compounds in mild conditions. *Journal of agricultural and food chemistry*, vol. 48, no. 9, pp. 3761–3766.

[33] Pripis-Nicolau, L., de Revel, G., Marchand, S., Beloqui, A.A. and Bertrand, A. (2001). Automated HPLC method for the measurement of free amino acids including cysteine in musts and wines; first applications. *Journal of the Science of Food and Agriculture*, vol. 81, no. 8, pp. 731–738.

[34] Ribéreau-Gayon, P., Dubourdieu, D., Donèche, B., Lonvaud, A., Glories, Y. and Maujean, A. (2000). Handbook of Enology, Vol. 1: The Microbiology of Wine and Vinifications; Vol. 2: The Chemistry of Wine Stabilization and Treatments.

[35] Rohman, A. and Man, Y.B. (2010). Fourier transform infrared (FTIR) spectroscopy for analysis of extra virgin olive oil adulterated with palm oil. *Food Research International*, vol. 43, no. 3, pp. 886–892.

[36] Savitzky, A. and Golay, M.J.E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, vol. 36, no. 8, pp. 1627–1639.

[37] Schieberle, P. and Hofmann, T. (1996). Identification of the key odorants in processed ribose-cysteine Maillard mixtures by instrumental analysis and sensory studies. *Special Publications of the Royal Society of Chemistry*, vol. 197, pp. 175–181.

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

[39] Shen, D., Wu, Q., Sciarappa, W.J. and Simon, J.E. (2012). Chromatographic fingerprints and quantitative analysis of isoflavones in Tofu-type soybeans. *Food Chemistry*, vol. 130, no. 4, pp. 1003–1009.

[40] Singleton, V.L. (1987). Oxygen with phenols and related reactions in musts, wines, and model systems: Observations and practical implications. *American Journal of Enology and Viticulture*, vol. 38, no. 1, pp. 69–77.

[41] Son, H.S., Hwang, G.S., Ahn, H.J., Park, W.M., Lee, C.H. and Hong, Y.S. (2009). Characterization of wines from grape varieties through multivariate statistical analysis of NMR spectroscopic data. *Food Research International*, vol. 42, no. 10, pp. 1483–1491.

[42] Thuy, P.T., Elisabeth, G., Pascal, S. and Claudine, C. (1995). Optimal Conditions for the Formation of Sotolon from. alpha.-Ketobutyric Acid in the French" Vin Jaune". *Journal of agricultural and food chemistry*, vol. 43, no. 10, pp. 2616–2619.

[43] Waterhouse, A.L. and Laurie, V.F. (2006). Oxidation of wine phenolics: A critical evaluation and hypotheses. *American Journal of Enology and Viticulture*, vol. 57, no. 3, pp. 306–313.

[44] Zhang, J., Cui, M., He, Y., Yu, H. and Guo, D. (2005). Chemical fingerprint and metabolic fingerprint analysis of Danshen injection by HPLC–UV and HPLC–MS methods. *Journal of pharmaceutical and biomedical analysis*, vol. 36, no. 5, pp. 1029–1035.

Chapter 4

GSA-PCA

Published as:
Jacobson, D., Emerton, G. (2012). GSA-PCA: gene set generation by principal component analysis of the Laplacian matrix of a metabolic network. BMC Bioinformatics, 13(1), 197.

4.1 Abstract

Background Gene Set Analysis (GSA) has proven to be a useful approach to microarray analysis. However, most of the method development for GSA has focused on the statistical tests to be used rather than on the generation of sets that will be tested. Existing methods of set generation are often overly simplistic. The creation of sets from individual pathways (in isolation) is a poor reflection of the complexity of the underlying metabolic network. We have developed a novel approach to set generation via the use of Principal Component Analysis of the Laplacian matrix of a metabolic network. We have analysed a relatively simple dataset to show the difference in results between our method and the current state-of-the-art pathway-based sets.
Results The sets generated with this method are semi-exhaustive and capture much of the topological complexity of the metabolic network. The semi-exhaustive nature of this method has also allowed us to design a hypergeometric enrichment test to determine which genes are likely responsible for set significance. We show that our method finds significant aspects of biology that would be missed (i.e. false negatives) and addresses the false positive rates found with the use of simple pathway-based sets.
Conclusions The set generation step for GSA is often neglected but is a crucial part of the analysis as it defines the full context for the analysis. As such, set generation methods should be robust and yield as complete a representation of the extant biological knowledge as possible. The method reported here achieves this goal and is demonstrably superior to previous set analysis methods.

4.2 Background

Gene Set Analysis (GSA) has proven to be a useful approach to microarray analysis. The underlying principle of GSA is that aggregate scores are assigned to each Gene Set based on all the individual gene scores within that set. There have been several different methods proposed to assign scores to gene sets [21, 19, 20, 10, 11, 17, 25, 26]. Of the approaches published to date, Gene Set Enrichment Analysis (GSEA) [20] [25] seems to have become the most commonly used. Of issue though is the fact that GSEA is based on a modified Kolmogorov-Smirnov test. This test can exhibit a lack of sensitivity; is difficult to employ in practical use, and requires at least 1000 permutations to be run. However, it has recently been found [15] that a one-sample Z-test can be very effective with gene sets for detecting shifts from the mean (sets that collectively show up or down regulation of their constituent genes). Unfortunately, this will not identify gene sets that have a balance of both up and down regulated genes as there will not be the requisite shift from the mean but in statistical terms is rather a change in scale. However, a chi-squared test can be used to good effect to detect such changes in scale and thus find gene sets that exhibit a mixture of up and down regulation [15]. Furthermore, Irizarry et al. [15] have shown that the use of a combination of the computationally simple and rapid Z-test and chi-squared methods outperform GSEA. Dinu et al. [6] have extended the Significance Analysis of Microarrays to Gene Set Analysis (SAM-GS). Of further interest is the method described by Efron and Tibshirani [7] which uses a max-mean statistic to target gene sets with only a fraction of the genes differentially expressed and the approach of Falcon and Gentleman [8] which takes into account the fact that overlap exists between different gene sets. A good review of the various statistical approaches has been written by Goeman and Bühlmann [9].

4.2.1 Gene set generation

Given the discussion above it is clear that considerable effort has been made to apply different statistical methods to GSA, however all of the methods are highly dependent on the very first step: the predefinition of the sets of genes to be analysed. The theoretical combinatorial space for gene sets is quite large and is defined by the binomial distribution of the number of genes in the genome and the size of the desired set:

$$\binom{genes}{setsize} = \frac{genes!}{(genes - setsize)!} \tag{4.2.1}$$

Thus, if one wanted to create all of the possible unique sets with 8 members for the $\sim$ 6000 genes present in the yeast genome, there would be *(6000 choose 8)* $= 4.1 \times 10^{25}$ sets. This is clearly an infeasible number of sets to

generate, much less evaluate. Instead, methods to date have used extant biological knowledge to generate relatively small numbers of sets to be evaluated. One of the common approaches taken for set generation is to simply place the genes involved in a specific pathway into a set. This approach suffers from the fact that pathways are merely human abstractions that are useful for visualisation and interpretation, as they can serve as mnemonic devices for areas of metabolism. However, in isolation, single pathway sets do not reflect the continuously connected nature of biological networks. The metabolic network of *Saccharomyces cerevisiae* (and the location of some of the gene sets found by PCA) can be seen in Figure 4.1. It is clear that this is a complex, interconnected network and, as such, any attempt to use simple pathway representations of it will inevitably be an incomplete representation of the underlying network. We therefore propose that many "pathway sets" are by definition rather arbitrary and incomplete and gene expression patterns may therefore be potentially missed due to improper/incomplete set generation. We suggest that a method that semi-exhaustively partitions the network into overlapping sets would be a better approach to set generation. In order to achieve this we have devised two algorithms that use the Principal Component Analysis of a Laplacian matrix of a metabolic network to do gene set generation. We have also devised a hypergeomtric test to determine which of the genes in the sets identified by gene set analysis are likely to be driving set selection. We have used the resulting gene sets to analyse a publicly available microarray dataset and compare the results obtained from our algorithms (and their respective parameters) to each other as well as to results obtained with traditional pathway sets.

It is important to note that our intent in this paper is not to compare GSA methods to standard parametric statistical approaches (such as t-tests) as there is ample literature on GSA to show its usefulness in difficult datasets (in which there are many orthogonal factors at play which can make a dataset difficult to analyse with other approaches). Rather, our intent is to select a relatively simple dataset with which to make the point that the existing state of the art GSA methods that use isolated pathways are very prone to miss significant amounts of the signal (as they are not capturing the entire metabolic set space) and to report insignificant genes as they are simply associated with a set that is deemed to be collectively significant. As such, both the false negative and false positive (due to "passenger" genes) rates of GSA are high when using isolated pathways. We believe that we have shown that we can address these issues with our method. To demonstrate this we have chosen a straightforward, publicly available dataset with a simple perturbation for which this can easily be demonstrated. We have used this dataset to simply highlight the differences in the results generated by our method rather than to do a full blown biological interpretation of the microarray results.

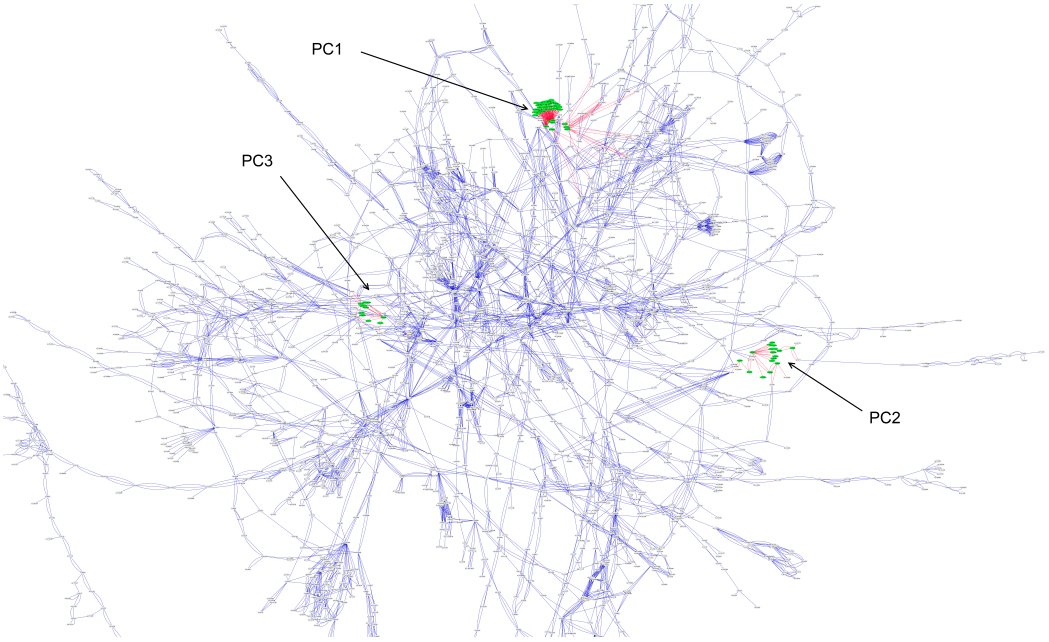


Figure 4.1: Location of the sets of nodes derived from the first three principal components of the Laplacian matrix of the metabolic network, topographically depicted on the metabolic network itself.

4.3 Methods

4.3.1 Affymetrix probeset to yeast genome mapping

The sequences for each of the individual probes of the Affymetrix Yeast 2.0 Genechip were mapped to the Yeast Genome by the use of blastn [2]. A Perl program was written to perform the following tasks: 1) extract 100% identity matches (over the full length of the probe) from the blastn results; 2) assemble the probes into probesets and 3.) model the resultant probeset-to-gene relationships as an Affymetrix-probeset-to-Genome graph.

4.3.2 Single pathway set creation

In order to compare our method to how GSA of pathways has been done in the past it was necessary to create single pathway specific sets. A Perl program was written to parse the XML files downloaded on June 27, 2011 from KEGG [16] and the genes listed in each pathway file were used to create a simple set for each pathway. This is analogous to how pathway sets have been created for GSA previously.

4.3.3 Metabolic network and Laplacian matrix creation

A Perl program was written to parse the XML files downloaded on June 27, 2011 from KEGG [16]. Nodes were created for compounds, reactions and genes, and edges created between genes and the reactions they are involved in as well as between compounds and the reactions they are substrates for or products of. The result of this process can be seen in Figure 4.1 as visualised by Cytoscape [24]. The resulting metabolic network was used for all subsequent set generation. A reference structure within the Perl program which reflected the nodes and edges in the graph was used to identify adjacency and degree parameters for each node and thus generate the corresponding Laplacian matrix.

4.3.4 Principal component analysis

Principal component analysis of the Laplacian matrix was done in R with the prcomp function. Qlucore v2.2 (Lund, Sweden) was also used for PCA model creation and visualisation of principle components during the exploratory phase of algorithm development.

4.3.5 Set theoretic analysis

Set theoretic analysis (intersects, differences, and sizes thereof) of the gene sets and pathway results sets was done in Perl with the use of the Set::Scalar library v 1.25 [13].

4.3.6 Graph theoretic analysis

Graph theoretic analysis was done in Perl with the use of the Graph library v. 0.94 [12].

4.3.7 Threshold-based set creation algorithm

As became clear from examining the PCA score plots, as well as plots of scores across all components, a gene or group of genes may have different scores in different principal components. As such, we decided to create sets at a number of different thresholds to investigate whether this approach would give more or less sensitivity in gene set analysis. Thus for each principal component the positive scores were compared against a series of integer thresholds (1 through 10) and if the score at a principal component was greater than the threshold it was added to a set created for that principal component. A similar procedure was followed for the negative scores at each principal component with the score required to be less than the negation of the integer threshold. Genes in these sets were then mapped to Affymetrix probeset ids with the use of the aforementioned Affymetrix-probeset-to-Genome graph (described elsewhere in

the Methods section), and the matched probes substituted for the genes in the set. Sets that contained more than, or equal to, five probeset ids were kept for further analysis. The resulting sets were subsequently printed out in the .gmt set format used by Efron and Tibshiran (2007) [7]. This algorithm was implemented in Perl.

4.3.8 Step-function-based set creation algorithm

In a separate effort to determine the effects of groups of genes clustering at distinct score ranges within each component on gene set creation and performance an algorithm for set creation was created as follows. An initial empty score-range set was created as an array and held in memory. Scores at each component were sorted in numerical rank order. For positive scores greater than one, neighbouring rank order scores were subtracted from one another and if the difference was < 1 they were added to the existing score-range set, if the difference was ≥ 1 the existing score-range set was closed and a new one (for the next score range) was created and the new gene added to it. This process was repeated across all of the scores in each principle component. A similar procedure was used for the negative scores less than negative one. This algorithm was implemented in Perl.

4.3.9 Gene set analysis with newly generated sets

We used the sets created by the Laplacian PCA method described above for the max-mean method by Efron and Tibshiran (2007) [7]. In order to test the new sets on a dataset that would likely have a limited number of subtle changes on metabolism we selected a dataset that examined the effect of an O-glycosylation inhibitor, OGT2468, on gene expression deposited by Javier Arroyo. They used *Saccharomyces cerevisiae* strain SEY6210 and analysed the global transcriptome in the absence of the OGT2468 (but with the corresponding amount of DMSO, 0.1%) and in the presence of 0.1 uM of OGT2468. They report that "yeast cells exposed to OGT2468 in YPD growth medium show a significant inhibition of mating, filamentation and induction of cell wall compensatory mechanism." The resulting microarray data was downloaded from the Gene Expression Omnibus (series id GSE12193) [1]. Specifically the microarrays used were those for the DMSO control (GSM306567, GSM306565, GSM306569) vs. the cells treated with 0.1 uM of OGT2468 (GSM306573, GSM306577, GSM306581). The microarray data was normalised in R with the RMA method [5] and the resulting log2 transformed values used for GSA. GSA was performed on this data with the following settings: resp. type = "Two class unpaired", nperms = 1000, minsize = 2 and FDR cut = 0.05. The GSA analysis for the largest number of sets (3481 Threshold 1 sets) ran in 2.3 minutes on a single CPU Intel Xeon E5620 2.40 GHz CPU. The genes found in the sets created at PCA score thresholds 1 through 10 which were

determined to be differentially expressed by GSA were tested with the hypergeometric test described below. The genes selected by this test were then used to create genes-only and compounds-containing pathway-centric graphs as described above. These graphs were visualised in Cytoscape. Fold change values were calculated as the ratio of the average expression values for the samples containing 0.1 uM of OGT2468 to the DMSO controls, and used as node attributes to colour the gene nodes in Cytoscape. The negative reciprocal was taken of ratios less than 1.

4.3.10 Hypergeometric enrichment test to determine genes most responsible for set selection

In order to determine which genes identified by GSA on our newly derived sets are drivers we implemented a hypergeometric test. The probability (p) of obtaining any such a set of values is given by the hypergeometric distribution:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} \tag{4.3.1}$$

Where $\binom{n}{k}$ is the binomial distribution, a = Number of times the gene is found in significant sets; b = Number of times all other genes are found in significant sets; c = Number of times the gene is found in non-significant sets; d = Number of times all other genes are found in non-significant sets, and $n = a + b + c + d$. A Perl program was written to parse each of the gene set analysis results as well as the original sets used for the analysis after thresholding, and from these two sources calculate a, b, c, d and n. The two tailed Fisher module of the Text::NSP Perl package [4] was used to test for significance using these values and multiple hypothesis testing corrected for with the Holm-Bonferroni method [14]. Genes with a q-value less than or equal to 0.1 were reported as significantly enriched (i.e. drivers) and included in a pathway-centric network reconstruction for visualisation and interpretation.

4.3.11 Pathway-centric network reconstruction for visualisation and interpretation

As was mentioned in the introduction, pathways are really human abstractions of subgraphs of a metabolic network that are particularly useful as mnemonic devices for contextual visualisation. Unfortunately, visualising the network with all of the reaction, compound, gene and pathway nodes present is overwhelming and as such difficult to interpret. If a biologist can see that the genes selected are part of a well-known pathway it helps them to interpret what part of the metabolic network they are examining. By evaluating the results in this linked metabolic context, one is able to see relationships between areas of metabolism that simply would not be apparent by looking at lists of genes or

lists of pathways. Thus, we have created two types of visualisations in order to better show the metabolic context of the results. The first visualisation just contains the significant genes and the pathways that they are associated with. This is a useful and relatively simple way to visualise the results. However, there are cases where related pathways are affected but do not show up as connected without the compounds being included in the network. Inclusion of all of the compounds leads to a very complex figure so we have chosen to simply create a single edge between pathways if they share one or more compounds.

Raw KEGG XML files actually pre-group all genes, reactions and compounds into these contextualised pathways. In order to provide these sorts of visualisation cues a pathway-centric view of the metabolic network was therefore constructed as follows: A Perl program was written to parse the KEGG XML files such that a node was created for each pathway as defined in KEGG. An edge was created between the pathway node and each gene or compound that is associated with that pathway. A gene determined to be significant by GSA and the subsequent hypergeometric test was used as a seed for a breadth first search of the pathway-centric graph with a radius of one. This was done iteratively for each gene and the union taken of the resulting subgraphs (for examples of outputs see Figure 4.2). A similar procedure was followed for the compound-linked-pathways view with the additional step of the creation of a single edge between pathways that shared one or more compound.

4.3.12 Genes-only and compound-containing graphs

Two different graphs were then created, one with and one without compounds. For the genes-only graph, compounds and their associated edges were simply removed from the graph. The compound-containing graph was constructed by removing all compounds with a degree less than two, such that compounds only served to link pathway nodes together. The genes-only view of the graph is easier to visualise and interpret whereas the graph containing pathway-linking compounds was useful in showing the connectedness (or lack thereof) of the subgraphs containing differentially expressed genes.

4.3.13 Gene ontology and funcat enrichment analysis

Genes that were found by the threshold 1 method but not by pathway GSA were checked for GO Enrichment by GOEAST [27] and FunCat Enrichment at MIPS [22].

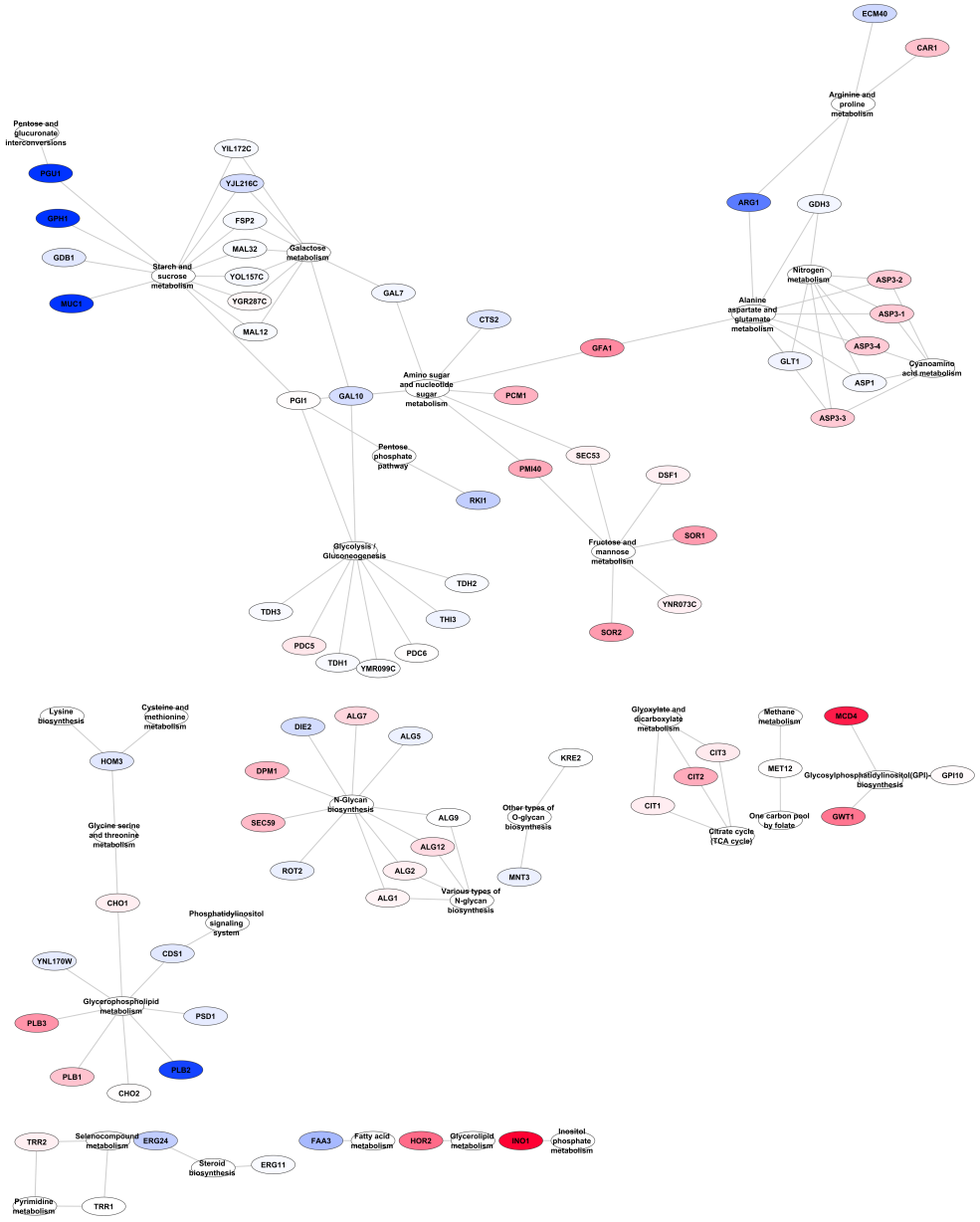


Figure 4.2: Genes from PCA scores threshold = 1 derived sets found to be significantly differentially expressed in their pathway-centric context. Nodes are coloured different intensities of blue (decrease) or red (increase) based on the fold change between treatment and control.

4.4 Results and discussion

4.4.1 Graph representation of a metabolic network

Mathematically, a metabolic network can be represented as a graph, $G = \{V, E\}$, where V is a set of n nodes and E a set of e edges (connections)

between nodes. Let $A(G) = A$ be the adjacency matrix of G such that each element A_{ij} is assigned a value of one if the corresponding nodes are adjacent and zero if they are not. The graph can be further described by a Laplacian transform of the adjacency matrix. The Laplacian matrix $L := (l_{i,j})_{n \times n}$ is defined as:

$$l_{i,j} := \begin{cases} deg(v(i)) & \text{if } i = j, \\ -1 & \text{if } i \neq j \text{ and } v(i) \text{ is adjacent to } v(j), \\ 0 & \text{otherwise} \end{cases} \tag{4.4.1}$$

where $deg(v(i))$ denotes the degree of $v(i)$, i.e. the number of edges incident to v(i). Thus, the Laplacian Matrix is the difference between the diagonal Degree Matrix (D) and the Adjacency Matrix (A).

$$L = D - A \tag{4.4.2}$$

A number of graph theoretic properties of a graph can be derived from its Laplacian matrix and the eigenvector and eigenvalues thereof, including the number of connected components in the graph; its algebraic connectivity (Fielder value); its spectral gap, etc. In fact, the PCA of a graph and spectral graph clustering have been linked previously by Saerens et al. [23].

4.4.2 Principle component analysis

Given a matrix, one can use multivariate statistical methods such a Principal Component Analysis (PCA) to try and find correlative relationships amongst the vectors. PCA is a bilinear modelling method which gives a visually interpretable overview of the most salient information in large, multidimensional datasets. By plotting the principal components it is possible to view statistical relationships between different variables in complex datasets and detect and interpret object groupings, similarities or differences, as well as the relationships between the different variables [18]. As described in the Methods section a graph was created from the KEGG database [16] and a Laplacian matrix derived from said graph. The Laplacian matrix produced, while not a typical object-variable data matrix, may still be analysed with multivariate methods. It was hypothesised that a principal component model would enumerate groups of nodes within the graph with similar topological structure, with the thought that similar columns within the Laplacian matrix would explain a certain amount of 'variance' in the matrix. Accordingly, we then performed PCA on the Laplacian matrix with the hopes of finding an exhaustive set of structures within the graph that could be used for gene set generation.

4.4.3 Gene set generation with score thresholds and step-functions

It was observed that the scores for each principal component often generated discontinuous clusters of objects. Two algorithms for set generation were therefore developed: one based on a step function that took the score discontinuities into account; and another that specified several predetermined thresholds. The sensitivity levels of both these methods were subsequently compared. These algorithms were used to generate gene sets from the genes found in each of the 2656 principal components (see Methods).

4.4.4 Gene set analysis with new gene sets

We used the sets created by the Laplacian PCA method described in the Methods section to perform the max-mean method of GSA [7]. In order to test the new sets on a dataset that would likely have a limited number of subtle changes on metabolism, we selected a dataset that examined the effect of an O-glycosylation inhibitor. The analysis was done as described in the Methods section and the results described below.

4.4.5 Hypergeometric enrichment test to determine genes most responsible for set selection

One of the difficulties faced in Gene Set Analysis is that it is unclear which of the genes within a set found to be significantly and collectively different are most responsible for that difference. This means that each significant gene set likely has a subset of genes that really account for the difference that is detected ("drivers"); and a separate subset of genes that do not significantly contribute to the greater set's difference ("passengers"). Fortunately, due to the semi-exhaustive nature of our set creation algorithm we have the ability to test for the likelihood of set members being drivers or passengers. If a gene is found in a number of significant sets at a considerably higher frequency than one would expect to see at random then it is more likely that the gene in question is a driver. In order to determine which genes identified by GSA using our newly derived sets are drivers we implemented a hypergeometric test as described in the Methods section. Those genes considered to be drivers were included in a pathway-centric network reconstruction for visualisation and interpretation (see Methods) as seen in Figures 4.2, 4.3 and 4.4.

High numbers of components are required to model a metabolic Laplacian matrix First, we noticed upon examining the Laplacian PCA model that it needed 2655 principal components to explain all of the variance in the matrix (the same number of total 'variables' in the matrix). This is unusual as PCA models are normally quite efficient at reducing the dimensionality of a

dataset. In this case we believe that it suggests that the model is likely semi-exhaustively explaining local structures in the graph. Additional file 4.9.1 is a plot of the percentage of variance explained by each principle component. It is easy to see from this plot that the variance being modelled is spread out quite broadly over the 2655 components with almost all of the components individually explaining less than 0.08% of the variance.

4.4.6 Individual components model local areas of the graph

In order to test this hypothesis visually we extracted nodes from each principle component and examined their locations in the metabolic network. Figure 4.1 is an overview of the metabolic network with the significant nodes found in the first three principal components (according to an imposed score threshold of 1), identified in green and their adjacent edges identified in red. It is clear that the first three principal components are identifying distinct, localised structures in the graph. With the hypothesis that genes that are closely related to one another in the metabolic network are likely to be co-regulated we believe that each principal component in the model is a candidate for one or more sets of genes to be tested by GSA. To examine this further we looked at the graph structures being located by many of the principal components. We found that the principal components are finding graph structures with highly connected sets of genes that will likely be good candidates for GSA. We continued this analysis through many higher components of the model to confirm that this was indeed occurring throughout a broad range of components. To confirm this observation the distance between all genes in each set was determined and the average distance within each set calculated to be 8.4. Given that there is a distance of one between a gene and the reaction it is associated with and a distance of two (reaction to compound to next reaction) between reactions, this means that in each set, on average, each gene is being associated with genes involved the first, second or third neighbouring reactions. The distances between all genes in the metabolic network have been calculated and their sorted distribution is shown in Additional file 4.9.2. The red arrow indicates the average intra-set gene distance. As such, the sets contain only about 20% of the possible gene pair distances. Thus, it appears that sets are modelling relatively local topologies in the network.

4.4.7 Semi-exhaustive nature of PCA-Graph gene set creation

We examined the distribution of PCA scores for each gene in the model across all of the principal components. Almost all genes participate in multiple principal components. This indicates that sets generated from the principal com-

ponents from the PCA of the Metabolic Laplacian matrix will be reasonably topologically exhaustive, that is to say covering the combinatorial space as constrained by the graph in an overlapping fashion. To confirm this we examined the number of times that each gene was found in sets created by a principal component score threshold of one. Additional file 4.9.3: shows the rank order distribution of the number of sets each gene is a member of. Only two genes belong to only one set and some genes belong to as many as 336 different sets. As such, it is clear that in GSA almost every gene would be tested in combination with many other groups of genes, which gives our method a higher likelihood of finding co-regulated sets of genes. We checked the location of the 20 genes that belong to the fewest gene sets and found that they typically are either located in the extreme leaf nodes of the large network or in the small disconnected subgraphs. Intuitively, this makes sense as members of the outer extremities of the large network and the disconnected subgraphs will be part of fewer graph structure variants and therefore occur in fewer principal components; thus belonging to fewer sets derived from the principal component scores.

4.4.8 Degree of overlap amongst sets

In order to determine the level of overlap between the sets generated by this approach an all-against-all comparison of the sets was done by way of set theoretic intersects. For the 3481 sets generated at a score threshold of one, an all-against-all comparison is comprised of 10,753,203 set intersects. The number of sets intersecting with each individual set was calculated and plotted (Additional file 4.9.4). As can be seen from Additional file 4.9.4 many of the sets do have intersects with one another ranging from as few as 2 to as many as 2702. In order to achieve a semi-exhaustive coverage of the graph's local topology this sort of overlap is desirable, as long as the degree of overlap is not so high that the sets become effectively redundant. In order to determine the degree of the overlap amongst sets the size of each intersection was calculated followed by the number of set intersections of each respective size (Additional file 4.9.4). Of the 10,753,203 set intersects performed, 8,225,500 showed no shared genes at all, 1,125,959 shared one gene, and 519,333 shared 2 genes. As can be seen in Additional file 4.9.4, the number of set intersections with higher degrees of overlap drops very quickly. This would appear to be a very desirable result as it appears that the Laplacian PCA is yielding sets that thoroughly cover local topological structures in the graph without introducing an excessive level of overlap, i.e. redundancy, in the sets. As such, the set intersection space is actually quite sparse as 80% of the potential set intersections show no overlap at all. This again emphasises that local topological structures in the graph are being modelled by PCA as one would not expect there to be overlaps of subgraphs (and the sets generated from them) that are topologically separated from one another.

4.4.9 Set size

The sizes of the sets vary at each principal component depending both on the region of the graph being modelled by that component and the magnitude of the score threshold employed. Additional file 4.9.5 shows the distribution of set sizes when a score threshold of +/− 1, 5 or 10 is used. The set sizes range from 5 to 92 members with an average set size of 27 genes.

4.4.10 Hypergeometric test

The hypergeometric test resulted in a ten-fold reduction in the number of genes from the threshold 1 sets up to a forty-nine-fold reduction from the threshold 5 sets, thus simplifying the resulting network that needed to be visualised and analysed.

4.4.11 Results from sets derived from different PCA score thresholds

Sets created at different PCA thresholds naturally have different set sizes and compositions and, as such, may have slightly different sensitivities in finding differential changes in some portions of the metabolic network. The threshold-based set generation algorithm produced 3481, 2703, 1487, 745, 331, 168, 80, 49 and 35 sets for threshold 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 respectively for a total of 11,216 sets.

In order to evaluate the efficacy of sets created at different thresholds, a set theoretic approach was used to determine the effects of these thresholds employed during set generation on the pathways identified by GSA. Lists (sets) of pathways with significant differences were generated from the results of GSA performed with sets generated with PCA score thresholds 1 through 7 (see Methods). Thresholds 6 through 10 yielded no results after the application of the hypergeometric test. This is likely due to the relatively small number of genes and sets generated at the higher threshold levels. As such, there are insufficient differences in global and individual set frequencies for the hypergeometric test to discriminate between driver and passenger genes (i.e. to generate sufficiently low p-values). The set theoretic difference was then determined between the pathway sets resulting from PCA scores thresholds 2 through 5 as compared to the pathway set found using a PCA scores threshold of 1.

All but one of the pathways found at thresholds 2 through 5 were also found with a threshold of 1, the one exception being one gene in Aminoacyl-tRNA biosynthesis found at threshold 3. As expected, the threshold 1 results contained many pathways not found at the other thresholds, specifically: 11, 17, 22 and 24 more pathways were found with threshold 1 than with thresholds 2, 3, 4 and 5 respectively. However, as the different thresholds contain

subsets of the pathways and genes found in threshold 1 they can be used as a 'zoom-in, zoom-out' method to view focused portions of the differentially changed network. Figure 4.2 contains the results of sets generated with a PCA scores threshold of 1. Figure 4.3 is a slightly different view of the results, containing edges between pathways that share one or more compounds. Although somewhat more complex, it clearly shows the linkages between genes in the result sets. Of note is the connections between the pathways seen at the top of the figure that show the connections between N-Glycan biosynthesis, the Phosphatidylinositol signaling system, Inositol phosphate metabolism, Glycerophospholipid metabolism, Glycerolipid metabolism and Glycosylphosphatidylinositol (GPI)-anchor biosynthesis (all but one of which were missed by the single-pathway method) which can be seen in more detail in Figure 4.4.

4.4.12 Comparison of threshold-based sets to step-function-based sets

The step-function-based algorithm generated 1704 gene sets. The lower number of sets generated is likely due to the fact that the step-function identified many groupings with less than five members and as such did not qualify as a set for GSA purposes. The GSA results from sets generated by the use of a step-function on PCA scores was compared to the results generated at a score threshold of 1. Similar to the threshold comparisons the step-function identified three pathways (Cysteine & methionine metabolism, TCA cycle and Methane metabolism) not found by the threshold 1 method. The threshold 1 method also found 18 pathways that were not found by the step-function method. This likely means that the entire positive or negative branch of a principle component is accurately modelling a topological structure and that subgraphs within that topology are not generally needed to increase the sensitivity of gene sets to be used for GSA.

4.4.13 Comparison of threshold 1 sets to single pathway sets

As has been discussed above, threshold 1 sets, with very few exceptions, give the most complete view of differential changes occurring in the metabolic network. In order to compare this new set generation method to previous approaches we created single pathway sets (see Methods) which are the type of sets that have traditionally been used in the past. We ran GSA on the 69 single pathway sets described above and compared the results to those generated by the use of threshold 1 sets. The results of the single pathway results can be seen in Figure 4.5. The single pathway sets only identified 6 pathways as opposed to the 30 identified by the threshold 1 method. Furthermore, because there is comparatively little overlap in the single pathway sets it is not possible

to use a hypergeometric test to determine which of the genes are drivers of the set statistic and which are simply passengers. Furthermore, with single pathway sets it is much harder to determine how the sets are related to each other within the metabolic network. It is clear that the results from Threshold 1 as presented in Figures 4.2, 4.3 and 4.4 are much more comprehensive that those from single pathway sets as presented in Figure 4.5.

Single pathway sets found 20 (true positives) of the 76 genes found to be differentially expressed by the Threshold 1 method, thus it missed 56 genes (false negatives) found by the Threshold 1 method. Furthermore, it included 111 apparent passenger genes (false positives) that were present presumably because the pathway set containing them was found to be significant and not because they contributed to the signal.

The results of the threshold 1 analysis could certainly be presented as lists of genes or as lists of different pathways. However, it is precisely this isolated pathway mode of thinking that we are pointing out has limitations, the effects of which have been noted elsewhere in the literature: "The classical method of metabolic engineering, identifying a rate-determining step in a pathway and alleviating the bottleneck by enzyme overexpression, has motivated much research but has enjoyed only limited practical success. Intervention of other limiting steps, of counter-balancing regulation, and of unknown coupled pathways often confounds this direct approach [3]."

Thus, it would appear that there have been many attempts at metabolic engineering in which researchers have not taken into account the fact that pathways don't exist in isolation but are rather all interconnected. As such, many attempts at metabolic engineering fail as they are based on an overly simplistic model. Therefore, we contend that seeing the genes in the context of the metabolic network is a more accurate way to portray and understand what is really occurring in metabolism. Pathway names are very useful mnemonic devices to remind one what area of metabolism is involved but should not be used to artificially isolate gene functions from one another.

Our method finds differentially expressed genes involved in N-Glycan biosynthesis, the Phosphatidylinositol signalling system, Inositol phosphate metabolism, Glycerophospholipid metabolism and Glycerolipid metabolism that were missed by the single-pathway method. Furthermore, it provides the visual context for how these pathways are interlinked and how they are linked to Glycosylphosphatidylinositol (GPI)-anchor biosynthesis (the one related pathway which the single pathway method does find). These findings and the relationship between them would seem to be an important aspect of the biology that was missed by the single-pathway approach.

In addition, in order to determine the central biological themes that our method found but were overlooked by the single pathway GSA method, all of the genes that were found by our method and not found with single pathway GSA were subjected to GO Enrichment analysis and FunCat Enrichment analysis as described in the Methods section. Although it is beyond the scope

of this paper to do a full biological interpretation of these genes, we have discussed them briefly below and included the full GO and FunCat Enrichment results as Supplementary Material.

4.4.14 FunCat enrichment

The general categories of the functions that our method finds that would have been missed previously include, as identified by FunCat enrichment: amino acid metabolism, nitrogen, sulphur and selenium metabolism, carbohydrate metabolism, lipid, fatty acid and isoprenoid metabolism, secondary metabolism, glycolysis and gluconeogenesis, tricarboxylic-acid pathway (citrate cycle, Krebs cycle, TCA cycle), metabolism of energy reserves (e.g. glycogen, trehalose), complex cofactor/co-substrate/vitamin binding and protein modification (N-glycosylation) [Additional file **??**]. Simply looking at the GO Enrichment diagrams gives one a good sense of how much pertinent biology was found with our method that was missed by pathway GSA.

4.4.15 Gene Ontology enrichment

It is clear that we are finding core biological themes in N-linked protein glycosylation, which itself is interesting given that it was O-linked glycosylation that was inhibited. In addition, lipid, phospholipid and glycerophospholipid biosynthesis are affected which links nicely with the enrichment for GPI anchor biosynthesis. There are further indications that branched chain and aromatic amino acid metabolism is affected as well as the TCA cycle and redox metabolism. Furthermore, it is clear that there is an effect on a number of genes involved in the cell wall, ER and plasma membranes, as one would expect to see with the inhibition of protein glycosylation which would be likely to affect a number of integral membrane proteins [Additional file 4.9.6, Additional file 4.9.7 and Additional file 4.9.8].

4.4.16 Differentially express genes

A list of the differentially expressed genes found by our method, including their systematic and gene names as well as their descriptions has been included as supplementary material [Additional file 4.9.9].

4.4.17 Gene sets and software availability

The gene sets (affy probeset ids) generated by threshold 1 are included as supplementary material [Additional file 4.9.10]. Software used for the analysis is available upon request to the first author.

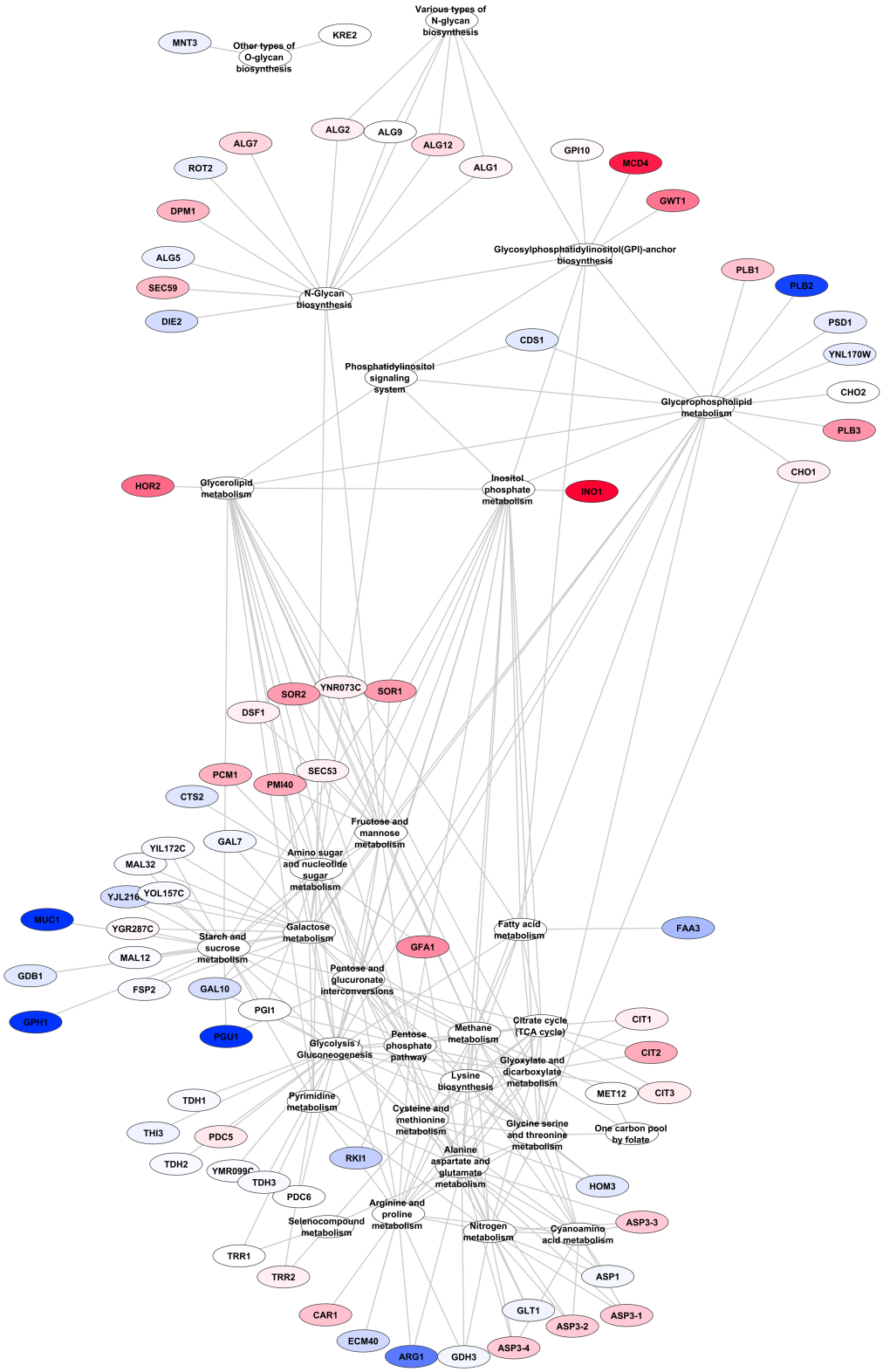


Figure 4.3: Genes from PCA scores threshold = 1 derived sets found to be significantly differentially expressed in their pathway-centric including edges between pathways that share at least one compound. Node colouring as described in Figure 4.2.

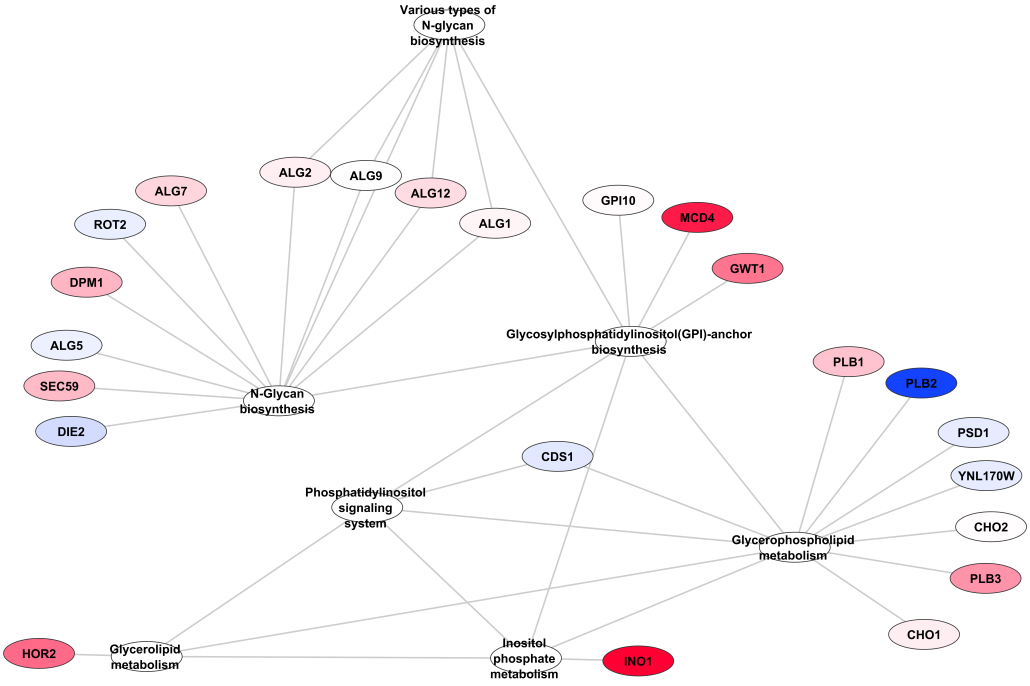


Figure 4.4: Zoom in of the Genes and Pathways from Figure 4.3 that are involved in Glycerolipid metabolism, Glycerophospholipid metabolism, Glycosylphosphatidylinositol (GPI)-anchor biosynthesis, Inositol phosphate metabolism, N-Glycan biosynthesis, and the Phosphatidylinositol signaling system. Node colouring as described in Figure 4.2.

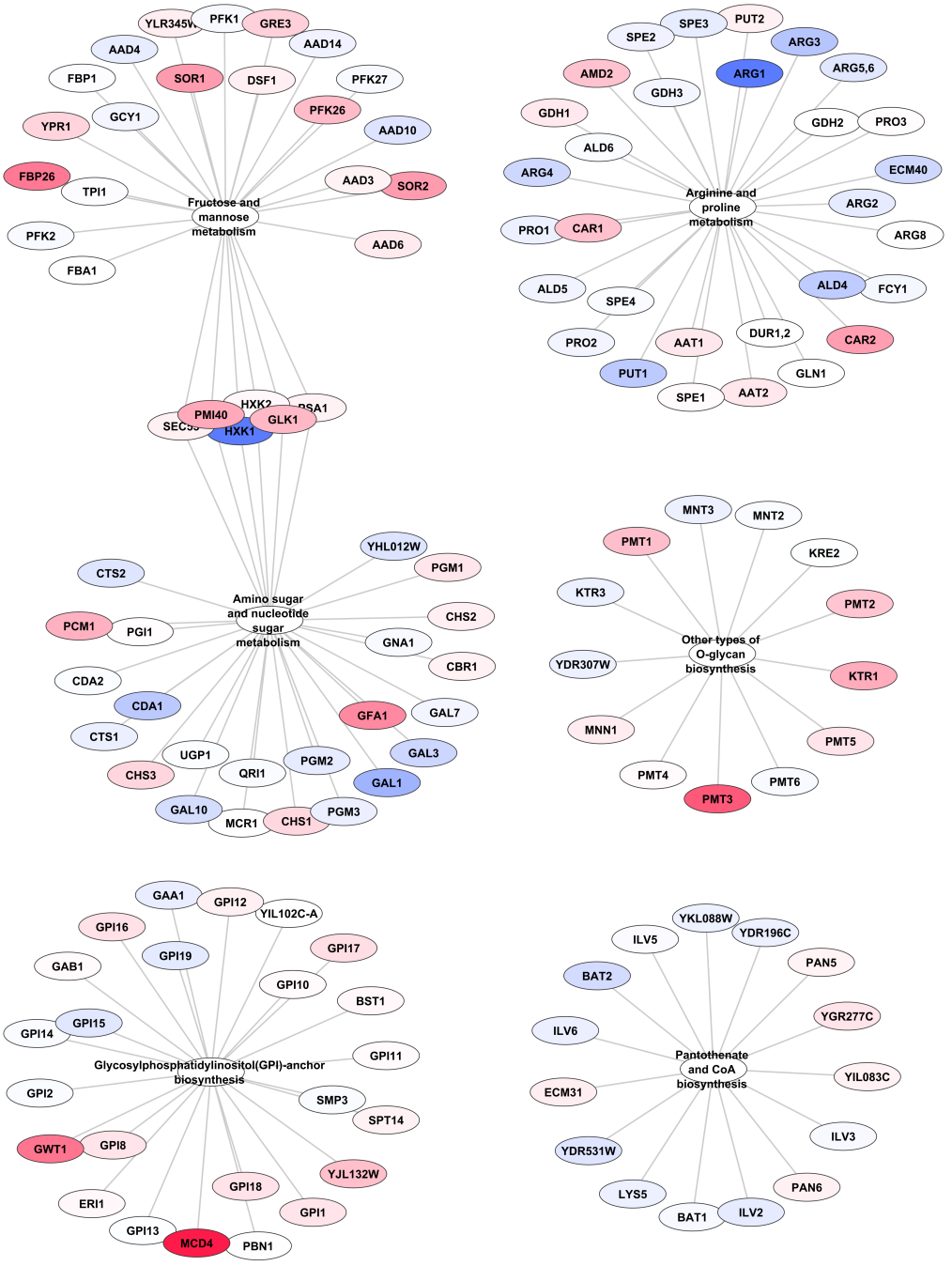


Figure 4.5: Results from single pathway sets. Node colouring as described in Figure 4.2.

4.5 Conclusions

To our knowledge there is no report in the literature of the use of Principal Component Analysis of the Laplacian matrix of a metabolic graph for any purpose, including set generation for Gene Set Analysis. As such, it appears that this is a novel method with which to find local topological structures in metabolic networks. We have shown that sets generated from the metabolic networks are semi-exhaustive in that there are many partial set overlaps but the degree of overlap is relatively low. The fact that each gene is a member of many sets allowed us to devise a hypergeometric enrichment test to determine which genes were likely to be driving the set statistic and which were likely to simply be passengers and could thus be pruned away from the results set. We have further shown that the structure represented by each signed half of each principal component (greater than or equal to a score threshold of 1) is adequate for set generation. Further stratification of each principal component, whether by threshold or step-function methods did not significantly increase sensitivity. However, the thresholding method did prove to be useful as a 'zoom-in, zoom-out' function for biological interpretation of the results. When compared to traditional pathway sets this method appears to be much more sensitive as it is a better representation of the underlying complexity of a metabolic network. Furthermore, the method applied here allows one to see the full context of the genes likely to be driving the set statistics rather than as simply lists of pathways each containing an unknown number of driver and passenger genes.

References

[1] (2008). Microarray Data Downloaded from the Gene Expression Omnibus: GSE12193. *http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE12193*.

[2] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990 October). Basic local alignment search tool. *Journal of molecular biology*, vol. 215, no. 3, pp. 403–10. ISSN 0022-2836.

[3] Bailey, J.E., Sburlati, A., Hatzimanikatis, V., Lee, K., Renner, W.A. and Tsai, P.S. (2002 September). Inverse metabolic engineering: a strategy for directed genetic engineering of useful phenotypes. *Biotechnology and bioengineering*, vol. 79, no. 5, pp. 568–79. ISSN 0006-3592.

[4] Banerjee, S. and Pedersen, T. (2003). The design, implementation, and use of the Ngram Statistics Package. *Computational Linguistics and Intelligent Text Processing*, pp. 370–381.

[5] Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P. (2003 January). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)*, vol. 19, no. 2, pp. 185–93. ISSN 1367-4803.

[6] Dinu, I., Potter, J.D., Mueller, T., Liu, Q., Adewale, A.J., Jhangri, G.S., Einecke, G., Famulski, K.S., Halloran, P. and Yasui, Y. (2007 January). Improving gene set analysis of microarray data by SAM-GS. *BMC bioinformatics*, vol. 8, p. 242. ISSN 1471-2105.

[7] Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *The Annals of Applied Statistics*, pp. 107–129.

[8] Falcon, S. and Gentleman, R. (2007 January). Using GOstats to test gene lists for GO term association. *Bioinformatics (Oxford, England)*, vol. 23, no. 2, pp. 257–258. ISSN 1367-4811.

[9] Goeman, J.J. and Bühlmann, P. (2007 April). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics (Oxford, England)*, vol. 23, no. 8, pp. 980–987. ISSN 1367-4811.

[10] Goeman, J.J., Oosting, J., Cleton-Jansen, A.-M., Anninga, J.K. and van Houwelingen, H.C. (2005 May). Testing association of a pathway with survival using gene expression data. *Bioinformatics (Oxford, England)*, vol. 21, no. 9, pp. 1950–7. ISSN 1367-4803.

[11] Goeman, J.J., van de Geer, S.A., de Kort, F. and van Houwelingen, H.C. (2004 January). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics (Oxford, England)*, vol. 20, no. 1, pp. 93–9. ISSN 1367-4803.

[12] Hietaniemi, J. (). Graph. *CPAN: http://search.cpan.org/˜jhi/Graph-0.94/*.

[13] Hietaniemi, J. (). Set::Scalar. *CPAN: http:// search.cpan.org/dist/Set-Scalar/*.

[14] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70.

[15] Irizarry, R.A., Wang, C., Zhou, Y. and Speed, T.P. (2009 December). Gene set enrichment analysis made simple. *Statistical methods in medical research*, vol. 18, no. 6, pp. 565–75. ISSN 1477-0334.

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

[17] Kim, S.-Y. and Volsky, D.J. (2005 January). PAGE: parametric analysis of gene set enrichment. *BMC bioinformatics*, vol. 6, p. 144. ISSN 1471-2105.

[18] Mardia, K.V., Kent, J.T. and Bibby, J.M. (1980). Multivariate analysis.

[19] Mootha, V.K., Lindgren, C.M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrå le, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D. and Groop, L.C. (2003 July). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, vol. 34, no. 3, pp. 267–73. ISSN 1061-4036.

[20] Pavlidis, P., Lewis, D.P. and Noble, W.S. (2002 January). Exploring gene expression data with class scores. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 474–85.

[21] Pavlidis, P., Qin, J., Arango, V., Mann, J.J. and Sibille, E. (2004 June). Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochemical research*, vol. 29, no. 6, pp. 1213–22. ISSN 0364-3190.

[22] Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Güldener, U., Mannhaupt, G., Münsterkötter, M. and Mewes, H.W. (2004 January). The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic acids research*, vol. 32, no. 18, pp. 5539–45. ISSN 1362-4962.

[23] Saerens, M., Fouss, F., Yen, L. and Dupont, P. (2004). The principal components analysis of a graph, and its relationships to spectral clustering. *Machine Learning: ECML 2004*, pp. 371–383.

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

[25] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P. (2005 October). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–50. ISSN 0027-8424.

[26] Tian, L., Greenberg, S.A., Kong, S.W., Altschuler, J., Kohane, I.S. and Park, P.J. (2005 September). Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 38, pp. 13544–9. ISSN 0027-8424.

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

4.6 Competing interests

The authors declare that they have no competing interests.

4.7 Authors' contributions

DJ conceived of the study, created the graph and Laplacean matrix used in this paper, wrote the R code to perform the PCA and the Perl code for set generation, set analysis, and hypergeometric testing, wrote the R script to perform the GSA analysis, did the subsequent interpretation of both the PCA and GSA results, created the figures and wrote the manuscript. GE performed the initial PCA of a Laplacian matrix in Matlab and provided the PCA scores that DJ used to develop the initial prototype, participated in discussions about the potential applications of this approach and edited the manuscript. All authors read and approved the final manuscript.

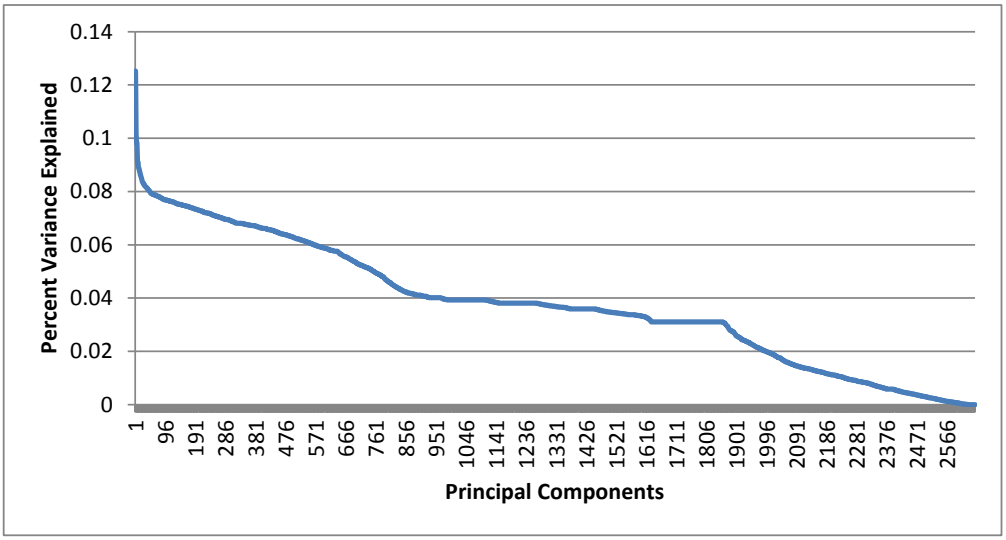
4.8 Acknowledgements

We would like to thank the members of the Computational Biology Group and the students and staff in general at the IWBT for support and discussion.

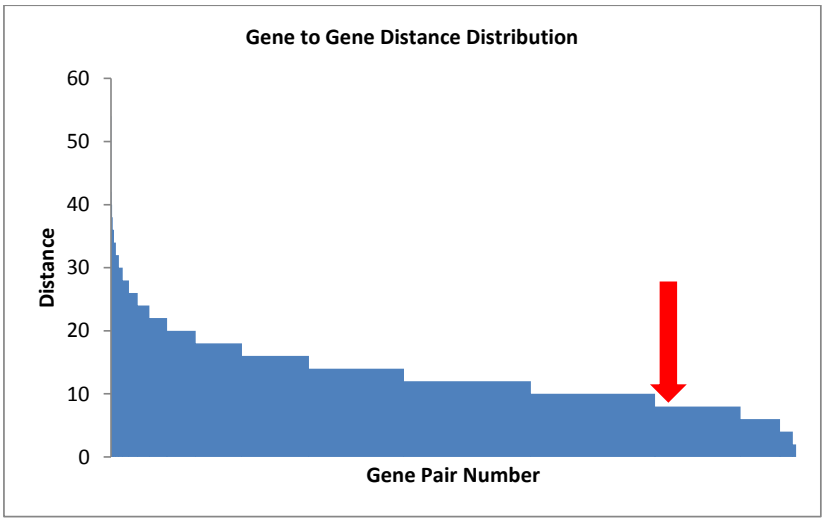
We would particularly like to thank Piet Jones for useful discussions about the hypergeometric test. We would like to gratefully acknowledge the funding provided by Winetech, THRIP, and the South African National Research Foundation Bioinformatics & Functional Genomics Programme.

4.9 Additional files

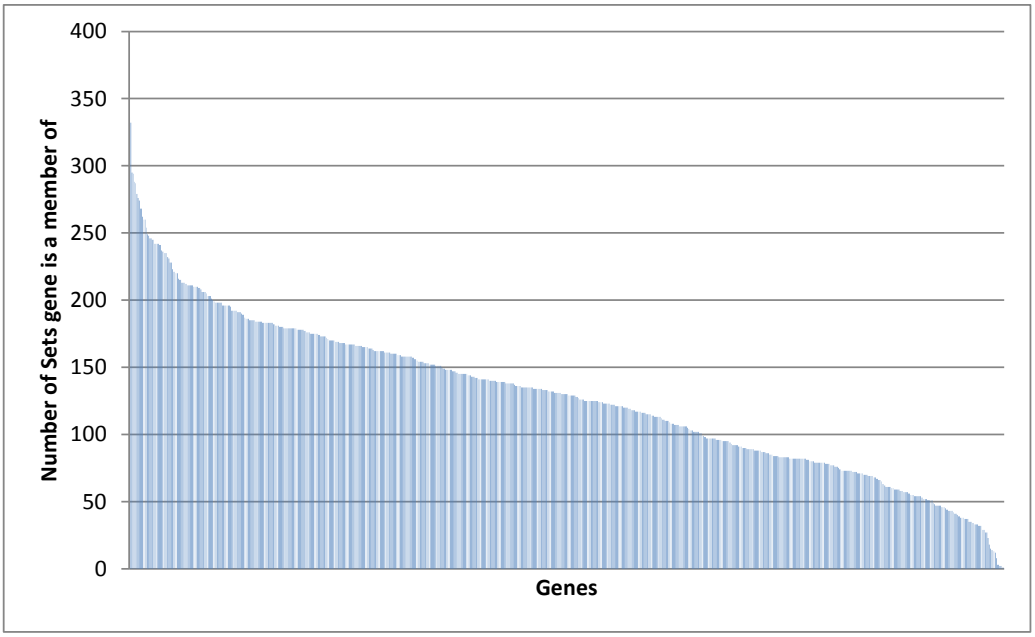
4.9.1 Percentage Variance Explained by each principal component.



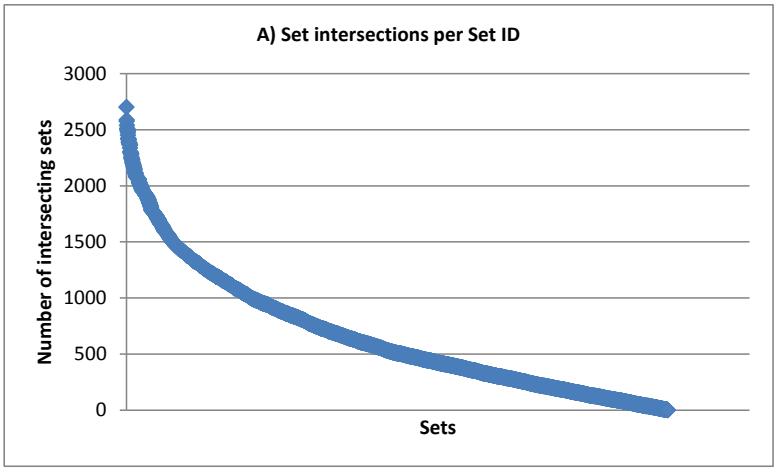
4.9.2 Gene to Gene Distance Distribution.

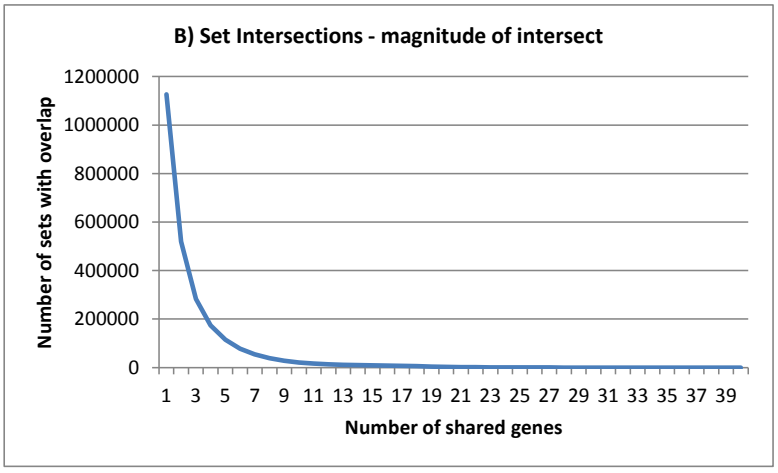


4.9.3 Rank order distribution of the number of sets that each gene is contained in.

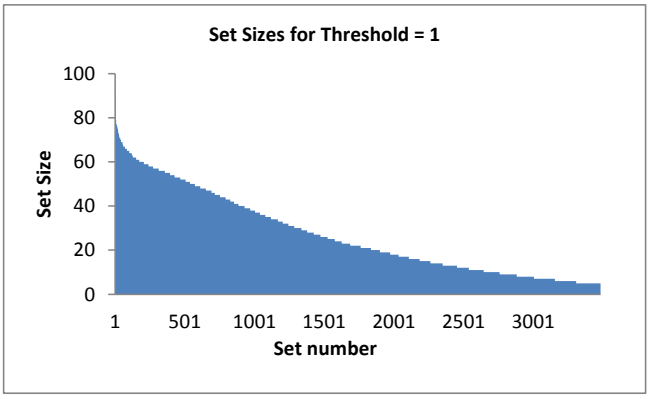


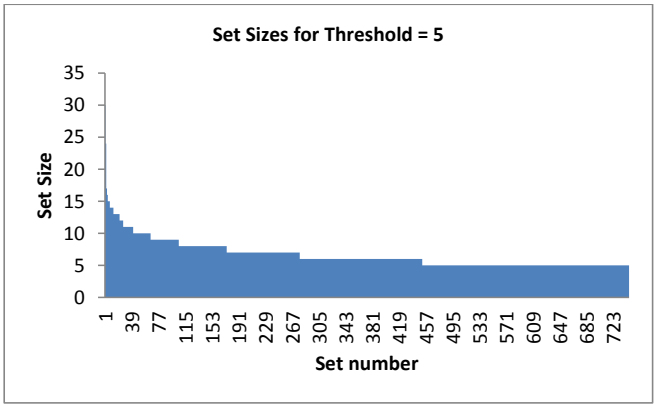
4.9.4 (A) number of other gene sets that intersect with each gene set. (B) Number of occurrences of gene overlaps between sets as found by set intersections.

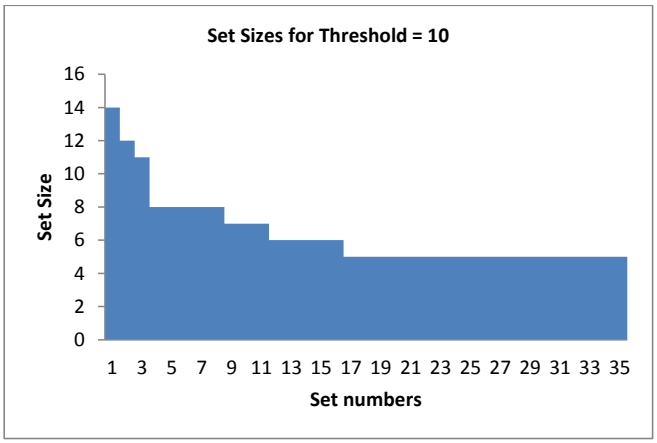




4.9.5 Set sizes generated at each positive and negative arm of each principal component at thresholds 1, 5 and 10.







4.9.6 Gene ID, Name and Description of metabolic genes found to be differentially expressed by GSA-PCA.

Systematic Name	Gene Name	Description
YBR110W	ALG1	Mannosyltransferase, involved in asparagine-linked glycosylation in the endoplasmic reticulum (ER)
YNR030W	ALG12	Alpha-1,6-mannosyltransferase localized to the ER
YGL065C	ALG2	Mannosyltransferase that catalyzes two consecutive steps in the N-linked glycosylation pathway
YPL227C	ALG5	UDP-glucose:dolichyl-phosphate glucosyltransferase, involved in asparagine-linked glycosylation in the endoplasmic reticulum
YBR243C	ALG7	UDP-N-acetyl-glucosamine-1-P transferase, transfers Glc-Nac-P from UDP-GlcNac to Dol-P in the ER in the first step of the dolichol pathway of protein asparagine-linked glycosylation
YNL219C	ALG9	Mannosyltransferase, involved in N-linked glycosylation
YOL058W	ARG1	Arginosuccinate synthetase, catalyzes the formation of L-argininosuccinate from citrulline and L-aspartate in the arginine biosynthesis pathway
YDR321W	ASP1	Cytosolic L-asparaginase, involved in asparagine catabolism
YLR155C	ASP3-1	Cell-wall L-asparaginase II, involved in asparagine catabolism
YLR157C	ASP3-2	Cell-wall L-asparaginase II, involved in asparagine catabolism
YLR158C	ASP3-3	Cell-wall L-asparaginase II, involved in asparagine catabolism
YLR160C	ASP3-4	Cell-wall L-asparaginase II, involved in asparagine catabolism
YPL111W	CAR1	Arginase, responsible for arginine degradation, expression responds to both induction by arginine and nitrogen catabolite repression
YBR029C	CDS1	Phosphatidate cytidylyltransferase (CDP-diglyceride synthetase)
YER026C	CHO1	Phosphatidylserine synthase, functions in phospholipid biosynthesis
YGR157W	CHO2	Phosphatidylethanolamine methyltransferase (PEMT), catalyzes the first step in the conversion of phosphatidylethanolamine to phosphatidylcholine during the methylation pathway of phosphatidylcholine biosynthesis
YNR001C	CIT1	Citrate synthase, catalyzes the condensation of acetyl coenzyme A and oxaloacetate to form citrate
YCR005C	CIT2	Citrate synthase, catalyzes the condensation of acetyl coenzyme A and oxaloacetate to form citrate, peroxisomal isozyme involved in glyoxylate cycle
YPR001W	CIT3	Dual specificity mitochondrial citrate and methylcitrate synthase
YDR371W	CTS2	Protein similar to Ashbya gossypii sporulation-specific chitinase
YGR227W	DIE2	Dolichyl-phosphoglucose-dependent alpha-1,2 glucosyltransferase of the ER, functions in the pathway that synthesizes the dolichol-linked oligosaccharide precursor for N-linked protein glycosylation, has a role in regulation of ITR1 and INO1

YPR183W	DPM1	Dolichol phosphate mannose (Dol-P-Man) synthase of the ER membrane, catalyzes the formation of Dol-P-Man from Dol-P and GDP-Man
YEL070W	DSF1	Deletion suppressor of mpt5 mutation
YMR062C	ECM40	Mitochondrial ornithine acetyltransferase, catalyzes the fifth step in arginine biosynthesis
YHR007C	ERG11	Lanosterol 14-alpha-demethylase, catalyzes the C-14 demethylation of lanosterol to form 4,4''-dimethyl cholesta-8,14,24-triene-3-beta-ol in the ergosterol biosynthesis pathway
YNL280C	ERG24	C-14 sterol reductase, acts in ergosterol biosynthesis
YIL009W	FAA3	Long chain fatty acyl-CoA synthetase, has a preference for C16 and C18 fatty acids
YJL221C	FSP2	Protein of unknown function, expression is induced during nitrogen limitation
YBR019C	GAL10	UDP-glucose-4-epimerase, catalyzes the interconversion of UDP-galactose and UDP-D-glucose in galactose metabolism
YBR018C	GAL7	Galactose-1-phosphate uridyl transferase, synthesizes glucose-1-phosphate and UDP-galactose from UDP-D-glucose and alpha-D-galactose-1-phosphate in the second step of galactose catabolism
YPR184W	GDB1	Glycogen debranching enzyme containing glucanotranferase and alpha-1,6-amyloglucosidase activities, required for glycogen degradation
YAL062W	GDH3	NADP(+)-dependent glutamate dehydrogenase, synthesizes glutamate from ammonia and alpha-ketoglutarate
YKL104C	GFA1	Glutamine-fructose-6-phosphate amidotransferase, catalyzes the formation of glucosamine-6-P and glutamate from fructose-6-P and glutamine in the first step of chitin biosynthesis
YDL171C	GLT1	NAD(+)-dependent glutamate synthase (GOGAT), synthesizes glutamate from glutamine and alpha-ketoglutarate
YPR160W	GPH1	Non-essential glycogen phosphorylase required for the mobilization of glycogen, activity is regulated by cyclic AMP-mediated phosphorylation, expression is regulated by stress-response elements and by the HOG MAP kinase pathway
YGL142C	GPI10	Integral membrane protein involved in glycosylphosphatidylinositol (GPI) anchor synthesis
YJL091C	GWT1	Protein involved in the inositol acylation of glucosaminyl phosphatidylinositol (GlcN-PI) to form glucosaminyl(acyl)phosphatidylinositol (GlcN(acyl)PI), an intermediate in the biosynthesis of glycosylphosphatidylinositol (GPI) anchors
YER052C	HOM3	Aspartate kinase (L-aspartate 4-P-transferase)
YER062C	HOR2	One of two redundant DL-glycerol-3-phosphatases (RHR2/GPP1 encodes the other) involved in glycerol biosynthesis
YJL153C	INO1	Inositol 1-phosphate synthase, involved in synthesis of inositol phosphates and inositol-containing phospholipids

YDR483W	KRE2	Alpha1,2-mannosyltransferase of the Golgi involved in protein mannosylation
YGR292W	MAL12	Maltase (alpha-D-glucosidase), inducible protein involved in maltose catabolism
YBR299W	MAL32	Maltase (alpha-D-glucosidase), inducible protein involved in maltose catabolism
YKL165C	MCD4	Protein involved in glycosylphosphatidylinositol (GPI) anchor synthesis
YPL023C	MET12	Protein with methylenetetrahydrofolate reductase (MTHFR) activity in vitro
YIL014W	MNT3	Alpha-1,3-mannosyltransferase, adds the fourth and fifth alpha-1,3-linked mannose residues to O-linked glycans during protein O-glycosylation
YIR019C	MUC1	GPI-anchored cell surface glycoprotein (flocculin) required for pseudohyphal formation, invasive growth, flocculation, and biofilms
YEL058W	PCM1	Essential N-acetylglucosamine-phosphate mutase
YLR134W	PDC5	Minor isoform of pyruvate decarboxylase, key enzyme in alcoholic fermentation, decarboxylates pyruvate to acetaldehyde, regulation is glucose- and ethanol-dependent, repressed by thiamine, involved in amino acid catabolism
YGR087C	PDC6	Minor isoform of pyruvate decarboxylase, decarboxylates pyruvate to acetaldehyde, involved in amino acid catabolism
YBR196C	PGI1	Glycolytic enzyme phosphoglucose isomerase, catalyzes the interconversion of glucose-6-phosphate and fructose-6-phosphate
YJR153W	PGU1	Endo-polygalacturonase, pectolytic enzyme that hydrolyzes the alpha-1,4-glycosidic bonds in the rhamnogalacturonan chains in pectins
YMR008C	PLB1	Phospholipase B (lysophospholipase) involved in lipid metabolism, required for deacylation of phosphatidylcholine and phosphatidylethanolamine but not phosphatidylinositol
YMR006C	PLB2	Phospholipase B (lysophospholipase) involved in phospholipid metabolism
YOL011W	PLB3	Phospholipase B (lysophospholipase) involved in phospholipid metabolism
YER003C	PMI40	Mannose-6-phosphate isomerase, catalyzes the interconversion of fructose-6-P and mannose-6-P
YNL169C	PSD1	Phosphatidylserine decarboxylase of the mitochondrial inner membrane, converts phosphatidylserine to phosphatidylethanolamine
YOR095C	RKI1	Ribose-5-phosphate ketol-isomerase, catalyzes the interconversion of ribose 5-phosphate and ribulose 5-phosphate in the pentose phosphate pathway
YBR229C	ROT2	Glucosidase II catalytic subunit required for normal cell wall synthesis
YFL045C	SEC53	Phosphomannomutase, involved in synthesis of GDP-mannose and dolichol-phosphate-mannose
YMR013C	SEC59	Dolichol kinase, catalyzes the terminal step in dolichyl monophosphate (Dol-P) biosynthesis

YJR159W	SOR1	Sorbitol dehydrogenase
YDL246C	SOR2	Protein of unknown function
YJL052W	TDH1	Glyceraldehyde-3-phosphate dehydrogenase, isozyme 1, involved in glycolysis and gluconeogenesis
YJR009C	TDH2	Glyceraldehyde-3-phosphate dehydrogenase, isozyme 2, involved in glycolysis and gluconeogenesis
YGR192C	TDH3	Glyceraldehyde-3-phosphate dehydrogenase, isozyme 3, involved in glycolysis and gluconeogenesis
YDL080C	THI3	Probable alpha-ketoisocaproate decarboxylase, may have a role in catabolism of amino acids to long-chain and complex alcohols
YDR353W	TRR1	Cytoplasmic thioredoxin reductase, key regulatory enzyme that determines the redox state of the thioredoxin system, which acts as a disulfide reductase system and protects cells against both oxidative and reductive stress
YHR106W	TRR2	Mitochondrial thioredoxin reductase involved in protection against oxidative stress, required with Glr1p to maintain the redox state of Trx3p
YGR287C	YGR287C	Isomaltase (alpha-D-glucosidase)
YIL172C	YIL172C	Putative protein of unknown function with similarity to glucosidases
YJL216C	YJL216C	Protein of unknown function, similar to alpha-D-glucosidases
YMR099C	YMR099C	Glucose-6-phosphate 1-epimerase (hexose-6-phosphate mutarotase), likely involved in carbohydrate metabolism
YNL170W	YNL170W	Dubious open reading frame
YNR073C	YNR073C	Putative mannitol dehydrogenase
YOL157C	YOL157C	Putative protein of unknown function

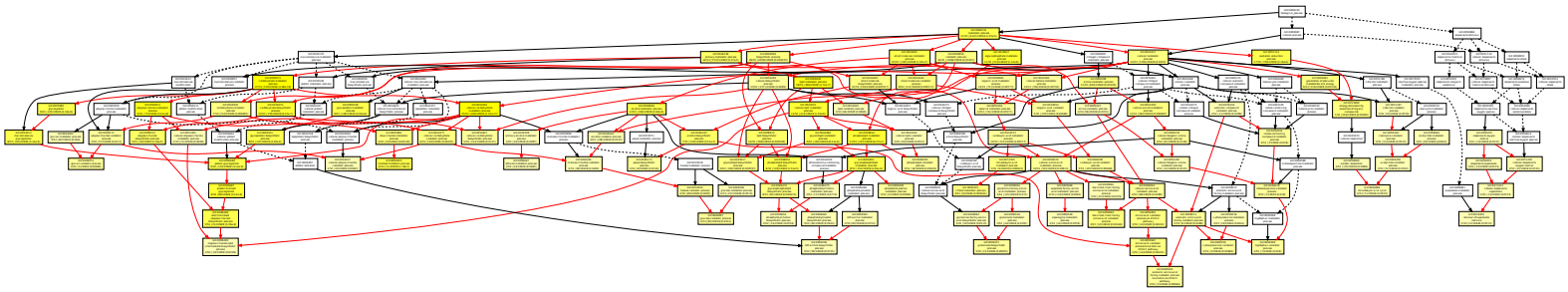
4.9.7 FunCat Enrichment of Genes Missed by Pathway-based GSA and found by GSA-PCA

FUNCTIONAL CATEGORY	P-VALUE
01 METABOLISM	3.26E-32
01.01 amino acid metabolism	7.72E-06
01.01.03 assimilation of ammonia, metabolism of the glutamate group	0.008386305
01.01.03.02 metabolism of glutamate	0.001066048
01.01.03.02.01 biosynthesis of glutamate	0.000288172
01.01.06 metabolism of the aspartate family	1.64E-06
01.01.06.01 metabolism of aspartate	5.92E-09
01.01.06.02 metabolism of asparagine	5.92E-09
01.01.06.02.02 degradation of asparagine	4.83E-11
01.01.06.04 metabolism of threonine	0.077961284
01.01.06.05 metabolism of methionine	0.038937459
01.01.09 metabolism of the cysteine - aromatic group	0.516028301
01.01.09.02 metabolism of serine	0.126586982
01.02 nitrogen, sulfur and selenium metabolism	1.94E-06
01.03 nucleotide/nucleoside/nucleobase metabolism	0.60290255
01.03.07 deoxyribonucleotide metabolism	0.004924993
01.04 phosphate metabolism	0.522461311
01.05 C-compound and carbohydrate metabolism	1.46E-22
01.05.02 sugar, glucoside, polyol and carboxylate metabolism	1.32E-16
01.05.02.04 sugar, glucoside, polyol and carboxylate anabolism	1.75E-08
01.05.02.07 sugar, glucoside, polyol and carboxylate catabolism	3.01E-15
01.05.03 polysaccharide metabolism	4.15E-09
01.05.03.02 peptidoglycan metabolism	0.000425229
01.05.03.02.04 peptidoglycan anabolism	0.000425229
01.05.06 C-2 compound and organic acid metabolism	0.077961284

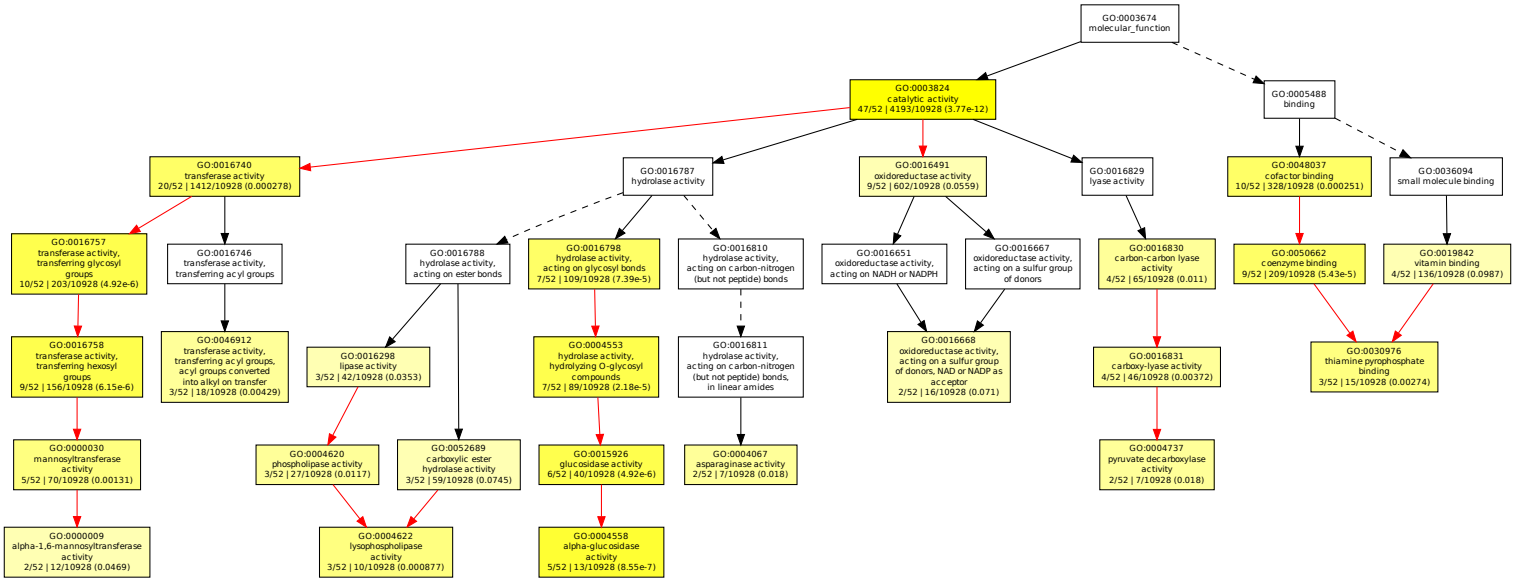
01.05.06.07 C-2 compound and organic acid catabolism	0.077961284
01.05.13 transfer of activated C-1 groups	0.165177731
01.05.13.03 tetrahydrofolate-dependent C-1-transfer	0.110664964
01.05.25 regulation of C-compound and carbohydrate metabolism	0.685447192
01.06 lipid, fatty acid and isoprenoid metabolism	1.30E-07
01.06.02 membrane lipid metabolism	1.88E-09
01.06.02.01 phospholipid metabolism	6.26E-09
01.06.02.02 glycolipid metabolism	0.126586982
01.06.06 isoprenoid metabolism	0.051871779
01.06.06.11 tetracyclic and pentacyclic triterpenes (cholesterin, steroids and hopanoids) metabolism	0.040995065
01.06.10 regulation of lipid, fatty acid and isoprenoid metabolism	0.209281393
01.07 metabolism of vitamins, cofactors, and prosthetic groups	0.43541859
01.07.01 biosynthesis of vitamins, cofactors, and prosthetic	0.259331964
01.07.07 regulation of the metabolism of vitamins, cofactors, and	0.044077638
01.20 secondary metabolism	0.032401996
01.20.01 metabolism of primary metabolic sugar derivatives	0.026680334
01.20.01.01 metabolism of secondary monosaccharides	0.017865484
01.20.19 metabolism of secondary products derived from glycine,	0.012257963
02 ENERGY	1.42E-10
02.01 glycolysis and gluconeogenesis	0.015592064
02.04 glyoxylate cycle	0.077961284
02.07 pentose-phosphate pathway	0.194840541
02.10 tricarboxylic-acid pathway (citrate cycle, Krebs cycle, TCA	0.002571949
02.16 fermentation	0.066110196
02.16.01 alcohol fermentation	0.00578709
02.19 metabolism of energy reserves (e.g. glycogen, trehalose)	3.73E-11
10 CELL CYCLE AND DNA PROCESSING	0.999953307
10.03 cell cycle	0.998018921
10.03.01 mitotic cell cycle and cell cycle control	0.984755316
11 TRANSCRIPTION	0.999976969
11.02 RNA synthesis	0.997600998
11.02.03 mRNA synthesis	0.995714342
11.02.03.04 transcriptional control	0.990460183
11.02.03.04.01 transcription activation	0.31601095
14 PROTEIN FATE (folding, modification, destination)	0.603484972
14.04 protein targeting, sorting and translocation	0.924459449
14.07 protein modification	0.044918176
14.07.01 modification with fatty acids (e.g. myristylation,	0.034946899
14.07.02 modification with sugar residues (e.g. glycosylation,	7.16E-09
14.07.02.01 O-directed glycosylation, deglycosylation	0.126586982
14.07.02.02 N-directed glycosylation, deglycosylation	2.82E-09

16 PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)	0.970086967
16.21 complex cofactor/cosubstrate/vitamine binding	0.000154247
16.21.07 NAD/NADP binding	1.47E-05
20 CELLULAR TRANSPORT, TRANSPORT FACILITIES AND	0.999964767
20.01 transported compounds (substrates)	0.99608177
20.01.10 protein transport	0.720954304
30 CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION	0.883551445
30.01 cellular signalling	0.838505291
30.01.09 second messenger mediated signal transduction	0.346363869
30.01.09.09 fatty acid derivatives mediated signal transduction	0.035417494
32 CELL RESCUE, DEFENSE AND VIRULENCE	0.119257158
32.01 stress response	0.105620542
32.01.01 oxidative stress response	0.086835185
32.01.03 osmotic and salt stress response	0.413918344
32.01.11 nutrient starvation response	7.36E-06
32.07 detoxification	0.282752973
32.07.01 detoxification involving cytochrome P450	0.035417494
32.07.07 oxygen and radical detoxification	0.209281393
34 INTERACTION WITH THE ENVIRONMENT	0.986954775
34.07 cell adhesion	0.110664964
34.07.01 cell-cell adhesion	0.086246226
42 BIOGENESIS OF CELLULAR COMPONENTS	0.961098988
42.01 cell wall	0.301138115
42.16 mitochondrion	0.790505437
43 CELL TYPE DIFFERENTIATION	0.782339234
43.01 fungal/microorganismic cell type differentiation	0.782339234
43.01.03 fungal and other eukaryotic cell type differentiation	0.782339234
43.01.03.05 budding, cell polarity and filament formation	0.77966155
43.01.03.09 development of asco- basidio- or zygospore	0.782592703
99 UNCLASSIFIED PROTEINS	0.999999353

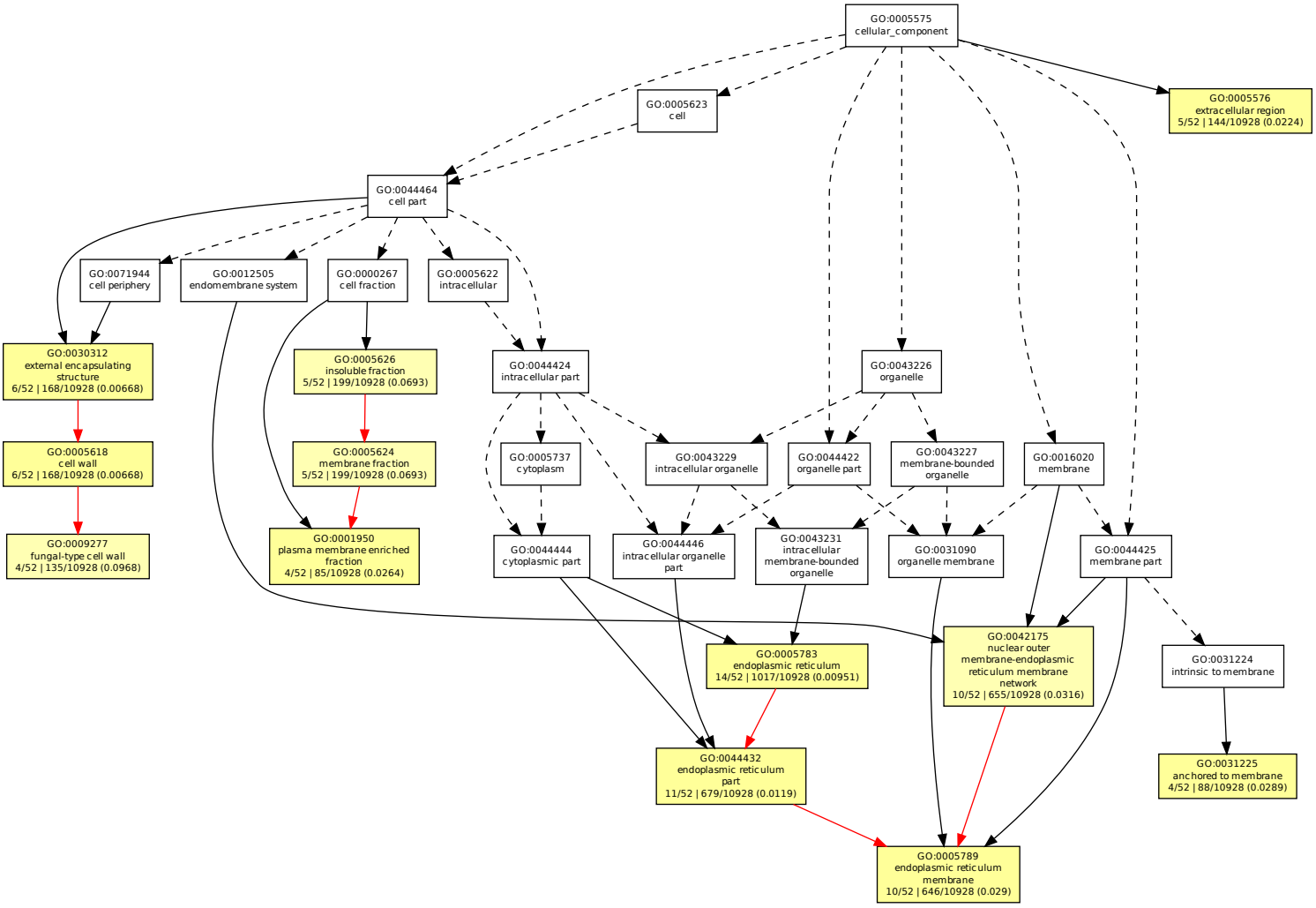
4.9.8 GO Enrichment (Biological Process) of the False Negatives from traditional GSA pathway sets.



4.9.9 GO Enrichment (Molecular Function) of the False Negatives from traditional GSA pathway sets.



4.9.10 GO Enrichment (Cellular Location) of the False Negatives from traditional GSA pathway sets.



Chapter 5

Network-based analysis for cross-species microarray experiments

5.1 Abstract

Background: The analysis of data from spotted potato cDNA microrarrays cross-hybridised with tobacco transcripts presented a number of challenges: the gene to probe relationships are ambiguous, the original annotation was sparse and tobacco is not a model organism so there was relatively little information with which to contextualise the results.
Results: These problems were addressed with network-based models for: cross-hybridisation of probes to transcripts; phylogenomic-based Gene Ontology, Interpro functional motif and descriptive annotation; projection of differentially expressed probes onto pathway and protein-protein interaction models from other species and a cross-species co-expression network. Prior to the use of the methods described here the results from the analysis of this experiment were too sparse for meaningful interpretation or publication. However, the use of these network-based methods led to the successful analysis of the data set, yielding considerable insights into the functions associated with the differentially expressed genes resulting from the over-expression of a grapevine defence gene in tobacco. The functions of these differentially expressed genes were shown to be involved in metabolism, cell walls, signalling pathways, photosynthesis, stress response and defence response pathways amongst others.
Conclusions: The network structures described in this paper proved to be useful in the analysis and interpretation of this challenging experimental data set. The context provided by GO annotation, functional motif description, pathway projections, protein-protein interaction networks and cross-species co-expression networks proved to be invaluable in understanding the underlying biology that was reflected in this data set. Thus, by embracing and modelling

complexity with networks one is able to successfully overcome the challenges presented by such data sets.

5.2 Background

5.2.1 Experimental Background

Polygalacturonase-inhibiting proteins (PGIPs) are extracellular proteins found in plants that can recognise and inhibit endopolygalacturonases (ePGs) from fungal pathogens [6]. ePGs hydrolyse homogalacturonan in the plant cell wall as part of a fungal pathogen's mechanism to gain entry into the cell. Two ePGs are required for *Botrytis cinerea* to be fully virulent [22, 13]. Federici *et al.* proposed that PGIPs, in addition to inhibiting ePGs, may also activate defence signalling pathways and, as such, may form part of a plant's innate immune system [10].

Tobacco has previously been used to study plant-pathogen interactions and has very little PGIP activity against Botrytis ePGs [12, 26]. It has been shown that the transgenic over-expression of *Vitis vinifera* PGIP 1 (*VvPGIP1*) in tobacco plants results in an elevated level of PGIP enzyme activity and confers resistance to *B. cinerea* infection [12].

In order to investigate these resistant transgenic lines of tobacco further the global gene expression of two such transgenic lines (*VvPGIP1* lines 37 and 45) was compared to that of wild type (SR1) with the use of the TIGR 10K potato microarray. A paper describing the biological interpretation of the experimental results has been published [2]. This paper is intended to explain in considerably more detail the computational methods that were developed in order to properly analyse this most challenging data set.

5.2.2 Computational Approach

The analysis of the experiment described above was made difficult by the potential ambiguity in probeset-to-transcript hybridizations (as this was a cross-species hybridisation approach), incomplete GO annotations of the genes of the two plant species (potato and tobacco) involved and a paucity of expression data concerning tobacco as it is not a model organism. It is important to note that at the time this work was done neither the tobacco nor the potato genome had been sequenced. Thus, the only sequence resources were the EST collections for these two species. The TIGR 10K potato microarray was annotated roughly ten years ago, prior to the sequencing of all but one plant genome (*Arabidopsis thaliana*). As such, the annotation for this microarray was sparse and GO enrichment analysis was largely uninformative as it only yielded three enriched GO terms (all in the same hierarchy). As there was no gene expression data available for tobacco at the time (as there was no tobacco

array available) it was impossible to put our experimental results in the context of other tobacco expression results. Furthermore, it was desirable to provide as much biological context for the experimental results as possible in order to aid in their interpretation. To address these challenges we built a simple model for cross-hybridisation, re-annotated the tobacco and potato expressed sequence tag (EST) collections via phylogenomic inference and provided biological context for our results by mapping them onto pathway databases, protein-protein interaction networks, a functional domain classification database (Interpro) and the creation of cross-species correlation networks. Prior to the use of these approaches the results from the analysis of this experiment were too sparse for meaningful interpretation or publication.

5.3 Methods

5.3.1 Probe Specificity

Two studies have looked at cross-hybridisation of transcripts from the cytochrome P450 gene family to probes for genes in the family. Both studies concluded that there is effectively no significant cross-hybridisation when the sequences are less than 80% identical to one another [25, 9].

BLAST [3] was used to map all of the extant tobacco ESTs to the potato probesets they would likely hybridise to on the TIGR 10K potato microarray. A threshold of 80% identity over at least a 100 bp region was used to identify transcripts that would be likely to hybridise to probes on the potato microarray. The result was stored as a network denoting the relationships between potential transcripts and probesets.

5.3.2 EST Translation

Annotated Genome sequences were available for neither *S. tuberosum* nor *N. tabacum* and thus we were reliant on EST datasets for annotation based analysis of the microarray. However, as is often the case, the annotation is incomplete and out of date for the EST datasets for *S. tuberosum* (from which the probesets were designed) and *N. tabacum*. Prot4EST [23] was modified in order to run on a high performance computing architecture and the EST sequences from both *S. tuberosum* and *N. tabacum* translated into their corresponding protein sequences. Tobacco EST assemblies from the PlantGDB database [7] and potato probe sequences were processed separately. Briefly, a protein target database was created containing the protein translations of the nine plant genomes available at the time from PLAZA [18]. A perl program was written that split the EST sequences into 100 query files and created the necessary Torque submit scripts to run 100 parallel submissions of each step of the Prot4EST translation pipeline.

5.3.3 Gene Family Determination

The PLAZA [18] protein translations (fasta format) and associated GO and Interpro [11] annotation for the complete genomes of *Ostreococcus lucimarinus, Chlamydomonas reinhardtii, Physcomitrella patens, Sorghum bicolor, Oryza sativa, Vitis vinifera, Populus trichocarpa, Carica papaya* and *Arabidopsis thaliana* as represented in PLAZA were downloaded.

5.3.3.1 Potato

The PLAZA protein sequences were combined with the translated *S. tuberosum* probe sequences (which were EST clones) and filtered for quality and a minimum length of 10 amino acids. This fasta file was then indexed for blast searches. An all-against all blastp query (e-value threshold of 10^{-5}) was done as follows: in order to parallelise the blastp queries a perl program was written which split the fasta file into 100 equal size fasta files and created Torque-based submit scripts for each fasta file, and one master script to submit all of the queries to a distributed compute cluster.

5.3.3.2 Tobacco

Similar to potato, the translated *N. tabacum* ESTs were combined with the PLAZA protein sequences, filtered and a parallel all-against-all blastp query prepared and performed.

5.3.3.3 OrthoMCL

OrthoMCL [14] was used to create orthologous clusters of all of the proteins resident in PLAZA with the translated *S. tuberosum* and *N. tabacum* ESTs. A perl programm was written to create submit scripts so that the blastp results could be parsed into OrthoMCL format in parallel. Orthologs and paralogs were determined and their relationships normalised with the use of OrthoMCL. The resulting similarity network was clustered into orthogous groups with the MCL algorithm with an inflation value of 2.0.

In order to spot check the quality of the orthologous groups detected, multiple alignments for all clusters associated with probesets of interest were created using MUSCLE [8] and visualized with Jalview [24].

5.3.4 Microarray Annotation by Phylogenomic Inference

GO and Interpro annotations found for members of the resultant orthologous clusters were projected onto the *S. tuberosum* and *N. tabacum* proteins in the cluster. Subsequently, with the use of the probeset-to-transcript network described above, annotations from both *S. tuberosum* and *N. tabacum* were

assigned to the corresponding probesets. This was done with a custom built perl program and the workflow can be seen in Figure 5.1.

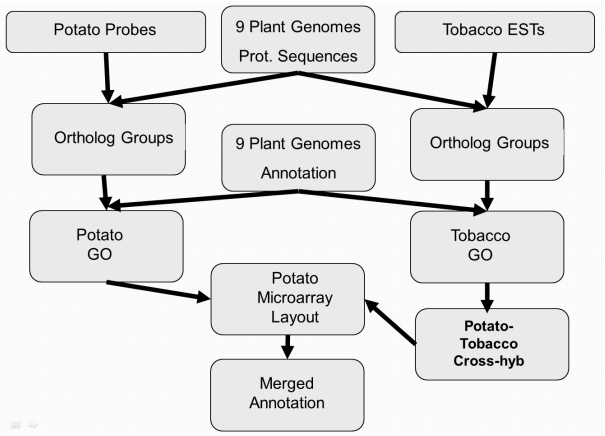


Figure 5.1: The annotation pipeline. Orthologous groups of 9 plant genomes and either potato or tobacco ESTs were created. GO Annotation from the members of the orthologous clusters was then mapped to the potato microarray probes.

5.3.4.1 Cluster Annotation Expression Network

A network structure was created to represent all of the resulting clusters, their associated annotation (GO terms, Interpro motifs, gene descriptions, etc), their assignment to probes and their corresponding expression values. An edge was created between a cluster and a probe if at least one of the tobacco transcripts in the cluster shared 80% identity with the probe over at least a 100 base pair segment. The resulting network structures were stored in XGMML and Cytoscape [19] was used for visualisation and querying. The Enhanced Search Plugin for Cytoscape makes this annotation network searchable with complex boolean queries as well as by annotation field.

5.3.5 GO and Interpro Domain Enrichment

Significantly differentially expressed probesets between SR1 and *VvPGIP1* lines 37 and 45 transgenic lines were determined by Limma [20]. The projected annotation for the differentially expressed probesets was then analysed for Enrichment of Gene Ontology Terms with GOEAST [27] and Interpro domains with FuncAssociate [4].

5.3.6 Cross-Species Pathway Projections

The orthologous clusters were used to map probesets to *A. thaliana* identifiers. These ids were then used to map the probesets (and their differential expression values) onto *A. thaliana* pathways represented in KEGG, Mapman and MetNET for further contextual analysis and visualisation. The workflow for this process can be seen in Figure 5.2.

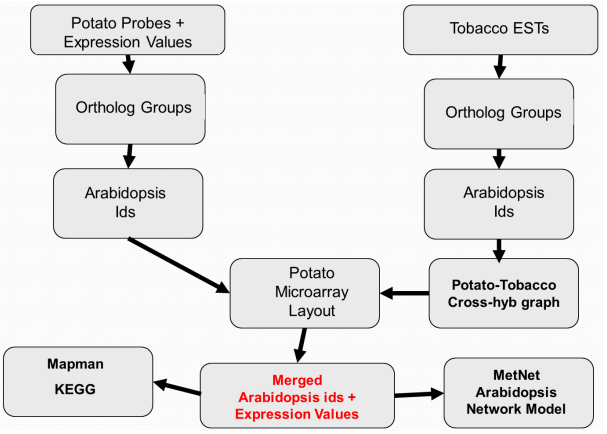


Figure 5.2: The cross-apecies pathway projection pipeline.

5.3.7 Cross-Species Coexpression Analysis

BLAST was used to map all of the sequences associated with *A. thaliana* ids onto specific probesets on the *A. thaliana* Affymetrix microarray which resulted in a gene-to-probeset network. The tobacco microarray probeset to ortholgous cluster network of differentially expressed genes described above was then used to produce a mapping of the *S. tuberosum* probesets to *A. thaliana* probesets. The workflow for this process can be seen in Figure 5.3.

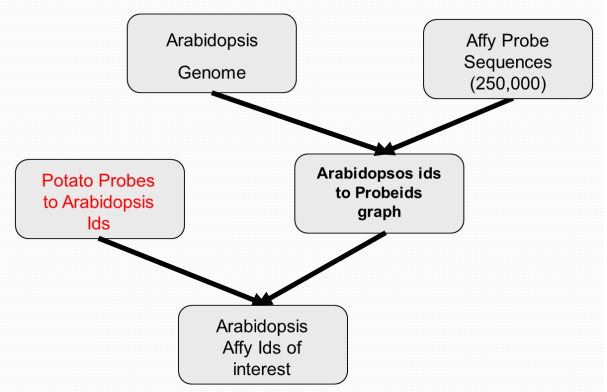


Figure 5.3: Cross-apecies microarray mapping workflow.

A perl script was written to download all of the expression data (50 probesets at a time) from over 1700 *A. thaliana* Affymetrix microarrays normalised together by CressExpress [21]. Vectors of each probeset in the Affymetrix microarrays were created (with the use of a custom built perl script) which included expression information from each microarray in the collection. Each of the differentially expressed *S. tuberosum* probesets were then analysed for coexpression in *A. thaliana* using these vectors and a Hierarchical Clustering Algorithm [5] based on Pearson correlations. Thresholds were employed to select the clusters representing the highest levels of coexpression and each cluster analysed for biological significance.

The workflow for this process can be seen in Figure 5.4.

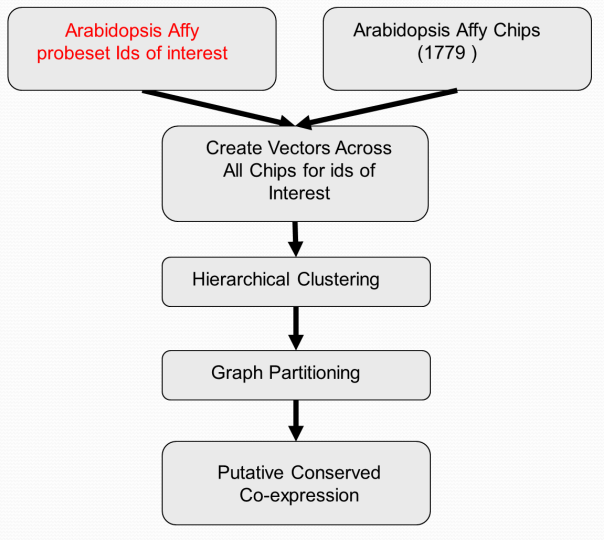


Figure 5.4: Cross-species coexpression network construction.

5.4 Results and Discussion

5.4.1 Cross-hybridisation Network

In an ideal scenario the transcripts from one gene would only hybridise to one probe on a microarray. However, even on microarrays carefully constructed from fully sequenced genomes, all microarrays have instances where the transcripts from multiple genes will be likely to hybridise to the same probe. Given that this experiment involved a cross-species hybridisation with relatively long probe lengths it was necessary to create a model which would include the high likelihood of transcript to probe ambiguity. As described in the Methods section a cross-hybridisation network model was built in order to reflect which tobacco transcripts (as represented by ESTs) were likely to cross-hybridise

with which potato probes. As can be seen in Figure 5.5, there is indeed a considerable number of probes that will hybridise to more than one transcript. This fact was used as the realistic basis for microarray annotation.

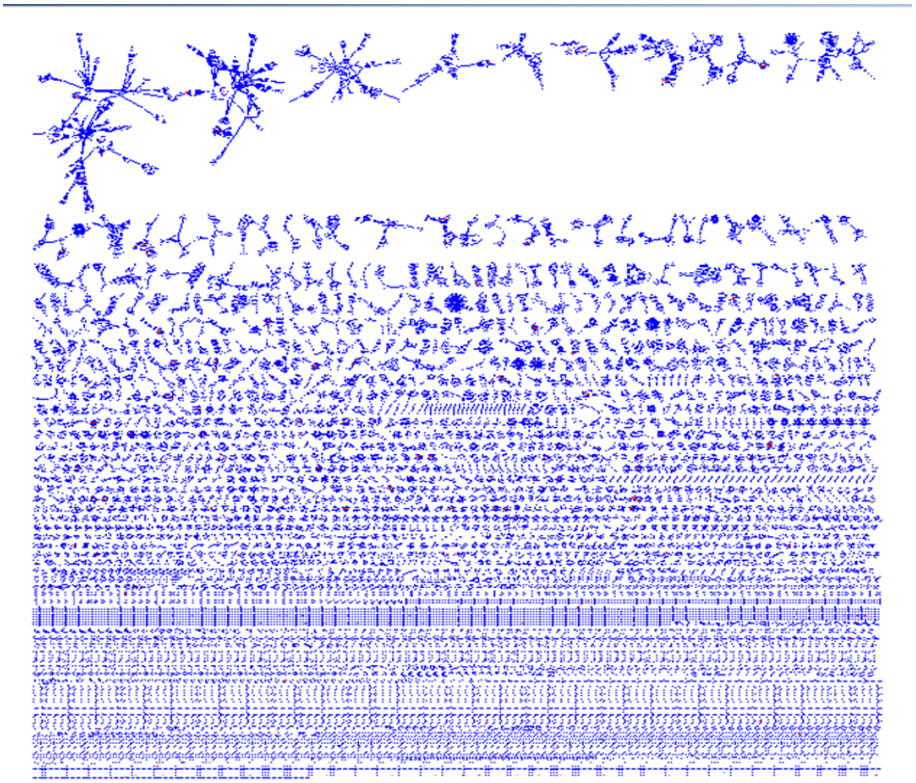


Figure 5.5: Potato probe to tobacco transcript cross-hybidisation network. Potato probes were mapped to tobacco ESTs with the use of BLAST. An edge between a probe and an EST was created if they shared 80% identity over 100 contiguous base pairs. The more complex the network cluster is the more ambiguity there is between probes and transcripts. The most ambiguous probes are seen at the top of the figure and the least ambiguous at the bottom.

5.4.2 Functional Motif Enrichment

Interpro enrichment analyses found several enriched functional motifs/domains from the wild-type vs transgenic lines differential gene list as can be seen in Table 5.1. Of immediate note are the functional domains that have roles in cell wall metabolism, including Beta-glucanase, Glycoside hydrolase, Xyloglucan endo-transglycosylase and Xyloglucan endotransglucosylase/hydrolase. Of further interest in the context of plant defence is the enrichment for lipooxygenase as lipooxygenase gene expression has previously been shown to respond to pathogens, abscisic acid, and methyl jasmonate [15].

Table 5.1: Interpro functional motifs found to be enriched in the genes in tobacco up or down regulated by the overexpresion of *VvPGIP1*. The functional motif shown in bold was the only function identified from this experiment prior to the use of the network-based methods described in this paper.

Interpro ID	**Functional Motif**
IPR008264	Beta-glucanase
IPR008263	Glycoside hydrolase, family 16, active site
IPR010713	Xyloglucan endo-transglycosylase, C-terminal
IPR000757	Glycoside hydrolase, family 16
IPR016455	**Xyloglucan endotransglucosylase/hydrolase**
IPR012269	Aquaporin
IPR000425	Major intrinsic protein
IPR013373	Archaeal flagellin, N-terminal related
IPR008985	Concanavalin A-like lectin/glucanase
IPR013320	Concanavalin A-like lectin/glucanase, subgroup
IPR020829	Glyceraldehyde 3-phosphate dehydrogenase, catalytic domain
IPR002818	ThiJ/PfpI
IPR002130	Peptidyl-prolyl cis-trans isomerase, cyclophilin-type
IPR015891	Cyclophilin-like
IPR001983	Translationally controlled tumour-associated TCTP
IPR011057	Mss4-like
IPR010417	Embryo-specific 3
IPR008976	Lipase/lipooxygenase, PLAT/LH2

5.4.3 GO Enrichment

When GO enrichment was performed on the differentially expressed probes with the original microarray annotation precisely three terms (hydrolases, galactosidase activity and beta-galactosidase activity) in the same hierarchy were found to be enriched [1]. This low number of terms is directly attributable to the paucity of annotation present in the original microarray annotation file.

The results of the GO Enrichment analysis performed with the newly created annotation can be seen in the supplemental material of [2]. The difference is quite striking as over three hundred GO terms were found to be enriched and a number of clear themes appear in the analysis with the new annotation. As was noted in the paper from this work [2] there were four broad groups affected by *VvPGIP1* over-expression, namely: stress defence signalling; photosynthesis; cell wall biogenesis and organization and carbon metabolism. Other, more specific groups of interest are associated with glucan and polysaccharide metabolic processes, water transport as well as response to auxin and brassinosteroid stimuli and to cyclopentones. Cyclopentones are of interest as Jasmonic acid (involved in defence signalling cascades) synthesis requires several cyclopentenone precursors which may also play roles similar

to Jasmonic acid *in vivo*. The differences in genes involved in photosynthesis, glycolysis, energy transfer and associated with the chloroplast and mitochondrial organelles reflect the differences in the metabolic state brought about by *VvPGIP1* over-expression [2].

5.4.4 Annotation Network Visualisation

As described in the Methods section an integrated network structure containing GO terms, the orthologous relationships and both the descriptive and Interpro annotation of all of the orthologs was constructed. This allows one to both browse and search through the GO, Interpro (motifs and domains) and descriptive annotations of the differentially expressed genes via their orthologous plant genes. For example, one can take a GO term of interest and see all of the tobacco genes mapping to it and see the annotation of all of their orthologs from other plant species. It also allows us to see which GO terms associate with each other via differentially expressed genes. A simple chain of events could be as follows. One starts by visualizing the entire annotation network as seen in Figure 5.6. It is then possible to select all of the nodes that have annotation (gene description, GO terms, Interpro motif descriptions) that contain the work *"defense"*. Nodes containing *"defense"* somewhere in their annotation will be highlighted in yellow as seen in Figure 5.7. A breadth-first search can then be performed in Cytoscape in order to select all of the nodes connected to the nodes selected by the query. Because this network is composed of discrete subgraphs for each probe this will simply select all of the relevant probes and its associated orthologous clusters. A new smaller network that just contains the selected subgraphs can then be created from the selected nodes as seen in Figure 5.8. One can zoom in on any of the subgraphs in order to evaluate cross-hybridisation and the associated annotation as can be seen in Figure 5.9. In this subgraph one can see that an upregulated probe (red) links to three different ortholgous clusters. Two of the clusters (top and bottom right) are from the tobacco ortholgous cluster analysis and represent two distinct gene families, both of which contain at least one tobacco transcript (EST) likely to cross-hybridise to the potato probe. The third orthologous cluster (left) is from the potato orthologous cluster analysis. It shares many orthogs with the tobacco cluster on the bottom right indicating that the two analyses found largely the same family. The nodes in the network can be selected manually or, as mentioned previously, can be queried with any set of terms or ids. When nodes are selected by any means the annotation associated with them can be seen in a data panel as seen in Figure 5.10.

5.4.4.1 Example of Annotation Network Used for Data Mining

The presence of the Interpro functional motif information allows one to expand the exploratory analysis. For example, there was a particular interest

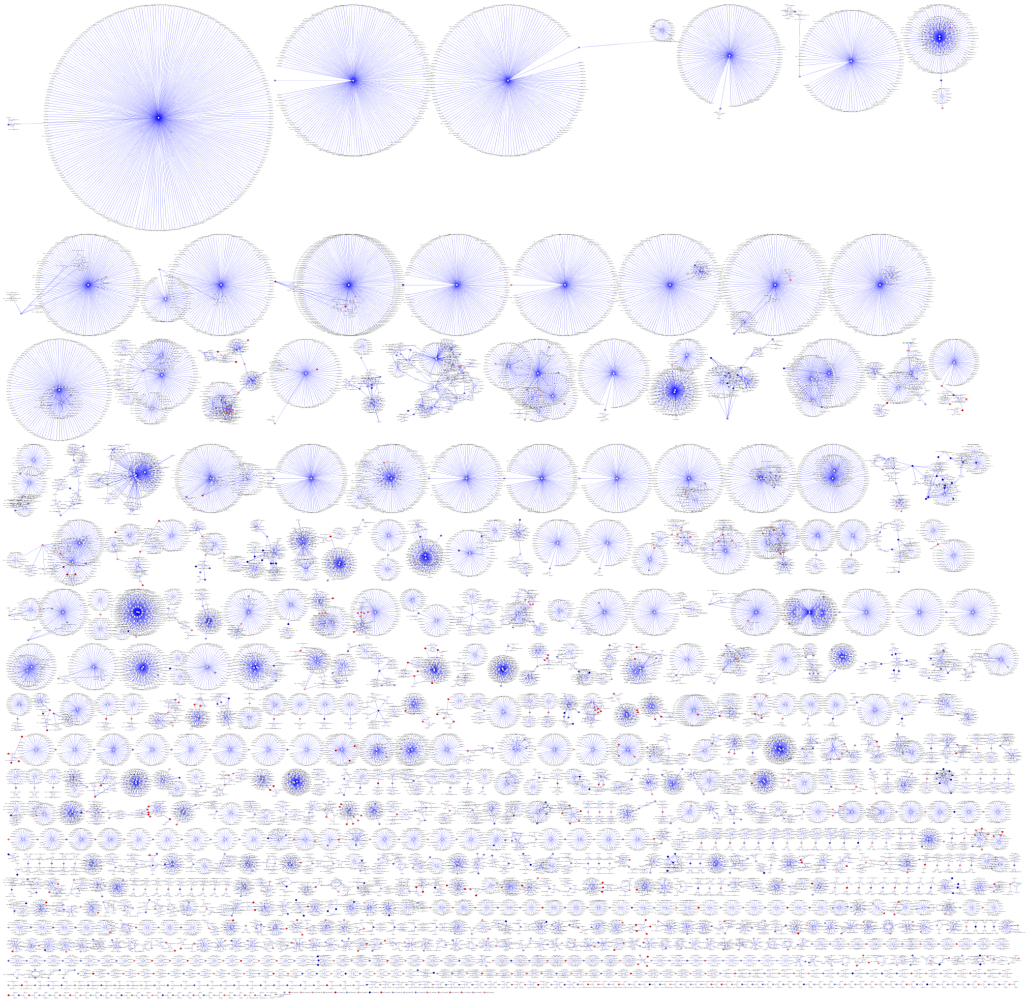


Figure 5.6: Whole expression annotation network. Differentially expressed potato probes are linked to the orthologous clusters of the tobacco transcripts likely to cross-hybiridise with them as well as ortholgous potato clusters that contain the probe sequence. Orhtologous genes are annotated with GO, Interpro and gene descriptions. Probes are coloured red or blue scaled to the positive or negative fold change respectively.

in understanding more about lignin associated genes in this dataset as lignin has previously been linked to pathogen defence [16, 17]. If one searches the Interpro database for any domain/motif mentioning lignin one finds the motifs listed in Table 5.2.

It is then possible to take those Interpro ids and search the integrated annotation network to yield a set of genes from the orhtologous clusters that contain one or more of these motifs/domains. Subsequently, a breadth-first search of the network will find the associated information, including several tobacco transcripts (ESTs) and a new network can be created (Figure 5.11)

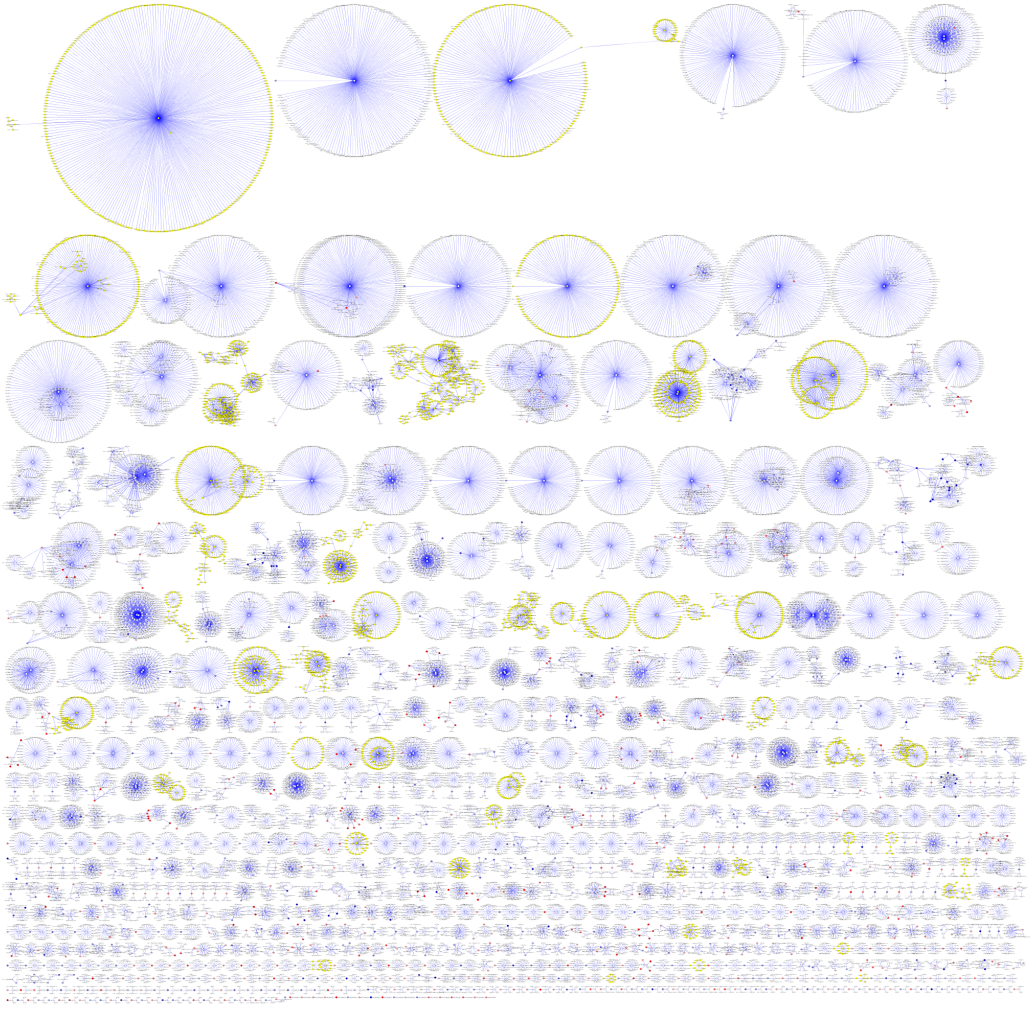


Figure 5.7: Annotation network defense query. The annotation network can be searched for any term. In this figure all nodes that have the word "defense" in their annotation have been selected and are shown in yellow.

with just the selected subgraphs as described above.

For each of these tobacco ESTs one can see the annotation of their orthologs and so infer likely functions: Plant peroxidase, Fungal lignin peroxidase, Haem peroxidase (plant/fungal/bacterial), Catalase-peroxidase haem and Plant ascorbate peroxidase.

Furthermore, one can learn more from the descriptive annotation that was integrated for the orthologs. For example, for the ascorbate peroxidase activity the following was obtained from the annotation of an ortholog:

> Encodes a cytosolic ascorbate peroxidase APX1. Ascorbate peroxidases are enzymes that scavenge hydrogen peroxide in plant cells. Eight types of APX have been described for *Arabidopsis*: three cytosolic (*APX1*,

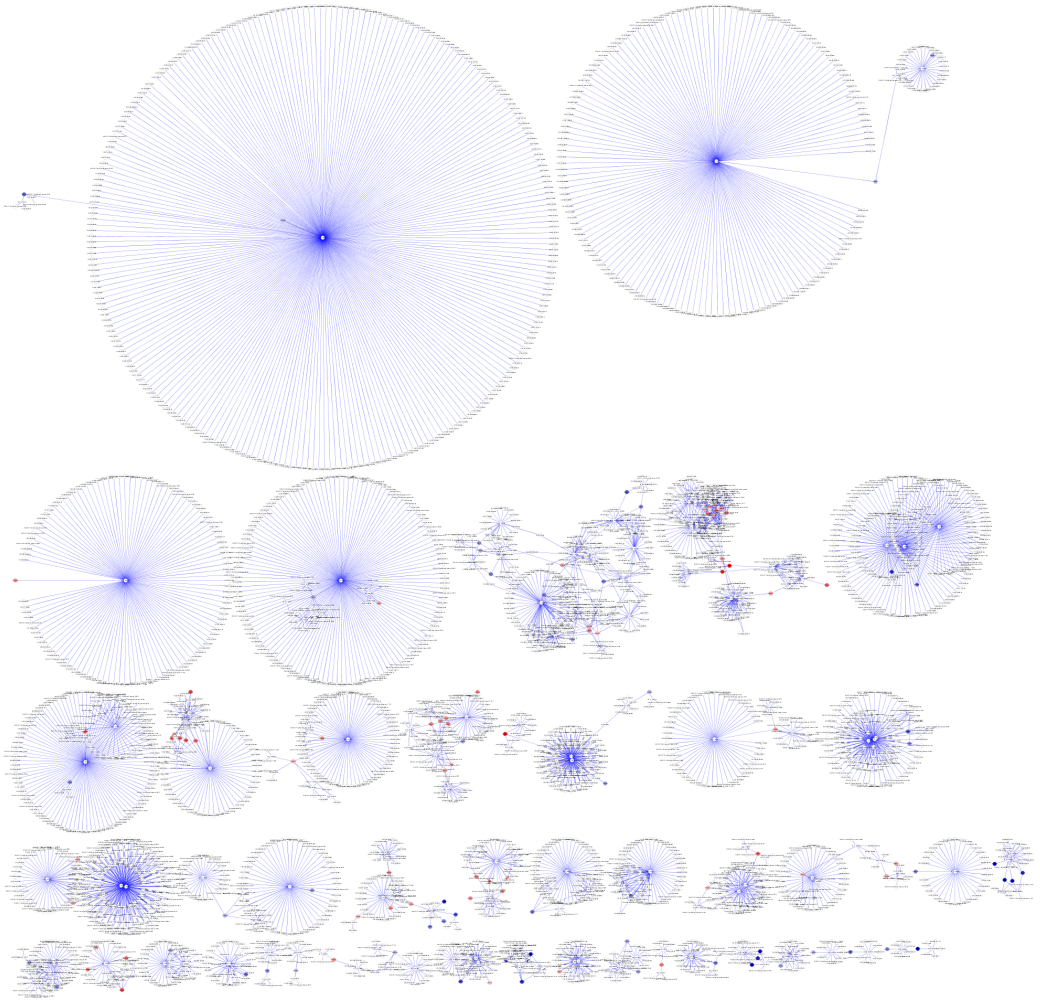


Figure 5.8: Annotation network defense subgraph. After querying for "defense" a breadth-first search was performed to select the entire annotation cluster and probesets associated with "defense" annotation and a subgraph created.

APX2, APX6), two chloroplastic types (*stromal sAPX, thylakoid tAPX*), and three microsomal (*APX3, APX4, APX5*) isoforms. At least part of the induction of heat shock proteins during light stress in *Arabidopsis* is mediated by H_2O_2 that is scavenged by *APX1*. Expression of the gene is down-regulated in the presence of paraquat, an inducer of photoxidative stress.;*APX1* (ASCORBATE PEROXIDASE 1, MATERNAL EFFECT EMBRYO ARREST 6)

In addition one can see all of the GO Terms associated with these orthologous clusters as shown in Table 5.3, all of which is easily accessible. As such, one can now take advantage of the the Gene Ontology, the motifs/activities delineated by Interpro and the textual summaries found in the descriptive an-

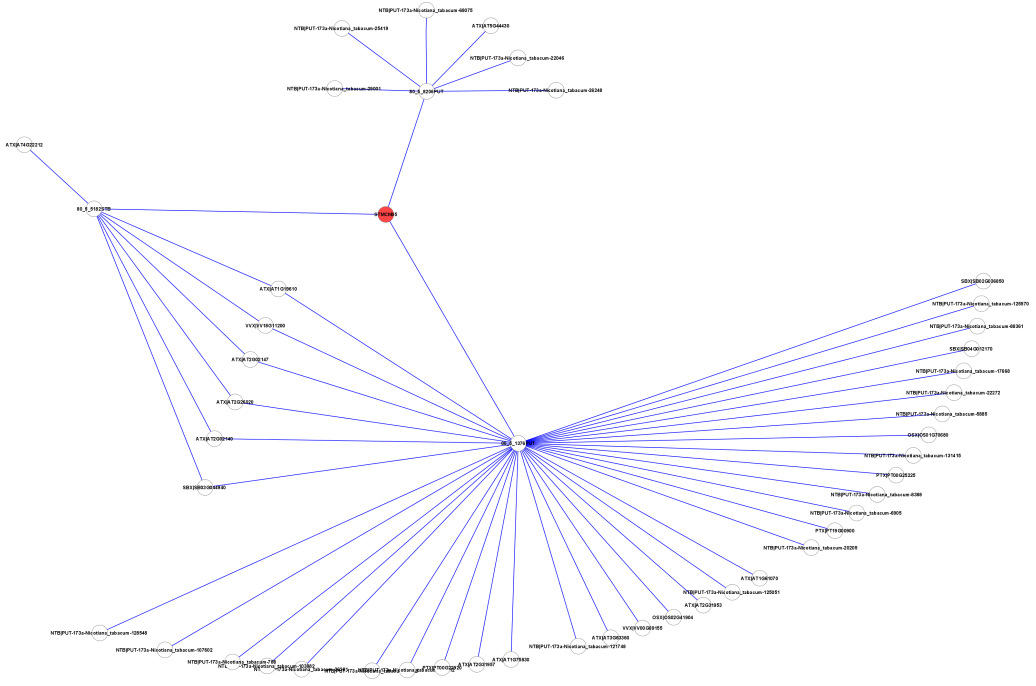


Figure 5.9: Individual cluster. An upregulated probe (red) links to three different ortholgous clusters. Two of the clusters (top and bottom right) are from the tobacco ortholgous cluster analysis and represent two distinct gene families, both of which contain at least one tobacco transcript (EST) likely to cross-hybridise to the potato probe. The third orthologous cluster (left) is from the potato orthologous cluster analysis. It shares many orthogs with the tobacco cluster on the bottom right indicating that the two analyses found largely the same family.

notation that will cover items that may not be covered by pathway databases.

If one were to explore the gene annotation for one of the genes (Peptidyl-prolyl cis-trans isomerase CYP18-3) containing one of these motifs in a bit more detail it becomes apparent that it has previously been associated with disease resistance:

> Peptidyl-prolyl cis-trans isomerase CYP18-3; PPIases accelerate the folding of proteins. It catalyses the cis-trans isomerisation of proline imidic peptide bonds in oligopeptides. Triggers the activation of bacterial *AvrRpt2* protease activity upon infection by *P. syringae*. Activated *AvrRpt2* confers virulence in plant lacking the *RPS2* resistance gene. In plants expressing *RPS2*, the *AvrRpt2*-mediated degradation of *RIN4* activates *RPS2*, which induces hypersensitive response (HR) and plant resistance.

This information, examined collectively, could indicate that *VvPGIP1* over-

Table 5.2: Interpro functional motifs associated with lignin.

Interpro ID	**Domain Name**
IPR001621	Fungal lignin peroxidase
IPR002016	Haem peroxidase, plant/fungal/bacterial
IPR019793	Peroxidases heam-ligand binding site
IPR019794	Peroxidase, active site
IPR002529	Fumarylacetoacetase, C-terminal-like
IPR014159	Protocatechuate 4,5-dioxygenase, alpha subunit
IPR014165	4-carboxy-4-hydroxy-2-oxoadipate aldolase/oxaloacetate decarboxylase
IPR012785	Protocatechuate 3,4-dioxygenase, beta subunit
IPR012786	Protocatechuate 3,4-dioxygenase, alpha subunit
IPR012733	4-hydroxybenzoate 3-monooxygenase
IPR015920	Cellobiose dehydrogenase, cytochrome
IPR008729	Phenolic acid decarboxylase, bacterial
IPR009880	Glyoxal oxidase, N-terminal
IPR008960	Carbohydrate-binding domain family 9-like
IPR012814	Pyranose oxidase
IPR012743	4-coumarate–CoA ligase
IPR020875	3-phenylpropionate/cinnamic acid dioxygenase, alpha subunit
IPR017391	COBRA-like
IPR011234	Fumarylacetoacetase, C-terminal-related
IPR017761	Laccase

Table 5.3: Gene Ontology terms associated with ortholgous clusters associated with lignin-related functional motifs.

apoplast	biological regulation
biological process	carbohydrate metabolic process
cell wall	cellular carbohydrate metabolic process
cellular glucan metabolic process	cellular polysaccharide metabolic process
chloroplast	cytoplasmic part
external encapsulating structure	extracellular region
glucan metabolic process	hydrolase activity, acting on glycosyl bonds
hydrolyzing O-glycosyl compounds	intracellular membrane-bounded organelle
intracellular organelle	intracellular organelle part
L-ascorbate peroxidase activity	licheninase activity
membrane	membrane-bounded organelle
mitochondrion	molecular function
organelle	organelle part
peptidyl-amino acid modification	plasma membrane
plastid	polysaccharide metabolic process
response to abiotic stimulus	response to cadmium ion
response to chemical stimulus	response to cold
response to endogenous stimulus	response to heat
response to inorganic substance	response to metal ion
response to organic substance	response to osmotic stress
response to salt stress	response to stimulus
response to stress	response to temperature stimulus
thylakoid	transferring glycosyl groups
transferase, transferring hexosyl groups	vacuole
xyloglucan:xyloglucosyl transferase	

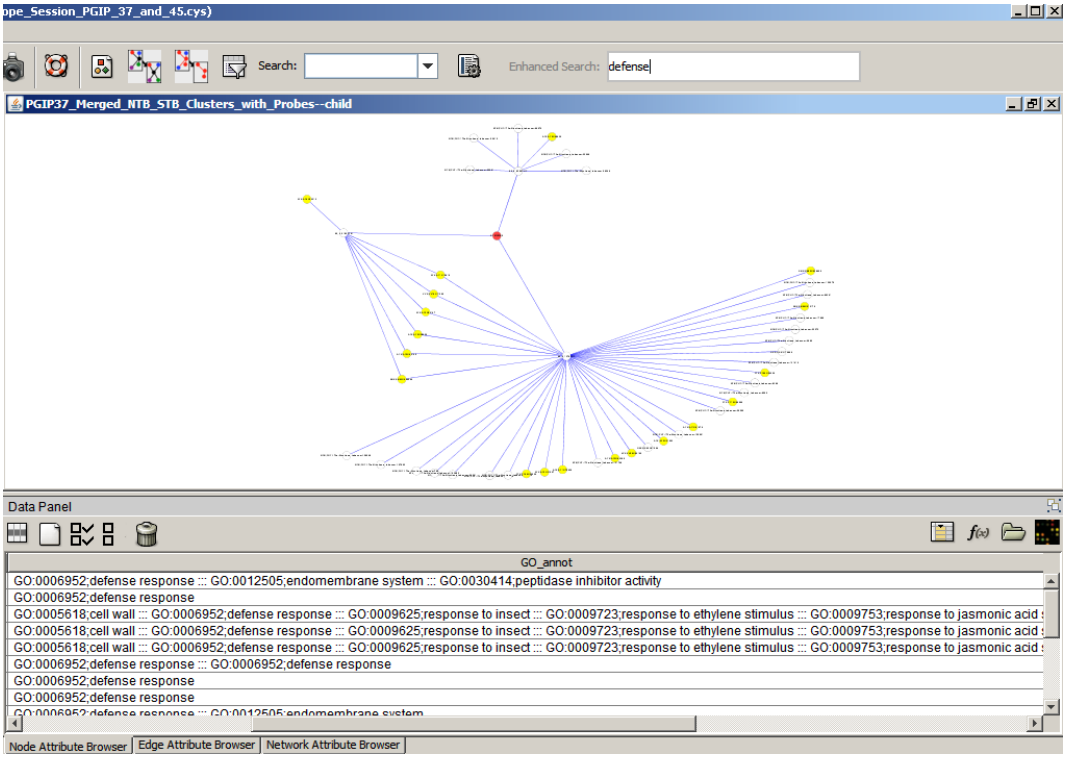


Figure 5.10: Query and annotation interface. The annotation can be queried with any term, or set of terms, fields and boolean operators. Nodes matching the query will be highlighted in yellow and the annotation displayed in the lower pannel. Selected annotation can be exported into a text file for further analysis if needed.

expression starts a signalling cascade that primes the host for a defence response, an idea that was pursued in the paper resulting from this work [2].

5.4.5 Biological Context: Integrating potato probes into extant Pathways and Networks from other Species

5.4.5.1 Metabolic Networks and Pathogen Response Pathways

One of the benefits of the orthologous clusters is that it allows one to pull out *Arabidopsis* ids for each of the differentially expressed spots on the chip. As described in the Methods section this proved to be useful in doing mappings of the probe expression onto pathway databases (Metnet, Mapman, KEGG).

As an example of the usefulness of pathway projection, Mapman visualisation was used in order to get an overview of the changes happening in different areas and functions of the cell. Two examples of this can be seen in Figure 5.12 and Figure 5.13 which show Mapman overviews of the probe data projected onto Mapman *Arabidopsis* bins and diagrams for defence responses and light reactions respectively. In Figure 5.12 it is immediately apparent that there

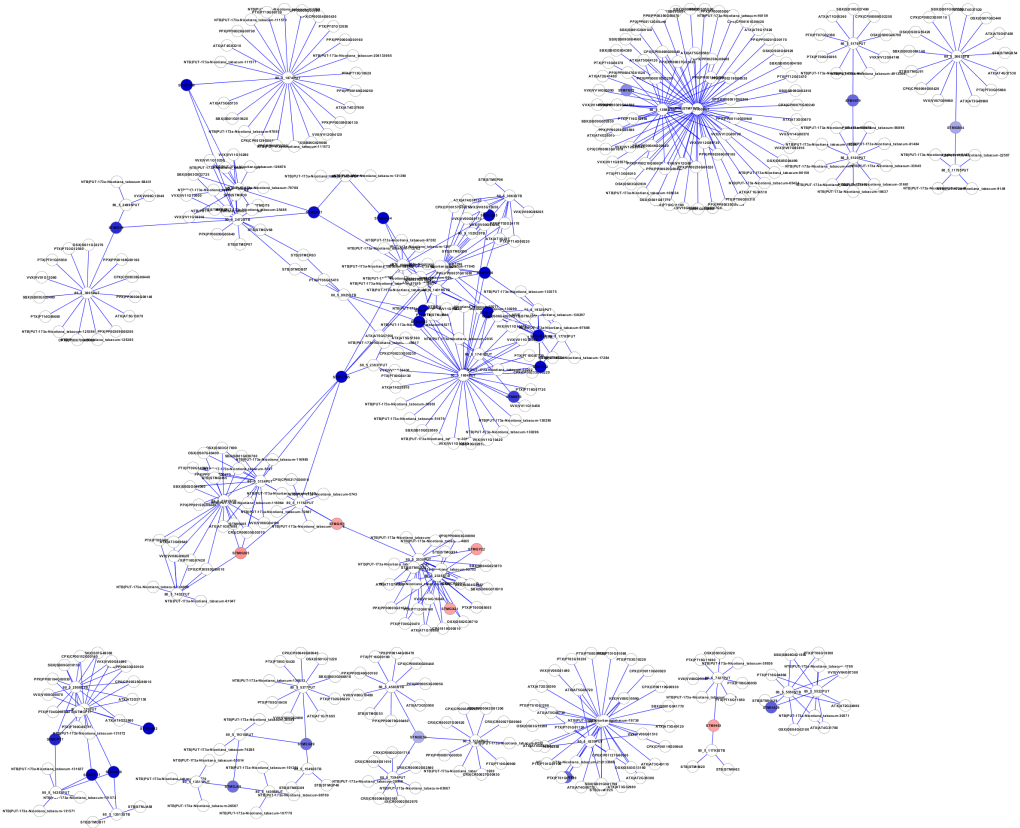


Figure 5.11: Subgraphs associated with Interpro functional motifs that are involved in lignin.

are a large number of changes in genes associated with the cell wall as well as signalling and stress response pathways. Figure 5.13 shows what appears to be transcriptional shifts that will likely change the stoichiometry of the pathways involved in light reactions and energy metabolism.

5.4.5.2 Protein-Protein Interaction Networks

By mapping the probes into protein-protein interaction (PPI) networks one can look for potential associations of the differentially expressed genes which yields different types of associations than one might find in metabolic pathways. For example, the PPIs show an association between the following four proteins:

UVB-resistance protein-like (At5g16040);

Small Ras-like GTP-binding nuclear protein, GTP-binding protein involved in nucleocytoplasmic transport. Required for the import of pro-

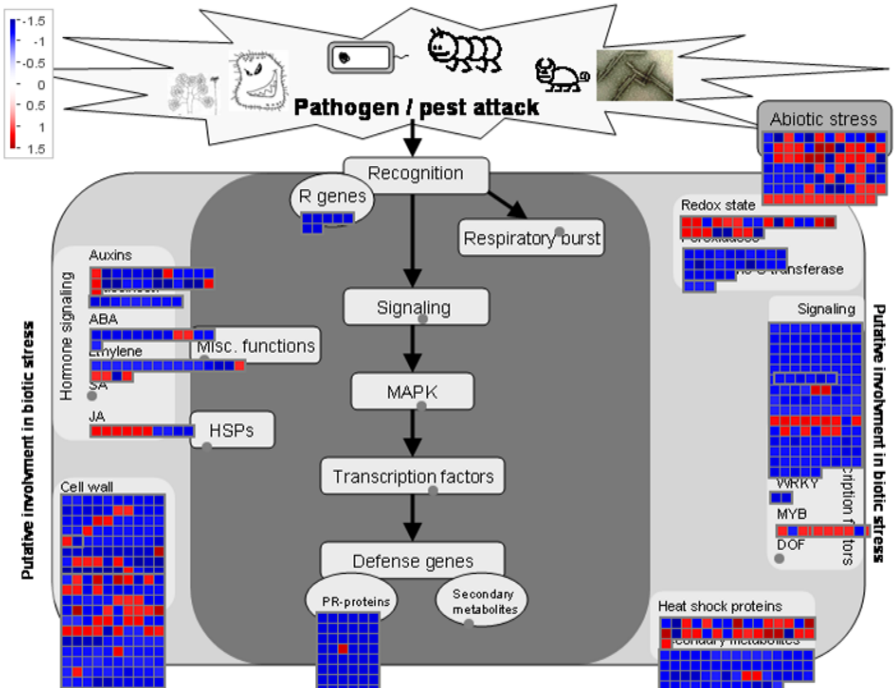


Figure 5.12: Cross-species pathway projection for Mapman visualisation of defence responses.

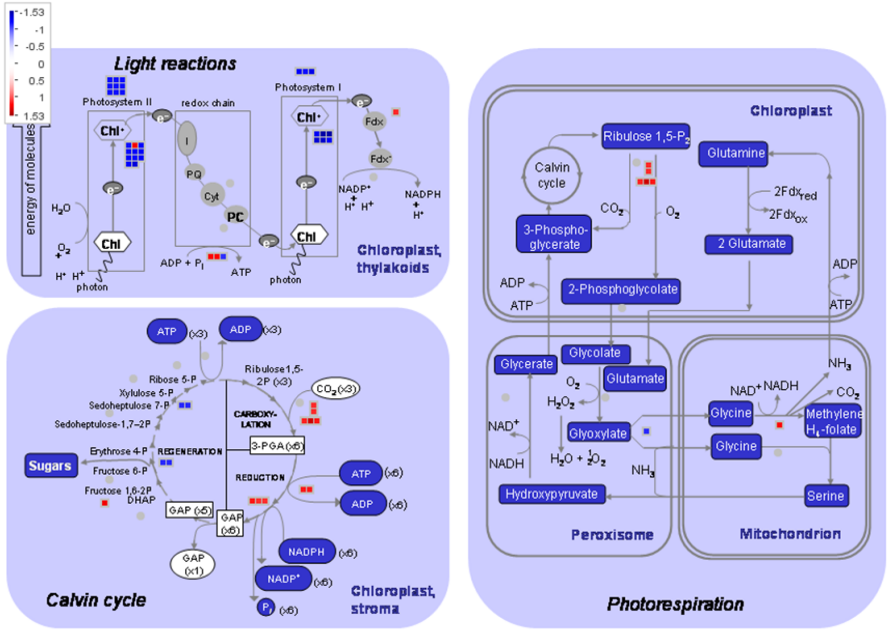


Figure 5.13: Cross-species pathway projection for Mapman visualisation of light reactions.

tein into the nucleus and also for RNA export. Involved in chromatin con-

Table 5.4: Functional motifs associated with orthologous clusters associated with differentially expressed genes identified as associated in a protein-protein interaction network.

AT hook, DNA-binding motif	Cytochrome P450
HMG-I/HMG-Y, DNA-binding	IMP dehydrogenase/GMP reductase
Lipocalin	Lipocalin/cytosolic fatty-acid binding protein
Cysteine peptidase active site	Pleckstrin homology
Pleckstrin homology-type	Ran GTPase
Ras	Ras GTPase
Ras small GTPase, Rab type	Ras small GTPase, Ras type
Small GTPase, Rho type	Small GTP-binding protein
Regulator of chromosome condensation, RCC1	Regulator of chromosome condensation/beta-lactamase-inhibitor protein II

densation and control of cell cycle (by similarity);

GTP-binding nuclear protein Ran-3, GTP-binding protein involved in nucleocytoplasmic transport. Required for the import of protein into the nucleus and also for RNA export. Involved in chromatin condensation and control of cell cycle (by similarity);

GTP-binding nuclear protein Ran-2, GTP-binding protein involved in nucleocytoplasmic transport. Required for the import of protein into the nucleus and also for RNA export. Involved in chromatin condensation and control of cell cycle (by similarity).

When these are used to query the annotation network one finds two clusters that contain forty-eight tobacco transcripts (given that there are roughly eight splice variants per gene this is likely to correspond to roughly six genes). These clusters map to several interesting Interpro functional motifs seen in Table 5.4.

The descriptive Annotation from three of the genes yields the following:

UV-B-specific signalling component that orchestrates expression of a range of genes with vital UV-protective functions. Located in the nucleus and the cytosol. Associates with chromatin via histones. UV-B light promotes URV8 protein accumulation in the nucleus. UVR8 (UVB-RESISTANCE 8);

Similar to regulator of chromosome condensation (RCC1) family protein [*A. thaliana*] (TAIR:AT5G16040.1), similar to hypothetical protein [Cleome spinosa] (GB:ABD96878.1), contains InterPro domain Regulator of chromosome condensation, *RCC1* (InterPro:IPR000408), contains InterPro domain Regulator of chromosome condensation/beta-lactamase-inhibitor protein II (InterPro:IPR009091), Ran GTPase binding / chro-

Table 5.5: Gene Ontology terms associated with ortholgous clusters associated with lignin-related functional motifs.

extracellular region	cell wall
mitochondrion	plasma membrane
response to stress	response to osmotic stress
signal transduction	response to abiotic stimulus
response to salt stress	response to inorganic substance
response to metal ion	membrane
heme binding	signaling process
signaling	signal transmission
external encapsulating structure	response to chemical stimulus
organelle	membrane-bounded organelle
intracellular organelle	intracellular membrane-bounded organelle
organelle part	cytoplasmic part
intracellular organelle part	response to cadmium ion
tetrapyrrole binding	apoplast
regulation of cellular process	response to stimulus
biological regulation	

matin binding;

A member of RAN GTPase gene family. Encodes a small soluble GTP-binding protein. Likely to be involved in nuclear translocation of proteins. May also be involved in cell cycle progression; RAN3, GTP binding.

The Associated GO terms with these clusters can be seen in Table 5.5. The combination of the annotation, the Interpro functional motifs in Table 5.4 and the GO terms in Table 5.5 may indicate that one is looking at part of a stress-related signalling cascade that reaches from the cell wall down into the nucleus. This further buttresses the hypothesis discussed previously that the over-expression of *VvPGIP1* starts a signalling cascade that primes the host for a defence response that was pursued in the paper resulting from this work [2].

5.4.6 Cross-Species Coexpression Analysis

As described in the methods section a hierarchical correlation network was created from over 1700 expression profiles of *A. thaliana* genes that are orthologous to genes associated with differentially expressed potato probes. Figure 5.14 shows the subgraphs of the network that result when all edges that have a Pearson weight of less than 0.8 are deleted. Each of these subgraphs then represent sets of *A. thaliana* genes that are putatively co-regulated across many conditions. As such, these are hypothetical regulatory modules. These networks contain both the Affymetrix probesets as well as the *A. thaliana*

genes which will hybridise to the probesets.

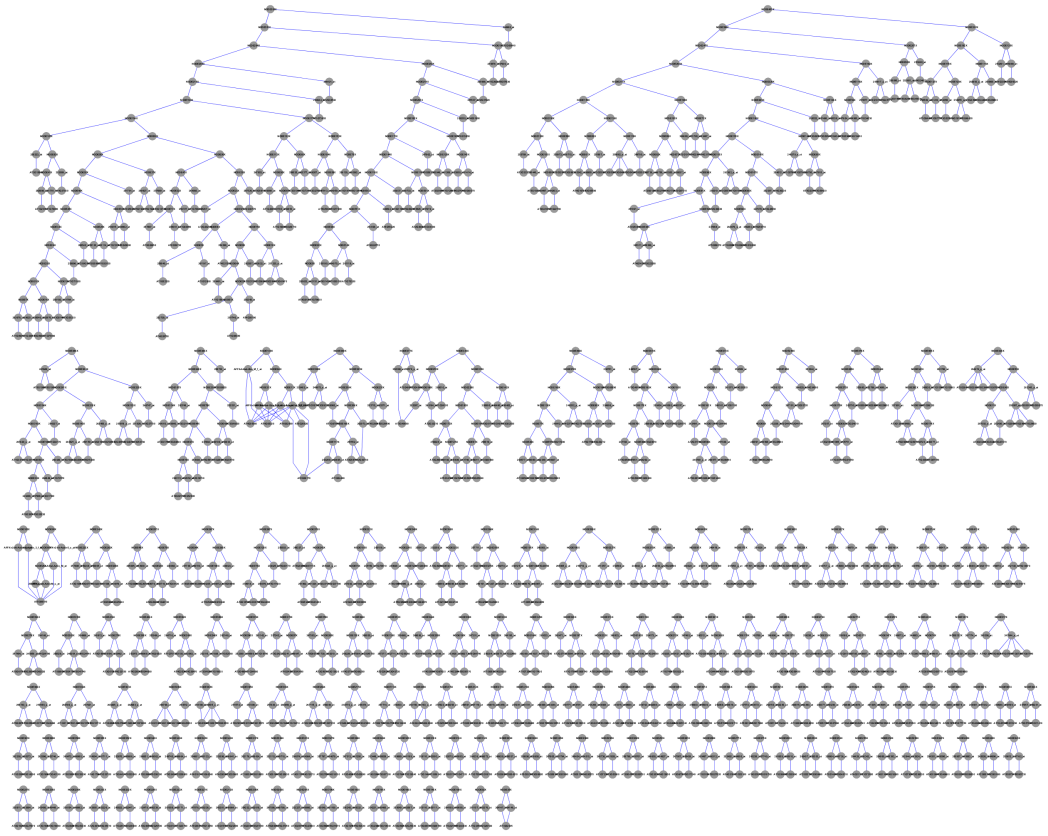


Figure 5.14: Subgraphs of the hierachical correlation network resulting from a 0.8 Pearson threshold applied to a network constructed from the expression profiles of *A. thaliana* genes that are orthologous to genes associated with differentially expressed potato probes.

The *A. thaliana* genes in the network are annotated with GO, gene descriptions and Interpro functional motifs. As such, the network can be queried and subgraphs that contain the query term(s) will be highlighted in yellow. Such queries reveal that the top left cluster contains genes involved in photosynthesis and/or located in the chloroplast and thylakoid membrane. This would seem to indicate that *VvPGIP1* over-expression in tobacco causes central shifts in carbon fixation and energy metabolism via a mechanism that is strongly conserved across species. The top right cluster contains many genes associated with ribosomes which may suggest that *VvPGIP1* over-expression in tobacco may have some affect on protein synthesis/translational regulation via an evolutionarily conserved mechanism.

One can of course zoom in on particular clusters and examine them in considerably more detail. As defensins (anti-microbial peptides) are also of interest a query was performed for defensins which found one particluar subgraph.

The entire subgraph was selected with a breadth-first search and examined in more detail. This subgraph can be seen in Figure 5.15 where one can see the hierachical branch points (NODE393X, NODE272X, NODE169X, etc), the Affymetrix probesets (246252_s_at, 258751_at, etc) and the genes that will hybridise to those probes (AT4G...). The fewer intervening branch points the more correlated probesets are, however, by definition, all of the probesets in this subgraph are at least 0.8 Pearson correlated.

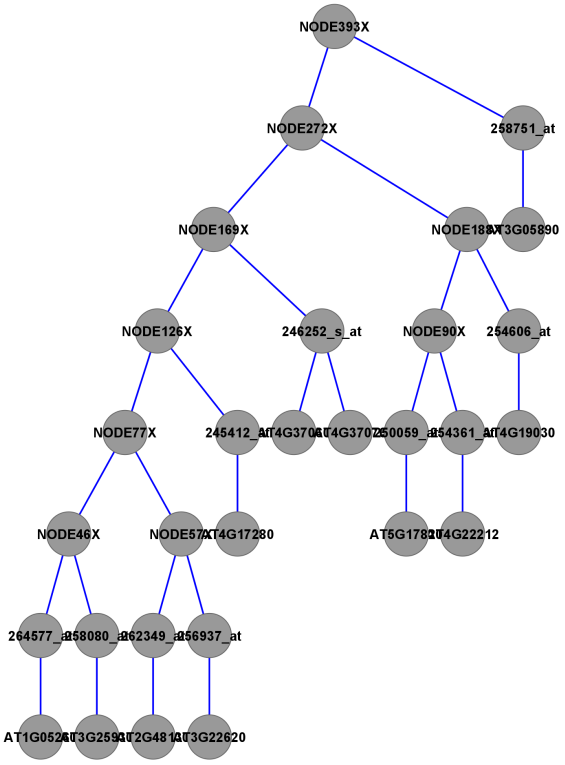


Figure 5.15: Correlation network subgraph containing a defensin gene.

The hierachical network was also integrated with the annotation network described previously. An overview of this integrated network can be seen in Figure 5.16. As shown in Figure 5.17, zooming in on this network structure allows one to see that it contains the hierarchical branch point nodes, the Affimetrix probesets, the *Arabidopsis* genes whose transcripts they will hybridise to, the orthologous clusters that the *Arabidopsis* genes are members of and the potato probes associated with the orthologous clusters as described previously.

Examining the GO terms associated with this putative co-expression subgraph reveals the groups of terms shown in Table 5.6. These terms would seem to suggest that the expression of this defensin is coordinated with that of stress response gene and water transporters as well as with genes associated with the

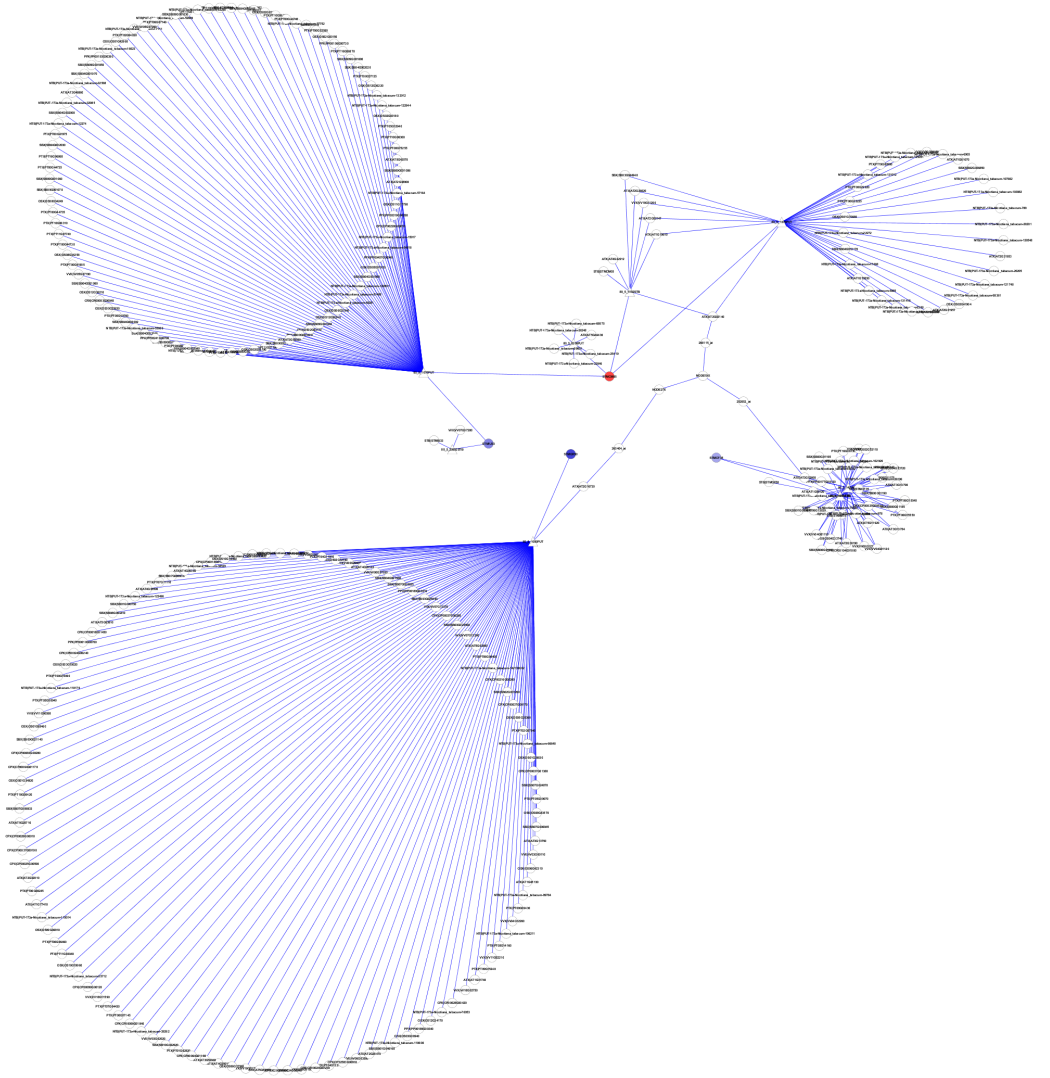


Figure 5.16: Correlation network subgraph integrated with the annotation network.

membrane and cell wall and peroxidase activities. This would be consistent with a readiness to respond to the oxidative stress that is commonly associated with pathogen infection. This hypothesis is currently being evaluated experimentally.

5.4.7 Validation

Since this work was done, a tobacco-specific microarray has become available. Manufactured by Agilent, the new microarray has over 40,000 probes and comparatively little ambiguity when EST collections are matched to probes. This new tobacco microarray has been used for a large follow-up study comparing

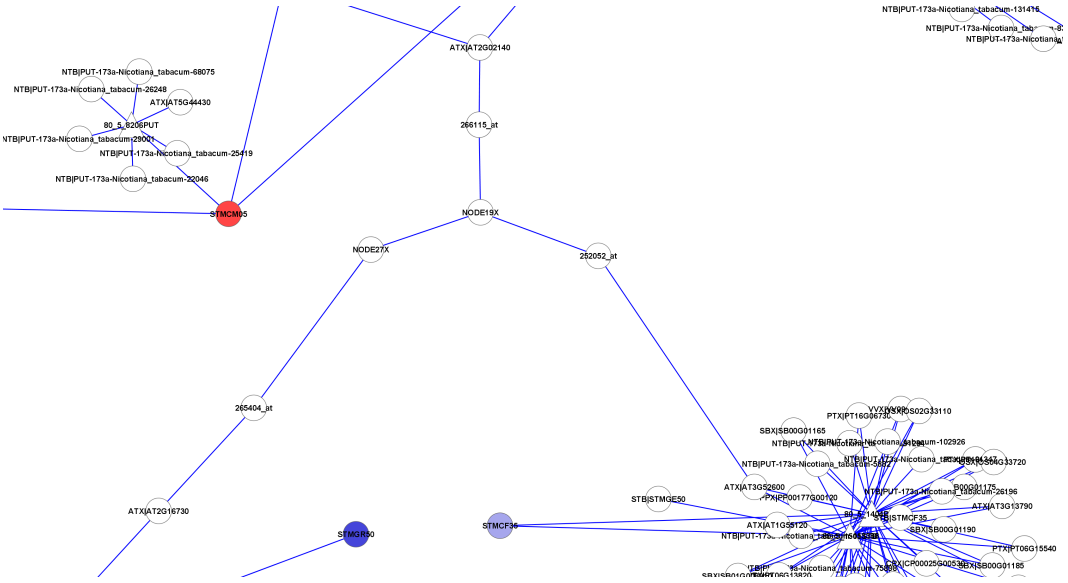


Figure 5.17: Zoomed in view of the integrated correlation annotation network which contains the hierarchical branch point nodes, the Affimetrix probesets, the *A. thaliana* genes that they will hybridise to, the orthologous clusters that the *A. thaliana* genes are members of and the potato probes associated with the orthologous clusters.

the *Botrytis* infection time courses of *VvPGIP1* line 37 vs wild type (SR1). The time zero time points of the two time courses replicates the experiment done with the potato microarray, that is a comparison of the gene expression levels in transgenic line 37 over-expressing *VvPGIP1 vs.* those in the wild type (SR1). GO enrichment and other analysis of the new experiment confirms the patterns and trends seen with the potato microarray (data not shown).

Table 5.6: Gene Ontology terms from putatively co-regulated cluster containing a defensin.

GO Terms
defense response
water channel activity
response to desiccation
response to oxidative stress
response to stress
cell wall
plant-type cell wall
membrane
plasma membrane
anchored to membrane
integral to membrane
lipid binding
lipid metabolic process
lipid transport
peroxidase activity

5.5 Conclusions

Instead of running away from the challenges of hybridisation ambiguity and a lack of annotation, a choice was made to embrace, and indeed, model it. The network structures described in this paper proved to be quite useful in the analysis and interpretation of this challenging experimental data set. The context provided by GO annotation, functional motif description, pathway projections, protein-protein interaction networks and cross-species co-expression networks proved to be invaluable in understanding the underlying biology that was reflected in this data set.

These results have been confirmed with a larger, more specific tobacco microarray. This is actually remarkable as the new microarray has four times as many probes and comparatively little ambiguity, yet tells the same story, even though the experiments were done several years apart. This would seem to suggest that the cross-hybridisation model combined with phylogenomic-based annotation is indeed a robust model for capturing the information necessary to interpret cross-species microarray results.

5.6 Authors Contribution

All computational work described in this paper was performed by DJ. EA and MA provided biological context for the data analysis. Authors' contributions to the associated biological paper [2] are described therein.

5.7 Competing Interests

The authors declare that they have no competing interests.

5.8 Acknowledgements

We would like to thank the South African National Research Foundation and Winetech for funding and our co-authors from [2] for data generation.

References

[1] Alexandersson, E. (2011). Personal Communication. Tech. Rep..

[2] Alexandersson, E., Becker, J.V., Jacobson, D., Nguema-Ona, E., Steyn, C., Denby, K.J. and Vivier, M.a. (2011 January). Constitutive expression of a grapevine polygalacturonase-inhibiting protein affects gene expression and cell wall properties in uninfected tobacco. *BMC research notes*, vol. 4, no. 1, p. 493. ISSN 1756-0500.

[3] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990 October). Basic local alignment search tool. *Journal of molecular biology*, vol. 215, no. 3, pp. 403–10. ISSN 0022-2836.

[4] Berriz, G.F., Beaver, J.E., Cenik, C., Tasan, M. and Roth, F.P. (2009 November). Next generation software for functional trend analysis. *Bioinformatics (Oxford, England)*, vol. 25, no. 22, pp. 3043–4. ISSN 1367-4811.

[5] de Hoon, M.J.L., Imoto, S., Nolan, J. and Miyano, S. (2004 June). Open source clustering software. *Bioinformatics (Oxford, England)*, vol. 20, no. 9, pp. 1453–4. ISSN 1367-4803.

[6] De Lorenzo, G., D'Ovidio, R. and Cervone, F. (2001 January). The role of polygalacturonase-inhibiting proteins (PGIPs) in defense against pathogenic fungi. *Annual review of phytopathology*, vol. 39, pp. 313–35. ISSN 0066-4286.

[7] Duvick, J., Fu, A., Muppirala, U., Sabharwal, M., Wilkerson, M.D., Lawrence, C.J., Lushbough, C. and Brendel, V. (2008 January). PlantGDB: a resource for comparative plant genomics. *Nucleic acids research*, vol. 36, no. Database issue, pp. D959–65. ISSN 1362-4962.

[8] Edgar, R.C. (2004 August). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, vol. 5, p. 113. ISSN 1471-2105.

[9] Evertsz, E.M., Au-Young, J., Ruvolo, M.V., Lim, A.C. and Reynolds, M.A. (2001 November). Hybridization cross-reactivity within homologous gene families on glass cDNA microarrays. *BioTechniques*, vol. 31, no. 5, pp. 1182, 1184, 1186 passim. ISSN 0736-6205.

[10] Federici, L., Di Matteo, A., Fernandez-Recio, J., Tsernoglou, D. and Cervone, F. (2006 March). Polygalacturonase inhibiting proteins: players in plant innate immunity? *Trends in plant science*, vol. 11, no. 2, pp. 65–70. ISSN 1360-1385.

[11] Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R.D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A.F., Selengut, J.D., Sigrist, C.J.A., Thimma, M., Thomas, P.D., Valentin, F., Wilson, D., Wu, C.H. and Yeats, C. (2009 January). InterPro: the integrative protein signature database. *Nucleic acids research*, vol. 37, no. Database issue, pp. D211–5. ISSN 1362-4962.

[12] Joubert, D.A., Slaughter, A.R., Kemp, G., Becker, J.V.W., Krooshof, G.H., Bergmann, C., Benen, J., Pretorius, I.S. and Vivier, M.A. (2006 December). The grapevine polygalacturonase-inhibiting protein (VvPGIP1) reduces Botrytis cinerea susceptibility in transgenic tobacco and differentially inhibits fungal polygalacturonases. *Transgenic research*, vol. 15, no. 6, pp. 687–702. ISSN 0962-8819.

[13] Kars, I., Krooshof, G.H., Wagemakers, L., Joosten, R., Benen, J.A.E. and van Kan, J.A.L. (2005 July). Necrotizing activity of five Botrytis cinerea endopolygalacturonases produced in Pichia pastoris. *The Plant journal : for cell and molecular biology*, vol. 43, no. 2, pp. 213–25. ISSN 0960-7412.

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

[15] Melan, M.A., Dong, X., Endara, M.E., Davis, K.R., Ausubel, F.M. and Peterman, T.K. (1993 March). An Arabidopsis thaliana lipoxygenase gene can be induced by pathogens, abscisic acid, and methyl jasmonate. *Plant physiology*, vol. 101, no. 2, pp. 441–50. ISSN 0032-0889.

[16] Minic, Z., Rihouey, C., Do, C.T., Lerouge, P. and Jouanin, L. (2004 June). Purification and characterization of enzymes exhibiting beta-D-xylosidase activities in stem tissues of Arabidopsis. *Plant physiology*, vol. 135, no. 2, pp. 867–78. ISSN 0032-0889.

[17] Nicholson, R.L. and Hammerschmidt, R. (1992). Phenolic compounds and their role in disease resistance. *Annual review of phytopathology*, vol. 30, no. 1, pp. 369–389.

[18] Proost, S., Van Bel, M., Sterck, L., Billiau, K., Van Parys, T., Van de Peer, Y. and Vandepoele, K. (2009 December). PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *The Plant cell*, vol. 21, no. 12, pp. 3718–31. ISSN 1532-298X.

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

[20] Smyth, G. (2005). Limma: linear models for microarray data. *Bioinformatics and computational biology solutions using R and Bioconductor*, pp. 397–420.

[21] Srinivasasainagendra, V., Page, G.P., Mehta, T., Coulibaly, I. and Loraine, A.E. (2008 July). CressExpress: a tool for large-scale mining of expression data from Arabidopsis. *Plant physiology*, vol. 147, no. 3, pp. 1004–16. ISSN 0032-0889.

[22] ten Have, A., Mulder, W., Visser, J. and van Kan, J.A. (1998 October). The endopolygalacturonase gene Bcpg1 is required for full virulence of Botrytis cinerea. *Molecular plant-microbe interactions : MPMI*, vol. 11, no. 10, pp. 1009–16. ISSN 0894-0282.

[23] Wasmuth, J.D. and Blaxter, M.L. (2004 November). prot4EST: translating expressed sequence tags from neglected genomes. *BMC bioinformatics*, vol. 5, p. 187. ISSN 1471-2105.

[24] Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M. and Barton, G.J. (2009 May). Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics (Oxford, England)*, vol. 25, no. 9, pp. 1189–91. ISSN 1367-4811.

[25] Xu, W., Bak, S., Decker, A., Paquette, S.M., Feyereisen, R. and Galbraith, D.W. (2001 July). Microarray-based analysis of gene expression in very large gene families: the cytochrome P450 gene superfamily of Arabidopsis thaliana. *Gene*, vol. 272, no. 1-2, pp. 61–74. ISSN 0378-1119.

[26] Xu, Z.-S., Xiong, T.-F., Ni, Z.-Y., Chen, X.-P., Chen, M., Li, L.-C., Gao, D.-Y., Yu, X.-D., Liu, P. and Ma, Y.-Z. (2009 January). Isolation and identification of two genes encoding leucine-rich repeat (LRR) proteins differentially responsive to pathogen attack and salt stress in tobacco. *Plant Science*, vol. 176, no. 1, pp. 38–45. ISSN 01689452.

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

5.9 Authors Contribution

All work described in this chapter was performed by DJ. Author's contributions to the associated paper are described therein.

5.10 Supplemental Information

Alexandersson et al. BMC Research Notes 2011, 4:493
http://www.biomedcentral.com/1756-0500/4/493

BMC Research Notes

RESEARCH ARTICLE Open Access

Constitutive expression of a grapevine polygalacturonase-inhibiting protein affects gene expression and cell wall properties in uninfected tobacco

Erik Alexandersson[1,4†], John VW Becker[1,5†], Dan Jacobson[1], Eric Nguema-Ona[1], Cobus Steyn[1], Katherine J Denby[2,3] and Melané A Vivier[1*]

Abstract

Background: Polygalacturonase-inhibiting proteins (PGIPs) directly limit the effective ingress of fungal pathogens by inhibiting cell wall-degrading endopolygalacturonases (ePGs). Transgenic tobacco plants over-expressing grapevine (*Vitis vinifera*) *Vvpgip1* have previously been shown to be resistant to *Botrytis* infection. In this study we characterized two of these PGIP over-expressing lines with known resistance phenotypes by gene expression and hormone profiling in the absence of pathogen infection.

Results: Global gene expression was performed by a cross-species microarray approach using a potato cDNA microarray. The degree of potential cross-hybridization between probes was modeled by a novel computational workflow designed in-house. Probe annotations were updated by predicting probe-to-transcript hybridizations and combining information derived from other plant species. Comparing uninfected *Vvpgip1*-overexpressing lines to wild-type (WT), 318 probes showed significant change in expression. Functional groups of genes involved in metabolism and associated to the cell wall were identified and consequent cell wall analysis revealed increased lignin-levels in the transgenic lines, but no major differences in cell wall-derived polysaccharides. GO enrichment analysis also identified genes responsive to auxin, which was supported by elevated indole-acetic acid (IAA) levels in the transgenic lines. Finally, a down-regulation of xyloglucan endotransglycosylase/hydrolases (XTHs), which are important in cell wall remodeling, was linked to a decrease in total XTH activity.

Conclusions: This evaluation of PGIP over-expressing plants performed under pathogen-free conditions to exclude the classical PGIP-ePG inhibition interaction indicates additional roles for PGIPs beyond the inhibition of ePGs.

Introduction

Polygalacturonase-inhibiting proteins (PGIPs) are extracellular leucine-rich repeat (eLRR) proteins present in plants with recognition and inhibition capabilities towards fungal endopolygalacturonases (ePGs; [1]). ePGs are capable of hydrolyzing the homogalacturonan component of plant cell wall pectin and are among the first enzymes to be secreted during fungal infection. These enzymes play a major role in the virulence of several phytopathogenic fungal and bacterial species [2-5]. Most notably, two ePGs of the necrotrophic fungus *Botrytis cinerea* are required for its full virulence on different plant hosts [4,5].

PGIPs differentially inhibit ePGs from not only a diverse range of fungi, but also different ePG isoforms from the same fungus [1]. This is also well illustrated for grapevine VvPGIP1 and ePGs from *Aspergillus niger* and *B. cinerea*, where differential inhibition towards these ePGs was observed in *in vitro* assays [6]. Further *in vitro* evidence suggests that the interaction and resultant inhibition of ePG by PGIP leads to prolonged existence of molecules with the ability to up-regulate the plant's defense response [7]. Thus, it has been suggested

* Correspondence: mav@sun.ac.za
† Contributed equally
[1]Institute for Wine Biotechnology, Department of Viticulture and Oenology, Faculty of AgriSciences, Stellenbosch University, Stellenbosch, South Africa
Full list of author information is available at the end of the article

that PGIPs protect plants from fungal infection not only by inhibiting fungal macerating enzymes and thereby directly limiting tissue damage, but also by switching on plant defense signaling pathways [8].

Although these *in vitro* experiments have contributed significantly to our understanding of the PGIP-ePGs interaction, recent findings have highlighted the need to better understand the *in planta* roles of PGIP. For example, it was shown that over-expression of *Vvpgip1* reduces the symptoms of BcPG2 from *B. cinerea* in tobacco leaves, without any evidence for an *in vitro* interaction [9]. From this work it appeared that the *in vivo* environment provided a context for this specific PGIP-ePG pair that could not be created *in vitro*. It was proposed that VvPGIP1 might bind to pectin as was shown for bean PGIPs [10] and that VvPGIP1 did not directly inhibit BcPG2, but perhaps rather shielded the most exposed and vulnerable positions in the pectin, thereby indirectly protecting against ePG actions [9].

Numerous studies have reported that high levels of *pgip* gene expression reduce the susceptibility towards *B. cinerea*, confirming the importance of PGIP in plant defense. These include over-expression of the pear *pgip* gene in tomato [11] and grapevine [12], *Arabidopsis pgip* genes over-expressed in *Arabidopsis* [13], a bean *pgip* gene in tobacco [14] and the grapevine *pgip* gene in tobacco [6]. In contrast, silencing of *Arabidopsis Atpgip1* led to enhanced susceptibility [15]. Recently, Veronico et al. [16] reported that the expression level and pattern of a native pea *Pvpgip* could be linked to the degree of resistance to cyst nematode infections. Interestingly, no ePG transcripts could be derived from the cyst nematode nor was a correlation seen between PGIP expression and a native pea ePGs, suggesting that PGIPs have a role in plant-pathogen interactions outside of the classical PGIP-ePG inhibition. More light was shed on the *in planta* role of PGIPs when Kanai et al. [17] presented data from *Arabidopsis* knock-out mutants and over-expressing lines, indicating that *Atpgip1* transcripts prolonged seed germination by influencing pectin degradation in the seed coat. Furthermore, the authors presented evidence that *Atpgips* are under the control of ABI5, a bZIP-type transcription factor that binds to ABRE elements.

Tobacco is a commonly used species to monitor plant-pathogen interactions and is suitable for PGIP over-expression studies since it has negligible PGIP activity against *Botrytis* ePGs [6,18]. We have previously demonstrated that transgenic tobacco plants over-expressing *Vvpgip1* leading to an increased PGIP enzyme activity are less susceptible to *B. cinerea* infection in both detached leaf and whole-plant time-course fungal infection assays [6]. These lines are considered to be PGIP-specific resistant lines (i.e. the resistance phenotype could be correlated with VvPGIP1 over-expression, activity levels as well as ePG-inhibition profiles) and as such provide a valuable genetic resource to study the possible role(s) of PGIPs in plant defence.

To this end, global gene expression, hormone profiling and subsequently, cell wall analysis were conducted on two *Vvpgip1* over-expressing lines with previously characterized resistance phenotypes in the absence of an infecting pathogen. The presence of PGIP, under these pathogen-free conditions, caused altered expression of genes related to a range of mechanisms, including primary metabolism, cell wall organization and metabolism, water transport, photosynthesis and defense responses. An *in silico* cross-species co-expression analysis revealed that many of these gene families were also co-expressed in *Arabidopsis*. Analysis of cell wall components of the plants over-expressing *Vvpgip1* showed a composition similar to wild-type (WT) with only a slight decrease in rhamnose content. However, the change in gene expression was accompanied by higher lignin content, increased level of auxin and, following *Botrytis* infection, a stronger jasmonic acid response. Furthermore, transcriptional down-regulation of a group of xyloglucan endotransglycosylase/hydrolases (XTHs) with key roles in cell wall restructuring and remodeling led to a reduced XTH enzyme activity in the uninfected transgenic tobacco. Taken together, these findings suggest that altered PGIP expression has an effect on the cell wall structure, also affecting fundamental mechanisms such as primary metabolism.

Materials and methods

Plant material and growth conditions

Transgenic *Nicotiana tabacum* SR1 (Petit Havana) *Vvpgip1* line 37 and 45 described in [6] and WT plants were grown in a mixture of soil and peat moss (Jiffy Products International AS, Norway) at 24°C and 55% relative humidity with a light intensity of 120 μmol m^2 s^{-1} over a 16 h light period. Plants were supplemented with liquid organic fertilizer every two weeks (Nitrosol®, Fleuron (Pty) Ltd, South Africa). Leaf material from leaf positions three to five, where leaf three is the youngest and first fully expanded leaf, from healthy 6 to 8-week-old transgenic and control tobacco plants was flash frozen in liquid nitrogen and stored at -80°C for RNA extraction, lignin analyses and phytohormone profiling. For hormone profiling during *Botrytis* infection, *B. cinerea* pathogenic cultures were prepared as described in [6]. Plants were grown in 100% relative humidity and infections performed with four inoculation spots per leaf of 5 μL of a *B. cinerea* spore suspension (1×10^3 spores in a 50% grape juice medium per spot). Infections were allowed to progress for 0, 18, 24 and 30 h before tissue immediately surrounding and including the infection

spots (15 mm diameter) were harvested (separate plants per infection time point were infected and harvested to eliminate wound-response effects). For cell wall component analysis and XTH activity assays, transgenic lines were first established on MS medium supplemented with kanamycin [19] and then transferred to soil and grown under natural light conditions at a controlled temperature. Leaves from leaf position 3, 4 and 5 were harvested when the plants reached the six leaf stage.

RNA extraction and microarray analyses

For microarray analysis leaves of the same age and position from individuals of two transgenic lines were compared to the same WT plant. RNA from 0.5 to 1 g finely ground plant material was extracted using a sodium perchlorate extraction buffer (5 M sodium perchlorate, 0.3 M Tris-HCl pH 8.3, 8.5% polyvinylpolypyrollidone (PVPP), 2.0% PEG 4000, 1.0% β-mercaptoethanol, 1.0% SDS). After shaking for 30 min at room temperature samples were centrifugated and plant debris was removed by passing the supernatant through a syringe plugged with cotton wool. Several phenol/chloroform extractions were performed before precipitating the RNA with 2.5 M LiCl at -20°C overnight. The pellet was washed with 70% ethanol and the resuspended RNA was purified using the RNeasy Mini Kit (Qiagen GmbH, Hilden, Germany). RNA integrity was ensured on 1.2% formaldehyde gels and purity was determined by 260/230 and 260/280 absorbance ratios (>2).

Twenty-five μg RNA was used in each cDNA synthesis reaction in a total volume of 30 μL. Before denaturation at 70°C for 10 min, 2 μL of oligo d(T) primers (500 μg/mL) were added. After denaturation, first strand buffer and DTT were added according to the manufacturer's recommendation (Invitrogen, Carlsbad, CA, USA). A 2:3 (aa-dUTP:dTTP) aminoallyl-dNTP (Ambion, Austin, TX, USA) mix was added (0.5 mM each of dATP, dCTP and dGTP; 0.2 mM aa-dUTP and 0.3 mM dTTP) before incubation at 46°C for 2 min, after which 200 U of SuperScript III Reverse Transcriptase (Invitrogen, Carlsbad, CA, USA) was added. The same amount of enzyme was added following incubation for 4 h at 46°C, after which cDNA synthesis proceeded overnight.

RNA was hydrolyzed by 10 μL 1 M NaOH and 0.5 M EDTA and incubated at 65°C for 15 min. To neutralize the pH, 10 μL 1 M HCl solution was added before unincorporated aminoallyl dUTP and free amines were removed using the RNeasy Mini Kit. cDNA was quantified using a Nanodrop and dried down to volumes of less than 1 μL in a vacuum dryer. Five μL of 0.1 M Na_2CO_3, pH 9.0, was added to the cDNA, and the mixture incubated at 37°C for 10 min. Cy3 or Cy5 (Amersham Biosciences, Buckinghamshire, UK) ester was added and the coupling allowed to proceed for 1 h in the dark at room temperature. Uncoupled dyes were removed by purification with the RNeasy Mini Kit. Probe labeling efficiency of Cy-esters was estimated by measuring the absorbance at 550 nm and 650 nm. Similar amounts of labeled samples were hybridized on TIGR 10 K potato microarrays (version 3). Pre-hybridization was done under lifterslips (Erie Scientific, Portsmouth, NH, USA) with a pre-warmed solution containing 5X saline-sodium citrate (SSC) buffer, 0.1% SDS and 1% BSA at 42°C for at least 30 min. The slides were then washed in five times of deionized water and finally briefly submerged in ethanol before short centrifugation. The combined Cy-labeled probes (28 μL) were mixed with 30 μL of a 2X hybridization buffer (50% formamide, 5X SSC and 0.2% SDS), 1 μL each of COT1 DNA (1 μg/uL) and poly(A)-DNA (12 μg/μL) added, for a total volume of 60 μL, denatured at 90°C for 3 min, and subsequently applied to the slide using lifterslips. Slides were incubated for 16 h at 42°C in hybridization chambers (ArrayIt, Telechem International, Sunnyvale, CA, USA) and successively washed in low stringency (2X SSC, 0.5% SDS; heated to 55°C), medium stringency 0.5X SSC) and high stringency (0.05X SSC) wash buffers for 5 min each and then briefly submerged in ethanol prior to short centrifugation. Slides were scanned with an Axon GenePix 4000A scanner using the GenePix 5.1 software (Molecular Devices, Sunnyvale, CA, USA).

Three microarray slides in total were hybridized: two slides using *Vvpgip1* line 37 and WT, with a dye swap included to account for dye bias; and one slide with *Vvpgip1* line 45 and WT. The result files were analyzed using the Limma package in R 2.9.1 [20]. Spots flagged as "bad" by the GenePix software were removed from further analysis. Background correction was done by the normexp method [21] with an offset of 50 to avoid negative intensities and normalization was then done by print-tip loess. The duplicateCorrelation function was used to estimate a common value for within-array duplicated spots [21]. Finally, fold changes and standard errors were obtained by fitting a linear model to each gene and standard errors were smoothed by empirical Bayes. Genes with a p-value below 0.05 after false discovery rate (FDR) control were regarded as significant. The microarray data was deposited in GEO (GSE26324).

Real-time quantitative PCR

For RT-qPCR analyses RNA was DNase treated (Qiagen GmbH, Hilden, Germany). cDNA was synthesized using SuperScript III Reverse Transcriptase according to the manufacturer's specifications (Invitrogen, Carlsbad, CA, USA) using 1 μg of RNA. Both oligo d(T) and random primers were added to obtain full length cDNA. cDNA samples were diluted 1:25 in dH_2O before 5 μL of

sample was added to 15 μL of LightCycler FastStart DNA Master SYBR Green I mix, prepared according to the manufacturer's recommendations (Roche Diagnostics GmbH, Mannheim, Germany) with a final primer concentration of 500 nM. RT-qPCR was performed using the LightCycler Instrument (Roche Applied Science).

Transcript specific primers were designed using Primer Express (Applied BioSystems) with default settings. Primer sequences for the tobacco genes xyloglucan endotransglycosylase (*XTH*, Genbank Acc AB017025.1) and tubulin used as reference gene (*TUB*, Genbank Acc AB052822) were: *XTH* forward 5'-AGTCCAAGTTTG-TAACACC-3' and reverse 5'-TCTGTCCTTAGTG-CATTCTG-3', amplification product 175 bp; *TUB* forward 5'-TCTGGCTGCTCTGGAAA-3' and reverse 5'-GCATACAAGACACCATCAAAT-3', amplification product 197 bp. cDNA amplification conditions were as follows: denaturation at 95°C for 10 min, followed by 45 cycles of denaturation, 95°C for 10 s; primer annealing at 58°C for 10 s and primer extension at 72°C for 8 s, during which data acquisition was performed. Melting curve analysis was performed by increasing the temperature by 0.1°C/s in the interval 65°C to 95°C. PCR efficiencies for each sample were calculated using LinRegPCR software [22]. The efficiencies were used to calculate relative expression in a mathematical model [23].

XTH activity determination

To determine the XTH activity in tobacco leaves, a dot-blot assay based on the method described in [24] was used. Whatman 3MM paper was coated with 1% (w/v) Tamarind seed xyloglucan (Megazyme) dissolved in aqueous 0.5% (w/v) 1,1,1-trichloro-2-methylpropan-2-ol (Sigma-Aldrich, Steinheim, Germany). Eight leaves were harvested per leaf position and frozen in liquid nitrogen. Two leaves from two individual plants were pooled to ensure adequate amounts of material. This was done for all the lines tested and for leaf position three, four and five. Four extractions were done per biological repeat and each extract's activity was measured in triplicate as described above. Boiled extracts served as negative control, while a batch of cauliflower extract served as positive control for each assay [24].

Total protein was extracted from 20 mg of freeze-dried material for 6 h in 1 mL 50 mM NaOAc, pH 5.5, with gentle agitation at 8°C. The supernatant was collected after centrifugation at 9600*g* for 20 min at 4°C. The protein concentration was determined by the Bradford method [25] with bovine serum albumin as standard. 10 μl of each extract containing 6 nmol of the SR-conjugate (XLLGol-SR) was spotted on the 3MM paper at 4°C and then incubated at room temperature for 12 h between cellulose acetate sheets and several tissue papers to apply an even load. It was then washed for 4 h in 100 mL of 50% ethanol, rinsed in acetone and dried.

The fluorescence was measured with the IVIS® 100 Imaging System (Caliper Life Sciences) using the DsRed filter (570 nm excitation, 615 nm emission) with a 1 s exposure. The Living Image software (Caliper Life Sciences, Hopkinton, MA, USA) was used to identify regions of interest (ROI) and quantify the total efficiency per cm^2 for each fluorescent spot. Background fluorescence was subtracted from all values, which were then normalized to total protein.

Cell wall analysis

Cell wall materials were extracted according to a protocol modified from [26]. Briefly, frozen tobacco leaves were immersed in liquid nitrogen and ground with a mortar and a pestle into a fine powder. Ground material was subsequently boiled with 80% ethanol for 20 min, washed in methanol:chloroform (1:1) for 24 h and finally washed with methanol. Tobacco tissues were left to incubate in methanol:chloroform (1:1) for another 24 h, due to the high level of lipids and others fatty materials (e.g. waxes) present. The remaining material (also called alcohol-insoluble residues, AIR) was dried in an oven at 70°C.

Total cell wall monosaccharide composition was determined on transgenic tobacco over-expressing *Vvpgip1* line 37 and WT. AIR (2 to 4 mg) from different leaf position (3 to 5) were hydrolyzed using 2 M trifluoroacetic acid (2 M TFA, 2 h at 110°C), followed by a 18 h methanolysis at 80°C with dry 2 M methanolic HCl. The generated methyl glycosides were converted into their TMS derivatives at 80°C and separated by gas chromatography (Hewlett Packard 5890 series II). The gas chromatographer was equipped with a flame ionization detector. The oven temperature program was 2 min at 120°C, 10°C/min to 160°C, and 1.5°C/min to 220°C and then 20°C/min to 280°C. Monosaccharides were identified based on their retention time and quantified by determination of their peak areas. The GC-FID was calibrated by using a range of increasing concentration of a mixture of our nine standard sugars and the sugar composition was expressed in molar percentage of monosaccharide. Myo-inositol (90 μL of 1 mg/mL solution) was used as an internal standard. In the final analysis, glucose was removed because of likely contamination of the cell wall fractions by starch-derived glucose.

Histochemical lignin assays and lignin quantification

The lignin content of the WT and *Vvpgip1* lines 37 and 45 was estimated in leaf sections with potassium

permanganate staining followed by transmission electron microscopy according to [27]. Additionally, the lignin content in the stems of the same lines was estimated by staining with a solution of phloroglucinol according to [28].

Quantification of lignins in the WT and transgenic lines *Vvpgip1* line 37 and 45 was done using the acetyl bromide method as described in [29], with some modifications as described below. Dried tobacco leaf material (leaves three to five of eight week old plants) was utilized to isolate AIR, consisting of cell wall material. Leaf tissue was ground in liquid nitrogen and extracted twice with 80% aqueous methanol following homogenization. Following centrifugation at 12 000x*g* for 10 min, the pellets were washed three times with 96% ethanol and twice with a solution of 96% ethanol:hexane (2:1). The resulting AIR pellets were dried overnight at 70°C. Five to ten mg of the AIR was used to determine the percentage of lignin contained therein. The AIR was washed with 25% acetyl bromide (in acetic acid), after which it was incubated in 1 mL of the same solution at 70°C for 30 min. The mixture was cooled to room temperature, and 0.9 mL of NaOH and 0.1 mL of hydroxylamine hydrochloride (0.1 M) added. The volume was subsequently adjusted to 10 mL with acetic acid. The solution was incubated overnight and the absorbance measured at 280 nm with a procedural blank. The lignin content of the samples was calculated as follows: % lignin content = (absorbance × 100)/(SAC × AIR (g l^{-1})); where SAC is the specific absorption coefficient of lignin, for which the value of 20 gl^{-1} cm^{-1} was used.

Analysis of phytohormones using GC/MS

Leaf tissue flash frozen in liquid nitrogen was extracted according to the method of Schmelz et al. [30], with some modifications as described below. Approximately 100 mg of tissue was ground to a fine powder prior to the addition of extraction solvent (n-propanol/water/HCl) and internal standard (*o*-anisic acid), of which 30 ng per sample was added. Samples were vortexed to ensure homogenization prior to partitioning with dichloromethane. For the conversion of phytohormone acids to their corresponding methyl esters, the organic phase (dichloromethane/propanol) was derivatized in 4 mL glass vials for 30 min using 4 μL of a 2 M trimethylsilyldiazomethane solution in hexane. The activity of the derivatization agent was subsequently quenched with 4 μL of a 2 M acetic acid solution, also in hexane.

Vapor phase extraction of the derivatized organic phase proceeded according to [30], with the exception that commercially available Super Q filters were used (Analytical Research Systems, Inc., Gainesville, FL, USA). Briefly, the derivatized sample was evaporated at 70°C and passed through a Super Q filter under a N_2 flow of 500 mL/min. To ensure complete vaporization of less volatile compounds the vial was subsequently heated to 200°C for 2 min while passing the vapor through the filter.

The analytes were eluted from the Super Q filter with 150 μL CH_2Cl_2 and analyzed by a Trace Gas Chromatograph (GC) (ThermoFinningan, Milan, Italy) coupled to a Trace Mass Spectrometer (MS) (Thermo MassLab, Manchester, UK). GC/MS conditions were amended from that described in [30]. Briefly, 2 μL of the dichloromethane (BDH, Poole, England) eluent was injected in the split/splitless injector of the GC, operated in the splitless mode (purge time 3.5 min, 50 mL/min) at 280°C. Compounds were separated on a Factor Four VF5-MS capillary column (Varian, Palo Alto, CA, USA) with dimensions 30 m 1. × 0.25 mm i.d. × 0,25 μm f.t. Flow of the carrier gas (Helium) through the column was 0.7 mL/min in the constant flow mode. The oven program used was 40°C, hold 1 min, ramp 15°C/min to 250°C, hold 5 min, ramp 20°C/min, and hold 2 min. In order to avoid carryover a post-run was performed after each analysis at 280°C under a head-pressure of 300 kPa. The temperature of the MS interface was kept at 280°C and the source at 200°C. The MS-detector was operated in Electron Impact (EI) mode at 70 eV and Selected Ion Monitoring (SIM) mode. The electron multiplier voltage was set at 500 V. Three carboxylic acid methyl ester analytes were detected and quantified using SIM with retention times and ion mass to charge ratios (m/z) as follows: methyl salicylate (8.39 min, m/z 92, 120, 152); methyl jasmonate (12.38 min, m/z 83) and methyl indole-3-acetate (13.81 min, m/z 130). The internal standard, *o*-anisic acid methyl ester was eluted at 9.70 min with m/z 92, 120 and 152.

For quantification the internal standard method was used. Calibration curves were constructed for each analyte over the range from 2 to 200 ng.mL^{-1}. The regression equations and their correlation coefficients obtained for SA, MeJA and IAA are detailed respectively: y = 0.0370x+0.2511 (r^2 = 0.9924), y = 0.0087x +0.0303 (r^2 = 0.9954) and y = 0.0120x + 0.0238 (r^2 = 0.9964). The limit of quantification (LOQ) was established to be 2 pg for all analytes. Salicylic acid, indole-3-acetic acid, methyl jasmonate, *o*-anisic acid, trimethylsilyldiazomethane, hexane and 1-propanol were purchased from Sigma-Aldrich (Steinheim, Germany).

Bioinformatics workflow for probe evaluation and annotation

The interpretation of the microarray results was challenged by the potential ambiguity in probe to transcript hybridizations and incomplete annotations of the genes of the two plant species involved.

Probe Specificity and annotation

BLAST was used to map all of the extant *N. tabacum* ESTs (PlantGDB-derived unique transcripts version number 173a; [31]) to the probes they would likely hybridize to on the TIGR 10 K potato microarray and the result stored as a graph denoting the relationships between potential transcripts and probes. A threshold of 80% identity over at least a 100 bp region was used to identify transcripts that would be likely to hybridize to microarray probes.

As annotated genome sequences are available for neither *S. tuberosum* nor *N. tabacum*, EST datasets were used for annotation based analysis. However, the annotation was incomplete and out of date for both the EST datasets for *S. tuberosum* (from which the probes were designed) and *N. tabacum*. Prot4EST [32], which encompasses BLASTX [33], ESTScan [34] and DECODER [35] was modified in order to run on 100 × 2.83Ghz cores high performance computing architecture (Stellenbosch University) and the EST sequences from both *S. tuberosum* and *N. tabacum* translated into their corresponding protein sequences. The PLAZA [36] protein translations and associated GO and Interpro annotation for the complete genomes of *Ostreococcus lucimarinus, Chlamydomonas reinhardtii, Physcomitrella patens, Sorghum bicolor, Oryza sativa, Vitis vinifera, Populus trichocarpa Carica papaya* and *Arabidopsis thaliana* were downloaded. BLASTP and OrthoMCL [37] were used to create orthologous clusters based on sequence similarity of all proteins resident in PLAZA with the translated *S. tuberosum* and *N. tabacum* ESTs. GO and Interpro annotations found for members of the resultant orthologous clusters were projected onto the orthologous members for *S. tuberosum* and *N. tabacum* proteins. Subsequently, with the use of the probe-to-transcript graph described above, annotations from both *S. tuberosum* and *N. tabacum* were assigned to the corresponding probes.

A graph structure was created to represent all of the resulting clusters, their associated annotation, assignment to probes and corresponding expression values. The resulting graph structure stored in XGMML and Cytoscape [38] were used for visualization and querying. In order to create Additional file 1, a Perl program was written to parse the graph structure, extract all connected components (e.g. clusters described above), and parse the BLASTP results to extract the most similar protein in *Arabidopsis* or rice (if no *Arabidopsis* match was found).

GO Enrichment

Significantly differentially expressed probes were determined by Limma as described above. The projected annotation for the differentially expressed probes was then analyzed for Enrichment of Gene Ontology Terms by GOEast using default settings [39].

Pathway and cross species co-expression analysis

BLAST was subsequently used to map all of the sequences associated with *Arabidopsis* gene identifiers onto specific probesets on the ATH1 Affymetrix microarray (thus yielding a direct map of *S. tuberosum* orthologous clusters to *Arabidopsis* probesets). A program was written in Perl to download the expression data from over 1700 *Arabidopsis* Affymetrix microarrays available at CressExpress [40]. Vectors of each probeset in the Affymetrix microarrays were created, including expression information from each microarray in the collection. Each of the differentially expressed *S. tuberosum* orthologous clusters were then analyzed for co-expression in *Arabidopsis* using these vectors and a Hierarchical Clustering Algorithm testing Pearson coefficient correlations thresholds of 0.05, 0.1 or 0.2 to select the clusters representing the highest levels of co-expression.

Visualization of affected genes divided into metabolic pathways or other processes was done by MapMan using a mapping file previously developed for the potato TIGR 10 K microarray [41].

Statistical analysis

Excel (Microsoft) and GenStat (VSN International) were used to generate Student's t-tests and ANOVAs as indicated in the text.

Results and discussion

Transgenic plants from various species with increased levels of polygalacturonase-inhibiting proteins (PGIPs) are known to have better protection against pathogenesis by *B. cinerea* and Pierce's disease [11-14], whereas *Arabidopsis* plants with *pgip* silencing show increased susceptibility towards *Botrytis* infection [15]. Clearly, PGIP expression levels affect pathogen infection and host-pathogen interaction.

The two *Vvpgip1*-overexpressing tobacco lines *Vvpgip1* line 37 and 45, which are the focus of this study, were selected for further profiling from a population of transgenic lines considered to have PGIP-specific resistance phenotypes, since *pgip* over-expression, PGIP activity and ePG inhibition correlated with resistance against *B. cinerea* as shown Joubert et al. [6]. In order to shed more light on the possible processes associated to PGIP expression *in planta* and the role of PGIP leading to increased pathogen resistance, gene expression and hormone profiling was conducted. These analyses together with biochemical analyses of the cell wall confirmed that healthy un-infected PGIP transgenic plants have altered gene expression, hormone levels and cell wall structure.

Global gene expression profiling

When the study was initiated, no microarray platform for tobacco was available. Instead, genome-wide expression in tobacco plants was monitored by the TIGR 10 K potato microarray, since these two species in the *Solanaceae* family have a high degree of sequence conservation and the heterologous hybridization system using microarray probes for potato ESTs has previously been used and evaluated [42]. Unfortunately, the latest annotation for the potato microarray was created in 2006, and thus did not contain the latest information on gene function. Further complicating matters is the potential ambiguity introduced by cross-hybridization that is bound to occur under heterologous hybridization.

To address these issues, the degree of cross-hybridization of tobacco transcripts to the potato probes was modeled by sequence similarity analysis of all of the tobacco ESTs as compared to the sequences reported for each probe on the microarray. This was achieved by generating a graph structure based on probe sequences and tobacco ESTs, which was used to estimate probe specificity and redundancy. Figure 1 gives an example overview of the graph structure, which were also used to update and re-annotate the microarray's probe annotations based on the tobacco transcripts likely to hybridize to a certain probe and by creating orthologous clusters based on the comparisons to nine sequenced plant genomes included in PLAZA [36]. By this approach a table of GO terms associated to each probe was generated and subsequently used for GO term enrichment analysis (Additional file 2). This file can be downloaded for future use when analyzing gene

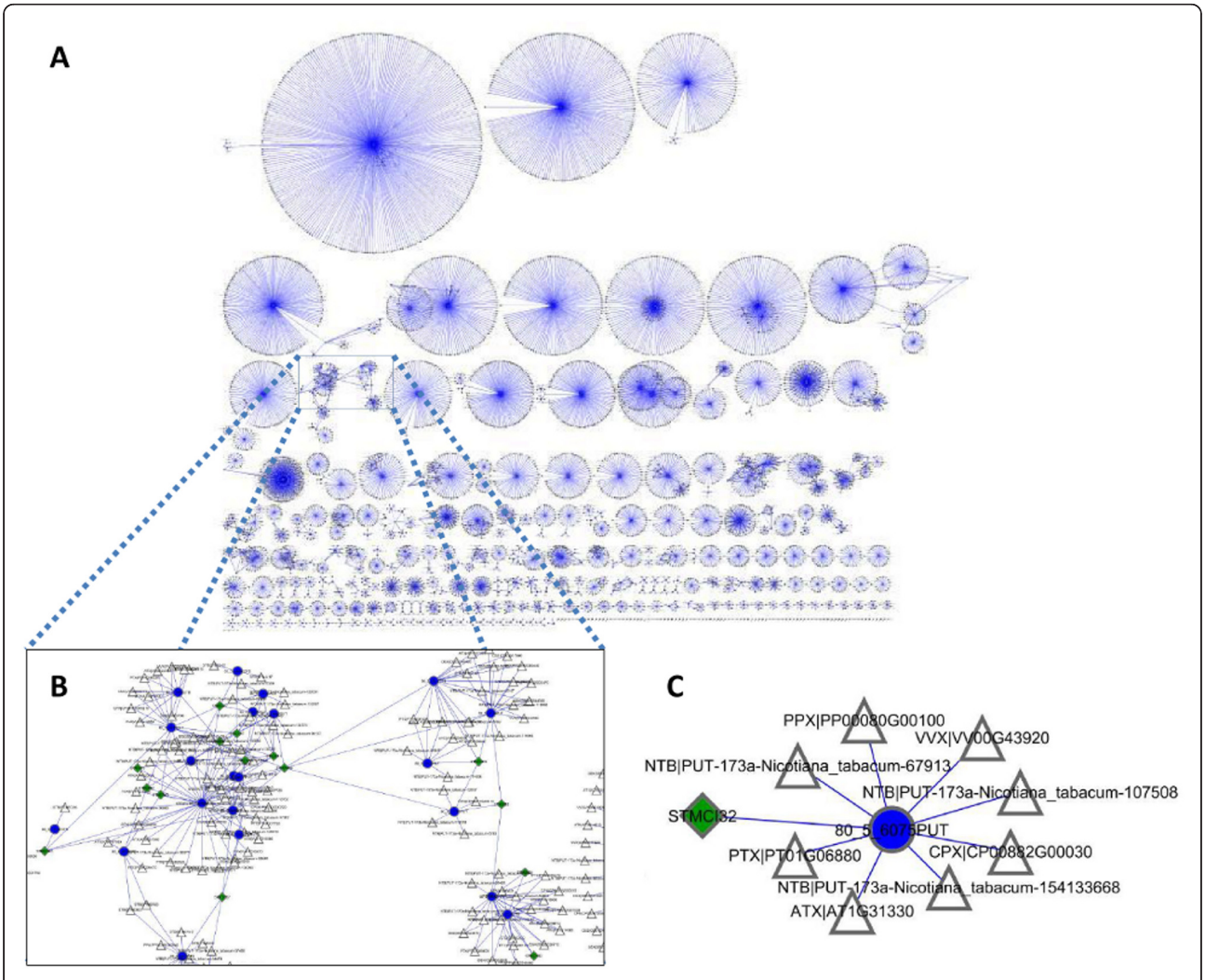


Figure 1 Example of the graph structure. The graph structure was used to assess possible cross-hybridization and probe redundancy as well as to improve annotation of probes. **a**. Graph structure of probes with significant change in signal in *Vvpgip1* line 37. **b**. Detailed view of part of the graph structure showing differentially expressed probes in *Vvpgip1* line 37. **c**. Example of an orthologous cluster. Probes are marked as *green squares, blue circles* represent a unique identifier for the orthologous cluster and matching transcripts are included as unfilled triangles.

expression data generated by the TIGR 10 K potato (version 3) microarray. Heterologous hybridization on arrays has been estimated before [43,44]. However, we are unaware of a study that also takes into account the affect of cross-hybridization on the probe annotation of the microarray and subsequent GO term enrichment analysis.

By processing expression data in Limma a total of 318 probes were found to be differentially expressed with a false discovery rate (FDR) of ≤ 0.05 in both *Vvpgip1* line 37 and 45 in comparison to WT (Table 1). Because the two transgenic lines most likely have different genotypic backgrounds due to the independent insertion events during transformation, we regard this approach to be technically more correct than considering the two lines as biological replicates by combining the expression data prior to Limma analysis. It is also a more conservative approach since considering the lines as biological replicates would almost double the number of probes (735) showing significant differential expression. A complete list of probes with significant difference in expression is included as Additional file 1 and a selection in Additional File 3. The majority of the probes were down-regulated, whereas only 58 were up-regulated in comparison to WT plants (Table 1). The remaining 41 probes showed opposite regulation in the two lines. The oppositely regulated probes can be a result of positional effects related to the insertion of the *Vvpgip1* expression cassette in the tobacco genome and thus reflect true differences between the transgenic lines. Alternatively, it might arise from rapid fluctuation in the abundance of certain transcripts caused by PGIP over-expression.
Differentially expressed probes were analyzed by examining the graph structure of orthologous clusters and by examining top hits based on sequence identity (Figure 1; Additional file 1). Because of the probe ambiguity and the likelihood that many probes bind different transcripts, the exact number of differentially expressed genes monitored could not be determined. In fact, since the tobacco genome still remains un-sequenced, probe redundancy and specificity can only be estimated at this point. However, based on the graph structure we approximate that the 318 differentially expressed probes correspond to ca 250 genes.

Table 1 Differentially expressed probes in *Vvpgip1* line 37 and 45 in comparison to WT (FDR, $p < 0.05$)

Vvpgip1 line 37	*Vvpgip1* line 45	Probes
Down	Down	219
Down	Up	16
Up	Down	25
Up	Up	58
	Total	**318**

A manual categorization of clusters and probes were done with the help of the MIPS 2.0 database [45], and probes were consequently divided into functional groups according to Additional file 1. Probes for which no sequence information was available are also listed if a significant change in signal intensity was observed. More than 15% of all clusters identified could not be functionally classified since the level of identity to other sequences was too low or no informative annotation could be inferred from related genes.

Even if the changes in expression were generally subtle there was an over-representation of certain functional groups among the differentially expressed genes when dividing genes by GO terms or MapMan categories. GO term enrichments were based on probes with a significant change in signal in both transgenic lines in comparison to the wild-type (WT). The full graphical representation of enriched terms can be found in Additional file 4.

Four broad groups of affected probes fall under the categories of cell wall biogenesis and organization, carbon metabolism, photosynthesis and stress defense signaling, whereas more specific groups of interest would be glucan and polysaccharide metabolic processes, water transport as well as response to auxin and brassinosteroid stimuli and to cyclopentones. Biosynthesis of jasmonic acid (JA) requires several cyclopentenone precursors and these have been suggested to be able to fulfill some of the JA roles *in vivo* [46]. Among the molecular function categories are glycolysis, energy transfer, cell wall components and water channel activity. Enrichment of cellular components linked to the chloroplast and mitochondrial compartments further strengthen the picture of differences in metabolism between the PGIP over-expressing lines and WT.

Functional groups of differentially expressed genes between the two transgenic lines and WT plants were visualized in MapMan using a mapping file previously developed specifically for the potato TIGR 10 K microarray [41]. As could be expected from the GO enrichment analysis results, the largest changes were seen in different metabolic pathways. A concordance between the lines was seen for groups of genes linked to cell wall modification and degradation, glycolysis, starch synthesis and photosystem-light reactions (Additional file 5).

In order to examine whether the changes in gene expression coupled to an augmented level of *Vvpgip1* expression seen in this study were similar to the expression pattern observed in other plant systems, the co-expression of *Arabidopsis* genes associated to the orthologous clusters based on sequence similarity were investigated. Strikingly, several gene families affected in the transgenic lines were also found to be co-expressed in *Arabidopsis*, e.g. members of the XTHs, peroxidases

UDP epimerases and JAZs (Additional file 1; Additional file 3). The similarities in transcriptional patterns related to *pgip* expression suggest that there is a degree of conservation of these effects between the two species. Furthermore, these similarities between our transgenic system and *Arabidopsis* PGIP expression strengthen the notion that the changes seen in expression pattern is due to *Vvpgip1* over-expression *per se* and not due to other general properties of PGIPs, such as possible indirect effects resulting from the fact that VvPGIP1 is a secreted, apoplastic protein as discussed as further discussed in Summary and Conclusions.

Hormone profiling

As part of the general comparison between the transgenic lines with resistance phenotypes to the susceptible controls, a hormone profile of salicylic acid (SA), indole-acetic acid (IAA) and jasmonic acid (JA) was established (Figure 2). Under the non-infecting conditions used in this study, IAA levels were statistically significantly increased in the transgenic lines, SA levels were slightly lower in the transgenic lines compared to the WT, whereas JA levels were below the detection level in both transgenic lines and WT. Hormone profiling was also performed during a *Botrytis* infection time series (Figure 3). After infection, JA was detectable and both transgenic lines displayed higher JA levels than WT at 18 and 24 h after infection ($p = 0.045$ at 24 h) at the local lesions. For SA and IAA no definite differences were seen between the groups of transgenic lines and WT following *Botrytis* infection.

Comparing the hormone data with the gene expression analysis, no genes involved in SA biosynthesis were differentially expressed and there was no enrichment of GO terms related to SA regulation. Thus, in spite of SA's involvement in pathogen defense signaling, no strong link between *Vvpgip1* over-expression and SA and SA-related mechanisms could be established in uninfected or infected tissue. This is in line with an earlier report on PGIP-regulation in *Arabidopsis* [13].

In contrast to SA, indole-acetic acid (IAA) levels without *Botrytis* infection were statistically separable into two groups with the transgenic lines containing higher levels of IAA grouped together (Figure 2). Interestingly, corresponding to the increase in IAA levels there is an over-representation of genes responding to auxin stimulus in the transgenic lines from the microarray data. The down-regulation of the Aux/IAA transcriptional repressors of auxin induced genes, which belong to a large family of transcription factors, could possibly be linked to the over-representation of genes responding to auxin stimulus detected in the *pgip* over-expressing lines. Auxin promotes the degradation of Aux/IAA transcriptional repressors making it possible for auxin response factors (ARFs) to activate the transcription of auxin-responsive genes. Furthermore, increased auxin levels enhance the binding of Aux/IAA proteins to the F-box protein TIR1 leading to ubiquitination and degradation of the Aux/IAA proteins (reviewed in [44]). The expression level of several components of protein degradation and proteolysis were affected, e.g. cullin, a family involved in SCF E3 ubiquitin ligase complexes, ubiquitin

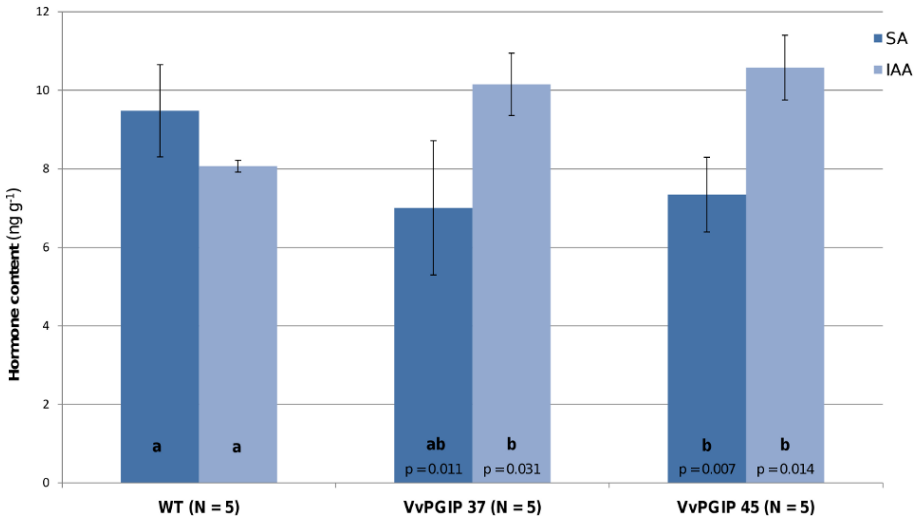


Figure 2 Phytohormone content. Salicylic acid (SA) and indole-acetic acid (IAA) content (ng g^{-1} fresh weight) measured as their corresponding methyl esters in WT and transgenic lines. The total pool of SA (*dark-blue bars*) present in plants as free and methylated forms were analyzed. Both SA and IAA levels (*light-blue bars*) were significantly different from WT in both transgenic lines as indicated by p-values (two-tailed Student's *t*-test). Statistical groups as determined by one-way ANOVA (multi-comparison Bonferroni test, $p < 0.05$) are given as letters. No jasmonic acid (total) was detected in the samples. *N* represents the number of biological repeats and *error bars* are given as two times the standard deviation.

proteins, ubiquitin-protein ligase and F-box proteins as well as a number of proteases. The high number of differentially expressed genes involved in protein degradation might be a consequence of changes in hormone signaling, as ubiquitination and targeting for proteosome degradation are common strategies in plant hormone signaling.

Under uninfected conditions JA levels were below the detection level. However, both transgenic lines had a stronger JA response following *Botrytis* infection than WT at the local lesions (Figure 3). On the transcriptional level under un-infecting conditions, *Vvpgip1* over-expression affected JAZ genes and cyclopentone responsive elements linked to JA regulation and biosynthesis. Like SA, JA was recently suggested to play an important role in signaling leading to Systemic Acquired Response (SAR; [45]). In *Vvpgip* line 37, expression of a LOX gene and a 12-oxophytodienoate-reductase were changed. These enzymes are involved in jasmonate biosynthesis and indicative of an association between PGIPs and JA-mediated signaling. The down-regulation of JAZ genes, which are repressors of JA signaling, indicates that JA signaling pathways might be 'primed' and could explain the quicker response in JA levels observed in the transgenic lines when challenged with a pathogen (Figure 3).

Related to ethylene signaling, s-adenosylmethionine synthetase (SAM), a member of a gene family involved in ethylene biosynthesis, was down-regulated together with an ethylene response factor (AtERF3) belonging to the B1 subfamily of the ERF/AP2 transcription factor family. The latter functions in adaptation to stress and has been shown to be induced by ethylene, JA and pathogens. AtERF1 is also a positive regulator of ET and JA signaling and a possible integration point in the cross-talk between ET and JA signaling pathways (reviewed in [47,48]).

PGIP expression influences the cell wall by lower XTH activity and increased lignin content

A large number of affected probes matched cell wall-associated genes indicating that cell wall modifications are taking place as an effect of PGIP over-expression. Many differentially expressed probes were linked to lignin and pectin metabolism.

Several tobacco xyloglucan endotransglycosylases (XTHs) representing members of XTH Group I and II were markedly down-regulated. Xyloglucan is the most abundant hemicellulose in dicotyledonous plants and plays a central role in the structure of plant cell walls by cross-linking cellulose microfibrils [49,50] and XTH enzymes are believed to be important for regulation of cell wall strength, extensibility and tissue integrity [51]. XTH Group I and II have *in vitro* been shown to mediate transglucosylation between xyloglucans, in contrast to group III that catalyzes xyloglucan endohydrolysis [52]. It should, however, be noted that the isoforms are grouped according to phylogenetic relationship and that the enzyme activity of all members have not yet been determined.

Because of the important role in cell wall remodeling and marked down-regulation, XTH expression and enzyme activity were investigated further. Indeed, a decrease in XTH expression around the same levels observed in the microarray analysis could be confirmed by RT-qPCR for a tobacco XTH gene belonging to the class I subfamily (Figure 4a). The down-regulation was also confirmed in two additional VvPGIP1 over-expressing lines tested (data not shown). Moreover, a dot-blot enzyme activity assay showed that the general transcriptional down-regulation of XTHs led to a decrease in total XTH activity in leaves of both lines (Figure 4b). Thus, the transcriptional regulation of XTHs had a clear effect on XTH activity, which strengthens the idea that PGIPs have a direct or indirect effect on cell wall modification possibly leading to changes in xyloglucan metabolism.

In tobacco leaves, down-regulation of an *XTH* (*Nt*XET-1) with consequent reduced XTH activity, was previously reported to result in a shift towards xyloglucan with a higher molecular weight, resembling that of older leaves [53]. The authors noted that the cell walls may be strengthened by the reduced turnover and hydrolysis of xyloglucan and it was suggested that the resultant wall strengthening may hold implications for plant-pathogen interactions. XTH activity also increases during fruit ripening and the expression of *V. vinifera* VvXTH1 reaches a maximum at the fully ripe stage when berry softening occurs [54]. Interestingly, *VvXTH1* expression in grape berries is inversely correlated to *VvPGIP1* expression, which instead steadily declines until grape berries reach the fully ripe stage [55]. This inversely correlated expression resembles the correlation between *Vvpgip1* and *XTHs* in the transgenic tobacco. Related to xyloglucan modification, a beta-D-xylosidase with the closest sequence identity to *Arabidopsis* AtXyl4, which is involved in the hydrolysis of the xylan backbone [56], was down-regulated.

Many affected cell wall-associated genes were involved in either lignin or pectin metabolism. The role of lignification in pathogen defense is well documented [57] and cell walls with increased lignin content provide the plant with an effective physical barrier against phytopathogens [58]. Cinnamoyl-CoA reductases (CCR), caffeoyl-CoA O-methyltransferases and beta-glucosidases were differentially expressed and are all involved in monolignol biosynthesis or modification. Monolignol bricks are exported to the cell wall, and then assembled to lignins *in muro* by laccases and peroxidases [59] and changed

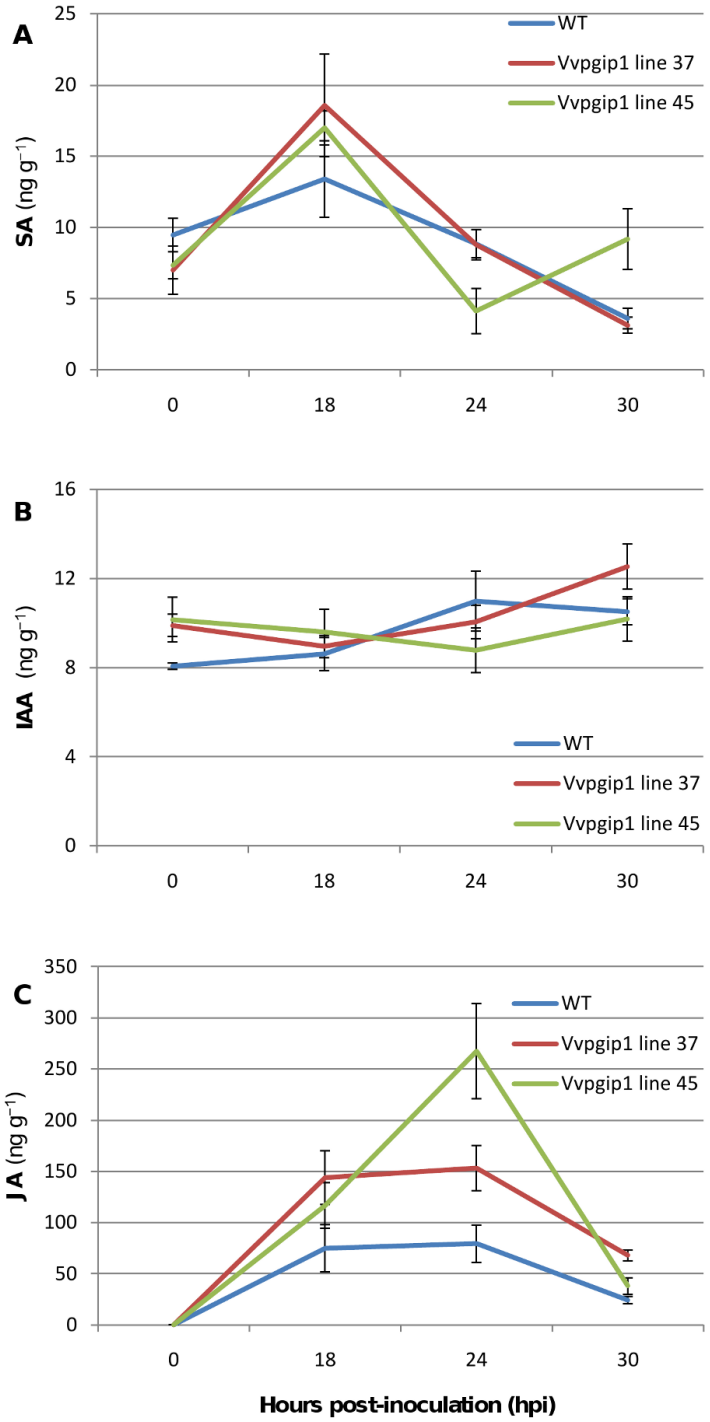


Figure 3 Phytohormone profiling following *Botrytis* infection. The hormones (**a**) salicylic acid (SA) (**b**) indole-acetic acid (IAA) and (**c**) jasmonic acid (JA) were measured as their corresponding methyl esters. Two to three plants per time point were used. *Error bars* are given as two times the standard deviation.

expression was seen for members of the class III peroxidase family. Recently, it was shown by RNAi silencing that several enzymes in monolignol biosynthesis are crucial for defense against powdery mildew penetration in wheat [60]. In tobacco mutants, down-regulation of a CCR has been shown to lead to changes in the lignin profiles and the syringyl-guaiacyl (S/G) ratio, but not necessarily to altered lignin content [61].

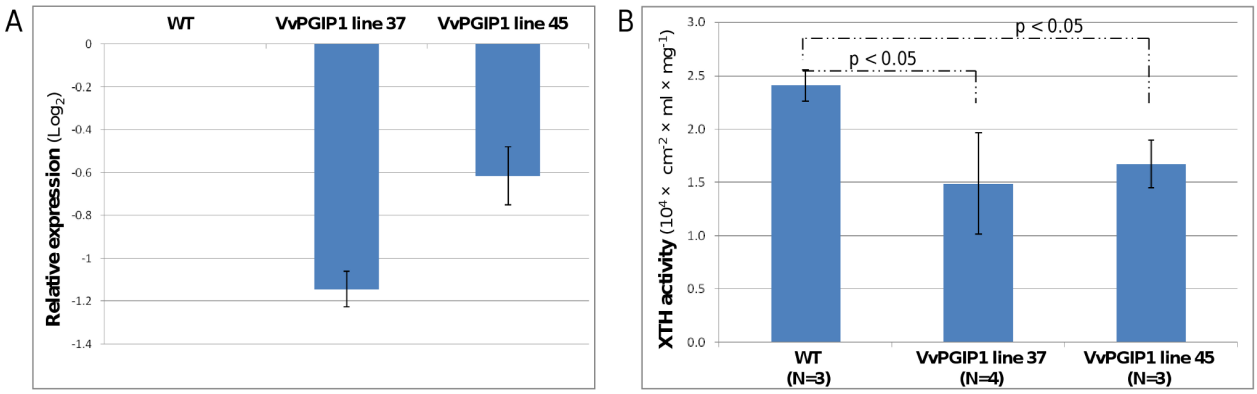


Figure 4 Expression and activity of XTH. **a**. Expression of *Nicotiana tabacum* xyloglucan endotransglycosylase (XTH, Genbank Acc AB017025.1) in *Vvpgip1* lines 37 and 45 relative to the untransformed control (WT). Standard deviation of the mean relative expression of four technical repeats per plant line is shown with *error bars*. **b**. Total XTH activity in tobacco leaves determined by a dot-blot enzyme activity assay. Leaves representing leaf position 3 was used for the assay. A two-tailed Student's *t-test* showed that $p < 0.05$ for both transgenic lines in comparison to WT. N denotes the number of biological repeats.

Because of the changes seen in gene expression and the known importance of lignin in plant defense, the lignin content was determined. An increased deposition of lignin in leaf and stem tissue of the transgenic lines was observed and by absolute quantification an increased lignin content could be confirmed for *Vvpgip1* line 37 and a similar trend was seen in line 45 (Figures 5a, b and 6). The increase of lignin in the cell walls of the PGIP-

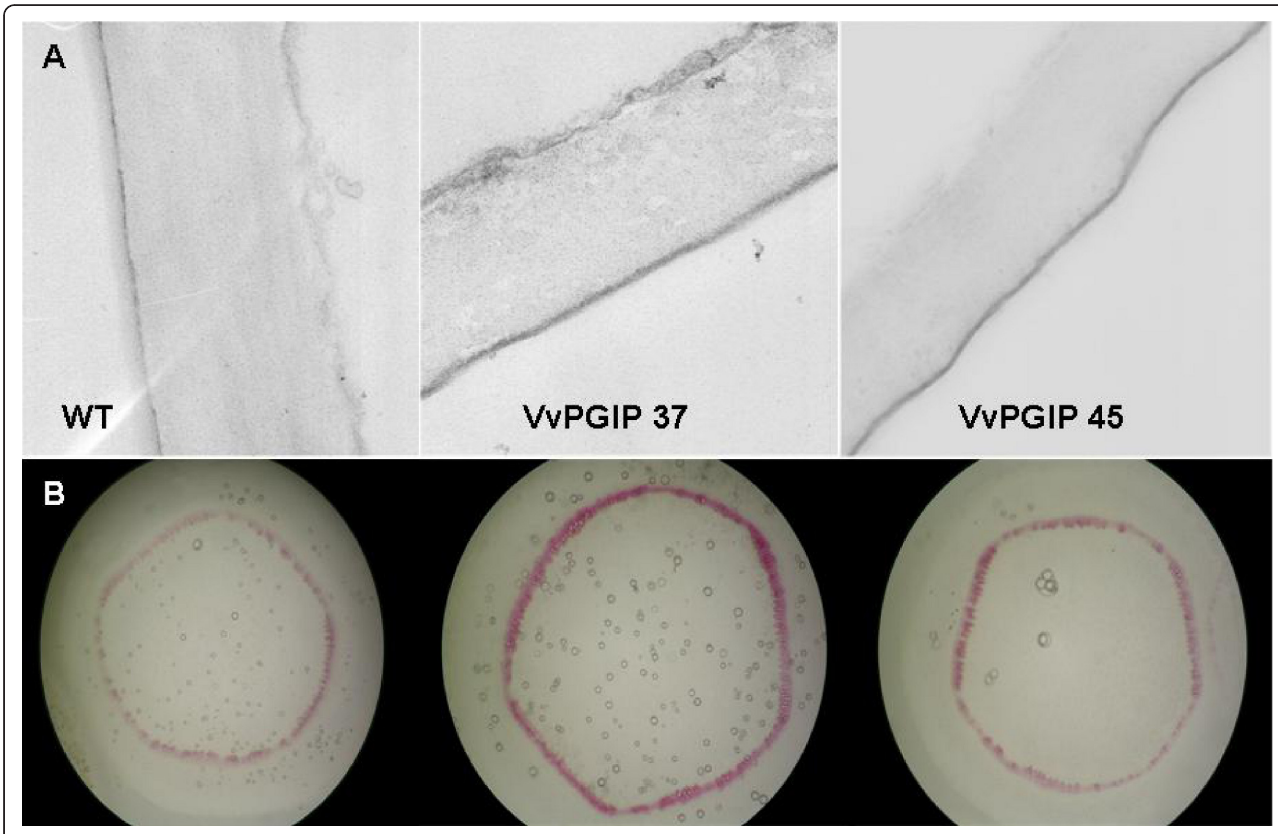


Figure 5 Lignin content determined by staining. Leaf and stem sections of WT and transgenic lines (*Vvpgip1* line 37 and 45) expressing the *Vvpgip1* transgene. **a**. Sections corresponding smaller parts of the leaf were stained with potassium permanganate and studied by transmission electron microscopy according to [27]. Lignin in the cell walls stain darker in the transgenic lines. **b**. Stem sections were hand cut with a razor and stained with a solution of phloroglucinol according to Ruzin [28]. The intensity of the pink-red stain indicates increased lignin deposition in the secondary xylem. Stem sections were taken from the same internode of plants of the same growth stage; leaves for embedding and sectioning were from the same position on plants of the same age (6-8 leaf stage).

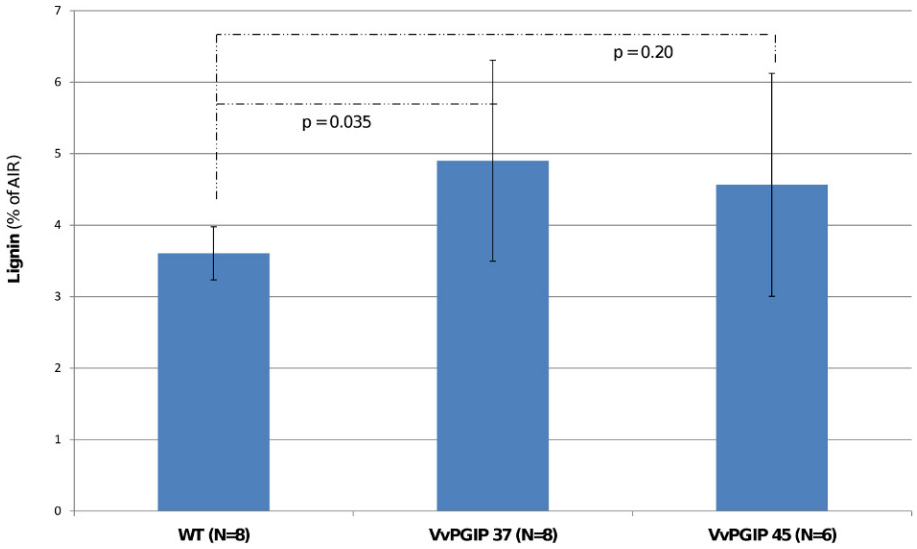


Figure 6 Cell wall lignin content in leaves. The lignin content of untransformed control (WT) and transgenic *Vvpgip1* line 37 and 45 is expressed as the percentage contained in the alcohol-insoluble residue (AIR). Lignin content was determined with the acetyl bromide lignin method, as described in [29]. Statistical analyses (two-tailed Student's *t-test*) could separate the WT from the *Vvpgip1* line 37 (p = 0.035) but not *Vvpgip1* line 45, for which a similar trend was seen (p = 0.20). Standard deviation is given as *error bars*. N represents the number of biological repeats.

specific resistant lines should impact positively on pathogen resistance.

Even if most evidence for the influence of plant hormones on the plant cell wall is indirect, there is a strong belief that phytohormones influence cell wall biosynthesis and remodeling [62]. Auxin, for example, has an effect on cell wall structure and expansion, acting in combination with brassinosteroids, which was the only other hormone-related GO category enriched in the transgenic lines. Previous work indicates that increased IAA content can be coupled to regulation of *XTH* and the increased deposition of lignin, as was observed in the transgenic lines [63-65]. Catalá et al. [65] reported a down-regulation of a tomato *XTH* gene (*LeXET2*) by auxin and transgenic tobacco lines overproducing IAA exhibited increased lignin content and altered lignin composition [63]. Sitbon et al. [64] suggested that the increased lignin deposition may have resulted from increased peroxidase activity, brought about by increased IAA levels and in the transgenic lines, expression levels of several class III peroxidases changed. However, with the exception of a cytosolic ascorbate peroxidase and in line 45 one peroxidase similar to the *Arabidopsis* peroxidase 53 precursor, the peroxidases are all down-regulated. Whether these changes are a direct result of constitutive over-expression of *pgip* influencing IAA, remains to be seen. Alternatively, the changes seen in cell wall remodeling and metabolism reported here, possibly caused by the interaction of PGIPs in the cell wall, leads to increased IAA-levels.

Among the differentially expressed genes related to pectin biosynthesis and composition was a galacturonosyltransferase-like protein (GATL), which is closely related to other galacturonosyltransferase involved in pectin and/or xylan synthesis. Five *Arabidopsis* GATL mutants were recently characterized and shown to have altered pectin and hemicellulose properties [66]. A putative pectinesterase was also down-regulated. These enzymes modify the degree of methylesterification of pectic homogalacturonan affecting cell wall strengthening (reviewed in [67]). PGIP has been shown to directly interact with pectin *in vitro* by binding a negatively charged homogalacturonan motif [10]. Examining *Arabidopsis* PGIP knock-out and over-expressing lines, a recent report further strengthened the link between PGIP and pectin stability by showing that constitutive PGIP over-expression increased the amount of pectin deposited in the seed coat [17].

Several other down-regulated cell wall-associated genes were involved in sugar metabolism. UDP-glucose epimerases (4-UGE) are known to be involved in the channeling of activated galactose to arabinogalactan proteins [68] and to xyloglucan and pectins [26,69]. Other genes related to cell wall biosynthesis and *in muro* remodeling observed to have altered expression patterns were a beta-galactosidase and a cellulose synthase-like gene.

The overall change in the expression of genes related to cell wall composition and structure observed led to the analysis of the leaf cell wall content of eight

monosaccharides in *Vvpgip1* line 37. However, with the exception of rhamnose, which is associated to rhamnogalacturonan I (RG-I) pectin polysaccharides, the cell wall composition was similar to WT (Figure 7). The differential expression of several genes involved in cell wall remodeling and the confirmation that the composition did not change significantly suggest that cell wall organization, rather than composition is affected by the presence of the PGIP. Future studies will focus on the architecture and cross-linking of cell wall components, specifically with regards to pectin content and composition.

PGIP expression induces a shift in primary metabolism

Apart from genes related to the cell wall and hormone biosynthesis and signaling, several genes involved in primary metabolism showed a change in expression (Additional file 1). Among these were enzymes involved in glycolysis, the OPP pathway and the TCA cycle. Several genes involved in nitrogen metabolism and catabolism were also affected. Interestingly, some metabolic genes that could be linked to sink-source tissue regulation were identified, e.g. cell wall invertases and an L-asparaginase 4 precursor [70,71]. Cell wall invertases are believed to be key enzymes in the transaction between sink and source tissues and a down-regulation of activity leading to a decrease in hexose sugar availability is an indication of a transformation from sink to source (reviewed in [71]).

Defense mechanisms are energy intensive and during pathogen attack it is necessary for the plant to regulate metabolic pathways in order to deprive the pathogen of energy resources at the same time as it recruits energy for its own defense response (reviewed in [72]). Our results indicate that the *Vvpgip1* over-expressing lines have adjustments of the primary metabolism even before pathogen attack which might be indicative of a 'primed-like' state. Alternatively, the changes seen in primary metabolism might also be linked to the changes observed in cell wall remodeling and metabolism, which probably affects energy demand and biosynthesis of various building blocks needed. These aspects need further investigation, specifically also in the presence of an infecting pathogen.

Summary and conclusions

Taken together, there is clear evidence that transgenic lines over-expressing *Vvpgip1* have altered cell wall properties compared to their untransformed counterparts, even in the absence of pathogen infection. In addition, increased levels of auxin in uninfected tissue and an amplified JA response following *Botrytis* infection were seen. Interestingly, there were subtle changes in the transcription of a large number of genes representing different distinct functional sets, notably related to cell wall functions, primary metabolism and stress responses. Among the more markedly down-regulated genes were *XTH*s, and we could show that the decrease

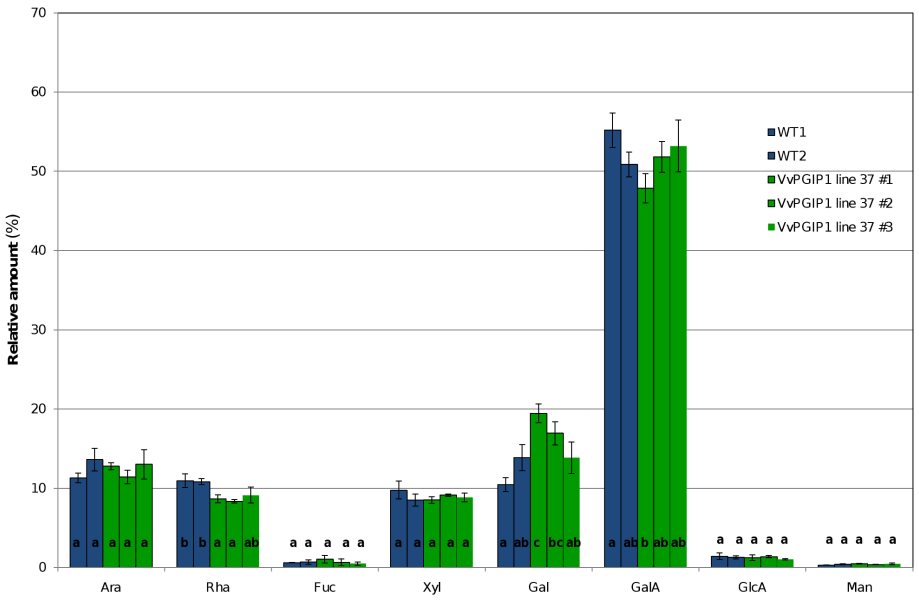


Figure 7 Cell wall component analysis. Total cell wall monosaccharide was determined in tobacco leaves (leaf position 3) for two WT and three *Vvpgip1* line 37 biological repeats. Letters represent statistical groups (within each metabolite) as determined by one-way ANOVA (multi-comparison Bonferroni test, $p < 0.05$).

in *XTH* expression led to a lowered XTH activity. The alteration of XTHs, together with an increased deposition of lignin also observed, are possible contributors to the reduced *Botrytis* susceptibility observed in these PGIP-specific resistant lines.

The current hypothesis is that PGIP's involvement in plant defense is limited to the inhibition of ePGs limiting tissue maceration and necrosis. Also, following inhibition of ePG by PGIP, it is believed that the lifetime of molecules with elicitor activity towards the activation of plant defenses are extended [7]. But, since the conditions primarily studied here were with the pathogen absent, neither the inhibition of ePG nor the extension of the lifetime of oligogalacturonides were involved. PGIP may directly influence defense responses in the plant possibly by strengthening the cell walls; whether by virtue of its structural features, which contains a LRR structure shared with many receptor involved in pathogen recognition or its integration in the cell wall.

However, at this stage we cannot exclude that the effects of the *Vvpgip1* over-expression observed are due to properties specific to VvPGIP1. Putatively, these could originate from more general properties of the protein, e.g., as an effect of its cellular transportation and presence in the apoplast. Still, the cross-species co-expression analysis shows that important genes like members of XTHs, peroxidases, UDP epimerases and JAZs are selectively co-expressed also with PGIPs in *Arabidopsis*, and thus gives some evidence that these groups of genes are specifically affected by the PGIP expression levels.

Lately, other evidence indicating a broader role of PGIPs has been presented. For example, studies have shown that PGIP-encoding gene regulates floral organ number in rice, that the expression of a pea PGIP affects the resistance to nematode invasion without the detection of ePGs in either organism and that PGIP expression influences seed imbibition [16,17,72]. In addition to the inhibition of ePGs and subsequent signaling events, these observations of PGIP's effects on plant development and pathogen resistance together with the work presented in this manuscript, are shedding new light on the *in planta* roles of PGIP.

Additional material

Additional file 1: Table of probes displaying differential expression. Table of all probes displaying differential expression. Differential expression (Log2) between control plants and *Vvpgip1*-transformed lines 37 and 45 (FDR <0.05). Probes are divided into orthologous clusters based on sequence similarity (see Materials and Methods). Functional category, name, an arbitrary set identification number and number of hybridizing transcripts are given for each orthologous cluster. The sequence similarity (%) and amino acid length of similar sequence for the top hit in *Arabidopsis* of each probe sequence is also included. If no match was found in *Arabidopsis* the top hit for rice was instead given. The descriptions presented are derived from the top hits. The level of co-expression of *Arabidopsis* genes associated to two or more of the orthologous clusters based on sequence similarity are given as Pearson coefficients correlations with different stringency (0.05, 0.1 or 0.2). Asterisk denotes that only a trend ($p < 0.15$) was seen in *Vvpgip1* line 45.

Additional file 2: Updated Gene Ontology (GO terms) associated with each probe on the TIGR 10 k microarray. Updated GO terms associated with each probe on the TIGR 10 k microarray. GO terms were updated with information retrieved from nine plant species with the help of a graph structure. The table is compatible for use with GOEast if saved as a tab-delimited file.

Additional file 3: A selection of differentially expressed genes. Probes are divided into orthologous clusters based on sequence similarity and expression ratios are given in Log2-scale. Top hits based on sequence similarity are given for either *Arabidopsis* or rice and includes the corresponding gene description. Related citations are given in the heading literature. Shaded orthologous clusters indicate that associated *Arabidopsis* genes based on sequence identity were co-expressed (Pearson coefficient correlation < 0.2) with one or more additional *Arabidopsis* genes linked to the orthologous clusters. Asterisks denote that only a similar trend for differential expression was observed in *Vvpgip* 1 line 45 (FDR < 0.15). For more details on constructions of orthologous clusters and cross-species co-expression analysis, see Materials and Methods.

Additional file 4: Gene ontology enrichments. GO enrichment of probes showing significant difference in signal intensity in *Vvpgip1* line 37 and 45 in comparison to WT. (A) Biological process (B) Cellular compartment and (C) Molecular function. The gene ontology maps were generated in GOEast [39]. Enriched terms are colored in yellow and the intensity of the color yellow denotes the level of enrichment. Red arrows stand for relationship between two enriched GO terms, black solid arrows stand for relationship between enriched and not enriched terms and black dashed arrows stand for relationship between two not enriched GO terms.

Additional file 5: MapMan overview of metabolic categories comparing *Vvpgip1* line 37 and 45. A MapMan mapping file adopted for the TIGR 10 K potato microarray was used [41]. Expression values were filtered after FDR-adjustment ($p < 0.05$).

Acknowledgements and funding

This work was supported by funding from the National Research Foundation (NRF), the Wine Industry Network of Expertise and Technology (Winetech), the South African Table Grape Industry (SATI) and the South African Technology and Human Resources for Industry Programme (THRIP). Post-doctorial funding for EA was provided by Carl Tryggers Stiftelse för Vetenskaplig Forskning, Sweden, and for ENO by Claude Leon Foundation, South Africa. Andreas G.J. Tredoux, Institute for Wine Biotechnology is acknowledged for technical support in optimizing the hormone profile method.

Author details

[1]Institute for Wine Biotechnology, Department of Viticulture and Oenology, Faculty of AgriSciences, Stellenbosch University, Stellenbosch, South Africa. [2]Department of Molecular and Cell Biology, University of Cape Town, Private Bag, Rondebosch, Cape Town, South Africa. [3]School of Life Sciences and Warwick Systems Biology Centre, University of Warwick, Wellesbourne Campus CV35 9EF, UK. [4]Department of Plant Protection Biology, Swedish Agricultural University, P.O. Box 102, SE-230 53 Alnarp, Sweden. [5]African Centre for Gene Technologies, Experimental Farm, University of Pretoria, Lynnwood Ridge, Pretoria, South Africa.

Authors' contributions

EA did the microarray data and statistical analyses. JVWB performed the microarrays, qPCR and hormone and lignin profiling experiments. DJ constructed the bioinformatical workflow and graph structures. ENO determined the cell wall composition and CS did the XTH enzyme activity

assay. KJD participated in the design of the microarray study. MAV conceived the study.
EA, JVWB, DJ and MAV drafted the manuscript. All authors read and approved of the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 19 July 2011 Accepted: 13 November 2011
Published: 13 November 2011

References

1. De Lorenzo G, D'Ovidio R, Cervone F: **The role of polygalacturonase-inhibiting proteins (PGIPs) in defense against pathogenic fungi.** *Annu Rev Phytopathol* 2001, **39**:313-335.
2. Rodriguez-Palenzuela P, Burr TJ, Collmer A: **Polygalacturonase is a virulence factor in Agrobacterium tumefaciens biovar 3.** *J Bacteriol* 1991, **173(20)**:6547-6552.
3. Shieh MT, Brown RL, Whitehead MP, Cary JW, Cotty PJ, Cleveland TE, Dean RA: **Molecular genetic evidence for the involvement of a specific polygalacturonase, P2c, in the invasion and spread of Aspergillus flavus in cotton bolls.** *Appl Environ Microbiol* 1997, **63(9)**:3548-3552.
4. ten Have A, Mulder W, Visser J, van Kan JA: **The endopolygalacturonase gene Bcpg1 is required for full virulence of Botrytis cinerea.** *Mol Plant Microbe Interact* 1998, **11(10)**:1009-1016.
5. Kars I, Krooshof GH, Wagemakers L, Joosten R, Benen JA, van Kan JA: **Necrotizing activity of five Botrytis cinerea endopolygalacturonases produced in Pichia pastoris.** *Plant J* 2005, **43(2)**:213-225.
6. Joubert DA, Slaughter AR, Kemp G, Becker JV, Krooshof GH, Bergmann C, Benen J, Pretorius IS, Vivier MA: **The grapevine polygalacturonase-inhibiting protein (VvPGIP1) reduces Botrytis cinerea susceptibility in transgenic tobacco and differentially inhibits fungal polygalacturonases.** *Transgenic Res* 2006, **15(6)**:687-702.
7. Cervone F, Hahn MG, De Lorenzo G, Darvill A, Albersheim P: **Host-Pathogen Interactions: XXXIII. A Plant Protein Converts a Fungal Pathogenesis Factor into an Elicitor of Plant Defense Responses.** *Plant Physiol* 1989, **90(2)**:542-548.
8. Federici L, Di Matteo A, Fernandez-Recio J, Tsernoglou D, Cervone F: **Polygalacturonase inhibiting proteins: players in plant innate immunity?** *Trends Plant Sci* 2006, **11(2)**:65-70.
9. Joubert DA, Kars I, Wagemakers L, Bergmann C, Kemp G, Vivier MA, van Kan JA: **A polygalacturonase-inhibiting protein from grapevine reduces the symptoms of the endopolygalacturonase BcPG2 from Botrytis cinerea in Nicotiana benthamiana leaves without any evidence for in vitro interaction.** *Mol Plant Microbe Interact* 2007, **20(4)**:392-402.
10. Spadoni S, Zabotina O, Di Matteo A, Mikkelsen JD, Cervone F, De Lorenzo G, Mattei B, Bellincampi D: **Polygalacturonase-inhibiting protein interacts with pectin through a binding site formed by four clustered residues of arginine and lysine.** *Plant Physiol* 2006, **141(2)**:557-564.
11. Powell AL, van Kan J, ten Have A, Visser J, Greve LC, Bennett AB, Labavitch JM: **Transgenic expression of pear PGIP in tomato limits fungal colonization.** *Mol Plant Microbe Interact* 2000, **13(9)**:942-950.
12. Aguero CB, Uratsu SL, Greve C, Powell AL, Labavitch JM, Meredith CP, Dandekar AM: **Evaluation of tolerance to Pierce's disease and Botrytis in transgenic plants of Vitis vinifera L. expressing the pear PGIP gene.** *Mol Plant Pathol* 2005, **6(1)**:43-51.
13. Ferrari S, Vairo D, Ausubel FM, Cervone F, De Lorenzo G: **Tandemly duplicated Arabidopsis genes that encode polygalacturonase-inhibiting proteins are regulated coordinately by different signal transduction pathways in response to fungal infection.** *Plant Cell* 2003, **15(1)**:93-106.
14. Manfredini C, Sicilia F, Ferrari S, Pontiggia D, Salvi G, Caprari C, Lorito M, De Lorenzo G: **Polygalacturonase-inhibiting protein 2 of Phaseolus vulgaris inhibits BcPG1, a polygalacturonase of Botrytis cinerea important for pathogenicity, and protects transgenic plants from infection.** *Physiol Mol Plant Pathol* 2005, **67**:108-115.
15. Ferrari S, Galletti R, Vairo D, Cervone F, De Lorenzo G: **Antisense expression of the Arabidopsis thaliana AtPGIP1 gene reduces polygalacturonase-inhibiting protein accumulation and enhances susceptibility to Botrytis cinerea.** *Mol Plant Microbe Interact* 2006, **19(8)**:931-936.
16. Veronico P, Melillo MT, Saponaro C, Leonetti P, Picardi E, Jones JT: **A polygalacturonase-inhibiting protein with a role in pea defence against the cyst nematode Heterodera goettingiana.** *Mol Plant Pathol* 2011, **12(3)**:275-287.
17. Kanai M, Nishimura M, Hayashi M: **A peroxisomal ABC transporter promotes seed germination by inducing pectin degradation under the control of ABI5.** *Plant J* 2010, **62(6)**:936-947.
18. Xu ZS, Xiong TF, Ni ZY, Chen XP, Chen M, Li LC, Gao DY, Yu XD, Liu P, Ma YZ: **Isolation and identification of two genes encoding leucine-rich repeat (LRR) proteins differentially responsive to pathogen attack and salt stress in tobacco.** *Plant Sci* 2009, **176**:38-45.
19. Murashige T, Skoog F: **A Revised Medium for Rapid Growth and Bio Assays with Tobacco Tissue Cultures.** *Physiol Plant* 1962, **15(3)**:473-497.
20. Smyth GK: **Limma: linear models for microarray data.** In *In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.).* Edited by: . Springer, New York; 2005:397-420.
21. Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A, Smyth GK: **A comparison of background correction methods for two-colour microarrays.** *Bioinformatics (Oxford, England)* 2007, **23(20)**:2700-2707.
22. Ramakers C, Ruijter JM, Deprez RH, Moorman AF: **Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data.** *Neurosci Lett* 2003, **339(1)**:62-66.
23. Pfaffl MW: **A new mathematical model for relative quantification in real-time RT-PCR.** *Nucleic Acids Res* 2001, **29(9)**:e45.
24. Fry SC: **Novel 'dot-blot' assays for glycosyltransferases and glycosyl hydrolases: optimisation for xyloglucan endotransglycosylase (XET) activity.** *Plant J* 1997, **11**:1141-1150.
25. Bradford MM: **A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding.** *Anal Biochem* 1976, **72**:248-254.
26. Nguema-Ona E, Andeme-Onzighi C, Aboughe-Angone S, Bardor M, Ishii T, Lerouge P, Driouich A: **The reb1-1 mutation of Arabidopsis. Effect on the structure and localization of galactose-containing cell wall polysaccharides.** *Plant Physiol* 2006, **140(4)**:1406-1417.
27. Fromm J, Rockel B, Lautner S, Windeisen E, Wanner G: **Lignin distribution in wood cell walls determined by TEM and backscattered SEM techniques.** *J Struct Biol* 2003, **143(1)**:77-84.
28. Ruzin SE: *Plant Microtechnique and Microscopy* New York: Oxford University Press; 1999.
29. de Ascensao AR, Dubery IA: **Soluble and wall-bound phenolics and phenolic polymers in Musa acuminata roots exposed to elicitors from Fusarium oxysporum f.sp. cubense.** *Phytochemistry* 2003, **63(6)**:679-686.
30. Schmelz EA, Engelberth J, Tumlinson JH, Block A, Alborn HT: **The use of vapor phase extraction in metabolic profiling of phytohormones and other metabolites.** *Plant J* 2004, **39(5)**:790-808.
31. Duvick J, Fu A, Muppirala U, Sabharwal M, Wilkerson MD, Lawrence CJ, Lushbough C, Brendel V: **PlantGDB: a resource for comparative plant genomics.** *Nucleic Acid Res* 2008, **36(Database issue)**:D959-965.
32. Wasmuth JD, Blaxter ML: **prot4EST: translating expressed sequence tags from neglected genomes.** *BMC Bioinformatics* 2004, **5**:187.
33. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acid Res* 1997, **25(17)**:3389-3402.
34. Lottaz C, Iseli C, Jongeneel CV, Bucher P: **Modeling sequencing errors by combining Hidden Markov models.** *Bioinformatics (Oxford, England)* 2003, **19(Suppl 2)**:ii103-112.
35. Fukunishi Y, Hayashizaki Y: **Amino acid translation program for full-length cDNA sequences with frameshift errors.** *Physiol Genomics* 2001, **5(2)**:81-87.
36. Proost S, Van Bel M, Sterck L, Billiau K, Van Parys T, Van de Peer Y, Vandepoele K: **PLAZA: a comparative genomics resource to study gene and genome evolution in plants.** *Plant Cell* 2009, **21(12)**:3718-3731.
37. Li L, Stoeckert CJ Jr, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13(9)**:2178-2189.
38. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, *et al*: **Integration of biological networks and gene expression data using Cytoscape.** *Nat Protoc* 2007, **2(10)**:2366-2382.
39. Zheng Q, Wang XJ: **GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis.** *Nucleic Acid Res* 2008, **36(Web Server issue)**:W358-363.

40. Srinivasasainagendra V, Page GP, Mehta T, Coulibaly I, Loraine AE: **CressExpress: a tool for large-scale mining of expression data from Arabidopsis.** *Plant Physiol* 2008, **147(3)**:1004-1016.
41. Rotter A, Usadel B, Baebler S, Stitt M, Gruden K: **Adaptation of the MapMan ontology to biotic stress responses: application in solanaceous species.** *Plant Methods* 2007, **3**:10.
42. Rensink WA, Lee Y, Liu J, Iobst S, Ouyang S, Buell CR: **Comparative analyses of six solanaceous transcriptomes reveal a high degree of sequence conservation and species-specific transcripts.** *BMC Genomics* 2005, **6**:124.
43. Renn SC, Aubin-Horth N, Hofmann HA: **Biologically meaningful expression profiling across species using heterologous hybridization to a cDNA microarray.** *BMC Genomics* 2004, **5(1)**:42.
44. Rise ML, von Schalburg KR, Brown GD, Mawer MA, Devlin RH, Kuipers N, Busby M, Beetz-Sargent M, Alberto R, Gibbs AR, *et al*: **Development and application of a salmonid EST database and cDNA microarray: data mining and interspecific hybridization characteristics.** *Genome Res* 2004, **14(3)**:478-490.
45. Schoof H, Ernst R, Nazarov V, Pfeifer L, Mewes HW, Mayer KF: **MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource for plant genomics.** *Nucleic Acid Res* 2004, **32(Database issue)**:D373-376.
46. Stintzi A, Weber H, Reymond P, Browse J, Farmer EE: **Plant defense in the absence of jasmonic acid: the role of cyclopentenones.** *Proc Natl Acad Sci USA* 2001, **98(22)**:12837-12842.
47. Gutterson N, Reuber TL: **Regulation of disease resistance pathways by AP2/ERF transcription factors.** *Curr Opin Plant Biol* 2004, **7(4)**:465-471.
48. Bari R, Jones JD: **Role of plant hormones in plant defence responses.** *Plant Mol Biol* 2009, **69(4)**:473-488.
49. McNeil M, Darvill AG, Fry SC, Albersheim P: **Structure and function of the primary cell walls of plants.** *Ann Rev Biochem* 1984, **53**:625-663.
50. Carpita NC, Gibeaut DM: **Structural models of primary cell walls in flowering plants: consistency of molecular structure with the physical properties of the walls during growth.** *Plant J* 1993, **3(1)**:1-30.
51. Saladie M, Rose JK, Cosgrove DJ, Catala C: **Characterization of a new xyloglucan endotransglucosylase/hydrolase (XTH) from ripening tomato fruit and implications for the diverse modes of enzymic action.** *Plant J* 2006, **47(2)**:282-295.
52. Rose JK, Braam J, Fry SC, Nishitani K: **The XTH family of enzymes involved in xyloglucan endotransglucosylation and endohydrolysis: current perspectives and a new unifying nomenclature.** *Plant Cell Physiol* 2002, **43(12)**:1421-1435.
53. Herbers K, Lorences EP, Barrachina C, Sonnewald U: **Functional characterisation of Nicotiana tabacum xyloglucan endotransglycosylase (NtXET-1): generation of transgenic tobacco plants and changes in cell wall xyloglucan.** *Planta* 2001, **212(2)**:279-287.
54. Nunan KJ, Davies C, Robinson SP, Fincher GB: **Expression patterns of cell wall-modifying enzymes during grape berry development.** *Planta* 2001, **214(2)**:257-264.
55. De Ascensao ARFDC: **Isolation and characterization of a polygalacturonase-inhibiting protein (PGIP) and its encoding gene from Vitis vinifera L .** *PhD thesis, Stellenbosch University.* 2001.
56. Minic Z, Rihouey C, Do CT, Lerouge P, Jouanin L: **Purification and characterization of enzymes exhibiting beta-D-xylosidase activities in stem tissues of Arabidopsis.** *Plant Physiol* 2004, **135(2)**:867-878.
57. Nicholson RL, Hammerschmidt R: **Phenolic compounds and their role in disease resistance.** *Annu Rev Phytopathol* 1992, **30**:369-389.
58. Ride JP: **Cell walls and other structural barriers in defense.** In *In: Biochemical Plant Pathology, J.A. Callow (Ed).* Edited by: . John Wiley and Sons, New York; 1983:215-236.
59. Kaneda M, Rensing KH, Wong JC, Banno B, Mansfield SD, Samuels AL: **Tracking monolignols during wood development in lodgepole pine.** *Plant Physiol* 2008, **147(4)**:1750-1760.
60. Bhuiyan NH, Selvaraj G, Wei Y, King J: **Gene expression profiling and silencing reveal that monolignol biosynthesis plays a critical role in penetration defence in wheat against powdery mildew invasion.** *J Exp Bot* 2009, **60(2)**:509-521.
61. Piquemal J, Lapierre C, Myton K, O'Connell A, Schuch W, Grima-Pettenati J, Boudet AM: **Down-regulation of Cinnamoyl-CoA Reductase induces significant changes of lignin profiles in transgenic tobacco plants.** *Plant J* 1998, **13**:71-83.
62. Sanchez-Rodriguez C, Rubio-Somoza I, Sibout R, Persson S: **Phytohormones and the cell wall in Arabidopsis during seedling growth.** *Trends in Plant Science* 2010, **15(5)**:291-301.
63. Sitbon F, Hennion S, Sundberg B, Little CH, Olsson O, Sandberg G: **Transgenic Tobacco Plants Coexpressing the Agrobacterium tumefaciens iaaM and iaaH Genes Display Altered Growth and Indoleacetic Acid Metabolism.** *Plant Physiol* 1992, **99(3)**:1062-1069.
64. Sitbon F, Hennion S, Anthony Little CH, Sundberg B: **Enhanced ethylene production and peroxidase activity in IAA-overproducing tobacco plants with increased lignin content and altered lignin deposition.** *Plant Sci* 1999, **141**:165-173.
65. Catala C, Rose JK, York WS, Albersheim P, Darvill AG, Bennett AB: **Characterization of a tomato xyloglucan endotransglycosylase gene that is down-regulated by auxin in etiolated hypocotyls.** *Plant Physiol* 2001, **127(3)**:1180-1192.
66. Kong Y, Zhou G, Yin Y, Xu Y, Pattathil S, Hahn MG: **Molecular analysis of a family of Arabidopsis genes related to galacturonosyltransferases.** *Plant Physiol* 2011, **155(4)**:1791-1805.
67. Pelloux J, Rusterucci C, Mellerowicz EJ: **New insights into pectin methylesterase structure and function.** *Trends Plant Sci* 2007, **12(6)**:267-277.
68. Andeme-Onzighi C, Sivaguru M, Judy-March J, Baskin TI, Driouich A: **The reb1-1 mutation of Arabidopsis alters the morphology of trichoblasts, the expression of arabinogalactan-proteins and the organization of cortical microtubules.** *Planta* 2002, **215(6)**:949-958.
69. Rosti J, Barton CJ, Albrecht S, Dupree P, Pauly M, Findlay K, Roberts K, Seifert GJ: **UDP-glucose 4-epimerase isoforms UGE2 and UGE4 cooperate in providing UDP-galactose for cell wall biosynthesis and growth of Arabidopsis thaliana.** *Plant Cell* 2007, **19(5)**:1565-1579.
70. Lea PJ, Sodek L, Parry MAJ, Shewry PR, Halford NG: **Asparagine in plants.** *Ann Appl Biol* 2007, **150**:1-26.
71. Wingler A, Roitsch T: **Metabolic regulation of leaf senescence: interactions of sugar signalling with biotic and abiotic stress responses.** *Plant Biol (Stuttgart, Germany)* 2008, **10(Suppl 1)**:50-62.
72. Jang S, Lee B, Kim C, Kim SJ, Yim J, Han JJ, Lee S, Kim SR, An G: **The OsFOR1 gene encodes a polygalacturonase-inhibiting protein (PGIP) that regulates floral organ number in rice.** *Plant Mol Biol* 2003, **53(3)**:357-369.

doi:10.1186/1756-0500-4-493
Cite this article as: Alexandersson *et al.*: Constitutive expression of a grapevine polygalacturonase-inhibiting protein affects gene expression and cell wall properties in uninfected tobacco. *BMC Research Notes* 2011 **4**:493.

Chapter 6

The Vineyard Yeast Microbiome, a Mixed Model Microbial Map

Published as:

Setati ME, Jacobson D, Andong U-C, Bauer F (2012) The Vineyard Yeast Microbiome, a Mixed Model Microbial Map. PLoS ONE 7(12): e52609. doi:10.1371/journal.pone.0052609

6.1 Abstract

Vineyards harbour a wide variety of microorganisms that play a pivotal role in pre- and post-harvest grape quality and will contribute significantly to the final aromatic properties of wine. The aim of the current study was to investigate the spatial distribution of microbial communities within and between individual vineyard management units. For the first time in such a study, we applied the Theory of Sampling (TOS) to sample grapes from adjacent and well established commercial vineyards within the same terroir unit and from several sampling points within each individual vineyard. Cultivation-based and molecular data sets were generated to capture the spatial heterogeneity in microbial populations within and between vineyards and analysed with novel mixed-model networks, which combine sample correlations and microbial community distribution probabilities. The data demonstrate that farming systems have a significant impact on fungal diversity but more importantly that there is significant species heterogeneity between samples in the same vineyard. Cultivation-based methods confirmed that while the same oxidative yeast species dominated in all vineyards, the least treated vineyard displayed significantly higher species richness, including many yeasts with biocontrol potential. The cultivatable yeast population was not fully representative of the more complex populations seen with molecular methods, and only the molecular data allowed discrimination amongst farming practices with multivariate and network analysis methods. Importantly, yeast species distribution

is subject to significant intra-vineyard spatial fluctuations and the frequently reported heterogeneity of tank samples of grapes harvested from single vineyards at the same stage of ripeness might therefore, at least in part, be due to the differing microbiota in different sections of the vineyard.

6.2 Introduction

Vineyards and grape berry surfaces provide a physical environment on which complex microbial communities comprising yeasts, bacteria and filamentous fungi establish themselves. In the wine industry, the species composition of these communities is of significant importance since the microbial species that are present on the berry may contribute to the fermentative process and therefore the aromatic properties of the resulting wine. This is of particular relevance in cases where the oenological practice includes spontaneous fermentations, as is the case in many wineries.

Data indicate that yeast populations on wine grapes increase from 102-103 cfu/g on immature berries to 103-106 cfu/g on mature berries. Yeast are spatially distributed over the grape berries and grape bunches, and also display temporal fluctuations in diversity over the course of grape berry development [6]-[26]. Species present on intact undamaged berries after veraison and until full ripeness have been reported to mainly belong to the group of oxidative basidiomycetous yeasts such as *Cryptococcus spp.*, *Rhodotorula spp.*, *Sporobolomyces spp.*, and *Filobasidium spp.*, as well as to the dimorphic ascomycetous black yeast, *Aureobasidium pullulans* [6], [26], [4]. In the vineyard environment, these yeasts are typically associated with the phyllosphere, grapes and soil [6]. The oxidative ascomycetous yeasts (e.g. *Candida spp.*, *Pichia spp.*, and *Metschnikowia spp.*), and the fermentative ascomycetous yeasts (e.g. *Hanseniaspora/Kloeckera spp.*) have been found to be present at low concentrations on undamaged berries and appear often localized in those areas of the grape surface where some juice might escape [9], [22]. The incidence of these yeasts on damaged grapes increases rapidly and 10 fold increases have been reported [4], [22]. In contrast, the most relevant fermentative wine yeast, *Saccharomyces cerevisiae* only occurs at concentrations of less than 10-100 cfu/g berry [17].

The density and diversity of the grape microbiota may be influenced by many factors including climatic conditions, diseases, insect pests and viticultural practices [7]-[8]. Recently, differences in yeast populations associated with grapes obtained from organic and conventional farms have been reported [12]-[38], thus alluding to the possible impact of farming methods on grape microbiota. However, in these studies microbial diversity was only analysed after grapes were crushed and blended, thus using the juice as auto-enrichment, and either after 70 g/L of sugar was consumed or in the middle and end of alcoholic fermentation, when many species have been eliminated due to the

high alcohol content. Such a strategy will have led to a significant enrichment of some species, and the elimination of many other species that were initially present on the grape. Furthermore, such an approach precludes a statistical validation of inter- and intra-vineyard variability.

In South Africa, wine grapes are produced using a range of farming methods from conventional to biodynamic farming. The majority of grapes are produced through what can be described as an intermediate scheme, the Integrated Production of Wine (IPW), which was established by the South African wine industry in 1998 [2]. This scheme embraces a more environmentally friendly farming system, including careful monitoring and understanding of diseases resulting in reduced input of biocides in the vineyard when compared with conventional farming [20]. The system also promotes the use of hay mulches and oats cover crops to improve soil moisture and fertility, as well as bait, ducks and other biocontrol strategies for pest control. However, integrated farming systems are not fully codified into rules, and do not have a regulated certification system [20]. In contrast, biodynamic farming is a specialised type of organic farming which prohibits any use of chemical fertilizers and pesticides as stipulated under the Demeter regulations [1]. In addition, biodynamic farming includes the use of specific fermented herbal and mineral preparations as compost additives and field sprays which are applied into the soil in animal organs e.g. bladder and cow horn [28].

Organic and biodynamic farming systems have been shown to enhance soil fertility and increase biodiversity [21]-[29]. In wheat plantations, microbial diversity has been found to be highest in biodynamic areas, followed by organically farmed and finally conventional plantations [21]. Although organic and biodynamic systems are globally becoming of increasing economic interest to wine producers, their impact on general vineyard health and wine quality has been the subject of relatively few studies. In particular, the impact of these practices on the vineyard ecosystem (including microbial diversity) is poorly understood.

The current study was aimed at evaluating microbial diversity associated with grapes obtained from conventional, biodynamic and integrated pest management vineyards, with a focus on epiphytic yeasts. The study also appears to be the first to assess intra-vineyard variability of microbial diversity. The data confirm previous results (on other crops) that biodynamic farming leads to a higher microbial diversity. It also shows that this diversity is unevenly distributed within individual vineyards, thus highlighting the importance of sampling multiple locations in the vineyard to assess the biodiversity of the ecosystem. From a wine making perspective, the data suggest that spatial fluctuations in microbial diversity might have a significant impact on downstream processes and analyses.

6.3 Materials and Methods

6.3.1 Vineyard Locations and Treatments

Cabernet sauvignon grape samples were collected from three directly adjacent vineyards. The vineyards, located in the Polkadraai region in Stellenbosch, South Africa (Figure S1), were carefully selected to allow conclusive assessment of the impact of farming practices on both intra- and inter-vineyard microbial biodiversity. In particular, the vineyards are positioned on the same slope and aspect, and were all established in the same period (1994 and 1995). All vineyards also use the same trellising system (Perold 4 wire), row width (2.5 m) and vine interspacing width (1.4 m). However, each vineyard has been managed consistently and over a long period through strongly divergent farming methods, referred to as "conventional" (33^o 57'41.50" S, 18^o 45'11.87" E elev 179 m), "Integrated production" (33^o57'40.65" S 18^o 45'08.23" E elev 184 m) and "biodynamic" (33^o57'39.33" S 18^o 45'13.46" E elev 183 m). The conventional and biodynamic vineyard had the same cabernet sauvignon rootstock (R101-14) while the integrated vineyard has rootstock R110-CS23A. Management practices were as follows (see Table S1) for details): The "biodynamic" vineyard, was converted to "biodynamic" farming principles in 2000, and certified by Demeter International in 2006. The vineyard was treated regularly with Kumulus (sulphur), nordox (copper oxide), striker (organic fungicide with chitosan) and lime for the protection of powdery mildew and downy mildew, from leaf-fall until full bloom. The "integrated production" vineyard has been managed through the integrated pest and vineyard management system since its inception, which includes the use of chicken manure, inoculation of mycorrhizae and Trichoderma spp. into the soil, as well as the use of oats as cover crops. Pest management consisted of a combination of fungicides including hyperphos (mono- and dipotassium hydrogen phosphate), dithane (ethylene bisdithiocarbamate), Kumulus (80% sulphur), acrobat MZ (dimethomorph/mancozeb), talendo (proquinazid), curzate (cymoxanil/mancozeb) and stroby (kresoximethyl); and insecticides such as vantex (pyrethroid) and delmathrin, based on recommendations from an annual evaluation of the vineyard as per IPW guidelines. In contrast, the vines in the conventional vineyard were treated with chemical fertilizers applied when necessary and the vines were consistently treated with a combination of fungicides including folpan (N-(trichloromethyl)thio) phthalimide, rootex (phosphorous acid), cumulus, dithane, acrobat, talendo, cungfu (copper hydroxide) and topaz (mono- and di-potassium salts of phosphorous acid), and different stages from leaf-fall to full bloom (Table S1). Sprays 1, 2 and 3 were applied with designer, a non-ionic sticker to improve the spread, coverage and retention of the fungicides and insecticides. No specific permits were required for the described field studies as they do not form part of protected land or conservation areas, and have not been reported to contain any endangered species. The three vineyards are

privately owned commercial entities consequently, permission to use them as a study site and to sample the grapes was granted independently by each of the owners.

6.3.2 Sampling Design

According to the Theory of Sampling (TOS) [16]-[15], the most efficient manner to sample a two-dimensional lot is to linearise it (aka to 'unfold' or to 'vectorize' it), into an elongated one-dimensional lot from which to extract increments at equidistant intervals [14], [18]. This approach is optimal with respect to capturing and characterizing the heterogeneity present within the lot, offering a way to derive a minimum number of increments needed (Q) if based on variographic analysis [36]; alternatively the number Q may reflect local logistical and/or economic constraints. From a sampling design perspective a vineyard block can be likened to a two-dimensional lot, where rows are easily unfolded into continuous series, in which panels (each containing 6 vines) make up a 'group'. At each vine location, the increments were defined to equal bunches. In the present study, one increment (bunch) was collected from each group, with groups regularly spaced throughout the unfolded linear lot. Thus in the conventional vineyard six rows (no.s 9, 11, 13, 15, 17 and 19) were sampled, where bunches were collected between panel 3, 7 and 11. In the biodynamic vineyard seven rows (no.s 1, 4, 7, 10, 13, 16, 19) were sampled while in the integrated vineyard only three rows were targeted (no.s 115, 117 and 119); here the bunches were collected from panels 1, 3, 5, 7, 9 and 11 respectively (Figure S1). Grape bunches were placed in sterile bags and transported to the laboratory and processed within 1 hour after harvest.

6.3.3 Pseudoreplicaton Test

In order to test for pseudoreplication effects the following approach was implemented in Perl. A Cartesian coordinate system was created for each of the three vineyards utilizing the fact that the row width is 2.5 meters and the panel width is 9 meters. Given this, each sampling point can be described as a two point vector and the distance between two sampling points can be calculated as follows:

$$Distance = \sqrt{dx^2 + dy^2} \tag{6.3.1}$$

Where dx is the difference between the x coordinates and dy is the difference between the y coordinates. For each possible pair of sampling points within each vineyard the Pearson correlation of species detected via ARISA analysis was plotted against the distance between the sampling points and the R^2 value calculated for each plot.

6.3.4 Yeast Enumeration and Isolation

Thirty undamaged berries were collected from each bunch of grapes by using scissors cleaned with 70% ethanol and placed in 250 ml sterile pre-weighed Erlenmeyer flasks. The berries were then washed with 50 ml of saline solution comprising 0.9% w/v NaCl and 0.2% (v/v) Tween 80 to release the microorganisms [30]. This step was carried out at 30^{o}C for 3 h with agitation on an Innova 5000 Gyrotory tier shaker (New Brunswick Scientific, Edison, New Jersey, USA) at 170 rpm. The washing solution was placed in 50 ml centrifuge tubes, followed by a centrifugation step at 5630 $\times$ g for 10 min. The pellet was re-suspended in 10 ml fresh solution and used for yeast enumeration and community profiling using automated ribosomal intergenic spacer analysis (ARISA). For yeast isolation and enumeration, decimal dilutions (10^{-1} to 10^{-3}) were prepared from the wash solutions, and 100 ul samples of each dilution were spread-plated in duplicate on Wallerstein nutrient agar (Sigma-Aldrich) supplemented with 34 mg/L chloramphenicol (Sigma-Aldrich) and 150 mg/L biphenyl (Riedel-deHaën, Seelze, Germany) to inhibit bacterial and mould growth, respectively. The plates were incubated at 30^{o}C and examined daily for growth until the colonies were easily distinguishable. Where possible, 4-6 representatives of each colony-morphology were isolated from plates with $\leq$ 250 colonies and purified through two rounds of streak plating onto fresh agar plates. In addition, unique but infrequent colonies that were observed on plates with $>$ 250 colonies were also isolated. The isolates were maintained in 20% (v/v) glycerol at -80^{o}C.

6.3.5 DNA Extraction and ARISA Fingerprinting

The yeast communities associated with grapes were analyzed using PCR and ARISA. The remaining wash solutions were centrifuged at 5630 $\times$ g for 10 min to collect microbial biomass. The pellet was re-suspended in lysis buffer and DNA was extracted as previously described by Hoffman [35]. The ITS1-5.8S-ITS2 rRNA region was amplified with the FAM labelled ITS1 primer (5'-TCCGTAGGTGAACCTGCGG-3') and ITS4 (5'-TCCTCCGCTTATTGATATGC-3'), using the Phire®Plant Direct PCR kit (FINNZYMES OY, Espoo, Finland) under the following conditions: an initial denaturation of 6 min at 98^{o}C, followed by 40 cycles of 98^{o}C for 20 s, 54^{o}C for 30 s, 72^{o}C 1 min, and a final extension of 10 min at 72^{o}C. The ARISA-PCR fragments were separated by capillary electrophoresis on an ABI3010xl Genetic Analyzer (Applied Biosystems, CA, USA) to obtain electropherograms of the different fragment lengths and fluorescent intensities. A ROX1.1 size standard was used [36]. The ARISA data was analysed using Genemapper 4.1 software (Applied Biosystems). A threshold of 50 fluorescent units was used to exclude background fluorescence. The software converted the fluorescence data into electropherograms, where the peaks represent fragments of different sizes, and the peak areas represent

the relative proportion of these fragments. The number of peaks in each electropherogram was interpreted as the OTU richness in the community. The fragment lengths and fluorescence for each sample were aligned using an Excel Macro. Only fragment sizes larger than 0.5% of the total fluorescence and between 300 and 1000 bp in length were considered for analysis. A bin size of 3 bp for fragments below 700 bp and 5 bp for fragments above 700 bp was employed to minimize the inaccuracies in the ARISA profiles [34]. All elution points in the electropherograms that did not contain a peak in at least one sample were removed with the use of a custom built Perl program. This process resulted in a matrix in which each row represented a sample and each column represented an OTU (species). Principal component analysis (PCA) of the ARISA profile matrix was performed in STATISTICA software Version 10 [13].

6.3.6 Vineyard Sampling Point Networks

An all-against-all comparison was done calculating the Pearson correlation between each and every sample vector in the ARISA matrix. As such, one is able to determine the correlation in population structure within and across vineyards. The relationships between samples were represented as a mathematical graph in order to form a correlation network with the nodes representing sampling point locations in each vineyard and the edges weighted with the Pearson correlations between the sampling point vectors. In order to select the highest correlations between sampling points a maximum spanning tree was created by transforming the edge weights into inverse correlations (by taking the difference between the number 1 and the absolute correlation values) and the subsequent use of a minimum spanning tree (mst) algorithm [3] on this inverse correlation network. A minimum spanning tree represents the shortest possible path through a graph and, as such, selects for the smallest inverse correlation (i.e highest correlation) pairs between all nodes in the network. The nodes were annotated with colours based on the vineyards the samples were taken from and the edge widths scaled with respect to the level of the original correlation values between the samples. The resulting network was visualized in Cytoscape [3].

6.3.7 OTU Probability Networks

A probability matrix was created by dividing each element of a sample vector by the sum of all of the elements in the vector. Thus each resulting element represented the probability of that sample containing that particular OTU. A probability network was then created by creating edges between each sample and the OTUs for which there was a probability value > 0 which were then used as edge weights. The sample nodes were annotated with colours based on the vineyards the samples were taken from. An edge-based spring embedded layout

algorithm was applied to the resulting network which was then visualized in Cytoscape [3].

6.3.8 Mixed-model Networks: Combining Correlation and Probability Networks

In order to represent the most probable microbial community structure of each sampling point a mixed-model network was developed as follows. The edge weights of the Vineyard Sampling point maximum spanning tree network described above were multiplied by 10 and the resulting re-weighted network Unioned with the probability network described above.

In order to select the highest probability edges between sampling points and OTUs a maximum spanning tree was created by transforming the edge weights into inverse values (by calculating the absolute difference between the number 1 and the edge weights) and the subsequent use of a minimum spanning tree (mst) algorithm [25] on this inverse edge-weighted network. After the mst algorithm was applied the original weights from both the correlation and probability networks were used as edge weights of the surviving edges. Edge thicknesses were then scaled with respect to the edge weights and sampling point node sizes were scaled with regard to degree (i.e. the number of edges incident to a node). OTU node sizes were scaled with regard to the probability of them occurring in the sample that they shared an edge with. The resulting network was then visualized in Cytoscape [3].

6.3.9 Molecular Yeast Identification

Selected colonies were picked from the plate by using a sterile inoculating loop and DNA was extracted from the colonies using the protocol for rapid isolation of yeast DNA [35]. The isolates were then identified by amplifying the ITS1-5.8S-ITS2 rRNA region using the ITS1 and ITS4. PCR amplifications were carried out in a final volume of 25 ul containing 0.25 uM of each primer, 1x PCR reaction buffer, 1 mM MgCl2, 200 mM dNTPs, 1U Takara Ex Taq™DNA polymerase (TaKaRa Bio Inc., Olsu, Shiga, Japan), 100 ng of DNA and sterilized de-ionized H_2O. The PCR reaction was carried out using the following conditions: initial denaturation at 94^oC for 2 min; 35 cycles of denaturing at 94^oC for 30 s; annealing at 54^oC for 45 s; an extension at 72^oC for 1 min; and a final extension step of 10 min at 72^oC. The PCR products were analysed by agarose gel electrophoresis; purified using the Zymoclean™Gel DNA recovery kit (Zymo Research Corporation, Irvine, CA, USA), following the manufacture's instruction, and then sequenced. The sequences obtained were assembled using BioEdit [19], and compared with sequences available in GenBank database available at the National Centre for Biotechnology Information (NCBI) http://www.ncbi.nlm.nih.gov/genbank/inde?x.html using the basic local alignment search tool (BLAST) algorithm [10]. Sequences which

displayed 98-99% identity to previously published species available at NCBI were binned into the same species. Sequences obtained in this study were deposited in NCBI GenBank database under accession numbers: JQ993367-JQ993394.

6.3.10 Statistical Analysis

The relative abundance of species was calculated as a proportion of a particular species in the samples based on colony counts and frequency of isolation. Species richness was assessed using the Menhinick's index while species evenness was assessed using the Pielou index [27]. The Shannon diversity index was used to assess the level of diversity in the three vineyards [12].

6.4 Results

6.4.1 Quantitative Analysis of Grape-associated Yeast Communities

The impact of farming systems on yeast population density was evaluated by culture-dependent methods following a 3 h rinsing of sound grape berries obtained from the conventional, biodynamic and integrated vineyards. The total yeast populations were higher in the biodynamic and conventional vineyard than in the integrated vineyard (Figure S2). The total yeast population ranged from $4 - 8 \times 10^4$ CFU/g on all vineyards, and the enumeration of cultivable population revealed no significant differences between the farming systems (P = 0,225).

6.4.2 Inter- and Intra-vineyard Variability of the Total Fungal Community

ARISA analysis was used to unravel fungal community structures associated with healthy/sound grapes in conventional, biodynamic and integrated pest management farming systems. Similar electropherograms were obtained from all the samples. Bands between 500 and 600 bp were dominant in all the vineyards, however, differences in fungal community structures were evident in the three vineyards. Principal components analysis (PCA) was performed on the ARISA profiles in order to evaluate inter-vineyard variation. Each vineyard could be differentiated on the basis of the ARISA fingerprints (Figure 6.1). The biodynamic and integrated vineyard could be separated on the first axis, with the integrated vineyard samples mainly clustered on the right hand side of the first factorial plane while the biodynamic vineyard samples clustered on the left hand side. In addition, the biodynamic and conventional vineyard could be further separated on the second axis which explained 31.2% of the

total variance. The conventional vineyard samples mainly clustered in the top plane while the biodynamic vineyard samples were located in the lower factorial plane (Figure 6.1). Community networks derived from the same ARISA data showed higher correlation between the biodynamic farming system and the integrated pest management system (Figure S3). The community network of the three farming systems comprised highly connected OTUs revealing significant overlap between the three systems (Figure 6.2), but also showing that there are several OTUs which are unique to specific farming systems. Once the link between microbial diversity and farming practices was established, we further explored intra-vineyard variability by evaluating the probability of certain OTUs being present in specific locations in the vineyard. The probabilistic species distribution patterns revealed interesting ecological patterns and for the first time confirmed intra-vineyard variability. For instance, in the integrated vineyard, row 117:panel 1 displayed a higher level of diversity, while row 115:panel 3 and row 117:panel 3, displayed the lowest diversity. Row 117:panel 7 comprised a unique fungal community which seemed more similar to the communities present in the conventional vineyard (Figure 6.3). In contrast, in the biodynamic and conventional vineyard, the level of diversity within the rows and panels were similar, however, the OTUs represented at each site differed, such that the likelihood of isolating certain species from specific locations were variable. For instance, peaks 182 (518 bp), 194 (545 bp) and 203 (568 bp) are strongly associated with row 4:panel (7-8) in the biodynamic vineyard and therefore, the probability of being isolated from this area is higher than with other sites. Similar observations were made for the conventional vineyard.

6.4.3 Qualitative Diversity Analysis

A total of 628 yeast isolates from the three vineyards were analysed. Eleven species representing 8 genera were isolated from the conventional vineyard; the yeast isolated from the integrated vineyard represented 8 genera and 9 species, while 17 species representing 12 genera were isolated from the biodynamic vineyard. The biodynamic vineyard displayed a higher species richness and biodiversity than both the conventional and integrated vineyard (Table 6.1). Species evenness below 1 was found in all the vineyards. The dimorphic ascomycetous black yeast-like fungus, *Aureobasidium pullulans*, was widely distributed in the three vineyards (Table 6.2). *Cryptococcus spp.* were the second most prevalent yeast group with *Cr. magnus, Cr. carnescens* and *Cr. oeirensis* present in the three vineyards, while *Cr. laurentii* was only isolated from the integrated and biodynamic vineyard. The red pigmented yeasts including *Sporobolomyces roseus* and *Rhodotorula spp.*, were also frequently isolated in the three vineyards. The biodynamic vineyard displayed some unique diversity of the minor yeast species including *Exophiala sp., Kazachstania sp., Sporisorium sp., Ustilago sp.* and *Meira sp.* (Table 6.2). However, these yeasts

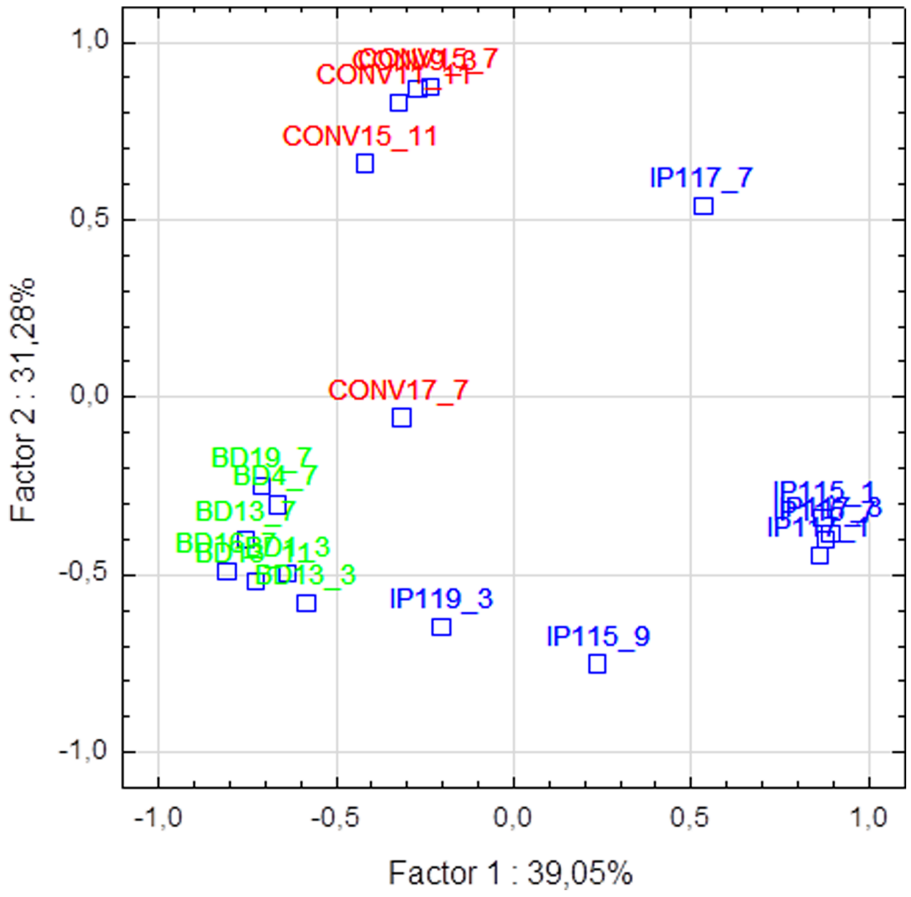


Figure 6.1: Principal component analysis based on fungal community structure assessed by ITS1-5.8S-ITS2 rRNA gene ARISA profiles. Biodynamic vineyard (Green), Conventional (Red), IPW (Blue).

were not evenly distributed within the vineyard. For instance, *S. roseus, Ustilago sp., Kazachstania sp.* and *R. diobovatum* were isolated from three of the 21 sampling sites. In addition, only one sampling site contained 9 of the 17 species isolated from the biodynamic vineyard. In the conventional vineyard, only *Cr. magnus*, and *Cr. oeirensis* were widely distributed in the vineyard, while *Rh. sloofiae* and *S. roseus* were isolated from 4 of the 18 sampling sites. *Rh. glutinis* and *Cr. magnus* were present in 6 of the 18 sites in the integrated vineyard, while *Issatchenkia terricola* and *Cr. oeirensis* were only retrieved from 1 sampling site. A community correlation network generated from culturable yeast diversity does not result in any obvious partitioning of the three vineyards (Figure 6.4).

6.4.4 Testing for Potential Pseudoreplication

The three adjacent vineyards used in this study were as similar as possible with regard to macro-environmental factors as described in the Materials and

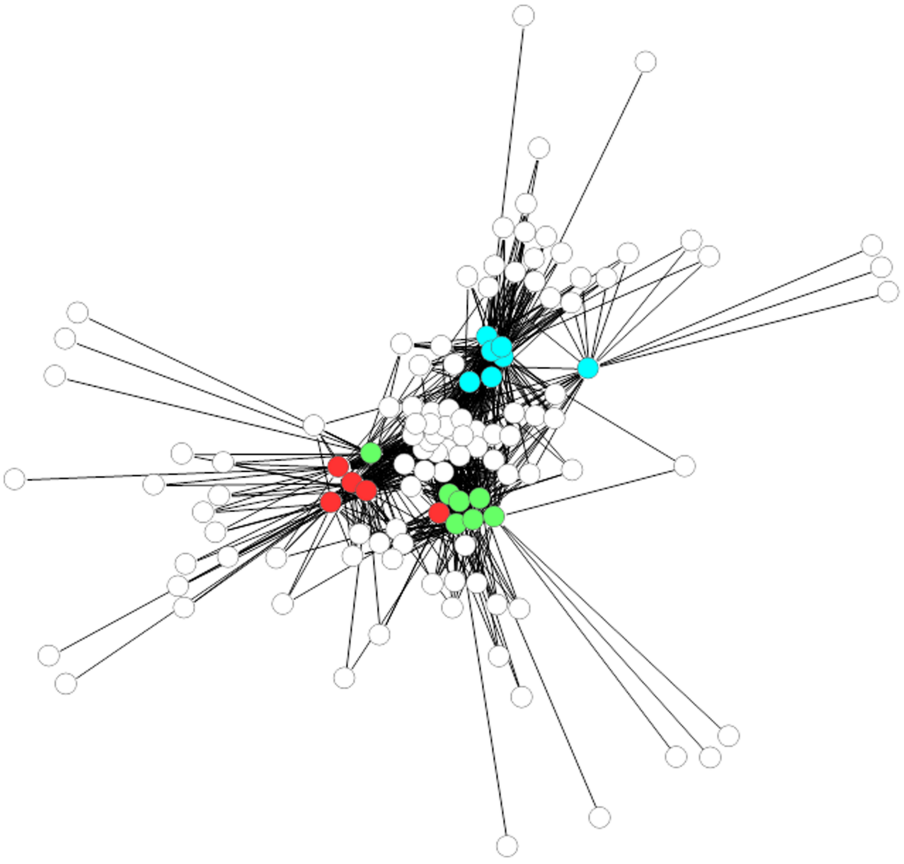


Figure 6.2: Probability network of OTU found at different sampling points. Sampling point nodes are coloured by farming practice: Biodynamic (Green), Conventional (Red) and IPW (Aqua). White nodes indicate OTUs common in the three vineyards.

Methods section. However, given that these were three different commercial vineyards which, as individual units, were managed with three different farming practices, the treatments, by definition, occurred within contiguous blocks. However, it is worth noting that the distances between sampling points within the Biodynamic and Conventional vineyards are often greater than the distance between the sampling points between those two vineyards. Nonetheless, pseudoreplication effects [11] from the sampling points within each vineyard are a possibility that we needed to address in order to ensure that environmental conditions [that have nothing to do with farming practice] present in each vineyard are not driving the selection of species. If this were the case one would expect to see a strong inverse correlation between the distance between sampling points and the species found in the samples. In order to test this, we devised a method which measures the distance between every possible sample pair within a vineyard and plots it against the species correlation values observed between those sample pairs. As can be seen in Figure S4, the R^2 values

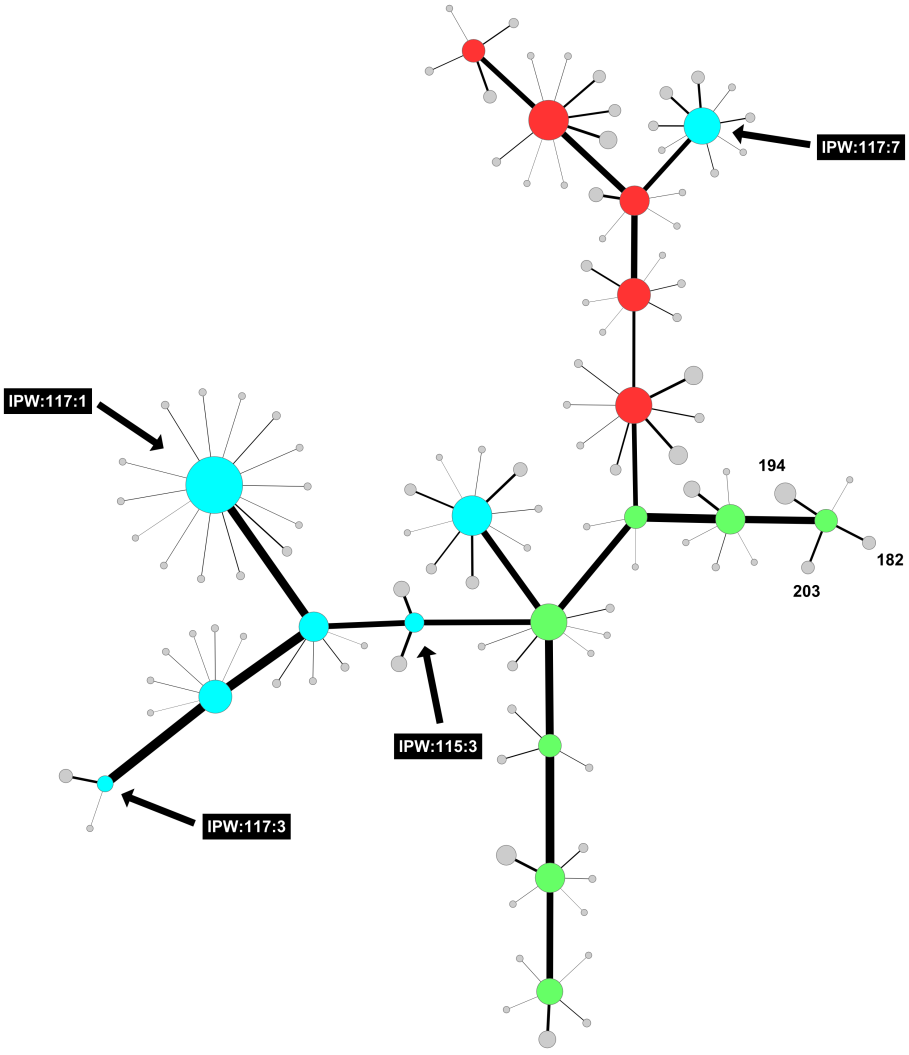


Figure 6.3: Mixed-Model Network: Sampling point Correlations and OTU probability distribution across samples. Sampling point nodes are coloured by farming practice: Biodynamic (Green), Conventional (Red) and IPW (Aqua). Sampling point node sizes are scaled by degree and OTU nodes by the probability of occurring in the adjacent sampling point. White nodes represent OTUs most likely to be isolated from a given sampling point.

for the conventional, biodynamic and IPW vineyards were 0.1063, -0.115 and 0.0106, respectively. As such, it appears that there is no correlative relationship between sample location within a vineyard and the species found there and thus apparently no pseudoreplication effect.

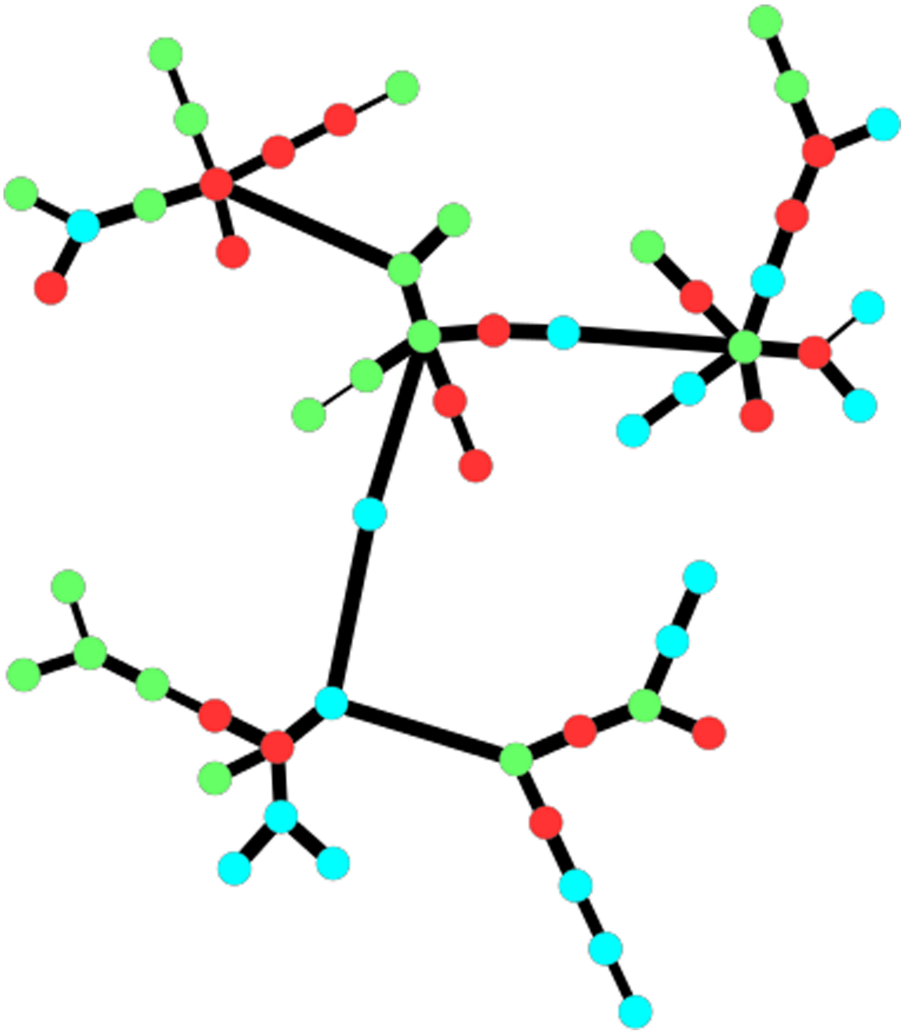


Figure 6.4: A correlation network of vineyard samples based on culturable yeast species. Nodes are coloured by farming practice: Biodynamic (Green), Conventional (Red) and IPW (Aqua).

Table 6.1: Ecological diversity indices determined using the yeast isolates obtained from the conventional (CONV), integrated (IPW) and biodynamic (BD) vineyard.

Vineyard	Menhinick's index (Species richness)	Pielou's index (Species evenness)	Shanon's Diversity Index (Species diversity)
CONV	0.96	0.5	1.20
IPW	1.06	0.63	1.45
BD	1.45	0.76	2.15

doi:10.1371/journal.pone.0052609.t001

Table 6.2: The occurrence and percentage distribution of yeasts associated with grape berries in the conventional (CONV), integrated (IPW) and biodynamic (BD) vineyards.

Species name	Accession number of closest relative	% Sequence identity	Yeast percentage distribution		
			CONV	IPW	BD
Aureobasidium pullulans	HM849057	99	70.4	63.2	52.5
Cryptococcus magnus	FN400937	100	9.2	7.9	6.6
Cryptococcus carnescens	EU149786	99	2.0	2.6	3.3
Cryptococcus oeirensis	AF444364	99	1.0	2.6	1.6
Cryptococcus laurentii	HM469461	100	–	2.6	2.5
Cryptococcus flavescens	FJ441026	100	–	–	3.3
Rhodotorula slooffiae	AF444589	99	4.1	–	4.1
Sporobolomyces roseus	AY015438	99	6.1	3.9	3.3
Cryptococcus saitoi	EU149781	99	–	–	0.8
Rhodosporidium diobovatum	HQ670682	99	–	–	5.7
Kazachstania sp.	AY582126	97	–	–	4.9
Pichia caribbica	EU568999	99	–	–	1.6
Candida parapsilosis	AB109228	99	–	–	1.6
Meira geulakonigii	GQ917051	99	–	–	1.6
Exophiala sp.	AB566310	97	–	–	1.6
Sporisorium sp.	AY344988	91	–	–	2.5
Ustilago sp.	AY740167	94	–	–	2.5
Candida sp.	FM178365	91	1.0	–	–
Saccharomycete sp.	FM178345	87	1.0	–	–
Bullera dendrophila	AF444443	94	3.1	–	–
Rhodotorula glutinis	HQ670677	99	–	7.9	–
Cryptococcus randhawii	AJ876528	97	–	2.6	–
Issatchenkia terricola	AY235808	99	1.0	6.6	–
Rhodotorula nothofagi	AY383749	99	1.0	–	–

doi:10.1371/journal.pone.0052609.t002

6.5 Discussion

The current study evaluated the impact of farming systems viz. conventional, integrated and biodynamic viticultural practices on grape associated yeast diversity. Due to their ease of manipulation, grape berries are a good model fruit with which to easily capture the diversity. Our focus was on sound grape berries as they provide a better reflection of vineyard diversity since damaged berries may result in the isolation of some fermentative yeasts which have been shown to be harboured and disseminated by fruit flies e.g. Drosophila sp. [7].

The yeast counts obtained in the current study were in the same order of magnitude (104-105 cfu/g) as previous reports on the density of yeast populations on healthy/sound grape berries [30]-[4], [24]. Molecular ecological networks based on data obtained from ARISA analysis were used to discern inter- and intra-vineyard variability. Our experimental results demonstrated that there were significant species overlaps between the three farming systems which is probably due to generalist fungal populations which are commonly present in vineyard settings. This finding could also be corroborated with cultivation-based methods, which revealed that the three vineyards shared certain common yeast species such as *Aureobasidium pullulans, Cyptococcus magnus, Sporobolomyces roseus*, and *Rhodotorula glutinis*. The yeast-like fungus *A. pullulans* was found to be the dominant yeast inhabiting the grape berry surface. This observation is consistent with previous studies which have applied culture-dependent methods as well as culture independent methods such as PCR-DGGE and FT-IR spectroscopy to monitor grape associated diversity [26], [9], [37]-[31]. In both culture-dependent and -independent approaches *A. pullulans* has been shown to account for 50-70% of the total population associated with undamaged grape berries. Other researchers have reported higher levels of this yeast-like fungus on organic vineyards than conventional vineyards. In contrast, our study shows a similar distribution of *A. pullulans* in the conventional, integrated and biodynamic vineyards. The dominance of this yeast-like fungus on grape surfaces has previously been attributed to its resistance to fungicides, the ability to detoxify $CuSO_4$ and the ability to compete against other fungi [32], [33]. Our data further shows that despite the overlap between the three farming systems, there is sufficient difference in the total fungal community composition to separate the three farming systems from each other. These differences could mainly be due to minor yeast species. For instance, the biodynamic vineyard displayed unique biodiversity which comprised members of the genera *Sporisorium, Meira* and *Exophiala*, which have never previously been associated with the vineyard environment. A higher number of yeasts with biocontrol potential including *Rhodosporidium diobovatum, Meira geulakoningii*, and *Cryptococcus laurentii* were isolated from the biodynamic vineyard. *M. geulakoningii* is a mite-associated yeast which has been shown to be active against different species of mites e.g. carmine spider mite (*Tetranychus cinnabarinus*) and citrus rust mite (*Phyllocoptruta oleivora*)

resulting in 100% mortality of the mites following treatment [23], [43]. This fungus could possibly be involved in suppressing mites such as *Tetranychus urticae* and *Colomerus vitis* which have been reported to be associated with grapevines in the Western Cape province of South Africa [44]. However, the distribution of this fungus and its actual role in the vineyard ecosystem needs to be investigated further. Other yeasts such as *Rh. diobovatum* and *Cr. laurentii* are also potential biocontrol agents against *B. cinerea*. This unique diversity could be due to the poor phytosanitary condition associated with the biodynamic vineyard, but it could also reflect the establishment of the natural enemies of different pests in the absence of pesticide application.

While previous research has alluded to spatial fluctuations within the vineyard, the extent to which such variations can occur with regard to grape berry associated microbiota has never been thoroughly investigated [6]. In addition to ruling out potential pseudoreplication effects from our sampling design, our data show that intra-vineyard variability can be attributed to considerable amounts of both inter- and intra-row spatial heterogeneity. This heterogeneity could be in part due to differences in immediate vine ecosystems and variation in inter-vine and intra-vine microclimates. For instance, the relative position of vines within the vineyard results in differences in the level of solar incoming radiation on the grape clusters, which in turn would affect the presence and proportion of pigmented yeasts such as member of the genera *Rhodotorula, Sporobolomyces* and *Rhodosporidium*. Ultimately, our study show that intravineyard variability is a significant factor, and may in some cases be higher than inter-vineyard differences even in cases of extreme treatment differences as applied to the blocks that were the subject of this study. This novel finding may lead to a reassessment of many previously published viticultural studies where the impact of vineyard treatments on wine composition was assessed. Indeed, this source of complexity has not been considered as a possible explanation for the observed heterogeneity of wines described in many such studies. Our data suggest that many differences may not derive from the differences in treatment, but rather differences in microbial diversity. The challenge in all field studies is the relatively large geo-spatial areas that need to be sampled and the logistical and financial limits to the number of samples that can be analysed. Often field studies take random samples from vineyards and, as such, may be reporting patterns that are not representative of the entire vineyard. The Theory of Sampling (TOS) has been developed over the past 50 years to deal with the sampling of large heterogeneous lots of material. From the results presented here it appears that viewing a vineyard as a large, heterogeneous two-dimensional lot, and using the TOS approach of lot-linearization and incremental sampling is an appropriate approach for such field studies in order to maximise the probability of collecting the most representative set of samples from a vineyard.

The current study shows unequivocally, that although culture-based methods generated interesting results regarding the microbial ecology of the vine-

yard, they were not an adequate approach to decipher the impact of farming systems on grape associated diversity probably due to the fact that this approach although not intentional, selects for certain groups of organisms, either due to their non-fastidious nature and rapid growth while excluding others whose cultivation requirements remain unknown. Overall, it was found that the sound grape berries are mainly colonized by oxidative yeasts, mainly *Aureobasidium pullulans* and *Cryptococcus spp.* These yeasts have been shown to occur on the surface of other parts of the phylloplane such as leaves and bark, as well as the soil [45]. Similar results have been reported in previous studies, and it is becoming more evident that although these yeasts are irrelevant to winemaking due their inability to ferment sugars or survive in wine, they represent the resident microbiota of grape berries [6], [5]. The biodynamic vineyard displayed higher diversity ($H' = 2.15$), while the conventional vineyard displayed the lowest diversity ($H' = 1.20$). However, the species evenness in the three vineyards was below 1, indicating the sparse distribution of the minor species. It could be speculated that the high diversity of the biodynamic vineyard is attributable to the fact that no fungicides except $CuSO_4$ are applied on the vineyard, however this needs to be investigated. Several studies have shown that fungicides do not have an impact on *Cryptococcus spp., Rhodotorula spp.* and *A. pullulans*, which may explain their higher frequency on all the vineyards [9], [33]. However, the impact of fungicides on other yeasts has not yet been investigated. Isolates of the genus *Kazachstania* were also obtained from the biodynamic vineyard. Although not frequently encountered in vineyard settings, the genus *Kazachstania's* association with wine grapes has been previously demonstrated [46]. Given the close proximity of the three vineyards, it could be speculated that there is limited cross-transfer of yeasts from one vineyard to the other especially regarding the minor yeasts. However, an in-depth analysis using culture-independent methods e.g. metagenomics would be best suited to provide further insight. In addition, it would be important to evaluate the microbial diversity over several vintages and at different grape ripening stages to confirm whether the distinction between the vineyards is persistent.

6.6 Supporting Information

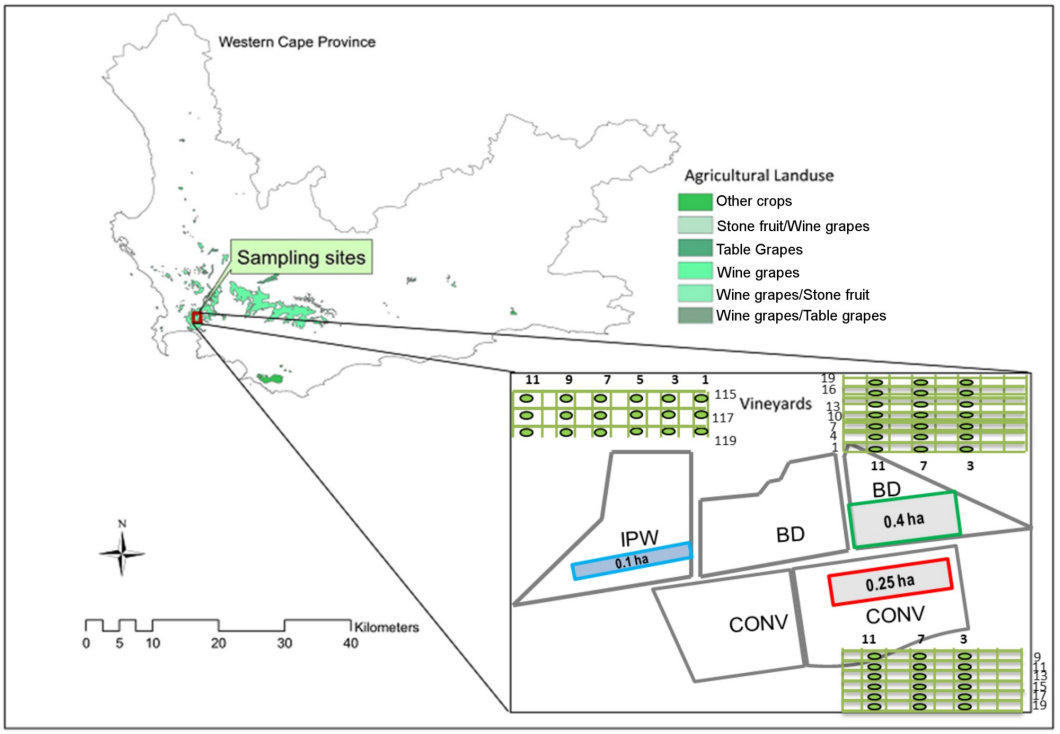


Figure S1: Geographic location of the study sites. IPW = integrated production of wine; BD = biodynamic; CONV = conventional.

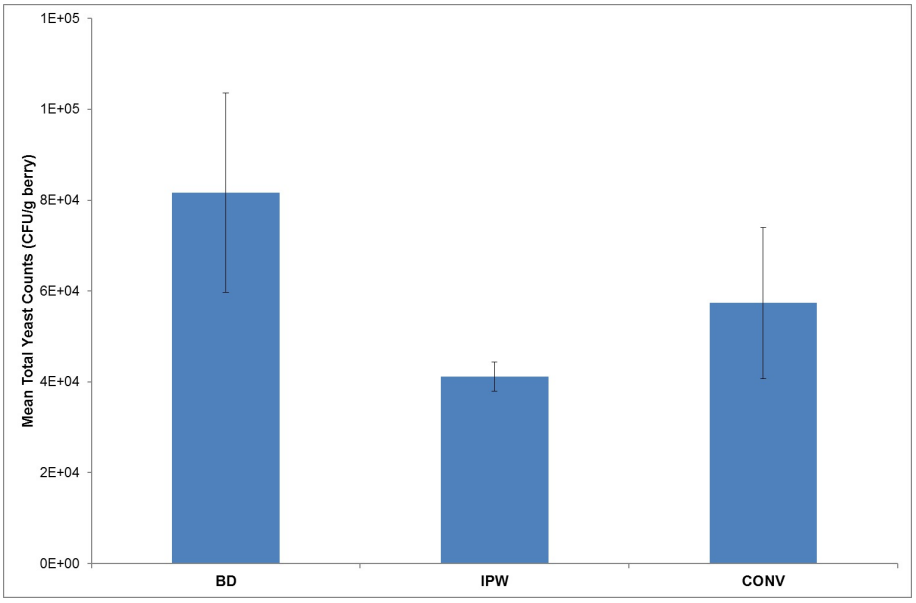


Figure S2: Total yeast populations enumerated on grape berry surfaces from biodynamic (BD), integrated production (IPW) and conventional (CONV) vineyards. The results were averaged from duplicate dilutions and are expressed as means Âś SE of total samples. Error bars represent the standard error of means.

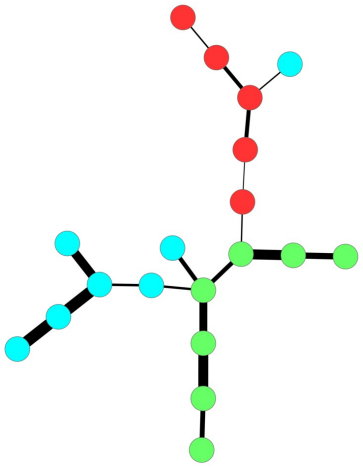


Figure S3: Correlation Network of Microbial Populations at different Sampling Points. Nodes are coloured by farming practice: Biodynamic (Green), Conventional (Red) and IPW (Aqua). Edge width is scaled to correlation value, so the thicker the edge the stronger the correlation.

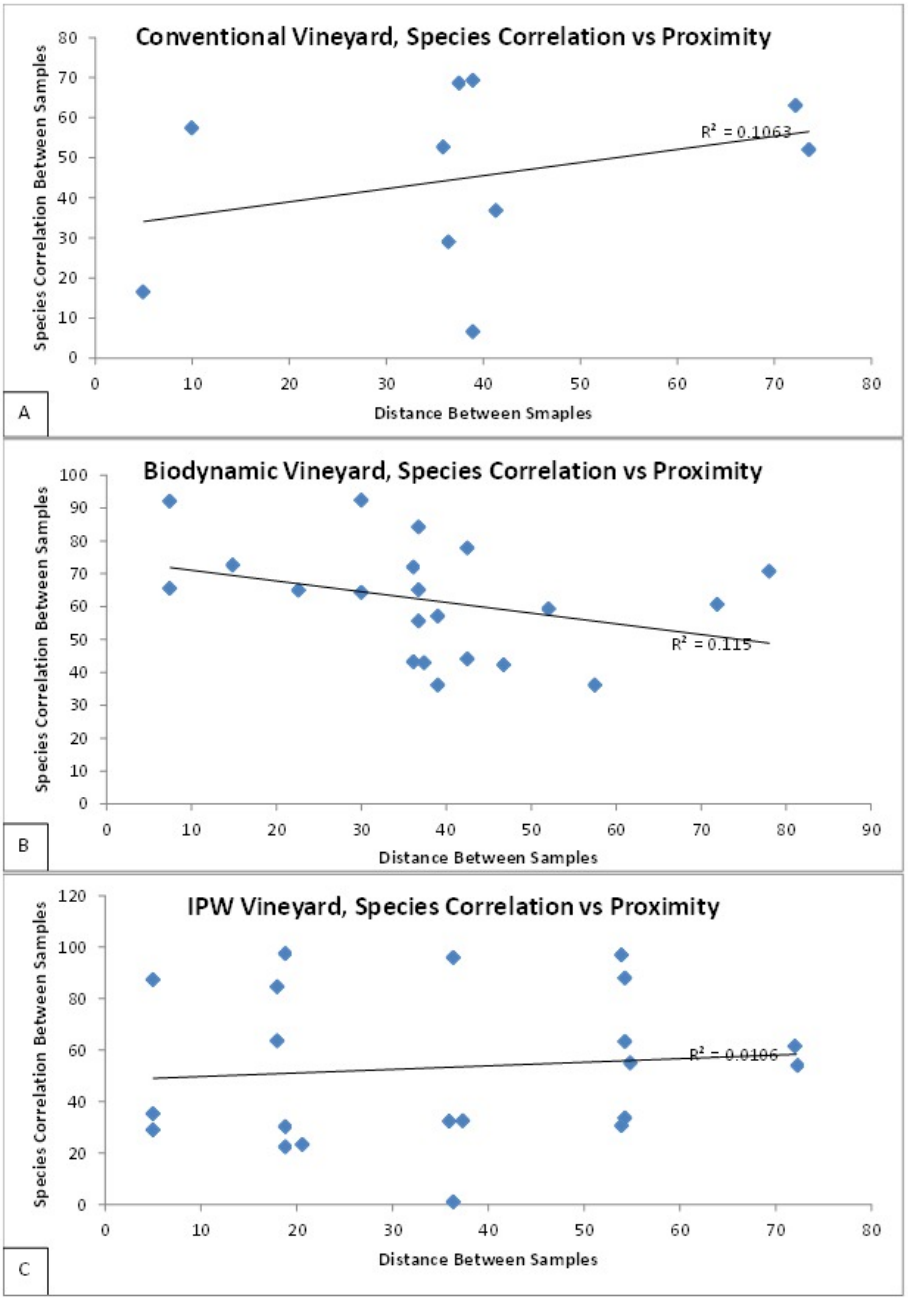


Figure S4: Species Correlation vs Spatial Distribution for each sample pair within vineyards for A) Conventional, B) Biodynamic and C) IPW vineyards.

Table S1: Spray programme for the biodynamic, conventional and integrated vineyard from leaf-fall till full bloom.

Spray	Biodynamic	Conventional	Integrated			
	Product	Dosage (g/ha)	Product	Dosage	Product	Dosage
1	Nordox Striker	390 450	Folpan Kumulus	312.5 g/ha 2.5 Kg/ha	Hyperphos Dithane Kumulus Acrobat	2 kg/ha 600 g/ha 3 kg/ha 1 kg /ha
2	Nordox Kumulus	390 600	Folpan Kumulus Rootex	312.5 g/ha 2.5 Kg/ha 750 ml/ha	Kumulus Acrobat	3 kg/ha 1 kg/ha
3	Nordox Kumulus	225 600	Dithane Kumulus Rootex	1 kg/ha 3 kg/ha 750 ml/ha	Kumulus Acrobat Talendo	3 kg/ha 1,5 kg/ha 1,5 kg/ha
4	Nordox Kumulus	390 600	Acrobat Talendo	1 kg/ha 25 g/ha	Hyperphos Curzate	3 kg/ha 1,5 kg/ha
5	Nordox Kumulus	390 600	Acrobat Talendo	1,5 kg/ha 25 g/ha	Talendo Hyperphos	1,5 kg/ha 4 kg/ha
6	Nordox Kumulus Lime	390 600 500	Cuprofix Kumulus	605 g/ha 3 kg/ha	Curzate Kumulus	2 kg/ha 3 kg/ha
7	Nordox Kumulus Lime	390 600 500	Dithane Topaz	2 kg/ha 30 g/ha	Hyperphos Dithane Stroby	4 kg/ha 1,2 kg/ha 150 g/ha
8			Dithane Topaz	2 kg/ha 30 g/ha	Dithane Stroby	1,2 kg/ha 150 g/ha

References

[1] (2010). Demeter production standards for the use of Demeter, Biodynamic and related trademarks. Available: www.demeter.net/node/494. Accessed 14 August 2012.

[2] (2010). Infruitec-Nietvoorbij, Agricultural Research Council. Integrated production of wine: Guidelines for farms. Available: http://www.ipw.co.za/guidelines.php. Accessed 14 June 2012.

[3] Altschul, S. (1997 September). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402. ISSN 13624962.

[4] Barata, A., González, S., Malfeito-Ferreira, M., Querol, A. and Loureiro, V. (2008 November). Sour rot-damaged grapes are sources of wine spoilage yeasts. *FEMS yeast research*, vol. 8, no. 7, pp. 1008–17. ISSN 1567-1356.

[5] Barata, A., Malfeito-Ferreira, M. and Loureiro, V. (2012 March). Changes in sour rotten grape berry microbiota during ripening and wine fermentation. *International journal of food microbiology*, vol. 154, no. 3, pp. 152–61. ISSN 1879-3460.

[6] Barata, A., Malfeito-Ferreira, M. and Loureiro, V. (2012 March). The microbial ecology of wine grape berries. *International journal of food microbiology*, vol. 153, no. 3, pp. 243–59. ISSN 1879-3460.

[7] Barata, A., Santos, S.C., Malfeito-Ferreira, M. and Loureiro, V. (2012 August). New insights into the ecological interaction between grape berry microorganisms and Drosophila flies during the development of sour rot. *Microbial ecology*, vol. 64, no. 2, pp. 416–30. ISSN 1432-184X.

[8] Barata, A., Seborro, F., Belloch, C., Malfeito-Ferreira, M. and Loureiro, V. (2008 April). Ascomycetous yeast species recovered from grapes damaged by honeydew and sour rot. *Journal of applied microbiology*, vol. 104, no. 4, pp. 1182–91. ISSN 1365-2672.

[9] Cadez, N., Zupan, J. and Raspor, P. (2010 August). The effect of fungicides on yeast communities associated with grape berries. *FEMS yeast research*, vol. 10, no. 5, pp. 619–30. ISSN 1567-1364.

[10] Combina, M., Mercado, L., Borgo, P., Elia, A., Jofré, V., Ganga, A., Martinez, C. and Catania, C. (2005 January). Yeasts associated to Malbec grape berries from Mendoza, Argentina. *Journal of applied microbiology*, vol. 98, no. 5, pp. 1055–61. ISSN 1364-5072.

[11] Comitini, F. and Ciani, M. (2008 September). Influence of fungicide treatments on the occurrence of yeast flora associated with wine grapes. *Annals of Microbiology*, vol. 58, no. 3, pp. 489–493. ISSN 1590-4261.

[12] Cordero-Bueso, G., Arroyo, T., Serrano, A., Tello, J., Aporta, I., Vélez, M.D. and Valero, E. (2011 January). Influence of the farming system and vine variety on yeast communities associated with grape berries. *International journal of food microbiology*, vol. 145, no. 1, pp. 132–9. ISSN 1879-3460.

[13] Dijkstra, E.W. (1959 December). A note on two problems in connexion with graphs. *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271. ISSN 0029-599X.

[14] Esbensen, K.H., Friis-Petersen, H.H., Petersen, L., Holm-Nielsen, J.B. and Mortensen, P.P. (2007 August). Representative process sampling âĂŤ in practice: Variographic analysis and estimation of total sampling errors (TSE). *Chemometrics and Intelligent Laboratory Systems*, vol. 88, no. 1, pp. 41–59. ISSN 01697439.

[15] Esbensen, K.H., Paoletti, C. and Minkkinen, P. (2012 February). Representative sampling of large kernel lots I. Theory of Sampling and variographic analysis. *TrAC Trends in Analytical Chemistry*, vol. 32, pp. 154–164. ISSN 01659936.

[16] Esbensen KH, M.P. (2004). 50 years of Pierre GyâĂŹs Theory of Sampling. Proceedings: First World Conference on Sampling and Blending (WCSB1). Tutorials on Sampling: Theory and Practise. *Chemometr Intell Lab Systems*, vol. 74, p. 236.

[17] Fleet, G. (2003 September). Yeast interactions and wine flavour. *International Journal of Food Microbiology*, vol. 86, no. 1-2, pp. 11–22. ISSN 01681605.

[18] Hoffman, C.S. (1997). Rapid isolation of yeast chromosomal DNA. *A: Current Protocols in Molecular Biology. Eds. FM Ausubel, R. Brent, RE Kingston, DD Moore, JG Seidman, JA Smith, i K. Struhl. John Willey & Sons. USA.*

[19] Hurlbert, S.H. (1984 June). Pseudoreplication and the Design of Ecological Field Experiments. *Ecological Monographs*, vol. 54, no. 2, p. 187. ISSN 00129615.

[20] Lamine, C. (2011 April). Transition pathways towards a robust ecologization of agriculture and the need for system redesign. Cases from organic farming and IPM. *Journal of Rural Studies*, vol. 27, no. 2, pp. 209–219. ISSN 07430167.

[21] Mäder, P., Fliessbach, A., Dubois, D., Gunst, L., Fried, P. and Niggli, U. (2002 May). Soil fertility and biodiversity in organic farming. *Science (New York, N.Y.)*, vol. 296, no. 5573, pp. 1694–7. ISSN 1095-9203.

[22] Nisiotou, A.A. and Nychas, G.-J.E. (2007 April). Yeast populations residing on healthy or botrytis-infected grapes from a vineyard in Attica, Greece. *Applied and environmental microbiology*, vol. 73, no. 8, pp. 2765–8. ISSN 0099-2240.

[23] Nisiotou, A.A. and Nychas, G.J.E. (2008). Kazachstania hellenica sp. nov., a novel ascomycetous yeast from a Botrytis-affected grape must fermentation. *International journal of systematic and evolutionary microbiology*, vol. 58, no. 5, pp. 1263–1267.

[24] Paz, Z., Burdman, S., Gerson, U. and Sztejnberg, A. (2007). Antagonistic effects of the endophytic fungus Meira geulakonigii on the citrus rust mite Phyllocoptruta oleivora. *Journal of applied microbiology*, vol. 103, no. 6, pp. 2570–2579.

[25] Pielou, E.C.J. (1966). The measurement of diversity in different types of biological collections. *Journal of theoretical biology*, vol. 13, pp. 131–144.

[26] Prakitchaiwattana, C.J., Fleet, G.H. and Heard, G.M. (2004 September). Application and evaluation of denaturing gradient gel electrophoresis to analyse the yeast ecology of wine grapes. *FEMS yeast research*, vol. 4, no. 8, pp. 865–77. ISSN 1567-1356.

[27] Raspor, P., Milek, D.M., Polanc, J., Mozina, S.S. and Cadez, N. (2006 May). Yeasts isolated from three varieties of grapes cultivated in different locations of the Dolenjska vine-growing region, Slovenia. *International journal of food microbiology*, vol. 109, no. 1-2, pp. 97–102. ISSN 0168-1605.

[28] Reeve, J.R., Carpenter-Boggs, L., Reganold, J.P., York, A.L. and Brinton, W.F. (2010 July). Influence of biodynamic preparations on compost development and resultant compost extracts on wheat seedling growth. *Bioresource technology*, vol. 101, no. 14, pp. 5658–66. ISSN 1873-2976.

[29] Reeve, J.R., Carpenter-Boggs, L., Reganold, J.P., York, A.L., McGourty, G. and McCloskey, L.P. (2005). Soil and winegrape quality in biodynamically and organically managed vineyards. *American journal of enology and viticulture*, vol. 56, no. 4, pp. 367–376.

[30] RENOUF, V., CLAISSE, O. and LONVAUD-FUNEL, A. (2005 October). Understanding the microbial ecosystem on the grape berry surface through numeration and identification of yeast and bacteria. *Australian Journal of Grape and Wine Research*, vol. 11, no. 3, pp. 316–327. ISSN 1322-7130.

[31] Sabate, J., Cano, J., Esteve-Zarzoso, B. and Guillamón, J.M. (2002 January). Isolation and identification of yeasts associated with vineyard and winery by RFLP analysis of ribosomal genes and mitochondrial DNA. *Microbiological research*, vol. 157, no. 4, pp. 267–74. ISSN 0944-5013.

[32] Schmid, F., Moser, G., Müller, H. and Berg, G. (2011 March). Functional and structural microbial diversity in organic and conventional viticulture: organic farming benefits natural biocontrol agents. *Applied and environmental microbiology*, vol. 77, no. 6, pp. 2188–91. ISSN 1098-5336.

[33] Schwartz, A. (1993). Occurrence of Natural Enemies of Phytophagous Mites. on Grape-vine Leaves Following. Application of Fungicides for Disease Con-trol. *S. Afr. J. Enol. Vitic*, vol. 14, no. 1, p. 16.

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

[35] Slabbert, E., Kongor, R.Y., Esler, K.J. and Jacobs, K. (2010 March). Microbial diversity and community structure in Fynbos soil. *Molecular ecology*, vol. 19, no. 5, pp. 1031–41. ISSN 1365-294X.

[36] Slabbert, E., Van Heerden, C.J. and Jacobs, K. (2010 July). Optimisation of automated ribosomal intergenic spacer analysis for the estimation of microbial diversity in fynbos soil. *South African Journal of Science*, vol. 106, no. 7/8. ISSN 1996-7489.

[37] Sztejnberg, A., Paz, Z., Boekhout, T., Gafni, A. and Gerson, U. (2004 November). A new fungus with dual biocontrol capabilities: reducing the numbers of phytophagous mites and powdery mildew disease damage. *Crop Protection*, vol. 23, no. 11, pp. 1125–1129. ISSN 02612194.

[38] Tofalo, R., Schirone, M., Telera, G.C., Manetta, A.C., Corsetti, A. and Suzzi, G. (2010 August). Influence of organic viticulture on non-Saccharomyces wine yeast populations. *Annals of Microbiology*, vol. 61, no. 1, pp. 57–66. ISSN 1590-4261.

6.7 Acknowledgments

We thank Dr. Etiene Slabbert (Department of Microbiology, Stellenbosch University) for his assistance with ARISA analyses. We would also like to thank Kim Esbensen for advice on the Theory of Sampling and Piet Jones and Debbie Weighill for useful discussions about the networks used in this study. We also thank Mr. Johan Reyneke, Mr. Deon Joubert as well as Mr and Mrs Emil and Sonet den Dulk for granting us permission to sample their vineyards.

Funding: This work was supported by Stellenbosch University (Subcom- B fund), WineTech and the National Research Foundation (THRIP programme). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

6.8 Author Contributions

Sampling Design: DJ, *Conceived and designed the experiments:* MES DJ FB. *Performed the experiments:* MES U-CA DJ. Analyzed the data: MES U-CA DJ. *Contributed reagents/materials/analysis tools:* MES DJ FB. *All Network-based analysis: DJ, Wrote the paper:* MES DJ FB.

6.8.1 Thesis Addendum

This chapter focuses on the analysis of a high dimensional microbiomics dataset of microbial populations, as defined by Automated Ribosomal Intergenic Spacer Analysis (ARISA), sampled within and across three different vineyards, each being managed with a distinct farming practice. In order to answer the research questions posed by this project we needed a multivariate method that could simultaneously: 1) determine if there was a high degree of heterogeneity across sampling sites within vineyards; 2) look for microbial population structures that were reflective of each farming practice and 3) detect if there were species that were present in all vineyards, regardless of farming practice. PCA could address the need to see if there were microbial population structures that reflected the farming practice but could not answer any of the other questions, nor did it quantitatively address the degree of correlations within and across farming practices. The solution that I derived involved the use of a maximum spanning tree of a Pearson correlation network, a probability network and the maximum spanning tree of a mixed model correlation-probability network.

The result was that with two figures, namely the probability network (Figure 6.2) and the mixed model network (Figure 6.3), we were able to answer all of the questions posed in a way that other multivariate data analysis tools were incapable of doing. Furthermore, this introduced the equivalent of the multivariate biplot into network analysis in which both samples and variables are represented in the same visualisation and the variables associated with the samples with which they are most associated. It appears that this mixed model approach is novel in the field of network analysis and has considerable advantages over existing multivariate analysis methods. Without the development of these methods the conclusions reached by this paper would not have been possible to come to nor to demonstrate so clearly.

None of this would have been possible without a proper sampling design with which to capture the patterns observed within and across vineyards. To achieve this I adapted the principles from the Theory of Sampling for use in vineyard microbiomics, something that also appears never to have been done before.

Chapter 7

Conclusions and Future Work

7.1 Concluding Remarks

7.1.1 Chemiomics

Chromatography has long been used for the separation of molecules enabling both the quantitative and qualitative analysis of constituents in complex mixtures. Targeted chromatography has been used extensively in the analysis of the chemical composition of wine, particularly with regard to finding compounds that contribute to flavour and aroma properties [22, 15, 1, 14, 3, 4, 2, 17]. However, one of the significant difficulties in gaining further understanding of the chemistry of wine from a systems perspective is the relative sparsity of known compounds for targeted analysis. Systemic insights will only be possible when a larger portion of the compounds in wine can be analysed simultaneously in order to identify which compounds are important to identify and study further.

In order to better understand the complex changes that occur in the volatile components during the aging of wine an untargeted GC-FID dataset of wines aged from between 2 and 60 years was generated by a collaborator in Portugal. The untargeted analysis of chromatographic datasets requires a significant amount of pre-processing in order to account for the differences between repeated chromatographic measurements. **The first aim of this thesis was to develop pre-processing methods using wavelets with which to create a refined alignment of chromatograms.**

The work presented in Chapter 3 addressed these pre-processing challenges by the development of an approach that combines correlation optimised warping (COW) [18] with wavelet-based alignment refinement. This combination of COW and wavelets yielded a finely aligned peak index suitable for use in network creation of untargeted chemiomics data. As such, the use and improvement of chromatographic alignments hold tremendous promise in the untargeted analysis of grapevine and microbial metabolites throughout the wine making process as well as the chemical compounds that reactions result from

bottle aging.

In un-targeted chemiomics most of the compounds detected are of unknown structure and function. As such, methods are needed with which to do internal contextualisation via the modelling of correlative relationships amongst samples and compounds in order to better understand the complex chemical mechanisms at play. **The second aim of this thesis was to develop multivariate and network-based methods with which to model un-targeted chemiomics datasets to show the correlative relationships amongst the compounds in order to further study the underlying chemical mechanisms occurring during wine aging.**

The work in presented in Chapter 3 demonstrated that this approach, as applied to GC-FID univariate data can be used to classify the age of Port wine and to predict potential molecules involved in this process by the deconvolution of time and the kinetics of different aging mechanisms. This approach was pursued with the hope that multivariate analysis and network reconstruction would be useful tools with which to study mechanisms related to a perturbation. The PCA score plots allow for sample classification and the visualization of larger peaks that correspond with aging. The PLS b coefficient plots provide the analytical chemist with a familiar set of patterns, namely virtual chromatograms which point out the larger peaks that appear to be associated with marker compounds for oxidation and Maillard mechanisms. The network reconstruction is very useful in visualizing the relationships between all the compounds detected via GC-FID and their changes in concentration over time. This view of the data should provide considerably more information in an effort to understand the probable kinetic contexts of the molecules represented by peaks in each chromatogram. Furthermore, it is possible to identify regions of the network that appear to be involved in the formation or consumption of target compounds. As such, the approach described here should indeed be a very powerful tool for the further study of mechanisms and kinetic networks in complex mixtures.

In summary, the approach reported in Chapter 3 enables one to: 1) minimize the cost of analysis; 2) perform sample classification and contextualization; 3) perform process monitoring of aging or other time series; 4) consider the prospect of building databases from the large amounts of univariate data already available; 5) screen for correlations with known mechanism markers; 6) select biomarkers for identification and further study and 7) explore the putative kinetic network for a greater understanding of the process being studied.

7.1.2 Gene Set Analysis

Most statistical methods currently used in the analysis of microarray-based gene expression datasets can not handle data generated with considerable variance between replicates due to the unknown orthogonal variance commonly present in field studies. Gene set analysis has been shown to be useful in this

regard [20, 19, 16, 5, 6, 10, 28, 30] , however, very little research has been done on gene set generation. **The third and fourth aims of this thesis were to explore new ways to generate gene sets to be used in gene set analyses and to address the high false positive and false negative rates of gene set analysis.**

Chapter 4 presents a novel method with which to generate gene sets via the use of PCA to decompose the Laplacian matrix of a metabolic network in order to semi-exhaustively explore the topology of the network. The overlapping nature of the resulting gene sets also allowed the development of a hypergeometric enrichment test to separate the driver genes from the passenger genes in gene set analysis and to then reconstruct a sparse biological network that represents the biology underlying the data. There only seems to be one report of the use of PCA on a network matrix [25]. The focus of their paper was on showing the relationship between PCA and spectral clustering, a link which is perhaps not surprising as both PCA and spectral graph theory are based on the eigenvalues, and eigenvectors of matrices. There appears to have been no report of the use of PCA for network topology discovery or set creation. As such, the work presented in Chapter 4 seems to quite novel and the resulting paper has been categorised as "Highly Accessed" by BMC Bioinformatics.

7.1.3 Network-based models for cross-species microarray analysis

Transcriptomic datasets present significant challenges involved in their biological interpretation. The most common output of most microarray analysis methods, a list of differentially expressed genes, is suboptimal for the generation of biological knowledge from the dataset. To create biological insights from such a dataset the experimental results need to be placed in the context of extant knowledge for further interpretation. **Thus, the fifth aim of this thesis was to develop methods to integrate as much annotation (extant knowledge) as possible about the genes and their biological context into network models which can be used for data visualisation and interpretation.**

In order to investigate *Botrytis-cinera*-resistant transgenic lines of tobacco [8] the global gene expression of two such transgenic lines (*VvPGIP1* lines 37 and 45) over-expressing a plant defence gene were compared to that of wild type (SR1) with the use of the TIGR 10K potato microarray by the plant biotechnology group at the IWBT. The analysis of data from this spotted potato cDNA microrarray cross-hybridised with tobacco transcripts presented a number of challenges: the gene to probe relationships are ambiguous, the original annotation was sparse and tobacco is not a model organism so there was relatively little information with which to contextualise the results. **The sixth aim of this thesis is to use phylogenomic information to de-**

termine orthologous relationships between tobacco, potato and nine sequenced plant genomes and, combined with a cross-hybridisation model, to develop a network-based framework for the analysis, annotation and interpretation of cross-species microarray data.

The work in Chapter 5 addressed the third and fourth aims with the use of network-based models for: cross-hybridisation of probes to transcripts; phylogenomic-based Gene Ontology, Interpro functional motifs and descriptive annotation; projection of differentially expressed probes onto pathways and protein-protein interaction models from other species and the creation of a cross-species co-expression network. These network-based methods led to the successful analysis of the data set, yielding insights into the differential expression of genes due to the over-expression of *Vvpgip1* in tobacco. This work has also led to the development of network-based models for gene duplication in an evolutionary framework in order to explore the relationships amongst species as regards to gene family content as well as the evolutionary relationships amongst gene families (not reported on in this thesis).

7.1.4 Mixed Model Networks for Microbiomics

Cultivation-based and molecular datasets were generated in order to investigate the spatial heterogeneity of microbial communities within vineyards and to determine if farming practices have an impact on microbial population structure across vineyards. The analysis of this microbiomic dataset needed the development of new methods with which to sample, delineate and interpret microbial population structures, their geospatial distribution and their responses to environmental conditions or perturbations.

More specifically, for this project we needed a multivariate method that could simultaneously: 1) determine if there was a high degree of heterogeneity across sampling sites within vineyards; 2) look for microbial population structures that were reflective of each farming practice and 3) detect if there were species that were present in all vineyards, regardless of farming practice. PCA could address the need to see if there were microbial population structures that reflected the farming practice but could not answer any of the other questions, nor did it quantitatively address the degree of correlations within and across farming practices. **Thus, the seventh aim of this thesis was to develop novel mixed-model networks, which combine sample correlations and microbial community distribution probabilities to use for the analysis of vineyard microbiomes.**

The solution presented in Chapter 6 involved the use of a maximum spanning tree of a Pearson correlation network, a probability network and the maximum spanning tree of a mixed-model correlation-probability network. This appears to be a novel approach as there does not seem to be any previous report in the literature of this type of mixed-model network. This method is also particularly interesting as it brings the idea of the *biplot* (which is commonly

used in multivariate data analysis to show the relationships amongst samples and variables) to the realm of network-based analysis. This method certainly merits further exploration and the paper in which this method was introduced has proven to be very popular (4,447 views in its first month of publication) and has gained considerable attention in the public media around the world, including the New York Times.

7.2 Future Work

The common thread through this thesis is the analysis of several types of -omics data and the need to model, analyse and visualise them in an integrated manner in order to extract the maximum amount of biological or chemical information and further our understanding of complex systems involved in wine making. Currently, multivariate approaches such as PCA and PLS are commonly applied to -omics datasets. Although very useful, these methods have a number of shortcomings. PCA and PLS, are, by definition, variance-based methods and, as such, will focus on maximizing the spread of samples or variables in their respective coordinate spaces without regard to how strong the relationships between the data points really are. The primary visualization method for PCA is a two dimensional plot along two different component axes. The two dimensional plots of objects or variables in principal coordinate space give a relativistic positioning of data points rather than a more quantitative measure of the strength of the relationships amongst them. When data points are projected to similar score or loading values along a component they are often interpreted as being correlated, however the strength of that putative correlation is completely unknown. Furthermore, in a two dimensional plot there is a limited amount of topology that can be represented and the analysis is purely visual in nature. Also of note is that PCA and PLS are designed to find linear patterns of covariance and, as such, are likely to miss non-linear patterns that are present in -omics wine datasets.

This thesis has made considerable use of networks to address the challenges inherent in several wine-related -omics datasets. The network-based approaches were able to address issues and answer research questions that PCA and PLS simply could not handle. In my opinion the network-based methods developed are a significant step forward when compared to existing methods. However, there is much more that can be done with networks than has been addressed to date. Future work will build on the approaches taken in this thesis and expand them with the use of network theory metrics and topology mining and further exploration of the intersects between network and multivariate analysis. Briefly, of interest includes the development of methods based on mixed-model networks, including mixtures of correlation metrics such as Spearman rank correlation [27], Weighted Rank Correlation [23], Kendall [9], Hoeffding's D measure [7], Theil-Sen [29], [26], Rank Theil-

Sen [13], Distance Covariance [11], Maximal Information [24], Pearson [21], Gausian Graphs [12] and probability theory to define novel networks with which to do high throughput hypothesis generation. It will be particularly interesting to evaluate and compare the resulting networks with a range of network-theory-based metrics. In addition, the mining of these networks for topological modules with which to evaluate the networks as well as to infer biological mechanisms from would seem to be an angle worth pursuing. The use of multiple correlation metrics will allow us to address one of the current shortcomings of PCA, PLS and the Pearson coefficient-based networks used to date, that is the specificity for linear covariance or correlation discovery. A mixed-metric approach will allow us to find non-linear patterns that are likely to be resident in biological datasets. It will also be interesting to explore the use of the scores and loading values of multivariate techniques such as PCA as the edge metrics for network creation and analysis and to compare the resulting networks to those derived with the various correlation metrics listed above.

Finally, in order to get genome-wide coverage by gene sets it would seem that the use of principal components analysis of the adjacency and Laplacian matrices of extant biological networks (metabolic networks, protein-protein-interaction networks, genetic interaction networks, transcription factor networks and literature-mining-based networks) for set generation for gene set analysis in order to extract differential gene expression signatures out of complex datasets would be a fruitful avenue to pursue.

References

[1] e Silva, H., de Pinho, P., Machado, B.P., Hogg, T., Marques, J.C., Câmara, J.S., Albuquerque, F. and Silva Ferreira, A.C. (2008). Impact of forced-aging process on Madeira wine flavor. *Journal of agricultural and food chemistry*, vol. 56, no. 24, pp. 11989–11996.

[2] Escudero, A., Cacho, J. and Ferreira, V. (2000). Isolation and identification of odorants generated in wine during its oxidation: A gas chromatography–olfactometric study. *European Food Research and Technology*, vol. 211, no. 2, pp. 105–110.

[3] Ferreira, A.C.S., Barbe, J.C. and Bertrand, A. (2003). 3-Hydroxy-4, 5-dimethyl-2 (5 H)-furanone: a key odorant of the typical aroma of oxidative aged port wine. *Journal of agricultural and food chemistry*, vol. 51, no. 15, pp. 4356–4363.

[4] Ferreira, A.C.S., Hogg, T. and de Pinho, P.G. (2003). Identification of key odorants related to the typical aroma of oxidation-spoiled white wines. *Journal of agricultural and food chemistry*, vol. 51, no. 5, pp. 1377–1381.

[5] Goeman, J.J., Oosting, J., Cleton-Jansen, A.-M., Anninga, J.K. and van Houwelingen, H.C. (2005 May). Testing association of a pathway with survival using gene expression data. *Bioinformatics (Oxford, England)*, vol. 21, no. 9, pp. 1950–7. ISSN 1367-4803.

[6] Goeman, J.J., van de Geer, S.A., de Kort, F. and van Houwelingen, H.C. (2004 January). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics (Oxford, England)*, vol. 20, no. 1, pp. 93–9. ISSN 1367-4803.

[7] Hoeffding, W. (1948). A non-parametric test of independence. *The Annals of Mathematical Statistics*, vol. 19, pp. 546–557.

[8] Joubert, D.A., Slaughter, A.R., Kemp, G., Becker, J.V.W., Krooshof, G.H., Bergmann, C., Benen, J., Pretorius, I.S. and Vivier, M.A. (2006 December). The grapevine polygalacturonase-inhibiting protein (VvPGIP1) reduces Botrytis cinerea susceptibility in transgenic tobacco and differentially inhibits fungal polygalacturonases. *Transgenic research*, vol. 15, no. 6, pp. 687–702. ISSN 0962-8819.

[9] Kendall, M.G. (1948). Rank correlation methods.

[10] Kim, S.-Y. and Volsky, D.J. (2005 January). PAGE: parametric analysis of gene set enrichment. *BMC bioinformatics*, vol. 6, p. 144. ISSN 1471-2105.

[11] Kosorok, M.R. (2009 January). On Brownian Distance Covariance and High Dimensional Data. *The annals of applied statistics*, vol. 3, no. 4, pp. 1266–1269. ISSN 1932-6157. `arXiv:1010.0297v2`.

[12] Krumsiek, J., Suhre, K., Illig, T., Adamski, J. and Theis, F.J. (2011 January). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC systems biology*, vol. 5, no. 1, p. 21. ISSN 1752-0509.

[13] Kumari, S., Nie, J., Chen, H.-S., Ma, H., Stewart, R., Li, X., Lu, M.-Z., Taylor, W.M. and Wei, H. (2012 January). Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. *PloS one*, vol. 7, no. 11, p. e50411. ISSN 1932-6203.

[14] Lavigne, V., Pons, A., Darriet, P. and Dubourdieu, D. (2008). Changes in the sotolon content of dry white wines during barrel and bottle aging. *Journal of agricultural and food chemistry*, vol. 56, no. 8, pp. 2688–2693.

[15] Marchand, S., de Revel, G. and Bertrand, A. (2000). Approaches to wine aroma: release of aroma compounds from reactions between cysteine and carbonyl compounds in wine. *Journal of agricultural and food chemistry*, vol. 48, no. 10, pp. 4890–4895.

[16] Mootha, V.K., Lindgren, C.M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrå le, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D. and Groop, L.C. (2003 July). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, vol. 34, no. 3, pp. 267–73. ISSN 1061-4036.

[17] Moutounet, M., Rabier, P., Puech, J.L., Verette, E. and Barillere, J.M. (1989). Analysis by HPLC of extractable substances in oak wood. Application to a Chardonnay wine. *Sci. Aliments*, vol. 9, no. 1, pp. 35–51.

[18] Nielsen, N.-P.V., Carstensen, J.M. and Smedsgaard, J.r. (1998 May). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A*, vol. 805, no. 1-2, pp. 17–35. ISSN 00219673.

[19] Pavlidis, P., Qin, J., Arango, V., Mann, J.J. and Sibille, E. (2004 June). Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochemical research*, vol. 29, no. 6, pp. 1213–22. ISSN 0364-3190.

[20] Pavlidis P, Lewis DP, N.W. (2002). Exploring gene expression data with class scores. In: *Pac Symp Biocomput*, pp. 474–485.

[21] Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Philos. Trans. Royal Soc. London Ser. A*, vol. 187, pp. 253–318.

[22] Pripis-Nicolau, L., De Revel, G., Bertrand, A. and Maujean, A. (2000). Formation of flavor components by the reaction of amino acid and carbonyl compounds in mild conditions. *Journal of agricultural and food chemistry*, vol. 48, no. 9, pp. 3761–3766.

[23] Quade D, S.I. (1992). A survey of weighted rank correlation. In: P.K. Salama, PK, S.I. (ed.), *Order statistics and nonparametrics: theory and applications*, pp. 213–225. Elsevier Science Publishers B.V., Amsterdam.

[24] Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M. and Sabeti, P.C. (2011 December). Detecting novel associations in large data sets. *Science (New York, N.Y.)*, vol. 334, no. 6062, pp. 1518–24. ISSN 1095-9203.

[25] Saerens, M., Fouss, F., Yen, L. and Dupont, P. (2004). The principal components analysis of a graph, and its relationships to spectral clustering. *Machine Learning: ECML 2004*, pp. 371–383.

[26] Sen, P.K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, vol. 63, no. 324, pp. 1379–1389.

[27] Spearman, C. (1987). The proof and measurement of association between two things. *The American journal of psychology*, vol. 100, no. 3/4, pp. 441–471.

[28] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P. (2005 October). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–50. ISSN 0027-8424.

[29] Theil, H. (1992). A rank-invariant method of linear and polynomial regression analysis. *Henri Theil's Contributions to Economics and Econometrics*, pp. 345—-381.

[30] Tian, L., Greenberg, S.A., Kong, S.W., Altschuler, J., Kohane, I.S. and Park, P.J. (2005 September). Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 38, pp. 13544–9. ISSN 0027-8424.