Exploring the Topology of Complex Phylogenomic and Transcriptomic Networks

by

Deborah A. Weighill

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

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

Supervisor: Dr. D.A. Jacobson

December 2014

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

Declaration

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

Date: ....11/08/2014..........

Abstract

Exploring the Topology of Complex Phylogenomic and Transcriptomic Networks

D.A. Weighill

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

Thesis: MSc Wine Biotechnology (Computational Biology)

December 2014

This thesis involved the development and application of network approaches for the construction, analysis and visualization of phylogenomic and transcriptomic networks.

A co-evolutionary network model of grapevine genes was constructed based on three mechanisms of evolution. The investigation of local neighbourhoods of this network revealed groups of functionally related genes, illustrating that the multi-mechanism evolutionary model was identifying groups of potentially co-evolving genes.

An extended network definition, namely 3-way networks, was investigated, in which edges model relationships between triplets of objects. Strategies for weighting and pruning these 3-way networks were developed and applied to a phylogenomic dataset of 211 bacterial genomes. These 3-way bacterial networks were compared to standard 2-way network models constructed from the same dataset. The 3-way networks modelled more complex relationships and revealed relationships which were missed by the two-way network models.

Network meta-modelling was explored in which global network and node-by-node network comparison techniques were applied in order to investigate the effect of the similarity metric chosen on the topology of multiple types of networks, including transcriptomic and phylogenomic networks. Two new network comparison techniques were developed, namely PCA of Topology Profiles and Cross-Network Topological Overlap. PCA of Topology Profiles compares networks based on a selection of network topology indices, whereas Cross-Network Topological Overlap compares two networks on a node-by-node level, identifying nodes in two networks with similar neighbourhood topology and thus highlighting areas of the networks with conflicting topologies. These network comparison methods clearly indicated how the similarity metric chosen to weight the edges of the network influences the resulting network topology, consequently influencing the biological interpretation of the networks.

Uittreksel

Exploring the Topology of Complex Phylogenomic and Transcriptomic Networks

("Exploring the Topology of Complex Phylogenomic and Transcriptomic Networks")

D.A. Weighill

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

Tesis: MSc Wyn Biotegnologie

Desember 2014

Hierdie tesis hou verband met die ontwikkeling en toepassing van netwerk benaderings vir die konstruksie, analise en visualisering van filogenomiese en transkriptomiese netwerke.

'n Mede-evolusionrê netwerk model van wingerdstok gene is gebou, gebaseerd op drie meganismes van evolusie. Die ondersoek van plaaslike omgewings van die netwerk het groepe funksioneel verwante gene aan die lig gebring, wat daarop dui dat die multi-meganisme evolusionêre model groepe van potensieele mede-evolusieerende gene identifiseer.

'n Uitgebreide netwerk definisie, naamliks 3-gang netwerke, is ondersoek, waarin lyne die verhoudings tussen drieling voorwerpe voorstel. Strategieë vir weeg en snoei van hierdie 3-gang netwerke was ontwikkel en op 'n filogenomiese datastel van 211 bakteriële genome toegepas. Hierdie 3-gang bakteriële netwerke is met die standaard 2-gang netwerk modelle wat saamgestel is uit dieselfde datastel vergelyk. Die 3-gang netwerke het meer komplekse verhoudings gemodelleer en het verhoudings openbaar wat deur die tweerigting-netwerk modelle gemis is.

Verder is netwerk meta-modellering ondersoek waarby globalle netwerk en punt-vir-punt netwerk vergelykings tegnieke toegepas is, met die doel om die effek van die ooreenkoms-maatstaf wat gekies is op die topologie van verskeie tipes netwerke, insluitend transcriptomic en filogenomiese netwerke, te bepaal.

Twee nuwe netwerk-vergelyking tegnieke is ontwikkel, naamlik "PCA of Topology Profiles" en "Cross-Network Topological Overlap". PCA van Topologie Profiele vergelyk netwerke gebaseer op 'n seleksie van netwerk topologie indekse, terwyl Cross-netwerk Topologiese Oorvleuel vergelyk twee netwerke op 'n punt-vir-punt vlak, en identifiseer punte in twee netwerke met soortgelyke lokale topologie en dus lê klem op gebiede van die netwerke met botsende topologieë. Hierdie netwerk-vergelyking metodes dui duidelik aan hoe die ooreenkoms maatstaf wat gekies is om die lyne van die netwerk gewig te gee, die gevolglike netwerk topologie beïnvloed, wat weer die biologiese interpretasie van die netwerke kan beïnvloed.

Acknowledgements

I would like to acknowledge Carissa Blekker and Piet Jones for assistance in the translation of the Abstract to Afrikaans, Piet Jones and Armin Geiger for useful discussions, as well as Daniel Jacobson for supervision, input, advice and support.

Dedications

To Tom, Malley, Dan and Ralph - Husband, companion, friend and bar.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Aims

1.1 Network Models and Evolution

Evolution is a heterogeneous process which can occur through various mechanisms. Point mutations which occur in coding regions may cause an amino acid change which may change the functionality of the resulting protein. Genes can also undergo duplication or deletion [1]. Duplication results in multiple copies of the same gene, allowing divergence of the duplicated genes through further mutation, possibly even evolving new functions [1]. Evolution can also occur through the evolution of gene expression regulation [2]. This involves point mutations, duplications or deletions which occur in the regulatory regions of genes or in separate regulatory elements.

These various models and mechanisms of evolution have previously been studied in isolation through the use of network models. Networks are useful tools for the analysis of biological systems. Being inherently complex, biological systems require the simultaneous modelling of many different components in order to properly represent the system. Networks allow for this level of complexity to be represented in that they model the interactions and relationships between components of a complex system in a pairwise manner, and represent the whole underlying system in an abstract form [3]. In a sense, networks make use of the advantages of reductionism, quantifying relationships between components of a system on an individual, pairwise manner, but still account for the overall complexity of the system by reconnecting all the components through their pairwise relationships.

What makes networks particularly useful is that they not only provide a platform for representing complex systems, but also an intuitive approach for the visualization of complex systems in the form of nodes connected by edges, and, in addition, a wealth of analysis methods can be applied to data represented as a network, such as clustering algorithms [4] and topological descriptors [5].

Networks have previously been applied in the modelling of these various mechanisms of evolution. Specific types of networks, namely trees, have been widely used to model evolution through point mutation through the construction of phylogenetic trees. Phylogenomic networks on the other hand, model the evolutionary relationships between organisms [6] often based on gene family content, a measure based on the evolutionary mechanism of gene duplication. Networks have also been used in the field of transcriptomics, in which they can be used to represent similarities between the expression profiles of genes [7]. Evolution through gene expression regulation has been investigated through cross-species co-expression analysis, identifying modules of co-expressed genes conserved across species [8].

One very widely used network analysis method is the extraction of groups of highly connected nodes, called modules. Depending on what kind of objects and relationships the network is modelling, these modules can have many different interpretations and uses. For networks in which nodes represent genes and edges model the similarities between genes based on sequence similarity, modules of highly connected genes can be interpreted as gene families [9]. In gene co-expression networks where nodes represent genes and edges model the similarity between expression profiles of genes, modules of highly connected nodes represent groups of co-expressed genes which are potentially functionally related [7].

Networks are clearly very useful tools in representing, analysing and visualizing complex systems. Thus, the exploration and development of new types of network methods and new network-based approaches is a useful endeavour in biological data analysis.

1.2 Aims

This thesis focuses on the development and application of new network approaches for the analysis of omics datasets, in particular, genomic and transcriptomic datasets. These datasets are large and complex in nature, and require analysis and visual representation before biological interpretations can be extracted. The aims of this thesis were to investigate new network approaches which combine networks resulting from different data types, investigate extended network definitions apart from the standard network structure of modelling pairwise relationships, and develop methods for network metamodelling - the comparison of network models.

1. Network models have previously been applied separately to model different mechanisms of evolution, namely evolution by gene expression

regulation through cross-species co-expression analysis [8], evolution by point-mutation through Evolutionary Rate Covariation [10; 11] and evolution by gene duplication through gene family analysis [6; 12]. However, to our knowledge, a combined network model representing these three mechanisms of evolution simultaneously has not been created. The first aim was to: construct modules of co-evolving grapevine genes in terms of these three mechanism of evolution; determine an approach for combining these three types of network modules into a super-network; and mine this super-network for functional insights.

2. Hypergraphs [13] are generalized graphs which do not restrict the edges to only modelling pairwise relationships. To our knowledge, these structures have not yet been applied in the field of phylogenomics. The second aim was to: investigate and develop an extended network definition (3-way networks) based on that of a hypergraph in which edges in a network model the relationships between triplets of objects; to investigate and develop weighting and pruning strategies for 3-way networks; apply these 3-way networks to a phylogenomic dataset of 211 bacterial genomes; and to compare the resulting 3-way networks to standard 2-way network models.

3. The final aim was to explore network meta-modelling (the comparison of network models) and to develop new approaches for network comparison on a whole-network level and on a node-by-node level.

1.3 Summary

Networks are useful structures for the representation, analysis and visualization of complex systems, and have been successfully applied in various areas of biology. This Master's study involves the development and application of new network approaches in the fields of evolution, phylogenomics and transcriptomics, exploration of the application of extended network definitions, and tools and approaches for comparing network models. These approaches include network-based analysis, as well as utilization of the intuitive visualization techniques which accompany the use of network models.

Bibliography

[1] Zhang, J.: Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, vol. 18, no. 6, pp. 292–298, 2003.

[2] Carroll, S.B.: Evolution at two levels: on genes and form. *PLoS Biology*, vol. 3, no. 7, p. e245, 2005.

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

[4] van Dongen, S.: *Graph clustering by flow simulation.* Ph.D. thesis, University of Utrecht, 2000.

[5] Horvath, S. and Dong, J.: Geometric interpretation of gene coexpression network analysis. *PLoS Computational Biology*, vol. 4, no. 8, p. e1000117, 2008.

[6] Dagan, T.: Phylogenomic networks. *Trends in Microbiology*, vol. 19, no. 10, pp. 483–491, 2011.

[7] Aoki, K., Ogata, Y. and Shibata, D.: Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant and Cell Physiology*, vol. 48, no. 3, pp. 381–390, 2007.

[8] Movahedi, S., Van Bel, M., Heyndrickx, K.S. and Vandepoele, K.: Comparative co-expression analysis in plant biology. *Plant, Cell & Environment*, vol. 35, pp. 1787–1798, 2012.

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

[10] Clark, N.L., Alani, E. and Aquadro, C.F.: Evolutionary rate covariation reveals shared functionality and coexpression of genes. *Genome Research*, vol. 22, no. 4, pp. 714–720, 2012.

[11] Sato, T., Yamanishi, Y., Kanehisa, M. and Toh, H.: The inference of protein–protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics*, vol. 21, no. 17, pp. 3482–3489, 2005.

[12] Snel, B., Bork, P., Huynen, M. *et al.*: Genome phylogeny based on gene content. *Nature Genetics*, vol. 21, pp. 108–110, 1999.

[13] Zhou, D., Huang, J. and Schölkopf, B.: Learning with hypergraphs: Clustering, classification, and embedding. In: *Advances in Neural Information Processing Systems*, pp. 1601–1608. 2006.

Chapter 2

Literature Review

2.1 Introduction

Networks are useful tools for understanding complex systems, and have been widely used to represent and investigate complex systems across many fields, including biological networks, communication networks and citation networks [1]. Biological systems are inherently complex, their individual components almost never operating in isolation. Understanding biological systems thus cannot be achieved through pure reductionist approaches involving studying the components of the system in isolation [2]. Network theory has provided the tools necessary to represent and visualize systems as a whole, accounting for complexity, yet allowing for resolution on a local and global scale. This review will cover the basic underlying principles of network theory and its roots in graph theory, how networks can be constructed, weighted, pruned and clustered and the various methods and metrics needed to do so. Ways in which networks can be numerically described through topological descriptors will then be reviewed. Lastly, applications and uses of networks in the fields of phylogenomics and transcriptomics will be discussed.

2.2 Network Theory

2.2.1 Overview

Networks are very useful tools which have been used increasingly to represent complex systems. They involve a certain reductionist-like approach in that they allow one to break a system down into individual parts called nodes and model the relationships between nodes in a pairwise manner. These relationships are called edges and are represented as lines drawn between the nodes. The overall complex system is then reconstructed by piecing together the overall network of nodes connected by edges [2]. Since the whole system is pieced back together, networks are also non-reductionist and allow the system to be examined as a whole as opposed to examining all of the parts in isolation.

2.2.2 Basic Graph Theory

Network Theory has its roots in the mathematical field of graph theory. A network can be defined mathematically as a graph, which is a structure consisting of nodes connected by edges. This can be formalized as follows: A graph G is defined as

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

where V is a set of nodes and E is a set of edges [3]. Networks can be represented visually by drawing the nodes as circles and labelling them, and then connecting the nodes by drawing lines between them, representing the edges. For example, consider a graph where

$$V = \{A, B, C, D, E, F, G\},$$

$$E = \{\{A, G\}, \{A, B\}, \{B, G\}, \{G, F\}, \{F, C\}, \{C, D\}, \{D, E\}, \{E, C\}\}.$$

This graph is represented visually in Figure 2.1a.

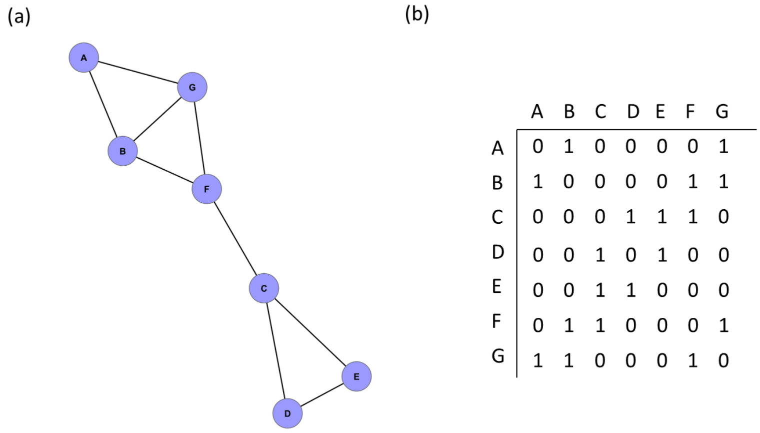
Another representation of a graph is the adjacency matrix. This is a numerical representation in which the rows and columns of the matrix represent nodes and each entry a_{ij} in the adjacency matrix is defined as [3]:

$$a_{ij} = \begin{cases} 1 & \text{if } \exists\ e_{ij} \in E \quad (2.2.2) \\ 0 & \text{if } \nexists\ e_{ij} \in E \quad (2.2.3) \end{cases}$$

Simply put, an entry in the adjacency matrix will be one if there is an edge present between the two corresponding nodes and 0 otherwise. The adjacency matrix of the graph in Figure 2.1a is shown in Figure 2.1b. This numerical representation of a graph is necessary to utilize computational algorithms on networks.

An extension to the definition of a graph is that of a weighted graph in which each edge is assigned a number, or weight [4]. This weight can be interpreted in many ways depending on what the weights are and how they were calculated. For example, the weights could represent a measure of similarity between the objects of interest (nodes) and thus quantify the similarity between pairs of objects in a system. Weighted graphs can also be represented in matrix form. The matrix associated with a weighted graph is called a weighted adjacency matrix. For each pair of nodes i and j, the entry a_{ij} in the weighted adjacency matrix A is the weight w_{ij} associated with the edge e_{ij} [5]. For an adjacency matrix A with associated with a graph G, each entry a_{ij} in A is defined as:

$$a_{ij} = \begin{cases} w_{ij} & \text{if } \exists\ e_{ij} \in E \quad (2.2.4) \\ 0 & \text{if } \nexists\ e_{ij} \in E \quad (2.2.5) \end{cases}$$



	A	B	C	D	E	F	G
A	0	1	0	0	0	0	1
B	1	0	0	0	0	1	1
C	0	0	0	1	1	1	0
D	0	0	1	0	1	0	0
E	0	0	1	1	0	0	0
F	0	1	1	0	0	0	1
G	1	1	0	0	0	1	0

Figure 2.1: **An Example Network.** (a) Visualization of the network drawn as nodes (circles) connected by edges (lines) (b) The corresponding unweighted adjacency matrix.

2.3 Similarity Metrics

2.3.1 Overview

Networks are often constructed to represent the similarities and relationships between objects within biological systems. Often, objects are represented as a vector of quantities. For example, when constructing gene co-expression networks, objects (genes) are represented by expression profiles (discussed further in Section 2.10). Networks are thus often constructed by performing an all-vs-all comparison of a set of objects of interest by calculating the similarity between all pairs of vectors representing the objects. In order to do this, similarity metrics are needed to provide a measure of similarity between two vectors. Various similarity metrics exist which all quantify different aspects of similarity.

2.3.2 Pearson Correlation Coefficient

Pearson's Correlation Coefficient was first introduced by Karl Pearson in 1895 [6] and is a very widely used correlation metric. Pearson's correlation coefficient r between two variables X and Y can be expressed as

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 \sum_I (Y_i - \bar{Y})^2}} \tag{2.3.1}$$

where $\bar{X}$ and $\bar{Y}$ are the means of variables X and Y respectively. Pearson's correlation coefficient takes on values between -1 and 1 and measures the linear association between two vectors [7]. Equation 2.3.1 can be expressed in an alternative form giving Pearson's correlation coefficient of vectors X and Y in terms of the covariance of the two vectors, scaled by their standard deviations (Equation 2.3.2).

$$r = \frac{\mathrm{Cov}(X,Y)}{S_X S_Y} \tag{2.3.2}$$

where $\mathrm{Cov}(X,Y)$ is the covariance of X and Y and S_X and S_Y are the standard deviations of X and Y respectively [7].

2.3.3 Spearman Correlation Coefficient

Spearman's Correlation Coefficient [8] r_s for variables X and Y has a formula similar to the Pearson Correlation Coefficient except that instead of using the actual values of the entries in the vectors, the ranks of the entries in the vectors are used. For vectors X and Y, let R_i denote the rank of value i in X, and let Q_i denote the rank of value i in Y. The Spearman Correlation Coefficient is then given by

$$r_s = \frac{\sum_i (R_i - \bar{R})(Q_i - \bar{Q})}{\sqrt{\sum_i (R_i - \bar{R})^2 \sum_i (Q_i - \bar{Q})^2}} \tag{2.3.3}$$

where $\bar{R}$ and $\bar{Q}$ are the means of rank variables R and Q respectively [9]. The Spearman Correlation Coefficient measures the monotonicity of two vectors, i.e. to what extent do the values in the vector increase as the values in the other vector increase. Unlike the Pearson Correlation Coefficient, it does not measure the extent of a linear relationship between the two vectors. [9].

2.3.4 Jaccard's Index

Jaccard's Index is a similarity index which was originally referred to as the "Coefficient of Community" [10]. It was developed to quantify the similarity between the plant species content of two areas. It is easily defined in terms of set intersects. Given two sets A and B, Jaccard's Index $J(A,B)$ is defined as [10]:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2.3.4}$$

Jaccard's Index can also be defined in terms of vectors. Let the two sets be two binary vectors, X and Y. Jaccard's Index $J(X,Y)$ can then be defined in terms of inner products as [11]:

$$J(X,Y) = \frac{\langle X,Y \rangle}{\langle X,X \rangle + \langle Y,Y \rangle - \langle X,Y \rangle} \tag{2.3.5}$$

Jaccard's Index takes on values in the range of 0 to 1.

In order to apply Jaccard's Index to non-binary vectors, a vector X of integers can easily be converted to a binary vector X_B as follows:

$$X_{Bi} = \begin{cases} 1 & \text{if } X_i \geq 1 \\ 0 & \text{if } X_i = 0 \end{cases} \tag{2.3.6}$$

2.3.5 Cosine

The Cosine similarity of two vectors X and Y simply involves taking the cosine of the angle between the two vectors (Equation 2.3.7),

$$\text{Cosine Similarity} = \cos(\Theta_{XY}) \tag{2.3.7}$$

where Θ_{XY} is the angle between vectors X and Y. This equation can also be written in inner-product form, in which the cosine of the angle between two vectors is expressed in terms of the inner product of the vectors, divided by their norms [12] (Equation 2.3.8).

$$\cos(\Theta_{XY}) = \frac{\langle X, Y \rangle}{||X||||Y||} \tag{2.3.8}$$

Cosine similarity takes on values between 0 and 1 [12], assuming that both vectors contain only positive values. This is the case with most biological data.

2.3.6 Sørensen Index

The Sørensen Index [13] (also known as the Dice Coefficient [14]) is a similarity index which was developed for ecological purposes and (similar to Jaccard's Index) is also based on set intersections. For two sets A and B the Sørensen Index $S(A, B)$ is defined as:

$$S(A, B) = \frac{|A \cap B|}{|A| + |B|} \tag{2.3.9}$$

Where $|A|$ is the number of elements in A and $|B|$ is the number of elements in B. The Sørensen Index can also be formulated in terms of vector algebra. For two binary vectors X and Y the Sørensen Index $S(X, Y)$ is defined as:

$$\begin{align} S(X, Y) &= \frac{2\langle X, Y \rangle}{\sum_i x_i + \sum_i y_i} \tag{2.3.10} \\ &= \frac{2\min(X, Y)}{\sum_i x_i + \sum_i y_i} \tag{2.3.11} \end{align}$$

where x_i is the i^{th} element of X and y_i is the i^{th} element of Y.

2.3.7 Czekanowski Index and Bray-Curtis Index

The Czekanowski Index is quantitative version of the Sørensen index. For vectors X and Y the Czekanowski Index is defined as [15]:

$$Cz = \frac{\sum_i 2\min(X_i, Y_i)}{\sum_i (X_i + Y_i)} \tag{2.3.12}$$

where X_i is the i^{th} element of X and Y_i is the i^{th} element of Y. The similarities between the forms of Equations 2.3.11 and 2.3.12 is easy to see, indicating the relationship between the Czekanowski Index and the Sørensen Index.

The Bray-Curtis [16] Index is often confused with the Czekanowski Index [15]. Although the Bray-Curtis Index has the same form as the Czekanowski Index (Equation 2.3.12) the underlying normalization assumptions are different. The Bray-Curtis Index assumes that all vectors are normalized by the total sum of each vector, i.e. the sum of all the entries in a vector is 1. Thus the Bray-Curtis Index $BC(X, Y)$ simplifies to [16; 15]:

$$\begin{aligned}
BC(X, Y) &= \frac{\sum_i 2\min(X_i, Y_i)}{\sum_i (X_i + Y_i)} && (2.3.13)\\
&= \frac{2\sum_i \min(X_i, Y_i)}{\sum_i X_i + \sum_i Y_i} && (2.3.14)\\
&= \frac{2\sum_i \min(X_i, Y_i)}{1 + 1} && (2.3.15)\\
&= \frac{2\sum_i \min(X_i, Y_i)}{2} && (2.3.16)\\
&= \min(X_i, Y_i) && (2.3.17)
\end{aligned}$$

2.3.8 Canberra Distance

The Canberra distance $Cb(X, Y)$ is a distance metric described as being the complement of Czekanowski's Index, and defined as [17]:

$$Cb(X, Y) = \frac{\sum_i |X_i - Y_i|}{\sum_i (X_i + Y_i)} \tag{2.3.18}$$

As mentioned, the Canberra distance is the complement of the Czekanowski Index [17]. This means that

$$Cb(X, Y) = 1 - Cz(X, Y) \tag{2.3.19}$$

or, equivalently that

$$Cz(X, Y) = 1 - Cb(X, Y) \tag{2.3.20}$$

This can be derived as follows:

$$1 - Cb(X,Y) = 1 - \frac{\sum_i |X_i - Y_i|}{\sum_i (X_i + Y_i)} \tag{2.3.21}$$

$$= \frac{\sum_i (X_i + Y_i) - \sum_i |X_i - Y_i|}{\sum_i (X_i + Y_i)} \tag{2.3.22}$$

$$= \frac{\sum_i ((X_i + Y_i) - |X_i - Y_i|)}{\sum_i (X_i + Y_i)} \tag{2.3.23}$$

Notice that Equation 2.3.23 has the same denominator as the Czekanowski Index in Equation 2.3.12. Thus, in order to show that $1-Cb(X,Y) = Cz(X,Y)$, we need to show that the numerators of Equation 2.3.23 and 2.3.12 are equal. To do this, consider the diagram in Figure 2.2. Assume that for a given i, $X_i > Y_i$. Then,

$$\sum_i ((X_i + Y_i) - |X_i - Y_i|) = 2Y_i. \tag{2.3.24}$$

Similarly, if for a given i, $Y_i > X_i$. Then,

$$\sum_i ((X_i + Y_i) - |X_i - Y_i|) = 2X_i. \tag{2.3.25}$$

Thus, combining the above two cases,

$$\sum_i ((X_i + Y_i) - |X_i - Y_i|) = 2\min(X_i, Y_i), \tag{2.3.26}$$

which is indeed the numerator of Equation 2.3.12. Thus, The Czekanowski Index $Cz(X,Y)$ is the complement of the Canberra distance $Cb(X,Y)$ related as $1 - Cb(X,Y) = Cz(X,Y)$.

2.3.9 Jaccardized Czekanowski Index

The Jaccardized Czekanowski Index [18] is a new similarity metric which attempts to formulate a quantitative version of Jaccard's Index in the same sense that the Czekanowski Index is a quantitative version of the Sørensen Index. The Jaccardized Czekanowski Index is derived as follows [18]: First, the Jaccard Index J is related to the Sørensen Index S by the following equation:

$$S = \frac{2J}{J+1} \tag{2.3.27}$$

Rearranging Equation 2.3.27 to make J the subject of the equation yields:

$$J = \frac{S}{2-S} \tag{2.3.28}$$

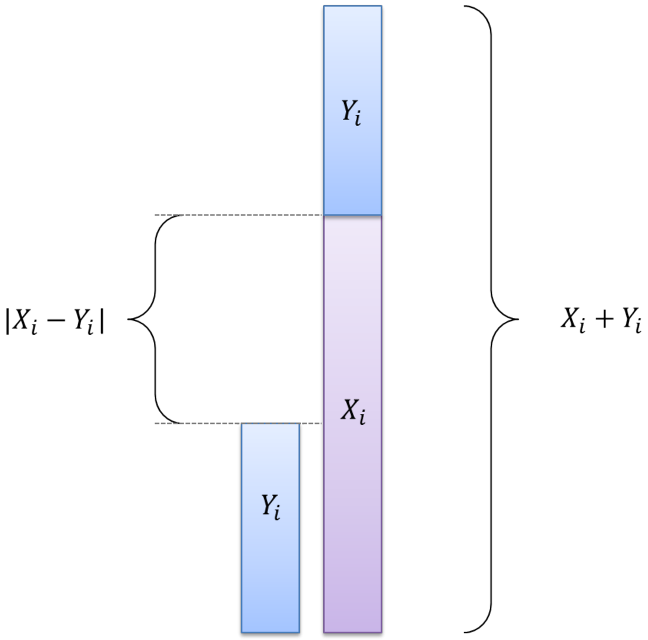


Figure 2.2: **Canberra Distance vs. Czekanowski Similarity** A visual aid in the relatedness of the Czekanowski similarity index and the Canberra distance.

Replacing the Sørensen Index S in Equation 2.3.28 with the Czekanowski Index Cz thus yields a quantitative version of Jaccard's Index called the Jaccardized Czekanowski Index:

$$JCz = \frac{Cz}{2 - Cz} \tag{2.3.29}$$

The Jaccardized Czekanowski Index was then found to not be novel, but is actually the same as the Ružička Index developed in 1958 [19].

2.3.10 Maximum Information Coefficient

The Maximum Information Coefficient (MIC) between two vectors X and Y is a similarity metric which, unlike the Pearson Correlation Coefficient, can detect non-linear correlations. The MIC is calculated as follows: Consider a set of ordered pairs (x_i, y_i) where x_i is the i^{th} value in X and y_i is the i^{th} value in Y. A partition is then created on the ordered pairs (x_i, y_i). This can be visualised as plotting a scatterplot of X vs Y, as drawing a grid $m \times n$ on this scatter plot, partitioning the points ((x_i, y_i) pairs) into blocks. Grids of different dimensions are drawn. Each grid results in a characteristic probability distribution of each variable, allowing the Mutual Information of the variables to be created. The Maximum Information Coefficient is the maximum Mutual

Information Coefficient obtained across all grids of all dimensions considered [20; 21].

2.4 Network Pruning Methods

When a network is constructed using a particular similarity metric to quantify similarity between objects, the result is an all-against-all complete network in which each node is connected to all other nodes by weighted edges. Network pruning methods are approaches for removing the lower-weighted or less significant edges in the network, thus ideally leaving behind only the significant, true relationships in the system and screening out noisy, low weighted relationships. Various pruning approaches will be outlined briefly below.

2.4.1 Hard and Soft Thresholding

One approach for network pruning is called thresholding. Zhang *et al.* proposed two types of thresholding, namely hard thresholding and soft thresholding [22]. Hard thresholding involves setting a minimum similarity cutoff and removing all edges with a weight lower than that cutoff. This reduces the number of edges in the network, and thus information can be lost [22]. Soft thresholding involves the use of a soft thresholding function, which increases the relative weight of highly weighted edges and decreases the relative weight of low-weighted edges. An example of a soft thresholding function f is [22]:

$$f(w_{ij}) = w_{ij}^{\beta} \tag{2.4.1}$$

This approach avoids loss of information, but does not help to reduce the number of edges which can be necessary when dealing with networks with a large number of nodes and edges.

2.4.2 Maximum Spanning Tree

Another method of network size reduction is to prune it to a backbone of maximum weight, namely a Maximum Spanning Tree (MST). MSTs can be calculated by first inverting the weights (similarity measures) of the edges, thus converting them into distance measures, and then applying a Minimum Spanning Tree (MiST) algorithm. A MiST is tree spanning all nodes of a given network which has minimum weight [23]. There are many algorithms for computing MiSTs. A common algorithm is Dijkstra's Algorithm, which constructs a MiST for a given network as follows [24]:

1. Select an arbitrary starting node a as the first node in the tree.
2. For each unvisited node x which is a neighbour to the tree, calculate the distance from a to x through the tree, with only one edge not present in the tree connecting x to the tree.

3. Add the node x with the shortest distance from a to the tree.

4. Repeat steps 2 and 3 until all nodes are added to the tree.

2.4.3 Disparity Filter

The Disparity filter is an alternative network pruning method aimed at extracting the network backbone of statistically significant edges i.e. those edges carrying a statistically significant proportion of the connectivity of a node [25]. The null model probability density function for weights of edges connected to a node of degree k is given by

$$p(x)dx = (k-1)(1-x)^{k-2}dx \tag{2.4.2}$$

For a given node i of degree k, each edge ij connecting i to a neighbour j has normalized weight p_{ij}. The probability α_{ij} of obtaining a weight larger than or equal to p_{ij} according to the null model is:

$$\alpha_{ij} = 1 - (k-1)\int_0^{p_{ij}} (1-x)^{k-2}dx \tag{2.4.3}$$

Edges for which α_{ij} is smaller than a chosen probability threshold are considered to carry a significant proportion of a node's weight and are included in the backbone [25].

2.5 Network Topology Measures

Once networks have been constructed for a certain set of objects of interest within a system using a particular similarity metric and have been pruned to select for most highly weighted edges, the networks will exhibit certain topologies. Network topology can be described quantitatively through a number of network properties or network measures [26]. These measures quantify local properties of individual nodes within a network as well as topological properties of the entire network as a whole.

2.5.1 Node-based Topology Measures

The following network measures are defined per node or per node pair for a given network, and include adjacency, connectivity, maximum adjacency ratio, topological overlap, TOM-connectivity, clustering coefficient, betweenness and efficiency, as defined below.

2.5.1.1 Adjacency

For two nodes i and j, the adjacency a_{ij} is the entry ij in the adjacency matrix of the network. In an unweighted network a_{ij} will be 1 if nodes i and j are connected by an edge and 0 otherwise. In a weighted network, a_{ij} will be equal to the strength of the connection (i.e. the edge weight) between nodes i and j [22].

2.5.1.2 Connectivity

The connectivity k_i for a node i is defined as [26]:

$$k_i = \sum_{j \neq i} a_{ij} \tag{2.5.1}$$

where a_{ij} is the adjacency of nodes i and j. It is an indication of how well connected a node is to the network. For an unweighted network the connectivity k_i of node i is the number of edges connected to node i, i.e. the degree of the node. For a weighted, the connectivity of node i it is the sum of the weights of the edges connected to node i.

2.5.1.3 Maximum Adjacency Ratio

The Maximum Adjacency Ratio (MAR_i) for a node i is an extension of the connectivity of a node and is defined as [26]:

$$\mathrm{MAR}_i = \frac{\sum_{j \neq i} (a_{ij})^2}{\sum_{j \neq i} a_{ij}} \tag{2.5.2}$$

MAR describes the extent to which a node has strong connections with its neighbours. Assuming that the network edges have weights between 0 and 1, the Maximum Adjacency Ratio obtains a maximum value of 1 when all the connections of a node have the maximum weight of 1 [26].

2.5.1.4 Topological Overlap

The Topological Overlap ω_{ij} between two nodes i and j quantifies how connected two nodes are by taking into consideration the direct connection between the nodes and indirect connection via neighbours of the nodes [27], and is defined as [22]:

$$\omega_{ij} = \frac{(\sum_u a_{iu} a_{uj}) + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}} \tag{2.5.3}$$

where a_{iu} and a_{uj} are adjacencies and k_i and k_j are the connectivities of nodes i and j respectively [22]. In an unweighted network, the term $\sum_u a_{iu} a_{uj}$ will equal the number of neighbours shared between nodes i and j. Consider two nodes i and j with $k_i < k_j$. For an unweighted network, the topological overlap

ω_{ij} will be equal to 1 if every neighbour of i is also a neighbour of j and if a_{ij} is equal to 1. Put simply, this means that for the topological overlap between two nodes i and j to be one, all neighbours of the node with smaller degree need to be neighbours of the node with larger degree, and the nodes i and j need to be directly connected. For the topological overlap to be zero, the nodes must not be connected and they must have no common neighbours [22].

2.5.1.5 TOM-based Connectivity

The TOM-connectivity of a node is based on the topological overlap between nodes and is defined as [22]:

$$k_i = \sum_{j \neq i} \omega_{ij} \tag{2.5.4}$$

where ω_{ij} is the topological overlap (Equation 2.5.3) between nodes i and j. A node will thus have a high TOM-Connectivity if it has a high topological overlap with its neighbours, i.e. a node is connected to and shares a lot of neighbours with its neighbours [22].

2.5.1.6 Clustering Coefficient

The Clustering Coefficient [28] for a node is a measure which indicates the local structure around the node, in particular how densely connected (cliquish) the node and its neighbours are [26]. For an unweighted network, the Clustering Coefficient C_i for a node i is defined as the number of edges present in the neighbourhood around node i over the total possible number of edges is that neighbourhood:

$$C_i = \frac{\sum_{l \neq i} \sum_{m \neq i,l} a_{il} a_{lm} a_{mi}}{k_i(k_i - 1)} \tag{2.5.5}$$

The Clustering Coefficient reaches its maximum value when each pair of a node's neighbours are connected to each other [26]. Zhang *et al.* (2005) extended the Clustering Coefficient to apply to weighted networks [22]:

$$C_i = \frac{\sum_{l \neq i} \sum_{m \neq i,l} a_{il} a_{lm} a_{mi}}{(\sum_{l \neq i} a_{il})^2 - \sum_{l \neq i} (a_{il})^2} \tag{2.5.6}$$

2.5.1.7 Betweenness

The Betweenness of a node i is the number of shortest paths between other pairs of nodes which run through node i [29]. This measure could indicate the importance of the node and how much it would affect the network should it be removed [29].

2.5.1.8 Efficiency

The Efficiency E_{ij} of a path between two nodes i and j is calculated as the inverse of the length of the shortest path between two nodes[30]:

$$E_{ij} = \frac{1}{d_{ij}} \tag{2.5.7}$$

where d_{ij} is the length of the shortest path between nodes i and j. The shorter the path between two nodes, the more efficient the path. If no path between nodes i and j exists in the graph, the distance d_{ij} between nodes i and j is defined to be $d_{ij} = \infty$ and thus the efficiency $E_{ij} = 0$ [30].

2.5.2 Global Network Topology Measures

The following network measures are global network measures which are calculated for a network as a whole and not on an individual node or node pair level, and include density, centralization, heterogeneity, path length and degree correlation.

2.5.2.1 Network Density

The Density D of a network is a quantification of how densely connected the network is. For an unweighted network, Network Density is defined as the fraction of the number of edges in the network divided by the total number of possible edges given the number of nodes [31]:

$$D = \frac{s}{n(n-1)} \tag{2.5.8}$$

where s is the number of edges in the network and n is the number of nodes in the network. Network density can easily be extended for weighted networks and can be calculated as the mean of all the off-diagonal entries in the adjacency matrix [32]:

$$\begin{align} D &= \frac{\sum_i k_i}{n(n-1)} \tag{2.5.9}\\ &= \frac{\sum_i \sum_{j\neq i} a_{ij}}{n(n-1)} \tag{2.5.10}\end{align}$$

where k_i is the connectivity of node i and a_{ij} is the entry ij in the adjacency matrix of the network.

2.5.2.2 Network Centralization

Network Centralization measures the extent to which there is a point in the network which is more central than all other points [33]. It obtains a maximum value of one when the network has a star topology (very centralized) and 0 if the connectivity of each node in the network is the same, for example a square [26]. The Centralization C of a network is defined as [32]:

$$C = \frac{n}{n-2}\left(\frac{k_{\text{max}}}{n-1} - D_N\right) \tag{2.5.11}$$

where n is the number of nodes in the network, k_{max} is the maximum connectivity of the network and D_N is the network density.

2.5.2.3 Network Heterogeneity

Network heterogeneity H quantifies how much the connectivity of the nodes in the network varies throughout the network in terms of the variance of the connectivities [31] and is defined as [32]:

$$H = \frac{\sqrt{\text{var}(k)}}{\text{mean}(k)} \tag{2.5.12}$$

where var(k) is the variance in the connectivity of the network and mean(k) is the mean connectivity of the network. A very heterogeneous network will have a large variation in the connectivities of the nodes whereas in a homogeneous network, connectivity will be evenly distributed throughout the network.

2.5.2.4 Path Length

The Path Length of a network is the average length of all shortest paths between pairs of vertices [28].

2.5.2.5 Degree Correlation

The Degree Correlation quantifies how correlated the degrees of neighbouring nodes are. Assortative networks arise if nodes of high degree are mostly connected to other nodes of high degree, whereas disassortative networks arise when nodes of high degree are mostly connected to nodes of low degree [29].

2.5.3 Measures Derived from Gene Co-expression Networks

Horvath and Dong (2008) used a gene co-expression network as a platform to develop network measures. The nodes of the network represented genes, and the edges represented co-expression of the genes across a number of microarray experiments. A selection of network significance measures were derived from these gene co-expression networks [26].

Gene Significance was defined as the correlation between the expression profile of a gene and some biological trait of interest. This measure could be used to identify genes potentially impacting a trait of interest. Network significance was then simply defined as the average gene significance. Hub Gene Significance was defined in order to quantify the relationship between gene (node) connectivity and gene significance and was defined as the gradient of the line obtained from linear regression of Gene Significance and Connectivity. Centroid Significance was defined as the significance of the centroid of the network. The centroid can be determined in a number of different ways, for example it can be defined as the node with the highest connectivity. Centroid Conformity was then defined per node as the weight of the edge between the node and the centroid [26].

2.6 Clustering Algorithms

Network modules can be defined as highly connected sub-graphs within a network [26]. There are a number of different methods for identifying network modules. Hierarchical clustering, K-means clustering, Markov clustering, Topological Overlap Clustering, Link Clustering, Graphlet-based clustering and Jaccard Clustering will be discussed below.

2.6.1 Hierarchical Clustering

Hierarchical clustering is a type of clustering which results in a tree like structure or dendrogram. Cutting the dendrogram at different levels then results in different clusterings of the set of data points [34]. There are two main classes of hierarchical clustering algorithms, namely agglomerative and divisive. Agglomerative clustering begins with each object in its own cluster and then iteratively merges closest clusters based on a distance measure until all points are in one cluster. Divisive clustering does the opposite, beginning with all data points in one cluster and iteratively dividing clusters in two until all points are in their own individual clusters. This approach is much more computationally intensive than agglomerative clustering and thus is much less used [34].

The general algorithm for agglomerative hierarchical clustering is as follows: All objects/data points begin in their own individual clusters. A distance matrix is then constructed specifying the distances between all pairs of clusters. The two clusters closest to each other (corresponding to the minimum entry in the distance matrix) are merged, and the distance matrix is updated to contain the distances between all pairs of the new clusters. This process of merging clusters and updating the distance matrix is repeated until all points are in one cluster [34].

There are various different hierarchical clustering methods which differ in the way they calculate the distance between clusters. These methods fall into two classes, namely linkage methods and geometric methods [34; 35]. When determining the distance between two clusters, linkage methods calculate the distances between all pairs of points within the two clusters in question, and chose the distance between those two clusters as either the minimum distance (in the case of single linkage clustering) or the maximum distance (in the case of complete linkage clustering) [34]. Geometric methods involve calculating the distance between the centroids of clusters [34].

2.6.2 K-means Clustering

K-means clustering is a clustering method developed by J. Macqueen in 1967 [36]. Unlike other clustering algorithms, it requires, as a parameter, the number of resulting clusters K, hence the name of the algorithm [36]. Consider a set of n objects to be clustered using k-means clustering, each object represented by a vector. The k-means algorithm begins by creating K initial centroids representing preliminary clusters. Each object to be clustered is then assigned to the cluster of its nearest centroid [37]. The value of each centroid $\bar{c}_k$ is then recalculated as the average of all objects in its cluster. This process of reassigning objects to their nearest centroids and recalculating centroids is then repeated until the cluster compositions no longer change [37].

2.6.3 Markov Clustering

The Markov Cluster Algorithm (MCL) is a graph based clustering algorithm which clusters the nodes of a graph into non-overlapping groups using a process called flow simulation [38; 39]. Clustering occurs through the execution of a series of matrix operations (namely expansion and inflation) performed on the adjacency matrix of the network. Random walks of increasing length are simulated though the network by the expansion operator. Walks of high probability are encouraged and walks of low probability are removed by the Inflation operator. This eventually results in groups of nodes connected by walks of high probability [39; 38]. This is illustrated by Figure 2.3 [38].
The MCL algorithm consists of the following steps:

1. The weighted adjacency matrix of the network to be clustered is normalized by column resulting in a stochastic matrix in which each entry ij represents the probability of travelling from node j to node i [39].

2. The expansion operator E where $E(A) = A \times A$ is applied to the matrix A.

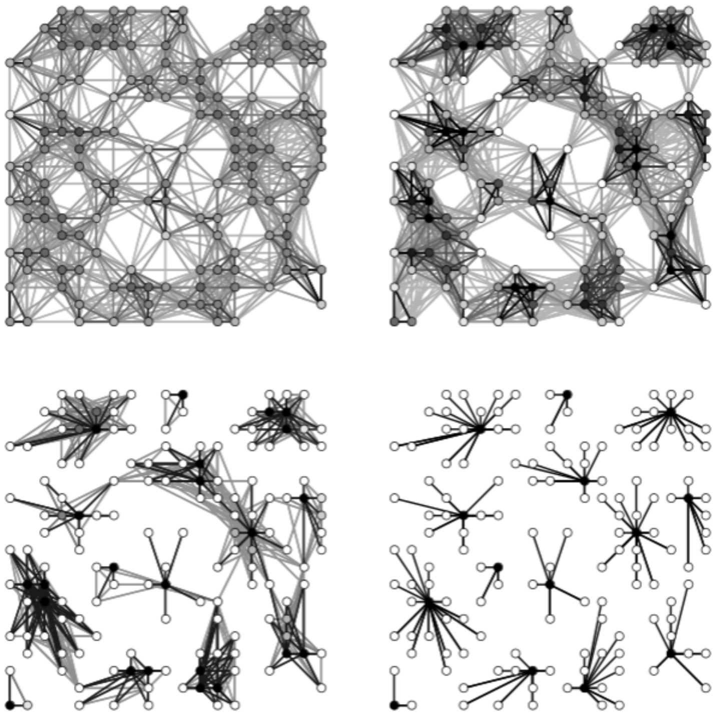


Figure 2.3: **Visualization of the MCL process.** Repeated rounds of expansion and inflation promote paths strong flow and remove paths of weak flow, resulting in clusters. (From [38].)

3. The inflation operator (r^{th} entry wise matrix power or Hadamard power) is applied to the resulting matrix, and the columns are renormalized making the matrix stochastic again.
4. Steps 2 and 3 are iteratively repeated until the matrix is *doubly idempotent*, i.e. further rounds of expansion and inflation have no effect on the matrix [38].

The parameter r (the power to which the entries in the matrix are raised in the Inflation operator) is called the inflation parameter and effects cluster granularity [38; 40]. A high inflation index will result in more, smaller clusters (high granularity) whereas a low inflation index will result in fewer, larger clusters (low granularity) [38; 40; 39].

2.6.4 Topological Overlap Clustering

Zhang and Horvath (2005) constructed modules in a gene co-expression network using what is called a Topological Overlap Matrix. Instead of defining modules as groups of genes (nodes) with highly correlated expression profiles, they defined modules as groups of genes (nodes) which have a large topological overlap. This was done by constructing the Topological Overlap Matrix (TOM) which contains the topological overlap values for each pair of nodes. Hierarchical clustering is then performed on this matrix where the distance between two nodes i and j is defined $1 - \omega_{ij}$, grouping genes into modules based on their topological proximity, not purely expression profile [22].

2.6.5 Link Clustering

Link clustering [41] is a network-based clustering algorithm which involves clustering the edges (links) of a graph into groups using hierarchical clustering. For two edges e_{ik} and e_{jk}, the similarity between those two edges $S(e_{ik}, e_{jk})$ is defined as the Jaccard overlap between the neighbourhoods of nodes i and j [41]:

$$S(e_{ik}, e_{jk}) = \frac{n(i) \cap n(j)}{n(i) \cup n(j)} \tag{2.6.1}$$

where $n(i)$ is the set of nodes containing node i and its neighbours. The edges are then clustered using hierarchical clustering. Each edge cluster gives rise to a node cluster containing the nodes connected by all the edges in the edge cluster. The node clusters resulting from this edge clustering approach can be overlapping, allowing a node to be present in more than one cluster [41].

2.6.6 Graphlet-Signature-Similarity-Based Clustering

Kuchaiev *et al.* (2011) developed the network analysis software package GraphCrunch 2, which clusters the nodes of a network into modules based on the local topology of the nodes, in particular, their graphlet signature. Graphlets (Figure 2.4 [42]) are unique, connected networks with a small number of nodes. The Graphlet Degree Vector or Signature of a node is a profile vector of a node's presence in graphlets in the network. Nodes are clustered into modules using the k-medioids algorithm (a variant of the k-means clustering algorithm) based on the similarity between their signature vectors [43].

2.6.7 Jaccard Clustering

Jaccard clustering involves the use of a modified Jaccard similarity metric to cluster a network into modules. For each pair of nodes i and j in a network, the modified Jaccard coefficient is calculated as

$$J_{ij} = \frac{C}{A + B - C} \tag{2.6.2}$$

where C is the number of nodes connected to both i and i, A is the number of nodes connected to node i and B is the number of nodes connected to node

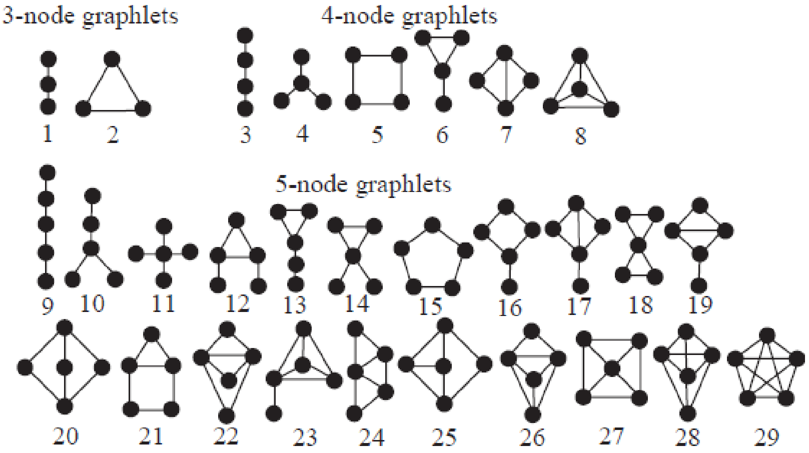


Figure 2.4: **Graphlets.** All possible 3, 4 and 5 node graphlets. (From [42].)

---

j. A threshold is set, and nodes connected by a Jaccard coefficient larger than this threshold are considered in the same cluster [44; 45].

2.7 Network Comparison and Network Overlap

2.7.1 Clustering Comparison

A clustering C is a partition of a set of objects consisting of non-overlapping sets of objects [46]. These sets are called clusters. There are many algorithms which generate clusterings on a data set which have been discussed above. There are various metrics which can be used to compare clusterings. Several metrics are based on counting pairs of elements and how often pairs of elements fall in the same cluster or in different clusters [47]. An example of a pair counting metric is the Jaccard index for clustering overlaps. The Jaccard overlap between two clusterings C_i and C_j is calculated as [47; 48]:

$$J(C_i, C_j) = \frac{N_{11}}{N_{11} + N_{01} + N_{10}} \tag{2.7.1}$$

where N_{11} is the number of pairs of elements (x, y) which are in the same cluster in C_i and C_j, N_{10} is the number of pairs of elements (x, y) which are in the same cluster in C_i but not C_j and N_{01} is the number of pairs of elements (x, y) which are in the same cluster in C_j but not C_i.

Other clustering overlap measures include those based on mutual information. These measures quantify the extent to which information about one clustering provides information about another clustering [48]. It is derived from the entropies of two clusterings as follows:

Let S denote the sample space of n objects. Let C_i and C_j denote clusterings of S. The Normalized Mutual Information between two clusterings C_i and C_j is defined as [47]:

$$NMI(C_i, C_j) = \frac{I(C_i, C_j)}{\sqrt{H(C_i)H_(C_j)}} \tag{2.7.2}$$

where $I(C_i, C_j)$ is the Mutual Information between clusterings C_i and C_j, $H(C_i)$ is the entropy of C_i and $H(C_j)$ is the entropy of C_j. The Mutual Information between two clusterings $I(C_i, C_j)$ and the entropies $H(C_i)$ and $H(C_j)$ and are defined as [48]:

$$I(C_i, C_j) = \sum_a \sum_b P(a, b) \log_2 \left(\frac{P(a, b)}{P(a)P(b)} \right) \tag{2.7.3}$$

$$H(C_i) = -\sum_a P(a) \log_2 (P(a)) \tag{2.7.4}$$

$$H(C_j) = -\sum_b P(b) \log_2 (P(b)) \tag{2.7.5}$$

where a is a cluster in clustering C_i and b is a cluster in clustering C_j. $P(a)$ is defined as $\frac{|a|}{n}$, $P(b)$ is defined as $\frac{|b|}{n}$ and $P(i,j)$ is defined as $\frac{|a \cap b|}{n}$.

Entropy (Equations 2.7.4 and 2.7.5) is a measure of the amount of uncertainty present in a clustering. This is best understood by the following thought experiment: Consider a clustering of n points, and consider picking an arbitrary point from any cluster. Assuming each point has an equal chance of being picked, the probability of the point being in cluster k of size n_k is $\frac{n_k}{n}$ [48]. If there is only one cluster in the clustering, then $\frac{n_k}{n} = 1$ causing the entropy (uncertainty) to be zero (Equations 2.7.4 and 2.7.5). Thus if there is only one cluster, there is no uncertainty/information present in the clustering. However, if the clustering contains more clusters with a more non-trivial probability distribution, the entropy (and information present in the clustering) increases. Mutual information is then derived from entropy, calculated as the information shared between two clusterings.

2.7.2 Network Profile Comparison

Another approach for comparing networks is implemented in a method called NetSimile [49]. This approach compares networks based on their topologies. For a set of networks to be compared, a selection of network topology measures are calculated for each network. These measures are compiled into a signature vector for each network. Network comparison then simply reduces to calculating the Canberra Distance between the network's signature topology vectors [49].

2.8 Orthology Detection

2.8.1 Overview

Orthology detection involves the detection of orthologs for a group of species [50]. The determination of these groups of equivalent genes called gene families is required before many phylogenomic analyses can be performed. Orthologs are defined as genes which arose through the process of speciation whereas paralogs are genes which arose through a process of gene duplication [51]. Together, orthologs and paralogs fall under the category of homologs - genes which are evolutionarily related and share a common ancestor [51]. Inparalogs specifically refer to paralogs which arose from a very recent gene duplication event and are thus still very similar. A variety of orthology detection methods exist to detect these evolutionary relationships between genes and generally fall into two classes, namely tree-based methods and graph-based methods [50]. The tree-based methods involve the construction of phylogenetic trees (e.g. LOFT [52], RIO [53], OrthologID [54]) or evolutionary distance matrices (e.g.

COCO-CL [55]) in order to detect orthologs and paralogs. Graph-based methods usually involve the quantification of sequence similarity between genes, construction of a network and subsequent clustering of the network using a clustering algorithm (e.g. OrthoMCL [56], TribeMCL [39], InParanoid [57]).

2.8.2 LOFT

LOFT (Levels of Orthology From Trees) is an orthology detection algorithm which discriminates between orthologous and paralogous relationships while attempting to retain some of the hierarchical nature of these relationships as they are detected from phylogenetic trees [52]. LOFT assigns numbers called "levels of orthology" to each gene, indicating hierarchical orthologous and paralogous relationships between then genes. In this case, LOFT is applied to already existing COGs (Clusters of Orthologous Groups) in order to illustrate the increased resolution provided by LOFT numbers.

The LOFT algorithm proceeds as follows [52]: A gene tree is constructed for an orthologous gene family (in this case COGs) by performing a multiple sequence alignment using Muscle [58] and then constructing a phylogeny using the neighbour joining method. The species overlap rule is used to label each node in the gene tree as either a speciation event or a gene duplication event: A node is declared to be a speciation event if the sets of species on each branch from that node are disjoint, i.e. they share no species. If the branches resulting from the node in question have any species in common, the node is declared to represent a gene duplication event. Levels of orthology numbers are then assigned to each all genes from the tree. Genes separated by only speciation events have the same LOFT numbers, whereas gene duplication events cause an extra level to be added to the LOFT numbers (Figure 2.5 [52]).

The LOFT numbers allow the hierarchical nature of orthologs detected from gene trees to be represented. LOFT also aids in the visualization of these orthologous relationships by colouring the different levels of orthology within a gene tree [52].

2.8.3 RIO

RIO (Resampled Inference of Orthologs) is an orthology detection method which detects orthologs of a given query sequence through the use of a phylogenetic tree [53]. RIO proceeds as follows: A query sequence of interest q is input and is assigned to a pfam family using HMMER. The query sequence is then aligned to the multiple sequence alignment of that family obtained from pfam. This multiple alignment is then bootstrap resampled a number of times (usually 100) and a phylogenetic tree constructed for each bootstrapped alignment. The resulting bootstrapped gene trees are then compared with a

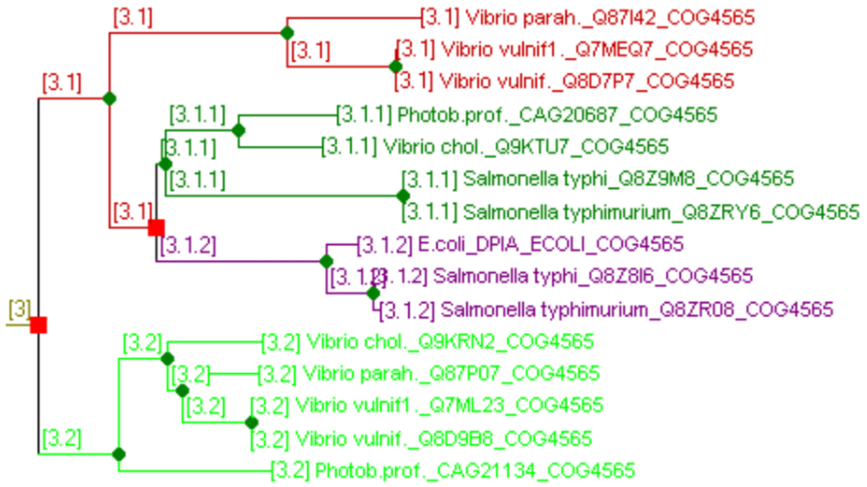

[3.1]
[3.1] Vibrio parah._Q87I42_COG4565
[3.1]
[3.1]
[3.1] Vibrio vulnif1._Q7MEQ7_COG4565
[3.1] Vibrio vulnif._Q8D7P7_COG4565
[3.1.1]
[3.1.1] Photob.prof._CAG20687_COG4565
[3.1.1]
[3.1.1] Vibrio chol._Q9KTU7_COG4565
[3.1.1]
[3.1.1] Salmonella typhi_Q8Z9M8_COG4565
[3.1]
[3.1.1] Salmonella typhimurium_Q8ZRY6_COG4565
[3.1.2]
[3.1.2] E.coli_DPIA_ECOLI_COG4565
[3.1.2]
[3.1.2] Salmonella typhi_Q8Z8I6_COG4565
[3]
[3.1.2] Salmonella typhimurium_Q8ZR08_COG4565
[3.2]
[3.2] Vibrio chol._Q9KRN2_COG4565
[3.2]
[3.2] Vibrio parah._Q87P07_COG4565
[3.2]
[3.2] Vibrio vulnif1._Q7ML23_COG4565
[3.2]
[3.2] Vibrio vulnif._Q8D9B8_COG4565
[3.2] Photob.prof._CAG21134_COG4565

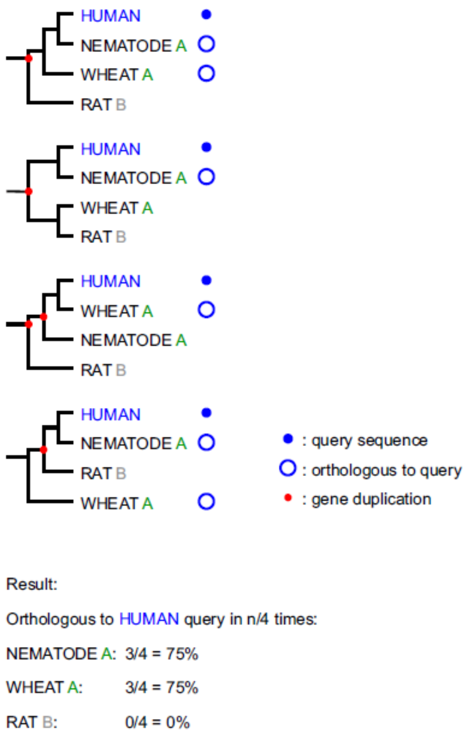
Figure 2.5: **Levels Of Orthology Numbers generated by LOFT.** Hierarchical LOFT numbers assigned to genes from a section of COG4565. Red square nodes represent gene duplication events and green diamond nodes represent speciation events are represented as green diamonds. (From [52].)

---

species tree to label each node as either a gene duplication event or a speciation event. Each gene g in the tree is then assigned a bootstrap score indicating the number of resampled trees in which g is orthologous to the query gene q (Figure 2.6) [53].

2.8.4 OrthologID

OrthologID [54] uses a combination of sequence similarity and tree-based approaches to detect orthologs from completely sequenced genomes. An all-vs-all BLAST is performed to calculate the sequence similarity between all pairs of genes from sequenced genomes, resulting in a network. Edges with an E-value greater than 1×10^{-20} are removed, as well as edges where the aligned region of the shorter sequence is less than 80% of the length of the longer sequence. Gene families are then considered as the connected components of this BLAST hit network [54]. Sequences within gene families are then aligned using MAFT and a maximum parsimony tree calculated using PAUP*. Orthologs are then identified and characteristic amino acids of each family identified using the CAST algorithm. These characteristic amino acids are highlighted in the OrthologID visual interface (Figure 2.7) [54].



Result:

Orthologous to HUMAN query in n/4 times:

NEMATODE A: 3/4 = 75%

WHEAT A: 3/4 = 75%

RAT B: 0/4 = 0%

Figure 2.6: **RIO Bootstrap Scores.** A simple example in which RIO is used to determine orthologs of a human gene using 4 bootstrap resamples. (From [53].)

2.8.5 COCO-CL

COCO-CL (Correlation Coefficient based Clustering) [55] is an orthology detection method developed to be run as a refinement step on already existing homologous gene families. It determines orthology based on the correlation between the evolutionary history of genes. Given the protein sequences of genes in a homologous gene family, COCO-CL proceeds as follows: A multiple sequence alignment is constructed using ClustalW and an evolutionary distance matrix D calculated from the resulting alignment. Each column V_i in D is thus a vector of evolutionary distances between protein i and all other proteins in the family. The Pearson Correlation Coefficient is then calculated between all pairs of vectors V_i and V_j resulting in a correlation matrix C in which each entry c_{ij} is the Pearson correlation coefficient between the evolutionary distance vectors V_i and V_j. Each correlation value c_{ij} is then replaced with the value $1 - c_{ij}$, converting them from similarity measures into distance measures, and the columns of the resulting correlation matrix (each column representing a gene) are clustered using single linkage hierarchical clustering, resulting in a tree [55]. The last step of the hierarchical clustering in which the

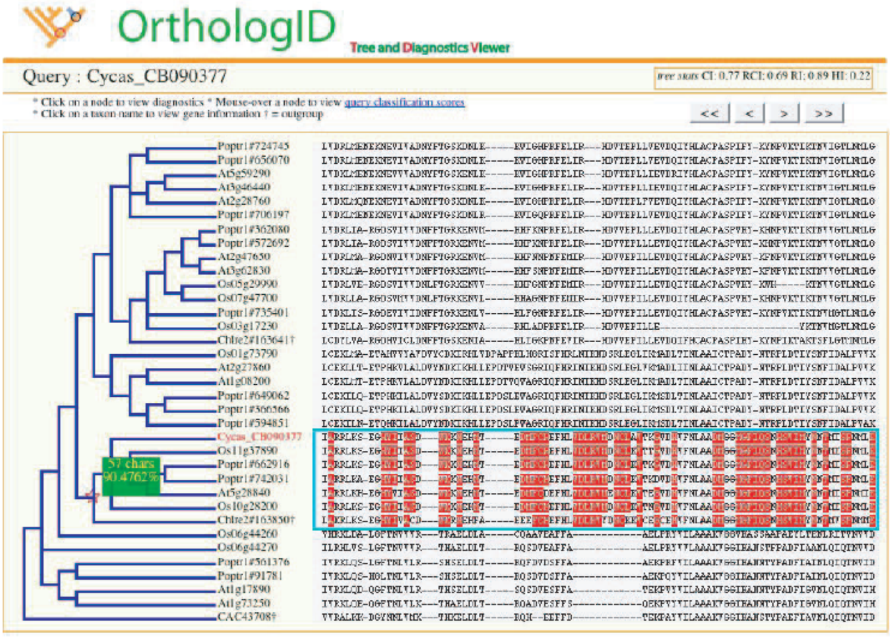


Figure 2.7: **OrthologID Interface.** The visual interface of OrthologID showing the charactieristic amino acids for a family highlighted in red. (Modified from [54].)

---

last two remaining clusters C_1 and C_2 are merged is used to label that node as either a speciation or gene duplication event, using a procedure very similar to the previously described species overlap rule. A score σ is calculated as:

$$\sigma = \frac{S}{\min(S_{C_1}, S_{C_2})} \tag{2.8.1}$$

where S_{C_1} is the number of species present in C_1, S_{C_2} is the number of species present in C_2 and S is the number of species present in both S_{C_1} and S_{C_2}. If $\sigma = 0$, the node is considered a speciation event. If $\sigma \approx 1$ then the node is considered a gene duplication event [55].

2.8.6 COGs

COGs (Clusters of Orthologous Groups) are groups of orthologous genes and paralogs which are orthologous to other genes within the cluster [59]. COGs are constructed for a set of genomes by comparing all pairs of proteins using BLAST. For each gene, the best BLAST hit in every other genome is determined. Triangles of best BLAST hits are then formed as the building blocks of the COGs (Figure 2.8A). COGs are then formed by merging triangles which share a common edge. COGs are available as a resource in the

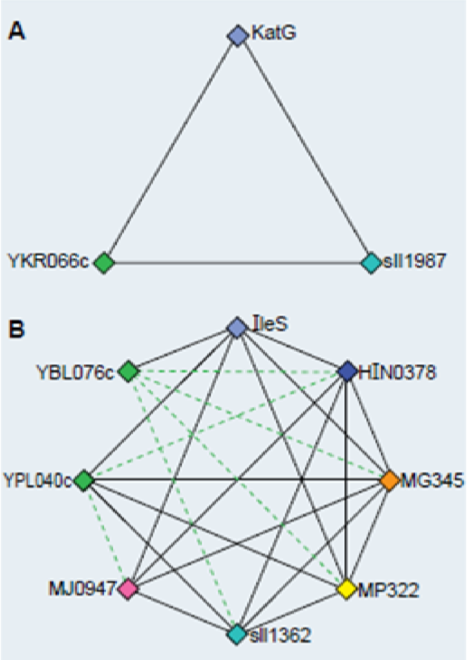


Figure 2.8: **Clusters of Orthologous Groups** (a) A Best Hit Triangle, the building block of COGs (b) COG constructed by merging triangles. (Modified from [59].)

COG database [60] which has been updated to include eukaryotic genomes [61], archaea [62; 63] and viruses [64].

2.8.7 TribeMCL

TribeMCL is protein clustering algorithm which constructs homologous protein families based on sequence similarity [39]. TribeMCL takes as input the sequences of a set of proteins to be clustered into gene families, which are then compared pairwise using BLAST. E-values of reciprocal BLAST matches are then averaged, the negative logarithm taken and a similarity matrix constructed in which each row and each column represents a protein and each entry ij is the sequence similarity S_{ij} between protein i and protein j calculated as:

$$S_{ij} = -\log_{10}(\text{E-value}_{av}) \tag{2.8.2}$$

where E-value$_{av}$ is the average reciprocal BLAST E-value between proteins i and j. Columns of this matrix are then normalized, thus turning it into a column stochastic matrix. This matrix can be viewed as a probability network in which each node represents a protein and edges represent similarities/transition probabilities between proteins. The Markov Cluster Algorithm

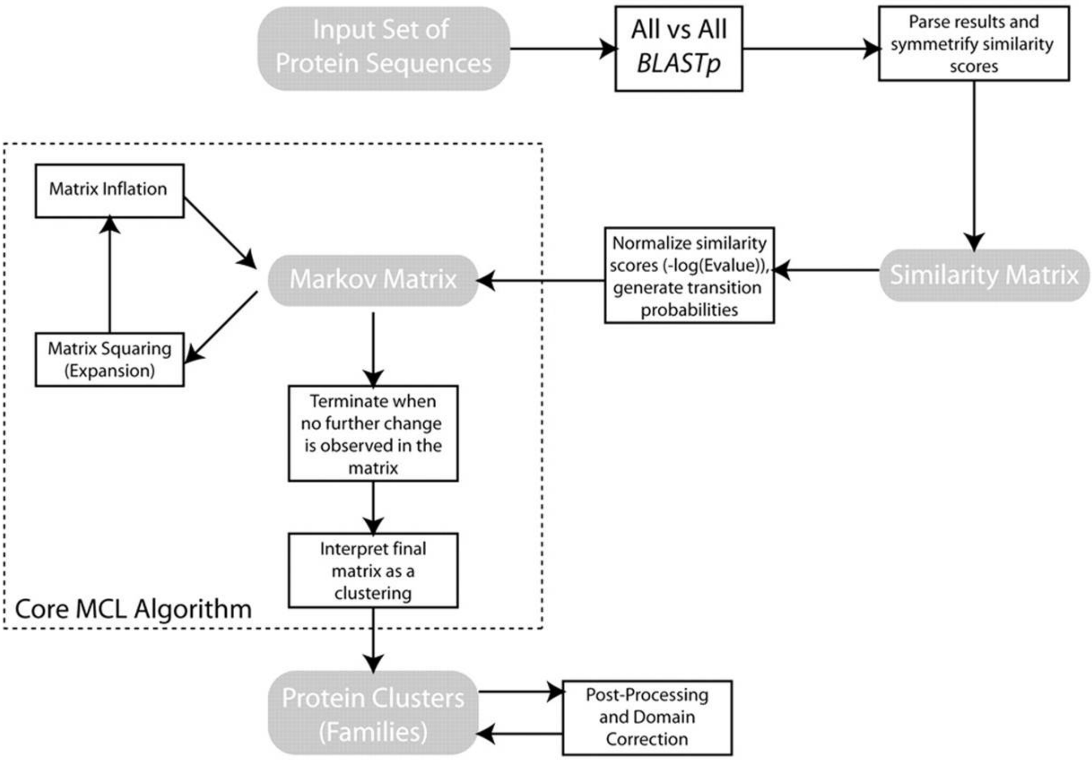

Input Set of Protein Sequences
All vs All BLASTp
Parse results and symmetrify similarity scores
Similarity Matrix
Normalize similarity scores (-log(Evalue)), generate transition probabilities
Markov Matrix
Matrix Inflation
Matrix Squaring (Expansion)
Terminate when no further change is observed in the matrix
Interpret final matrix as a clustering
Core MCL Algorithm
Protein Clusters (Families)
Post-Processing and Domain Correction

Figure 2.9: **TribeMCL** Flow diagram of the TribeMCL pipeline. (From [39].)

(MCL - see Section 2.6.3) [38] is then used to cluster this network into modules. Each resulting module is interpreted as a protein family. This process is summarized in Figure 2.9 [39].

2.8.8 InParanoid and MultiParanoid

InParanoid constructs orthologous protein families consisting of orthologs and inparalogs between the genomes of two species [65]. Sequence similarity between all pairs of proteins in two genomes is computed using BLAST, and the bit scores of reciprocal hits are averaged resulting in a similarity score for each pair of proteins. A bitscore cutoff of 50 and and a percentage overlap cutoff of 50% is applied. Orthologous families are then constructed as follows: For each protein, the best reciprocal BLAST hit is determined. These best reciprocal hits are considered seed orthologs for the clusters and inparalogs are added into these clusters if their similarity score to one of the seed orthologs is greater than or equal to the similarity score of the two seed orthologs. This is illustrated in Figure 2.10 [65]. Various rules are then applied to merge/delete/divide overlapping clusters (summarised in Figure 2.11 [65]) resulting in the final clusters.

Relative confidence scores ranging from 0% to 100% are assigned to inparalogs quantifying their relative similarity to the main ortholog where 100% is assigned to the main ortholog and 0% is assigned to the inparalog in the cluster that is furthest away from the main ortholog. This is illustrated in Figure 2.12 [65].

MultiParanoid was written as an extension of InParanoid to enable orthologous families to be constructed across multiple genomes as opposed to only two genomes [57]. Orthologous groups of proteins are constructed for all pairs of species using InParanoid, and clusters sharing seed orthologs are merged.

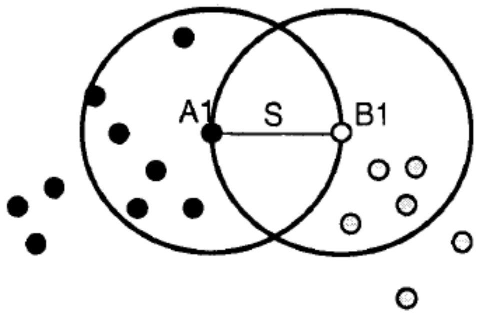


Figure 2.10: **InParanoid Clustering.** Seed orthologs (A1 and B1) form the centers of clusters and inparalogs clustered around orthologs if they have a similarity score greater than or equal to the similarity between the two seed orthologs. (From [65].)

2.8.9 OrthoMCL

OrthoMCL [56] is an orthology detection algorithm similar to TribeMCL in that it uses sequence similarity to construct a BLAST hit network and clusters this using MCL. It also however uses some of the best hit principles introduced in InParanoid to prune the network before clustering.

The OrthoMCL pipeline is summarised in Figure 2.13 [56]. The algorithm begins by taking as input the protein sequences for a set of completely sequenced genomes. An all-vs-all BLAST is performed to quantify the sequence similarity between all pairs of proteins. Potential orthologs and inparalogs are then identified from the BLAST E-values [56; 66]. Inparalogs are identified as intraspecies matches above the E-value and Percentage Match cutoffs who have the best match with each other when compared to matches with all other proteins

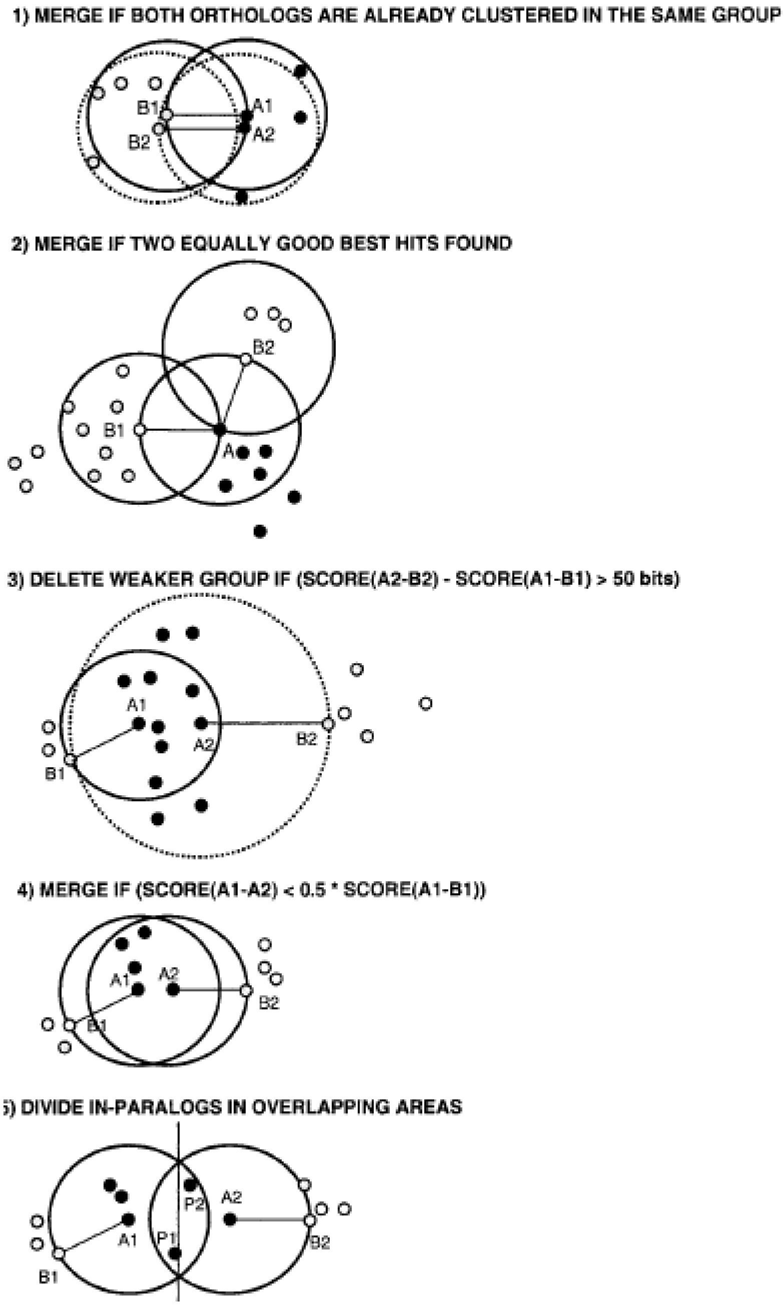

1) MERGE IF BOTH ORTHOLOGS ARE ALREADY CLUSTERED IN THE SAME GROUP
B1
B2
A1
A2
2) MERGE IF TWO EQUALLY GOOD BEST HITS FOUND
B2
B1
A
3) DELETE WEAKER GROUP IF (SCORE(A2-B2) - SCORE(A1-B1) > 50 bits)
A1
A2
B2
B1
4) MERGE IF (SCORE(A1-A2) < 0.5 * SCORE(A1-B1))
A1
A2
B2
B1
5) DIVIDE IN-PARALOGS IN OVERLAPPING AREAS
P2
A2
A1
P1
B2
B1

Figure 2.11: **InParanoid Cluster Merging.** Approaches for merging, deleting or dividing overlapping clusters around seed orthologs. (From [65].)

from all other species. Orthologs are identified as inter-species matches between two species which are above the E-value and percentage match cutoffs and have the best match with each other when compared with matches with proteins from the other species in question. Coorthologs are identified as genes

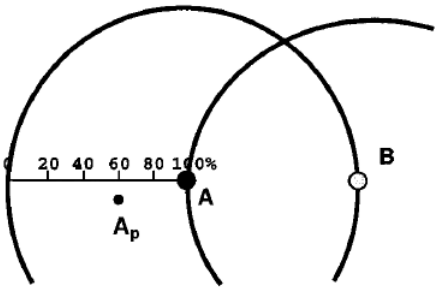


Figure 2.12: **InParanoid Confidence Scores.** Range of confidence scores assigned to inparalogs indicating their relative similarity to the main ortholog. (From [65].)

in different species which are linked by a composition of orthology and inparalogy as illustrated in Figure 2.14. Inparalog, ortholog and coortholog pairs are then assigned a weight calculated as follows:

$$\text{Weight} = -\log_{10}(E\text{-}value_{av}) \tag{2.8.3}$$

where $E\text{-}value_{av}$ is the average of the reciprocal E-values of the two proteins in question. Inparalog weights are normalized by the average weight for that particular species whereas ortholog and coortholog weights are normalized by the average weight for that particular species pair [56; 66]. The resulting pairs of orthologs, inparalogs and coorthologs are then represented as a network and clustered using MCL. The resulting clusters are interpreted as orthologous protein families.

2.8.10 Orthology Detection Method Comparison

A comparison study of various orthology detection methods was performed by Chen *et. al* in 2007 using latent class analysis [67]. Differences in the results of a selection of tree-based and graph-based orthology detection methods were analysed and the false negative (FN) and false positive (FP) rates of the methods were estimated. The resulting sensitivity and specificity of the methods is shown in Figure 2.15 [67]. In general, orthology detection methods seemed to exhibit a trade-off between sensitivity and specificity. Tree-based methods (for example, RIO) had a higher FN rate but a low FP rate, whereas homology/graph based methods (for example, KOG) showed the opposite - a lower FN rate and a higher FP rate. OrthoMCL and InParanoid showed the most optimal combination of sensitivity and specificity, being closest to the origin in the sensitivity-specificity plot (Figure 2.15) [67].

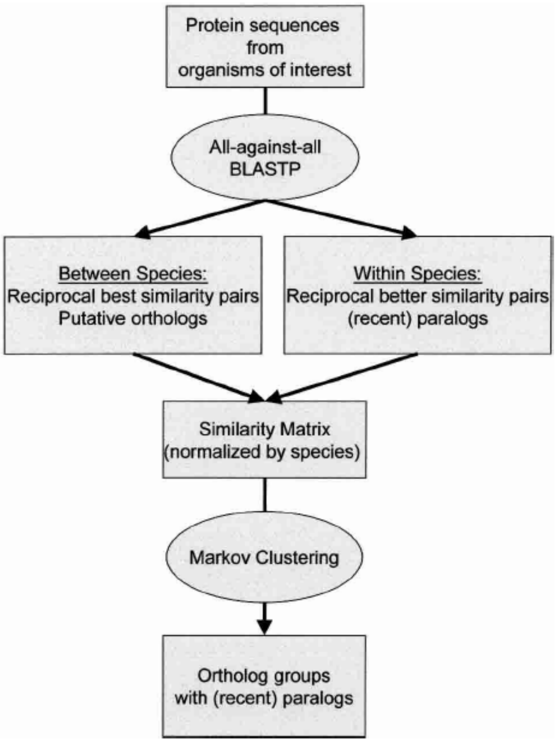

Protein sequences from organisms of interest
All-against-all BLASTP
Between Species: Reciprocal best similarity pairs Putative orthologs
Within Species: Reciprocal better similarity pairs (recent) paralogs
Similarity Matrix (normalized by species)
Markov Clustering
Ortholog groups with (recent) paralogs

Figure 2.13: **OrthoMCL Pipeline.** Flow diagram for the OrthoMCL pipeline. (From [56]).

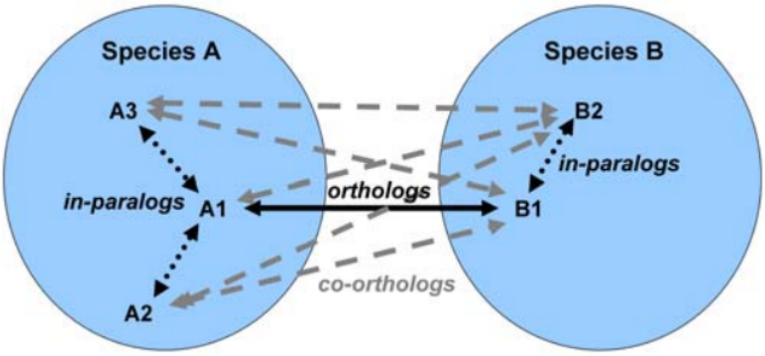
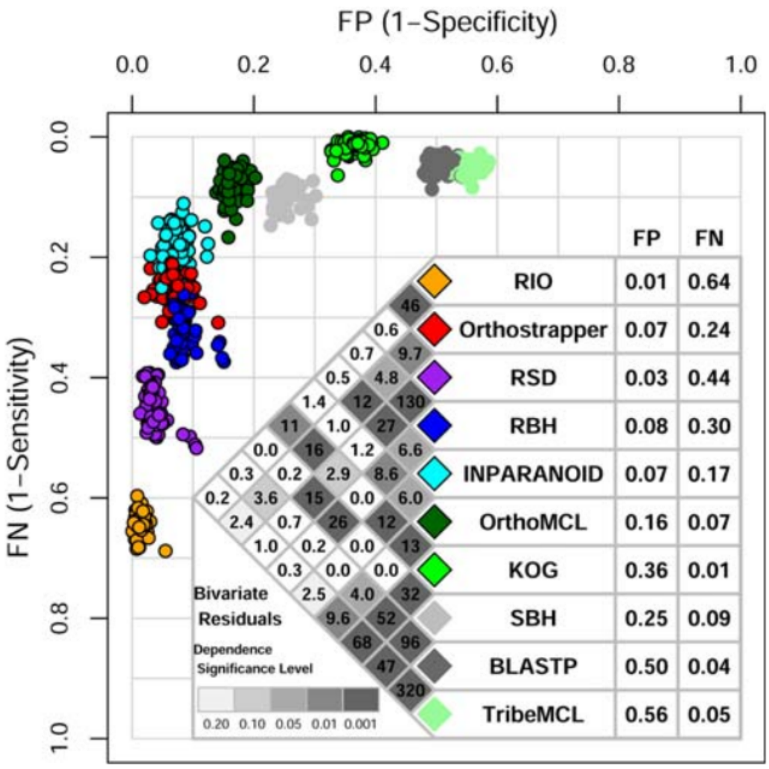


Figure 2.14: **Coorthologs.** Coorthologs refined as genes in two different species which are connected transitively through an orthologous relationship, indicated by solid black lines, and an inparalogous relationship,indicated by dotted black lines. (From [67].)



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

Figure 2.15: **Estimated Sensitivity and Specificity.** The false negative (FN) rate and false positive (FP) rate of different orthology detection methods estimated using latent class analysis. (From [67].)

2.9 Whole-Genome Phylogenomic Networks

2.9.1 Overview

Phylogenomics is a field which can be seen as a combination of genomics and the study of evolution [68]. It involves the investigation of the evolutionary relationships between organisms based on whole genome information as opposed to a single or a small subset of genes [69]. Examples of different types of phylogenomic networks including whole genome phylogenies, networks of gene sharing and co-evolution networks will be discussed below.

2.9.2 Whole-Genome Phylogenies and Gene Sharing Networks

Phylogenies are a specific type of network (directed acyclic graphs, or trees) which are used to represent the evolutionary relationships between species [70]. Traditionally, phylogenies were constructed using a single ortholog from each species. This results in a phylogeny which reflects the evolutionary history of a single gene within the species in question and not the evolutionary relationships between each species as a whole. Thus, using different genes results in different phylogenetic trees [71]. With the increased availability of fully sequenced genomes, phylogenies can be constructed based on whole genomes (for example, see [72]). Constructing phylogenies using genome-scale data has been found to result in more accurate species trees [71]. There are several ways in which whole-genome scale phylogenies have been constructed. Phylogenies can be constructed from whole genome data by determining a set of genes, concatenating the sequences into a “super-gene alignment" [73] and then constructing a species tree from the resulting concatenated alignment. Alternatively, a species tree can be constructed from each aligned set of orthologs, resulting in a species tree for each ortholog family, and then calculating a consensus tree [73].

Genome-scale phylogenies have also been constructed based on gene family content of the species in question. This is done by constructing gene families across the species, constructing a binary vector for each species consisting of 1’s and 0’s where a 1 represents the presence of that particular gene family in the species and a 0 represents the absence of the gene family in the species. These vectors are then compiled into a matrix of gene family content and a phylogeny is constructed from this matrix [74] [75].

Phylogenies can also be constructed based on shared gene content of the species in question. Similarity between pairs of genomes is defined as the fraction of shared orthologs between those two genomes and a standard phylogeny construction algorithm such as the Neighbour Joining method is then applied [76]. SHOT is an example of a webserver which constructs phylogenies based on shared gene content [77]. This method was applied to 50 complete genomes across the tree of life and reflected a topology very similar to that of the SSU rRNA phylogeny. However, the particular local topology around *C. elegans*, *Homo sapiens* and *Drosophila melanogaster* was more similar to the traditional phylogeny constructed from morphological characteristics and phylogenies constructed from protein data than it was to the SSU rRNA phylogeny [77].

The process of evolution does not always lend itself to a tree-like structure (vertical inheritance) because of phenomena such as hybridization, duplications and lateral gene transfer [69]. This more complicated view of evolution consisting of vertical and horizontal inheritance is better represented by a network [69].

Gene sharing networks are phylogenomic networks which model the similarity between species in terms of how many genes are shared between pairs of species [69]. Kloesges *et al.* used a network-based approach to investigate shared gene content of 329 proteobacteria [78]. Genomes for proteobacteria were obtained and genes were clustered into gene families based on sequence identity using TribeMCL. Similarity between species (nodes of the network) was then calculated as the number of gene families present in both species.

Halary *et al.* [79] constructed a gene-sharing network in which each node represented a genome (either cellular, plasmid and phage genome) and each edge represented shared gene content between the two genomes it connected (Figure 2.16 [79]). This network revealed that shared gene content is mostly separate between what they call different "DNA vehicles" (either a cellular chromosome (green nodes), plasmid (purple nodes) or phage (red nodes) genome). As illustrated in Figure 2.16 [79], it can be seen that most connections are between DNA vehicles of the same type.

2.9.3 Co-evolution Networks

Gene co-evolution networks have been used in the prediction of protein-protein interactions [80]. Interacting proteins, be they physically interacting or simply functionally related, are thought to co-evolve. This signature, called Evolutionary Rate Covariation (ERC), is identified as similarities between the phylogenetic trees of interacting proteins across species [81]. The original method to detect ERC is called the Mirror Tree Method [82]. In short, the Mirror Tree Method calculates ERC between two proteins as a value between 0 and 1. Given two proteins, orthologs of those proteins are determined using BLAST and multiple sequence alignments are constructed for each ortholog group. Distance matrices are calculated from these alignments, containing the

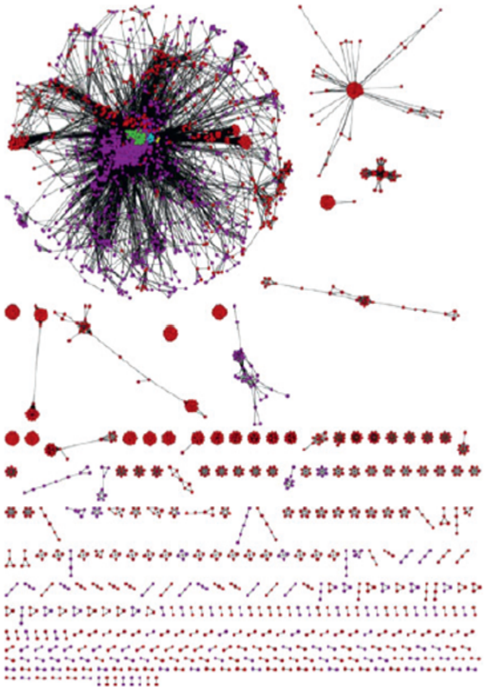


Figure 2.16: **Shared Gene Network.** Network of gene sharing for various organisms. Each node represents a genome. Green nodes represent cellular genomes, purple nodes represent plasmid genomes and red nodes represent phage genomes. Nodes are connected based on shared gene content. (From [79].)

---

evolutionary distances between all pairs of proteins within each alignment. Co-evolution of the two protein sequences is then determined as the Pearson Correlation of the two distance matrices [82]. This method gives rise to a number of false positives. Thus, a number of adapted Mirror Tree methods have been developed. One such adaptation, called ContextMirror, looks at each pair of potentially co-evolving proteins in the context of the complete co-evolutionary network of all proteins under consideration, instead of just the co-evolution signature between two individual proteins (Figure 2.17) [80]. The method begins in the same fashion as the original Mirror Tree Method, as for each protein in a species, orthologous proteins in other species are identified. A multiple sequence alignment is then constructed for each protein family and an evolutionary distance matrix is constructed from each multiple alignment. The similarity between the phylogenetic trees of two proteins is then calculated as the Pearson Correlation Coefficient of their two distance matrices. In the original Mirror Tree Method, these correlation values would be taken as the ERC value. ContextMirror however includes an extra step in which

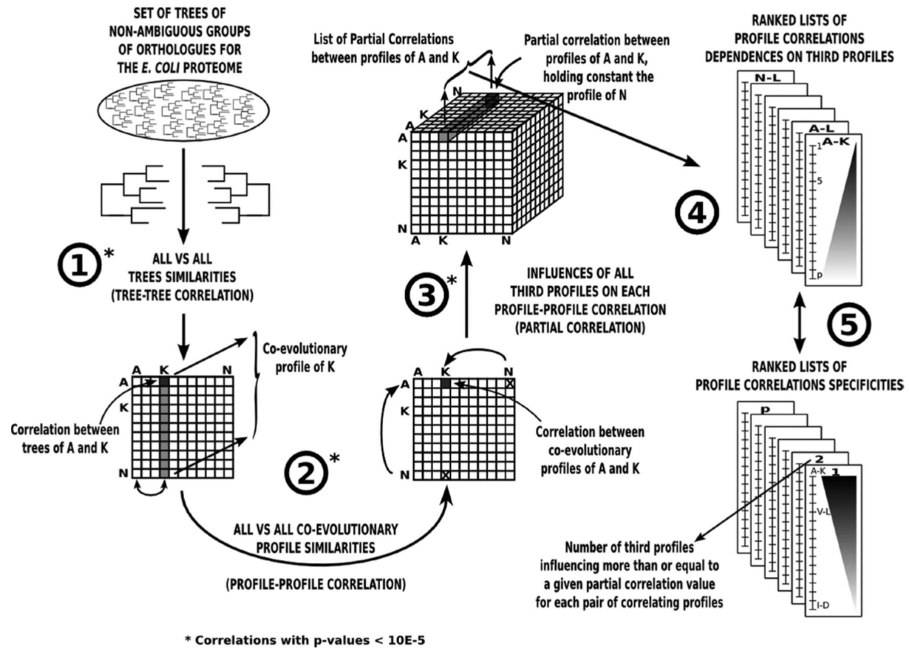


Figure 2.17: **ContextMirror Method.** Flow diagram for the ContextMirror Method. (From [80].)

a co-evolutionary network is constructed where the nodes represent proteins and the edges represent the correlation between their evolutionary distance matrices. This co-evolutionary network can be represented as a square matrix in which each row/column represents a protein and each entry represents the correlation between the distance matrices of the proteins in the respective row/column. A second round of correlation is then performed, calculating the Pearson Correlation Coefficient between all pairs of rows of this matrix. This calculates the co-evolutionary signal between two proteins by looking at the similarity between their co-evolutionary contexts within the whole co-evolutionary network, and not simply the similarity between their individual phylogenetic trees [80]. This method resulted in a higher accuracy in the prediction of protein-protein interactions than the original Mirror Tree Method [80].

2.10 Transcriptomic Co-expression Networks

2.10.1 Overview

Grouping genes together based on similarity between their expression profiles has been found to result in groups of genes with similar function [83]. The underlying assumption is that genes which are co-expressed are potentially co-regulated and thus may be involved in related functions. Networks have become widely used in gene co-expression analysis and a selection of examples of the use of network methods in gene co-expression analysis is discussed in this section.

2.10.2 Weighted Gene Co-expression Network Analysis

Zhang and Horvath (2005) introduced a pipeline for network-based analysis of gene co-expression called Weighted Gene Co-expression Network Analysis (WGCNA) [22]. Gene expression data is often structured as a matrix in which each row i represents a gene and each column j represents an experimental condition and each entry ij represents the expression level of gene i under condition j. Each row thus represents the expression profile of a different gene. The first step in constructing a gene co-expression network is to choose a definition of expression profile similarity and select a similarity metric to quantify that definition of similarity [22]. A common similarity metric used for gene co-expression analysis is the Pearson Correlation Coefficient. This similarity metric is then used to calculate the expression profile similarity between all pairs of genes. A thresholding strategy must then be chosen, which can involve hard thresholding or soft thresholding. The result is a square matrix in which rows and columns represent genes and each entry ij represents the thresholded similarity between the expression profiles of gene i and gene j. This matrix can be viewed as a gene co-expression network in which nodes represent genes and edges represent expression profile similarity between the genes.

The co-expression network is then clustered into groups of nodes exhibiting topological similarity within the network. This is done by constructing the Topological Overlap Matrix [22] in which each entry ij is the topological overlap ω_{ij} between genes i and j in the co-expression network. After converting each entry in the topological overlap matrix from a similarity measure to a distance measure ($d_{ij} = 1 - \omega_{ij}$) the resulting topological overlap network is clustered into modules of co-expressed genes using hierarchical clustering [22].

2.10.3 Cross-Species Co-expression

Gene co-expression analysis can be combined with phylogenomic information to compare gene expression modules across species [84]. This involves constructing a gene co-expression network for each species of interest. Clustering of the resulting network will result in gene expression modules. Gene family information can then be used to link genes in different species through homologous or orthologous relationships, thus providing a link between co-expression modules in different species (Figure 2.18) [84]).

One such approach, which uses sequence-based homology information to al-

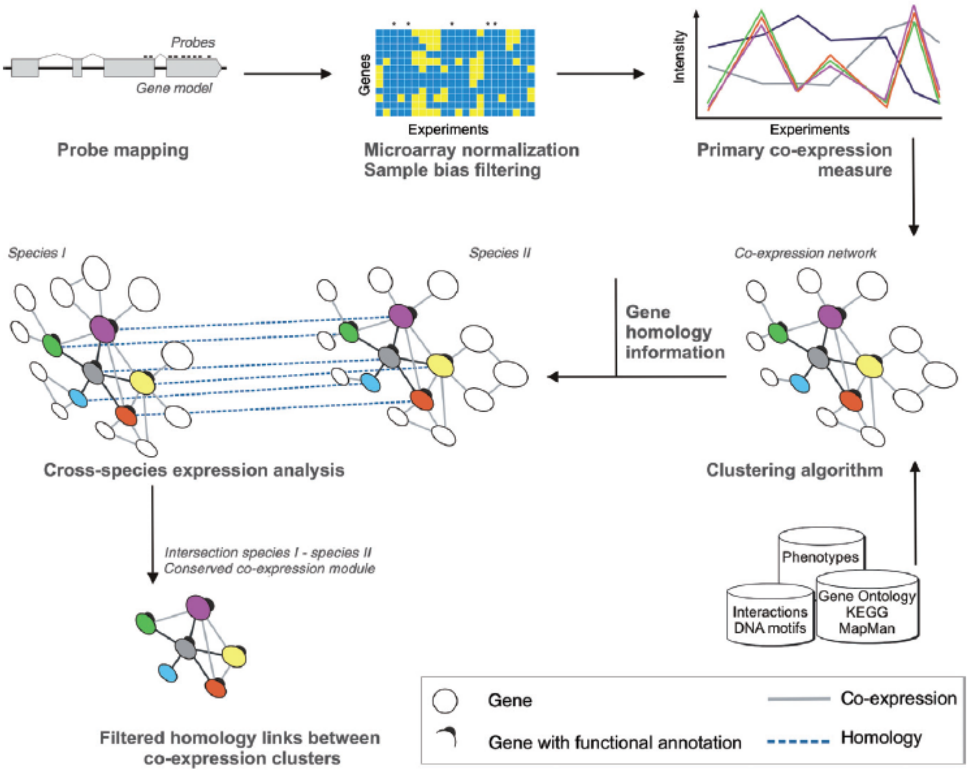


Figure 2.18: **Co-expression Module Comparison.** General pipeline for the determination of co-expression modules conserved across species. (From [84].)

low co-expression module comparison across 7 plant species, is PlaNet [85]. For each plant species, a gene co-expression network was constructed where the similarity measure used to quantify similarity between expression profiles was the Highest Reciprocal Rank, determined from Pearson Correlation Coefficients [85]. The co-expression network was then clustered into co-expression modules. Co-expression modules were compared across species using the NetworkComparer pipeline [85]. This involves linking the expression modules through genes in the same gene family (Pfam). Given a particular gene of interest in a particular species, the gene family from Pfam containing that gene is determined. The Node Vacinity Network (NVNs) is then determined for each gene in the Pfam family. NVNs are then compared by how similar their Pfam content is. Similar NVNs are then used to identify similar co-expression modules across species [85].

2.10.4 Joint Clustering of Co-expression Networks

Various approaches exist which use additional data besides gene expression profile similarity to cluster genes into functional modules. Ulitsky and Shamir

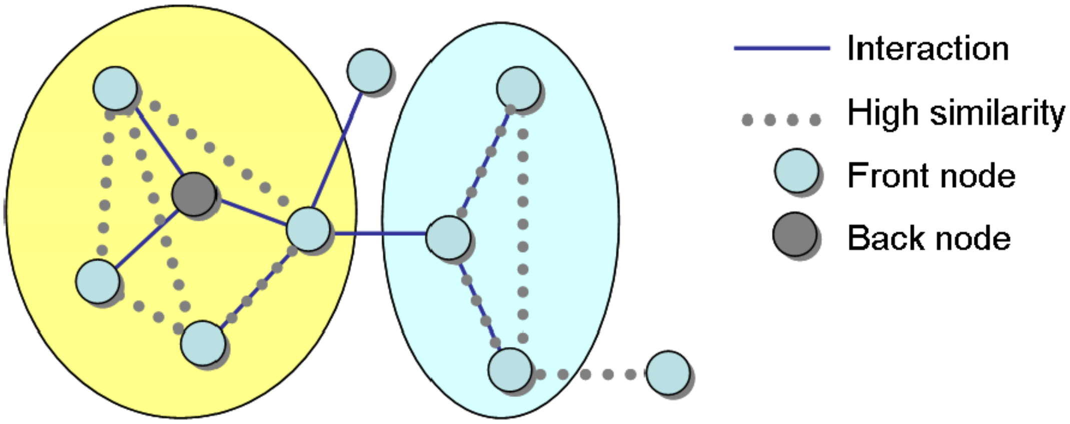


Figure 2.19: **MATISSE Gene Modules.** Nodes are clustered into modules based on topological connectedness in the interaction network indicated by solid black lines and expression profile similarity indicated by dotted grey lines. (From [86].)

(2007) described an approach in which they identified network modules by using a combination of topological measures and high throughput data. Their method, MATISSE (Module Analysis via Topology of Interactions and Similarity SEts) makes use of an interaction network (for example a gene interaction network) and its topology as well as gene co-expression similarity information and uses a statistical approach to identify modules as connected subnetworks of the interaction network in which the genes also have a high similarity. This ensures that the genes within a module will have similar regulation, and also be topologically connected in terms of interaction with other genes (Figure 2.19) [86].

MATISSE was tested on the combined protein-protein and protein-DNA interaction network of the yeast *S. cerevisiae*. Expression profiles of genes under different environmental conditions related to osmotic stress were also used. Similarity scores between genes were calculated as the Pearson Correlation between the expression profiles of the genes. MATISSE, as well as clustering by expression similarity was used to identify modules within the interaction network. Functional annotation enrichment was then performed, and modules constructed using MATISSE had higher enrichment than modules constructed purely from expression data [86].

Another approach for the clustering of gene expression networks using extant biological information is called Co-clustering [87]. This approach involved the development of a joint distance metric which considered the distance between two genes in a metabolic network and the distance between two genes based on their expression profile similarity. The distance $d_m(i, j)$ between two genes in a metabolic pathway was defined as the shortest path between the nodes in the KEGG metabolic network. The distance $d_e(i, j)$ between two genes i and j based on expression profile similarity was defined was $d_e = 1 - p(i, j)$ where $p(i, j)$ is the Pearson Correlation Coefficient between the expression profiles of gene i and gene j. The combined distance $\Delta(i, j)$ function was then defined as [87]:

$$\Delta(i,j) = 1 - 0.5 \times \left(\frac{1}{1 + e^{-a(d_m(i,j)-b)}} + \frac{1}{1 + e^{-c(d_e(i,j)-d)}} \right) \tag{2.10.1}$$

Hierarchical clustering was then used to cluster genes into functional modules based on the above joint distance metric [87].

2.11 Conclusions

It is evident that networks are very versatile structures. Nodes can represent any object of interest, and the edges can model many different types of relationships between objects of interest. These relationships can be quantified in many different ways through the use of different similarity metrics. Network structure and topology can also be described in a quantitative manner, facilitating the comparison of networks. Once networks have been created, there are a variety of pruning and clustering approaches to extract information from the networks, be those groups of similar nodes (resulting from clustering) or only the most highly weighted edges in the system (resulting from pruning). These network structures have been successfully used to represent complex biological systems. They have proven to be useful tools for the fields of phylogenomics and transcriptomics in providing representation, analysis methods and visualization strategies for these complex systems.

Bibliography

[1] Barabási, A.-L. and Albert, R.: Emergence of scaling in random networks. *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

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

[3] Gross, J.L. and Yellen, J.: *Handbook of graph theory.* CRC press, 2003.

[4] Clark, J. and Holton, D.: *A first look at graph theory.* World Scientific Pub Co Inc, 1991.

[5] Wallis, W.: *A beginner's guide to graph theory.* Birkhauser, 2007.

[6] Pearson, K.: Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, vol. 58, no. 347-352, pp. 240–242, 1895.

[7] Rodgers, J.L. and Nicewander, W.A.: Thirteen ways to look at the correlation coefficient. *The American Statistician*, vol. 42, no. 1, pp. 59–66, 1988.

[8] Spearman, C.: The proof and measurement of association between two things. *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.

[9] Pinto da Costa, J. and Soares, C.: A weighted rank measure of correlation. *Australian & New Zealand Journal of Statistics*, vol. 47, no. 4, pp. 515–529, 2005.

[10] Jaccard, P.: The distribution of the flora in the alpine zone. 1. *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[11] Lipkus, A.H.: A proof of the triangle inequality for the tanimoto distance. *Journal of Mathematical Chemistry*, vol. 26, no. 1-3, pp. 263–265, 1999.

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

[13] Sørensen, T.: A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biologiske skrifter*, vol. 5, pp. 1–34, 1948.

[14] Dice, L.R.: Measures of the amount of ecologic association between species. *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[15] Yoshioka, P.M.: Misidentification of the bray-curtis similarity index. *Marine Ecology Progress Series*, vol. 368, pp. 309–310, 2008.

[16] Bray, J.R. and Curtis, J.T.: An ordination of the upland forest communities of southern wisconsin. *Ecological Monographs*, vol. 27, no. 4, pp. 325–349, 1957.

[17] Lance, G. and Williams, W.: Computer programs for hierarchical polythetic classification ("similarity analyses"). *The Computer Journal*, vol. 9, no. 1, pp. 60–64, 1966.

[18] Schubert, A.: Measuring the similarity between the reference and citation distributions of journals. *Scientometrics*, vol. 96, no. 1, pp. 305–313, 2013.

[19] Schubert, A. and Telcs, A.: A note on the jaccardized czekanowski similarity index. *Scientometrics*, vol. 98, no. 2, pp. 1397–1399, 2014.

[20] 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.: Detecting novel associations in large data sets. *Science (New York, N.Y.)*, vol. 334, no. 6062, pp. 1518–24, 2011.

[21] 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.: Detecting novel associations in large data sets - Supplementary Material. `http://www.sciencemag.org/content/334/6062/1518/suppl/DC1`. Accessed February 2013.

[22] Zhang, B., Horvath, S. *et al.*: A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, pp. 5144–6115, 2005.

[23] Graham, R.L. and Hell, P.: On the history of the minimum spanning tree problem. *Annals of the History of Computing*, vol. 7, no. 1, pp. 43–57, 1985.

[24] Dijkstra, E.W.: A note on two problems in connexion with graphs. *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[25] Serrano, M.Á., Boguñá, M. and Vespignani, A.: Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences*, vol. 106, no. 16, pp. 6483–6488, 2009.

[26] Horvath, S. and Dong, J.: Geometric interpretation of gene coexpression network analysis. *PLoS Computational Biology*, vol. 4, no. 8, p. e1000117, 2008.

[27] Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. and Barabási, A.-L.: Hierarchical organization of modularity in metabolic networks. *Science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

[28] Watts, D.J. and Strogatz, S.H.: Collective dynamics of 'small-world' networks. *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[29] Reijneveld, J.C., Ponten, S.C., Berendse, H.W. and Stam, C.J.: The application of graph theoretical analysis to complex networks in the brain. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, vol. 118, no. 11, pp. 2317–31, 2007.

[30] Latora, V. and Marchiori, M.: Efficient behavior of small-world networks. *Physical Review Letters*, vol. 87, no. 19, p. 198701, 2001.

[31] Snijders, T.A.: The degree variance: An index of graph heterogeneity. *Social Networks*, vol. 3, no. 3, pp. 163–174, 1981.

[32] Dong, J. and Horvath, S.: Understanding network concepts in modules. *BMC Systems Biology*, vol. 1, p. 24, 2007.

[33] Freeman, L.C.: Centrality in social networks conceptual clarification. *Social Networks*, vol. 1, no. 3, pp. 215–239, 1979.

[34] Xu, R., Wunsch, D. *et al.*: Survey of clustering algorithms. *Neural Networks, IEEE Transactions On*, vol. 16, no. 3, pp. 645–678, 2005.

[35] Murtagh, F. and Contreras, P.: Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 86–97, 2012.

[36] MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, p. 14. California, USA, 1967.

[37] Steinley, D.: K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, vol. 59, no. 1, pp. 1–34, 2006.

[38] van Dongen, S.: *Graph clustering by flow simulation*. Ph.D. thesis, University of Utrecht, 2000.

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

[40] Van Dongen, S.: Graph clustering via a discrete uncoupling process. *SIAM Journal On Matrix Analysis and Applications*, vol. 30, no. 1, pp. 121–141, 2008.

[41] Ahn, Y.-Y., Bagrow, J.P. and Lehmann, S.: Link communities reveal multiscale complexity in networks. *Nature*, vol. 466, no. 7307, pp. 761–4, 2010.

[42] Pržulj, N., Corneil, D.G. and Jurisica, I.: Modeling interactome: scale-free or geometric? *Bioinformatics*, vol. 20, no. 18, pp. 3508–3515, 2004.

[43] Kuchaiev, O., Stevanović, A., Hayes, W. and Pržulj, N.: Graphcrunch 2: Software tool for network modeling, alignment and clustering. *BMC Bioinformatics*, vol. 12, no. 1, p. 24, 2011.

[44] Riley, D.R., Angiuoli, S.V., Crabtree, J., Hotopp, J.C.D. and Tettelin, H.: Using sybil for interactive comparative genomics of microbes on the web. *Bioinformatics*, vol. 28, no. 2, pp. 160–166, 2012.

[45] Sybil: Documentation. `http://sybil.sourceforge.net/documentation.html`, 2012. Accessed March 15, 2013.

[46] Meilă, M.: Comparing clusterings: an axiomatic view. In: *Proceedings of the 22nd International Conference on Machine learning*, pp. 577–584. ACM, 2005.

[47] Wagner, S. and Wagner, D.: *Comparing clusterings: an overview.* Universität Karlsruhe, Fakultät für Informatik, 2007.

[48] Meilă, M.: Comparing clusterings-an information based distance. *Journal of Multivariate Analysis*, vol. 98, no. 5, pp. 873–895, 2007.

[49] Berlingerio, M., Koutra, D., Eliassi-Rad, T. and Faloutsos, C.: A scalable approach to size-independent network similarity. Available: http://arxiv.org/pdf/1209.2684.pdf.

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

[51] Koonin, E.: Orthologs, paralogs, and evolutionary genomics 1. *Annual Review Genetics*, vol. 39, pp. 309–338, 2005.

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

[53] Zmasek, C. and Eddy, S.: Rio: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, vol. 3, no. 1, p. 14, 2002.

[54] Chiu, J., Lee, E., Egan, M., Sarkar, I., Coruzzi, G. and DeSalle, R.: Orthologid: automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics*, vol. 22, no. 6, pp. 699–707, 2006.

[55] Jothi, R., Zotenko, E., Tasneem, A. and Przytycka, T.: Coco-cl: hierarchical clustering of homology relations based on evolutionary correlations. *Bioinformatics*, vol. 22, no. 7, pp. 779–788, 2006.

[56] Li, L., Stoeckert, C. and Roos, D.: Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome Research*, vol. 13, no. 9, pp. 2178–2189, 2003.

[57] Alexeyenko, A., Tamas, I., Liu, G. and Sonnhammer, E.: Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*, vol. 22, no. 14, pp. e9–e15, 2006.

[58] Edgar, R.: Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[59] Tatusov, R., Koonin, E. and Lipman, D.: A genomic perspective on protein families. *Science*, vol. 278, no. 5338, pp. 631–637, 1997.

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

[61] Tatusov, R., Natale, D., Garkavtsev, I., Tatusova, T., Shankavaram, U., Rao, B., Kiryutin, B., Galperin, M., Fedorova, N. and Koonin, E.: The cog database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*, vol. 29, no. 1, pp. 22–28, 2001.

[62] Makarova, K.S., Sorokin, A.V., Novichkov, P.S., Wolf, Y.I. and Koonin, E.V.: Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biol Direct*, vol. 2, p. 33, 2007.

[63] Wolf, Y.I., Makarova, K.S., Yutin, N. and Koonin, E.V.: Updated clusters of orthologous genes for archaea: a complex ancestor of the archaea and the byways of horizontal gene transfer. *Biology Direct*, vol. 7, p. 46, 2012.

[64] Kristensen, D.M., Waller, A.S., Yamada, T., Bork, P., Mushegian, A.R. and Koonin, E.V.: Orthologous gene clusters and taxon signature genes for viruses of prokaryotes. *Journal of Bacteriology*, vol. 195, no. 5, pp. 941–950, 2013.

[65] Remm, M., Storm, C. and Sonnhammer, E.: Automatic clustering of orthologs and in-paralogs from pairwise species comparisons1. *Journal of Molecular Biology*, vol. 314, no. 5, pp. 1041–1052, 2001.

[66] Chen, F., Mackey, A., Stoeckert Jr, C. and Roos, D.: Orthomcl algorithm. `http://orthomcl.org/orthomcl/about.do`, 2003. Accessed March 12, 2012.

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

[68] Eisen, J.A. and Fraser, C.M.: Phylogenomics: intersection of evolution and genomics. *Science*, vol. 300, no. 5626, pp. 1706–1707, 2003.

[69] Dagan, T.: Phylogenomic networks. *Trends in Microbiology*, vol. 19, no. 10, pp. 483–491, 2011.

[70] Cavalli-Sforza, L.L. and Edwards, A.W.: Phylogenetic analysis. models and estimation procedures. *American Journal of Human Genetics*, vol. 19, no. 3 Pt 1, p. 233, 1967.

[71] Rokas, A., Williams, B.L., King, N. and Carroll, S.B.: Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, vol. 425, no. 6960, pp. 798–804, 2003.

[72] Fitzpatrick, D.A., Logue, M.E., Stajich, J.E. and Butler, G.: A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evolutionary Biology*, vol. 6, no. 1, p. 99, 2006.

[73] Gadagkar, S., Rosenberg, M. and Kumar, S.: Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, vol. 304, no. 1, pp. 64–74, 2005.

[74] Hughes, A.L., Ekollu, V., Friedman, R. and Rose, J.R.: Gene family content-based phylogeny of prokaryotes: the effect of criteria for inferring homology. *Systematic Biology*, vol. 54, no. 2, pp. 268–76, 2005.

[75] Montague, M.G. and Hutchison, C.a.: Gene content phylogeny of herpesviruses. *PNAS*, vol. 97, no. 10, pp. 5334–9, 2000.

[76] Snel, B., Bork, P., Huynen, M. *et al.*: Genome phylogeny based on gene content. *Nature Genetics*, vol. 21, pp. 108–110, 1999.

[77] Korbel, J.O., Snel, B., Huynen, M.A. and Bork, P.: Shot: a web server for the construction of genome phylogenies. *Trends in Genetics*, vol. 18, no. 3, pp. 158–162, 2002.

[78] Kloesges, T., Popa, O., Martin, W. and Dagan, T.: Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. *Molecular Biology and Evolution*, vol. 28, no. 2, pp. 1057–1074, 2011.

[79] Halary, S., Leigh, J.W., Cheaib, B., Lopez, P. and Bapteste, E.: Network analyses structure genetic diversity in independent genetic worlds. *Proceedings of the National Academy of Sciences*, vol. 107, no. 1, pp. 127–132, 2010.

[80] Juan, D., Pazos, F. and Valencia, A.: High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proceedings of the National Academy of Sciences*, vol. 105, no. 3, pp. 934–939, 2008.

[81] Clark, N.L., Alani, E. and Aquadro, C.F.: Evolutionary rate covariation reveals shared functionality and coexpression of genes. *Genome Research*, vol. 22, no. 4, pp. 714–720, 2012.

[82] Pazos, F. and Valencia, A.: Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Engineering*, vol. 14, no. 9, pp. 609–614, 2001.

[83] Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D.: Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, vol. 95, no. 25, pp. 14863–14868, 1998.

[84] Movahedi, S., Van Bel, M., Heyndrickx, K.S. and Vandepoele, K.: Comparative co-expression analysis in plant biology. *Plant, Cell & Environment*, vol. 35, pp. 1787–1798, 2012.

[85] Mutwil, M., Klie, S., Tohge, T., Giorgi, F.M., Wilkins, O., Campbell, M.M., Fernie, A.R., Usadel, B., Nikoloski, Z. and Persson, S.: Planet: combined sequence and expression comparisons across plant networks derived from seven species. *The Plant Cell Online*, vol. 23, no. 3, pp. 895–910, 2011.

[86] Ulitsky, I. and Shamir, R.: Identification of functional modules using network topology and high-throughput data. *BMC Systems Biology*, vol. 1, p. 8, 2007.

[87] Hanisch, D., Zien, A., Zimmer, R. and Lengauer, T.: Co-clustering of biological networks and gene expression data. *Bioinformatics*, vol. 18, no. suppl 1, pp. S145–S154, 2002.

Chapter 3

Multi-Mechanism Co-Evolutionary Networks Reveal Functionally-Related Genes

D. Weighill, M. Vivier and D. Jacobson

As submitted for publication in PLoS Computational Biology, currently undergoing peer review.

3.1 Abstract

Gene evolution does not occur through a single, simple, homogeneous process but rather occurs through several interconnected mechanisms, all of which are subjected to selective pressure. To date, no genome-wide study has yet been performed which investigates all three of these mechanisms of evolution simultaneously. Networks are useful structures for representing complex relationships and are thus ideal for modelling biological systems. Networks are also convenient for the integration of multiple omics data types, since they can be readily clustered, merged and compared. This study uses a module-based network construction approach in order to investigate plant gene co-evolution based on three different mechanisms of evolution, using both genomic and transcriptomic data.

Here we present the use of networks to identify and represent sets of protein encoding grapevine genes that appear to be co-evolving by one or more evolutionary mechanisms. Analysis of the resulting networks shows that gene co-evolution takes place in the scope of both biotic and abiotic selective pressures that seem to select for the co-evolution of genes that have related molecular and/or biological functions.

This network-based view of evolution gives new insights into the likely evolutionary relationships between genes, explains how selective pressures can act on pleiotropic networks of interconnected functions and provides new opportunities for the inference of functions for unannotated genes. This appears to be the first genome-wide study to address these three mechanisms of evolution simultaneously. Plant functions which were previously known to be functionally related appear to be co-evolving via several different mechanisms. The network-based method presented here provides a novel way to view the evolution of an organism and leads to new functional and evolutionary insights.

3.2 Author Summary

Plants form a crucial part of the world's ecosystems and food web. How evolution affects plants thus impacts both our local and global environments as well as our food supply. Here we report a unique network-based approach to study the evolution of plants whilst considering three evolutionary mechanisms simultaneously. 793,113 genes across 26 plant genomes were analysed for patterns of co-evolution via one or more mechanisms and the results presented as a co-evolution network. This network-based view of evolution gives new insights into the likely evolutionary relationships amongst genes and provides new opportunities for the inference of functions for unannotated genes. Genome-wide evolution underpins virtually every biological phenomena. Thus, this approach has implications in nearly every branch of plant biology.

3.3 Introduction

Grapevine is one of the most widely cultivated plants in the world [1] and its domestication and importance to human heritage stretches back to Neolithic times [2]. Thus, in addition to the traits selected for during the evolution of wild predecessors by environmental factors, domesticated grapevine has been under human selective pressure for the past 10,000 years. Since domestication, grapevine has often been propgated vegetatively, and thus its genome has been more likely to accumulate deletions and insertions [3]. As such, the *Vitis vinifera* genome has a considerable number of large gene families which have probably been stabilised by human selective pressure and vegetative propagation.

3.3.1 Evolutionary Mechanisms

Gene evolution does not occur through a single, simple, homogeneous process but rather occurs through several overlapping mechanisms [4]. For simplicity's sake, consider three interconnected mechanisms of evolution, namely evolution by point mutations, by gene duplication or deletion and evolution by gene regulation.

Point mutations that are nonsynonymous substitutions, insertions or deletions within the coding regions of a gene can result in the change of an amino acid at that particular site, potentially causing a change in the functionality of the resulting protein. In the presence of a selective pressure, point mutations that provide a selective advantage may become fixed, whereas those that are detrimental may be lost.

Evolution can occur on the scale of gene families through gene duplication and deletion that has previously been described as gene birth-and-death evolution [5]. Genes are “born" through gene duplication, and “die" either by being deleted from the genome or from accumulating so many detrimental point mutations that they are no longer functional [5] and become pseudogenes [6]. Gene duplication can occur in several ways, including unequal crossing over, retroposition and whole chromosome or genome duplication [6]. Duplicate genes are an important driver of evolution as one member of the pair of duplicated genes can perform its original function while the other is free to diverge, possibly providing an evolutionary advantage to the organism. Some gene duplication events lead to the duplicated gene becoming fixed if there is a specific advantage to having more of that gene product [6]. Subfunctionalization can also occur when both the original and duplicated gene diverge, resulting in each gene performing a part of the function of the original gene [6]. Alternatively, duplicate genes can also diverge to give rise to new functions [6], giving the organism the ability to perform a function that it could not perform before.

Finally, evolution can also occur by altered regulation of gene expression, rather than by a change in the number or function of genes. This can arise from point mutations in gene regulatory segments, intron-exon splice sites [4] or mutations that will affect the stability of mRNA [7]. A good example of the convergence of evolutionary mechanisms in grapevine is the evolution of white berry cultivars. An ancestral gene duplication resulted in two copies of the MYBA gene which regulate anthocyanin production. Subsequent retrotransposon insertion in one MYBA gene and point mutations in the other inactivated these regulatory elements and thus prevent anthocyanin biosynthesis resulting in white berries [8]. Alternatively, as we have shown previously, this can occur with changes in the expression levels of specific transcription factors [9]. Gene expression can also be affected by microRNAs and epigentic factors. Thus, as discussed above, mutations occurring in non-coding, regulatory regions of DNA are also potential drivers of evolution, even though they do not directly affect protein functionality.

3.3.2 Modeling Evolutionary Mechanisms

3.3.2.1 Point Mutations

Point-mutation-based models of evolution that result in nonsynonymous substitutions, insertions or deletions in coding regions have previously been used to predict protein-protein interactions. The Mirror-Tree method was first applied as a method for predicting protein-protein interactions by Pazos and Valencia [10] for which they used the metric of co-evolution developed by Goh *et al.* [11]. In order to identify putative co-evolving genes across a set of gene families, the Mirror-Tree method involves constructing a multiple sequence alignment for each family, calculating evolutionary distances between the genes in each alignment and unfolding the resulting evolutionary distance matrix into a phylogenetic profile for each gene. Gene co-evolution is then calculated as the Pearson correlation coefficient between the phylogenetic profiles [10]. This measure of co-evolution, sometimes called Evolutionary Rate Covariation (ERC), was then used to predict whether proteins were potentially interacting or not. Protein pairs with a high ERC value are thought to be potentially interacting. This method was found to have a high sensitivity but low specificity, producing a large number of false positives [10; 12]. An adapted version of the Mirror-Tree method was proposed by introducing a projection operator to remove information related to phylogenetic relationships between species from the phylogenetic vectors that resulted in a significant decrease in the amount of false positives [12]. ERC has also been shown to not only indicate protein interactions but also related protein functions [13] and thus suggests that ERC can also be used as a measure of protein co-evolution.

3.3.2.2 Gene Family Profiles

Gene family profiles are vectors indicating the presence of gene families across species. These gene family profiles have been used to construct phylogenies based on gene family content [14] and correlation or similarity between these profiles has been used to identify proteins that are potentially functionally related [15; 16]. However, these profiles have previously been viewed as simple binary vectors that indicate the presence or absence of a gene family in a species but leaving out any information about the size of a gene family in each species. Our current study used gene family profiles containing the size of each gene family in each of 26 plant genomes and calculated the co-evolution of gene families as the correlation between these non-binary gene family profiles.

3.3.2.3 Co-expression

Co-expressed genes are often functionally related and can be defined as the similarity (often Pearson Correlation) between the expression profiles of pairs of genes [17]. Microarray data available from many experiments can be combined to construct expression profiles across a wide variety of experimental conditions. Modules of co-expressed genes can be detected by constructing a gene co-expression network and then applying a clustering algorithm [17].

3.3.2.4 Co-evolution

Correlation of genome variants, be it gene duplications, point mutations or regulatory changes, would seem to be likely to occur as a result of a selective pressure. For example, biotic stress will be very likely to act as a selective pressure on many genes in the complex set of defense response pathways. As such, one would expect to see several correlations between the evolution of genes involved in plant defense. In addition, genome variants, regardless of the mechanism by which they are created, may be viewed as a selective pressure in their own right. For instance, variants that are selected for in bacteria that lead to drug resistance are often associated with a fitness penalty to the bacteria. However, they are selected for under a more extreme selective pressure (cell death) by the presence of a drug. It is often found that the fitness penalty incurred in obtaining drug resistance is a selective pressure for mutations in other genes that in turn compensate for the fitness penalty of the first mutation [18]. On a limited scale, such compensatory evolution (via point mutations) has also been shown to occur in plants [19; 20]. In this fashion one could expect that many genome variants will show correlations with each other as one variant becomes a selective pressure for another variant. Our primary hypothesis in this study is that such evolutionary correlations exist and are reflected in the genomes and transcriptomes of plants. Furthermore, we hypothesize that correlations that reflect co-evolution will be focused around (and informative of) specific molecular and biological functions. As such, the genome-wide study of co-evolution within and across different mechanisms is likely to not only provide new evolutionary insights but could also serve as an important source of functional inference for gene functions and further elucidate pleiotropic patterns present in plant genomes.

With this in mind, we have defined gene co-evolution in terms of three broad mechanisms of evolution. Genes can be co-evolving in terms of (1) having similar point mutation rates, (2) being members of gene families that are correlated across species and (3) having similar gene expression profiles. These three models of co-evolution have been addressed previously to some extent but typically in isolation [21; 12; 17]. Our study aims to investigate these three models simultaneously.

To our knowledge, these three mechanisms of evolution (point mutations, gene duplication/deletion and gene expression regulation) have not been simultaneously investigated in a genome-wide study of gene co-evolution. This study aims to investigate the co-evolution of grapevine genes using a module-based network model encompassing all three mechanisms of evolution. We believe that this approach can be useful for both functional inference for sets of genes as well as providing better understanding of the evolutionary patterns in any organism. In this paper we have focussed on grapevine as it is an economically important species that is not a model organism. One of the challenges of studying a non-model organism is that there is typically less functional information known about genes and thus genome-wide hypothesis generation and experimental result interpretation is more difficult. As such, this method's ability to bring together multiple lines of evidence for gene co-evolution and mine the information available from other sequenced plant genomes in order to do functional inference is particulary attractive in a non-model organism such as grapevine.

3.4 Results and Discussion: Evolutionary Mechanism Interplay

As shown in Figure 3.1, modules of co-evolving genes in grapevine (*V. vinifera*) were constructed based on three mechanisms of evolution. Gene co-expression modules were constructed in order to represent genes that are potentially co-evolving in terms of their regulation of gene expression. Gene family correlation modules were constructed that represent gene families with distinct molecular functions that nevertheless appear to be co-evolving through gene duplication and deletion. Lastly, evolutionary rate covariation (ERC) modules were constructed that represent genes that are co-evolving in terms of having co-varying point mutation rates [12]. For each pair of evolutionary mechanisms an overlap network was constructed in which the nodes represent co-evolution modules and the edges represent enriched overlaps between these evolutionary modules. Overlaps were quantified by the Jaccard index [22] and enrichment was determined using the hypergeometric test with multiple hypothesis correction. The resulting networks were visualized in Cytoscape [23]. Figure S1 shows the overlap network between the ERC modules and the co-expression modules, Figure S2 shows the overlap network between the ERC modules and the gene family modules and Figure S3 shows the overlap network of the co-expression modules and the gene family modules. In each case, blue nodes represent co-expression modules, pink nodes represent ERC modules and yellow nodes represent gene family modules. Node sizes are scaled by the number of genes in that particular co-evolution module.

In each of these networks a node represents a set of genes that are co-evolving in terms of a single mechanism of evolution and an edge represents a set of genes that are co-evolving in terms of two mechanisms of evolution. For ex-

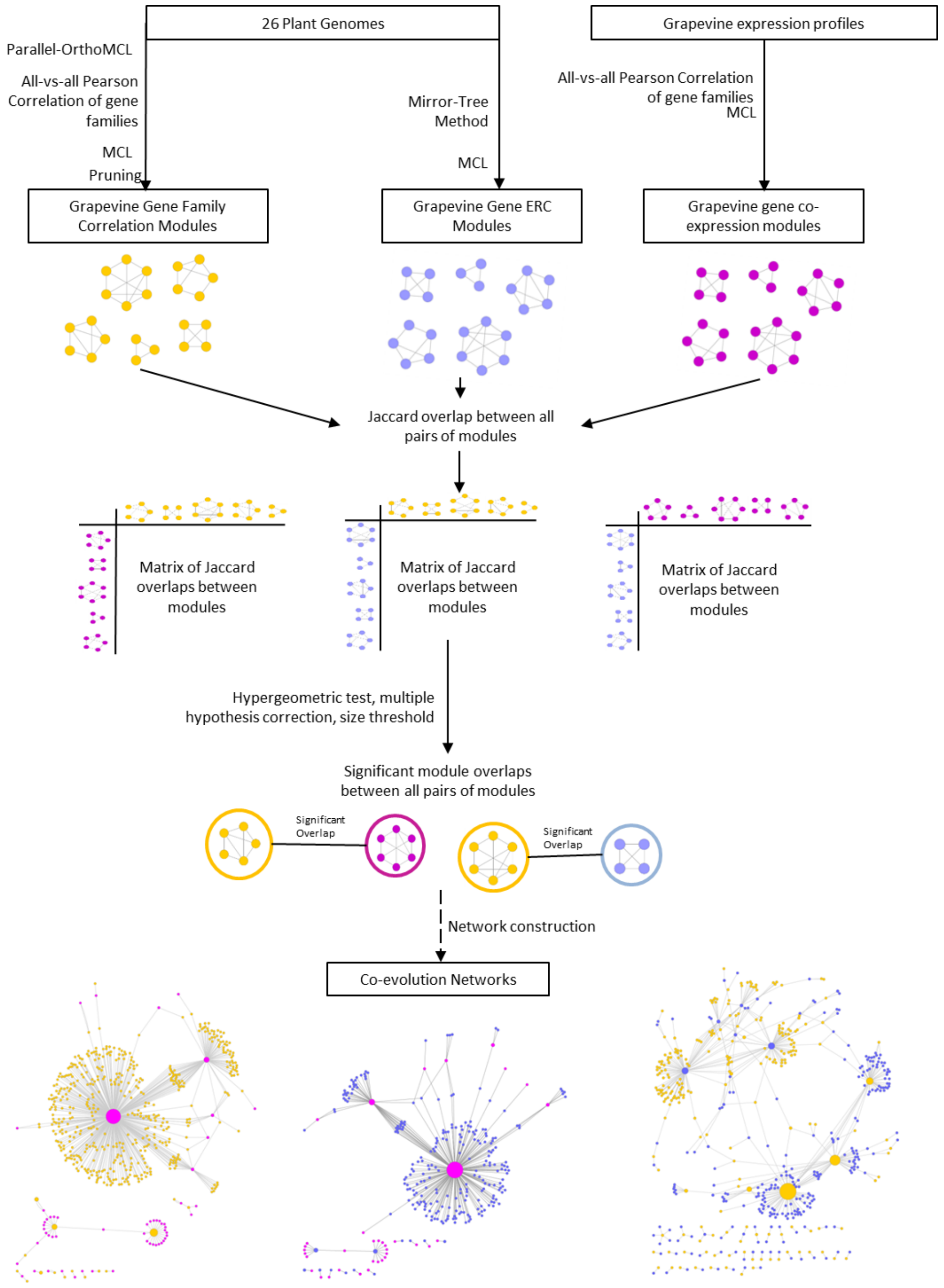

26 Plant Genomes
Grapevine expression profiles
Parallel-OrthoMCL
All-vs-all Pearson Correlation of gene families
MCL
Pruning
Mirror-Tree Method
MCL
All-vs-all Pearson Correlation of gene families
MCL
Grapevine Gene Family Correlation Modules
Grapevine Gene ERC Modules
Grapevine gene co-expression modules
Jaccard overlap between all pairs of modules
Matrix of Jaccard overlaps between modules
Matrix of Jaccard overlaps between modules
Matrix of Jaccard overlaps between modules
Hypergeometric test, multiple hypothesis correction, size threshold
Significant module overlaps between all pairs of modules
Significant Overlap
Significant Overlap
Network construction
Co-evolution Networks

Figure 3.1: **Methods Summary.** Summary of the workflow used for constructing the co-evolution networks.

ample, an edge between a co-expression module and a gene family correlation module represents a set of genes that appear to be co-evolving in terms of their transcriptional regulation, but are also members of gene families that appear to be co-evolving through gene duplication or deletion. The presence of edges signifies genes that are potentially co-evolving in terms of more than one mechanism of evolution, showing an interplay and overlap between different evolutionary mechanisms. However, the presence of large hub nodes connected to many other nodes suggests that genes that are co-evolving in terms of one mechanism can be split into several subsets that are also co-evolving in terms of the other mechanisms of evolution.

In order to investigate the interplay between all three mechanisms of evolution simultaneously, a combined co-evolution network (Figure 3.2A) was created by merging the three separate co-evolution networks. This scale free network (see Figure S4), which contains 19,137 grapevine genes, was visualised, queried and explored in an interactive Cytoscape session as a method for hypothesis generation. The number of gene pairs that are co-evolving in terms of different evolutionary mechanisms are shown in Figure 3.2B. The number of genes that are present in all three kinds of co-evolutionary modules, in two kinds of modules and in single kinds of modules are summarised in Figure 3.2C. The Gene Ontology (GO) [24] is commonly used to provide a unified view of the functional annotations of gene products and, as such, includes over 39,000 standardised terms that are used to annotate the molecular functions, biological processes and cellular locations of genes in sequenced genomes. The genes in each co-evolution module were annotated with Gene Ontology terms thus allowing the functions of potentially co-evolving genes to be investigated. Local topologies of the network were investigated by selecting a node and performing a breadth-first search with a distance of one or two hops from the starting node. For all nodes present in a subnetwork found by such a breadth-first search, the genes were extracted and GO term enrichment was performed on that subset of genes (node enrichment view) using GOEAST [25]. Enrichment was also performed on the set of genes present in the edges of the same subnetwork (edge enrichment view).

The GOEAST enrichment results are shown in Figures S7-S16. Several subnetworks were found to be enriched in molecular and biological functions that are known to functionally interact within the cell, such as defense-related functions or synthesis and transport functions. The fact that functionally interactive genes were found to be enriched in specific neighbourhoods of the network is indicative that our model is identifying functionally interacting genes that appear to be co-evolving around shared selective pressures. From an evolutionary point of view this makes sense as abiotic and biotic selective pressures would be expected to exert influence on the evolution of a range of functionally interactive genes. Four such subnetworks and their locations in the co-evolutionary network are shown in Figure S5, and will be discussed as examples.

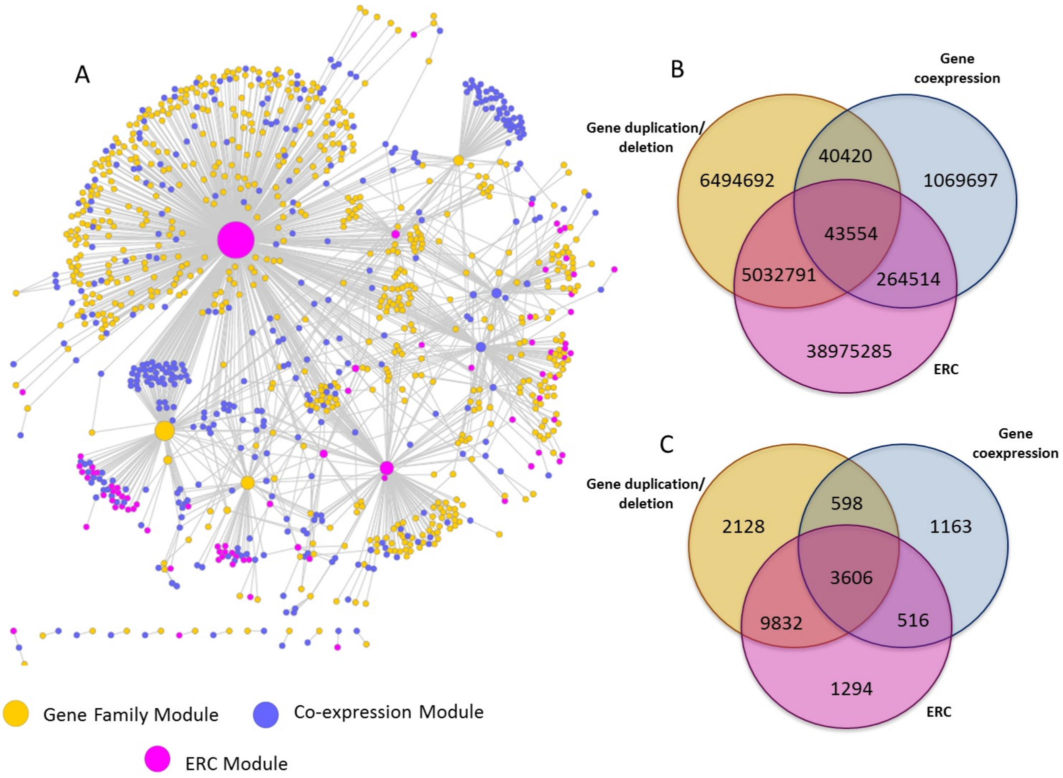


Figure 3.2: **Combined Co-evolution Network.** (A) Network of significant overlaps between all three types of co-evolution modules. This network was constructed by merging the networks from Supplementary Figures S2, S3 and S4. Each node represents a module of potentially co-evolving genes: Blue nodes represent gene co-expression modules, pink nodes represent ERC modules and yellow nodes represent gene family modules. Edges (lines between nodes) represent significant overlaps between co-evolution modules. (B) Number of grapevine gene pairs that are co-evolving in terms of 1, 2 and 3 mechanisms of evolution. (C) Number of grapevine genes that are evolving in terms of 1, 2 and 3 mechanisms of evolution.

3.4.1 Defense response co-evolution

Subnetwork 1. The first subnetwork (Figure S5 A) consists of a central gene family correlation module (yellow node) intersecting with several co-expression modules (blue nodes) and one ERC module (pink node). The node enrichment view of this subnetwork showed enrichment in several GO terms, including defense response, apoptosis and cellulose biosynthetic process. The edge enrichment view of this subnetwork also showed enrichment in apoptotic process and cellular glucan metabolic process. Thus the node and edge enrichment views showed much of the same enriched functions, even though the subset of genes in the edges is considerably smaller than the subset of genes in the nodes. Cell wall construction and defense-related functions were also enriched in this subnetwork when excluding the gene family module as well as when examining the gene family module in isolation. As such, the enrichment in this subnetwork is not being driven by a single node. The functions enriched in this subnetwork seem to be related to a theme of plant defense. Apart from known defense responses, such as apoptosis, cell wall construction functions, including cellulose biosynthesis, are also enriched in this subnetwork, which would also seem to fit into the plant defense theme as plants are known to modify their cell walls as a defense response to invading microbes. Examples of such modifications include the cross-linking of structural proteins present in the cell wall [26] and lignification [27]. Although the function of ligninification was not enriched in this subnetwork based on GO terms (and thus did not appear in the GO term enrichment) several genes were found to be orthologs of *A. thaliana* laccase-like genes, which are known to be involved in lignin biosynthesis. Thus, even though it was not seen in the GO enrichment view of this subnetwork, ligninification is present in this subnetwork and fits in with the defense theme. Furthermore, it has previously been shown that changes in cellulose biosynthesis can activate lignin synthesis and defense responses through jasmonate, ethylene and other signaling pathways [28] and enhances disease resistance [29].

Subnetwork 2. The second subnetwork (Figure S5 B) consists of a central co-expression module (blue node) surrounded by several gene family modules (yellow nodes), ERC modules (pink nodes) and one other co-expression module. The node enrichment view of this subnetwork showed several enriched functions, including several GO terms related to chromosome organization and nucleosome assembly, cell wall macromolecule catabolic process, monoterpene biosynthetic process, terpenoid transport, chitin catabolic process and response to other organisms. The edge enrichment view of this subnetwork also showed enriched functions in terpenoid transport and nucleosome assembly and extended the set of functions found in the node enrichment view to include cellulose biosynthetic process, plant hypersensitive response, defense response to virus, response to bacterium, response to ethylene stimulus, response to jasmonic acid stimulus, response to salicylic acid and response to various other compounds.

In order to investigate whether a combined set of functions enriched in both the node view and the edge view of a subnetwork would reveal a more complete and detailed set of functions, MultiGOEAST [25] was used to combine and compare the enrichment results from the edge enrichment and node enrichment views of this subnetwork (Figure S11). Yellow rectangles represent GO terms that are enriched in both the edge and the node view, green rectangles represent GO terms only enriched in the edge view and red rectangles represent GO terms only enriched in the node view. From Figure S11, it can be seen that although the enriched functions are sometimes overlapping, the node view and the edge view provide many related but non-overlapping functions. Clearly, a more complete picture emerges by interpreting the nodes and edges as combined sets of enriched functions.

The main functions found in this subnetwork and how these functions relate to each other are summarised in Figure S6 A. This subnetwork also appears to have a defense-related theme. Direct defense-related functions include defense responses to viruses and bacteria, as well as plant-type hypersensitive response. Cell wall construction functions were also enriched in this subnetwork and, as explained for subnetwork 1, these could also be defense-related. Chitin catabolic process could be an indication of plant chitinases, which are known to be involved in plant defense against fungi by digesting the chitin in fungal cell walls [30; 31]. Response to various signalling hormones were also enriched, including jasmonic acid, ethylene and salicylic acid. All three of these hormones are known to be involved in activating plant defense responses [32]. The pathways of these three hormones are also known to interact, causing crosstalk [32].

Terpenoid synthesis and terpenoid transport functions are both enriched in this subnetwork and jasmonic acid is also known to stimulate the production of terpenoids [33]. Because of this functional link between jasmonic acid and terpenoid synthesis, it is perhaps not surprising that these two functions seem to be co-evolving. Terpenoids also play a role in defense response and disease resistance in plants [34; 35], possibly further explaining the presence of terpenoid synthesis-related functions in the same co-evolution module as other defense-related functions. It also makes sense that terpenoid synthesis and terpenoid transport functions could be co-evolving, since a change in the regulation of terpenoid synthesis could itself act as a selective pressure with regards to the terpenoid transport needs of a cell. It is notable that the grapevine genome contains a very large family of terpene synthases with 69 putatively functional genes [36]. This is roughly twice as many terpene synthases than has been seen in any other plant genome to date [37].

Epigenetic functions such as nucleosome assembly, DNA packaging and DNA conformational change were enriched in this subnetwork along with defense and stress related functions. These epigenetic and defense functions were also enriched in this subnetwork when excluding the central co-expression module, and when looking at the central co-expression module in isolation. Epigenetic regulation is known to play a role in gene expression regulation under stress conditions [38]. Thus, all the functions mentioned above, which, according to our model, are co-evolving via three different mechanisms, are known to be functionally related in plants with evidence from previous studies.

3.4.2 Protein translation, transport and degradation co-evolution

Subnetwork 3. The third subnetwork consists of a central ERC module intersecting with several co-expression modules and gene family modules (Figure S5 C). Using biological process GO enrichment, the node enrichment view of this group includes many functions related to the control of protein levels and localisation in the cell, including proteolysis, gene expression, translation, tRNA metabolism, vesicle-mediated transport and protein transport. Edge enrichment adds to this theme and shows enrichment in protein deneddylation, intracellular protein transport and ribosome biogenesis. From the cellular component GO hierarchy, the edge enrichment view also includes "clathrin adaptor complex" that forms part of the coat of protein transport vesicles and aids in the protein transport process [39]. The combined node and edge enrichment views thus suggest that functions related to protein synthesis, transport and degradation appear to be co-evolving. Ribosome biogenesis and tRNA metabolism are clearly functionally linked as they are the machinery required for protein translation. There is literature evidence to suggest that there is a functional link between protein degradation and translation, which is related to the quality control of proteins in the Endoplasmic Reticulum [40]. The combination of the node and edge enrichment views also provides a more complete picture than either of the two views in isolation. This suggests that components of these processes could have co-evolved via separate mechanisms whilst specific key elements may have co-evolved via multiple evolutionary mechanisms.

3.4.3 Stress response and developmental gene co-evolution

Subnetwork 4. The fourth subnetwork consists of a central gene family module surrounded by several co-expression modules and ERC modules (Figure S5 D). The node enrichment view includes functions such as pentacyclic triterpenoid biosynthetic process, (1-3)-beta-D-glucan biosynthetic process, terpenoid transport, plant-type cell wall modification, plant-type cell wall biogenesis, various response functions including response to hormone stimulus, response to chitin, response to water deprivation, response to salt stress, response to light stimulus, as well as several developmental functions. The edge enrichment view complements the node enrichment view with functions related to trehalose biosynthesis, xylan catabolism, defense response to fungus, and hyperosmotic stress.

The combined node and edge enrichment views seem to indicate that functions related to abiotic and biotic stress responses appear to be co-evolving. The main related functions enriched in this subnetwork are summarised in Figure S6 B. Again, the two functions of terpenoid synthesis and terpenoid transport are enriched in the same neighbourhood, suggesting the evolutionary link between these two functions. Terpenoids are thought to play a role in plant defense responses, potentially explaining why the terpenoid related functions appear alongside defense-related functions [34; 35]. The (1-3)-beta-D-glucan biosynthetic process could also be a defense-related process as callose, a beta-1,3-glucan polymer, is a component of the papilla, a cell wall apposition that is a protective layer formed by a plant at the site of a fungal attack [41; 42]. Trehalose biosynthesis is also related to plant defense, as it elicits plant defense responses [43] and is also involved in abiotic stress response [43]. In addition, trehalose has been associated with resistance to salt stress and drought stress [44]. This is also supported by the enrichment of response to salt stress and response to water deprivation functions in the same neighbourhood as trehalose biosynthesis. Cell wall modification is also related to plant defense response as plants are known to modify their cell walls in order to strengthen them against pathogenic attacks including structural protein crosslinking and lignification [26; 27].

A plant is effectively constantly monitoring its environment and responding appropriately, whether that response is growth or defence or stress response. When plants respond to biotic and abiotic stress it is known that, although the origins of the signals may be distinct, there is considerable overlap in the resulting signalling cascades. Plant defence responses often involve modification of the cell wall. Similarly, the growth and development of a plant also, by necessity, requires changes to the cell walls of tissues that are expanding and or transforming into different developmental stages. Terpenoids are also known to play roles in both plant defence, signalling and development. It would seem likely then that there could be considerable overlap in the signaling machinery to be used in this process of environmental monitoring and responses. This would be consistant with what we are seeing in Subnetwork 4 which contains elements of all of these functions. This concept of the co-evolution of signaling cross-talk is explored further below with gene-based subnetworks.

In all of the subnetworks discussed above, the node enrichment and edge enrichment views give non-identical yet related and often quite complementary information about enriched functions. Looking at the node enrichment view or the edge enrichment view in isolation would not give as complete a picture of co-evolving and related functions as the combined view does. This suggests that there are certain cellular functions that co-evolve through distinct individual mechanisms of evolution, while some functions are co-evolving via multiple evolutionary mechanisms.

3.4.4 Gene-based subnetworks

There seems to be a mixture of functions related to biotic and abiotic stress in subnetwork 4, suggesting crosstalk between these responses. There is literature evidence that supports the idea of interactions and crosstalk between wounding, biotic and abiotic stresses and hormone responses in plants, often through the hormones abscisic acid, salicylic acid, jasmonic acid and ethylene [45; 46]. Although these hormone response functions are not individually enriched in this subnetwork, there are modules within this subnetwork that are annotated with these hormones that may well explain the presence of both biotic and abiotic stress response functions. In order to explore this further, a selection of GO-terms related to these functions were chosen and a GO-gene network was constructed in which nodes represent genes in subnetwork 4 annotated with at least 2 of these GO terms and edges linked genes to these GO-terms (Figure S17). The different categories of functions (e.g. hormone responses and signalling, abiotic stress responses, abiotic stress responses and actual physical defense responses) group together well within this network and seem to form a quasi-pathway of stress response, starting with general hormone signalling at the top of the network, followed by more specific (abiotic or biotic) stress response signalling (which includes crosstalk) and ending with actual cellular responses such as cell wall modification or cell death at the bottom of the network. The presence of genes linking these different functions supports the hypothesis of the co-evolving stress response functions present in subnetwork 4. Figure S18 shows a subnetwork of this quasi-pathway within 2 edges of the GO-term “defense response to bacteria". This is a zoomed in view of the genes that are linking the different GO functions and thus putatively accounting for the apparent crosstalk. Many of these genes are kinases involved in signalling, such as the cystein rich receptor-like protein kinase CRK10 (VV00G16870, VV03G01940, VV02G09070 and VV10G07230), which links "defense response to bacterium" to "response to salicylic acid stimulus, and the histidine kinase AHK4 (VV01G05500 and VV01G04890), which links defense responses (to bacteria and fungus) to response to abiotic stresses (salt stress and water deprivation). Transcription factors are also present in this network causing crosstalk between functions, including WRKY40 (VV09G00130), which connects response to wounding, response to salicylic acid stimulus and response to bacteria and fungus, and MYB91 (VV08G14530), which connects response to salicylic acid, auxin and jasmonic acid, response to bacteria and fungi as well as response to salt stress. Another view of this network is shown in Figure S19, in which genes are connected to the modules in which they are present. From this network one can see that these genes are well connected through all three types of co-evolution modules, again suggesting considerable co-evolution of these cross-talking stress response functions. A combined view constructed by merging the gene-go network and the gene-module network is shown in Figure S20.

Functionally related subnetwork linkages The defense-related theme seemed to be present in multiple subnetworks, namely subnetworks 1, 2 and 4 (Figures S4 A, B and D, respectively). In order to investigate whether these subnetworks had some spatial link in the combined network, the subnetworks were concatenated and mapped back to the combined overlap network. When looking at the location of these subnetworks in the combined overlap network, they were adjacent. (Figure 3.3). It is thus not surprising that they share some functional similarities.

In order to better understand the functional interactions amongst subnetworks 1, 2 and 4 (Figure 3.3), GO-terms that were shared between the nodes present in the sections linking the three adjacent subnetworks together were identified and used to create a GO-gene network as described above. The resulting network is shown in Figure 3.4. InterPro annotations and descriptions of *A. thaliana* orthologs were then assigned to each gene in this network. If one examines the GO term nodes in Figure 3.4, it is apparent that the main functions linking the three subnetworks are defense- and stress-related functions found in each of the individual subnetworks, including cellulose biosynthesis, apoptosis, response to jasmonic acid, defense response, response to wounding and response to salt stress. This explains the presence of these related functions in these three separate subnetworks and indicates that edges within this co-evolutionary network provide plausible functional links. At the top of Figure 3.4, genes are annotated as being involved in responses to various stresses (abiotic and biotic) and response to hormones. At the bottom, the genes are associated with an actual defense response, namely apoptosis. Thus, this network seems to represent a broader putative pathway that begins with response to various stresses followed by the subsequent signalling cascade leading to the end result: apoptosis as a defense response. This is supported by the InterPro and *A. thaliana* ortholog description annotations. Genes near the top of the network are annotated as protein kinases, such as MEKK1 and AHK4 (VV12G05950 and VV12G09480 respectively), which are known to be involved in signal transduction, and a histone deacetylase HDA6 (VV17G01160), which could play a role in epigenetic regulation of transcription. Also present are genes annotated as specific transcription factors related to plant defense, including WRKY33 and WRKY40 (VV08G06390 and VV09G00130 respectively). At the bottom of the network, many of the genes are annotated as coding for disease-resistant proteins. For example, VV07G02940, VV12G06830 and VV00G25990 all contain the NB-ARC domain, which the Interpro annotation indicates is involved in signalling related to apoptosis. Grapevine gene IDs are Plaza *V. vinifera* IDs and gene names are those of the closest *Arabidopsis* ortholog.

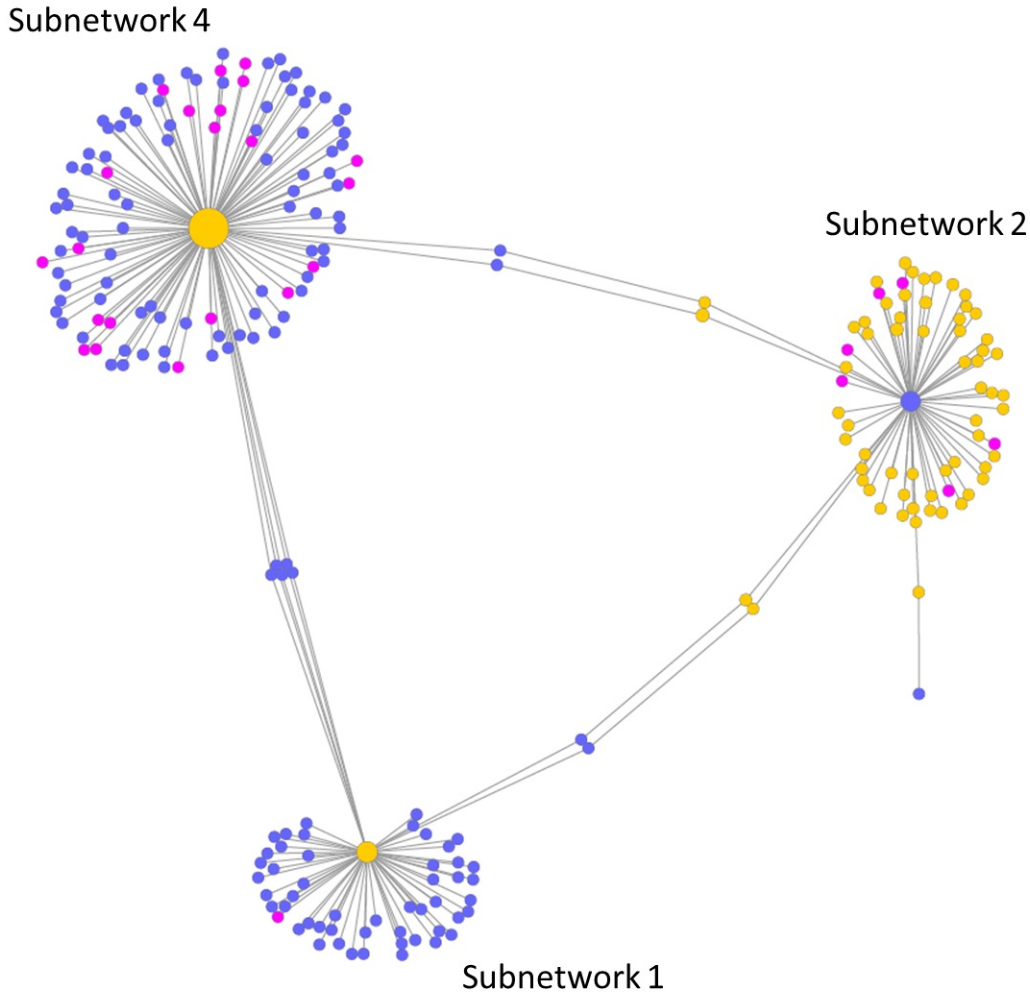


Figure 3.3: **Linked Subnetworks.** Subnetworks 1, 2 and 4 combined to produce a connected network. All three of these subnetworks showed enrichment in defense-related functions and are in close proximity in the combined overlap network.

3.4.5 Networks as Models of Evolution

Networks are very useful mathematical structures that are capable of representing complex biological relationships and have been used successfully for a wide variety of tasks, including the identification of gene families [47; 21] and in studying gene expression [48; 49; 50]. Networks consist of a set of nodes (drawn as circles in our networks) representing a set of objects of interest and a set of edges (drawn as straight lines between two nodes) representing relationships between these objects. In our study we have used nodes to represent mod-

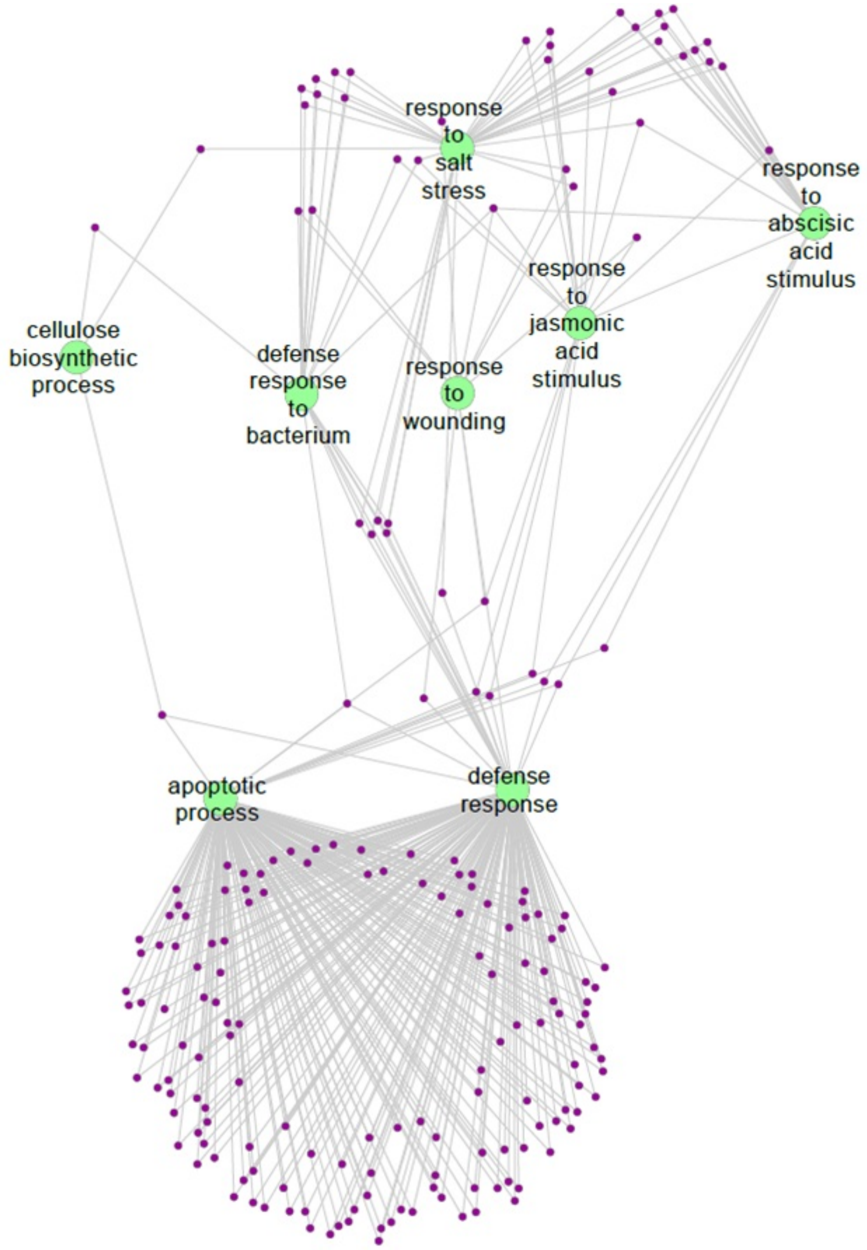

response to salt stress
response to abscisic acid stimulus
response to jasmonic acid stimulus
cellulose biosynthetic process
defense response to bacterium
response to wounding
apoptotic process
defense response

Figure 3.4: **Quasi-pathway.** Network of the main functions present in the links between subnetworks 1, 2 and 4 and the genes associated with those functions. Purple nodes represent genes, green nodes represent GO terms. Edges link GO terms to genes associated with those terms or link GO terms in close proximity in the GO hierarchy.

---

ules (sets) of potentially co-evolving genes based on a particular mechanism of evolution. Edges represent overlaps between these modules, thus indicating a possible interplay between different evolutionary mechanisms. Networks have proven to be an excellent tool for a study of evolution as they can easily be clustered, merged, compared, queried and manipulated both mathematically and visually. Furthermore, as an abstract model, networks seamlessly handle the use of multiple data types. GO enrichment performed on local network topologies of the combined co-evolution network was used to reveal the main functions of co-evolving grapevine genes.

A module-based network construction approach allowed for three different mechanisms of evolution to be combined into a single evolutionary model for grapevine. GO annotation of the network and enrichment analysis of subnetworks allowed for putatively co-evolving functions to be identified. The combination of the node and edge enrichment views provided more complete pictures of co-evolving functions than either of the two views in isolation and points towards functional elements that are predominantly co-evolving via a single evolutionary mechanism and a subset of key genes that are co-evolving by multiple mechanisms. Many of the functions that appear to be co-evolving in the four subnetworks used as examples in this paper are already known to be functionally related and have substantial literature support. This would seem to support the validity of this network-based method as a useful model of evolution. This network model can easily be extended to other species and can be used for further hypothesis generation in a systematic fashion across all network topologies. As such, this approach has the potential to significantly improve our understanding of the evolution of grapevine as well as other species.

It is possible that there is bias in the portion of the evolutionary network attributable to gene expression patterns as many of the transcriptome studies in grapevine have focused on biotic or abiotic stress responses. Furthermore the majority of the publicly available transcriptome data for grapevine was produced with the Affymetrix microarray which covers roughly half of the genes in the grapevine genome. These are the limitations of working with non-model species for which the data is more limited. However, the approach taken here combines data from multiple perspectives and the networks discussed include genes that were identified as co-evolving by other mechanisms besides gene regulation and, as such, we believe that the networks are reasonably robust to these sorts of potential biases.

Frequently, thirty percent of protein-encoding genes found in the genomes of eukaryotes are either unannotated or annotated as proteins of unknown function. This represents a significant challenge to our quest to understand an organism as a complex system of interacting molecular and biological functions. However, as we have shown, the network topologies in our evolutionary model correspond to related/interacting functions. Thus, this evolutionary network can be used as a new method by which to do guilt-by-association/functional inference hypothesis generation for the 3600 grapevine genes of unknown function present in the network. This should, for example, help in determining phenotypes to look for in gene deletion/silencing or over-expression studies of genes of unknown function.

3.5 Materials and Methods

3.5.1 Co-evolution Module Construction

Modules of co-evolving grapevine genes were constructed separately assuming three models of evolution, namely evolution by gene duplication (gene family correlation modules), evolution by gene expression regulation (gene co-expression modules) and evolution by point mutations (Evolutionary Rate Covariation modules). A summary of the workflow used to construct these modules is shown in Figure 3.1.

3.5.1.1 Gene Family Correlation Modules

26 translated plant genomes were obtained, of which 25 (including grapevine) were downloaded from Plaza (version 2.5) [51] and the potato genome was downloaded from the Solanaceae Genomics Resource [http://solanaceae.plantbiology.msu.edu/]. Gene families were constructed across these 26 plant genomes using our newly developed Parallel-OrthoMCL (described elsewhere), a parallel version of the OrthoMCL software package [21], allowing gene families to be identified across much larger sets of genomes. A species-family matrix (SF-matrix) was constructed in which the columns represented plant species and rows represented gene families, such that entry ij was the number of genes in gene family i present in species j. Gene families that were correlated across species were then determined by calculating the Pearson correlation coefficient between all pairs of rows of the SF-matrix using mcxarray [52] and applying an absolute threshold of 0.8. Modules of correlated gene families were then constructed by clustering the resulting thresholded network using MCL [52]. These modules were subsequently pruned to remove all non-grapevine genes, resulting in modules of correlated grapevine gene families.

3.5.1.2 Gene Co-expression Modules

472 microarray experiments using the Grapevine Affymetrix Genechip were downloaded from Gene Expression Omnibus and processed using RMA [53]. Co-expression was then calculated as the Pearson correlation coefficient between the expression profiles of the probes [52]. An absolute threshold of 0.8 was applied and the resulting thresholded network was clustered using MCL [52] in order to create co-expression modules. Probes were mapped to their

corresponding genes in order to produce modules of co-expressed grapevine genes.

3.5.1.3 Evolutionary Rate Covariation Modules

Modules of genes with similar evolutionary rates were constructed using the Mirror-Tree method adapted with the projection operator [12]. A 'minimal gene family' was constructed around each grapevine gene by selecting the best ortholog or co-ortholog of that gene in each of the 26 plant species used in Parallel-OrthoMCL. A minimum family size threshold of 5 was applied. The amino acid sequences of the resulting gene families were aligned using MUSCLE [54]. The evolutionary distances between genes within each family were calculated using the ProtDist program from the PHYLIP package [55], resulting in a distance matrix for each grapevine gene. The distance matrices were then unfolded into phylogenetic vectors v_i. All phylogenetic vectors were then normalised by their standard deviation, and the average phylogenetic vector was calculated as:

$$v_{av} = \frac{1}{m} \sum_{i=1}^{m} \frac{v_i^s}{||v_i^s||} \tag{3.5.1}$$

where v_i^s is a phylogenetic vector normalised by standard deviation and $||.||$ is the euclidean norm [12]. The projection operator in equation 3.5.2 was applied to each of the original phylogenetic vectors.

$$\epsilon_i = v_i - v_{av}\langle v_i, v_{av}\rangle \tag{3.5.2}$$

The evolutionary rate covariation between grapevine genes was then calculated as the Pearson correlation coefficient between all pairs of the projected phylogenetic vectors ϵ_i using mcxarray [52]. An absolute threshold of 0.9 was applied, after which the thresholded network was clustered using MCL [52]. The resulting clusters represent modules of grapevine genes with similar evolutionary rate covariation signatures.

3.5.2 Module Overlap and GO Enrichment

For each pair of evolutionary mechanisms, the module overlap was calculated between all pairs of modules using the Jaccard Index. For two sets A and B, the Jaccard Index J_{AB} is defined as the size of the intersection of the two sets, divided by the size of the union of the two sets (Equation 3.5.3).

$$J_{AB} = \frac{|A \cap B|}{|A \cup B|} \tag{3.5.3}$$

In the case of overlaps between ERC modules and gene family modules, inparalogs were excluded from intersections. This resulted in an overlap matrix for each pair of evolutionary mechanisms, in which the columns represented co-evolutionary modules from one mechanism, rows represented co-evolutionary modules from another mechanism, and entry ij represented the jaccard overlap between modules i and j. The right tailed Fisher exact test was used to identify significant module overlaps. This was performed using a customized Perl program which made use of the Text::NSP::Measures::2D::Fisher Perl module from CPAN (http://www.cpan.org/). When testing the null hypothesis "module i is not enriched in module j, the p-value was calculated as:

$$p = \frac{\binom{R}{x}\binom{T-R}{C-x}}{\binom{T}{C}} \tag{3.5.4}$$

where x is entry ij, R is the sum of row i, C is the sum of column j and T is the sum of all entries in the matrix. The Holm-Bonferroni method was used for multiple hypothesis correction [56]. An intersection cut-off of 2 was the applied. Networks were then constructed from the significant module overlaps and visualized in Cytoscape [23]. A combined co-evolution network was constructed by merging the three co-evolution networks constructed for pairs of evolutionary mechanisms. GO terms for the grapevine genes were downloaded from Plaza (version 2.5) [51] and mapped onto their corresponding modules. GO enrichment was performed on subsets of genes in local neighbourhoods of the combined co-evolution network using GOEAST [25]. For each subnetwork in question, two sets of genes were extracted using a customized Perl script, specifically, the genes present in the nodes as well as the genes present in the edges. GO term enrichment was performed on each of these sets of genes, called node enrichment and edge enrichment, respectively [25]. MultiGOEAST was used to compare the GOEAST results from the node enrichment and edge enrichment views of a subnetwork.

3.5.3 Module-GO-Term Network Construction

A GO-term network was constructed to investigate the functions present in the nodes and edges linking these subnetworks 1, 2 and 4. (Figure S21). Each module was connected to nodes representing GO terms associated with that module. GO terms were also linked if they were within a distance of 2 from each other in the GO hierarchy.

3.5.4 Gene-GO-Term Network Construction

The main functions linking subnetworks 1, 2 and 4 were selected as the central GO term nodes in the Module-GO-Term network (Figure S21), namely response to salt stress, cellulose biosynthetic process, defense response to bacterium, response to wounding, response to jasmonic acid stimulus, response to abscisic acid stimulus, apoptotic process and defense response. Genes which were present in subnetworks 1, 2 and 4 which were annotated with at least 2 of these terms were then selected. A network was constructed in which each node represented either a gene (purple nodes) or GO-terms (light green nodes) (Figure 3.4). GO-terms were connected to genes annotated with that term, and GO-terms were also linked if they were within a distance of 2 from each other in the GO hierarchy. Grapevine InterPro annotations as well as the *Arabidopsis* gene descriptions were downloaded from Plaza (version 2.5) [51]. Grapevine genes were annotated with the descriptions of their *Arabidopsis* orthologs, as determined by Parallel-OrthoMCL. The genes in this network were then assigned InterPro annotations and gene descriptions where possible.

3.6 Acknowledgments

The authors would like to acknowledge The Center for High Performance Computing and the Stellenbosch High Performance Computing Cluster for computing resources, as well as the South African National Research Foundation Technology (www.nrf.ac.za) and Human Resources Programme and Winetech. The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

The authors like to thank Armin Geiger and Piet Jones for useful discussions and Philip Young, Erik Alexandersson, Tiaan Heunis, Klaas Vandepoele, Ray Everngam and Bruce Jacobson for reading the manuscript.

3.7 Author Contributions

D Jacobson conceived of, designed and supervised the study. D Weighill wrote the code and interpreted the resulting networks. D Jacobson and D Weighill discussed the results and co-wrote the manuscript. M Vivier critically read the manuscript and gave editorial input.

3.8 Competing Interests

The authors declare that they have no competing financial interests.

3.9 Supplementary Material

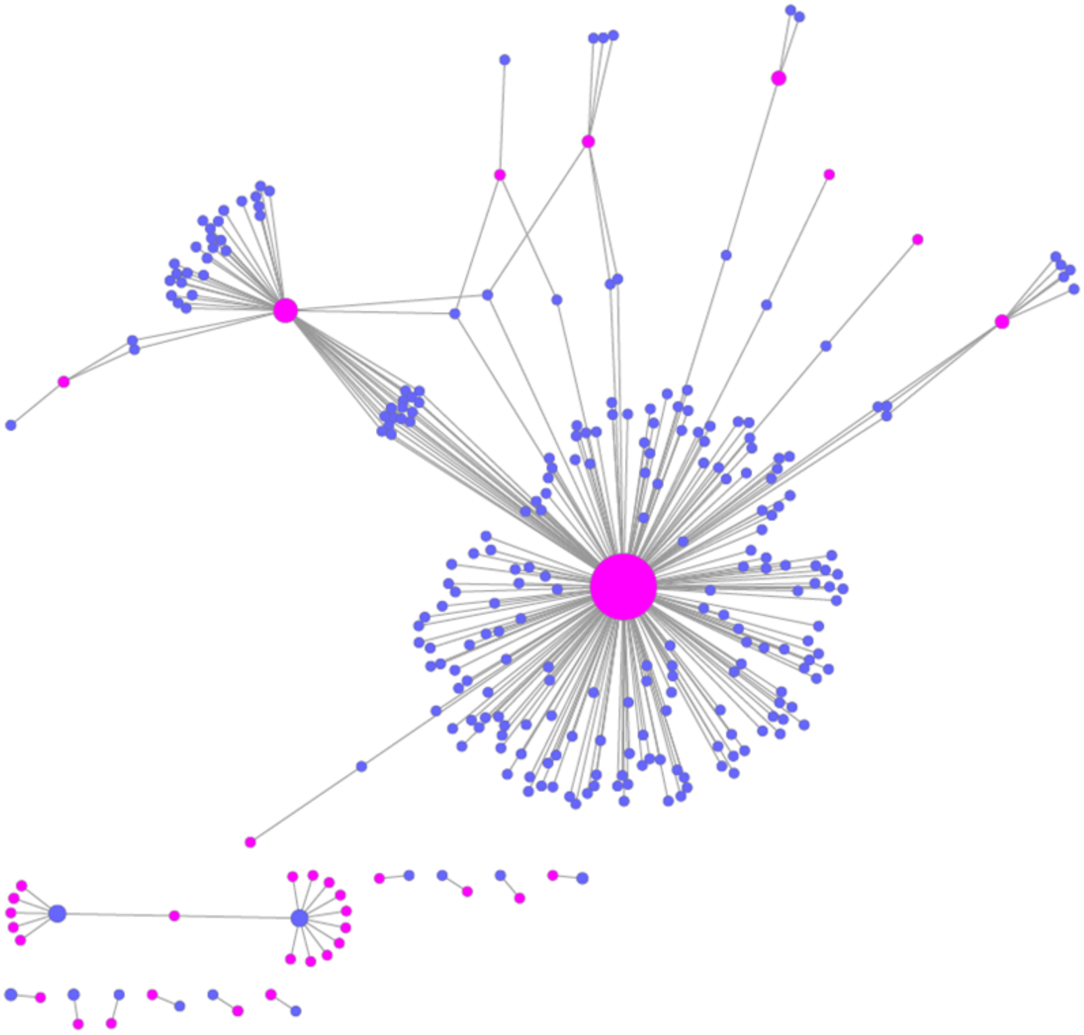


Figure S1: **Co-expression-ERC Co-evolution Network.** Network of significant overlaps between co-expression modules (blue nodes) and ERC modules (pink nodes). Edges represent significant overlaps between the co-expression modules and the ERC modules.

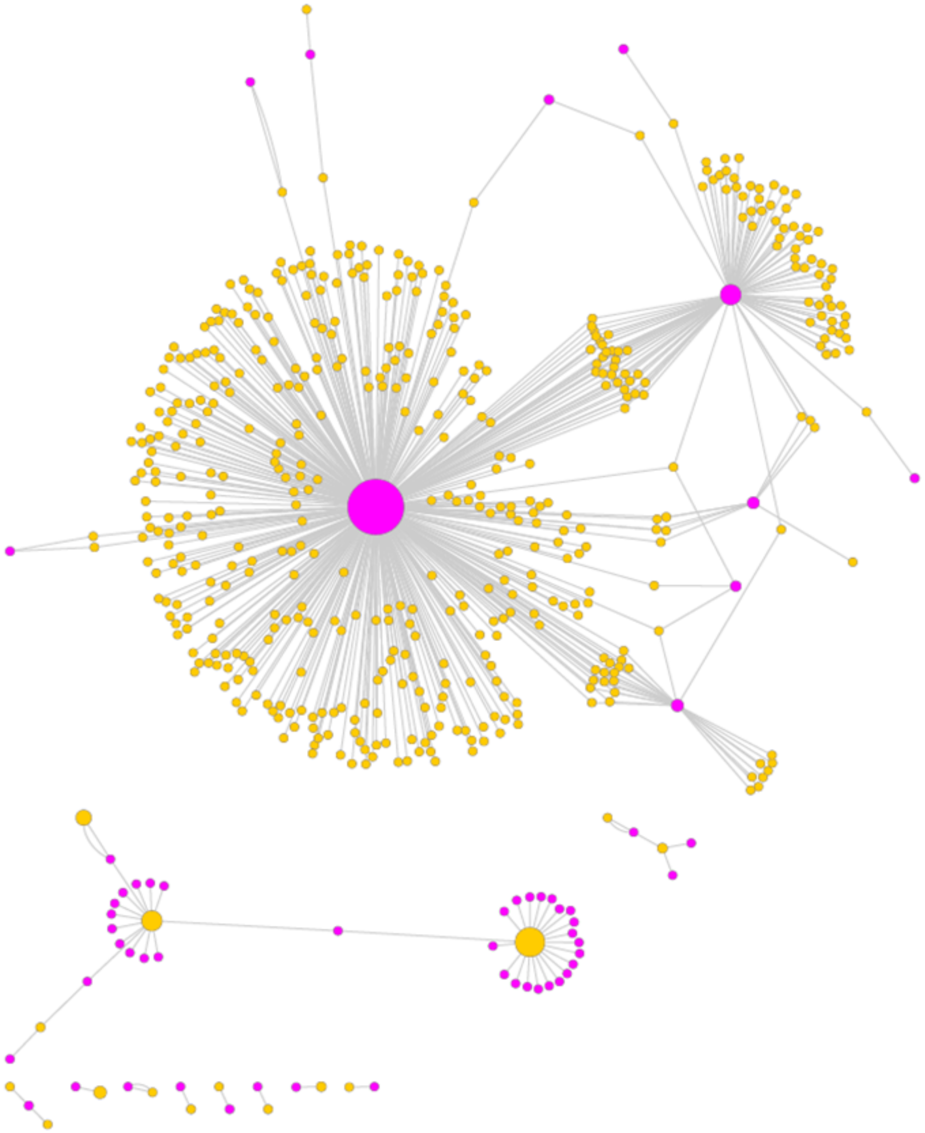


Figure S2: **ERC-Gene family Co-evolution Network.** Network of significant overlaps between ERC modules (pink nodes) and gene family correlation modules (yellow nodes). Edges represent significant overlaps between the ERC modules and gene family modules.

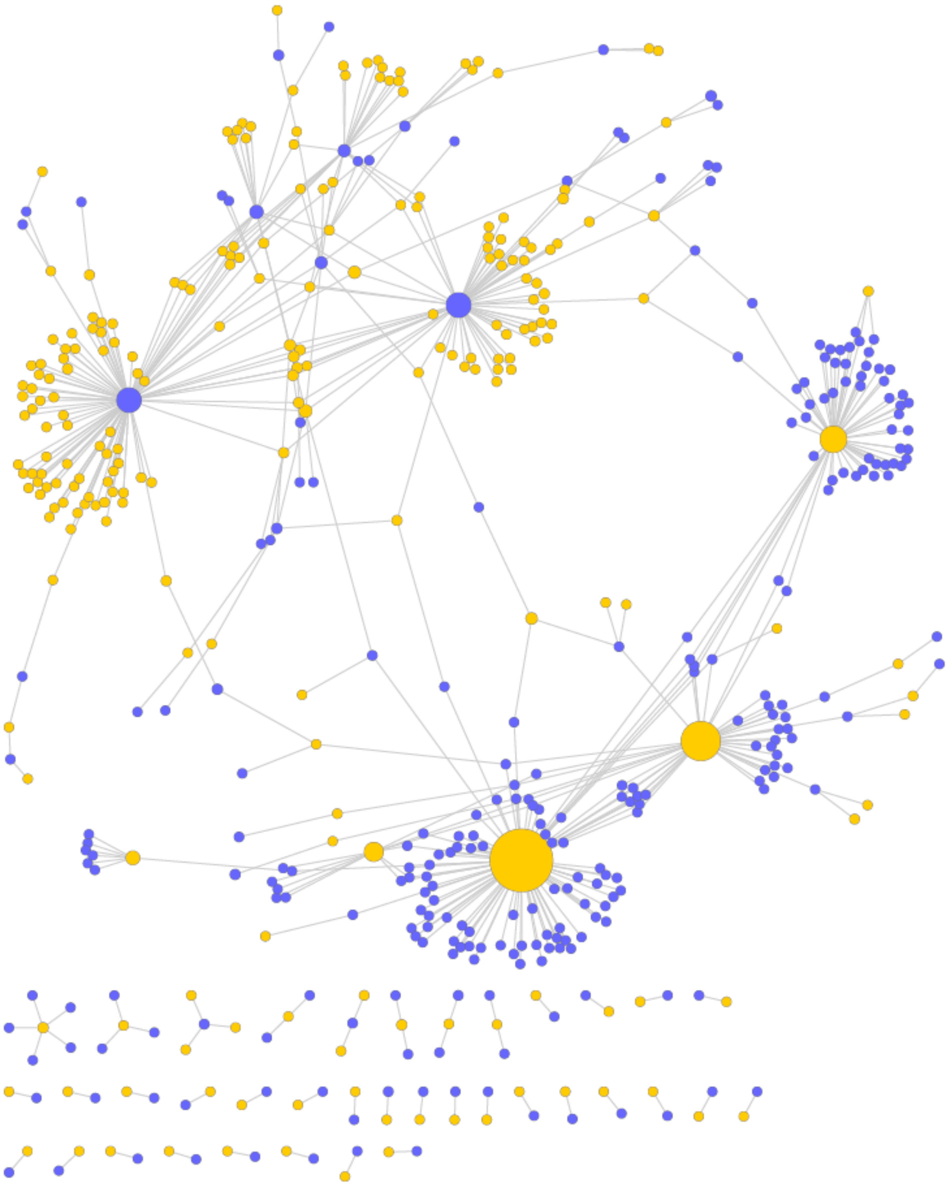


Figure S3: **Co-expression-Gene family Co-evolution Network.** Network of significant overlaps between co-expression modules (blue nodes) and gene family correlation modules (yellow nodes). Edges represent significant overlaps between the co-expression modules and gene family modules

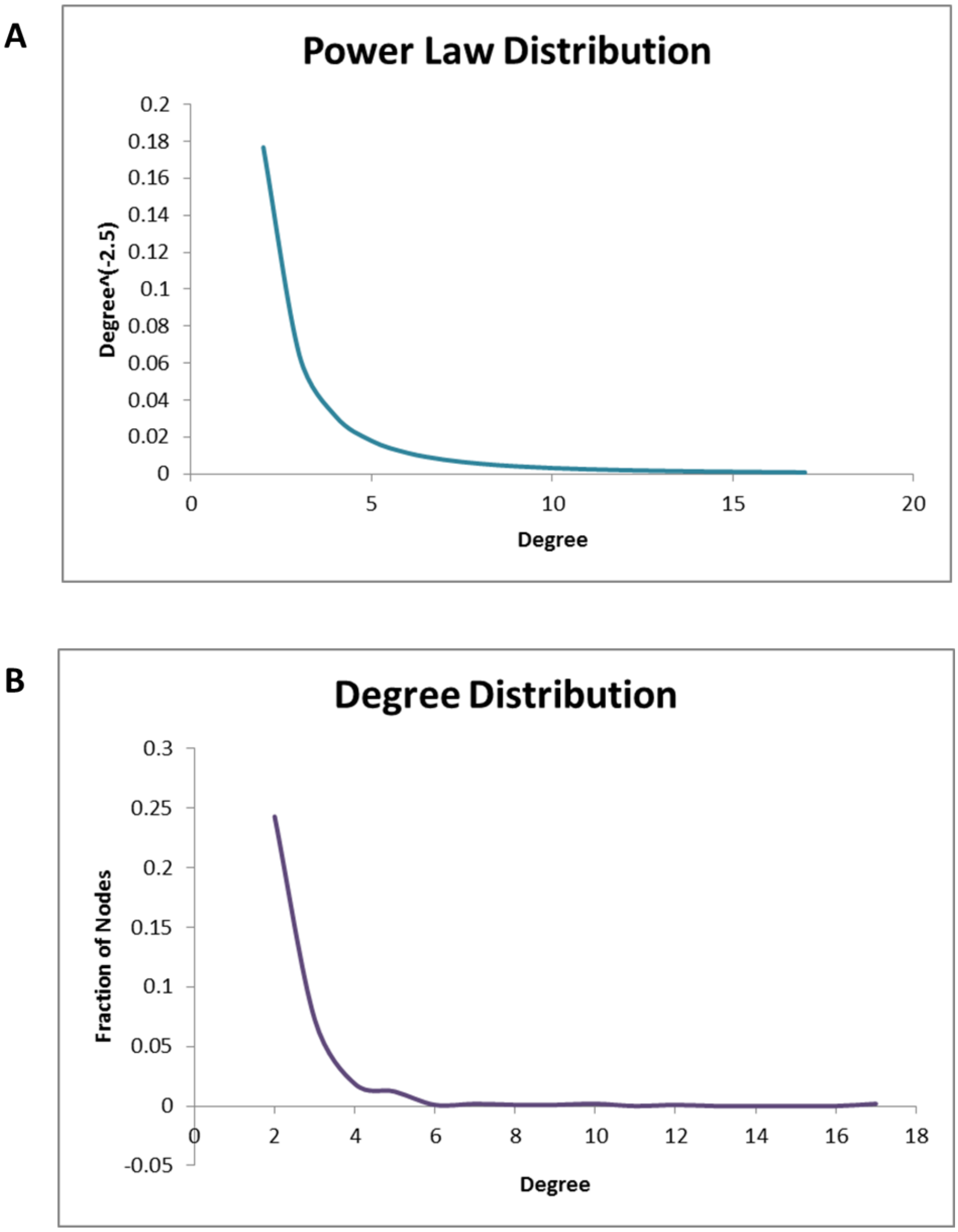


Figure S4: **Distributions** (A) Power Law Distrbution (B) Degree Distribution for the combined co-evolution network. A property of scale-free networks is that their degree distribution follows a power-law distribution. This figure illustrates that the combined co-evolution network is scale-free, since its degree distribution is similar to the power law distribution.

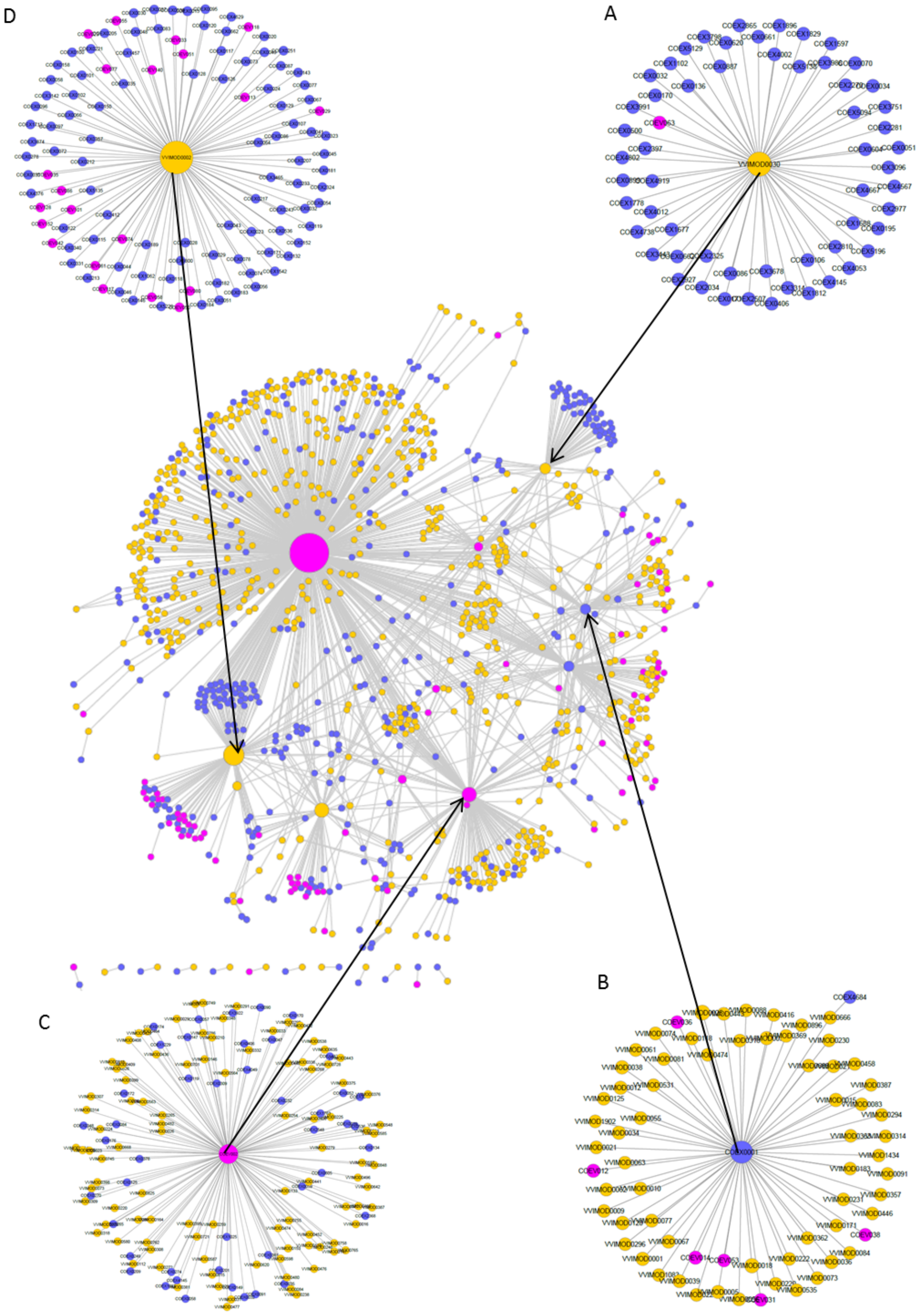


Figure S5: **Subnetworks.** (A) Subnetwork 1 consists of a central gene family correlation module (yellow node) intersecting with several co-expression modules (blue nodes) and one ERC module (pink node). (B) Subnetwork 2 consists of a central co-expression module (blue node) surrounded by several gene family modules (yellow nodes), ERC modules (pink nodes) and one other co-expression module. (C) Subnetwork 3 consists of a central ERC module intersecting with several co-expression modules and gene family modules. (D) Subnetwork 4 consists of a central gene family module surrounded by several co-expression modules and ERC modules.

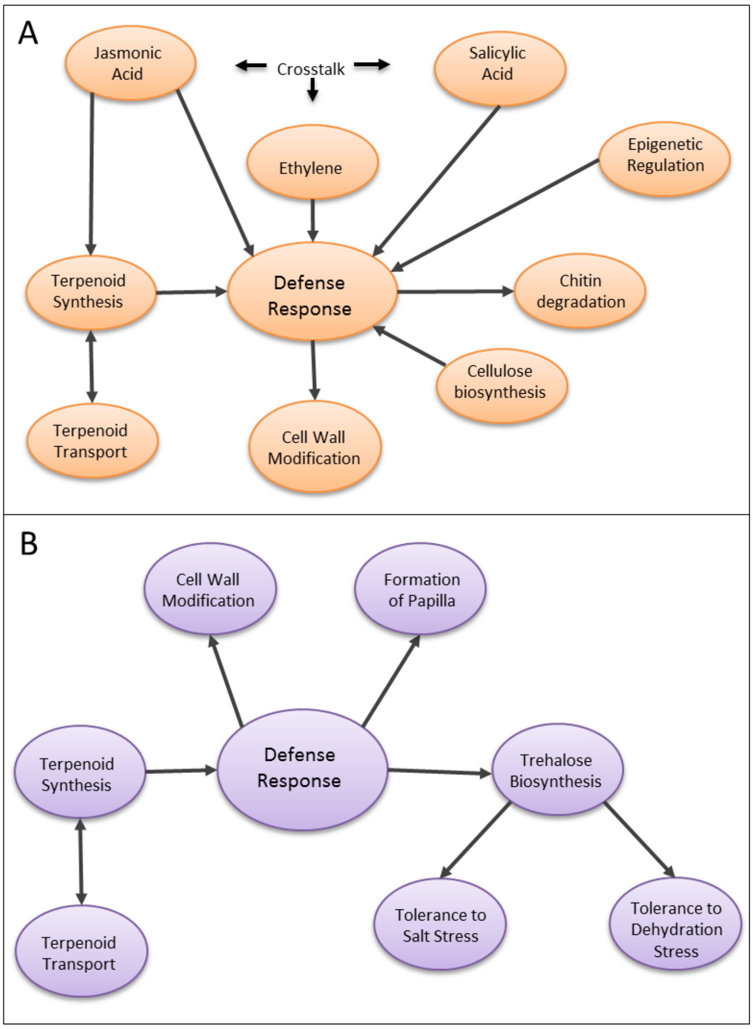


Figure S6: **Co-evolving Functions.** Summary of the related functions which are enriched in (A) subnetwork 2 and (B) subnetwork 4. Arrows indicate relationships for which there is previous literature evidence as referred to in the text.

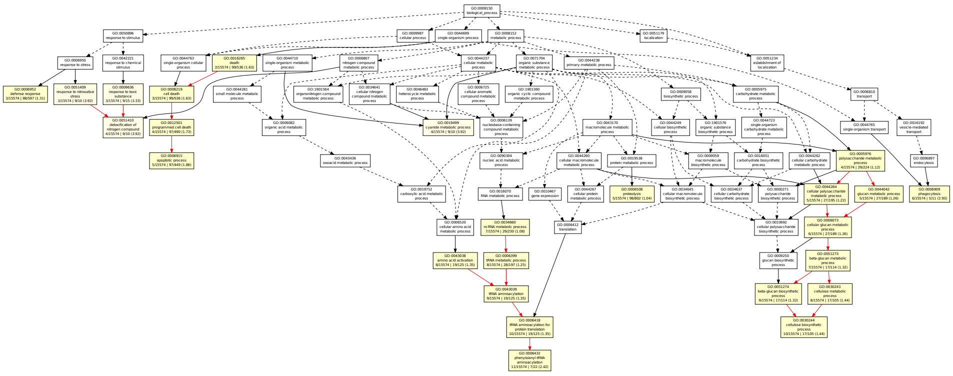


Figure S7: **Node Enrichment View: Subnetwork 1.** GOEAST results for the node enrichment view of subnetwork 1. Yellow rectangles indicate enriched GO terms. Arrows indicate relationships between terms in the Gene Ontology and are red if both terms are enriched, black if one of terms connected by the arrow is enriched or dashed if nether term connected is enriched.

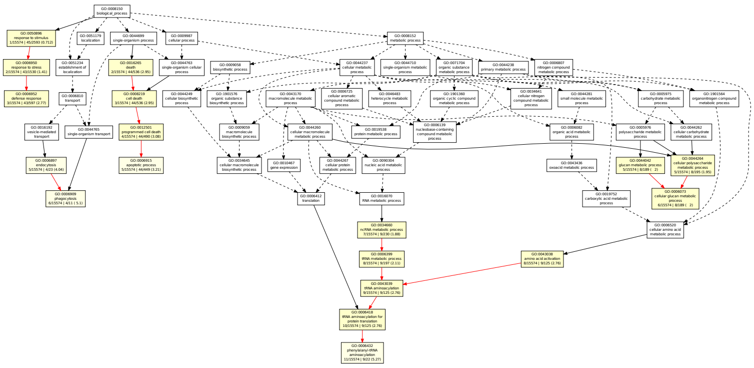


Figure S8: **Edge Enrichment View: Subnetwork 1.** GOEAST results for the node enrichment view of subnetwork 1. Yellow rectangles indicate enriched GO terms. Arrows indicate relationships between terms in the Gene Ontology and are red if both terms are enriched, black if one of terms connected by the arrow is enriched or dashed if nether term connected is enriched.

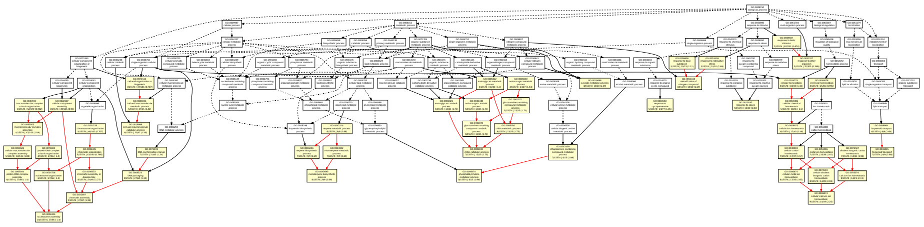


Figure S9: **Node Enrichment View: Subnetwork 2.** GOEAST results for the node enrichment view of subnetwork 2. Yellow rectangles indicate enriched GO terms. Arrows indicate relationships between terms in the Gene Ontology and are red if both terms are enriched, black if one of terms connected by the arrow is enriched or dashed if nether term connected is enriched.

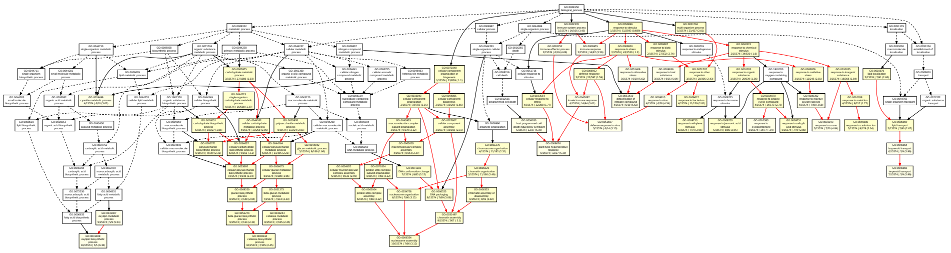


Figure S10: **Edge Enrichment View: Subnetwork 2.** GOEAST results for the node enrichment view of subnetwork 2. Yellow rectangles indicate enriched GO terms. Arrows indicate relationships between terms in the Gene Ontology and are red if both terms are enriched, black if one of terms connected by the arrow is enriched or dashed if nether term connected is enriched.

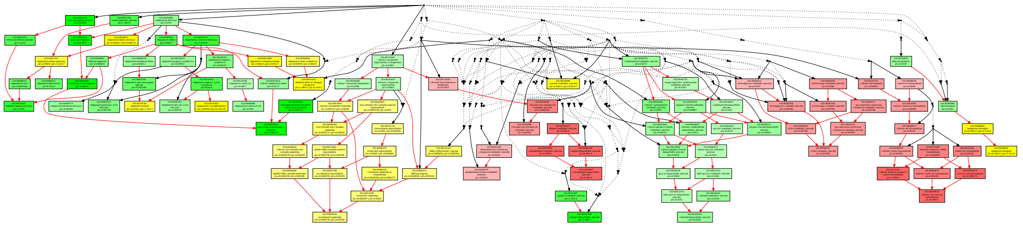


Figure S11: **Combined Enrichment View: Subnetwork 2.** Multi-GOEAST results combining the node enrichment view and edge enrichment view of subnetwork 2. Yellow rectangles represent GO terms which are enriched in both the edge and the node view, green rectangles represent GO terms only enriched in the edge view and red rectangles represent GO terms only enriched in the node view. Arrows indicate relationships between terms in the Gene Ontology and are red if both terms are enriched, black if one of terms connected by the arrow is enriched or dashed if nether term connected is enriched.

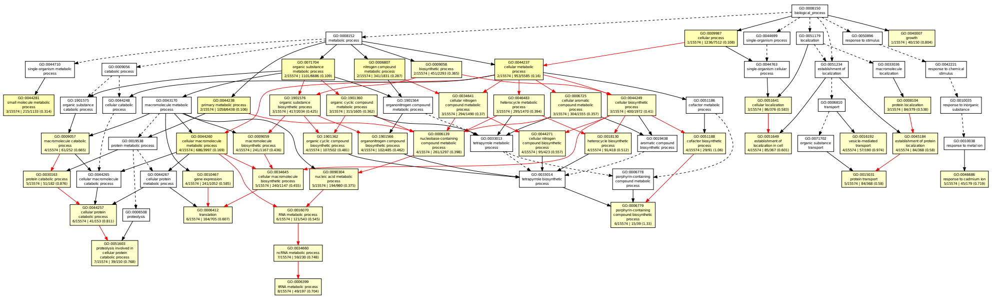


Figure S12: **Node Enrichment View: Subnetwork 3.** GOEAST results for the node enrichment view of subnetwork 3. Yellow rectangles indicate enriched GO terms. Arrows indicate relationships between terms in the Gene Ontology and are red if both terms are enriched, black if one of terms connected by the arrow is enriched or dashed if nether term connected is enriched.

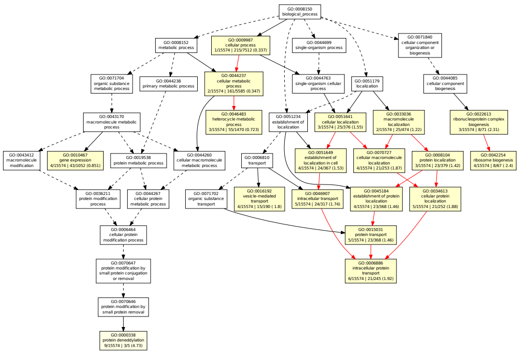

GO:0008150 biological_process
GO:0008152 metabolic process
GO:0009987 cellular process 1/15574 | 215/7512 (0.337)
GO:0044699 single-organism process
GO:0071840 cellular component organization or biogenesis
GO:0071704 organic substance metabolic process
GO:0044238 primary metabolic process
GO:0044237 cellular metabolic process 2/15574 | 161/5585 (0.347)
GO:0044763 single-organism cellular process
GO:0051179 localization
GO:0044085 cellular component biogenesis
GO:0043170 macromolecule metabolic process
GO:0046483 heterocycle metabolic process 3/15574 | 55/1470 (0.723)
GO:0051234 establishment of localization
GO:0051641 cellular localization 3/15574 | 25/376 (1.55)
GO:0033036 macromolecule localization 2/15574 | 25/474 (1.22)
GO:0022613 ribonucleoprotein complex biogenesis 3/15574 | 8/71 (2.31)
GO:0043412 macromolecule modification
GO:0010467 gene expression 4/15574 | 43/1052 (0.851)
GO:0019538 protein metabolic process
GO:0044260 cellular macromolecule metabolic process
GO:0006810 transport
GO:0051649 establishment of localization in cell 4/15574 | 24/367 (1.53)
GO:0070727 cellular macromolecule localization 4/15574 | 21/253 (1.87)
GO:0008104 protein localization 3/15574 | 23/379 (1.42)
GO:0042254 ribosome biogenesis 4/15574 | 8/67 (2.4)
GO:0036211 protein modification process
GO:0044267 cellular protein metabolic process
GO:0071702 organic substance transport
GO:0016192 vesicle-mediated transport 4/15574 | 15/190 (1.8)
GO:0046907 intracellular transport 5/15574 | 24/317 (1.74)
GO:0045184 establishment of protein localization 4/15574 | 23/368 (1.46)
GO:0034613 cellular protein localization 5/15574 | 21/252 (1.88)
GO:0006464 cellular protein modification process
GO:0015031 protein transport 5/15574 | 23/368 (1.46)
GO:0070647 protein modification by small protein conjugation or removal
GO:0006886 intracellular protein transport 6/15574 | 21/245 (1.92)
GO:0070646 protein modification by small protein removal
GO:0000338 protein deneddylation 9/15574 | 3/5 (4.73)

Figure S13: **Edge Enrichment View: Subnetwork 3.** GOEAST results for the node enrichment view of subnetwork 3. Yellow rectangles indicate enriched GO terms. Arrows indicate relationships between terms in the Gene Ontology and are red if both terms are enriched, black if one of terms connected by the arrow is enriched or dashed if nether term connected is enriched.

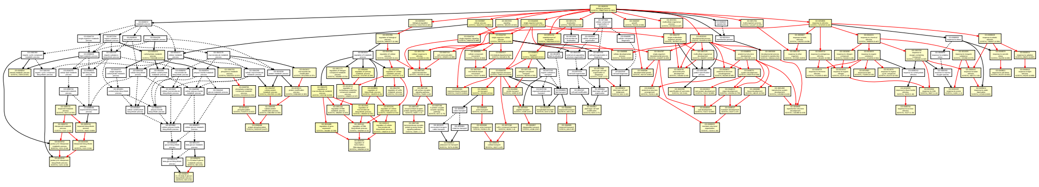


Figure S14: **Node Enrichment View: Subnetwork 4.** GOEAST results for the node enrichment view of subnetwork 4. Yellow rectangles indicate enriched GO terms. Arrows indicate relationships between terms in the Gene Ontology and are red if both terms are enriched, black if one of terms connected by the arrow is enriched or dashed if nether term connected is enriched.

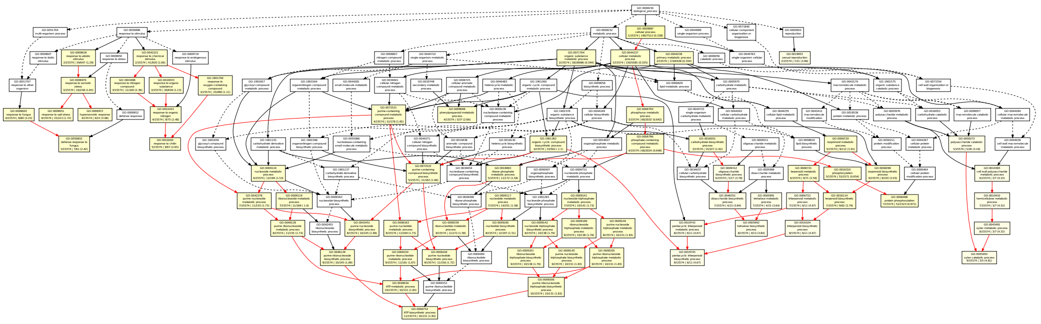


Figure S15: **Edge Enrichment View: Subnetwork 4.** GOEAST results for the node enrichment view of subnetwork 4. Yellow rectangles indicate enriched GO terms. Arrows indicate relationships between terms in the Gene Ontology and are red if both terms are enriched, black if one of terms connected by the arrow is enriched or dashed if nether term connected is enriched.

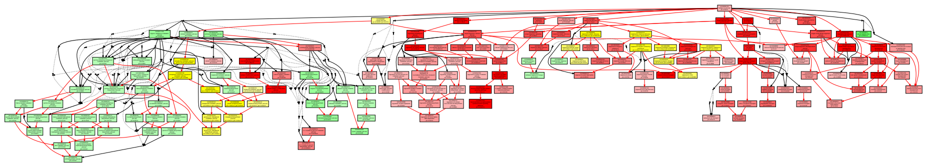


Figure S16: **Combined Enrichment View: Subnetwork 4.** Multi-GOEAST results combining the node enrichment view and edge enrichment view of subnetwork 4. Yellow rectangles represent GO terms which are enriched in both the edge and the node view, green rectangles represent GO terms only enriched in the edge view and red rectangles represent GO terms only enriched in the node view. Arrows indicate relationships between terms in the Gene Ontology and are red if both terms are enriched, black if one of terms connected by the arrow is enriched or dashed if nether term connected is enriched.

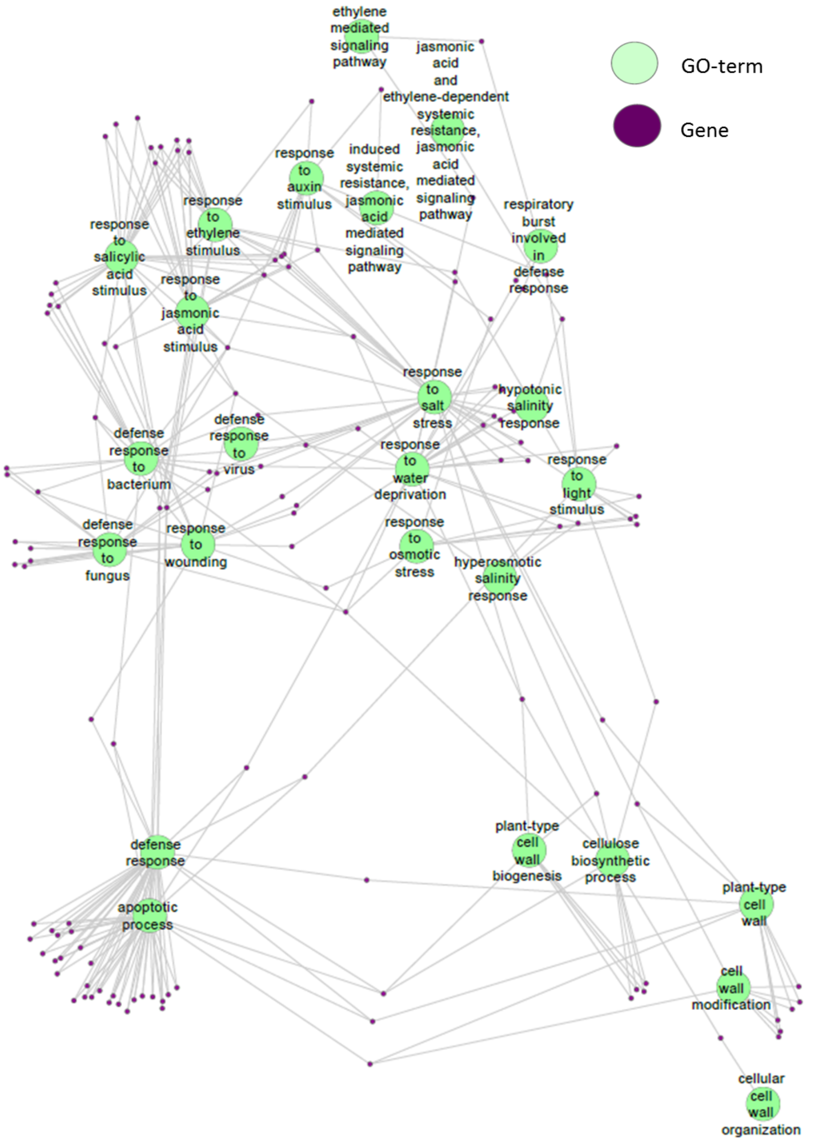

GO-term
Gene
ethylene mediated signaling pathway
jasmonic acid and ethylene-dependent systemic resistance, jasmonic acid mediated signaling pathway
response to auxin stimulus
induced systemic resistance, jasmonic acid mediated signaling pathway
response to ethylene stimulus
response to salicylic acid stimulus
respiratory burst involved in defense response
response to jasmonic acid stimulus
response to salt stress
hypotonic salinity response
defense response to bacterium
defense response to virus
response to water deprivation
response to light stimulus
defense response to fungus
response to wounding
response to osmotic stress
hyperosmotic salinity response
defense response
plant-type cell wall biogenesis
cellulose biosynthetic process
apoptotic process
plant-type cell wall
cell wall modification
cellular cell wall organization

Figure S17: **Hormone Crosstalk Quasi-Pathway** Network of genes present in subnetwork 11 (purple nodes) which are connected to at least 2 of selected GO-terms (light green nodes). This network indicates the crosstalk between biotic and abiotic stress responses through hormone signalling on a gene level, as suggested by the enrichment in subnetwork 4.

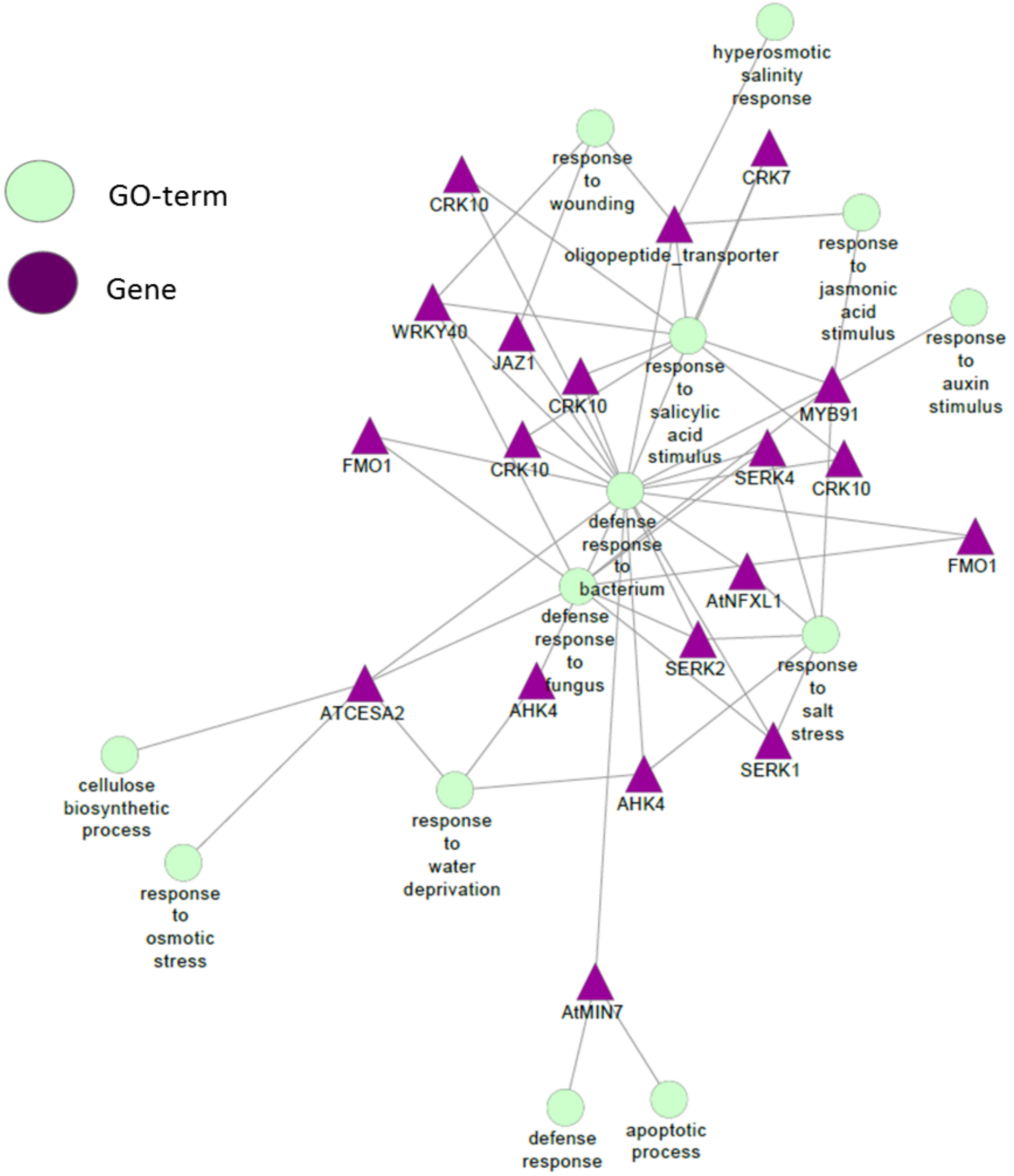

GO-term
Gene
hyperosmotic salinity response
response to wounding
CRK7
CRK10
oligopeptide_transporter
response to jasmonic acid stimulus
response to auxin stimulus
WRKY40
JAZ1
CRK10
response to salicylic acid stimulus
MYB91
FMO1
CRK10
SERK4
CRK10
defense response to bacterium
FMO1
AtNFXL1
defense response to fungus
SERK2
response to salt stress
ATCESA2
AHK4
SERK1
cellulose biosynthetic process
response to water deprivation
AHK4
response to osmotic stress
AtMIN7
defense response
apoptotic process

Figure S18: **Breadth first search** Subnetwork of Figure S17, constructed by selecting all nodes within a breadth first search of length 2 from the node "defense response to bacteria".

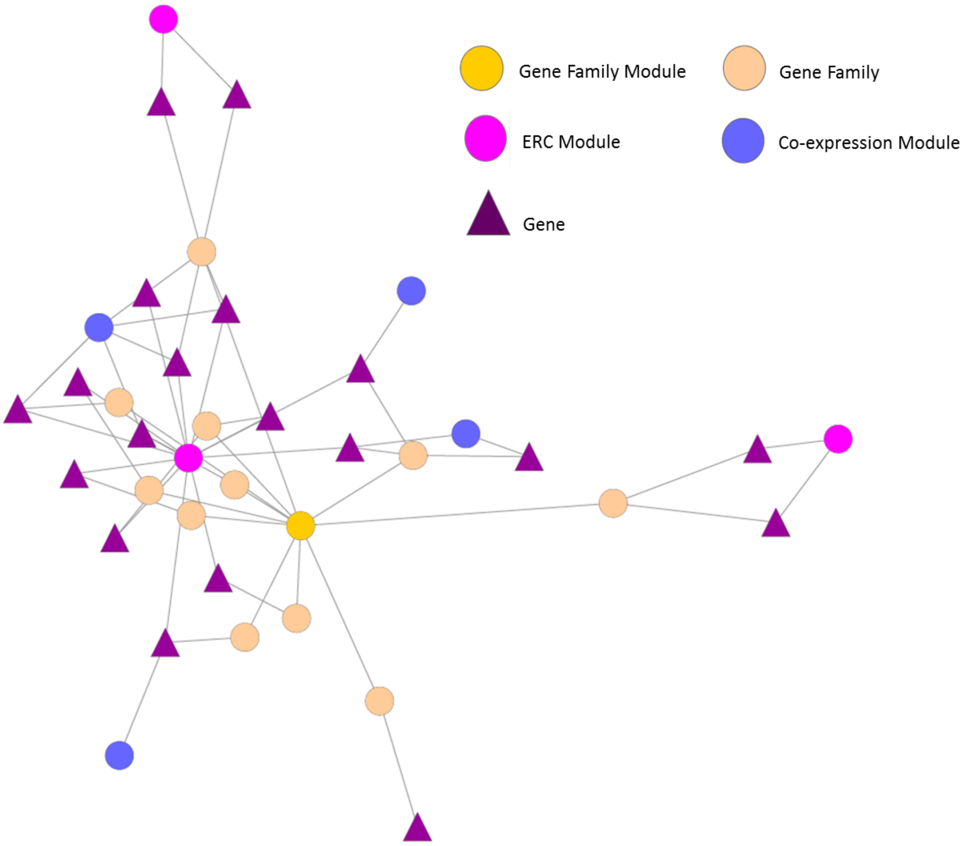


Figure S19: **Crosstalk Gene-Module Network** Network of genes (purple nodes) from Figure S18 connected to co-evolution modules in which they are present. In the case of gene family modules, genes are connected to the gene families in which they are present, and the gene families are connected to the gene family modules in which they are present. Yellow nodes represent gene family modules, light orange nodes represent gene families, blue nodes represent co-expression modules an pink nodes represent ERC modules.

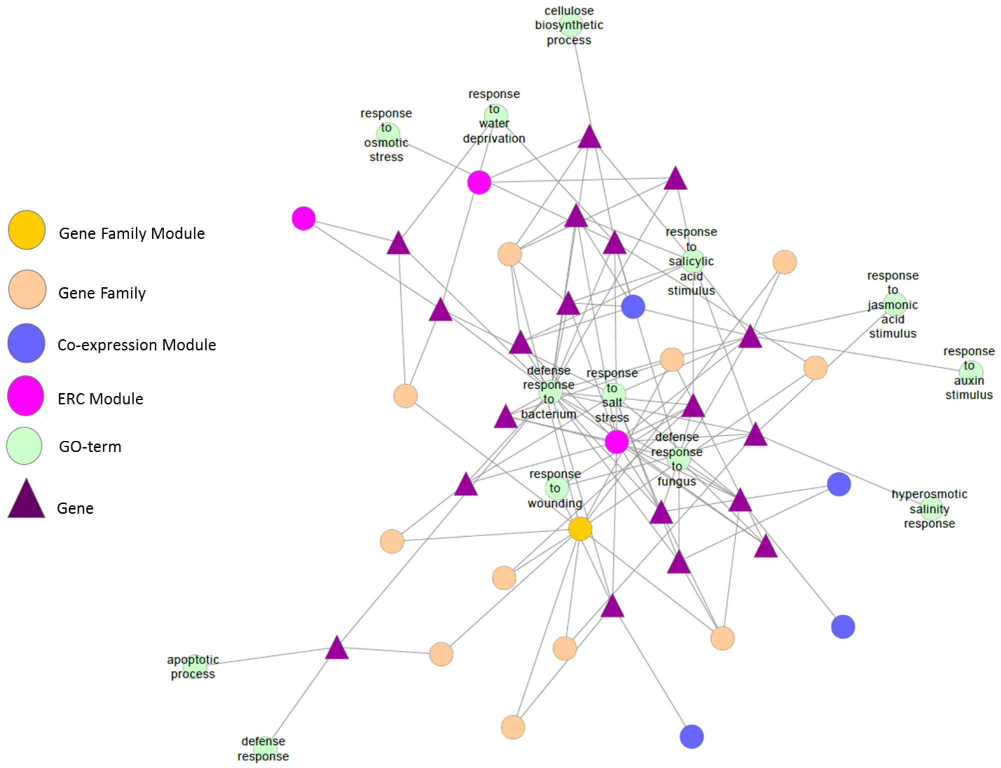

cellulose biosynthetic process
response to water deprivation
response to osmotic stress
Gene Family Module
Gene Family
Co-expression Module
ERC Module
GO-term
Gene
response to salicylic acid stimulus
response to jasmonic acid stimulus
response to auxin stimulus
defense response to bacterium
response to salt stress
defense response to fungus
response to wounding
hyperosmotic salinity response
apoptotic process
defense response

Figure S20: **Merged Crosstalk Network** Merged gene-go network from Figure S18 and gene-module network from Figure S19. Purple nodes represent genes, light green nodes represent GO-terms, yellow nodes represent gene family modules, light orange nodes represent gene families, blue nodes represent co-expression modules an pink nodes represent ERC modules

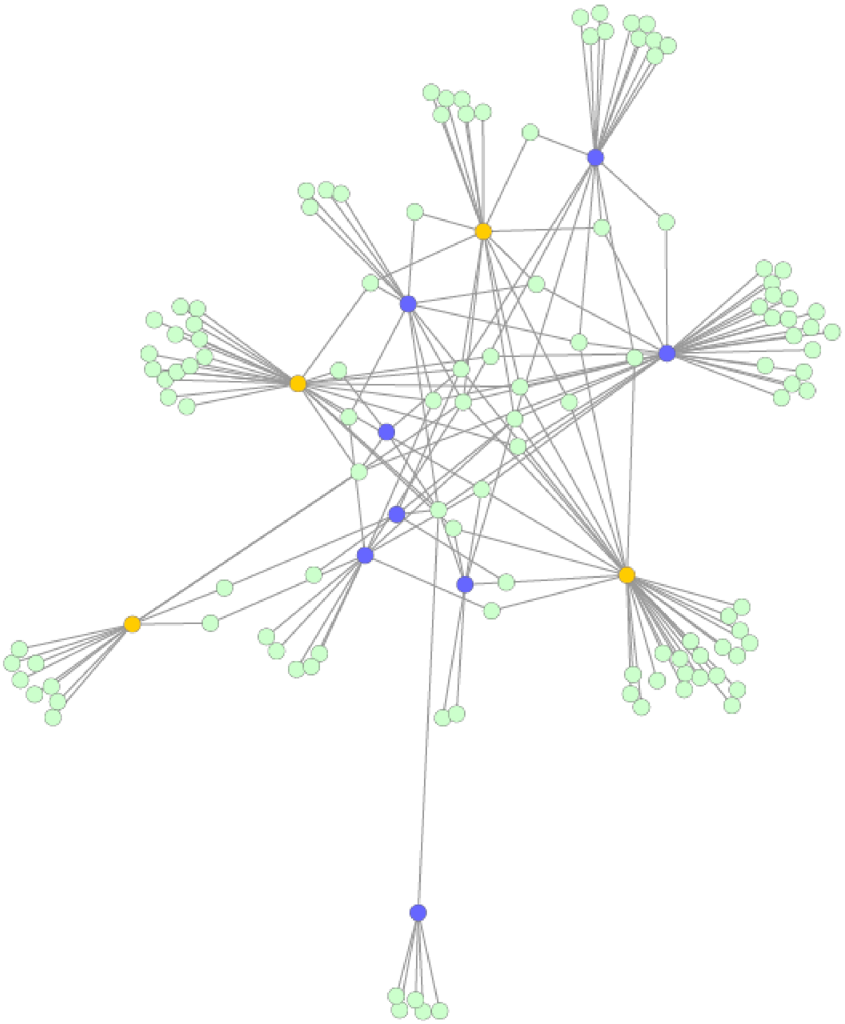


Figure S21: **Linking Nodes.** Network of nodes linking subnetworks 1, 2 and 4. Yellow and blue nodes represent co-expression and gene-family modules respectively, and green nodes represent GO-terms.

Table S1: **Plant Genomes.** List of plant genomes used for gene family construction, and their associated three letter code for OrthoMCL analysis, common names and genome download source.

Species Name	Code	Common Name	Source
Lotus japonicus	lja	Legume	Plaza
Medicago truncatula	mtr	Nitrogen Fixating Legume	Plaza
Glycine max	gma	Soybean	Plaza
Malus domestica	mdo	Apple	Plaza
Fragaria vesca	fve	Strawberry	Plaza
Manihot esculenta	mes	Cassava	Plaza
Ricinus communis	rco	Caster Oil Plant	Plaza
Populus trichocarpa	ptr	Black Cottonwood Tree	Plaza
Arabidopsis thaliana	ath	Arabidopsis thaliana	Plaza
Arabidopsis lyrata	aly	Arabidopsis lyrate	Plaza
Carica papaya	cpa	Papaya	Plaza
Theobroma cacao	tca	Cocoa Tree	Plaza
Vitis vinifera	vvi	Grapevine	Plaza
Oryza sativa ssp. japonica	osa	Rice	Plaza
Oryza sativa ssp. indica	osi	Rice	Plaza
Brachypoium distachyon	bdi	Grass	Plaza
Sorghum bicolor	sbi	Sorghum	Plaza
Zea mays	zma	Maize	Plaza
Selaginella moellendorffii	smp	Lycophyte	Plaza
Physcomitrella patens	ppa	Moss	Plaza
Ostreococcus lucimarinus	olu	Algae	Plaza
Ostreococcus tauri	ota	Algae	Plaza
Micromonas sp. RCC299	mrc	Algae (picophytoplankton)	Plaza
Volvox carteri	vca	Algae	Plaza
Chlamydomonas reinhardtii	cre	Algae	Plaza
Solanum tuberosum	pot	Potato	SGR

Bibliography

[1] Roubelakis-Angelakis, K.A.: *Grapevine Molecular Physiology and Biotechnology*. Springer, 2009.

[2] McGovern, P.E., Hartung, U., Badler, V.R., Glusker, D.L. and Exner, L.J.: The beginnings of winemaking and viniculture in the ancient near east and egypt. *Expedition*, vol. 39, no. 1, pp. 3–21, 1997.

[3] Zharkikh, A., Troggio, M., Pruss, D., Cestaro, A., Eldrdge, G., Pindo, M., Mitchell, J.T., Vezzulli, S., Bhatnagar, S., Fontana, P. *et al.*: Sequencing and assembly of highly heterozygous genome of *Vitis vinifera* L. cv Pinot Noir: Problems and solutions. *Journal of Biotechnology*, vol. 136, no. 1, pp. 38–43, 2008.

[4] Carroll, S.B.: Evolution at two levels: on genes and form. *PLoS Biology*, vol. 3, no. 7, p. e245, 2005.

[5] Nei, M. and Rooney, A.: Concerted and birth-and-death evolution of multigene families. *Annual Review of Genetics*, vol. 39, p. 121, 2005.

[6] Zhang, J.: Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, vol. 18, no. 6, pp. 292–298, 2003.

[7] Chamary, J. and Hurst, L.D.: Evidence for selection on synonymous mutations affecting stability of mrna secondary structure in mammals. *Genome Biology*, vol. 6, no. 9, p. R75, 2005.

[8] Walker, A.R., Lee, E., Bogs, J., McDavid, D.A., Thomas, M.R. and Robinson, S.P.: White grapes arose through the mutation of two similar and adjacent regulatory genes. *The Plant Journal*, vol. 49, no. 5, pp. 772–785, 2007.

[9] Rossouw, D., Jacobson, D. and Bauer, F.F.: Transcriptional regulation and the diversification of metabolism in wine yeast strains. *Genetics*, vol. 190, no. 1, pp. 251–61, January 2012.

[10] Pazos, F. and Valencia, A.: Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Engineering*, vol. 14, no. 9, pp. 609–614, 2001.

[11] Goh, C.-S., Bogan, A.A., Joachimiak, M., Walther, D. and Cohen, F.E.: Coevolution of proteins with their interaction partners. *Journal of Molecular Biology*, vol. 299, no. 2, pp. 283–293, 2000.

[12] Sato, T., Yamanishi, Y., Kanehisa, M. and Toh, H.: The inference of protein–protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics*, vol. 21, no. 17, pp. 3482–3489, 2005.

[13] Clark, N.L., Alani, E. and Aquadro, C.F.: Evolutionary rate covariation reveals shared functionality and coexpression of genes. *Genome Research*, vol. 22, no. 4, pp. 714–720, 2012.

[14] Montague, M.G. and Hutchison, C.a.: Gene content phylogeny of herpesviruses. *PNAS*, vol. 97, no. 10, pp. 5334–9, 2000.

[15] Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O.: Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *PNAS*, vol. 96, no. 8, pp. 4285–4288, 1999.

[16] Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. and Eisenberg, D.: A combined algorithm for genome-wide prediction of protein function. *Nature*, vol. 402, no. 6757, pp. 83–86, 1999.

[17] Movahedi, S., Van Bel, M., Heyndrickx, K.S. and Vandepoele, K.: Comparative co-expression analysis in plant biology. *Plant, Cell & Environment*, vol. 35, pp. 1787–1798, 2012.

[18] Handel, A., Regoes, R.R. and Antia, R.: The role of compensatory mutations in the emergence of drug resistance. *PLoS Computational Biology*, vol. 2, no. 10, p. e137, October 2006. ISSN 1553-7358.

[19] Rampey, R.a., Baldridge, M.T., Farrow, D.C., Bay, S.N. and Bartel, B.: Compensatory mutations in predicted metal transporters modulate auxin conjugate responsiveness in Arabidopsis. *G3 (Bethesda, Md.)*, vol. 3, no. 1, pp. 131–41, January 2013. ISSN 2160-1836.

[20] van Dijk, A.D.J. and van Ham, R.C.H.J.: Conserved and variable correlated mutations in the plant MADS protein network. *BMC Genomics*, vol. 11, no. 1, p. 607, January 2010. ISSN 1471-2164.

[21] Li, L., Stoeckert, C. and Roos, D.: Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome Research*, vol. 13, no. 9, pp. 2178–2189, 2003.

[22] Real, R. and Vargas, J.M.: The probabilistic basis of jaccard's index of similarity. *Systematic Biology*, vol. 45, no. 3, pp. 380–385, 1996.

[23] Shannon, P., Markiel, A., Ozier, O., Baliga, N., Wang, J., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.

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

[25] Zheng, Q. and Wang, X.-J.: Goeast: a web-based software toolkit for gene ontology enrichment analysis. *Nucleic Acids Research*, vol. 36, no. suppl 2, pp. W358–W363, 2008.

[26] Bradley, D.J., Kjellbom, P. and Lamb, C.J.: Elicitor-and wound-induced oxidative cross-linking of a proline-rich plant cell wall protein: a novel, rapid defense response. *Cell*, vol. 70, no. 1, pp. 21–30, 1992.

[27] Hématy, K., Cherk, C. and Somerville, S.: Host–pathogen warfare at the plant cell wall. *Current Opinion in Plant Biology*, vol. 12, no. 4, pp. 406–413, 2009.

[28] Cano-Delgado, A., Penfield, S., Smith, C., Catley, M. and Bevan, M.: Reduced cellulose synthesis invokes lignification and defense responses in Arabidopsis thaliana. *The Plant Journal*, vol. 34, no. 3, pp. 351–362, May 2003. ISSN 0960-7412.

[29] Hernández-Blanco, C., Feng, D.X., Hu, J., Sánchez-Vallet, A., Deslandes, L., Llorente, F., Berrocal-Lobo, M., Keller, H., Barlet, X., Sánchez-Rodríguez, C., Anderson, L.K., Somerville, S., Marco, Y. and Molina, A.: Impairment of cellulose synthases required for Arabidopsis secondary cell wall formation enhances disease resistance. *The Plant Cell*, vol. 19, no. 3, pp. 890–903, March 2007. ISSN 1040-4651.

[30] Fritig, B., Heitz, T. and Legrand, M.: Antimicrobial proteins in induced plant defense. *Current Opinion in Immunology*, vol. 10, no. 1, pp. 16–22, 1998.

[31] Punja, Z.K. and Zhang, Y.-y.: Plant chitinases and their roles in resistance to fungal diseases. *Journal of Nematology*, vol. 25, no. 4, p. 526, 1993.

[32] Derksen, H., Rampitsch, C. and Daayf, F.: Signaling cross-talk in plant disease resistance. *Plant Science*, 2013.

[33] Martin, D.M., Gershenzon, J. and Bohlmann, J.: Induction of volatile terpene biosynthesis and diurnal emission by methyl jasmonate in foliage of norway spruce. *Plant Physiology*, vol. 132, no. 3, pp. 1586–1599, 2003.

[34] Aharoni, A., Jongsma, M.A. and Bouwmeester, H.J.: Volatile science? metabolic engineering of terpenoids in plants. *Trends in Plant Science*, vol. 10, no. 12, pp. 594–602, 2005.

[35] Giuliano, G., Tavazza, R., Diretto, G., Beyer, P. and Taylor, M.A.: Metabolic engineering of carotenoid biosynthesis in plants. *Trends in Biotechnology*, vol. 26, no. 3, pp. 139–145, 2008.

[36] Martin, D.M., Aubourg, S., Schouwey, M.B., Daviet, L., Schalk, M., Toub, O., Lund, S.T. and Bohlmann, J.: Functional annotation, genome organization and phylogeny of the grapevine (vitis vinifera) terpene synthase gene family based on genome assembly, flcdna cloning, and enzyme assays. *BMC Plant Biology*, vol. 10, no. 1, p. 226, 2010.

[37] Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C. *et al.*: The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, vol. 449, no. 7161, pp. 463–467, 2007.

[38] Chinnusamy, V. and Zhu, J.-K.: Epigenetic regulation of stress responses in plants. *Current Opinion in Plant Biology*, vol. 12, no. 2, pp. 133–139, 2009.

[39] Bonifacino, J.S. and Lippincott-Schwartz, J.: Coat proteins: shaping membrane transport. *Nature Reviews Molecular Cell Biology*, vol. 4, no. 5, pp. 409–414, 2003.

[40] Pedrazzini, E., Giovinazzo, G., Bielli, A., de Virgilio, M., Frigerio, L., Pesca, M., Faoro, F., Bollini, R., Ceriotti, A. and Vitale, A.: Protein quality control along the route to the plant vacuole. *The Plant Cell Online*, vol. 9, no. 10, pp. 1869–1880, 1997.

[41] Collinge, D.B.: Cell wall appositions: the first line of defence. *Journal of Experimental Botany*, vol. 60, no. 2, pp. 351–352, 2009.

[42] Underwood, W.: The plant cell wall: a dynamic barrier against pathogen invasion. *Frontiers in Plant Science*, vol. 3, 2012.

[43] Fernandez, O., Béthencourt, L., Quero, A., Sangwan, R.S. and Clément, C.: Trehalose and plant stress responses: friend or foe? *Trends in Plant Science*, vol. 15, no. 7, pp. 409–417, 2010.

[44] Iordachescu, M. and Imai, R.: Trehalose biosynthesis in response to abiotic stresses. *Journal of Integrative Plant Biology*, vol. 50, no. 10, pp. 1223–1229, 2008.

[45] Fujita, M., Fujita, Y., Noutoshi, Y., Takahashi, F., Narusaka, Y., Yamaguchi-Shinozaki, K. and Shinozaki, K.: Crosstalk between abiotic and biotic stress responses: a current view from the points of convergence in the stress signaling networks. *Current Opinion in Plant Biology*, vol. 9, no. 4, pp. 436–442, 2006.

[46] Cheong, Y.H., Chang, H.-S., Gupta, R., Wang, X., Zhu, T. and Luan, S.: Transcriptional profiling reveals novel interactions between wounding, pathogen, abiotic stress, and hormonal responses in arabidopsis. *Plant Physiology*, vol. 129, no. 2, pp. 661–677, 2002.

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

[48] Horvath, S. and Dong, J.: Geometric interpretation of gene coexpression network analysis. *PLoS Computational Biology*, vol. 4, no. 8, p. e1000117, 2008.

[49] Reijneveld, J.C., Ponten, S.C., Berendse, H.W. and Stam, C.J.: The application of graph theoretical analysis to complex networks in the brain. *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology*, vol. 118, no. 11, pp. 2317–31, 2007.

[50] Jacobson, D. and Emerton, G.: GSA-PCA: gene set generation by principal component analysis of the Laplacian matrix of a metabolic network. *BMC Bioinformatics*, vol. 13, no. 1, p. 197, August 2012. ISSN 1471-2105.

[51] Proost, S., Van Bel, M., Sterck, L., Billiau, K., Van Parys, T., Van de Peer, Y. and Vandepoele, K.: Plaza: a comparative genomics resource to study gene and genome evolution in plants. *The Plant Cell Online*, vol. 21, no. 12, pp. 3718–3731, 2009.

[52] van Dongen, S.: *Graph clustering by flow simulation*. Ph.D. thesis, University of Utrecht, 2000.

[53] Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, vol. 4, no. 2, pp. 249–264, 2003.

[54] Edgar, R.: Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[55] Felsenstein, J.: Phylip - phylogeny inference package (version 3.2). *Cladistics*, vol. 5, no. 3, pp. 164–166, 1989.

[56] Holm, S.: A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, , no. 6, pp. 65–70, 1979.

Chapter 4

3-way Networks: Application of Hypergraphs for Modelling Increased Complexity in Comparative Genomics

D. Weighill and D. Jacobson

As submitted for publication in PLoS Computational Biology, currently undergoing peer review.

4.1 Abstract

We present and develop the theory 3-way networks, a type of hypergraph in which each edge models relationships between triplets of objects as opposed to pairs of objects as done by standard network models. We explore approaches of how to prune these 3-way networks, illustrate their utility in comparative genomics and demonstrate how they find relationships which would be missed by standard 2-way network models using a phylogenomic dataset of 211 bacterial genomes.

4.2 Author Summary

Genomes contain the information underlying the molecular functions of an organism. One way to compare the entire genomes of different organisms is to compare their gene-family content profiles which is effectively a comparison of their functional potential. Standard networks, when used to model phylogenomic similarities, are not capable of capturing some of the underlying complexity of the relationships between genomes. In order to address this, we have developed a new three-way similarity metric and constructed three-way networks modelling the relationships between 211 bacterial genomes. We find that such three-way networks find cross-species genomic similarities that would have been otherwise missed by simpler models such as standard networks.

4.3 Introduction

Network models are a useful reductionist approach for modelling complex systems. Networks involve representing a collection of objects as nodes, and representing relationships between those objects as edges. Thus, networks model a system in a pairwise manner, breaking a system down into individual parts (nodes), modelling relationships between pairs of these individual parts (edges) and then reconstructing the system as a network [1]. However, modelling a system based on only pairwise relationships biases the model against more complex relationships that may exist in the system. To this end, we introduce a new ternary network definition, namely 3-way networks based on the concept of hypergraphs. 3-way networks model the relationships between triplets of objects instead of pairs of objects. The concept of calculating the similarity between objects three at a time is not a novel concept [2; 3; 4] and general hypergraphs [5] have previously been used in certain areas of biology, including metabolic modelling, gene expression and RNA interaction studies [6; 7; 8; 9; 10]. However, to our knowledge, this is the first time that the concept of 3-way networks has been applied in the field of comparative genomics.

In this study, we develop the theory around 3-way networks in terms of abstract definition, weighting 3-way networks and pruning 3-way networks. We develop a new 3-way metric for the weighting of 3-way edges. We then apply a 3-way network model to a set of 211 bacterial genomes, modelling the similarities between the bacteria on a whole genome scale, (based on gene family content), and compare the resulting 3-way networks to those obtained using standard 2-way network models.

4.4 Results and Discussion

4.4.1 Definition of 3-way Networks

A network, or graph, G is an ordered pair, defined as

$$G = (V, E) \tag{4.4.1}$$

where $V = \{v_1, v_2, ..., v_n\}$ is a set of n nodes and $E = \{e_1, e_2, ..., e_m\}$ is a set of m edges [11]. In this case, nodes represent a certain set of objects of interest and edges can be interpreted as relationships between these objects. In particular, edges represent pairwise relationships and thus are defined (for an undirected network) as pairs of nodes. With the aim of modelling higher order relationships than simply pairwise relationships, we define 3-way networks as network models of ternary relationships, i.e. relationships between triplets of objects. 3-way networks are defined by replacing the previous definition of an edge as a set of 2 nodes by a set of 3 nodes. Thus a 3-way network is a type of hypergraph [5]. This can be formalized with the following definition:

Definition 1. *A 3-way network is a graph* $G = (V, E)$ *where* $V = \{v_1, v_2, ..., v_n\}$ *is the set of nodes and* $E = \{e_1, e_2, ..., e_m\}$ *is the set of edges. Each edge* e_i *is defined as a set of 3 nodes,* $e_i = \{v_x, v_y, v_z\}$ *where* $x, y, z \in \{1, 2, 3, ..., m\}$.

Graphically, each 3-way edge is a line connecting 3 nodes, which can be interpreted as a relationship between 3 objects. An example of a 3-way network with 5 nodes, $V = \{v_1, v_2, v_3, v_4, v_5\}$ and 2 edges, $E = \{e_1, e_2\} = \{\{v_1, v_2, v_3\}, \{v_3, v_4, v_5\}\}$ is shown in Figure 4.1a.

4.4.2 Weighted 3-way Networks

4.4.2.1 3-way Sørensen Index

In a normal 2-way network, each edge can be assigned a weight indicating the strength of the relationship between the two nodes the edge is connecting. This concept can easily be extended to a 3-way network, in which an edge weight will indicate the strength of the relationship between the 3 nodes the edge is connecting. For a 3-way network, this requires a similarity metric which quantifies the similarity between 3 objects at a time. Assuming that each object is represented by a vector, a similarity metric which quantifies the similarity between 3 vectors is needed. The Sørensen Index [12] is a similarity metric which quantifies the overlap between the features of pairs of objects. Let X and Y be two objects and let each object be viewed as a set of features. The Sørensen Index $S(X, Y)$ is defined as:

$$S_2(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|} \tag{4.4.2}$$

where $|X|$ is the number of features of object X, $|Y|$ is the number of features of object Y and $|X \cap Y|$ is interpreted as the number of features shared by object X and object Y [13]. If objects are represented by vectors, the Sørensen Index between two vectors X and Y can be expressed as:

$$S_2(X, Y) = \frac{2\sum_i \min(X_{Bi}, Y_{Bi})}{\sum_i (X_{Bi} + Y_{Bi})} \tag{4.4.3}$$

where X_B and Y_B are binary vectors defined as:

$$X_{Bi} = \begin{cases} 1 & \text{if} X_i \geq 1 \\ 0 & \text{if} X_i = 0 \end{cases} \tag{4.4.4}$$

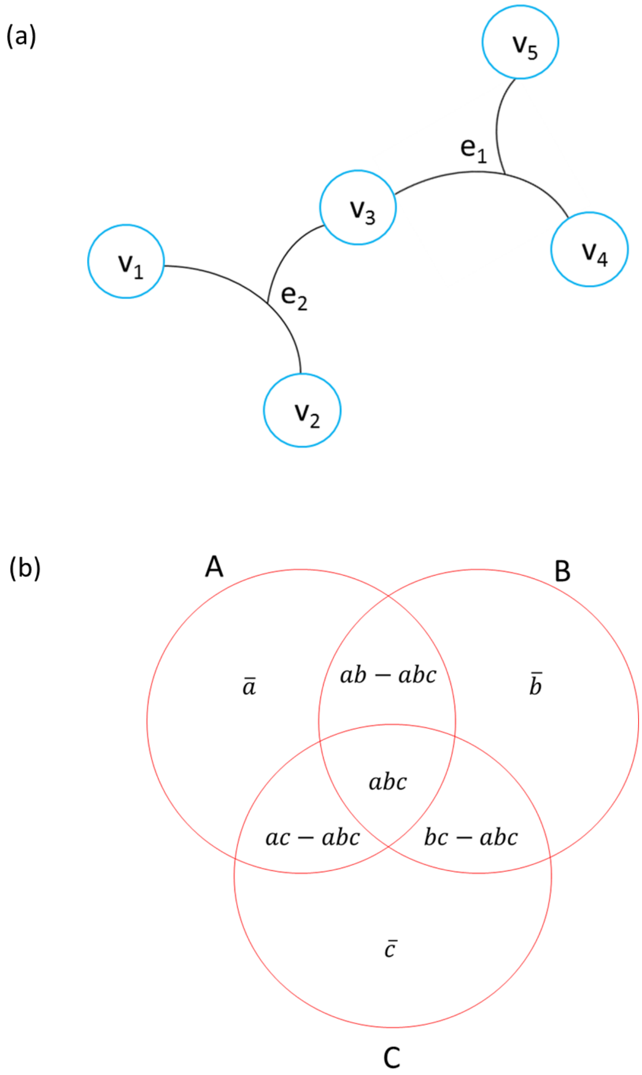


Figure 4.1: **3-way Edges and Intersections** (a) A small, 3-way network consisting of 5 nodes v_1, v_2, v_3, v_4 and v_5 and two 3-way edges e_1 and e_2. Edge e_1 connects nodes v_3, v_4 and v_5 and edge e_2 connects nodes v_1, v_2 and v_3. (b) Venn diagram for a 3-way intersection of species. a is the number of families present in species A, b is the number of families present in species B, c is the number of families present in species C, ab is the number of families present in species A and species B, ac is the number of families present in species A and species C, bc is the number of families present in species B and species C, abc is the number of families present in species A, B and C, $\bar{a}$ is the number of families present only in species A, $\bar{b}$ is the number of families present only in species B and $\bar{c}$ is the number of families present only in species C.

$$Y_{Bi} = \begin{cases} 1 & \text{if } Y_i \geq 1 \\ 0 & \text{if } Y_i = 0 \end{cases} \tag{4.4.5}$$

An extension of the Sørensen Index exists for calculating the similarity between triplets of objects. This metric was originally developed for quantifying the similarity between the species content of different biological samples [2]. Generally, for each triplet of objects A, B, and C, each represented by a vector, the three-way Sørensen index can be defined as:

$$S_3(ABC) = \frac{3}{2}\left(\frac{ab + ac + bc - abc}{a + b + c}\right) \tag{4.4.6}$$

where a is the number of features present in object A, b is the number of features present in object B, c is the number of features present in object C, ab is the number of features present in object A and object B, ac is the number of features present in object A and object C, bc is the number of features present in object B and object C and abc is the number of features present in object A, B and C [2]. These variables can be visualized on a venn diagram (Figure 4.1b).
The 3-way Sørensen Index can also be expressed in vector format as follows:

$$S_3(X, Y, Z) = \frac{\frac{3}{2}\sum_i (\min(X_{Bi}, Y_{Bi}) + \min(X_{Bi}, Z_{Bi}) + \min(Y_{Bi}, Z_{Bi}) - \min(X_{Bi}, Y_{Bi}, Z_{Bi}))}{\sum_i (X_{Bi} + Y_{Bi} + Z_{Bi})} \tag{4.4.7}$$

4.4.2.2 3-way Czekanowski Index

A quantitative version of the Sørensen Index is called the Czekanowski Index [14]. For two vectors X and Y, the Czekanowski Index is defined as:

$$C_2(X, Y) = \frac{2\sum_i \min(X_i, Y_i)}{\sum_i (X_i + Y_i)} \tag{4.4.8}$$

Notice that the equation is the same as that of the Sørensen Index in vector format, except that the original vectors are used and not binary vectors. The Czekanowski Index thus considers the size of the overlaps between features of an object and not simply the presence or absence of features. Using the same structure as the 3-way Sørensen Index, we extended the Czekanowski Index to a 3-way form. For 3 vectors X, Y and Z, we have defined the 3-way Czekanowski Index between the three vectors as:

$$C_3(X, Y, Z) = \frac{\frac{3}{2}\sum_i (\min(X_i, Y_i) + \min(X_i, Z_i) + \min(Y_i, Z_i) - \min(X_i, Y_i, Z_i))}{\sum_i (X_i + Y_i + Z_i)} \tag{4.4.9}$$

4.4.3 Pruning 3-way Networks

Many approaches used to prune edges from a network such as Maximum Spanning Tree (MST) algorithms and clustering algorithms are designed for standard 2-way networks and are not directly applicable to 3-way networks. However, certain approaches are easily transferable to 3-way networks, namely thresholding and best-edge selection.

4.4.3.1 Thresholding

Thresholding can easily be transferred from a standard 2-way network to a 3-way network. Thresholding is one of the simplest ways to prune any network. A threshold is set and edges with a weight below the chosen threshold are removed. In order to determine a justifiable threshold for a 3-way Sørensen network we have developed the following theorem:

Theorem 1. *If the intersection of three objects abc is zero (i.e. there is no feature present in all three objects), then* $S_3(ABC) \leq \frac{3}{4}$.

Proof. If $abc = 0$, then

$$\begin{aligned} S_3(ABC) &= \frac{3}{2}.\frac{ab+ac+bc-abc}{a+b+c} \\ &= \frac{3}{2}.\frac{ab+ac+bc-abc}{2(ab+ac+bc)+\bar{a}+\bar{b}+\bar{c}} \\ &= \frac{3}{2}.\frac{ab+ac+bc}{2(ab+ac+bc)+\bar{a}+\bar{b}+\bar{c}} \end{aligned}$$

There are two cases to consider.

Case 1: If $\bar{a}$, $\bar{b}$ and $\bar{c}$ are all equal to 0, then

$$\begin{aligned} S_3(ABC) &= \frac{3}{2}.\frac{ab+ac+bc}{2(ab+ac+bc)+\bar{a}+\bar{b}+\bar{c}} \\ &= \frac{3}{2}.\frac{ab+ac+bc}{2(ab+ac+bc)} \\ &= \frac{3}{2}.\frac{1}{2} \\ &= \frac{3}{4} \end{aligned}$$

Thus if $abc = 0$ and $\bar{a}$, $\bar{b}$ and $\bar{c}$ are all equal to 0 then $S_3(ABC) = \frac{3}{4}$.

Case 2: If $\bar{a}$, $\bar{b}$ and $\bar{c}$ are all greater than zero 0 (they cannot be less than zero, since there cannot be a negative number of features associated with an object), then

$$\begin{aligned} 2(ab+ac+bc)+\bar{a}+\bar{b}+\bar{c} &> 2(ab+ac+bc) \\ \text{Therefore, } S_3(ABC) &= \frac{3}{2}.\frac{ab+ac+bc}{2(ab+ac+bc)+\bar{a}+\bar{b}+\bar{c}} \\ &< \frac{3}{2}.\frac{ab+ac+bc}{2(ab+ac+bc)} \\ &= \frac{3}{4} \end{aligned}$$

Thus if $abc = 0$ and $\bar{a}$, $\bar{b}$ and $\bar{c}$ are all greater than zero 0, $S_3(ABC) < \frac{3}{4}$. Combining these two cases, we can conclude that if $abc = 0$, $S_3(ABC) \leq \frac{3}{4}$. □

A similar thresholding strategy can be adopted for the 3-way Czekanowski Index. We need the following:

Lemma 1. *Given integers a, b and c, the following relation holds:*

$$\min(a,b)+\min(a,c)-\min(a,b,c) \leq a \qquad (4.4.10)$$

We now prove a theorem similar to Theorem 1, but relating to the 3-way Czekanowski Index.

Theorem 2. *Given 3 species X, Y, and Z, if there is no gene family present in all 3 species, then $C_3(XYZ) \leq \frac{3}{4}$.*

Proof. If there is no gene family present in all 3 species X, Y and Z, then $\sum_i \min(X_i, Y_i.Z_i) = 0$. Therefore,

$$\begin{aligned} C_3(X,Y,Z) &= \frac{\frac{3}{2}\sum_i(\min(X_i,Y_i)+\min(X_i,Z_i)+\min(Y_i,Z_i)-\min(X_i,Y_i,Z_i))}{\sum_i(X_i+Y_i+Z_i)} \\ &= \frac{\frac{3}{2}\sum_i(\min(X_i,Y_i)+\min(X_i,Z_i)+\min(Y_i,Z_i))-\sum_i(\min(X_i,Y_i,Z_i))}{\sum_i(X_i+Y_i+Z_i)} \\ &= \frac{\frac{3}{2}\sum_i(\min(X_i,Y_i)+\min(X_i,Z_i)+\min(Y_i,Z_i))}{\sum_i(X_i+Y_i+Z_i)} \end{aligned}$$

Using Lemma 1, this can be expanded as:

$$\begin{aligned} C_3(X,Y,Z) &= \frac{\frac{3}{2}\sum_i(\min(X_i,Y_i)+\min(X_i,Z_i)+\min(Y_i,Z_i))}{\sum_i(X_i+Y_i+Z_i)} \\ &\leq \frac{\frac{3}{2}\sum_i(\min(X_i,Y_i)+\min(X_i,Z_i)+\min(Y_i,Z_i))}{2\sum_i(\min(X_i,Y_i)+\min(X_i,Z_i)+\min(Y_i,Z_i))} \\ &= \frac{3}{4} \end{aligned}$$

Thus if $\sum_i \min(X_i, Y_i.Z_i) = 0$, then $C_3(X,Y,Z) \leq \frac{3}{4}$. □

Thus the minimum justifiable threshold for 3-way Sørensen and 3-way Czekanowski networks is 0.75.

4.4.3.2 Best Edges

Another simple way to prune a network is to select for each node, the best x edges connected to that node, i.e. select the x edges with the highest weight for each node. This is easily done by taking a list of all edges connected to a given node, ranking them by weight from highest to lowest, and then selecting the top x edges. This approach does not depend on the definition of the edge. It is directly transferable from the concept of a 2-way network to the concept of a 3-way network.

4.4.4 Phylogenomic Networks of Bacterial Genomes

Gene families were calculated across a dataset consisting of 211 bacterial genomes using TribeMCL [15] and gene family content profiles constructed for each bacterial species. Various phylogenomic 2-way similarity, 3-way similarity and gene family enrichment networks were then constructed in order to investigate the relationships between the bacterial species based on gene family content and to compare the effect of 3-way networks as opposed to 2-way networks. These networks are described below. In each network, nodes represent bacterial species and edges represent similarities between species based on 2-way or 3-way similarity between their gene family content profiles, or represent connections between species based on shared gene family enrichment.

4.4.4.1 3-way and 2-way Sørenesen Networks

The concept of 3-way networks was developed in order to attempt to model more complex relationships that would otherwise be missed by pairwise relationships. To this end, the definition of an edge was extended to represent a ternary relationship, i.e. a relationship between 3 nodes. In order to quantify these ternary relationships, a 3-way similarity metric was chosen, namely the Sørensen Index. This allowed “high order similarities" or similarities between more than two species to contribute to our interpretation. The 3-way Sørensen Index was used to quantify the similarity between all triplets of bacterial species, based on their gene family content. Applying a threshold of 0.76 allowed us to select for edges which we were sure had a contributing factor of the 3-way intersection and not simply a high intersection between pairs of species (See Theorem 1). This thresholded network can be seen in Figure S1. Large coloured nodes represent bacterial species and the combination of the small white nodes and the grey 2-way edges represent 3-way edges. Certain genera were selected and those bacterial species nodes coloured according to genus. (The default node colour was grey, thus grey nodes are not all in the same genus). The 3-way network was also pruned by selecting only the best and second best edge for each node. This best-edge 3-way Sørensen network can be seen in Figure 4.2.

Networks were also constructed using the standard 2-way Sørensen Index and pruned using a best edge approach and a Maximum Spanning Tree (MST) approach. For the best edge approach, the best and second best edges were selected for each node. The resulting network is shown in Figure 4.3a.

A Maximum Spanning Tree is a useful approach for sparsifying a network by isolating the 'backbone' of the network as the shortest tree spanning all nodes which has maximum weight. The Sørensen MST can be seen in Figure 4.3b.

The 3-way networks in Figure 4.2 and Figure S1 have an interesting structure. In each network, nodes of the same colour group together, indicating that the genera group together well. The network shown in Figure 4.2 especially seems to show an interesting middle ground between connectedness and modularity. There are generally many connections within genera, but also some connections between genera. In contrast to this is the 2-way Sørensen MST shown in Figure 4.3b. This view of the network shows no modularity as the genera do seem to group together, but there are no connections within the genera indicating how similar the species within genera are. The 2-way Sørensen best edge network (Figure 4.3a) was constructed by selecting only the best and second best edges for each node from the standard 2-way Sørensen network. It would appear that this 2-way best edge network is overly sparse, and does not give much information about the connectedness between genera. It would seem that the genera are also not as well grouped as in the 3-way best-edge network.

4.4.4.2 3-way and 2-way Czekanowski Networks

A new 3-way metric was developed called the 3-way Czekanowski Index. It is an extension of the standard 2-way Czekanowski Index [16] in the same way that the 3-way Sørensen Index [2] is an extension of the original 2-way Sørensen Index [13]. A 3-way network was constructed using the 3-way Czekanowski Index and pruned in the same way described above for the 3-way Sørensen network. The thresholded 3-way Czekanowski network and the best-edge 3-way Czekanowski network can be seen in Figures S3 and 4.4 respectively. Networks were also constructed using the standard 2-way Czekanowski Index and can be seen in Figure's 4.5a and 4.5b.

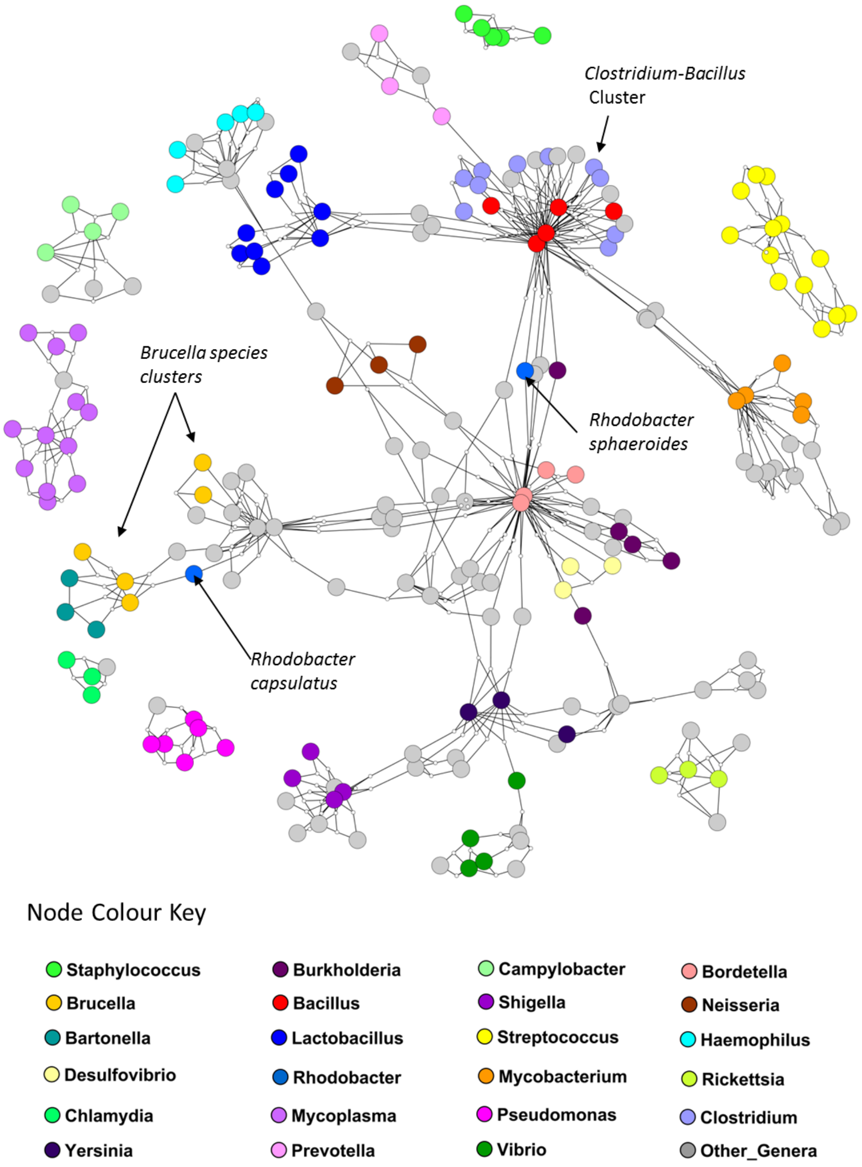


Figure 4.2: **Best-Edges 3-way Sørensen Network**. 3-way Sørensen network pruned by selecting the best and second best edge for each node. Nodes represent bacterial species and edges represent similarity between triplets of bacterial species based on gene family content, quantified using the 3-way Sørensen Index. Nodes are coloured according to genus. Default colour is grey.

4.4.4.3 Gene Family Enrichment Networks

In order to get another perspective on the relationships between the bacteria species based on gene families, a gene family enrichment network was constructed (Figure 4.6). In this network, large, coloured nodes represent bacterial species and small white nodes represent gene families which are enriched in more than one species as determined using Fisher's Exact Test [17]. Each gene family node is connected to the species in which the gene family is enriched. It can clearly be seen that the genera group together well in this network. Shared enriched families thus seem to be a competent measure of species similarity. This network also allows us to target gene families which seem to be distinguishing characteristics of small groups of species.

4.4.4.4 Network Comparison

The 3-way Sørensen networks often support the interpretations of the 2-way networks. However, in some cases, the 3-way networks give new information which differs from that of the 2-way networks. A selection of examples will be discussed.

***Clostridium-Bacillus* Cluster** The cluster of red and light blue nodes in the 3-way Sørensen network (Figure 4.2) and the 3-way Czekanowski network (Figure 4.4) consist of *Clostridium* species (light blue nodes) and *Bacillus* species (red nodes). Figure 4.7a and 4.7b show subnetworks containing these two clusters, and it is clear that, in both the Sørensen 3-way network and the Czekanowski 3-way network, there are a number of 3-way edges connecting species within and between those two genera. When looking at the same two genera in the 2-way Sørensen and 2-way Czekanowski networks (Figures 4.3a, 4.3b, 4.5a and 4.5b) there is no evidence of any particular link between these 2 genera. In the 2-way Sørensen MST (Figure 4.3b) the two genera are close together, but there are no edges between them. In the 2-way best edge Sørensen network (Figure 4.3a) these two genera are in two completely separate modules, giving no indication whatsoever that they are connected or similar. Similar patterns are seen in the 2-way Czekanowski MST (Figure 4.5b) and the 2-way best edge Czekanowski network (Figure 4.5a). When looking at the shared enriched gene family network (Figure 4.6) the *Clostridium* and *Bacillus* species are topologically close together. The *Clostridium* and *Bacillus* species as well as their neighbouring gene families were selected as a subnetwork from the family enrichment network and can be seen in Figure 4.7c. It is apparent that the *Clostridium* and *Bacillus* species share several enriched gene families. The 3-way Sørensen and 3-way Czekanowski networks seem to be picking up a relationship between the two genera which is not seen in the 2-way networks, which is further supported by the gene family enrichment data.

Gene families which were enriched in both genera, and present in at least 3 species were selected for further analysis. The genes in these gene families were then compared against all *Clostridium* and *Bacillus* proteins in NCBI using BLAST [18; 19]. Many of the genes identified were related to sporulation.

Clostridium and *Bacillus* species are known to sporulate and there is literature evidence for the conservation of various sporulation genes across these two genera [20]. Sporulation is a process which involves the production of endospores, which are dormant and highly resistent to environmental stresses [20]. Examples of genes in these gene families enriched in both *Bacillus* and *Clostridium* species were *abrB* and *gerKA*, which are known to be involved in sporulation in *Bacillus* species [21].

Another gene family enriched in both *Clostridium* and *Bacillus* species contained genes with polysaccharide deacetylase functions, in particular, the gene *pdaB*. There is literature evidence for the requirement of polysaccharide deacetylases for sporulation in *Bacillus subtilis*, in which *pdaB* mutants were unable to properly maintain their spores in the later stages of sporulation [22]. The *pdaA* gene has also been found to be neccesary for spore germination in *B. subtilis* [23]. The enrichment of this family in both *Clostridium* and *Bacillus* species along with the other sporulation families could suggest a similar role of deacetylases in the sporulation of *Clostridium* species.

We also found that another gene family enriched in both *Bacillus* and *Clostridium* species contained genes related to chemotaxis, namely a methyl accepting chemotaxis protein. Chemotaxis and sporulation are oppositely regulated processes and are both regulated by the major sporulation regulating protein Spo0A [24]. Thus, it would appear that even though *Bacillus* and *Clostridium* are quite distant phylogenetically, they share a set of sporulation related protein families which appear to be detected by 3-way networks, and are missed by simpler 2-way networks quantifying only 2-way relationships.

***Brucella* Partitioning** Species in the genus *Brucella* can be found as light orange nodes. In the Sørensen MST and the Czekanowski MST (Figures 4.3b add 4.5b respectively), this genus is split into two groups, one group containing *B. canis*, *B. abortis* and *B. ovis* (Group 1), and the other group containing *B. melitensis* and *B. suis* (Group 2) . These same separate groupings are also seen in the best-edge 3-way Sørensen network (Figure 4.2) and best-edge 3-way Czekanowski network (Figure 4.4). Thus using different 2-way and 3-way similarity metrics, the *Brucella* species partition in the same way. Figure 4.8a and b show the neighbourhoods within one 3-way edge of the *Brucella* species in the best edge Sørensen network and the best edge Czekanowski network respectively. Figure 4.8c is a subnetwork of the enrichment network (Figure 4.6) showing all nodes within a radius of 2 of the *Brucella* nodes. From Figure 4.8 the same groupings of the genus can be observed, thus this separation of the genus can be seen with whole gene family profiles, as well as with gene family enrichment. These groupings are different to the divergence previously found in the *Brucella* genus, in which *B. abortus* clustered nearer to *B. melitensis* and *B. suis* clustered nearer to *B. canis* [25]. These different groupings of the *Brucella* species could be due to the fact that the phylogeny constructed in [25] was based on SNPs and is therefore a point mutation-based view of evolution, whereas our phylogenomic networks are constructed with gene families, and are thus a gene duplication/deletion-based view of evolution.

From Figure 4.8 a and b, it can be seen that both the 3-way Sørensen and 3-way Czekanowski networks group *Brucella ovis*, *Brucella canis* and *Brucella abortus* with members of the *Bartonella genus.* This is supported by the gene family enrichment view in Figure 4.8c. Figure 4.8a and b also suggests a relationship between Group 2 *Brucella* species and *Ochrobactrum anthropi.* This is also seen in the gene family enrichment view. Of the 3-way networks, only the Czekanowski network suggests that Group 2 of *Brucella* species, namely *Brucella suis* and *Brucella melitensis* group together with members of the *Bordetella* genus. This is also seen in the gene family enrichment view in Figure 4.8c. None of the 2-way networks suggested this connection. The 2-way MSTs (Figures 4.3b and 4.5b) show the proximity of Group 1 to the*Bartonella* species and the proximity of Group 2 to *O. anthropi*, however they do not suggest the link between Group 2 *Brucella* species and *Bordetella* species. The 2-way best edge networks (Figures 4.3a and 4.5a) only show the connection between Group 2 and *O. anthropi.* They show none of the relationships suggested by 3-way networks between Group 1 and *Bartonella* species, and Group 2 and *Bordetella* species.

***Rhodobacter* Separation** Consider the genus *Rhodobacter* in the above networks (two medium blue nodes). In the Sørensen MST (Figure 4.3b) these two nodes are neighbours. This is also seen in the best edge Sørensen network (Figure 4.3a). However, in both Czekanowski 2-way networks (Figures 4.5b and 4.5a), these two *Rhodobacter* species are not neighbours. The 3-way Sørensen and 3-way Czekanowski networks (Figures 4.2 and 4.4) place these nodes quite far apart. Figure 4.9a and b show the neighbourhoods within one 3-way edge of *Rhodobacter* species in the 3-way Sørensen network and 3-way Czekanowski network respectively. From this figure, it can be seen that the nodes are in separate neighbourhoods. This is also seen in the enriched family view in Figure 4.9c. This figure shows the species which share at least one enriched family with *Rhodobacter* species. Both Sørensen and Czekanowski best edge 3-way networks thus pick up a separation between the two *Rhodobacter* species which is supported by the gene family enrichment data and not found by the 2-way Sørensen networks.

Combination View: *Rhodobacter* and *Brucella* species A further examination of Figures 4.8 and 4.9 shows that there seem to be overlaps between the *Brucella* groupings in Figure 4.8 and the *Rhodobacter* groupings in Figure 4.9. Figure 4.10 shows the neighbourhood around *Brucella* species and *Rhodobacter species* in (a) the 3-way best edge Czekanowski network and (b) the gene family enrichment network. Group 1 *Brucella* species cluster with *Bartonella* species and *Rhodobacter capsulatus* and Group 2 *Brucella* species cluster with *Bordetella* species, *Ochrobactrum athropi* and *Rhodobacter sphaeroides*. This amount of detail in groupings of species was not found in any of the 2-way networks.

Combined 2-way and 3-way Networks Merging the 3-way best edge Sørensen network (Figure 4.2) and the 2-way Sørensen MST (Figure 4.3b) results in an interesting network which is shown in Figure S4. This network combines the modularity of the 3-way network showing the connections within genera and a few cross-genera connections with the MST which shows the overall connections across genera. This combined 2-way and 3-way Czekanowkski network (Figure S5) was also constructed by merging the 3-way best edge Czekanowski network (Figure 4.4) and the 2-way Czekanowski MST (Figure 4.5b). These combination networks provide an interesting, “best of both worlds” view. They combine the connectedness and simplicity of an MST, which allows for no modularity, but forces all nodes to connected to the network, and the modularity and complex relationships provided by the 3-way networks which show a mixture of within-module connection and inter-module connections, and show relationships missed by standard 2-way networks.

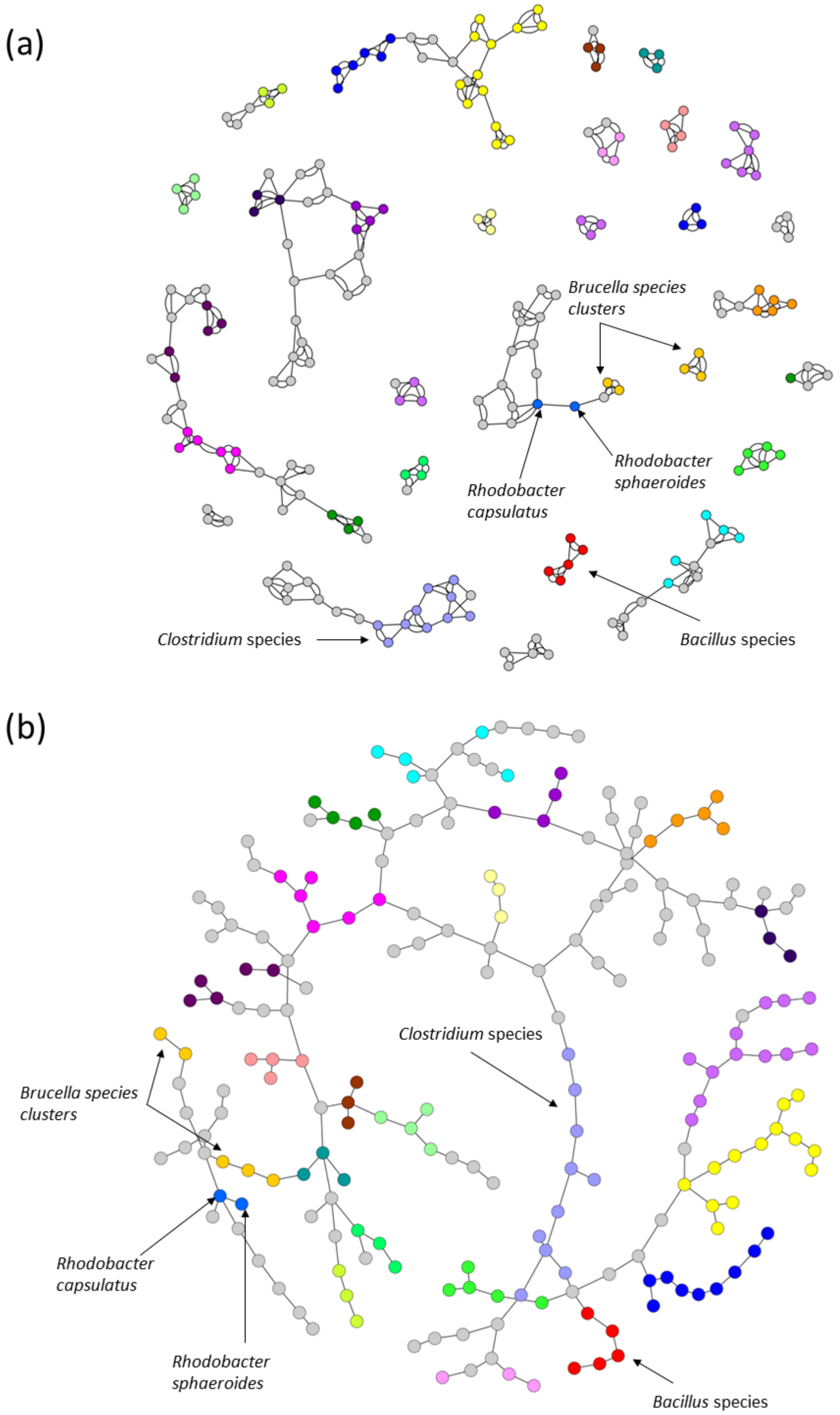


Figure 4.3: **2-way Sørensen Networks** (a) 2-way Sørensen Best Edges Network (b) Maximum Spanning Tree (MST) of the all-vs-all Sørensen network. Nodes represent bacterial species and edges represent similarity between pairs of bacterial species based on gene family content, quantified using the 3-way Sørensen Index. Nodes are coloured according to genus. The same node colour key as in Figure 2 applies.

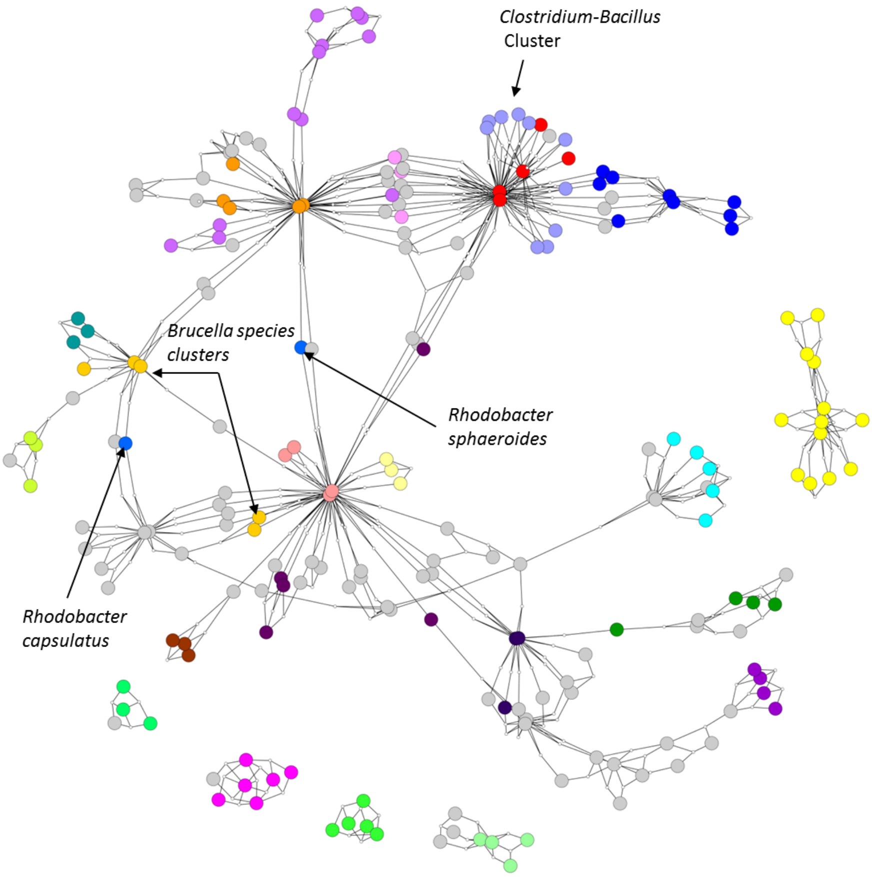


Figure 4.4: **Best-Edges 3-way Czekanowski Network** 3-way Czekanowski network pruned by selecting the best and second best edge for each node. Nodes represent bacterial species and edges represent similarity between triplets of bacterial species based on gene family content, quantified using the 3-way Czekanowski Index. Nodes are coloured according to genus. The same node colour key as in Figure 2 applies.

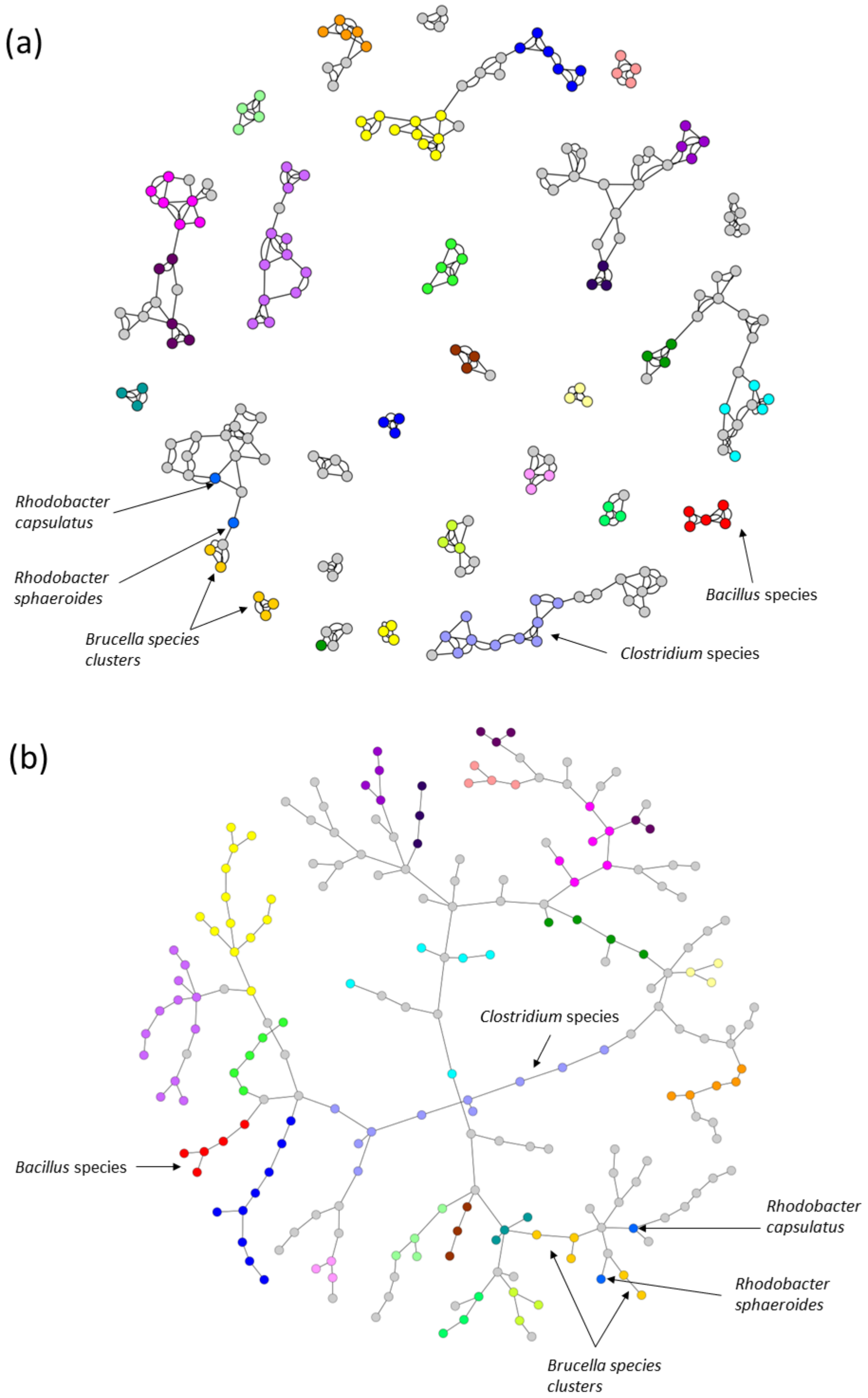


Figure 4.5: **2-way Czekanowski Networks** (a) 2-way Czekanowski Best Edges Network (b) Maximum Spanning Tree (MST) of the all-vs-all Czekanowski network. 3-way Sørensen network pruned by selecting the best and second best edge for each node. Nodes represent bacterial species and edges represent similarity between pairs of bacterial species based on gene family content, quantified using the 3-way Sørensen Index. Nodes are coloured according to genus. The same node colour key as in Figure 2 applies.

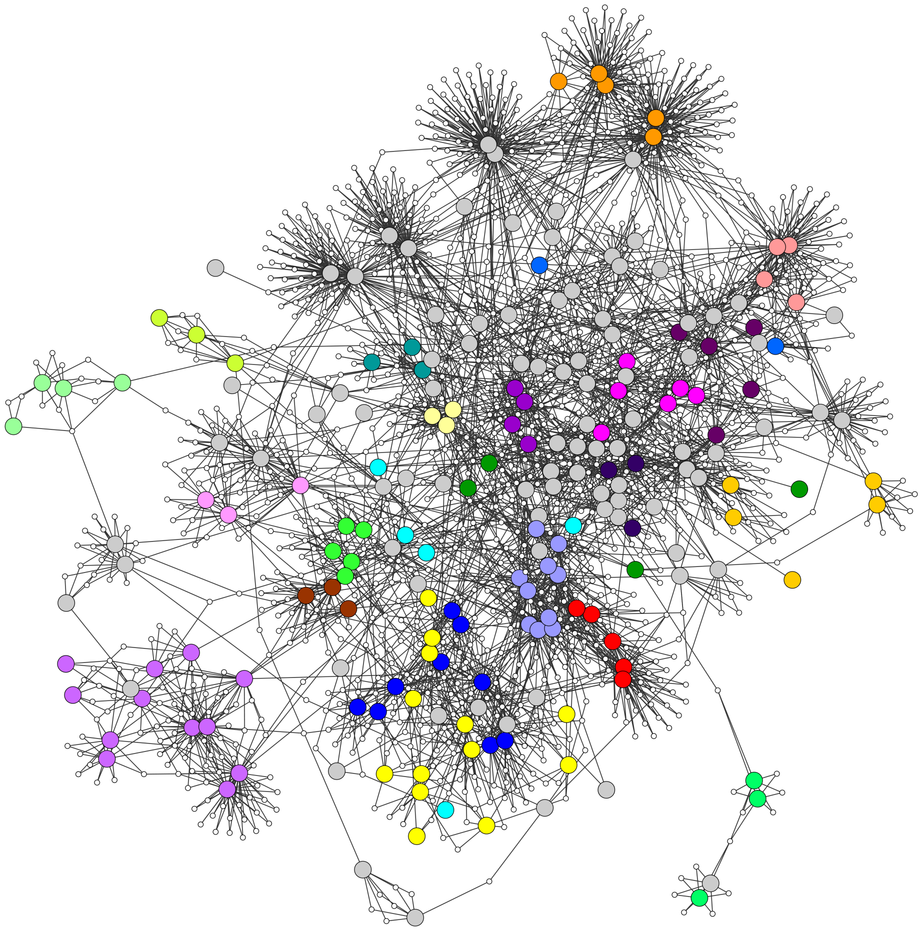


Figure 4.6: **Shared Enriched Families** Network of bacteria species connected through shared enriched gene families. Small, white nodes represent gene families, coloured nodes represent bacterial species coloured by genus. Edges connect gene families to species in which they are enriched.

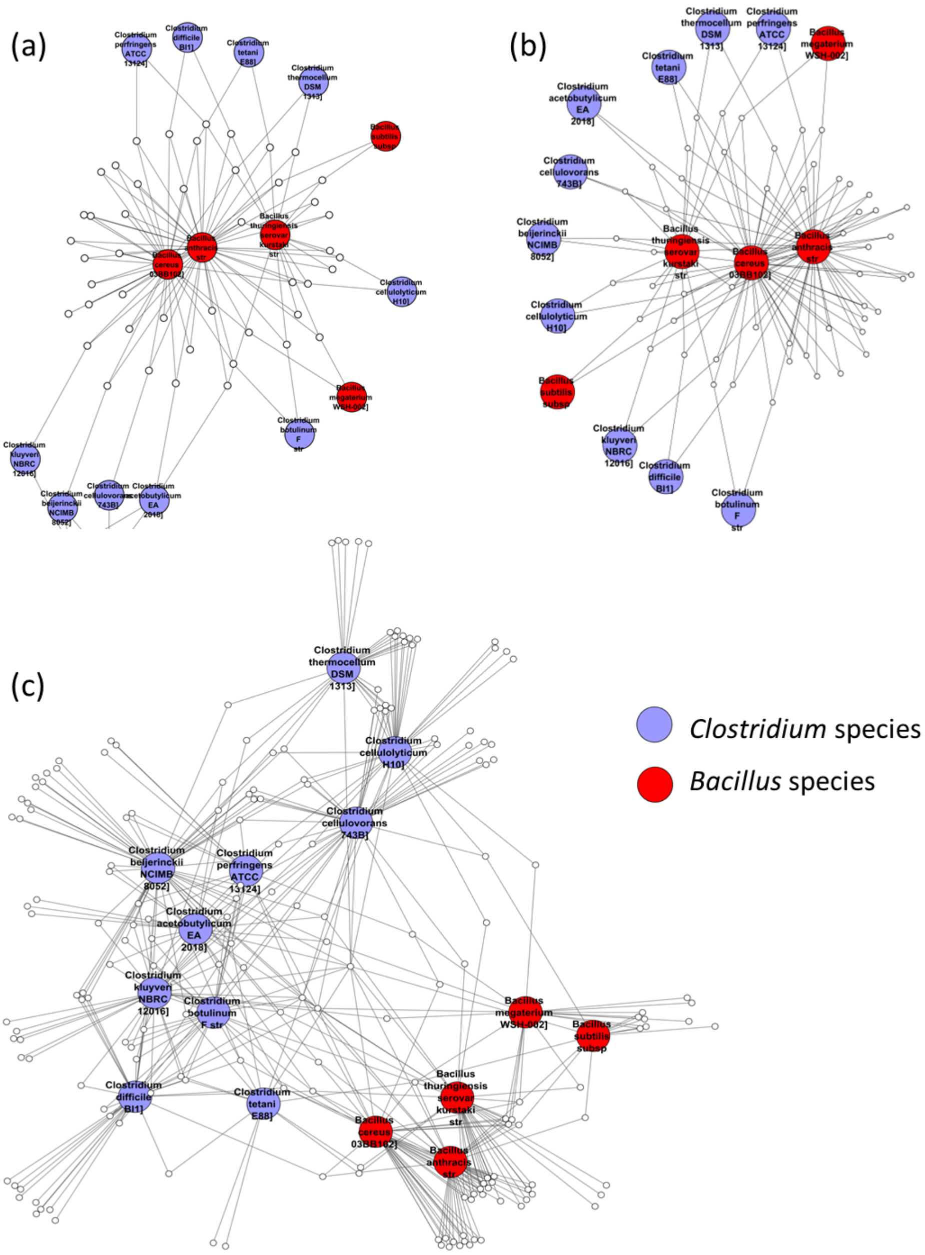

(a)
Clostridium perfringens ATCC 13124]
Clostridium difficile BI1]
Clostridium tetani E88]
Clostridium thermocellum DSM 1313]
Bacillus subtilis subsp
Bacillus thuringiensis serovar kurstaki str
Bacillus anthracis str
Bacillus cereus 03BB102]
Clostridium cellulolyticum H10]
Bacillus megaterium WSH-002]
Clostridium botulinum F str
Clostridium kluyveri NBRC 12016]
Clostridium beijerinckii NCIMB 8052]
Clostridium cellulovorans 743B]
Clostridium acetobutylicum EA 2018]
(b)
Clostridium thermocellum DSM 1313]
Clostridium perfringens ATCC 13124]
Bacillus megaterium WSH-002]
Clostridium tetani E88]
Clostridium acetobutylicum EA 2018]
Clostridium cellulovorans 743B]
Clostridium beijerinckii NCIMB 8052]
Bacillus thuringiensis serovar kurstaki str
Bacillus cereus 03BB102]
Bacillus anthracis str
Clostridium cellulolyticum H10]
Bacillus subtilis subsp
Clostridium kluyveri NBRC 12016]
Clostridium difficile BI1]
Clostridium botulinum F str
(c)
Clostridium thermocellum DSM 1313]
Clostridium cellulolyticum H10]
Clostridium cellulovorans 743B]
Clostridium beijerinckii NCIMB 8052]
Clostridium perfringens ATCC 13124]
Clostridium acetobutylicum EA 2018]
Clostridium kluyveri NBRC 12016]
Clostridium botulinum F str
Clostridium difficile BI1]
Clostridium tetani E88]
Bacillus megaterium WSH-002]
Bacillus subtilis subsp
Bacillus thuringiensis serovar kurstaki str
Bacillus cereus 03BB102]
Bacillus anthracis str
Clostridium species
Bacillus species

Figure 4.7: **Clostridium and Bacillus subnetwork**. Subnetworks containing the *Clostridium* and *Bacillus* species selected from (a) 3-way best edge Sørensen Network (b) 3-way best edge Czekanowski Network (c) Gene family enrichment network.

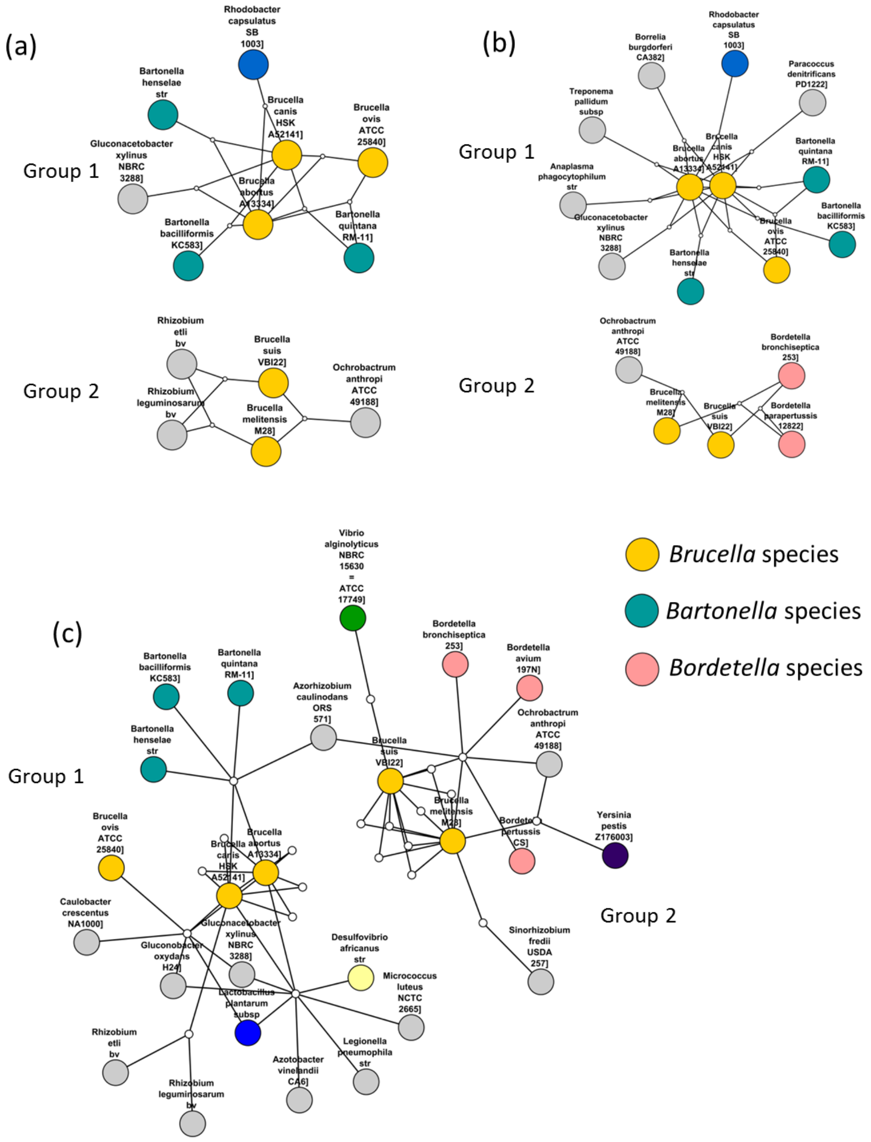

(a)
Group 1
Rhodobacter capsulatus SB 1003]
Bartonella henselae str
Brucella canis HSK A52141]
Brucella ovis ATCC 25840]
Gluconacetobacter xylinus NBRC 3288]
Brucella abortus A13334]
Bartonella bacilliformis KC583]
Bartonella quintana RM-11]
Group 2
Rhizobium etli bv
Brucella suis VBI22]
Ochrobactrum anthropi ATCC 49188]
Rhizobium leguminosarum bv
Brucella melitensis M28]
(b)
Group 1
Rhodobacter capsulatus SB 1003]
Borrelia burgdorferi CA382]
Paracoccus denitrificans PD1222]
Treponema pallidum subsp
Brucella canis HSK A52141]
Brucella abortus A13334]
Bartonella quintana RM-11]
Anaplasma phagocytophilum str
Gluconacetobacter xylinus NBRC 3288]
Brucella ovis ATCC 25840]
Bartonella bacilliformis KC583]
Bartonella henselae str
Group 2
Ochrobactrum anthropi ATCC 49188]
Bordetella bronchiseptica 253]
Brucella melitensis M28]
Brucella suis VBI22]
Bordetella parapertussis 12822]
(c)
Vibrio alginolyticus NBRC 15630 = ATCC 17749]
Brucella species
Bartonella species
Bordetella species
Bordetella bronchiseptica 253]
Bordetella avium 197N]
Bartonella bacilliformis KC583]
Bartonella quintana RM-11]
Azorhizobium caulinodans ORS 571]
Ochrobactrum anthropi ATCC 49188]
Bartonella henselae str
Brucella suis VBI22]
Group 1
Brucella melitensis M28]
Brucella ovis ATCC 25840]
Brucella abortus A13334]
Brucella canis HSK A52141]
Bordetella pertussis CS]
Yersinia pestis Z176003]
Caulobacter crescentus NA1000]
Gluconacetobacter xylinus NBRC 3288]
Group 2
Gluconobacter oxydans H24]
Desulfovibrio africanus str
Sinorhizobium fredii USDA 257]
Micrococcus luteus NCTC 2665]
Lactobacillus plantarum subsp
Rhizobium etli bv
Legionella pneumophila str
Azotobacter vinelandii CA6]
Rhizobium leguminosarum bv

Figure 4.8: **Clustering within *Brucella* genus**. Subnetworks containing *Brucella* species constructed by selecting *Brucella* species and all neighbouring species nodes from (a) 3-way best edge Sørensen Network (b) 3-way best edge Czekanowski Network (c) Gene family enrichment network.

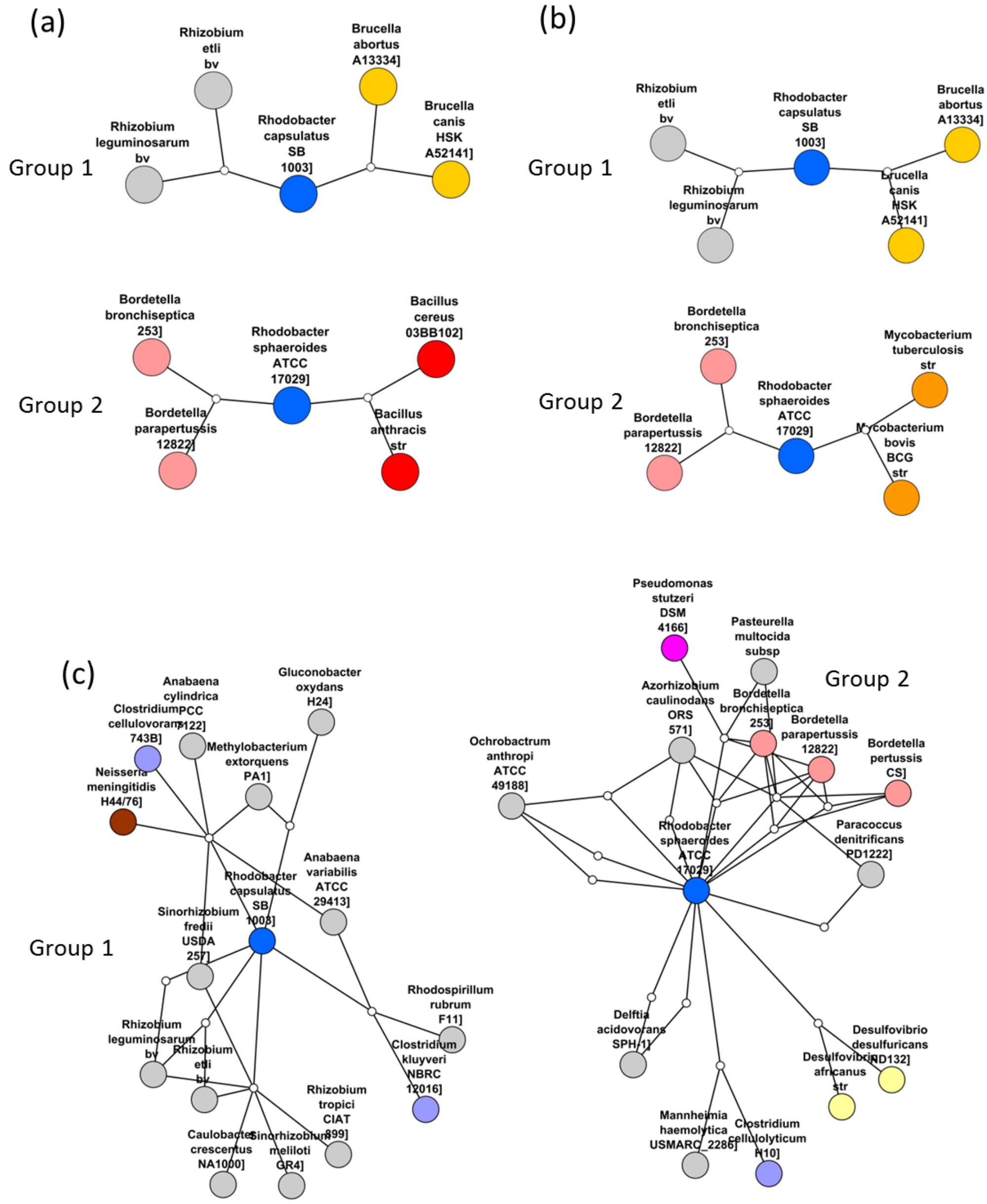

(a)
Group 1
Rhizobium etli bv
Rhizobium leguminosarum bv
Rhodobacter capsulatus SB 1003]
Brucella abortus A13334]
Brucella canis HSK A52141]
Group 2
Bordetella bronchiseptica 253]
Bordetella parapertussis 12822]
Rhodobacter sphaeroides ATCC 17029]
Bacillus cereus 03BB102]
Bacillus anthracis str
(b)
Group 1
Rhizobium etli bv
Rhizobium leguminosarum bv
Rhodobacter capsulatus SB 1003]
Brucella abortus A13334]
Brucella canis HSK A52141]
Group 2
Bordetella bronchiseptica 253]
Bordetella parapertussis 12822]
Rhodobacter sphaeroides ATCC 17029]
Mycobacterium tuberculosis str
Mycobacterium bovis BCG str
(c)
Group 1
Anabaena cylindrica PCC 7122]
Clostridium cellulovorans 743B]
Gluconobacter oxydans H24]
Methylobacterium extorquens PA1]
Neisseria meningitidis H44/76]
Anabaena variabilis ATCC 29413]
Rhodobacter capsulatus SB 1003]
Sinorhizobium fredii USDA 257]
Rhodospirillum rubrum F11]
Rhizobium leguminosarum bv
Rhizobium etli bv
Clostridium kluyveri NBRC 12016]
Rhizobium tropici CIAT 899]
Caulobacter crescentus NA1000]
Sinorhizobium meliloti GR4]
Group 2
Pseudomonas stutzeri DSM 4166]
Pasteurella multocida subsp
Azorhizobium caulinodans ORS 571]
Bordetella bronchiseptica 253]
Bordetella parapertussis 12822]
Bordetella pertussis CS]
Ochrobactrum anthropi ATCC 49188]
Rhodobacter sphaeroides ATCC 17029]
Paracoccus denitrificans PD1222]
Delftia acidovorans SPH-1]
Desulfovibrio desulfuricans ND132]
Desulfovibrio africanus str
Mannheimia haemolytica USMARC_2286]
Clostridium cellulolyticum H10]

Figure 4.9: **Separation of *Rhodobacter* species**. Subnetworks containing *Rhodobacter* species constructed by selecting *Rhodobacter* species and all neighbouring species nodes from (a) 3-way best edge Sørensen Network (b) 3-way best edge Czekanowski Network (c) Gene family enrichment network.

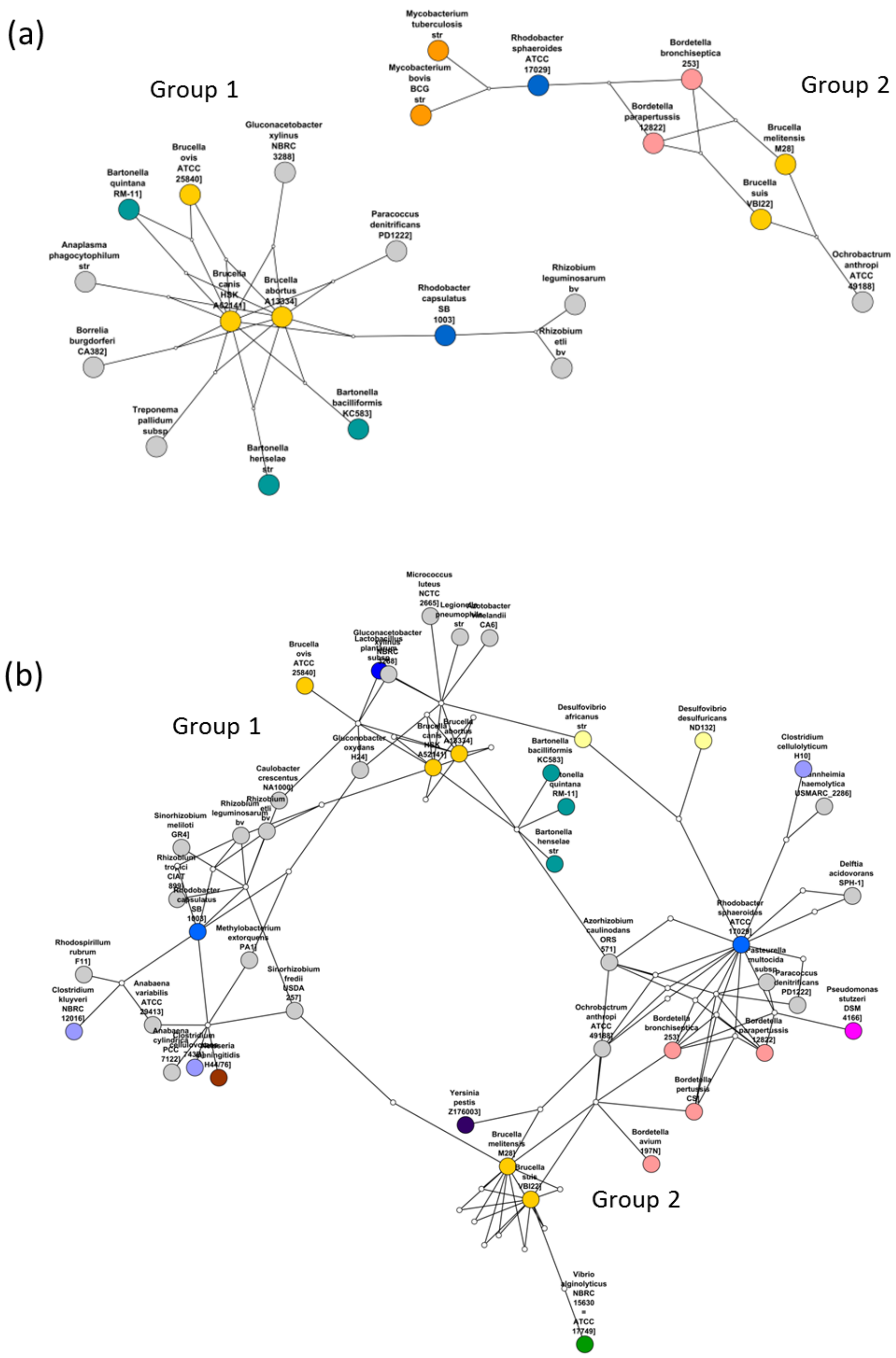

(a)
Group 1
Group 2
Mycobacterium tuberculosis str
Mycobacterium bovis BCG str
Rhodobacter sphaeroides ATCC 17029]
Bordetella bronchiseptica 253]
Bordetella parapertussis 12822]
Brucella melitensis M28]
Brucella suis VBI22]
Ochrobactrum anthropi ATCC 49188]
Gluconacetobacter xylinus NBRC 3288]
Brucella ovis ATCC 25840]
Bartonella quintana RM-11]
Anaplasma phagocytophilum str
Brucella canis HSK A52141]
Brucella abortus A13334]
Paracoccus denitrificans PD1222]
Rhodobacter capsulatus SB 1003]
Rhizobium leguminosarum bv
Rhizobium etli bv
Borrelia burgdorferi CA382]
Treponema pallidum subsp
Bartonella henselae str
Bartonella bacilliformis KC583]
(b)
Group 1
Group 2
Micrococcus luteus NCTC 2665]
Legionella pneumophila str
Azotobacter vinelandii CA6]
Gluconacetobacter xylinus NBRC 3288]
Lactobacillus plantarum subsp
Brucella ovis ATCC 25840]
Gluconobacter oxydans H24]
Brucella canis HSK A52141]
Brucella abortus A13334]
Desulfovibrio africanus str
Desulfovibrio desulfuricans ND132]
Clostridium cellulolyticum H10]
Mannheimia haemolytica USMARC_2286]
Bartonella bacilliformis KC583]
Bartonella quintana RM-11]
Bartonella henselae str
Caulobacter crescentus NA1000]
Rhizobium etli bv
Rhizobium leguminosarum bv
Sinorhizobium meliloti GR4]
Rhizobium tropici CIAT 899]
Rhodobacter capsulatus SB 1003]
Methylobacterium extorquens PA1]
Sinorhizobium fredii USDA 257]
Rhodospirillum rubrum F11]
Clostridium kluyveri NBRC 12016]
Anabaena variabilis ATCC 29413]
Anabaena cylindrica PCC 7122]
Clostridium cellulovorans 743B]
Neisseria meningitidis H44/76]
Delftia acidovorans SPH-1]
Rhodobacter sphaeroides ATCC 17029]
Azorhizobium caulinodans ORS 571]
Pasteurella multocida subsp
Paracoccus denitrificans PD1222]
Pseudomonas stutzeri DSM 4166]
Ochrobactrum anthropi ATCC 49188]
Bordetella bronchiseptica 253]
Bordetella parapertussis 12822]
Bordetella pertussis CS]
Bordetella avium 197N]
Yersinia pestis Z176003]
Brucella melitensis M28]
Brucella suis VBI22]
Vibrio alginolyticus NBRC 15630 = ATCC 17749]

Figure 4.10: ***Rhodobacter* and *Brucella* species**. Subnetworks containing *Brucella* and *Rhodobacter* species constructed by selecting *Brucella* and *Rhodobacter* species and all neighbouring species nodes from (a) 3-way best edge Czekanowski Network (b) Gene family enrichment network.

4.5 Conclusions

3-way networks were explored for their use in comparative genomics and their utility in modelling more complex relationships. These networks, when used to model the phylogenomic relationships between 211 bacterial species revealed relationships between the species which were not found when using standard 2-way network models. A potential limitation of this approach of using 3-way networks is their combinatorial complexity. With larger datasets, calculating the similarity between all possible triplets of objects will require a large amount of compute power. However, with the appropriate High Performance Computing resources, these networks will be a useful tool for comparative genomics in order to model and reveal complex relationships.

4.6 Materials and Methods

4.6.1 Bacterial Gene Family Construction

Gene families were constructed using the TribeMCL pipeline [15]. An all-vs-all protein BLAST [18] was performed on the translated genomes of 211 bacterial species. The Perl script `orthomclBlastParser` from the OrthoMCL package [26] was then used to parse the Blast results in order to select only the best Blast match per gene pair. For each gene pair ab, a score S_{ab} was calculated as [15]:

$$S_{ab} = \log_2\left(\frac{E_{ab} + E_{ba}}{2}\right) \tag{4.6.1}$$

where E_{ab} and E_{ba} are the E-values for the reciprocal BLAST hits between gene a and gene b. This resulted in a network in which each node represented a gene and each edge ab represented the similarity between the two nodes (a and b) which it connects, weighted by the similarity score S_{ab}. MCL was then applied using an inflation value of 2 to cluster the network into gene families [27]. From the resulting gene families, a matrix was constructed called the Species-Family (SF) matrix, in which the rows represented bacterial gene families constructed using TribeMCL, and columns represented bacterial species, and each entry ij represented the number of genes in gene family i present in species j.

4.6.2 3-way Network Construction

The 3-way Sørensen Index and the 3-way Czekanowski Index was used to quantify the similarity between each triplet of species. Let X_i and Y_i and Z_i represent the i^{th} element in columns X, Y and Z of the SF-matrix (i.e. the number of members of gene family i in species X species Y and species Z respectively. Let X_B, Y_B and Z_B be the binary vectors associated with vectors X, Y and Z respectively. For each triplet of species (X, Y, Z) the Sørensen

Index was calculated using Equation 4.4.7 and the Czekanowski Index was calculated using Equation 4.4.9. This resulted in a Sørensen 3-way network and a Czekanowski 3-way network. Using Theorem 1, any threshold set above 0.75 will exclude any 3-way relationships with no 3-way intersection contribution. Thus, a threshold of 0.76 was applied to each network and visualized in Cytoscape [28]. These networks can be seen in Figures S1 and S2. Cytoscape can only visualize 2-way networks in the sense that it can only handle edges connecting 2 nodes. To our knowledge, no visualization software exists for 3-way networks. Thus, the 3-way network had to be transformed such that it could be visualized in Cytoscape. To do so, each 3-way-edge was represented by a node with degree 3, connected to the bacterial species nodes which the 3-way-edge connected. In the transformed network, each node thus either represented a bacterial species or a 3-way edge (referred to as an 'edge-node'). A close-up of these 3way-edges can be seen in Figure S3.

A best-edge approach was also used to prune the 3-way networks. For each bacterial species node, the best and second best edges (edges with the highest and second highest weight) were selected. A network was constructed and transformed into a format which can be visualized in Cytoscape as described above. The resulting networks can be seen in Figures 4.2 and 4.4.

4.6.3 2-way Network Construction

The standard 2-way Sørensen and 2-way Czekanowski Indices were used to quantify the similarities between all pairs of species. Let X_i and Y_i represent the i^{th} element in column X and column Y in the SF-matrix (i.e. the number of members of gene family i in species X and species Y respectively. Let X_B be the binary vector associated with vector X and Y_B be the binary vector associated with vector Y. For each pair of species (X, Y) the Sørensen Index was calculated using Equation 4.4.3 and the Czekanowski Index was calculated using Equation 4.4.8. These networks were pruned using two approaches, namely a Maximum Spanning Tree and best edge selection. Each Maximum Spanning Tree was calculated by converting the network from a similarity network into a distance network by inverting the edge weights i.e. for each edge weight w the inverted edge weight w' was calculated as

$$w' = 1 - w.$$

A Minimum Spanning Tree algorithm was then applied to the distance network using the Dijkstra algorithm from the Graph Perl Module (Jarkko Hietaniemi, http://www.cpan.org/). This was performed by using the Perl program for MST construction as used in [29]. For best edge selection, the best and second best edge for each node was selected based on edge weight. These pruned networks were visualized in Cytoscape [28] and can be seen in Figures 4.3a, 4.3b, 4.5a and 4.5b.

4.6.4 Combined 2-way and 3-way Network Construction

For both the Sørensen Index and the Czekanowski Index, the union of the 3-way best-edge network and the 2-way MST was calculated, resulting in a combined network model. These can be seen in Figures S4 and S5.

4.6.5 Gene Family Enrichment

Fisher's exact test [17], followed by Holm-Bonferroni multiple hypothesis correction [30] was used to determine enrichment of gene families within species. This was performed using a customized Perl program which made use of the Text::NSP::Measures::2D::Fisher Perl module from CPAN (http://www.cpan.org/). A p-value cutoff of 0.05 was used. Gene families which were enriched in more than one species (so-called shared-enriched families) were selected and a new network was constructed in which each node represented either a bacterial species or a gene family, and each edge connected a gene family to bacterial species in which it was enriched. The species were coloured according to their genera. The network was visualized in Cytoscape [28] (Figure 4.6).

4.7 Acknowledgements

The authors would like to acknowledge The Center for High Performance Computing and the Stellenbosch High Performance Computing Cluster for computing resources, as well as the South African National Research Foundation Technology (www.nrf.ac.za) and Human Resources Programme and Winetech. The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

4.8 Author's contributions

D Weighill and D Jacobson conceived of and designed the method, D Weighill wrote the code and created the networks, D Weighill and D Jacobson discussed and interpreted the networks, D Weighill drafted the manuscript, D Jacobson critically revised and edited the manuscript.

4.9 Competing Interests

The authors declare that they have no competing financial interests.

4.10 Supplementary Material

The following supplementary material contains a proof of Lemma 1, and several supplementary Figures.

4.10.1 Lemma 1

Given integers a, b and c, the following relation holds:

$$\min(a,b) + \min(a,c) - \min(a,b,c) \leq a \tag{4.10.1}$$

Proof. There are 6 cases to consider.

Case 1: If $a \leq b \leq c$ then:

$$\begin{aligned}\min(a,b) + \min(a,c) - \min(a,b,c) &= a + a - a\\ &= a\\ &\leq a\end{aligned}$$

Case 2: If $a \leq c \leq b$ then:

$$\begin{aligned}\min(a,b) + \min(a,c) - \min(a,b,c) &= a + a - a\\ &= a\\ &\leq a\end{aligned}$$

Case 3: If $b \leq c \leq a$ then:

$$\begin{aligned}\min(a,b) + \min(a,c) - \min(a,b,c) &= b + c - b\\ &= c\\ &\leq a\end{aligned}$$

Case 4: If $b \leq a \leq c$ then:

$$\begin{aligned}\min(a,b) + \min(a,c) - \min(a,b,c) &= b + a - b\\ &= a\\ &\leq a\end{aligned}$$

Case 5: If $c \leq b \leq a$ then:

$$\begin{aligned}\min(a,b) + \min(a,c) - \min(a,b,c) &= c + b - c\\ &= b\\ &\leq a\end{aligned}$$

Case 6: If $c \leq a \leq b$ then:

$$\begin{aligned}\min(a,b) + \min(a,c) - \min(a,b,c) &= c + a - c \\ &= a \\ &\leq a\end{aligned}$$

□

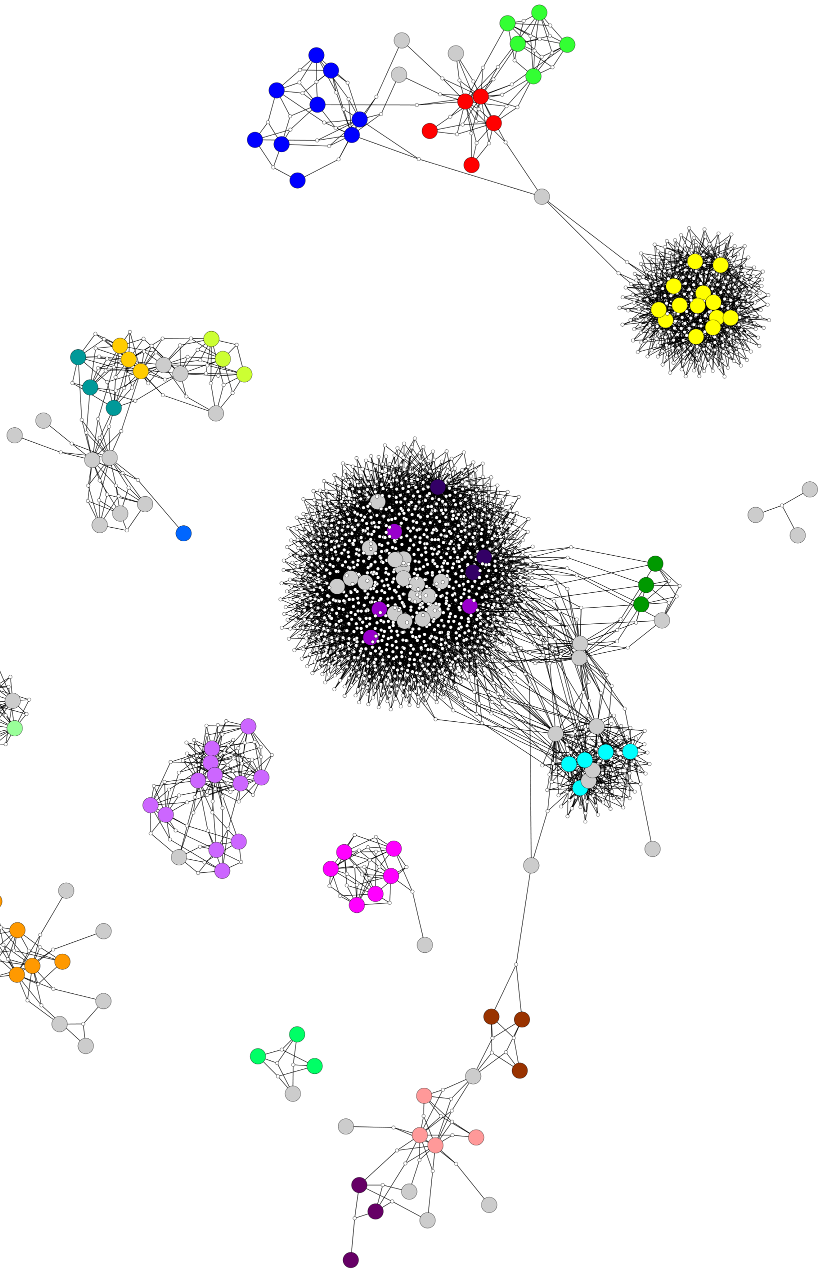


Figure S1: **Thresholded 3-way Sørensen Network** Network constructed by setting a 0.76 threshold for the 3-way Sørensen Network, and removing all 3-way edges below this threshold.

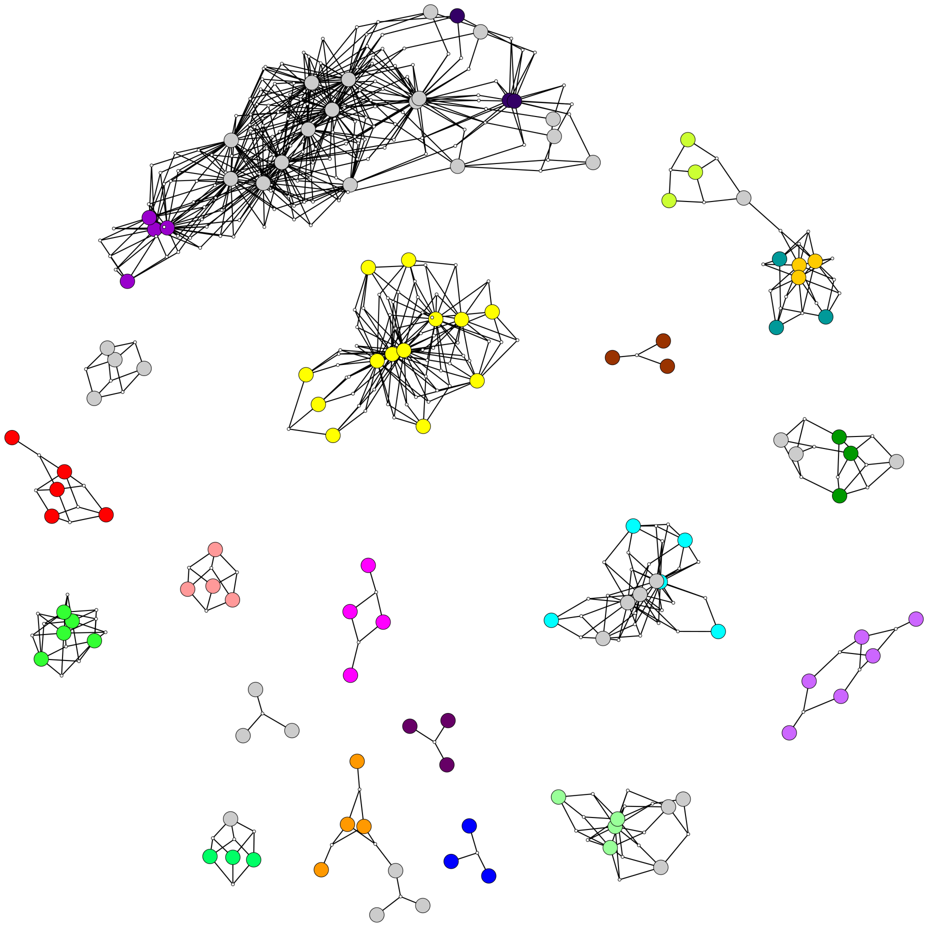


Figure S2: **Thresholded 3-way Czekanowski Network**Network constructed by setting a 0.76 threshold for the 3-way Czekanowski Network, and removing all 3-way edges below this threshold.

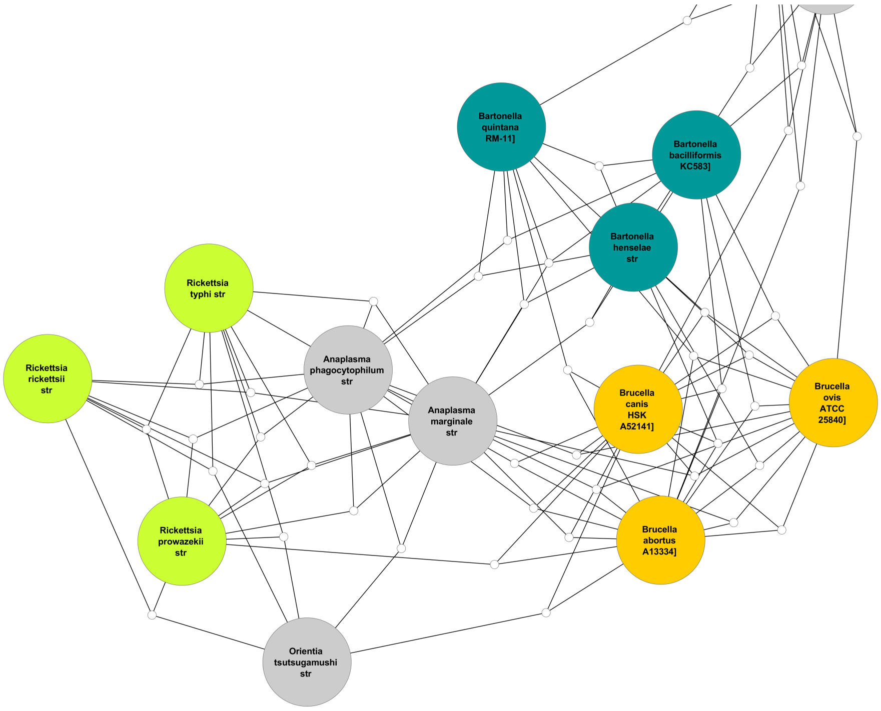

Bartonella quintana RM-11]
Bartonella bacilliformis KC583]
Bartonella henselae str
Rickettsia typhi str
Rickettsia rickettsii str
Anaplasma phagocytophilum str
Anaplasma marginale str
Brucella canis HSK A52141]
Brucella ovis ATCC 25840]
Rickettsia prowazekii str
Brucella abortus A13334]
Orientia tsutsugamushi str

Figure S3: **3-Way Edges** Close-up of a section of the thresholded 3-way network showing the 3-way edges. Large, coloured nodes represent bacterial species, whereas small white nodes and their respective 3 edges represent 3-way edges connecting the bacterial nodes.

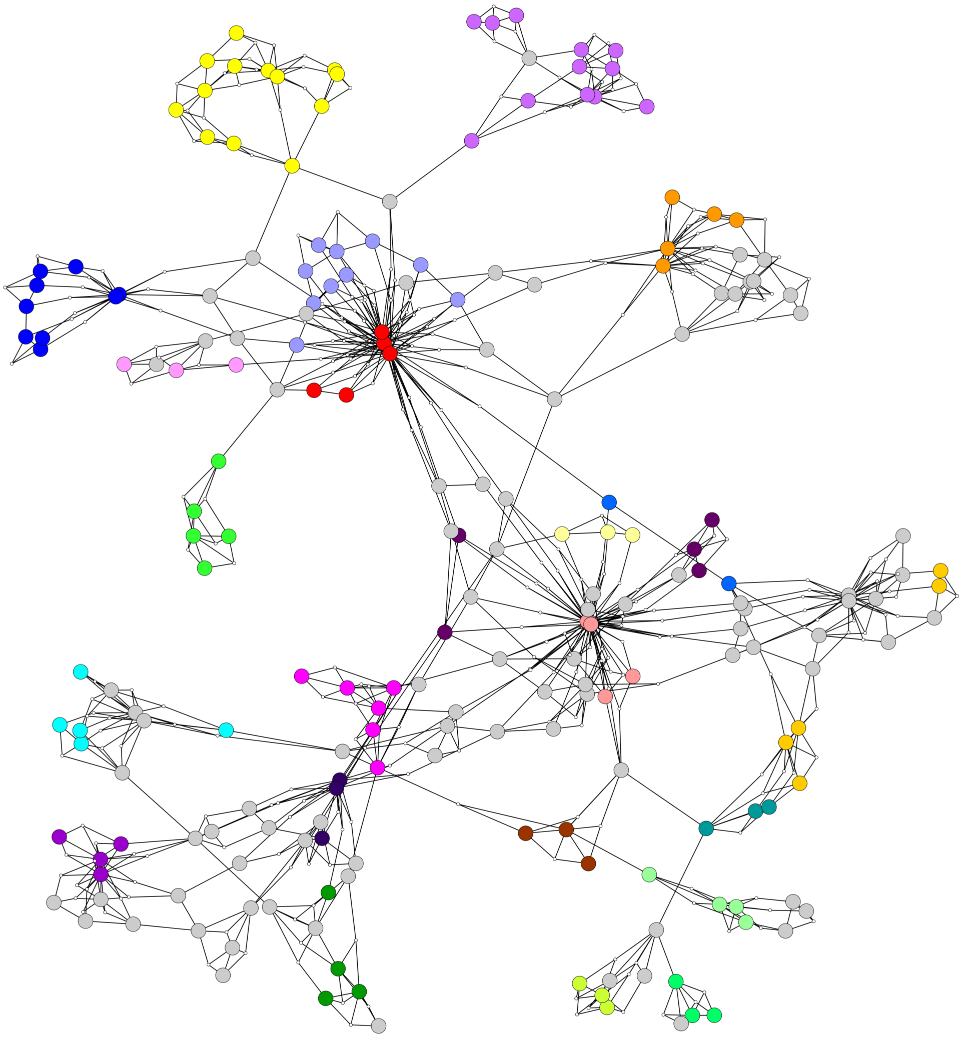


Figure S4: **Union Sørensen MST and Sørensen 3-way Best Edge Network**. Network constructed by taking the union of the Sørensen 3-way Best Edge Network (Figure 2) and the Sørensen MST (Figure 3b).

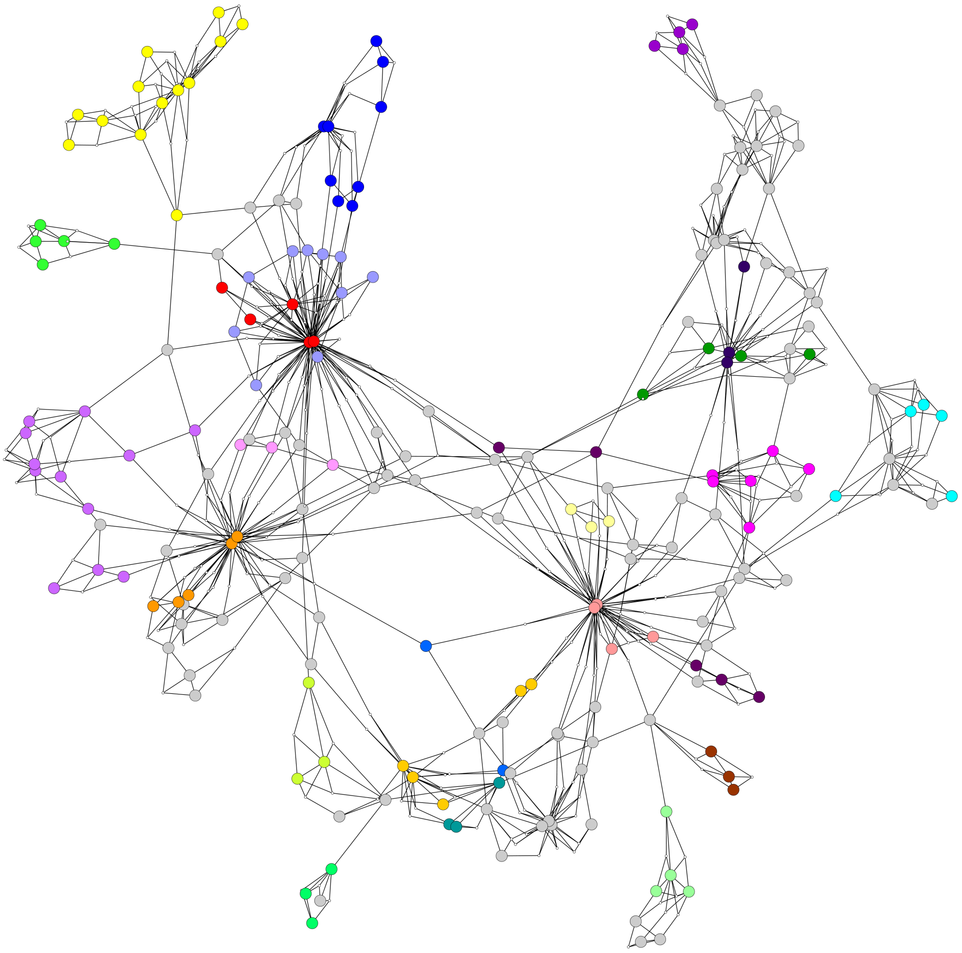


Figure S5: **Union Czekanowski MST and Czekanowski 3-way Best Edge Network** Network constructed by taking the union of the Czekanowski 3-way Best Edge Network (Figure 4) and the Czekanowski MST (Figure 5b).

Bibliography

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

[2] Diserud, O.H. and Ødegaard, F.: A multiple-site similarity measure. *Biology Letters*, vol. 3, no. 1, pp. 20–22, 2007.

[3] Santini, G., Soldano, H. and Pothier, J.: Use of ternary similarities in graph based clustering for protein structural family classification. In: *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*, pp. 457–459. ACM, 2010.

[4] Zhang, L., Gao, Y., Hong, C., Feng, Y., Zhu, J. and Cai, D.: Feature correlation hypergraph: Exploiting high-order potentials for multimodal recognition. 2013.

[5] Zhou, D., Huang, J. and Schölkopf, B.: Learning with hypergraphs: Clustering, classification, and embedding. In: *Advances in neural information processing systems*, pp. 1601–1608. 2006.

[6] Mithani, A., Preston, G.M. and Hein, J.: Rahnuma: hypergraph-based tool for metabolic pathway prediction and network comparison. *Bioinformatics*, vol. 25, no. 14, pp. 1831–1832, 2009.

[7] Seref, O., Brooks, J.P. and Fong, S.S.: Decomposition of flux distributions into metabolic pathways. *Computational Biology and Bioinformatics, IEEE/ACM Transactions On*, vol. 10, no. 4, pp. 984–993, 2013.

[8] Wang, Z., Zhu, X.-G., Chen, Y., Li, Y., Hou, J., Li, Y. and Liu, L.: Exploring photosynthesis evolution by comparative analysis of metabolic networks between chloroplasts and photosynthetic bacteria. *BMC Genomics*, vol. 7, no. 1, p. 100, 2006.

[9] Kim, S.-J., Ha, J.-W. and Zhang, B.-T.: Constructing higher-order mirna-mrna interaction networks in prostate cancer via hypergraph-based learning. *BMC Systems Biology*, vol. 7, no. 1, p. 47, 2013.

[10] Kim, S.-J., Ha, J.-W. and Zhang, B.-T.: Bayesian evolutionary hypergraph learning for predicting cancer clinical outcomes. *Journal of Biomedical Informatics*, 2014.

[11] Gross, J.L. and Yellen, J.: *Handbook of graph theory.* CRC press, 2003.

[12] Sørensen, T.: A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biologiske skrifter*, vol. 5, pp. 1–34, 1948.

[13] Wolda, H.: Similarity indices, sample size and diversity. *Oecologia*, vol. 50, no. 3, pp. 296–302, 1981.

[14] Yoshioka, P.M.: Misidentification of the bray-curtis similarity index. *Marine Ecology Progress Series*, vol. 368, pp. 309–310, 2008.

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

[16] Bray, J.R. and Curtis, J.T.: An ordination of the upland forest communities of southern wisconsin. *Ecological Monographs*, vol. 27, no. 4, pp. 325–349, 1957.

[17] Fisher, R.A.: The logic of inductive inference. *Journal of the Royal Statistical Society*, pp. 39–82, 1935.

[18] Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D. *et al.*: Basic local alignment search tool. *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[19] Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S. and Madden, T.L.: Ncbi blast: a better web interface. *Nucleic Acids Research*, vol. 36, no. suppl 2, pp. W5–W9, 2008.

[20] Traag, B.A., Pugliese, A., Eisen, J.A. and Losick, R.: Gene conservation among endospore-forming bacteria reveals additional sporulation genes in bacillus subtilis. *Journal of Bacteriology*, vol. 195, no. 2, pp. 253–260, 2013.

[21] Stragier, P. and Losick, R.: Molecular genetics of sporulation in bacillus subtilis. *Annual Review of Genetics*, vol. 30, no. 1, pp. 297–341, 1996.

[22] Fukushima, T., Tanabe, T., Yamamoto, H., Hosoya, S., Sato, T., Yoshikawa, H. and Sekiguchi, J.: Characterization of a polysaccharide deacetylase gene homologue (pdab) on sporulation of bacillus subtilis. *Journal of Biochemistry*, vol. 136, no. 3, pp. 283–291, 2004.

[23] Fukushima, T., Yamamoto, H., Atrih, A., Foster, S.J. and Sekiguchi, J.: A polysaccharide deacetylase gene (pdaa) is required for germination and for production of muramic δ-lactam residues in the spore cortex of bacillus subtilis. *Journal of Bacteriology*, vol. 184, no. 21, pp. 6007–6015, 2002.

[24] Paredes, C.J., Alsaker, K.V. and Papoutsakis, E.T.: A comparative genomic view of clostridial sporulation and physiology. *Nature Reviews Microbiology*, vol. 3, no. 12, pp. 969–978, 2005.

[25] Foster, J.T., Beckstrom-Sternberg, S.M., Pearson, T., Beckstrom-Sternberg, J.S., Chain, P.S., Roberto, F.F., Hnath, J., Brettin, T. and Keim, P.: Whole-genome-based phylogeny and divergence of the genus brucella. *Journal of Bacteriology*, vol. 191, no. 8, pp. 2864–2870, 2009.

[26] Li, L., Stoeckert, C. and Roos, D.: Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome Research*, vol. 13, no. 9, pp. 2178–2189, 2003.

[27] van Dongen, S.: *Graph clustering by flow simulation.* Ph.D. thesis, University of Utrecht, 2000.

[28] Shannon, P., Markiel, A., Ozier, O., Baliga, N., Wang, J., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[29] Setati, M.E., Jacobson, D., Andong, U.-C. and Bauer, F.: The vineyard yeast microbiome, a mixed model microbial map. *PloS One*, vol. 7, no. 12, p. e52609, 2012.

[30] Holm, S.: A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, , no. 6, pp. 65–70, 1979.

Chapter 5

Network Meta-Modelling: Similarity Metric Comparison

D. Weighill and D. Jacobson

5.1 Introduction

Meta-modelling involves creating models of models in order to compare the outcomes of a model when different parameters are used. Network models involve modelling the similarity between pairs of objects of interest [1]. A parameter of such a network model could be the similarity metric chosen to quantify the similarity between nodes in order to weight the edges. Many similarity metrics exist, and were developed to quantify different aspects of similarity. Thus, using different similarity metrics to construct a network model should result in different results and thus affect the end biological interpretation.

This study focuses on network meta-modelling, exploring a selection of approaches for the comparison of networks. In particular, network models of particular datasets constructed using different similarity metrics will be compared in order to investigate the effect the choice of similarity metric has on the resulting network topology.

5.2 Results and Discussion

5.2.1 Overview

Two types of datasets were used for the exploration of network comparison approaches. The first dataset on which Clustering and Network Topology Profile Comparisons were performed, was a large grapevine microarray dataset, consisting of 472 microarray experiments, each containing 16602 probesets. The co-expression networks generated from this dataset were very large, containing thousands of nodes and edges. The second type of dataset included the fully sequenced genomes of 71 fungi and 211 bacteria. The networks resulting from these two datasets were smaller and simpler, allowing visual inspection of the results of a new network comparison technique we developed, namely Cross-Network Topological Overlap.

5.2.2 Metric Comparison though Network Topology Profiles and Clustering Comparison

7 similarity metrics, namely the Pearson and Spearman Correlation Coefficients, Jaccard, Sørensen, Czekanowski, SPS indices and Euclidean similarity (Table 5.1) were used as measures for gene co-expression across several grapevine microarray experiments. The SPS (Stringent Proportional Similarity) Index is a metric we created by modifying the Czekanowski Index (also known as Proportional Similarity Index [2]) with the aim of creating a similarity metric which was still a quantitative overlap index like the Czekanowski Index, but is more stringent, in that vectors have to be more similar in quantitative overlap in order to achieve the same score as with the Czekanowski Index.

The distributions of the co-expression values for each metric are shown in Figure 5.1. It is evident that the different similarity metrics have very different distributions, however, certain patterns do come forward. The Jaccard and Sørensen distributions are similar. This can be expected, since both of these metrics are based on set overlaps. For two sets A and B, the set overlap formulation of the Sørensen Index $S_o(A, B)$ and the Jaccard Index $J(A, B)$ are defined as [5]:

$$S_o(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{5.2.1}$$

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{5.2.2}$$

The Sørensen and Jaccard Indices are related to each other by the following equation [5]:

$$S_o = \frac{2J}{J + 1} \tag{5.2.3}$$

This relationship is reflected in the distributions in that the Jaccard distribution is skewed, having a longer right tail than the Sørensen distribution.

The Pearson and Spearman distributions are very similar. This does seem

Table 5.1: **Similarity Metrics.** Definitions of similarity metrics. X and Y are vectors of length n. X_B and Y_B are the binary vectors associated with vectors X and Y respectively, R and Q are the rank vectors associated with vectors X and Y respectively, $D(X,Y)$ is the Euclidean distance between vectors X and Y and $\langle X,Y\rangle$ is the inner product of vectors X and Y.

Similarity Metric	Formula
Pearson Correlation [3]	$P(X,Y) = \frac{\sum_i (X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_i (X_i-\bar{X})^2 \sum_i (Y_i-\bar{Y})^2}}$
Spearman Correlation [4]	$S_p(X,Y) = \frac{\sum_i (R_i-\bar{R})(Q_i-\bar{Q})}{\sqrt{\sum_i (R_i-\bar{R})^2 \sum_i (Q_i-\bar{Q})^2}}$
Sørensen Index [5]	$S_o(X,Y) = \frac{2\sum_i \min(X_{B\,i},Y_{B\,i})}{\sum_i (X_{B\,i}+Y_{B\,i})}$
Jaccard Index [5; 6]	$J(X,Y) = \frac{\langle X,Y\rangle}{\langle X,X\rangle+\langle Y,Y\rangle-\langle X,Y\rangle}$
Czekanowski Index [7; 2]	$C_z(X,Y) = \frac{2\sum_i \min(X_i,Y_i)}{\sum_i (X_i+Y_i)}$
SPS Index	$SPS(X,Y) = 1 - \frac{1}{n}\sum_i \frac{\lvert X_i^2-Y_i^2\rvert}{X_i^2+Y_i^2}$
MIC [8]	Maximum Mutual Information
Euclidean Similarity	$E(X,Y) = 1 - \frac{D(X,Y)}{\max_{X,Y}(D(X,Y))}$

logical since both are correlation coefficients with similar formulation (Table 5.1) and both measure to what extent the elements of two vectors follow the same pattern, the difference being that Pearson measures the linear relationship between two vectors and Spearman, being less stringent, measures the monotonic relationship between two vectors.

The SPS and Czekanowski distributions are similar in that they follow the same pattern of inflection points, however, the SPS distribution is flatter, having less of a spike on the right side of the distribution, indicating that it is indeed more stringent and the Czekanowski Index.

5.2.2.1 Network Topology Profile Comparison

A network comparison method based on the principles of NetSimile [9] was developed, allowing the comparison of a set of networks in a pairwise man-

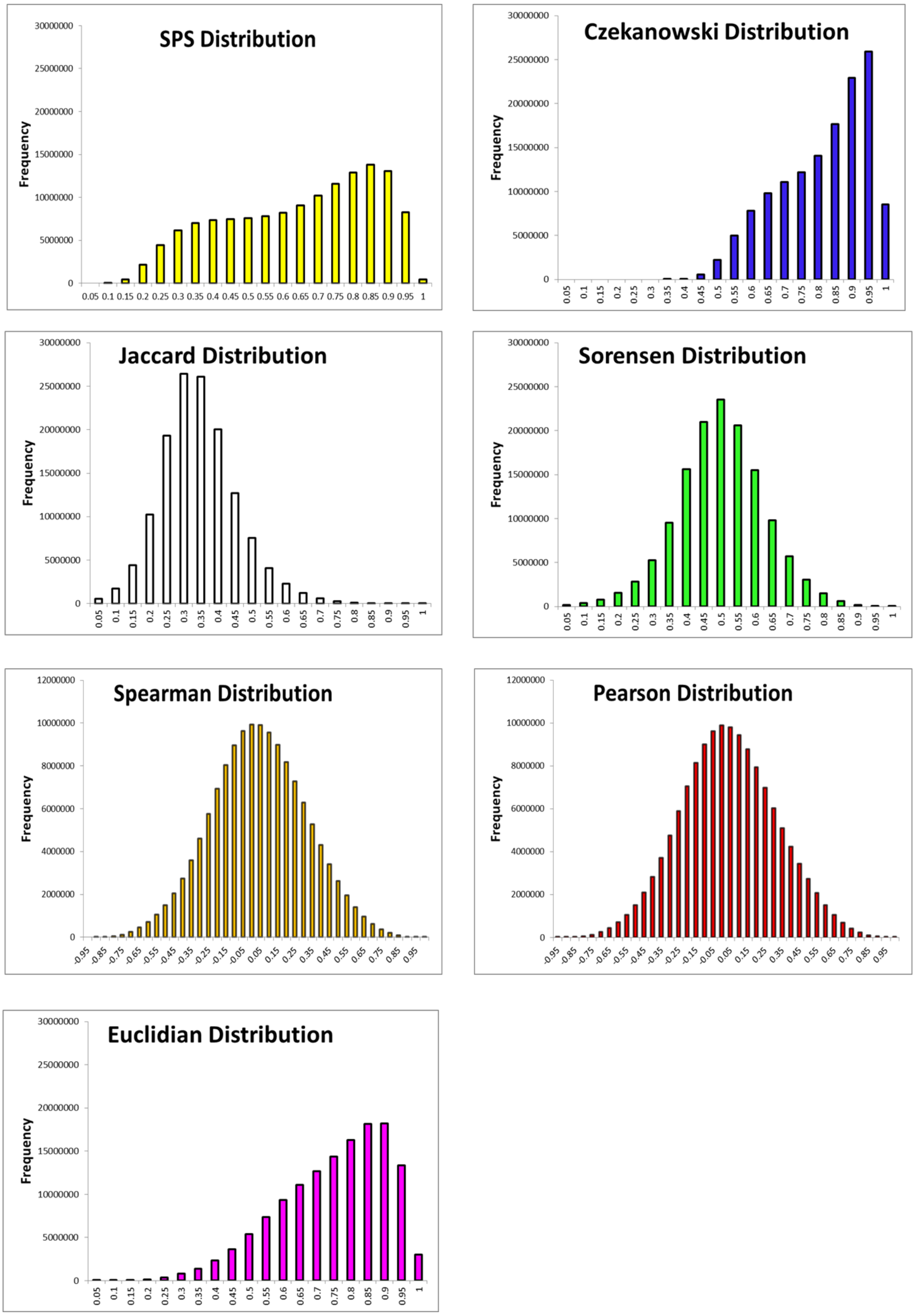


Figure 5.1: **Distributions.** Frequency distributions of co-expression values for each of the similarity metrics when applied to the grapevine microarray expression dataset.

ner. This method involved the calculation of several topology indices for each network. *Local indices* are calculated per node, and include clustering coefficients, connectivities, scaled connectivities and maximum adjacency ratios. *Global indices* are calculated for a network as a whole, and include maximum connectivity, density, centralization, heterogeneity and degree correlation (See Table 5.2). These topology indices form the variables in a topology profile for each network. Perl programs were written in order to calculate a series of local and global topology indices for a given set of networks and to construct topology profiles for these networks. Certain Perl programs made use of the Statistics::Basic Perl Module (Paul Miller, http://www.cpan.org/). The topology profiles form the rows of a matrix in which each row represents one of the input networks and each column represents a network topology index. Four different topology profile matrices were created with different sets of variables, namely:

1. Weighted global indices
2. Unweighted global indices
3. Weighted local indices
4. Unweighted local indices

These topology profiles can be further compared using multivariate methods such as Principal Components Analysis (PCA). In order to further investigate the relationships between and the effect of different similarity metrics on network topology, our network comparison method was used to compare grapevine co-expression networks generated using the 7 different similarity metrics. Each co-expression network was pruned to maintain only the top 1 % of edges. This pruning strategy was applied instead of a hard thresholding approach because the metrics have such varied distributions (Figure 5.1).

Global and local topology indices were then calculated for each network. This resulted in the 4 topology profile matrices described above, each of which was analysed with PCA. The score plot for the weighted local index matrix is shown in Figure 5.2a. SPS-metric and Czekanowski Index cluster together, as do Pearson and Spearman Correlation Coefficients and Søresen and Jaccard Indices. Intuitively, these groupings seem logical. Pearson and Spearman Corrlation are both correlation coefficients and are calculated in a similar manner, except that Spearman uses ranks instead of actual variable values. Sorensen and Jaccard Indices are both set overlap measures and are calculated in a similar manner and thus would be expected to be similar. Lastly, the SPS Index was derived from the Czekanowski Index and thus it makes sense that they are similar. The score plot for the weighted global index matrix is shown in Figure 5.2c. Similar groupings of metrics are seen in this score plot. It is

Table 5.2: **Network Topology Indices.** Definitions of network indices [10; 11], where i and j are nodes, a_{ij} is the adjacency of nodes i and j, S is the vector of degrees of all source nodes and T is the vector of degrees of all target nodes.

Topology Index	Definition
	Local Indices
Connectivity	$k_i = \sum_{j \neq i} a_{ij}$
Scaled Connectivity	$k_i^{\text{scaled}} = \frac{\sum_{j \neq i} a_{ij}}{k\text{max}}$
Maximum Adjacency Ratio	$\text{MAR}_i = \frac{\sum_{j \neq i} (a_{ij})^2}{\sum_{j \neq i} a_{ij}}$
Clustering Coefficient	$CC_i = \frac{\sum_{l \neq i} \sum_{m \neq i,l} a_{il} a_{lm} a_{mi}}{(\sum_{l \neq i} a_{il})^2 - \sum_{l \neq i} (a_{il})^2}$
	Global Indices
Maximum Connectivity	$k\text{max} = \max(k_i)$
Density	$D_N = \frac{\sum_i \sum_{j \neq i} a_{ij}}{n(n-1)}$
Centralization	$C_N = \frac{n}{n-2} \left(\frac{k\text{max}}{n-1} - D_N \right)$
Heterogeneity	$H_N = \frac{\sqrt{\text{var}(k)}}{\text{mean}(k)}$
Degree Correlation	$\text{Pearson}(S, T)$

interesting to note that the number of variables in the topology profile matrix in which the variables are local indices is vastly greater than that which the variables are global indices. Since local indices are calculated for each node and there are thousands of nodes in each network, the number of variables in the local index topology profile matrix is very large. However, global indices are calculated only once per network, thus there are only 5 variables in the global index topology profile matrix. It is interesting that even though there are far fewer variables in the global index topology profile matrix than the local index topology profile matrix, both give similar groupings in their respective PCA score plots.

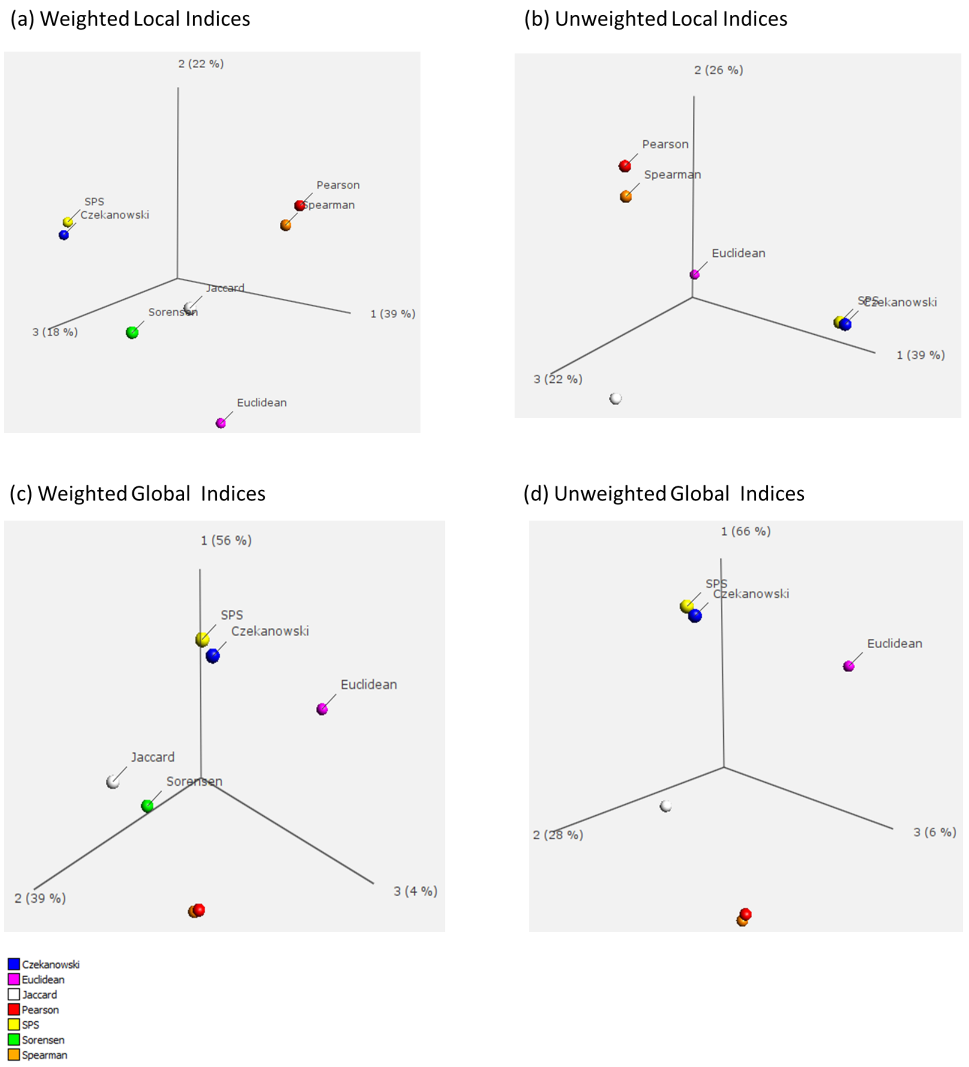
In general, the score plots resulting from PCA of the matrices with unweighted indices as variables (Figure 5.2b and 5.2d) have similar but tighter groupings than those resulting from PCA of matrices with weighted indices as variables (Figures 5.2a and 5.2c). The Jaccard and Sørensen scores are in fact identical in both score plots resulting from using unweighted indices as variables (Figures 5.2b and 5.2d).

5.2.2.2 Network Clustering Comparison

The 7 pruned similarity networks were all clustered using MCL [12] and the resulting clusterings were compared using three clustering comparison metrics, namely Average-Maximum Overlap, Jaccard Overlap and Normalized Mutual Information (see Methods). This resulted in three all-vs-all networks in which each node represented a similarity metric and each edge represented similarity between those two similarity metrics, based on how similar the clusterings of the two respective co-expression networks were. These three clustering comparison networks are show in Figure 5.3. All three clustering comparison approaches give similar results. From the thickness of the edges, it can be seen that the Pearson network clustering is most similar to the Spearman clustering, Jaccard is most similar to Sørensen, SPS is most similar to Czekanowski and Euclidean is quite different from all other metrics. These are the same groupings which were seen in the Score Plots resulting from PCA of the network topology profiles and suggested by the distributions of the metrics.

5.2.3 Metric Comparison through Network Merging and Cross-Network Topological Overlap

Phylogenomic networks were constructed in order to represent the evolutionary relationships and similarities between 71 fungal species and 211 bacterial species based on gene family content. For the fungal dataset, 8 similarity metrics were used to calculate the similarity between the gene family content of 71 fungal species. A similar procedure was performed to calculate the similarity between the gene family content of 211 bacterial species, using 7 different



(a) Weighted Local Indices

(b) Unweighted Local Indices

(c) Weighted Global Indices

(d) Unweighted Global Indices

Figure 5.2: **Score Plots: PCA of Topology Profiles.** Score plots resulting from PCA of the topology profile matrices in which variables are (a) weighted local topology indices, (b) unweighted local topology indices, (c) weighted global topology indices and (d) unweighted global topology indices. Scores of the Jaccard and Sørensen Index networks in (b) and (d) are identical, thus their points in the score plots are superimposed and cannot both be visualized or labelled.

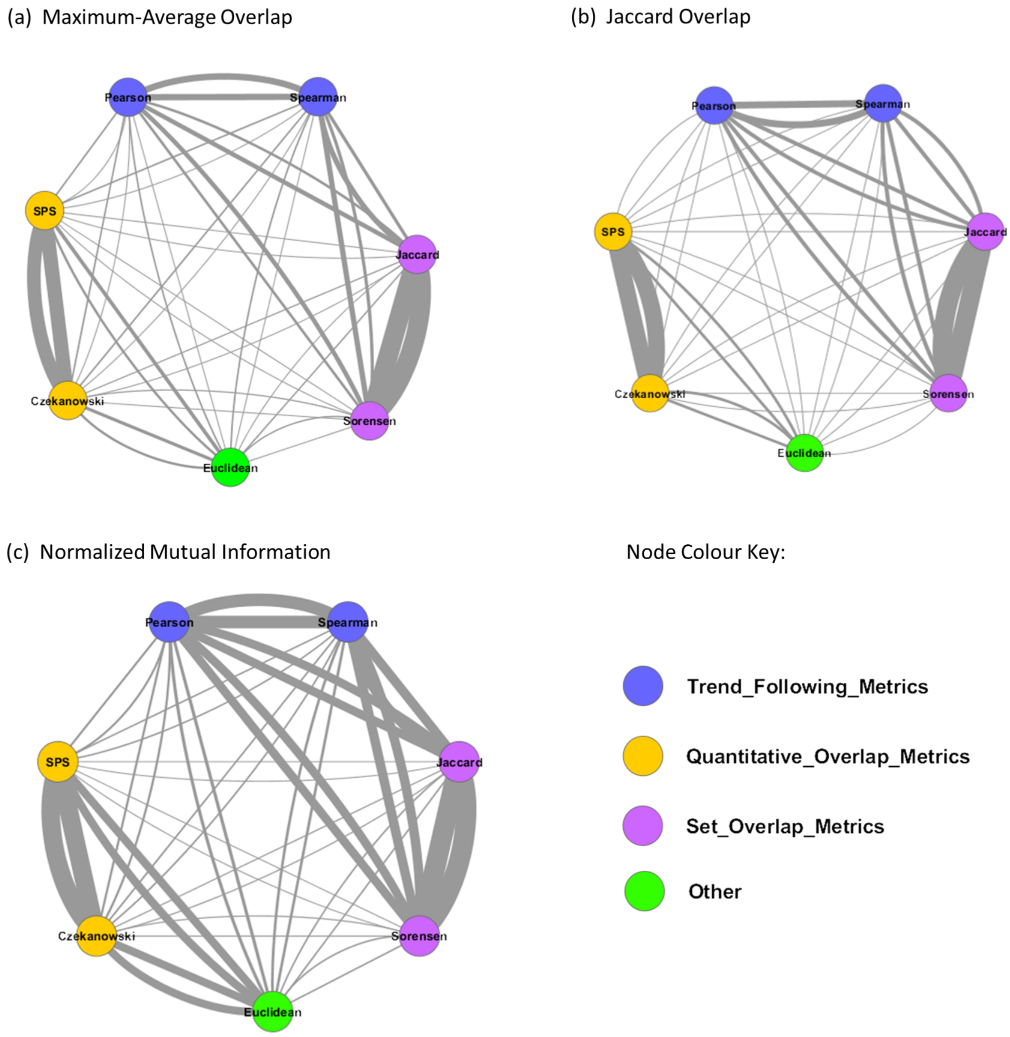


Figure 5.3: **Clustering Similarity.** Each node represents a network (in particular a gene-co-expression network) constructed using a particular similarity metric as the measure of gene co-expression. The similarity between these 7 similarity metrics (nodes) is quantified by calculating the similarity between the MCL clusterings of these networks through the use of (a) Maximum Average Clustering Overlap, (b) Jaccard Clustering Overlap and (c) Normalized Mutual Information between clusterings. Edge thickness corresponds to the weight of the edges based on the particular clustering similarity measure.

similarity metrics.

This resulted in 8 fungal MSTs and 7 bacterial MSTs in which each node represented a species (either fungal or bacterial) and each edge represented similarity between the gene family content of the two species the edge connected, quantified using a particular similarity metric. In the fungal MSTs (Figure 5.4), nodes were coloured according to high order taxonomic groupings, whereas in the bacterial MSTs (Figure 5.5), nodes were coloured according to genus. From the MSTs, it can be seen that, in general, all similarity metrics seem to group the species within their taxonomic groupings or genera. Thus, globally, the choice of similarity metric does not make much difference. However, locally, the choice of similarity metric does result in different topologies. In order to better visualize this, all fungal MSTs and all bacterial MSTs were merged into two Union MSTs, one for fungi (Figure 5.4i) and one for bacteria (Figure 5.5h). These merged views give a good visualization on how much the similarity metrics agree on a global and a local scale. The presence of multiple edges between nodes indicates that multiple similarity metrics place these two nodes adjacent in their respective MSTs. From the Union networks in Figures 5.4i and 5.5h, it can clearly be seen through the colour distributions that these similarity metrics generally agree on a global scale, grouping species within their taxonomic/genera groupings. However, the similarity metrics do differ on a local scale. This is illustrated by the connections between nodes which are present in only a few of the MSTs.

5.2.3.1 Cross-Network Topological Overlap

Topological Overlap [14] is a network measure which quantifies the extent to which two nodes within a network are connected through direct connections between the two nodes and indirect connections through shared neighbours of the two nodes. We extended this concept and introduce a formulation of Topological Overlap, called Cross-Network Topological Overlap, which calculates the topological overlap between nodes in different networks, quantifying the similarity between the neighbourhoods of two nodes in different networks (Figure 5.6). Selecting best-hits for a node in another network thus selects nodes which are topologically most similar to that node. This provides a node-by-node based approach for comparing networks.

Consider two networks, A and B. Let A_i denote the i^{th} node in network A and B_j denote the j^{th} node in network B. We define Cross-Network Topological Overlap (CNTO) in a directional manner. Let $CN_{TO}(A_i, B_j)$ be the CNTO of node A_i *onto* node B_j. Then, the two directional CNTOs are defined as:

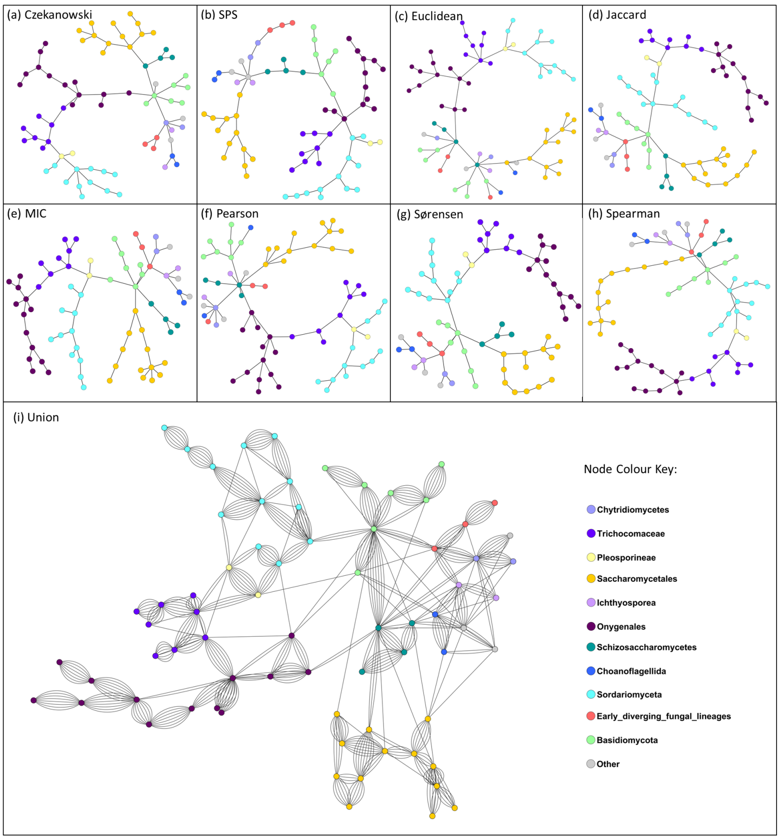


Figure 5.4: **Fungal Gene Family Content MSTs.** Each MST shows the similarity between the gene family content of fungal species, each calculated using a different similarity metric. In each network, each node represents a fungal species and each edge represents similarity between the gene family content of two species calculated using a different similarity metric, namely (a) Czekanowski Index (b) SPS Index (c) Euclidean Similarity (d) Jaccard Index (e) Maximum Information Coefficient (f) Pearson Correlation Coefficient (g) Sorensen Index (h) Spearman Correlation Coefficient. (i) shows a union of the MSTs in (a)-(h). Species nodes are coloured according to their taxonomic groupings. All networks were visualized in Cytoscape [13].

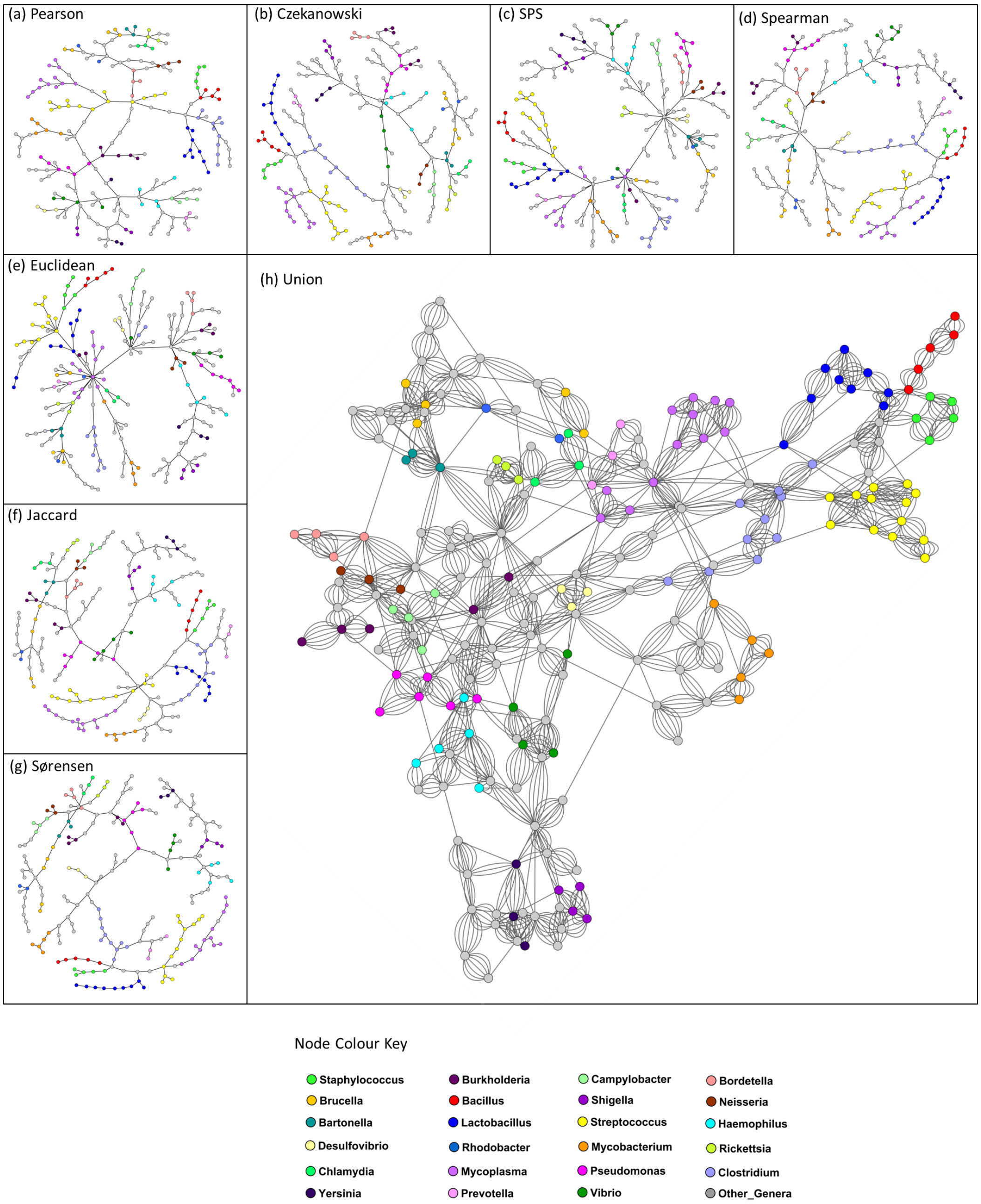


Figure 5.5: **Bacterial Gene Family Content MSTs.** Each MST shows the similarity between the gene family content of bacterial species, each calculated using a different similarity metric. In each network, each node represents a bacterial species and each edge represents similarity between the gene family content of two species calculated using a different similarity metric, namely (a) Pearson Correlation Coefficient (b) Czekanowski Index (c) SPS Index (d) Spearman Correlation Coefficient (e) Euclidean Similarity (f) Jaccard Index (g) Sorensen Index. (h) shows a union of the MSTs in (a)-(g). Species nodes are coloured according to their genus. All networks were visualized in Cytoscape [13].

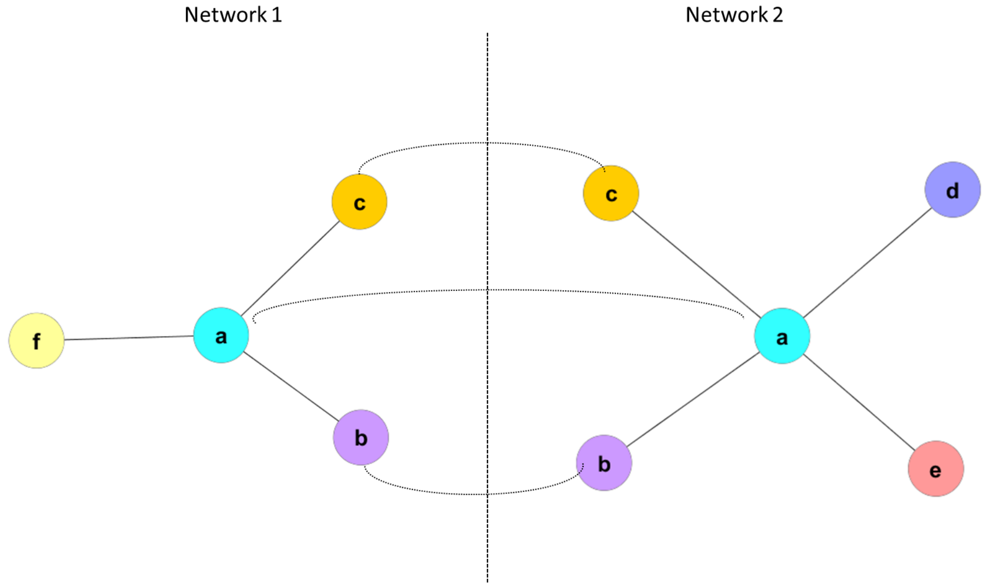


Figure 5.6: **Cross-Network Topological Overlap** Subnetworks of the neighbourhood of node i in two hypothetical networks are shown. Solid edges represent edges within a network, and dashed edges represent edges constructed to link each node with its corresponding node in the other network.

$$CN_{TO}(A_i, B_j) = \frac{n_{A_i,B_j} + d_{A_i,B_j}}{k_{A_i} + 1} \tag{5.2.4}$$

$$CN_{TO}(B_i, A_j) = \frac{n_{A_i,B_j} + d_{A_i,B_j}}{k_{B_j} + 1} \tag{5.2.5}$$

where n_{A_i,B_j} is the number nodes which are neighbours of both A_i and B_j, k_{A_i} is the connectivity of node A_i, k_{B_j} is the connectivity of node B_j and d_{A_i,B_j} defined by:

$$d_{A_i,B_j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{5.2.6}$$

Thus, the directed CNTO will be equal to 1 if the two nodes are in fact the same node, and if they share all their neighbours. The symmetrical CNTO of two nodes is then defined as the average of the two respective directional topological overlaps.

In order to investigate this new CNTO measure and how it can be used to compare networks, we applied it to compare the phylogenomic MSTs as these networks were simple and small enough that the output could be visually inspected.

CNTO was used to compare the Jaccard and Sørensen fungal MSTs, as well as the Pearson and Sørensen fungal MSTs. The Pearson and Sørensen networks were chosen for comparison because these metrics have very different definitions and were shown to result in very different network topologies when applied to the transcriptomic dataset. For a given pair of MSTs, A and B, the CNTO was calculated for all pairs of nodes i and j in which i is a node in A and j is a node in B. For each node i in a MST A, the node(s) in MST B with the highest topological overlap with node i were selected. Pairwise CNTO networks were then constructed. These networks contained two copies of each node, one from each of the two MSTs being compared, and each node is connected to the node(s) from the other MST with which they have the highest CNTO. The CNTO networks for the comparison of the Jaccard and Sørensen MSTs and the Pearson and Sørensen MSTs can be seen in Figure 5.7a and 5.7b respectively. These networks very clearly show the degree of similarity in the topologies of two networks. Figure 5.7a illustrates that the topologies of the Jaccard MST and the Sørensen MST are identical, since each node from the Jaccard MST (black bordered nodes) connects only to its corresponding node in the Sørensen MST (grey bordered nodes) with $CNTO = 1$. Figure 5.7b illustrates the similarities and differences in the topologies of the Pearson and Sørensen MSTs. Certain nodes are topologically similar between these two MSTs (illustrated by the pairs of nodes at the bottom of the network in Figure 5.7b), however, the disagreement of the two similarity metrics is shown largely in the top half of the network.

In order to illustrate how CNTO selects the most topologically similar node in another network, consider the three labelled nodes in Figure 5.7b. The network shows that the nodes in the Sørensen MST most topologically similar to the species node *Capaspora owczarzaki* in the Pearson MST are *Lodderomyces elongisporus* and *Schizosaccharomyces octosporus*. The position of *Capaspora owczarzaki* in the Pearson MST is illustrated in Figure 5.8a. The only information we have topologically about this node is that it is a neighbour of the node *Schizosaccharomyces japonicus*. Thus, logically, the most topologically similar nodes in the Sørensen MST should be neighbours of *Schizosaccharomyces japonicus*. Consider the Sørensen MST in Figure 5.8b. Neighbours of *Schizosaccharomyces japonicus* are either *Schizosaccharomyces octosporus*, *Lodderomyces elongisporus* or *Cryptococcus neoformans*. CNTO chose *Schizosaccharomyces octosporus* and *Lodderomyces elongisporus* as more topologically similar than *Cryptococcus neoformans*, since their degrees are lower, thus have a higher fraction of shared neighbours with *Capaspora owczarzaki* than *Cryptococcus neoformans* does.

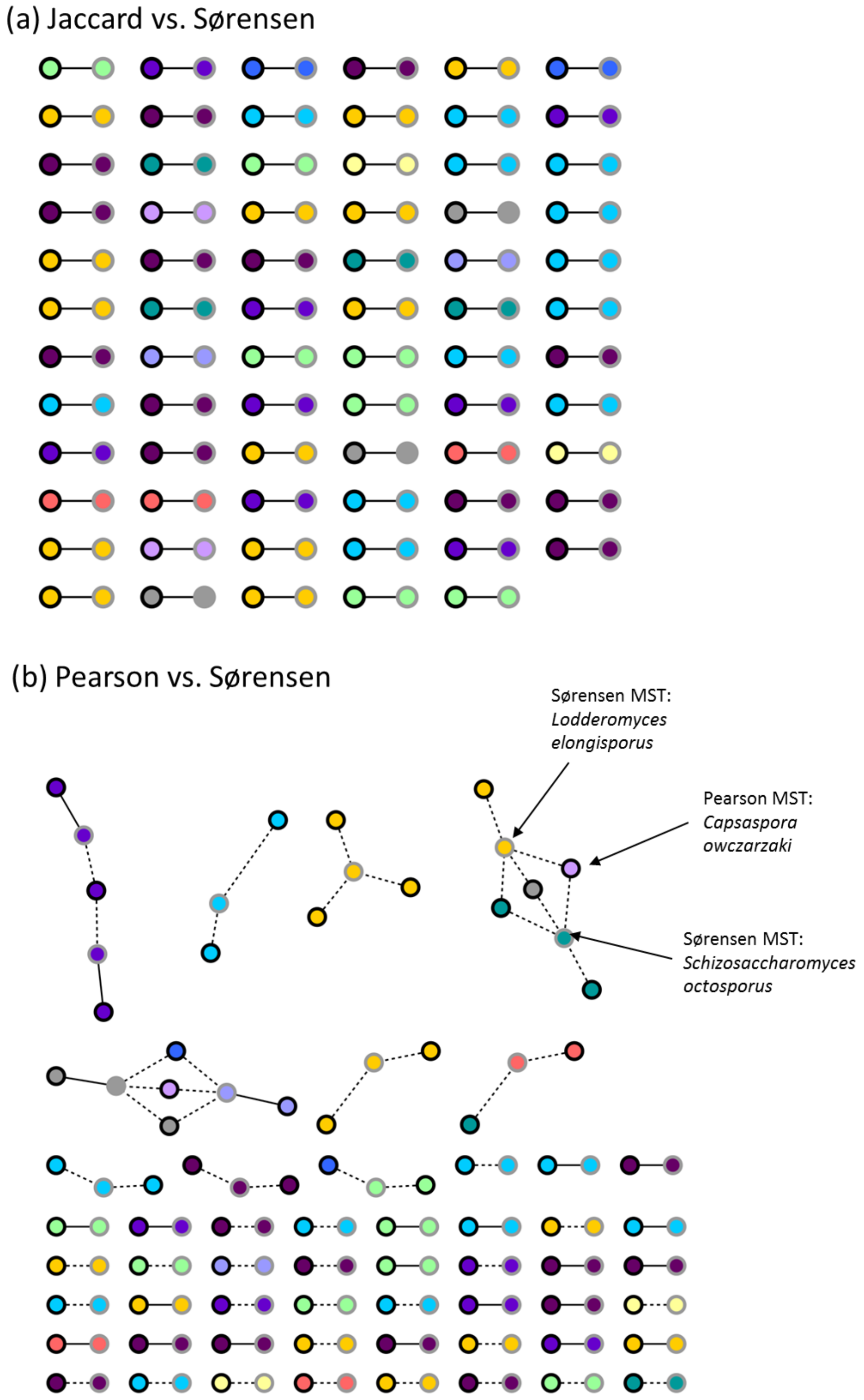

(a) Jaccard vs. Sørensen
(b) Pearson vs. Sørensen
Sørensen MST: *Lodderomyces elongisporus*
Pearson MST: *Capsaspora owczarzaki*
Sørensen MST: *Schizosaccharomyces octosporus*

Figure 5.7: **CNTO Networks: Comparison of Fungal MSTs.** CNTO networks resulting from the comparison of fungal MSTs. Each node represents a fungal species from a MST corresponding to one similarity metric and is connected to the node(s) in a MST from another metric with which it has the highest CNTO. (a) CNTO network from the comparison of Jaccard and Sørensen fungal MSTs. Black-bordered nodes represent fungal species nodes from the Jaccard MST and grey bordered nodes represent fungal species nodes from the Sørensen MST. (b) CNTO network from the comparison of Pearson and Sørensen fungal MSTs. Black-bordered nodes represent fungal species nodes from the Pearson MST and grey bordered nodes represent fungal species nodes from the Sørensen MST. Solid edges represent $CNTO = 1$ (nodes are identical and share all their neighbours) while dashed edges represent $CNTO < 1$.

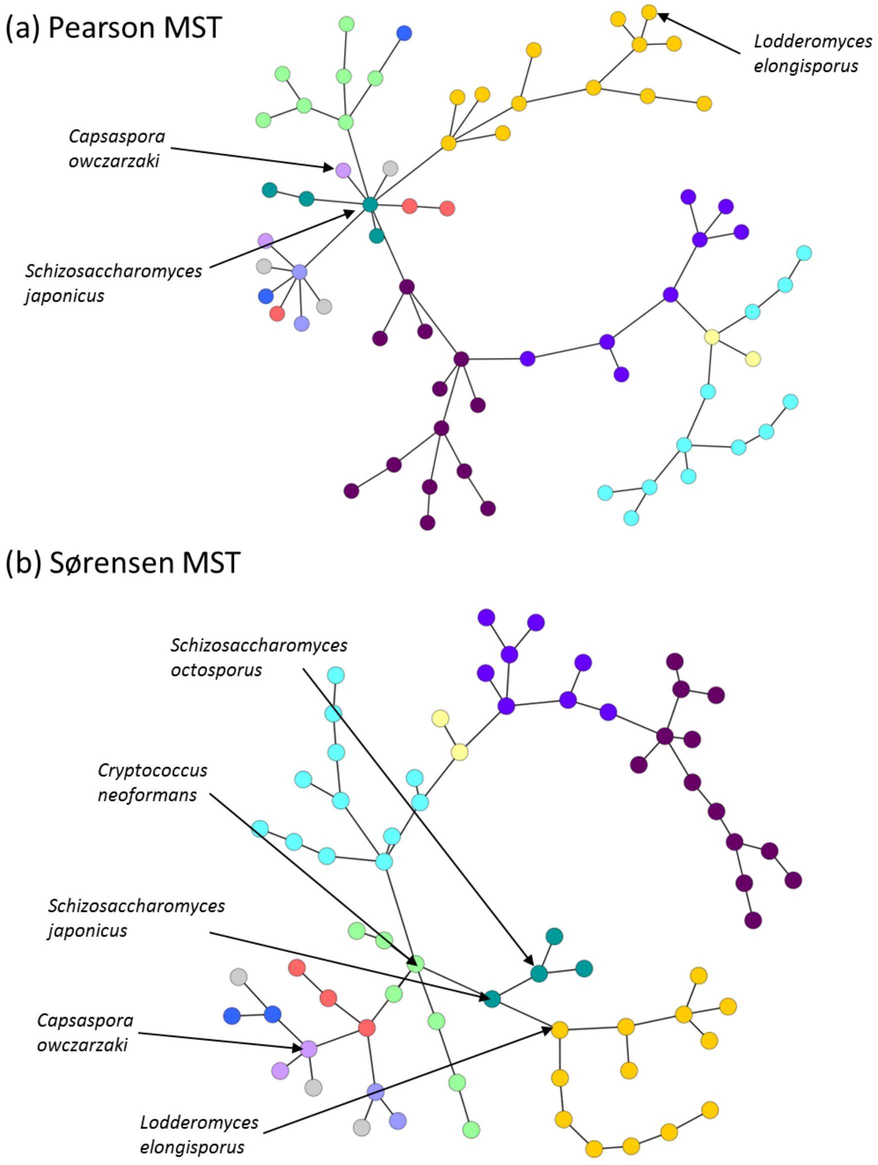

(a) Pearson MST
Lodderomyces elongisporus
Capsaspora owczarzaki
Schizosaccharomyces japonicus
(b) Sørensen MST
Schizosaccharomyces octosporus
Cryptococcus neoformans
Schizosaccharomyces japonicus
Capsaspora owczarzaki
Lodderomyces elongisporus

Figure 5.8: **Pearson and Sørensen Fungal MSTs** Fungal MSTs in which nodes represent fungal species and edges represent similarity between the gene family content of species quantified using (a) the Pearson Correlation Coefficient and (b) the Sørensen Index.

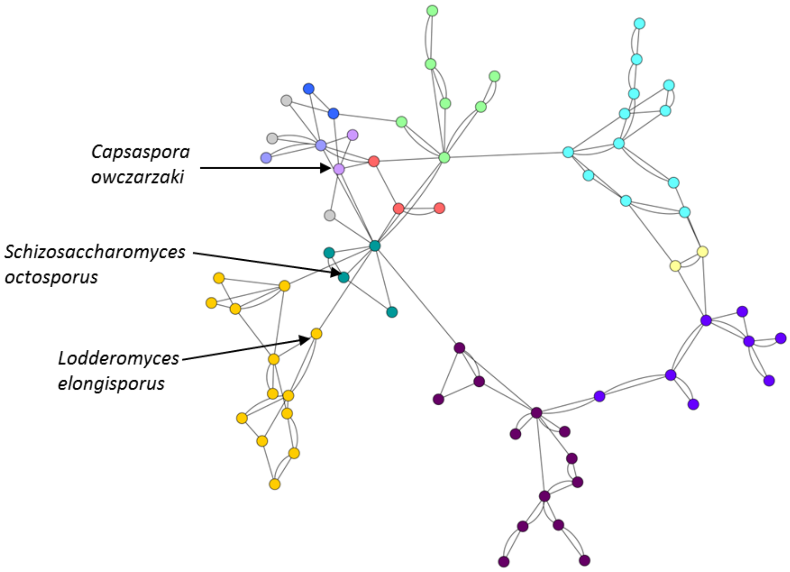


Figure 5.9: **Union of Pearson and Sørensen MSTs.** Merged Sørensen and Pearson fungal MSTs from Figure 5.8.

As illustrated, for a given node in a particular network, CNTO selects the node(s) in a corresponding network with the most similar topological surroundings in terms of fraction of shared neighbours. CNTO networks like those in Figure 5.7 reveal different information than would be gained from simply merging the two networks being compared. For example, consider the merged fungal Sørensen MST and Pearson MST shown in Figure 5.9. This merged view gives an indication of shared edges, but does not easily show which nodes are most topologically similar in terms of shared neighbours as is shown by the CNTO networks.

MSTs, in general, have very simple topologies. They have no cycles and are very minimalistic in topology. They were chosen as example networks to develop and explore this method of network comparison because of their simplicity and ease of visualization. In order to explore the results of this method on networks with more complex topology, the Pearson and Sørensen all-vs-all bacterial networks were pruned to maintain the top 2.5% of edges. The resulting networks can be seen in Figure 5.10. These networks have more complex topologies than the MSTs, having a much larger variance in node connectivities. CNTO was then calculated between all pairs of nodes in these two pruned networks, and a CNTO network constructed (Figure 5.11). This network clearly indicates the differences in the local topologies of nodes in the two pruned bacterial networks being compared.

Consider the labelled nodes in Figure 5.11. The species nodes in the Sørensen bacterial network (Figure 5.10b) most similar to *Lactobacillus acidophilus* in the Pearson bacterial network (Figure 5.10a) are *Enterococcus faecium* and *Lactococcus lactis*. The common and uncommon neighbours of these nodes are illustrated in Figure 5.12. On the left hand side of each panel is the node in question, *Lactobacillus acidophilus* connected to its neighbours in the Pearson bacterial network in (Figure 5.10a). On the right hand side of each panel is a node from the Sørensen bacterial network (Figure 5.10b) connected to its neighbours. The neighbours shared between *Lactobacillus acidophilus* from the Pearson network and the node from the Sørensen network in the right panel are enclosed in a rectangle. This Figure illustrates why the CNTO measure selected *Enterococcus faecium* and *Lactococcus lactis* in the Sørensen network as more topologically similar to *Lactobacillus acidophilus* in the Pearson network, than its equivalent node, *Lactobacillus acidophilus*, in the Sørensen network. As can be seen in Figure 5.12, *Lactobacillus acidophilus* from the Pearson network shares proportionally many more neighbours with *Enterococcus faecium* and *Lactococcus lactis* in the Sørensen network than with *Lactobacillus acidophilus* from the Sørensen network.

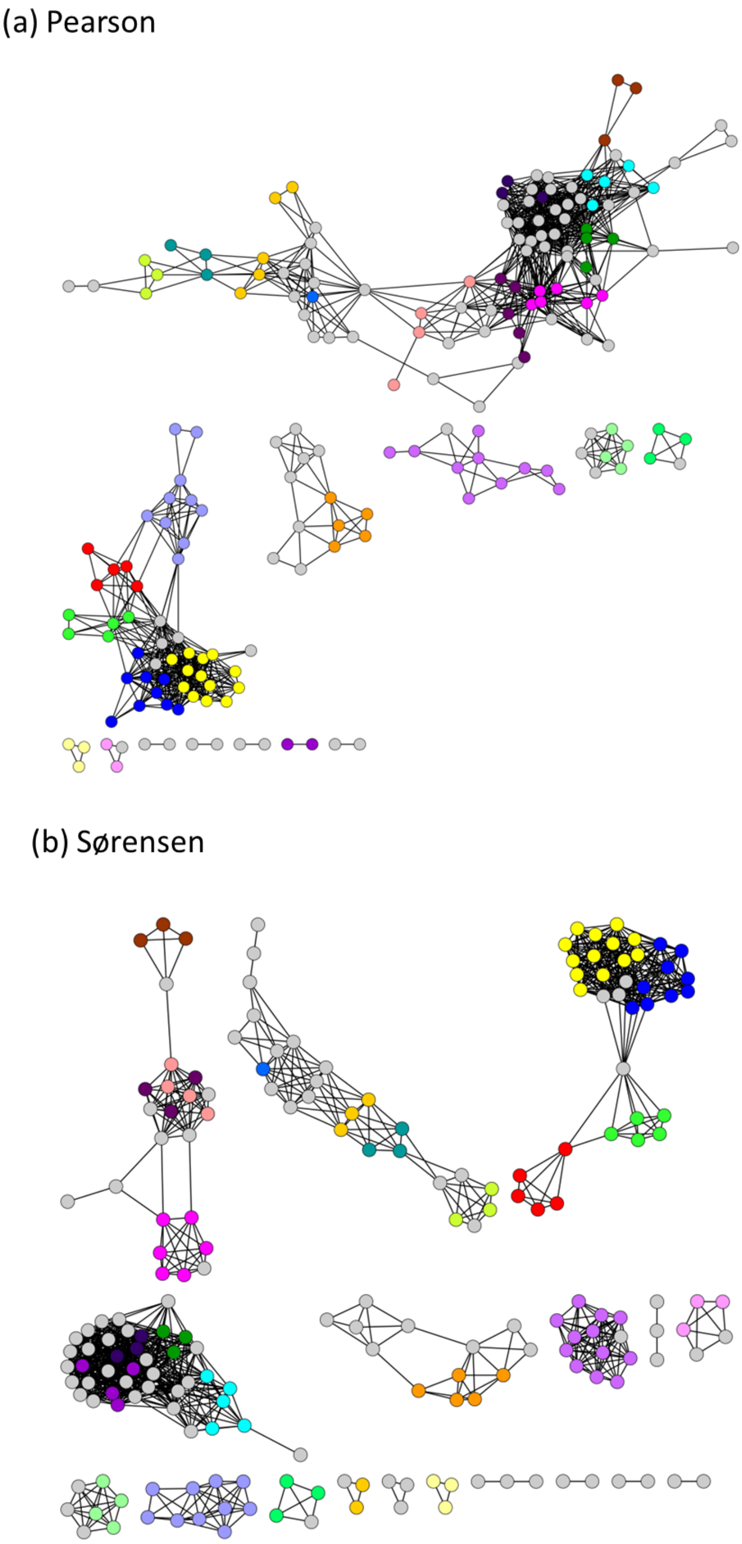


Figure 5.10: **Pearson and Sørensen Pruned Bacterial Networks** Pruned phylogenomic networks in which nodes represent bacterial species and edges represent similarity between the gene family content of bacterial species quantified using (a) the Pearson Correlation Coefficient and (b) the Sørensen Index. Nodes are coloured according to genus. These networks are pruned to maintain only the top 2.5% of edges.

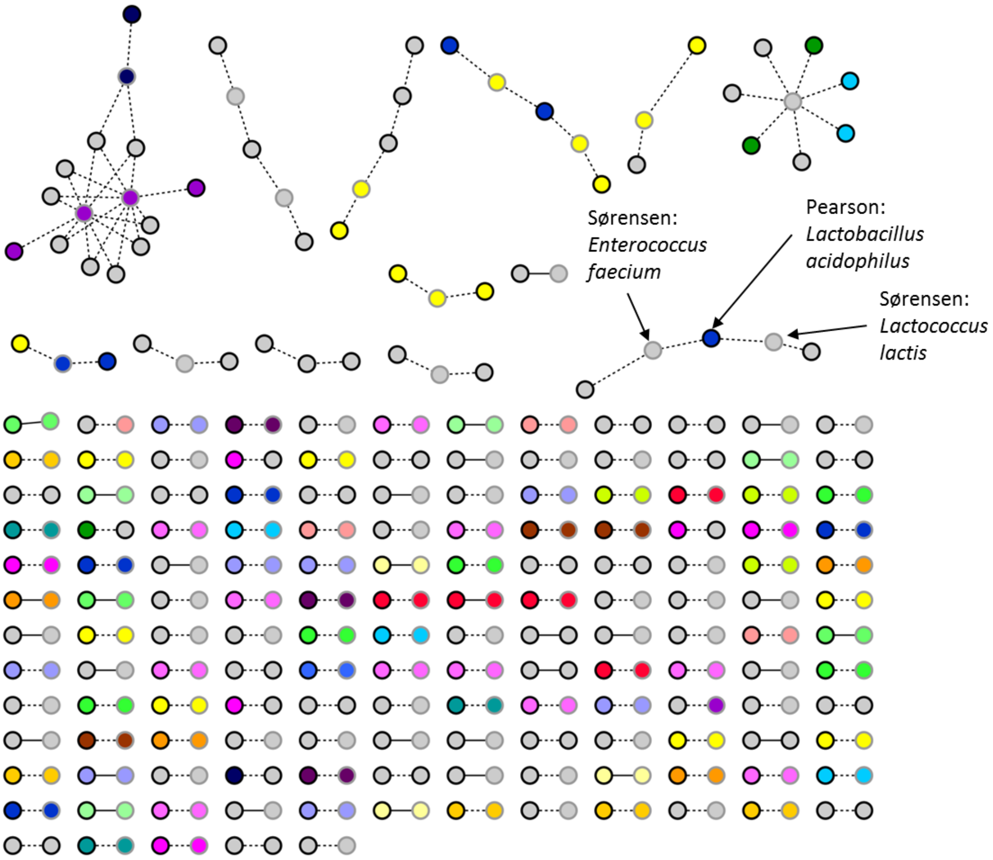


Figure 5.11: **CNTO Network: Pearson and Sørensen Bacterial Networks** Comparison of the pruned bacterial networks in Figure 5.10 through CNTO. Black bordered nodes represent nodes from the pruned Pearson bacterial network (Figure 5.10a) and grey bordered nodes represent nodes from the pruned Sørensen bacterial network (Figure 5.10b). Each node is connected to the node(s) in the other network with which it has the highest CNTO. Solid edges represent $CNTO = 1$ (nodes are identical and share all their neighbours) while dashed edges represent $CNTO < 1$.

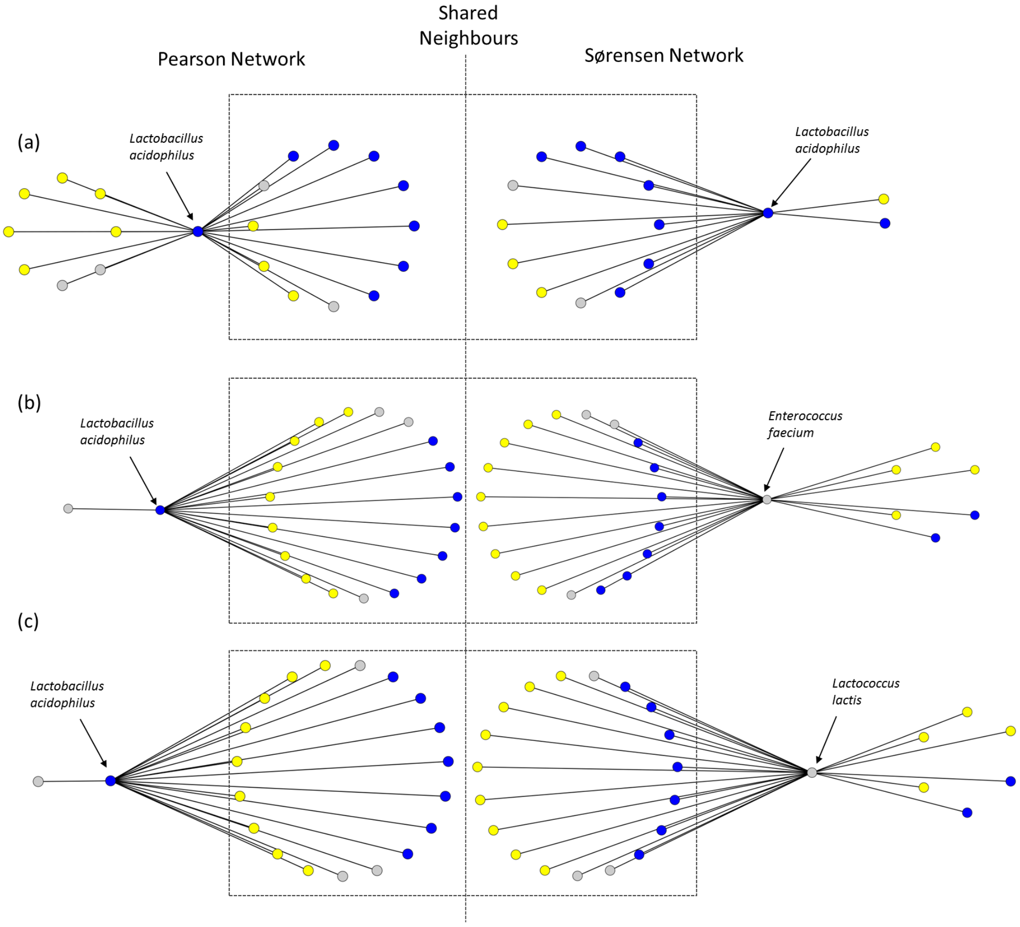

Shared Neighbours
Pearson Network
Sørensen Network
(a)
Lactobacillus acidophilus
Lactobacillus acidophilus
(b)
Lactobacillus acidophilus
Enterococcus faecium
(c)
Lactobacillus acidophilus
Lactococcus lactis

Figure 5.12: **Shared Neighbours** Neighbours of *Lactobacillus acidophilus* in the Pearson network, as shown in Figure 5.10a, which are shared with nodes in the Sørensen network, as shown in Figure 5.10b, are illustrated within this Figure. *Lactobacillus acidophilus* and its neighbours in the Pearson network are shown on the left hand side, nodes in the Sørensen network and their neighbours are shown on the right hand side, and neighbours shared between the node on the left and the node on the right are enclosed in rectangles.

5.3 Conclusions

In this study, different similarity metrics were applied to construct networks from three different datasets, and the effect of different similarity metrics on the resulting network topology was investigated through various network comparison approaches. Two new network comparison approaches and one existing approach were investigated, including PCA of network topology profiles, Cross Network Topological Overlap and Clustering Comparison [15]. It is evident from all of these investigations that the similarity metric chosen can have a large impact on the topology of the resulting network. These differences in network topology also carry through to the results of further analysis, such as clustering. The choice of similarity metric could thus greatly impact the resulting biological interpretation of the networks. A potential limitation of using network topology measures to compare networks is that certain topology measures, such as shortest path, are computationally time consuming to calculate and may become infeasible for very large networks. However, with the appropriate High Performance Computing resources, they can be applied to larger networks.

The fact that different similarity metrics will result in different biological interpretations can be exploited as an advantage. Since each similarity metric describes and quantifies a different aspect of similarity, the use of multiple similarity metrics will provide multiple perspectives on the data, each of which is valuable. An agglomerative approach in which many different similarity metrics are used to gain different perspectives and insights into a dataset is thus appealing.

Furthermore, with the Cross-Network Topological Overlap method that we have presented here, it is relatively easy to identify the portions of the network which are affected by the choice of similarity metric. This approach provides different information than would be gained from merging two or more networks being compared. CNTO specifically highlights areas of the networks with conflicting topologies in a node-based manner, connecting nodes to their most topologically similar nodes in another network, whereas network merging is an edge-based approach, simply revealing shared edges between the two networks.

5.3.1 Future Work

This ability of CNTO to highlight and zoom in on these areas of interest is a very useful attribute, especially when comparing large networks. In addition, CNTO potentially has broader applications to network comparisons in a wide variety of real-world networks, including communication networks, transport networks and social networks. This approach can be applied to compare any kind of network, and highlight the areas of these networks which have conflicting topologies. The further application of CNTO in the comparison of various types of networks is suggested for future work.

5.4 Methods

5.4.1 Metric Comparison though Network Topology Profiles and Clustering Comparison

5.4.1.1 Co-expression Similarity Network Construction

472 grapevine Affymetrix microarray experiments were downloaded from Gene Expression Omnibus and normalized using RMA [16]. In the resulting expression matrix E, the columns represented microarray experiments, rows represented probesets and each entry Xi represented the $\log_2$(expression) value of probeset X in experiment i. Seven metrics were then used to calculate the similarity (“co-expression") between all pairs of probesets.

Let X and Y denote rows of the expression matrix E corresponding to the expression profiles of genes x and y respectively. Let X_B and Y_B denote the binary vectors corresponding to X and Y, calculated as:

$$X_{Bi} = \begin{cases} 1 & \text{if } X_i \geq \bar{X} \\ 0 & \text{if } X_i = 0 \end{cases} . \tag{5.4.1}$$

where X_{Bi} is the i^{th} entry of X_B and $\bar{X}$ is the mean of X. Seven similarity metrics (defined in Table 5.1) were then calculated between all pairs of genes. The Pearson and Spearman Correlation Coefficients and Czekanowski, SPS and Euclidean Distance Indices were calculated using the original vectors, and Sørensen and Jaccard Indices were calculated using the binary vectors defined in Equation 5.4.1. The mcxarray program from MCL-Edge [12] was used to calculate the Pearson and Spearman correlation coefficients. Customized Perl scripts were written to calculate the other similarity metrics. This resulted in 7 similarity networks (one for each similarity metric) in which each node represented a probeset and each edge represented similarity between the expression profiles of the probesets the edge was connecting, according to a particular similarity metric. These similarity networks were subsequently pruned to maintain only the top 1 percent of edges, including reciprocal edges but not including self-loops.

5.4.1.2 Metric Distribution Construction

A distribution was constructed for each similarity metric using a bin size of 0.05. All similarity metrics with a range of 0 to 1 (Sørensen, Jaccard, Czekanowski and SPS Indices and Euclidean Similarity) thus had 20 bins. Pearson and Spearman Correlation Coefficients have a range of -1 to 1 and thus needed 40 bins.

5.4.1.3 Network Comparison through Topology Indices

For each of the 7 pruned co-expression networks, a series of network topology indices were calculated. Weighted versions of the topology indices use the actual similarity value as the weight of the edges, whereas unweighted topology indices do not acknowledge edge weights, only the presence or absence of edges in the pruned networks.

The weighted and unweighted versions of the following global (whole-network) indices were calculated for each of the co-expression networks:

1. Density
2. Centralization
3. Heterogeneity
4. Degree Correlation
5. Maximum Connectivity

The following weighted local (node-based) indices were calculated for each node in each network:

1. Clustering Coefficient
2. Scaled Connectivity
3. Connectivity
4. Maximum Adjacency Ratio

The same unweighted local indices were calculated for each network, with the exception of Maximum Adjacency Ratio, which is meaningless in the context of an unweighted network.

Topology profile matrices were then constructed in which each row represents one of the input networks and columns represent topology indices. Four topology profile matrices were constructed which the variables were weighted local indices, unweighted local indices, weighted global indices and unweighted global indices, respectively. PCA was performed on these matrices using Qlucore (Qlucore AB, 2008, [17]).

5.4.1.4 Network Comparison through Clustering Comparison

The pruned similarity networks were clustered using MCL [12]. This produced a clustering (a set of clusters) for each similarity network. Perl scripts were then written to compare all pairs of clusterings using three measures, namely Average-Maximum Overlap, Jaccard Clustering Overlap and Normalized Mutual Information.

Let C_i and C_j be two clusterings. Then, the Average-Maximum Overlap between clusterings C_i and C_j was calculated as follows: For each pair of clusters (a, b) where $a \in C_i$ and $b \in C_j$, the Jaccard Index was calculated as

$$J(a,b) = \frac{|a \cap b|}{|a \cup b|} \tag{5.4.2}$$

This results in the matrix in which the rows represent clusters from clustering C_i, columns represent clusters from clustering C_j and each entry (a, b) is the Jaccard overlap of a from clustering C_i and b from clustering C_j. The maximum value of each row is then taken, representing the “best hit" overlap for each cluster in clustering C_i. The average of these maxima is then taken, giving a score for how similar clustering C_i is to C_j. The matrix is then transposed and the process repeated, since this similarity score is not symmetric. A network was then created in which each node represented a co-expression network (constructed using a specific similarity metric) and each edge represented the Average-Maximum Overlap score between the clusterings of the two nodes (networks) the edge is connecting. The network was visualized in Cytoscape [13] and can be seen in Figure 5.3a.

The Jaccard clustering overlap between clusterings C_i and C_j was calculated as [18]:

$$J(C_i, C_j) = \frac{N_{11}}{N_{11} + N_{01} + N_{10}} \tag{5.4.3}$$

where N_{11} is the number of pairs of elements (x, y) which are in the same cluster in C_i and C_j, N_{10} is the number of pairs of elements (x, y) which are in the same cluster in C_i but not C_j and N_{01} is the number of pairs of elements (x, y) which are in the same cluster in C_j but not C_i. A network was created in which each node represented a co-expression network (constructed using a specific similarity metric) and each edge represented the Jaccard overlap between the clusterings of the two nodes (networks) which the edge is connecting. The network was visualized in Cytoscape [13] and can be seen in Figure 5.3b.

The normalized Mutual Information clustering overlap between clusterings C_i and C_j was calculated as [15]:

$$NMI(C_i, C_j) = \frac{\sum_a \sum_b P(a,b) \log_2 \left(\frac{P(a,b)}{P(a)P(b)} \right)}{\sqrt{\sum_a P(a) \log_2 (P_a) \sum_b P(b) \log_2 (P_b)}} \tag{5.4.4}$$

where a is a cluster in clustering C_i, b is a cluster in clustering C_j, $P(a)$ is defined as $\frac{|a|}{n}$, $P(b)$ is defined as $\frac{|b|}{n}$ and $P(a,b)$ is defined as $\frac{|a \cap b|}{n}$. The Normalized Mutual Information was calculated between all pairs of the 7 clusterings (one for each co-expression network) and a network was constructed in which each node represented a co-expression network (calculated using a specific similarity metric) and each edge represented the normalized mutual information between the clusterings of the two nodes (networks) connected by that edge. The resulting network was visualized in Cytoscape and can be seen in Figure 5.3c.

5.4.2 Metric Comparison through Network Merging and Cross-Network Topological Overlap

Two datasets, one consisting of the fully sequenced genomes of 71 fungal species (downloaded from the Broad Institute [http://www.broadinstitute.org/] and the *Saccharomyces* Genome Database [http://www.yeastgenome.org/download-data], and the other consisting of the fully sequenced genomes of 211 bacterial species (downloaded from NCBI, [http://www.ncbi.nlm.nih.gov/]) were used obtained, and gene families constructed.

5.4.2.1 Gene Family Construction

Gene families were constructed across 71 fungal species using a parallel version of OrthoMCL [19]. Gene families were constructed across the 211 bacterial species using TribeMCL [20]. TribeMCL constructs less stringent families than OrthoMCL does, however, TribeMCL is faster, and was thus chosen for the larger dataset of 211 bacterial genomes. In both cases, an inflation value of 2 was used during the MCL [12] clustering step. All families of size 2 or less were excluded from further analysis. From the resulting gene families, two matrices (named Species-Family Matrices or SF-matrices) of gene family content profiles were constructed, one containing fungal gene family profiles and the other containing bacterial gene family profiles. In both matrices, each column represented a species, each row represented a gene family and each entry ij represented the number of genes in gene family i present in species j.

5.4.2.2 Phylogenomic Network Construction and Pruning

The similarity between the gene family content of all pairs of fungal species was calculated using 8 different similarity metrics. Let X_i and Y_i represent the i^{th} element in column X and column Y in the SF-matrix (i.e. the number of members of gene family i in species X and species Y respectively. Let X_B be the binary vector associated with vector X and Y_B be the binary vector associated with vector Y, calculated as:

$$X_{Bi} = \begin{cases} 1 & \text{if} X_i \geq 1 \\ 0 & \text{if} X_i = 0 \end{cases} \tag{5.4.5}$$

Eight similarity metrics (defined in Table 5.1) were then used to calculate the similarity between the gene family content of all pairs of fungal species X and Y. Pearson and Spearman Correlation Coefficients, MIC, Euclidean Similarity and Czekanowski and SPS Indices were calculated using the original vectors, and the Sørensen and Jaccard Indices were calculated using the binary vectors defined in equation 5.4.5.

The same procedure was performed to calculated the similarity between all pairs of bacterial species, however, in this case, only 7 similarity metrics were used, as the MINE package which is used to calculate MIC failed to run on the bacterial dataset because of memory limitations.

The mcxarray program from MCL-Edge [12] was used to calculate the Pearson and Spearman correlation coefficients. The MINE Java program [8] was used to calculated the Maximum Information Coefficient.

Applying each of these similarity metrics yielded 8 all-vs-all similarity networks for the fungal dataset and 7 all-vs-all similarity networks for the bacterial dataset in which each node represented a species and each edge represented the similarity between the two species which the edge connected based on the particular similarity metric. The all-vs-all networks were then pruned by calculating a Maximum Spanning Tree (MST) for each similarity network using the Perl program for MST construction used in [21]. This Perl program calculates MSTs by converting each edge weight w from a similarity value to distance value $w' = 1 - w$ and calculating a Minimum Spanning Tree on the resulting distance network using the Dijkstra algorithm from the Graph Perl Module (Jarkko Hietaniemi, http://www.cpan.org/). The resulting fungal MSTs were visualized using Cytoscape [13] and are shown in Figure 5.4 and the bacterial MSTs are shown in Figure 5.5. The fungal species nodes were coloured by their taxonomic groupings determined using the NCBI Taxonomy Browser [22]. Bacterial species nodes were coloured according to genus. The default colour is grey, thus the colour grey does not indicate any specific genus or taxonomic grouping.

Two other pruned networks were created from the all-vs-all Sørensen and Pearson bacterial networks. Each of these networks were pruned by selecting the top 2.5% of edges (not including reciprocal edges or self-loops).

5.4.2.3 MST Merging

The two Union MSTs (Figures 5.4i and 5.5h) were constructed by merging all fungal and bacterial MSTs, respectively, using the Cytoscape Advanced Network Merge Plugin.

5.4.2.4 Cross-Network Topological Overlap Networks

The Cross-Network Topological Overlap was calculated between all pairs of nodes for a selection of pairs of networks, namely:

1. Jaccard Fungal MST vs. Sørensen Fungal MST
2. Pearson Fungal MST vs. Sørensen Fungal MST
3. Pearson Bacterial pruned network vs. Sørensen Bacterial pruned network (networks pruned to maintain only the top 2.5% of edges).

Pairs of nodes which shared no neighbours across two networks in question were excluded. For each node, the nodes in the other network with the highest topological overlap were selected, and the resulting CNTO networks were visualized in Cytoscape [13].

5.5 Acknowledgements

The authors would like to acknowledge The Center for High Performance Computing and the Stellenbosch High Performance Computing Cluster for computing resources, as well as the South African National Research Foundation Technology (www.nrf.ac.za) and Human Resources Programme and Winetech. The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

5.6 Author's contributions

D Weighill and D Jacobson conceived of and designed the methods, D Weighill wrote the code and created the networks, D Weighill and D Jacobson discussed and interpreted the networks, D Weighill drafted the manuscript, D Jacobson critically revised and edited the manuscript.

5.7 Competing Interests

The authors declare that they have no competing financial interests.

Bibliography

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

[2] Bloom, S.A.: Similarity indices in community studies: potential pitfalls. *Marine Ecology Progress Series*, vol. 5, no. 2, pp. 125–128, 1981.

[3] Rodgers, J.L. and Nicewander, W.A.: Thirteen ways to look at the correlation coefficient. *The American Statistician*, vol. 42, no. 1, pp. 59–66, 1988.

[4] Pinto da Costa, J. and Soares, C.: A weighted rank measure of correlation. *Australian & New Zealand Journal of Statistics*, vol. 47, no. 4, pp. 515–529, 2005.

[5] Schubert, A.: *Measuring the Similarity Between the Reference and Citation Distributions of Journals*, vol. 96, no. 1, pp. 305–313, 2013.

[6] Lipkus, A.H.: A proof of the triangle inequality for the tanimoto distance. *Journal of Mathematical Chemistry*, vol. 26, no. 1-3, pp. 263–265, 1999.

[7] Yoshioka, P.M.: Misidentification of the bray-curtis similarity index. *Marine Ecology Progress Series*, vol. 368, pp. 309–310, 2008.

[8] 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.: Detecting novel associations in large data sets. *Science (New York, N.Y.)*, vol. 334, no. 6062, pp. 1518–24, 2011.

[9] Berlingerio, M., Koutra, D., Eliassi-Rad, T. and Faloutsos, C.: A scalable approach to size-independent network similarity. Available: http://arxiv.org/pdf/1209.2684.pdf.

[10] Horvath, S. and Dong, J.: Geometric interpretation of gene coexpression network analysis. *PLoS Computational Biology*, vol. 4, no. 8, p. e1000117, 2008.

[11] Reijneveld, J.C., Ponten, S.C., Berendse, H.W. and Stam, C.J.: The application of graph theoretical analysis to complex networks in the brain. *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology*, vol. 118, no. 11, pp. 2317–31, 2007.

[12] van Dongen, S.: *Graph clustering by flow simulation.* Ph.D. thesis, University of Utrecht, 2000.

[13] Shannon, P., Markiel, A., Ozier, O., Baliga, N., Wang, J., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[14] Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. and Barabási, A.-L.: Hierarchical organization of modularity in metabolic networks. *Science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

[15] Wagner, S. and Wagner, D.: *Comparing clusterings: an overview.* Universität Karlsruhe, Fakultät für Informatik, 2007.

[16] Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, vol. 4, no. 2, pp. 249–264, 2003.

[17] Qlucore. `http://www.qlucore.com/1`, 2008. Accessed February 14, 2013.

[18] Meilă, M.: Comparing clusterings-an information based distance. *Journal of Multivariate Analysis*, vol. 98, no. 5, pp. 873–895, 2007.

[19] Li, L., Stoeckert, C. and Roos, D.: Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome Research*, vol. 13, no. 9, pp. 2178–2189, 2003.

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

[21] Setati, M.E., Jacobson, D., Andong, U.-C. and Bauer, F.: The vineyard yeast microbiome, a mixed model microbial map. *PloS One*, vol. 7, no. 12, p. e52609, 2012.

[22] Federhen, S.: The ncbi taxonomy database. *Nucleic Acids Research*, vol. 40, no. D1, pp. D136–D143, 2012.

Chapter 6

Conclusions and Future Work

6.1 Concluding Remarks

In this work, new network approaches were investigated and applied to genomic and transcriptomic datasets. In particular, approaches for merging networks into combined models, application of extended network definitions and network meta-modelling approaches for network comparison were investigated, developed and applied.

In addressing Aim 1 of the study, an approach for constructing a multi-mechanism co-evolutionary network was developed and applied, resulting in a multi-mechanism co-evolutionary network for grapevine. This combined co-evolutionary network was constructed by determining modules of co-evolving grapevine genes for the different mechanisms, (point mutations, gene duplication and deletion, and gene expression regulation) and subsequently linking these modules through module overlaps. To our knowledge, this was the first time these three mechanisms of evolution had been modelled simultaneously in a single network model. Exploration of local neighbourhoods of this combined co-evolutionary network revealed groups of functionally related genes, suggesting the success of the model in bringing together potentially co-evolving genes based on multiple mechanisms of evolution.

An extended network definition (3-way networks) of the standard network model was investigated in which edges were chosen to model relationships between triplets of objects instead of pairs of objects. This appears to be the first time this has been used in the field of phylogenomics. Approaches for constructing, weighting and pruning 3-way networks were investigated and applied to a phylogenomic dataset of 211 bacterial genomes. These 3-way networks were compared to standard 2-way phylogenomic networks constructed from the same dataset. The 3-way networks enabled the quantification and modelling of more complex relationships than possible with the 2-way networks, and revealed relationships missed by the standard network models. This development and application of 3-way networks to a phylogenomic dataset addressed Aim 2.

The last aim of this work was addressed through the exploration of network meta-modelling techniques and the development and application of two new network comparison approaches, namely PCA of network topology profiles and Cross-Network Topological Overlap (CNTO). PCA of network topology profiles provided a global approach for comparing a number of networks, whereas CNTO allowed for a node-by-node comparison of pairs of networks, highlighting areas of the networks with conflicting topologies. The application of these new network comparison approaches as well as already existing approaches (such as clustering comparison [1]) in network meta-modelling of similarity metrics allowed valuable insights to be gained into the effect of different similarity metrics on network topology. The development and application of network meta-modelling techniques fulfilled Aim 3 of this work.

6.2 Future Work

The work done in this Master's study has scope to be extended in many ways. We would like to extend the co-expression analysis to construct modules of co-expressed genes which are conserved across species. This will allow us to extend the model to contain many more species. We would also like to explore different, more advanced approaches for the construction of the ERC modules, such as ContextMirror [2].

The application of 3-way networks to larger phylogenomic datasets, as well as other types of data such as transcriptomic and microbiomic datasets is another target for future work. We would also like to further extend 3-way networks to quantify even higher order relationships, such as 4-way relationships and, eventually, n-way relationships. These models could then also be combined into general hypergraph network models in which edges can model the similarity between an arbitrary number of objects.

The meta-modelling approaches we developed are applicable to various kinds of networks. We would like to apply these methods to network comparisons from a variety of fields and investigate how these network meta-modelling techniques can give insights into the underlying systems the networks are modelling.

Bibliography

[1] Wagner, S. and Wagner, D.: *Comparing clusterings: an overview.* Universität Karlsruhe, Fakultät für Informatik, 2007.

[2] Juan, D., Pazos, F. and Valencia, A.: High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proceedings of the National Academy of Sciences*, vol. 105, no. 3, pp. 934–939, 2008.