Analyzing volatile compound measurements using traditional Multivariate techniques and Bayesian networks

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

Master of Arts
in
Statistics

at Massey University, Albany, New Zealand

Shweta Baldawa
2009

Abstract

The purpose of this project is to compare two statistical approaches, traditional multivariate analysis and Bayesian networks, for representing the relationship between volatile compounds in kiwifruit. Compound measurements were for individual vines which were progeny of an intercross. It was expected that groupings in the data (or compounds) would give some indication of the generic nature of the biochemical pathways. Data for this project was provided by the Flavour Biotech team at Plant and Food Research. This data contained many non-detected observations which were treated as zero and to deal with them, we looked for appropriate value of `c` for data transformation in `log(x+c)`. The data is 'large *p* small *n*' paradigm – and has much in common with data, although it is not as extreme as microarray. Principal component analysis was done to select a subset of compounds that retained most of the multivariate structure for further analysis. The reduced set of data was analyzed by Cluster analysis and Bayesian network techniques. A heat map produced by Cluster analysis and a graphical representation of Bayesian networks were presented to scientists for their comments. According to them, the two graphs complemented each other; both graphs were useful in their own unique way. Along with clusters of compounds, clusters of genotypes were represented by the heat map which showed by how much a particular compound is present in each genotype while the relation among different compounds was seen from the Bayesian networks.

Acknowledgments

I would like to sincerely thank my supervisor, Dr Beatrix Jones for her constant guidance and support right from the beginning through the end of this project. Her encouragement enabled me to widen my understanding of the subject.

I would like to extend my thanks towards Ross Atkinson and Robert Winz from the Flavour Biotech team at Plant and Food Research for allowing me to use their data and providing their comments on the results. I would also like to thank Foundation for Research, Science and Technology (contract C06X0403) for funding this project and Marsden Fast Start Grant to Beatrix Jones (MAU0501) for partial support in completion of this work.

I would like to specially thank my husband Jugal, whose enduring love and support enabled me to complete my thesis.

Finally, I would like to thank my sister, Neha for her help with proof reading.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Volatile organic compounds have high vapour pressures which is sufficient under normal conditions to significantly vaporize and enter the atmosphere. These are measured for kiwifruit to study the flavour and aroma in them. For example, the esters, ethyl butanoate and methyl benzoate were shown to increase sweet aroma and flavour (McMath et al. 1992), and E-hex-2-enal increased "characteristic kiwifruit aroma and flavour" (Young et al. 1995). Plant and Food Research scientists are interested in characterisation of aroma and flavour of kiwifruit. They conduct studies and measure compounds in fruit in the hope of receiving some indication of the generic nature of biochemical pathways.

1.2 Description of data

The dataset for this study was a compound database collected in 2002 from a mapping family planted at the Te Puke Research Centre. It came from a cross made in 1996 of two kiwifruit parents which originated from different parts of China. The female parent was called CK51_05, which was from a 1991 seed introduction from Henan province, and the male parent was called CK15_02, which was from a 1981 seed introduction from Guangxi province. Two unrelated parents with different genetic backgrounds and different phenotypes were deliberately chosen to generate heterozygosity, which could be useful for mapping. This family has 134 females and 137 males. In 2002 fruits were collected from 123 females that were fruiting, they were phenotyped for the usual fruit attributes and volatile organic compounds were measured by solvent extraction and GC/MS.

There were 275 volatile compounds grouped as acid, alcohol, aldehyde, ester, sulphur, ketone, lactone, monoterpene and hydrocarbon. The values recorded are concentration in nanograms per gram. The data contains many observations that are below the measurement threshold and are recorded as zeros. The most interesting compounds like

acid, alcohol, ester and monoterpene were picked from the above group of compounds as examples for the study.

For our study, we combined the biochemically related group of compounds, acid-alcohol and ester into one group. Hence, the focus of this thesis would be on two groups: namely, acid-alcohol-ester and monoterpene. Monoterpene has 45 compounds while acid-alcohol-ester has 194 compounds. 16 compounds were common in both groups.

The objective of this study is to compare two statistical approaches for representing the relationship between volatile compounds in kiwifruit. We produced a heat map by Cluster analysis to show by how much a particular compound is present in each genotype and we made a graphical representation of Bayesian networks to show the relation among different compounds.

Thesis outline

In Chapter 2, an overview is given on methods of data transformation, cluster analysis and Bayesian networks. Chapter 3 – 5 gives presentations of the approach and how the methods were selected and the analyses carried out. Chapter 3 describes the methodology used in transforming data by simulation. Chapter 4 elaborates on dimension reduction of transformed data via principal component analysis, and this reduced set of data is then used for cluster analysis. The focus of Chapter 5 is on Bayesian networks including the framework for parameter and structure learning and also details about the software BANJO used for its analysis. Chapter 6 presents results of Cluster analysis shown by a ‘heat map’ and a graphical representation of Bayesian networks produced with BANJO. Finally, Chapter 7 outlines the conclusions drawn from this research and offers suggestions for future work.

Chapter 2

Literature Review

2.1 Data transformation

Osborne (2002) discusses the assumption made by many statistical procedures about the variables being normally distributed. A significant violation of the assumption of normality can seriously increase the chances of the researcher committing either a Type I or II error depending on the nature of the analysis and the non-normality. Micceri (1989) points out that one reason, although not the only reason researchers utilize data transformations is improving the normality of variables. Since our data is highly skewed with many small or zero observations, it became imperative to transform the data as described in detail in Chapter 3.

Normality in the distribution of variables is not strictly required when principal component analysis is used descriptively, but it does enhance the analysis (Tabachnick & Fidell, 1996). Since multi-variate normality also implies linear relationships among pairs of variables, we tried to look for linear relationship between compounds via principal component analysis. In case of lack of linearity, data transformation can be done by taking the logarithm, square root, reciprocal, or some other function of the data.

Van den Boogaart, Tolosana-Delgado, and Bren (2006) observed that in compositional data, missing values are quite common. “Below detection limit” is the most-commonly found type of missing values. Since no full quantitative information is available such a censored value is actually treated as zero and data transformation is conducted to take care of non-normality. Rowan et al (2007) made a choice of c based on the minimum non-zero value, under the idea that 47 volatiles available from seedlings of an apple population might be there but just non-detectable. His data was $\log_{10}$ transformed, after adding half the minimum non-zero value (0.0005) for 21 of the 47 volatiles. We will be using the “`log(x+c)`” transformation rather than a sophisticated censored data approach.

Kennedy (2003) suggests that although the Box-Cox transformation is very popular, it has the disadvantage of breaking down when zero values are transformed (because the log of zero values is undefined).To avoid the difficulty with zeros in case of log x, Bartlett (1947) used log (1+x) transformation in place of log x as a logarithmic transformation. Berthouex and Brown (2002) mentioned log transformation being stronger than the square-root transformation. By "stronger", they meant the range of the transformed variables is relatively smaller for a log transformation than that for the square root. They expressed log (x+c) for the sample which contains some zero values and state that c is usually arbitrarily chosen to be 1 with larger values of c making the transformation less severe. We will examine in Chapter 3 how choice of c affects inference for linear relationships.

2.2 Cluster Analysis

According to Hastie, Tibshirani, & Friedman (2001), cluster analysis, also called data segmentation, has a variety of goals. All relate to grouping or segmenting a collection of objects into "clusters" such that those within each cluster are more closely related to one another than objects assigned to different clusters. Central to all of the goals of cluster analysis is the notion of the degree of similarity (or dissimilarity) between the individual objects being clustered. Gentleman, Hanhe & Huber (2006) points out the notion of agglomerative hierarchical clustering that merges clusters iteratively. This algorithm is easy to implement, and somewhat easy to interpret; often the resulting dendrogram will appear to indicate that there are groups in the data.

Yin, Yang, Yao, & Shi (2005) used the method proposed by Eisen et al (1998) to analyze the expression data of mouse sperm genes from the State Key Laboratory of Reproductive Biology of the Chinese Academy of Sciences to find out the hidden pattern in the gene data. The method uses cluster analysis to process the genome-wide expression data from DNA micro array hybridization. The cluster result is displayed in a dendrogram with each node indicating the merging of different sub-clusters. Our data is like gene expression data as it has high dimensional, inter-related measurements, it is potentially influenced by the genotypes of the individuals they have come from, and researchers are interested in identifying clusters of phenotypes and the genotypes associated with them.

Gentleman et al (2005) discussed heat maps as having the notion of rearranging the columns and rows to show structure in the data. The heat map is attached to the dendrogram in which different colours represent the different expression level of the genes. The rows and columns of the matrix are reordered so that similar rows are placed next to each other, and the same for columns. The orderings that are widely used are those derived from a hierarchical clustering. Yin et al (2005) attached heat map to the dendrogram to help biologists in finding out the genes with similar functions in a naturalistic way.

The heatmap function became available with the statistical R programming language release 1.7.0. It requires *Biobase* package from Bioconductor to display the output graphically. This function calculates distances between gene or sample profiles using Euclidean distance (Gentleman, Carey, Bates, Bolstad et al, 2004).

2.3 Bayesian Networks

According to Heckerman (1998), a Bayesian network for a set of variables $X = \{X_1,...., X_n\}$ is defined by,

(1) A network structure (DAG)

(2) Local probability distributions so that

$$P(x) = \prod_{i=1}^{n} P(x_i \mid pa(x_i)) \qquad (2.1)$$

We use X_i to denote both the variable and its corresponding node, and Pa_i to denote the parents of node X_i in a Bayesian network structure as well as the variables corresponding to those parents. A simple example of Bayesian network structure is given in Figure 2.1. To learn a Bayesian network, we have to choose the structure of the model and assess local probability distributions.

An example of simple Bayesian network structure can be shown with five compounds of monoterpene group from our data: cymene (C), menth-4-ol (M), terpinolene (T), piperitone (P) and linalool oxide.cis (L).

Figure 2.1: An example of a simple Bayesian network structure.

This network structure implies several conditional independencies:
$I(C;L), I(M;P \mid C,L), I(T;C,L,P \mid M), I(P;T,M,L \mid C)$, *and* $I(L;C,P)$.
The network structure also implies that the joint distribution has the product form,
$P(C,M,T,P,L) = P(C)P(M \mid C,L)P(T \mid M)P(P \mid C)P(L)$

Heckerman (1998) presented a tutorial on Bayesian networks, which discusses the case where a network structure is known. The physical joint probability distribution for a set of variables X can be encoded in some network structure B which can be written as,

$$p(X \mid \theta_s, B^h) = \prod_{i=1}^{n} p(x_i \mid pa_i, \theta_i, B^h) \qquad (2.2)$$

where θ_i is the vector of parameters for the distribution $p(x_i \mid pa_i, \theta_i, B^h)$ which is in fact a multinomial distribution in our case; θ_s is the vector of parameters $(\theta_1, ..., \theta_n)$, pa_i is set of parents for value x_i and B^h denotes the event that the physical joint probability distribution can be factored according to B.

Hartemink (2001) in his Ph.D. dissertation gave a detailed introduction to discretization techniques. As the amount of data available for reasoning about genetic regulatory networks is comparatively limited, he discussed the need to reduce the dimensionality of the modelling by discretizing variables into a small number of levels. Data discretized in three groups can represent non-linear relationship in case of multinomial distribution. To capture non-linear interactions if they are present and to deal with non-detected observations we transformed our data to discrete values using quantile discretization.

There has been an enormous amount of work done in the area of learning Bayesian-network structures from data and many authors have contributed their ideas in this area of research. In this study, we are using a software package that deal with graphical models called BANJO (Bayesian Network Inference with Java Objects).

Geiger & Heckerman (1995) show that the assumptions of parameter independence and likelihood equivalence imply that the priors for the parameters of any complete network structure must have a Dirichlet distribution. BANJO uses Dirichlet parameter priors as they give closed form solution. The conjugacy of the Dirichlet priors allows us to have the posterior probabilities in the same form as prior probabilities.

Several authors have discussed certain assumptions for deriving priors on network-structure. Buntine (1991) describes a set of assumptions that leads to a richer yet efficient approach for assigning priors. The first assumption is that the variables can be ordered and the second assumption is the presence or absence of possible arcs is mutually independent. An alternative approach, described by Heckerman et al (1995b) uses a prior network. His idea was to penalize the prior probability of any structure according to some measure of deviation between that structure and the prior network. For assigning priors to network-structure we assume a uniform prior over structures.

Madigan & York (1995) depicted in their article how Bayesian graphical models unify and simplify standard discrete data problems such as Bayesian log-linear modelling with either complete or incomplete data. They described two classes of graphical models: undirected decomposable and directed acyclic (DAG). In problems where some variables are obviously determined before others, or cause others, the directed graphs allow a natural representation of them. Undirected models, in contrast, are best suited to problems where the variables are determined simultaneously, or perhaps are both influenced by some variable that is not explicitly modelled. Our data determines the compounds simultaneously; however directed acyclic graphs are used as they define a larger class of models and can be interpreted as undirected graphs in BANJO.

Friedman et al (2000) proposed a new framework for discovering interactions between genes based on multiple expression measurements. This framework builds on the use of Bayesian networks. A method for recovering gene interactions from microarray data

was applied to the *S. cerevisiae* cell cycle measurements of Spellman et al (1998). Bayesian networks represent the dependence structure between expression levels of different genes (Pearl 1988). Our data has lots in common with gene expression data and consequently the fact that Bayesian networks have been used with expression data suggests they would be useful in our context.

The common approach to the problem of unknown network structure is to introduce a statistically motivated scoring function that evaluates each network with respect to the training data D, and to search for the optimal network (B, θ_B) according to this score. In this score, we evaluate the posterior probability of a graph given the data:

$Score(B:D) = P(B \mid D) = \frac{P(D \mid B)P(B)}{P(D)}$ where $P(D)$ is a constant and $P(D \mid B)$ can

be calculated as $P(D \mid B) = \int P(D \mid \theta_B, B)P(\theta_B \mid B)d\theta_B.$ (2.3)

where $P(D \mid \theta_B, B)$ is the likelihood of the data given the network (B, θ_B). Equation 2.3 is the marginal likelihood, which averages the probability of the data over all possible parameter assignments to B. The particular choice of priors $P(B)$ and $P(\theta_B \mid B)$ for each B determines the exact Bayesian score (Friedman et al, 2000). Equation 2.2 gives the Bayesian scoring metric when the structure prior $P(B)$ is uniform.

Hartemink et al (2002) concentrate on search methods that seek to maximise some scoring function that describes the ability of the network to explain the observed data. In a search context, the Bayesian scoring metric (BDe) derived by Heckerman et al (1995) is an especially common choice for the scoring function. Heuristic rather than exhaustive search strategies were considered since the identification of the highest-scoring model under the BDe for a given set of data is known to be NP- complete (Chickering, 1996). Commonly used local heuristic search algorithms include greedy hill-climbing, greedy random, Metropolis and simulated annealing. After implementing these search algorithms it was observed by Hartemink et al (2002) that simulated annealing consistently finds the highest scoring models among these algorithms. Out of many software packages, the one that uses simulated annealing is BANJO. A detailed explanation of this method is given in Chapter 5.

Chapter 3

Transformation methodology

3.1 Motivation

Data transformations are an important tool for the proper statistical analysis of data. There are an infinite number of transformations that can be used, but the common ones are square-root transformation for count data or the log transformation for size data. It is always important to decide which transformation to use prior to analyses.

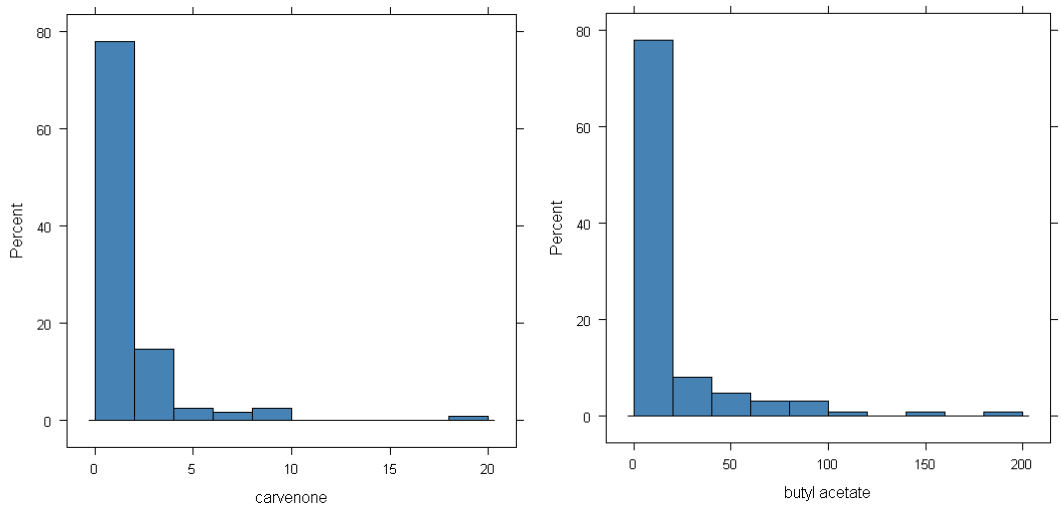


Figure 3.1: Histograms of carvenone and butyl acetate which belong to the group of 'monoterpene' and 'ester' in the compound database.

For example, as shown in Figure 3.1, the compounds carvenone and butyl acetate are non-normally distributed; they are present in large amount in fewer genotypes, in small amount in more number of genotypes and absent in many other genotypes. They also affect inference of relationships. We will discuss below in detail how to transform these compounds along with many others for further study.

Durbin et al (2002) considered log transformation of the form $\ln(\mu + c)$, where μ is the true expression level and c is some positive constant to stabilize the variance of microarray data expressed at high levels. Yamamura (1999) suggested that $c = 0.5$ is preferable to $c = 1$ in $\log_e(x + c)$ because a discrete distribution defined in {0, 1, 2…) is approximately described by a continuous distribution defined in $(0, \infty)$ if we use $c = 0.5$.

Since our data is highly skewed with many small or zero (missing) observations as shown in the example above, it was imperative to transform the data. As we cannot take the log of zero, we should add a constant to each number to make them non-zero. So we used a typical transformation for this scenario which was log(x+c), where x represented the measurements of different compounds for different genotypes. And c was chosen independently for the two groups, monoterpene and acid-alcohol-ester.

To illustrate the importance of the choice of c, we considered three values which were too small (0.0005), too large (10) and intermediate (0.5) for monoterpene in Figure 3.2. We look at the relationship between two compounds, namely carvenone and isopiperitenone, for these different choices of c.

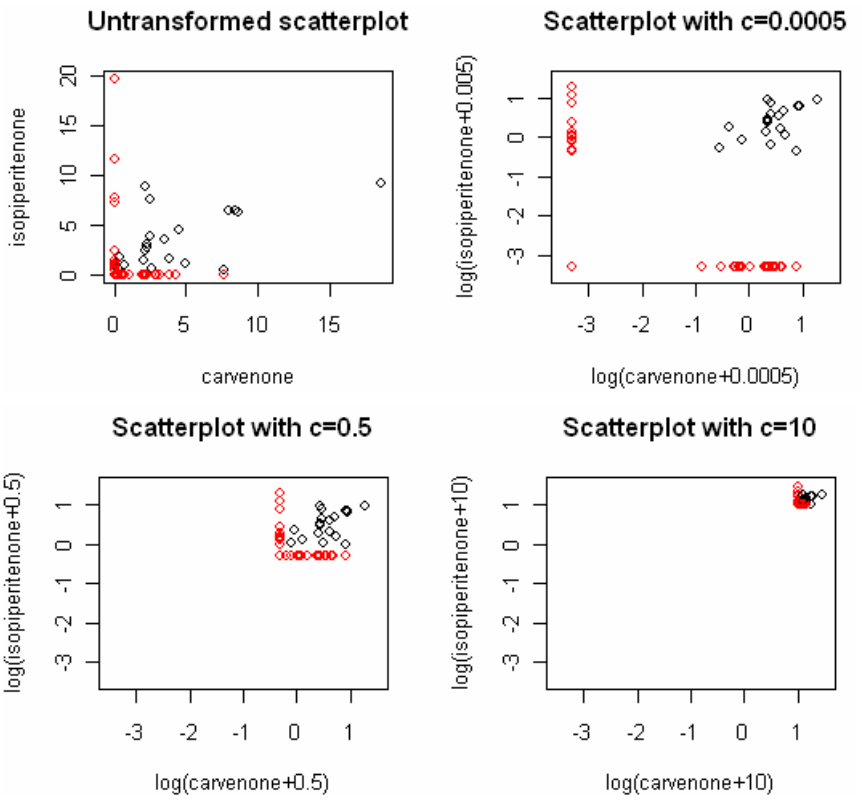


Figure 3.2: Plots of carvenone and isopiperitenone for different c values.

Similarly for acid-alcohol-ester, three values were considered. Too small and too large values were similar from before, but the intermediate value here was 2. Relationship between two compounds was observed, namely butyl acetate and ethyl pentanoate for the choice of c.

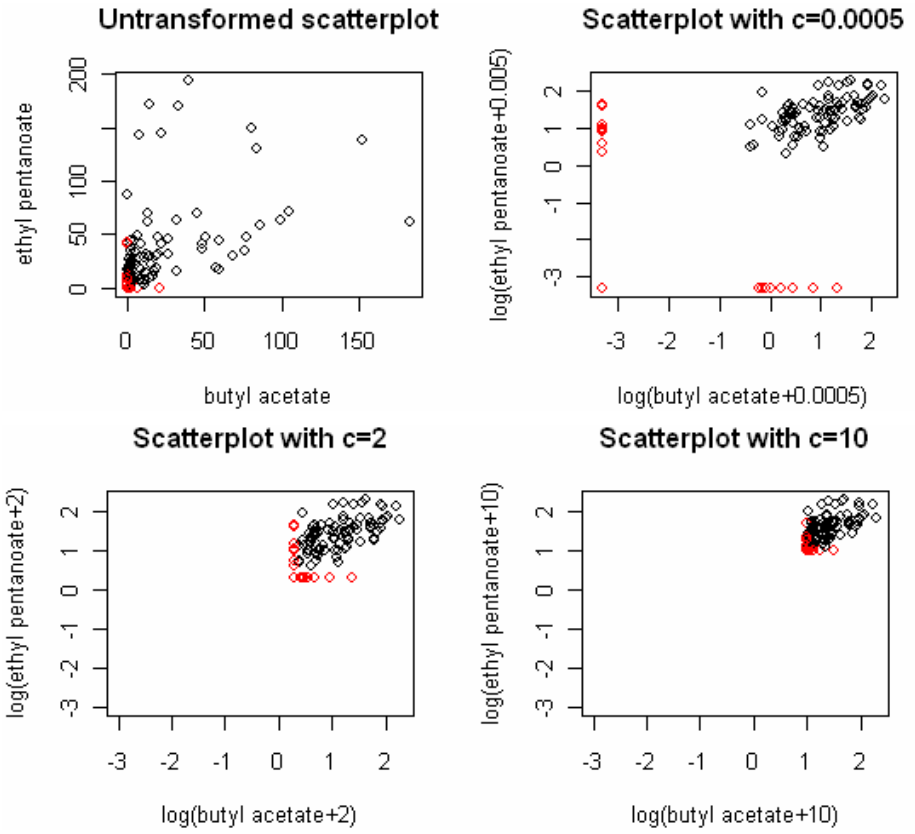


Figure 3.3: Plots of butyl acetate and ethyl pentanoate for different c values.

The untransformed scatterplot of the pair of compounds for both groups obscures the linear relationship as shown in Figure 3.2 and 3.3. Hence three plots were made for each group to observe the performance of c values. It was seen that c = 0.0005 had many outliers, which will act as points of leverage.Using c = 10 showed the differences at the lower end of the scale compressed for both groups. It was observed that c = 0.5 and c = 2 is roughly normally distributed for monoterpene and acid-alcohol-ester respectively. As observed, c might affect the behaviour at the low end of the data, and we will focus on this behaviour while selecting c. The best value of c is one that makes the data look normal.

3.2 Selection of constant by simulation

Since there are many non-detected observations which may be considered as zeros, the data was log transformed after adding a constant to each number to make them non-zero. The constant was selected by simulating 10,000 random normal samples of size 123 (no. of genotypes). Then gap1 was calculated as the difference between 2nd smallest and smallest value and gap2 was the difference between 3rd smallest and 2nd smallest value. There are no exact zeros for simulated data. In Fig 3.2 and 3.3, choice of c was

made to make the data look normal at the lower end of the scale and this was done by examining gap1 and gap2.

Different values for c like 0.1, 0.5, 1, 2, 3, 4 and 10 were selected to standardize the log(x+c) values for each compound and each value of c. Any value of c would be appropriate as we are looking for a reasonable c that will eliminate the problem of compression of scale as shown in Fig 3.2 and 3.3 with c=10. Then gap1 and gap2 values were computed to see which value of c is in accordance with gap1 and gap2 of the random normal samples. Gap1 and gap2 for the compounds were taken as the difference between the smallest unique measurements where the smallest values are all log(c). While all the compounds were transformed, only those compounds were used for selecting c which had at least 75% non-zero measurements which means compound measurements should be present in at least 92 genotypes out of 123.

For monoterpene group only 5 compounds were selected out of 45 to compute gap1 and gap2 for 0.1, 0.5, 1, 2, and 10 c values. And out of 194 compounds from acid-alcohol - ester group, 27 compounds were considered for 2, 3, 4 and 10 values of c. These were compared with gap1 and gap2 for random normal samples. Ideally gap1 and gap2 for the transformed data values will have a similar distribution to gap1 and gap2 for the simulated normal values. But a more detailed work can be carried on by conducting separate simulation for every compound and goal would be to find c which can work for all of them simultaneously. Also, since we are not really looking at smallest gap, our aim would be to find what 'gap i' looks like for normal data, where i is the number of non-detected observations.

The constant value chosen in Figure 3.4 after comparing gap1 and gap2 between the normal simulation and the transformed data for monoterpene group is 0.5. The upper quartile (Q3) of the transformed data and normal data looks similar for gap1, while for gap2 the median (Q2) of transformed data is approximately same as for the random normal samples. For gap1, though the maximum value of constant 0.1 is similar to the maximum value of normal data, the upper quartile (Q3) is not the same. All the values of c look compressed for gap2 except the value 0.5.

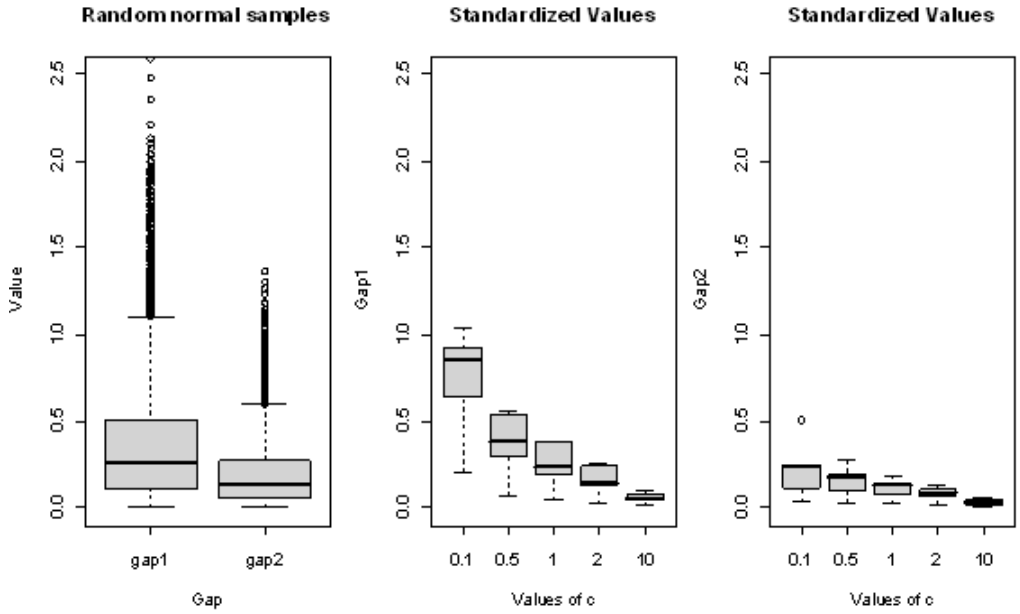


Figure 3.4: Box plot shows 5 data points of gap1 and gap2 for random normal samples and monoterpene group.

For acid, alcohol & ester group in Figure 3.5, the c value chosen is 2 after comparison. The lower quartile (Q1) for gap1 is same for both data values, though the median is closer to Q1 in transformed data than for the normal data. For gap2, the upper and lower quartile of transformed data is nearly equal to that of normal data. The upper quartile (Q3) for all other values of c is quite low as compared to the value 2 for both gap1 and gap2.

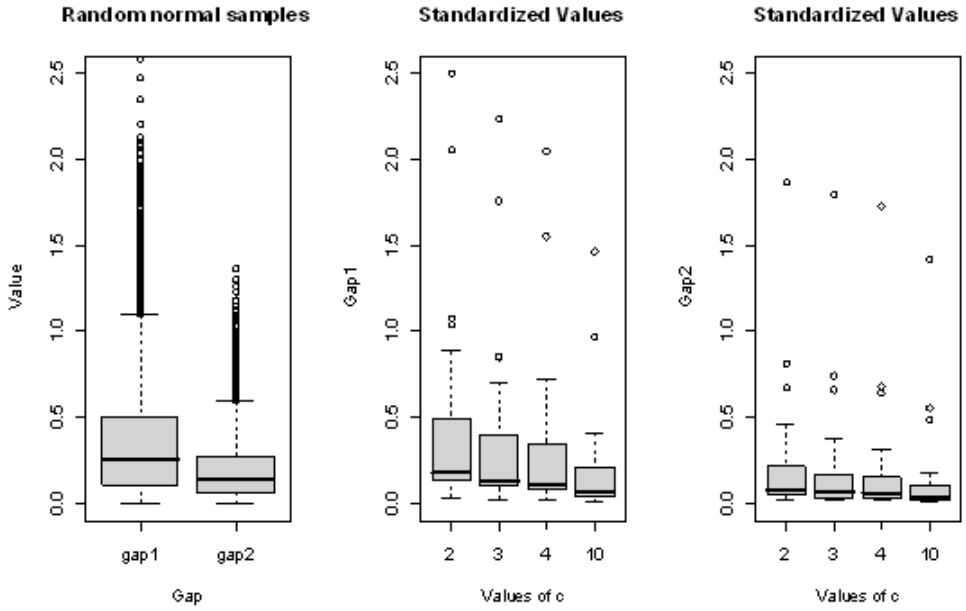


Figure 3.5: Box plot shows 5 data points of gap1 and gap2 for random normal samples and acid, alcohol & ester group.

The two groups have different values of `c` as derived from the simulations above. All compounds in the monoterpene group will be transformed by using `log (x+0.5)`, while for transforming compounds in acid-alcohol -ester group `log(x+2)` will be used. Applying the `log(x+c)` transformation to the compounds in Figure 3.1 makes them look normal with the exception of a bulge in the lower tail as shown in Figure 3.6. Carvenone has 84 observations which are non-detected; therefore we get negative values for those observations. While butyl acetate has 28 non-detected observations, it is present in small amount in most of the genotypes, and present in large amount in few of them. Therefore, its distribution looks skewed to the right to some extent. We cannot eliminate big "lump" of low value measurement.

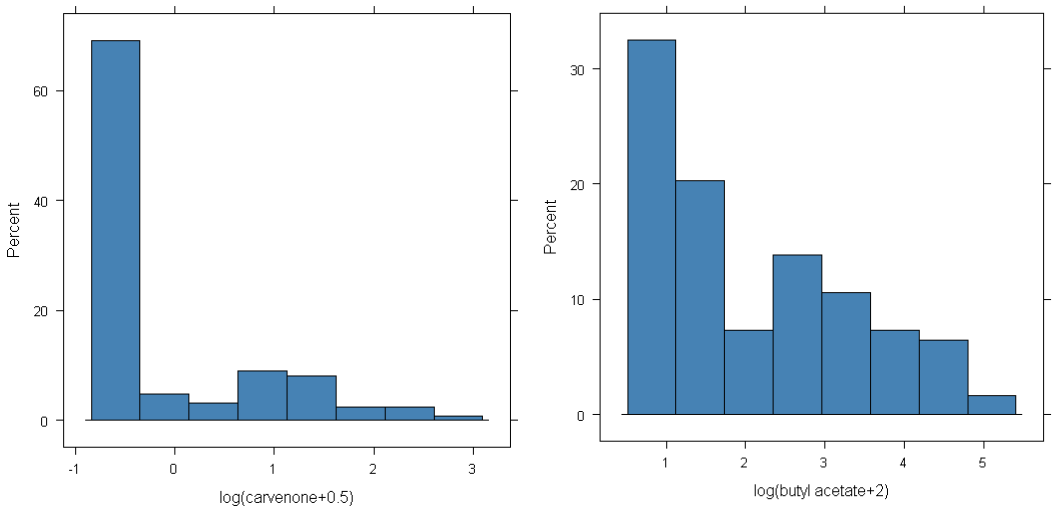


Figure 3.6: Histograms of transformed carvenone and butyl acetate which belong to the group of 'monoterpene' and 'ester' in the compound database.

We can see `log(x+c)` transformation on compound where it is more successful. Hex-E2-enol in Figure 3.7 looks perfectly normal as all observations are detected in that compound.

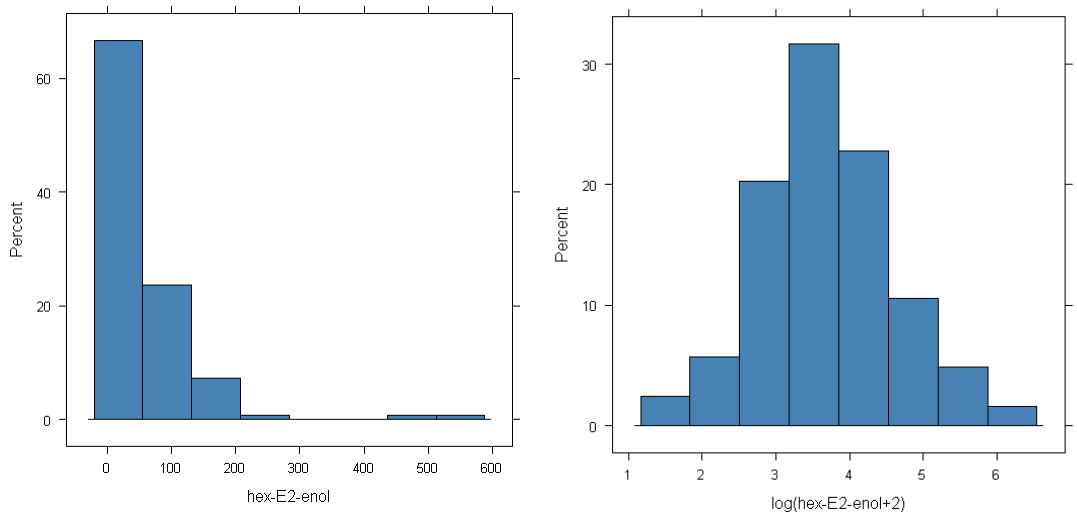


Figure 3.7: Histograms of untransformed and transformed hex-E2-enol which belong to the group of ‘alcohol’ in the compound database.

Chapter 4

Multivariate techniques

4.1 Principal component analysis

We are using Principal component analysis in an atypical way, to select subset of original variables that will retain the overall features, rather than to create new variables. Principal component analysis is a mathematical technique that reduces the dimensions of the data by transforming it to a new set of variables (the principal components) while retaining most of the variation in the data set. We select a subset of the compounds showing large variability across genotypes, or shared variability among compounds by looking at those compounds which load highly on the first several principal components. These compounds are thought to represent the variability in the flavour of kiwifruit, like for instance butanoates and acetates gives a very fruity flavour to the fruit. Also, the important kiwifruit volatiles are believed to be synthesized from a set of precursors like methionine, phenylalaine, and linolyl-CoA, resulting in sets of correlated compounds belonging to the same pathways.

Principal components (PCs) are uncorrelated and ordered such that the k th PC has the k th largest variance among all PCs. The k th PC can be interpreted as the direction that maximizes the variation of the projections of the data points such that it is orthogonal to the first k -1 PCs [Jolliffe, 2002]. The traditional approach is to use the first few PCs in data analysis since they capture most of the variation in the original data set, while the last few PCs are often assumed to capture only the residual 'noise' in the data.

Principal component analysis is performed on the symmetric covariance matrix or on the correlation matrix. These matrices are calculated from the data matrix. If the responses are highly variable and widely different in measurement units then it would be preferred to use the correlation matrix which also means to standardize the data first. However, if the measurement units are commensurable, then statistically it is more desirable to use the covariance matrix. Since all the compounds in our data were recorded in nanograms per gram and they were transformed using `log(x+c)`, principal

components for this were taken out from the covariance matrix. We do principal component analysis within groups, and the within group measurements have similar orders of magnitude.

We followed the procedure as below for selecting compounds for cluster analysis:

- We selected those PCs which accounted for 90% variation in the data (King and Jackson, 1999).
- Then we computed loadings for the selected components which defines the size of the contribution of each original variable to the PCs.
- The loadings in each PC were squared and only those were selected that fall above 75% of the highest loading value. This threshold value was chosen to get manageable number of variables.
- The compounds for which the highest loading appeared twice in different PC were considered only once.

The compounds were selected for monoterpene and acid-alcohol-ester group. For monoterpene group, 14 components accounted for 90% variation in the data. The loadings were squared for the 14 components; and for each component the highest loading value and values that fall above 75% of that loading value were looked for. For example, the highest loading value in component 4 was 0.28356, and we searched in for values greater than 0.21267 (75% of 0.28356). We found 1.3.3.Trimethyl.2.oxabicyclo.2.2.2.octan.5.one was the only compound that had loading value greater than 0.21267, and hence it was selected. Though, exo.2.hydroxycineole had the highest loading value in this component it was not selected because it was already picked up by the second component. This procedure was done for all the components and finally 16 compounds got selected from monoterpene group.

Similarly, for acid-alcohol-ester group 90% variation in the data was shown by 29 components. Again, loadings of those components were squared and 35 compounds were selected which were in relation to the highest loading value and values that fall above 75% of that loading value in each principal component. Hierarchical clustering was then performed on the selected compounds and the results were portrayed by producing a heat map, discussed in next section.

4.2 Cluster analysis

Cluster analysis, also referred to as a unsupervised learning method, is widely used for finding groups in data. We have restricted our attention here towards hierarchical clustering. These hierarchic techniques produce a *dendrogram* that starts with the calculation of the distances of each compound with respect to other compounds. Groups are then formed by agglomeration where one starts with each compound by being alone as an individual cluster, and in successive steps combine the pair of clusters that are closest to each other into one new cluster.

For our hierarchical clustering, the distance measure used between individual observations is the *Euclidean distance*. For example, the data for a cluster analysis consists of the values of p variables $X_1, X_2, ..., X_p$ for n objects. The Euclidean distance function can be written as,

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \tag{4.1}$$

where x_{ik} is the value of variable X_k for individual i and x_{jk} is the value of the same variable for individual j [Manly, 2005]. Also a distance measure between clusters or groups of observations is determined by *complete linkage*, which is also called farthest neighbour. In this method, the distances between clusters are determined by the greatest distance between any two objects in the different clusters. The choice of Euclidean distance and complete linkage is default in R (*hclust*). K-means clustering, which is the most common method of flat-partition-based clustering, is not considered here because a specified number of expected clusters k is difficult to determine.

As our data is like gene expression data as discussed in Section 2.2, for its analysis we used Bioconductor which is Open source development software. Bioconductor is based primarily on the R programming language. *Biobase* package is part of the Bioconductor project. It contains standardized data structures to represent gene expression data. The *ExpressionSet* class from *Biobase* package is designed to combine several different sources of information into a single convenient structure. It consists of several conceptually distinct parts which can be described as,

1. assay data - is a matrix of ‘expression’ values. The matrix has F rows and S columns, where F is the number of features and S is the number of samples.

2. phenotypic data - summarizes information about samples.
3. feature data - contains feature covariates specific to the experiment.

The *ExpressionSet* class was created for the monoterpene and acid-alcohol-ester groups. For the monoterpene group, S was 123 genotypes and F was 16 compounds. While for the acid-alcohol-ester group F was 35 compounds with the same number of genotypes. Phenotypic data had names of the genotypes while feature data had labels of compounds. After creating an *ExpressionSet* for both groups, we presented the output graphically by using *heatmap* function with colours taken from *RColorBrewer* package from CRAN (Comprehensive R Archive Network).

A heat map is a false colour display where the rows and the columns have been permuted to show interesting patterns. Eisen et al. (1998) presented the results of clustering (dendrogram), together with a heat map of gene expression values. Since then they have become a standard visualization method for this type of data. A heat map is a representation of normalized values, where the number of rows in the heat map is equal to the number of features (compounds in our case) and the number of columns is equal to the number of samples (genotypes). One can then colour code each rectangle representing the expression level of one feature in one sample.

Heat maps were produced for 16 compounds of monoterpene group and 35 compounds of acid-alcohol-ester group selected via Principal component analysis with 123 genotypes in both. Heat maps for the two groups are displayed and discussed in Chapter 6.

Chapter 5

Bayesian networks

5.1 Introduction

A Bayesian network is a graphical representation of a joint probability distribution, representing dependence and conditional independence relationships. The important features of Bayesian networks are,

- Bayesian networks are *directed acyclic graphs*, which mean their edges have direction and there are no directed loops within the graph.
- A joint probability distribution is not the collection of individual probabilities for each variable, but allowing the value of one variable to affect the value of another.
- Two variables are dependent if knowledge of one provides predictive value for other variable. On the other hand, independence means when knowledge of one variable provides no predictive value for other. Finally, conditional independence enables us to untangle the relationships amongst the variables within the network, values of which can be correlated in some manner and point out direct influence.

We hope to interpret the parent-child relationships in the graphical model of compounds as compounds that directly affect each other's levels, e.g. the parent compound is a chemical precursor of the child compound.

There are separate priors on the structure and the parameters. In Bayesian networks literature, the most commonly used class of priors are the Dirichlet priors over parameters (Spiegelhalter & Lauritzen, 1990, and Cooper & Herkovits ,1992). For each value combination of the parent variables a multinomial sampling is parameterized via a set of parameters $\theta_1,.....,\theta_k$ such that $\sum_i \theta_i = 1;\ \theta_i$ corresponds to the probability of the ith outcome. A Dirichlet distribution over this set of parameters is defined via a set of hyper parameters $\alpha_1,.....,\alpha_k$. Then, the generalization can be written as,

$$Dir(\theta \mid \alpha_1,.....,\alpha_k) = \frac{\Gamma(\alpha)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1} \tag{5.1}$$

If there is a data set D whose sufficient statistics are $N_1,.....,N_k$, then

$$P(\theta \mid D) = Dir(\theta \mid \alpha_1 + N_1,.....,\alpha_k + N_k) \tag{5.2}$$

The distribution of $\theta's$ for different nodes and different parent values are assumed to be independent. The above prior is used conditional on the network structure. BANJO selects a common value of 1 for the $\alpha's$.

5.2 Learning the Structure of Bayesian networks

The main approach to structure learning in Bayesian networks is to define a score that evaluates how well the dependencies or independencies in a structure match the data, and the task is to search for a structure that maximizes the score. The commonly used scoring function is the Bayesian scoring metric (BDe) which is defined in Equation (2.2). From that equation we see the score used to evaluate the quality of each network is an integral over possible values for the parameters. $P(B)$ is defined as a structure prior which is uniform in our case and plays a relatively minor role. $P(\theta_B \mid B)$ is a parameter prior which is Dirichlet as given in Equation 5.3 above.

The task of structural learning has been reduced to a search problem. The challenging part of this problem is that the size of the space of all structures is super exponential in the number of nodes, so an exhaustive enumeration of all the structures is not possible. Instead, researchers have considered heuristic search strategies that move around in the search space by iteratively performing small changes to the current structure [Jensen, 1996].

Commonly used local heuristic search algorithms include greedy search and simulated annealing. The details of heuristic search algorithms are presented as follows:

Greedy search

Greedy search is a simple heuristic search procedure. It chooses some initial structure which can be an empty structure, a randomly chosen structure, or a prior structure specified by the user and calculates the gain for adding or deleting an edge with the

restriction that resulting graph should be acyclic. It then performs the edge addition or deletion process with highest gain and use the resulting model as the current model. This algorithm chooses what looks locally best, rather than worrying about whether or not it will be best in the long run. In other words, when working with such algorithms we are not guaranteed to find a global optimal structure but only a local optimal structure.

Greedy search is usually performed with multiple restarts to escape local maxima. After a local maximum is found the search is reinitialized with a random structure. This reinitialization is then repeated for a fixed number of iterations, and the best structure found throughout the entire process is selected.

Simulated annealing

Definition given by (Black, 2009) is "Simulated Annealing is a technique to find a good solution to an optimization problem by trying random variations of the current solution. This technique stems from thermal annealing which aims to obtain perfect crystallizations by a slow temperature reduction to give atoms the time to attain the lowest energy state."

Simulated annealing is an improvement on greedy search as it has the potential to escape local minima. Instead of picking the best move, it picks a random move. If the move improves the situation, it is executed. At this point simulated annealing is the same as greedy search. Otherwise, the algorithm makes the move with some probability less than 1. The chance of getting stuck in a local minimum is greatly decreased by using simulated annealing as opposed to greedy search.

The term annealing comes from the field of metallurgy, where the basic procedure is to heat up a piece of metal and then cool it down in a controlled fashion. The atoms in a heated metal become unstuck from their initial positions with some level of internal energy. Then as cooling takes place, the atoms gradually configure themselves in states of lower internal energy. If the cooling is sufficiently slow, the final internal energy is lower than the initial internal energy, thus refining the crystalline structure and reducing defects (Chong and Zak, 2008). The parameter settings of initial value of temperature, cooling factor and computation time were chosen in the hope to find global optimum.

Simulated annealing was originally devised by Kirkpatrick et al (1983).The simulated annealing algorithm starts with model m_i chosen at random, and the error (or energy) $E(m_i)$ is computed. Then a new model m_j is obtained for which the error $E(m_j)$ is computed. If $\Delta E = E(m_j) - E(m_i)$, then m_j is always accepted as $\Delta E < 0$. However if $\Delta E > 0,$ the new model is accepted with probability

$$P_{accept} = \exp\left(-\frac{\Delta E}{T}\right), \tag{5.3}$$

where T is the temperature parameter. The above probability distribution is known as Boltzmann's distribution. The E used in BANJO is $-\log(BDe)$, making P_{accept} equivalent to a Metropolis Hastings algorithm when $T = 1$. The acceptance process is repeated several times at a constant temperature. Then the temperature is lowered following a cooling schedule and the process is repeated. The algorithm is stopped when the error does not change after a sufficient number of trials.

Geometric cooling is the most commonly used cooling schedule and is described as (Kirkpatrick et al., 1983)

$$T_{k+1} = \alpha T_k \tag{5.4}$$

If T_0 is the initial temperature, T_k is the temperature in iteration k, $\alpha < 1$ (typically in the range of 0.9-0.99) that controls the rate of temperature decline. It should be noted that, in the simulated annealing algorithms used in BANJO, reannealing (or tempering) – a sequence of cooling and reheating (increase in temperature) is considered instead of straight forward annealing. To avoid being caught at local minima, the temperature increases periodically than decreases monotonically (Misevičius, 2003). The reheating is applied in an advanced phase of the search, where the search algorithm is nearing to the convergence. Because the early convergence of the local search-based algorithms tends toward local minima, the reheating should allow the search algorithm to escape them with higher probability and increase the chance to reach the global optimum (or at least to reach better local minimum).

5.3 BANJO (Bayesian Network Inference with Java Objects)

The two groups of compounds namely; acid-alcohol-ester and monoterpene were discretized into three values:

1. Non-detected observations
2. Observations <= median value
3. Observations > median value

The median value was determined excluding non-detected observations. The two groups of compounds were analyzed separately, and the discretization was applied separately to each group of compounds. This discretized data was run in BANJO [Hartemink, 2005]. It is a software application and framework for structure learning of static and dynamic Bayesian networks. Since our data came from single point of time, the input data for BANJO was arranged according to the static data file.

The static settings file in BANJO 2 is organized into multiple sections, separated by dividing lines.

- As discretization was done on our data, we did not use the option of discretizing in BANJO.
- We did not specify an initial structure as we used an empty structure. The minimum and maximum Markov lags were at the same time-point so they were set to zero. The strength of prior was equivalent to one observation, that means $\alpha = 1.$
- We used Simulated Annealing as the heuristic search strategy combined with evaluation of a single random local move at each step. For each network structure an overall network's score is computed using the BDe metric and Metropolis-Hastings stochastic decision mechanism which determines whether the proposed network in the current search iteration will be accepted as the new current network for the next iteration, or if it will be rejected, in which case the search proceeds from the current network.
- Pre-compute log gamma and cache setting are the tuning options for the runtime memory requirements of BANJO. Default values were used as memory was not an issue with relatively small data.
- In Simulated Annealing, we experimented with the values for setting up the initial temperature when starting the search, the cooling factor, the reannealing temperature for "restarting" the search, the maximum number of accepted networks before adjusting the cooling factor, the maximum number of search

iterations before adjusting the cooling factor, and the minimum number of search iterations before reannealing (See Tables 5.1 to 5.3 below).

- The search was scheduled to run for a maximum time of an hour and the number of search iterations to be executed without checking the stopping criteria was set to 1000.
- We used raw correlations of the transformed but undiscretized data to show a parent variable's influence on a child. We also used the top scoring graph which gave the highest score and showed the number of networks examined, while other summary measures like influence scores, dot output and consensus graph are given in BANJO to get the post-processing results.

Nothing much was changed in the settings file for both the groups from above discussed sections, except the section on parameters used by specific search methods for simulated annealing. Different values of real number greater than zero were used for initial temperature and cooling factor with different time that a search is scheduled to run. The main reason for choosing different parameter settings was to get a good combination of values for finding the number of networks examined and a high score. A preliminary data of 26 ester compounds was used for this study. The tables below give the combination of values for three different time periods.

Table 5.1, shows that with initial temperature as 1000 and cooling factor as 0.1; we get high number of re-anneals, high score and scores computed. We can also get varying number of networks examined when different initial temperatures are compared with the same cooling factor. The bolding of numbers in Table 5.1- 5.3 indicate the most number of networks examined, high number of re-anneals, high score and most scores computed.

Number of re-anneals gives us confidence of exploring more regions of space and scores computed tells us about the computations performed according to the memory requirements of BANJO. We are interested here in looking at the combination of values which gives a high number of networks examined with high score. The highest cooling factor of 0.7 attained high score and examined the most networks, but there is not much difference among other cooling factors when number of networks examined is

compared with the same initial temperature. Also with high cooling factor less regions of space (number of re-anneals) is investigated.

Table 5.1: Combination of different initial temperature and cooling factor with maximum time of 15minutes

Initial temp	1000	1000	1000	10000	100000
Cooling factor	0.1	0.2	0.7	0.1	0.1
Reannealing temp	800	800	800	800	800
Max. accepted networks before cooling	2500	2500	2500	2500	2500
Max. proposed networks before cooling	10000	10000	10000	10000	10000
Min. accepted networks before re-annealing	500	500	500	500	500
Max. time	15min	15min	15min	15min	15min
Number of networks examined	90743000	92535000	**95757000**	74031000	86947000
Number of re-anneals	**2813**	2569	796	2295	2694
High Score	**-2303.1531**	-2303.1531	-2303.1531	-2303.6690	-2303.2869
(Node) Scores computed	**1985436**	1983745	1819115	1618568	1894824

Table 5.2 tells us the same thing as observed in Table 5.1. When same cooling factor (0.1) is run with different initial temperatures, the one value of temperature which examined the most networks, more number of re-anneals, high score and most scores computed is 1000. Again there is not much difference in the networks examined with high cooling factor of 0.7.

Table 5.2: Combination of different initial temperature and cooling factor with maximum time of 30 minutes

Initial temp	10000	10000	1000	100000
Cooling factor	0.1	0.7	0.1	0.1
Reannealing temp	800	800	800	800
Max. accepted networks before cooling	2500	2500	2500	2500
Max. proposed networks before cooling	10000	10000	10000	10000
Min. accepted networks before re-annealing	500	500	500	500
Max. time	30min	30min	30min	30min
Number of networks examined	161034000	**184901000**	173433000	168813000
Number of re-anneals	4992	1540	**5377**	5233
High Score	-2303.1531	-2303.1531	**-2303.1531**	-2303.2869
(Node) Scores computed	3513215	3496804	**3771262**	3667494

It was seen from Table 5.1 and 5.2 that cooling factor of 0.1 was the most efficient. So in Table 5.3, the same cooling factor with three different values of initial temperature was run for maximum time of an hour. Once again it was proved that initial temperature of 1000 with cooling factor of 0.1 gave the best result. In other words, the slower the cooling, the higher is the probability of finding the optimum solution. These set of values along with others were used in the settings file of BANJO for both groups of compounds.

Table 5.3: Combination of different initial temperature and cooling factor with maximum time of an hour

Initial temp	1000	10000	100000
Cooling factor	0.1	0.1	0.1
Reannealing temp	800	800	800
Max. accepted networks before cooling	2500	2500	2500
Max. proposed networks before cooling	10000	10000	10000
Min. accepted networks before re-annealing	500	500	500
Max. time	1hr	1hr	1hr
Number of networks examined	**322312000**	314443000	291905000
Number of re-anneals	**9994**	9749	9048
High Score	**-2303.1531**	-2303.1531	-2303.1531
(Node) Scores computed	**6987840**	6834576	6346358

The settings file for static Bayesian network with 16 variables in case of monoterpene group and 35 variables in case of acid-alcohol-ester group with 123 observations in both is given in Appendix. This file is then used to run in BANJO. Finally, when the maximum allotted search time is reached, BANJO prints out the search result which is also shown in Appendix.

BANJO supplies the obtained high-scoring Bayesian network for monoterpene group in the following form:

```
Network score: -1531.4331,first found at iteration 64627
16
 0  1  9
 1  1  12
 2  1  3
 3  1  5
 4  1  11
 5  1  14
 6  0
 7  1  1
 8  1  3
 9  0
10  1  9
11  1  10
12  1  14
13  1  11
14  1  10
15  0
```

The first line indicates the score (-1531.4331) and when it was first encountered (iteration 64627).

Line 2 indicates that the number of variables in the network is 16.

Line 3 to 18 (one for each of the 16 variables) first list the id of a variable, then the number of parents, and then a listing of the parents. E.g., “0 1 9” means that variable id = 0 has one parent which is id = 9. Similar explanation can be given for the high-scoring Bayesian network of acid-alcohol-ester group. The graphical representations of the obtained networks generated using the BANJO dot format output is detailed in results chapter.

Chapter 6

Results

6.1 Overview

The two groups of compounds, namely monoterpene and acid-alcohol-ester are analyzed using cluster analysis and Bayesian networks.16 compounds from the monoterpene group and 35 compounds from the acid-alcohol-ester group are chosen via principal component analysis to portray the results.

This chapter is divided into three parts. In the first part (Section 6.2), heat maps for the two groups are displayed and discussed. The focus of the second part (Section 6.3) is on the graphical representation of Bayesian networks produced with BANJO for the two groups of compounds. All 45 compounds from monoterpene group are also explored in this section. Finally, the comments made on the two analyses by Ross Atkinson and Robert Winz are summarized in Section 6.4.

6.2 Heat maps

The data is transformed by using `log(x+0.5)` for the monoterpene group and `log(x+2)` for the acid-alcohol-ester group before producing a heat map. The heat map shows the clustering of compounds by genotypes. Each row represents one compound selected by principal component analysis to share variability with other compounds; each column represents one genotype. The colour code in each rectangle represents the expression level of one compound in one genotype, with light colours for low data values and dark colours for high data values. The lightest colour represents non-detected observation.

Monoterpene group

Figure 6.1, reveals three clusters of compounds from the dendrogram on the right-hand side. The first cluster shows the compounds which are non-detected and present in very small amount, while the second cluster shows those which are present in small to medium amount. The third cluster represents three compounds which are present in large amount in most of the genotypes.

Three clusters of genotypes can be seen from the dendrogram on top of the Figure 6.1. The first and the third cluster look quite similar with the exception of three compounds (cineole, hydroxycineole and trimethyl oxabicyclo octan) which are present in large amounts in the third cluster of genotypes. The second cluster has very few non-detected compounds and more of compounds which are present in small amount.

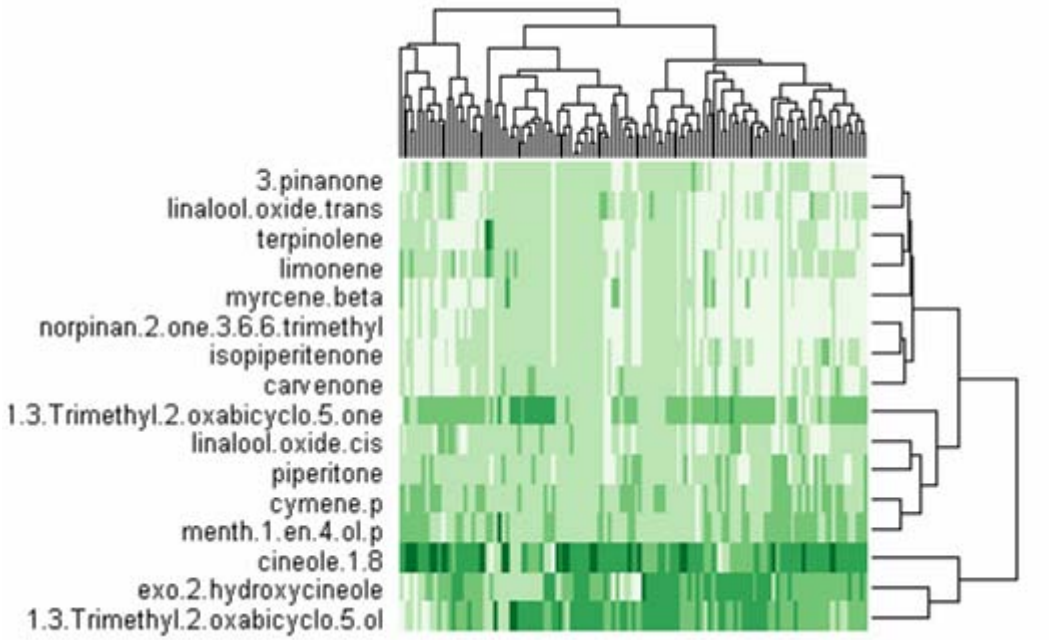


Figure 6.1: Heat map representing 123 genotypes and 16 monoterpene compounds.

Acid-alcohol-ester group

Figure 6.2 also reveals three clusters of compounds with a similar pattern. The first cluster shows the butanoates and benzoates with butyl hexanoate, which are present in most of the genotypes. This cluster has only compounds from the ester group. Ethyl butanoate is present in large amount as seen by the dark green colour, while other compounds are seen in medium to large amount.

Compounds that were present in a small number of genotypes are represented by second cluster. This cluster has many compounds which are non-detected and a few compounds which are present in small amount. In this cluster, octanoic acid is the only compound that belongs to the acid group; there are two compounds from the alcohol group and the rest are from the ester group.

The third cluster comprises compounds which are present in small to medium amount in most of the genotypes. This cluster has compounds mainly from alcohol group and a few from the ester group along with just one from acid group.

Three clusters of genotypes, not very obvious can be seen from the dendrogram on top of Figure 6.2. Two lines are added to the plot to indicate which individuals are grouped together. No distinctive feature can be pointed out from the first cluster. The second cluster of genotypes is the smallest of all the clusters. It shows one compound from ester group being methyl butanoate and two compounds from alcohol group being hex.E2.enol & hexanol are present in large amount, while methyl benzoate is non-detected in two genotypes of this cluster. The presence of highly correlated compounds like butyl hexanoate and hexyl butanoate in relatively small amount is shown by the third cluster. Benzyl alcohol is present in large amount in most genotypes of this cluster.

After cutting the tree, we looked for visual evidence in heat map for the number of clusters in genotypes. By simple Mendelian control, a major gene with 2 alleles heterozygous in both parents results in 1:2:1 ratio which gives three groups; so if a single gene was controlling the entire system we might expect to see one large cluster and two smaller ones of equal size, but it is not true in our case. We could have got more than three clusters of genotypes. However, the scientists pointed out in our discussion that this pattern may hold for a subset of the compounds, and the heat map would allow them to visualize this.

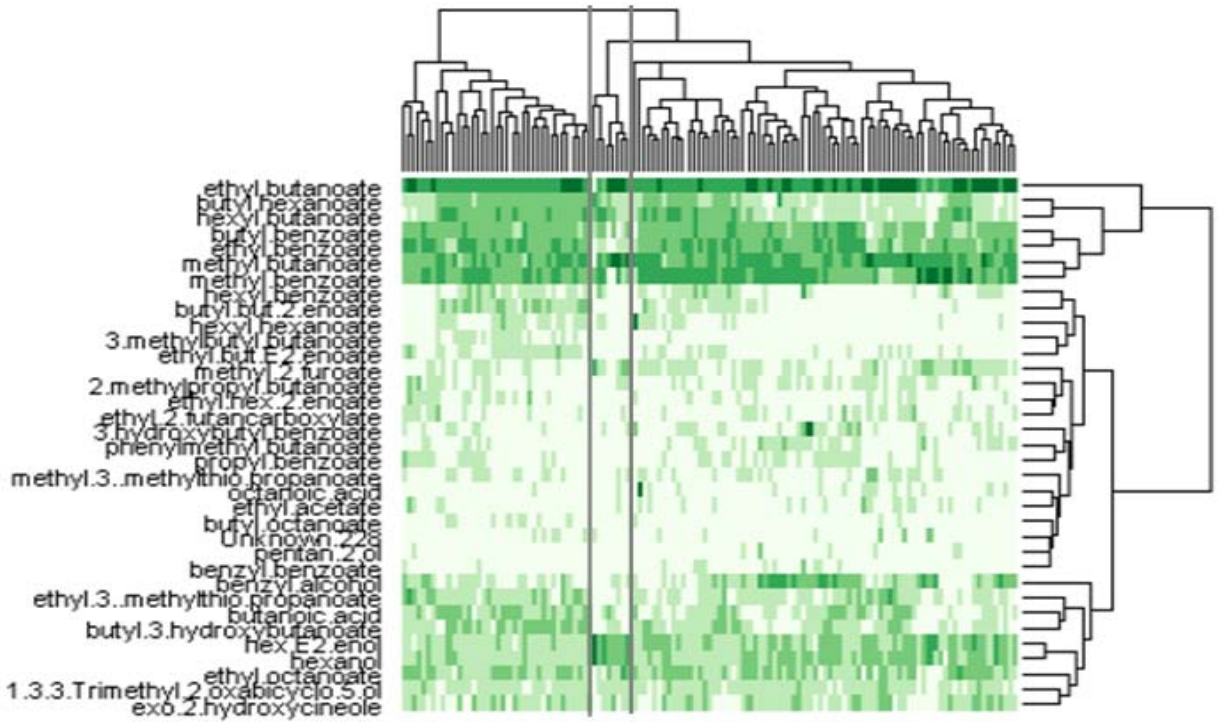


Figure 6.2: Heat map representing 123 genotypes and 35 compounds of acid-alcohol-ester.

6.3 Graphical representation of Bayesian networks

The two groups of compounds are discretized into three values. The non-detected observations are treated as not available/missing values defined as 1. A median value is calculated excluding non-detected observations. Then observations less than or equal to median value is defined as 2 and observations greater than median value is defined as 3. Out of this discretized data, compounds chosen by principal component analysis for the two groups along with full set of 45 compounds of monoterpene is then used to run in BANJO.

Monoterpene group

Three different graphs are presented below for this group,

- Figure 6.3: Biochemical pathway graph was provided by scientists from Plant and Food Research. It was constructed with current knowledge of biological pathways based on chemistry.
- Figure 6.4: Bayesian networks for all 45 compounds of monoterpene group.
- Figure 6.5: Bayesian networks for 16 compounds chosen by Principal component analysis.

Figure 6.3: Biochemical pathway known for monoterpene group.

Unknown.240 compound is not present in Fig 6.3, but its measurement is included in the analysis of monoterpene group and hence represented in Fig 6.4 and 6.5. Compounds like geranyl diphosphate and linoyl diphosphate that are not measured are depicted in Fig 6.3 as names without boxes.

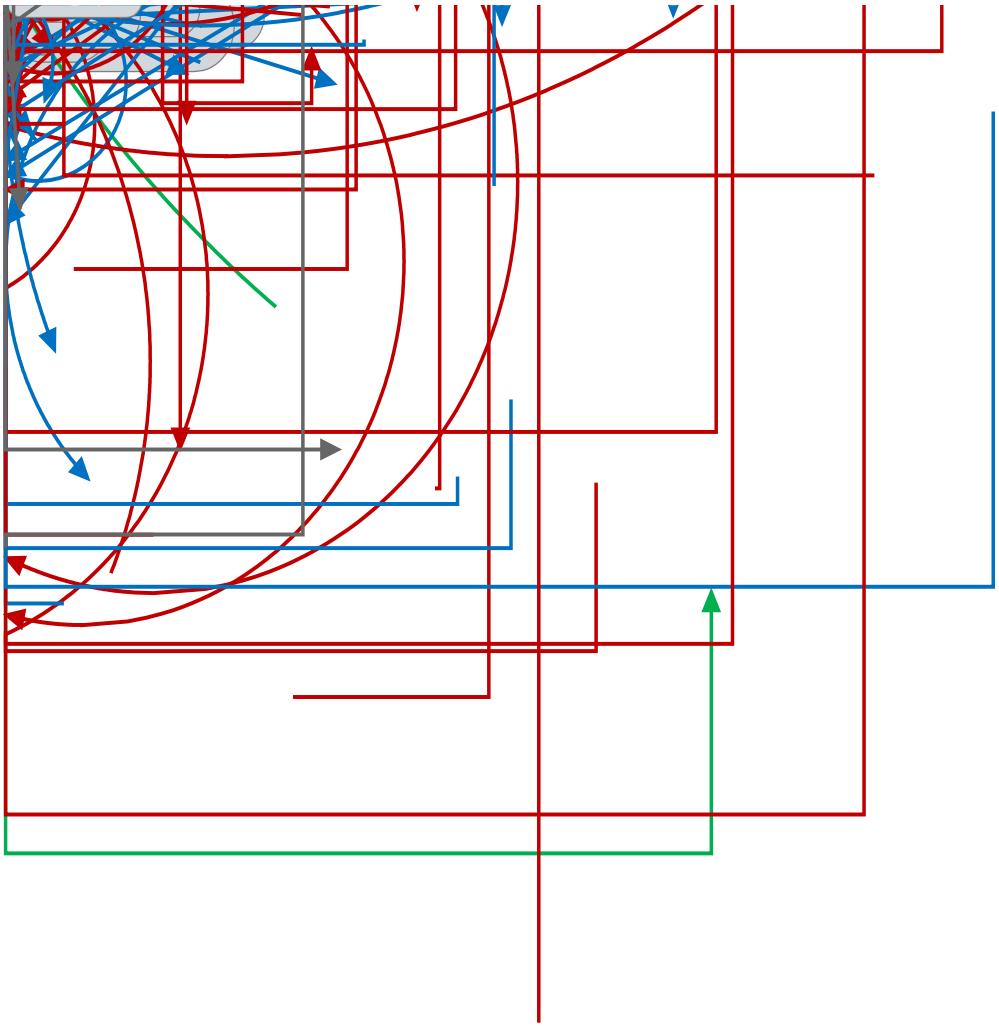


Figure 6.4: Graphical representation of Bayesian networks inference for monoterpene group with 45 compounds.

Table 6.1: Division of Bayesian networks graph for monoterpene group with 45 compounds into four levels and compounds present in that level.

Levels	**Compounds present in that level**
Start	geraniol cis and geranyl acetate
Middle	Myrcene-β sabinene hydrate, sabinene,3 thujen 2 one myrtenol, pinocarveol, α-pinene, β-pinene, pinocarvone, verbenol, borneol, 3-pinanone.
End	Linalool, epoxylinalool, linalool oxide.cis, linalool oxide.trans. Terpin hydrate, α-terpineol; 1, 3, 3-Trimethyl-2-oxabicyclo [2.2.2] octan-5-ol; 1, 3, 3-Trimethyl-2-oxabicyclo [2.2.2] octan-5-one; cineole 1.8-; exo-hydroxycineole. Limonene, carvenone, carveol, 1.8.menthadien-4-ol, isopiperitenone, carvone; 2,3-dihydrocarvone; piperitone. α-terpinene, β-terpinene, γ-terpinene, terpinolene, menth-1-en-4ol.p-, cymene-p; 1,3,8-p-methatriene; p-cymene-8-ol. Norpian-2-one3.6.6-trimethyl, berbenone, β-damascenone, 4oxo α damascene.
Unknown	Unknown.240

This Bayesian network was not readily interpretable. So in an attempt to see which compounds are present at various stages of ripeness we divided the compounds into four levels (Table 6.1). Compounds like geraniol cis and geranyl acetate are at the starting level. Middle level comprises of myrcene, sabinene and initial part of pinene synthases. All terminal nodes appear to be at the end level which consists of linalool, cineole, limonene, terpinene, and later part of pinene synthases.

Figure 6.4 shows the edges between same levels of compounds in red colour, the edges between compounds in the immediate next level in blue colour and the edges between compounds of starting level and end level in green colour. More red edges (42) than blue (28) means there is possibly greater correlation among end level compounds which could be due to heterogeneity in fruit ripeness. The least number of edges (4) is seen in green colour. In an additional effort to understand this group of compounds, we selected new set of compounds for this group via Principal component analysis as discussed earlier. This produces the simpler graph as seen in Figure 6.5.

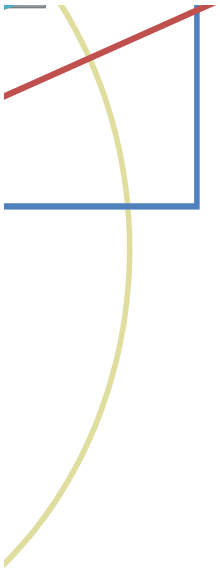


Figure 6.5: Graphical representation of Bayesian networks for monoterpene group with 16 compounds chosen by principal component analysis.

As referred in section 5.1, a Bayesian network is a graphical representation of a joint probability distribution, representing dependence and conditional independence relationships. For example, we see that cymene is the parent of piperitone and menth-1-en -4ol, indicating that all three are correlated; the fact that menth-1-en-4ol and piperitone don't have an edge directly between them indicates their level of correlation is fully explained by the fact that both are related to cymene. Here, we have created Bayesian networks for the compounds shown in the heat map. Where edges are present, we also give the correlation of the continuous transformed measurements. For monoterpenes, we show a graph representing the chemical pathways (Fig.6.5) in which each compound is involved as a point of comparison.

For the monoterpenes, we see that there are a few edges in common with the pathway graph. They are terpinolene & menth –1- en-4 ol and isopiperitenone & carvenone. However, in most of the graph high correlation is seen among compounds present when the fruit is fully ripened, rather than compounds in the same mechanistic pathway. The highest correlation is seen among terminal or near terminal compounds e.g. cymene.p and piperitone or terpinolene and menth1en4ol.

Acid-alcohol-ester group

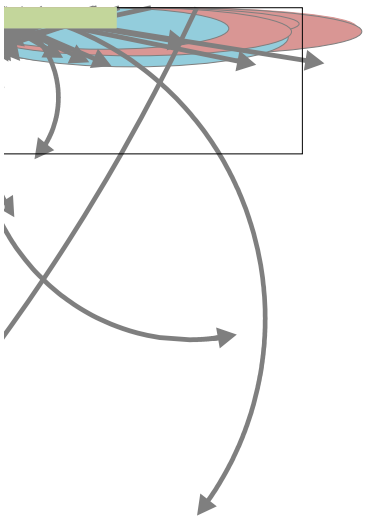


Figure 6.6: Graphical representation of Bayesian network for acid, alcohol & ester group with 35 compounds chosen by principal component analysis.

Compounds linked with different butanoates are highly correlated. Butyl-3hydroxybutanoate which belong to both alcohol and ester group is highly correlated

with butanoic acid, while butyl but2enoate which comes from the ester group is negatively correlated with octanoic acid. Strong correlation can be seen between hexyl butanoate and butyl hexanoate. Benzyl alcohol and phenyl methyl butanoate look quite closely related.

6.4 Comments made on the two analyses

Heat maps and graphical representation of Bayesian networks for the two groups of compounds were discussed with Ross Atkinson and Robert Winz from the Flavour Biotech team at Plant and Food Research. The simpler monoterpene graph with 16 compounds was shown to them. They suggested the biological pathway graph (Figure 6.3) based on hypothesized chemistry is not necessarily the truth. According to them, it was interesting to see from the Bayesian networks for the same group (Figure 6.5) that even though compound 1, 3, 3-Trimethyl-2-oxabicyclo [2.2.2] octan-5-ol was present in high number of genotypes, due to low variability it was not correlated with other compounds.

In the acid-alcohol-ester group, from Figure 6.6, it was seen that methyl butanoate and ethyl butanoate were the important compounds that gave characteristic aroma in gold and green kiwifruit. They thought the compounds were correlated due to ripeness heterogeneity rather than a chemical relationship.

The scientists noticed genotypic mapping can be done with the information provided by heat maps (Figure 6.1& 6.2). Genotypes resulting in similar compound levels can be chosen, as the dendrogram gives appropriate clusters of genotypes for both groups of compounds. They also recommended that compounds which are highlighted as either present or absent in a particular genotype cluster are likely to be under simple genetic control can be shown by the heat maps. This is an interesting conclusion; however trait mapping methods are beyond the scope of this thesis.

Feedback was taken from the two scientists on usefulness of the two graphs. According to them, both graphs were useful in their own sense. They were pleased to see direct interactions among different compounds from the Bayesian networks. At first, they thought the graph shows correlations but after a detailed explanation they were able to understand the concept of conditional independence.

It was advised that 1, 3, 3-Trimethyl-2-oxabicyclo [2.2.2] octan-5-ol and exo.2hydoxy cineole were technically alcohols; but they synthesized differently from the simpler alcohols. It was observed from the heat maps of acid-alcohol-ester and monoterpene group that the intensity of colour representing the levels of the two compounds is different in the two graphs. The presence of these compounds in genotypes does not actually change from one group to another; it is just that their levels relative to the other compounds in the group (monoterpenes or alcohols) is changing.

Adding further to the comments made on Figure 6.6, they believed ratio between methyl butanoate and ethyl butanoate was a good predictor of optimum maturity. The correlation of these two compounds with others gave them an idea to look in for the development of other compounds that could control flavour maturity. The compounds from acid group and compounds found in both ester and alcohol group together were pre-cursors to esters. Strong predictive ability could be seen from the Bayesian networks as one enzyme could be responsible for hexyl butanoate, hexyl benzoate and hexyl hexanoate.

Chapter 7

Conclusion and Discussion

7.1 Conclusion

This chapter focuses on concluding the results of this study. An analysis was done on two groups of volatile compounds, viz., acid-alcohol-ester and monoterpene. The problem of non-detected observations was tackled by transforming the data with different constant values. The goal was to get an indication of the presence of volatile compounds in each genotype, and to see a relationship between them. To reach this goal, Clustering and Bayesian networks were implemented on compounds chosen by Principal component analysis.

The heat map produced with cluster analysis for both groups of compounds revealed three clusters of compounds: compounds which are non-detected and present in very small amount, compounds which are present in small to medium amount and compounds which are present in large amount in most of the genotypes. Three clusters of genotypes from the monoterpene group and three clusters of genotypes from the acid-alcohol-ester group were also seen from the dendrogram of the heat map.

A graphical representation of Bayesian networks is produced with the software BANJO. For monoterpene group, in most of the graph high correlation is seen among compounds present when the fruit is fully ripened and highest correlation is seen among terminal or near terminal compounds. While for acid-alcohol-ester group, the graph shows compounds linked with different butanoates are highly correlated.

7.2 Discussion

The compound database for this study is similar to microarray data if we consider "large *p*, small *n*" paradigm, where p= number of compounds and n= number of genotypes in our case. Acid-alcohol-ester group has 194 compounds and 123 genotypes which means ($p > n$) and monoterpene has 45 compounds which means $n = 3p$ (still moderate relative to p). Also like microarray data, we identify similar individuals and pathways with correlated genes. We can cite an example of a paper written by Mishel et al (2003)

which is similar to our work. Their individual tumor samples are like our genotypes, and their molecular subtypes are like our correlated compounds.

The optimality criteria used in BANJO (BDe) tends to prevent overfitting based on the amount of data. Since we have a small amount of data, our network tends to be quite simple. For example, we discretized our observations into three values; if a node has two parent nodes, there are 9 possible combinations of the parent levels, resulting in 18 parameters governing the distribution of data at the child node. Thus if we increase the number of parents then the parameters cannot be precisely estimated. If one wishes to reduce the complexity of the network even more than the data suggest, BANJO allows one to put a cap on the number of parents that a node is allowed to have. Since we have a small amount of data, our network tends to be quite simple, and where we place the cap on the number of parents it has little effect. If the true relationships between the compounds are complex, we would need substantially more data to recover them. As our data came from one time point we conducted static Bayesian networks and while doing that we lost substantial information about edge directions and thus about possible causal interactions between the genes. Husmier (2003) suggested interactions between genes are not instantaneous, but its effect happens with a time delay after its cause.

Clustering provides a computationally cheap way to extract useful information out of large expression data sets. It only groups interacting genes together in a block, where the detailed form of the interaction patterns is lost. Therefore, probabilistic relationships between multiple interacting genes are represented by Bayesian networks. The structure of a Bayesian network describes the relationships between these genes in the form of conditional independence relations. Presence of many non-detected observations in our study does not violate assumptions of Bayesian networks and hence they are in the flow of the biochemical pathways as seen from the graphical representation. In hierarchical clustering method, the presence of many non-detected observations is quite obvious as they form a cluster with observations which are present in small amount as seen from the heat maps.

7.3 Future study

Two different correlation values can be worked out in future. The first set will give the correlation of a present/ absent indicator in genotypes; they will be shown under the

same pathway or under simple genetic control. The second set will give correlation values for those compounds which are present in both; correlated compounds could be related by some genetic factor, stage of ripeness, or consequence of chemical pathway.

It was pointed out from the Bayesian networks of acid-alcohol-ester group that chemically butanoic acid should be the parent of butyl 3hydroxybuatnoate. Therefore it may be less confusing to represent the edges as undirected in case of butanoic acid, as in many cases there are equivalent graphs with edges reversed. Our use of directed graphs is for computational convenience.

Appendix

BANJO settings file and search results for the two groups of compounds

Settings file for monoterpene group

```
-------------------------------------------------------------------------------
- Banjo                      Bayesian Network Inference with Java Objects -
- Release 2.0                                                 1 Apr 2007 -
- Licensed from Duke University                                          -
- Copyright (c) 2005-2007 by Alexander J. Hartemink                      -
- All rights reserved                                                    -
-------------------------------------------------------------------------------
- Project:                                                  mono16 example
- User:                                                             demo
- Dataset:                                        16-vars-123-observations
- Notes:                                    static bayesian network inference
-------------------------------------------------------------------------------
- Settings file:                             data/static/static.settings.txt
-------------------------------------------------------------------------------
- Input directory:                                        data/static/input
- Observations file:                                        static.data.txt
- Number of observations:                                               123
- Number of variables:                                                   16
- Discretization policy:                                               none
- Exceptions to the discretization policy:                             none
-------------------------------------------------------------------------------
- Initial structure file: (optional)
- 'Must be present' edges file:                          static.mandatory.str
- 'Must not be present' edges file:
-------------------------------------------------------------------------------
- Min. Markov lag:                                                       0
- Max. Markov lag:                                                       0
- Max. parent count:                                                     5
- Equivalent sample size for Dirichlet parameter prior:                1.0
-------------------------------------------------------------------------------
- Searcher:                                                       SimAnneal
- Proposer:                                                  RandomLocalMove
- Evaluator:                                    defaulted to EvaluatorBDe
- Cycle checker:                                          CycleCheckerDFS
- Decider:                                  defaulted to DeciderMetropolis
-------------------------------------------------------------------------------
- Pre-compute logGamma:                                                 no
- Cache:                                                         fastLevel2
-------------------------------------------------------------------------------
- Initial temperature:                                                1000
- Cooling factor:                                                      0.1
- Reannealing temperature:                                             800
- Max. accepted networks before cooling:                              2500
- Max. proposed networks before cooling:                             10000
- Min. accepted networks before reannealing:                           500
-------------------------------------------------------------------------------
- Output directory:                                       data/static/output
- Report file:                                             static.report.txt
- Number of best networks tracked:                                      1
- Max. time:                                                       1.0 h
- Min. networks before checking:                                    1000
- Screen reporting interval:                                       20.0 s
- File reporting interval:                                         10.0 m
-------------------------------------------------------------------------------
- Compute influence scores:                                            yes
- Compute consensus graph:                                             yes
- Create consensus graph as HTML:                                      yes
- Create 'dot' output:                                                 yes
- Location of 'dot':                C:/Program Files/ATT/Graphviz/bin/dot.exe
-------------------------------------------------------------------------------
```

Search results for monoterpene group

```
------------------------------------------------------------------------------
- Final report                                              Best network overall
------------------------------------------------------------------------------


Network score: -1531.4331, first found at iteration 64627
16
 0 1 9
 1 1 12
 2 1 3
 3 1 5
 4 1 11
 5 1 14
 6 0
 7 1 1
 8 1 3
 9 0
10 1 9
11 1 10
12 1 14
13 1 11
14 1 10
15 0
```

```
------------------------------------------------------------------------------
- Search Statistics
------------------------------------------------------------------------------

Statistics collected in searcher 'SearcherSimAnneal':
  Number of networks examined: 511855000
  Total time used: 1.0 h
  High score: -1531.4331, first found at iteration 64627
  Number of re-anneals: 6196

Statistics collected in proposer 'ProposerRandomLocalMove':
  Additions -- proposed:    172038478
  Deletions -- proposed:    169899528
  Reversals -- proposed:    169916993

Statistics collected in cycle checker 'CycleCheckerCheckThenApply':
  Additions -- considered:  172038478,  acyclic:  143316917
  Deletions -- no cyclicity test necessary
  Reversals -- considered:  169916993,  acyclic:  169559115

Statistics collected in evaluator 'EvaluatorBDe':
  Scores computed:          1801532
  Scores (cache)            placed         fetched
      with 0 parents:       16             331548499
      with 1 parents:       240            88585465
      with 2 parents:       1680           230176239
      with 3 parents:       1513782        202133
      with 4 parents:       251009         21040
      with 5 parents:       34805          1719

Statistics collected in decider 'DeciderMetropolis':
  Additions -- considered:  143316917,  better score:  4324793,  other accepted
  13461123
  Deletions -- considered:  169899528,  better score:  13494106,  other accepte
:  4291797
  Reversals -- considered:  169559115,  better score:  34513465,  other accepte
:  8140074
  Average permissivity:     0.06

Memory info after completing the search: Banjo is using 1 mb of memory
```

```
- Post-processing                                                Influence scores
```

```
Influence score for   (9,0) ->   (0,0)   .5534
Influence score for  (12,0) ->   (1,0)   .0000
Influence score for   (3,0) ->   (2,0)   .4206
Influence score for   (5,0) ->   (3,0)   .0000
Influence score for  (11,0) ->   (4,0)   .0000
Influence score for  (14,0) ->   (5,0)   .0000
Influence score for   (1,0) ->   (7,0)   .0000
Influence score for   (3,0) ->   (8,0)   .3115
Influence score for   (9,0) ->  (10,0)   .7188
Influence score for  (10,0) ->  (11,0)   .0000
Influence score for  (14,0) ->  (12,0)   .7126
Influence score for  (11,0) ->  (13,0)   .0000
Influence score for  (10,0) ->  (14,0)   .3135
```

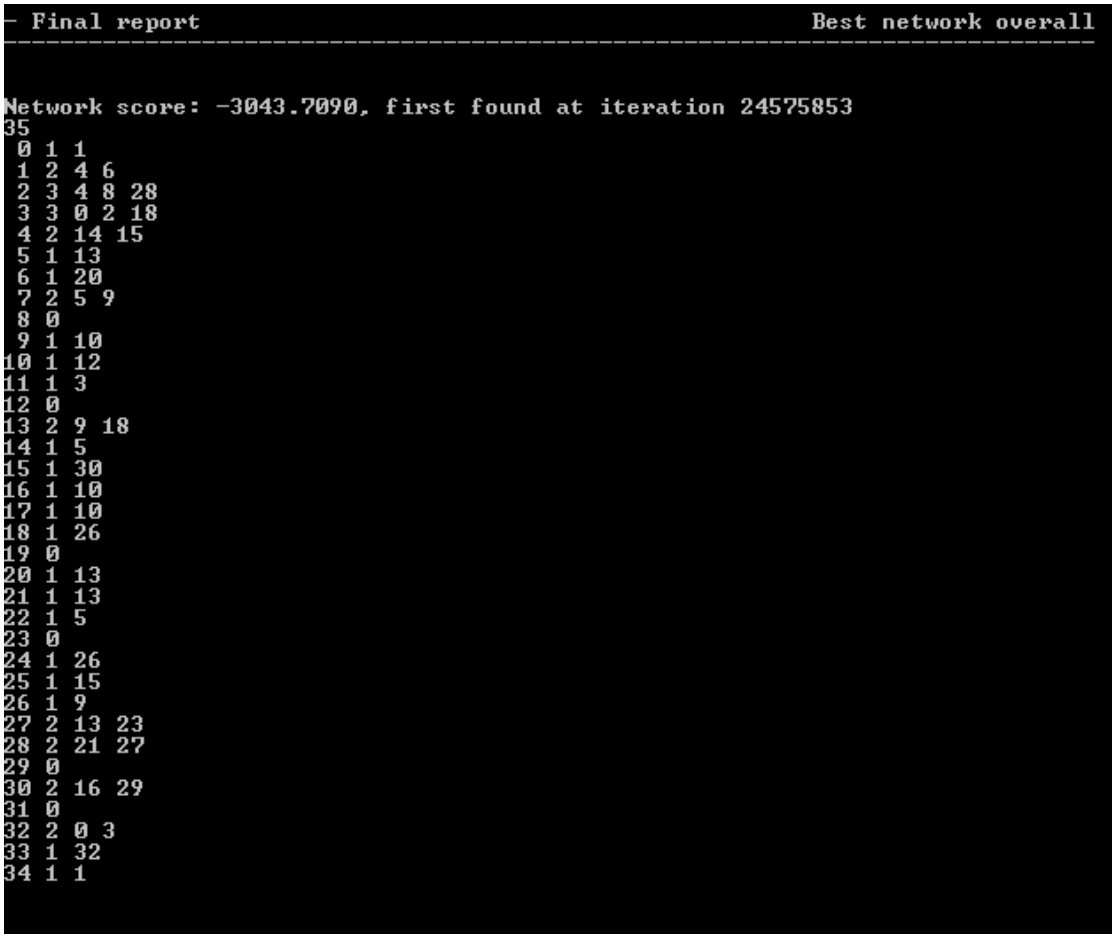
```
- Post-processing                                    DOT graphics format output
```

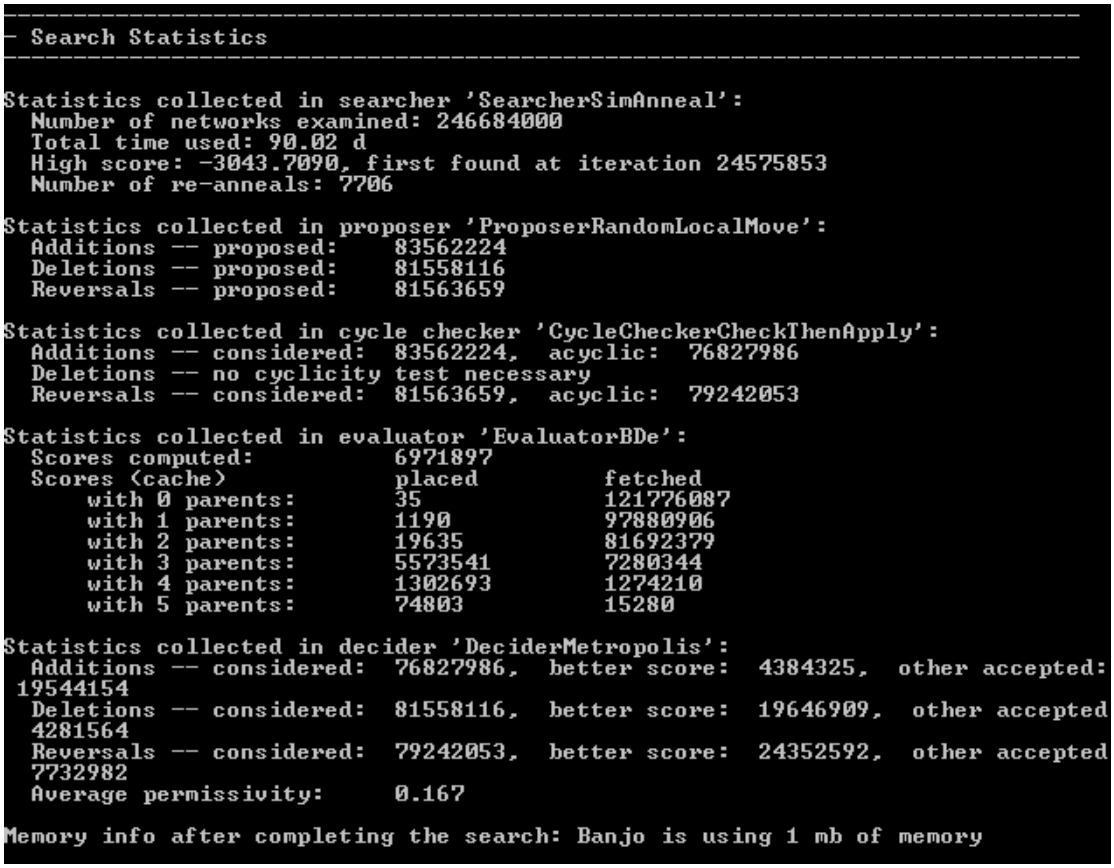
```
digraph abstract {

label = "Banjo Version 2.0.1\nHigh scoring network, score: -1531.4331\nProject:
mono16 example\nUser: demo\nDataset: 16-vars-123-observations\nNetworks searched
: 511855000";
labeljust="l";

    0 [label="piperitone"];
    1 [label="3pinanone"];
    2 [label="carvenone"];
    3 [label="isopiperitenone"];
    4 [label="1.3Trimethyl2oxabicyclo2octan5one"];
    5 [label="norpinan2one3.6trimethyl"];
    6 [label="1.3Trimethyl2oxabicyclo2octan5ol"];
    7 [label="cineole1.8"];
    8 [label="exo.2hydroxycineole"];
    9 [label="cymene.p"];
    10 [label="menth1en4ol.p"];
    11 [label="linalooloxide.cis"];
    12 [label="limonene"];
    13 [label="linalooloxide.trans"];
    14 [label="terpinolene"];
    15 [label="myrcenebeta"];

    9->0;
    12->1;
    3->2;
    5->3;
    11->4;
    14->5;
    1->7;
    3->8;
    9->10;
    10->11;
    14->12;
    11->13;
    10->14;
}
```

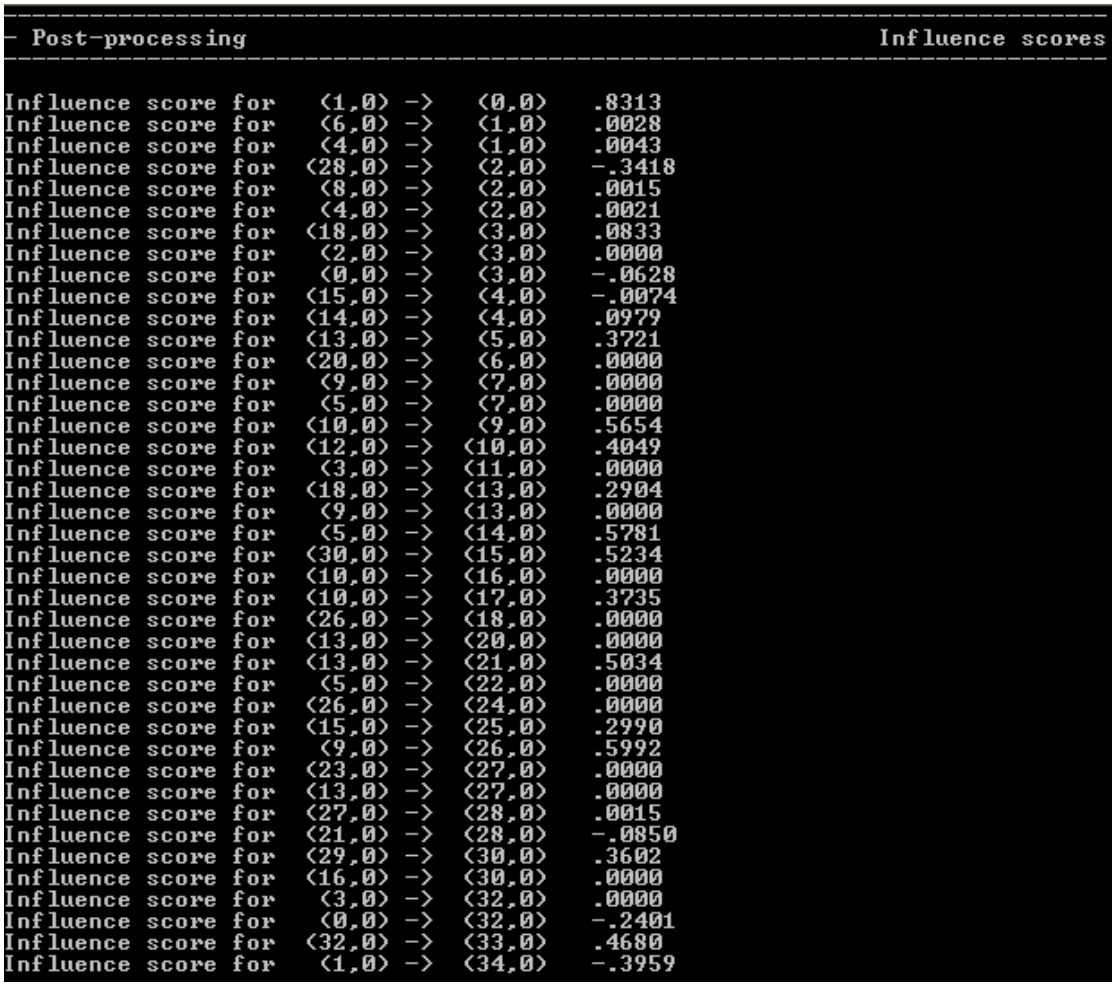
Settings file for acid-alcohol-ester group

```
-------------------------------------------------------------------------
- Banjo                     Bayesian Network Inference with Java Objects -
- Release 2.0                                                1 Apr 2007 -
- Licensed from Duke University                                         -
- Copyright (c) 2005-2007 by Alexander J. Hartemink                     -
- All rights reserved                                                   -
-------------------------------------------------------------------------
- Project:                                                  aae35 example
- User:                                                             demo
- Dataset:                                      35-vars-123-observations
- Notes:                                 static bayesian network inference
-------------------------------------------------------------------------
- Settings file:                           data/static/static.settings.txt
-------------------------------------------------------------------------
- Input directory:                                       data/static/input
- Observations file:                                      static.data.txt
- Number of observations:                                            123
- Number of variables:                                               35
- Discretization policy:                                           none
- Exceptions to the discretization policy:                         none
-------------------------------------------------------------------------
- Initial structure file: (optional)
- 'Must be present' edges file:                        static.mandatory.str
- 'Must not be present' edges file:
-------------------------------------------------------------------------
- Min. Markov lag:                                                    0
- Max. Markov lag:                                                    0
- Max. parent count:                                                  5
- Equivalent sample size for Dirichlet parameter prior:             1.0
-------------------------------------------------------------------------
- Searcher:                                                    SimAnneal
- Proposer:                                                RandomLocalMove
- Evaluator:                                  defaulted to EvaluatorBDe
- Cycle checker:                                           CycleCheckerDFS
- Decider:                                defaulted to DeciderMetropolis
-------------------------------------------------------------------------
- Pre-compute logGamma:                                              no
- Cache:                                                      fastLevel2
-------------------------------------------------------------------------
- Initial temperature:                                             1000
- Cooling factor:                                                   0.1
- Reannealing temperature:                                          800
- Max. accepted networks before cooling:                           2500
- Max. proposed networks before cooling:                          10000
- Min. accepted networks before reannealing:                        500
-------------------------------------------------------------------------
- Output directory:                                      data/static/output
- Report file:                                            static.report.txt
- Number of best networks tracked:                                    1
- Max. time:                                                      1.0 h
- Min. networks before checking:                                   1000
- Screen reporting interval:                                     20.0 s
- File reporting interval:                                       10.0 m
-------------------------------------------------------------------------
- Compute influence scores:                                         yes
- Compute consensus graph:                                          yes
- Create consensus graph as HTML:                                   yes
- Create 'dot' output:                                              yes
- Location of 'dot':               C:/Program Files/ATT/Graphviz/bin/dot.exe
-------------------------------------------------------------------------
```

Search results for acid-alcohol-ester group




- Search Statistics
Statistics collected in searcher 'SearcherSimAnneal':
Number of networks examined: 246684000
Total time used: 90.02 d
High score: -3043.7090, first found at iteration 24575853
Number of re-anneals: 7706
Statistics collected in proposer 'ProposerRandomLocalMove':
Additions -- proposed: 83562224
Deletions -- proposed: 81558116
Reversals -- proposed: 81563659
Statistics collected in cycle checker 'CycleCheckerCheckThenApply':
Additions -- considered: 83562224, acyclic: 76827986
Deletions -- no cyclicity test necessary
Reversals -- considered: 81563659, acyclic: 79242053
Statistics collected in evaluator 'EvaluatorBDe':
Scores computed: 6971897
Scores (cache) placed fetched
with 0 parents: 35 121776087
with 1 parents: 1190 97880906
with 2 parents: 19635 81692379
with 3 parents: 5573541 7280344
with 4 parents: 1302693 1274210
with 5 parents: 74803 15280
Statistics collected in decider 'DeciderMetropolis':
Additions -- considered: 76827986, better score: 4384325, other accepted:
19544154
Deletions -- considered: 81558116, better score: 19646909, other accepted
4281564
Reversals -- considered: 79242053, better score: 24352592, other accepted
7732982
Average permissivity: 0.167
Memory info after completing the search: Banjo is using 1 mb of memory


- Post-processing
Influence scores
Influence score for (1,0) -> (0,0) .8313
Influence score for (6,0) -> (1,0) .0028
Influence score for (4,0) -> (1,0) .0043
Influence score for (28,0) -> (2,0) -.3418
Influence score for (8,0) -> (2,0) .0015
Influence score for (4,0) -> (2,0) .0021
Influence score for (18,0) -> (3,0) .0833
Influence score for (2,0) -> (3,0) .0000
Influence score for (0,0) -> (3,0) -.0628
Influence score for (15,0) -> (4,0) -.0074
Influence score for (14,0) -> (4,0) .0979
Influence score for (13,0) -> (5,0) .3721
Influence score for (20,0) -> (6,0) .0000
Influence score for (9,0) -> (7,0) .0000
Influence score for (5,0) -> (7,0) .0000
Influence score for (10,0) -> (9,0) .5654
Influence score for (12,0) -> (10,0) .4049
Influence score for (3,0) -> (11,0) .0000
Influence score for (18,0) -> (13,0) .2904
Influence score for (9,0) -> (13,0) .0000
Influence score for (5,0) -> (14,0) .5781
Influence score for (30,0) -> (15,0) .5234
Influence score for (10,0) -> (16,0) .0000
Influence score for (10,0) -> (17,0) .3735
Influence score for (26,0) -> (18,0) .0000
Influence score for (13,0) -> (20,0) .0000
Influence score for (13,0) -> (21,0) .5034
Influence score for (5,0) -> (22,0) .0000
Influence score for (26,0) -> (24,0) .0000
Influence score for (15,0) -> (25,0) .2990
Influence score for (9,0) -> (26,0) .5992
Influence score for (23,0) -> (27,0) .0000
Influence score for (13,0) -> (27,0) .0000
Influence score for (27,0) -> (28,0) .0015
Influence score for (21,0) -> (28,0) -.0850
Influence score for (29,0) -> (30,0) .3602
Influence score for (16,0) -> (30,0) .0000
Influence score for (3,0) -> (32,0) .0000
Influence score for (0,0) -> (32,0) -.2401
Influence score for (32,0) -> (33,0) .4680
Influence score for (1,0) -> (34,0) -.3959

```
------------------------------------------------------------------------------
- Post-processing                                    DOT graphics format output
------------------------------------------------------------------------------

digraph abstract {

label = "Banjo Version 2.0.1\nHigh scoring network, score: -3043.7090\nProject:
aae35 example\nUser: demo\nDataset: 35-vars-123-observations\nNetworks searched:
 246684000";
labeljust="l";

    0 [label="ethylbutanoate"];
    1 [label="methylbutanoate"];
    2 [label="methylbenzoate"];
    3 [label="ethylbenzoate"];
    4 [label="butylbenzoate"];
    5 [label="hexylbutanoate"];
    6 [label="ethyloctanoate"];
    7 [label="butylhexanoate"];
    8 [label="methyl2furoate"];
    9 [label="butyl3hydroxybutanoate"];
    10 [label="ethyl3.methylthiopropanoate"];
    11 [label="2methylpropylbutanoate"];
    12 [label="ethyl2furancarboxylate"];
    13 [label="butylbut2enoate"];
    14 [label="hexylbenzoate"];
    15 [label="phenylmethylbutanoate"];
    16 [label="propylbenzoate"];
    17 [label="methyl3methylthiopropanoate"];
    18 [label="ethylbutE2enoate"];
    19 [label="3hydroxybutylbenzoate"];
    20 [label="ethylhex2enoate"];
    21 [label="3methylbutylbutanoate"];
    22 [label="hexylhexanoate"];
    23 [label="ethylacetate"];
    24 [label="butyloctanoate"];
    25 [label="benzylbenzoate"];
    26 [label="butanoicacid"];
    27 [label="octanoicacid"];
    28 [label="hexE2enol"];
    29 [label="hexanol"];
    30 [label="benzylalcohol"];
    31 [label="pentan2ol"];
    32 [label="1.3Trimethyl2oxabicyclo"];
    33 [label="exo2hydroxycineole"];
    34 [label="Unknown228"];

  1->0;
  4->1;
  6->1;
  4->2;
  8->2;
  28->2;
  0->3;
  2->3;
  18->3;
  14->4;
  15->4;
  13->5;
  20->6;
  5->7;
  9->7;
  10->9;
  12->10;
  3->11;
  9->13;
  18->13;
  5->14;
  30->15;
  10->16;
  10->17;
  26->18;
  13->20;
  13->21;
  5->22;
  26->24;
  15->25;
  9->26;
  13->27;
  23->27;
  21->28;
  27->28;
  16->30;
  29->30;
  0->32;
  3->32;
  32->33;
  1->34;
```

References

Bartlett, M. S., (1947). The use of transformation. *Biometric Bulletin,* 3, 39-52.

Berthouex, P.M. and Brown, L.C. (2002). *Statistics for Environmental Engineers* (2nd Edition). Lewis Publishers.

Bioconductor packages obtained from http//www.bioconductor.org

Buntine, W. (1991). Theory refinement on Bayesian networks. In *Proceedings of Seventh Conference on Uncertainty in Artificial Intelligence,* Los Angeles, CA, pages 52-60. Morgan Kaufmann.

Chickering, D.M.(1996). Learning Bayesian networks is NP-complete. In D.Fisher & H.-J.Lenz, eds, *Learning from Data: Arificial Intelligence and Statistics V,* chap.12, 121-130. Springer Verlag.

Chong, E. K. P. and Zak, S. H. (2008). *An Introduction to Optimization (Wiley-Interscience Series in Discrete Mathematics and Optimization)*. Wiley-Interscience, 3 edition.

Cooper, G.F. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning,* 9, 309-347.

Durbin, B.P., Hardin, J.S., Hawkins, D.M., and Rocke, D.M. (2002). A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18: S105-S110.

Eisen, M., Spellman, P., Brown, P. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS*, *95*:14863-14868.

Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000) Using Bayesian Networks to Analyze Expression Data. *Journal of Computational Biology,7(3-4):*601-620. doi:10.1089/106652700750050961.

Geiger, D. and Heckerman, D. (1995). A characterization of the Dirichlet distribution applicable to learning Bayesian networks. Technical Report MSR-TR-94-16, Microsoft Research, Redmond, WA.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 5, R80.

Gentleman, R. C., Carey, V., Huber, W., et al. (2005) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. New York: Springer.

Gentleman, R., Hahne, F. and Huber, W. (2006). Visualizing Genomic Data. *Bioconductor Project Working Papers.* Working Paper 10. http://www.bepress.com/bioconductor/paper10

Hartemink, A. (2001). Principled Computational Methods for the Validation and Discovery of Genetic Regulatory Networks._Massachusetts Institute of Technology, Ph.D. dissertation.

Hartemink, A.J. (2005). Bayesian Network Inference with Java Objects (BANJO). [BANJO algorithm web site at Duke University]. Available from: http://www.cs.duke.edu/~amink/software/banjo/

Hartemink, A., Gifford, D., Jaakkola, T. and Young, R. (2002) Combining Location and Expression Data for Principled Discovery of Genetic Regulatory Network Models._In *Pacific Symposium on Biocomputing 2002 (PSB02)*, Altman, R., Dunker, A.K., Hunter, L., Lauderdale, K., & Klein, T., eds. World Scientific: New Jersey. pp. 437–449.

Hastie,T., Tibshirani, R. and Friedman, J. (2001). *The elements of statistical learning: data mining, inference and prediction.* New York: Springer.

Heckerman, D. (1998). A Tutorial on Learning with Bayesian Networks. Technical Report MSR-TR-95-06, Microsoft Research, Redmond, Washington. March 1995 (revised Nov 1996). Available at ftp://ftp.research.microsoft.com/pub/tr/TR-95-06.PS

Heckerman, D., Geiger, D.and Chickering, D.M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20: 197-243.

Heckerman, D., Mamdani, A. and Wellman, M. (1995b). Real-world applications of Bayesian networks. *Communications of the ACM*, 38.

Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic bayesian networks. *Bioinformatics*, 19(17):2271-2282.

Jensen, F. (1996). *An Introduction to Bayesian Networks*. Springer.

Jolliffe, I.T. (2002). *Principal Component Analysis* (2^{nd} Edition). Springer, New York.

Kennedy, P. (2003). *A Guide to Econometrics* (5^{th} Edition).Cambridge, MA: The MIT Press.

King, J. R. and D. A. Jackson. (1999). Variable selection in large environmental data sets using principal components analysis. *Environmetrics* 10, 67–77.

Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. (1983). Optimization by simulated annealing. *Science* 220, 671-680.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, 63:215-232.

Manly, Bryan F.J. (2005). *Multivariate statistical methods: a primer*. Chapman & Hall/CRC.

McMath, K.L., Patterson, V.J., Young, H., Macrae, E.A. and Ball, R.D. (1992). Factors affecting the sensory perception of sweetness and acidity in kiwifruit. Acta Horticulturae 20, 489-500.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156-166.

Mischel, P. S., Shai, R., Shi, T., Horvath, S., Lu, K. V., Choe, G., Seligson, D., Kremen, T. J., Palotie, A., Liau, L. M., Cloughesy, T. F., and Nelson, S. F. (2003). Identification of molecular subtypes of glioblastoma by gene expression profiling. *Oncogene*, 22(15):2361-2373.

Misevičius, A. (2003). A Modified Simulated Annealing Algorithm for the Quadratic Assignment Problem. *Informatica* 14(4), 497-514.

Osborne, J. (2002). Notes on the use of data transformations. *Practical Assessment, Research & Evaluation*, 8(6). Retrieved May 6, 2009, from http://PAREonline.net/getvn.asp?v=8&n=6.

Paul E. Black, "simulated annealing", in *Dictionary of Algorithms and Data Structures* [online], Paul E. Black, ed., U.S. National Institute of Standards and Technology. 30 March 2009. (accessed TODAY) Available from: http://www.itl.nist.gov/div897/sqg/dads/HTML/simulatedAnnealing.html

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Francisco, Calif.

Rowan, D., Hunt, M., Alspach, P., Dimouro, A., Chagné, D., Weskett, R. and Volz, R. (2007). *Metabolic profiling of fruit volatiles in progeny of a 'Royal Gala' X 'Granny Smith' apple cross.* Metabolomics Society's 3rd Annual International Conference, 11-14 June 2007, Manchester, U.K. Pp. 68. (Poster abstract PMS022.).

Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D. and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *sacccharomyces cerevisiae* by microarray hybridization, *Molecular Biology of the Cell* 9, 3273–3297.

Spiegelhalter, D.J and Lauritzen, S.L. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20, 579-605.

Tabachnick, B. G. and Fidell, L. S. (1996). *Using Multivariate Statistics* (3rd Edition). HarperCollins College Publishers.

Van den Boogaart, K. G., Tolosana-Delgado, R. and Bren, M. (2006) Concepts for handling of zeros and missing values in compositional data, In: 2006 Annual Conference of the International Association for Mathematical Geology (IAMG), Liège, Belgium, September 3-8, 2006. (extended abstract and oral presentation).

Yamamura, K. (1999). Transformation using (x+0.5) to stabilize the variance oof populations. *Researches on Population Ecology,* 41: 229-234.

Yin, W., Yang, Q., Yao, S. and Shi, Y. (2005). Hierarchical Clustering Analysis from Genomic Dataset. In *Proceedings of the IEEE/WIC/ACM international Conference on intelligent Agent Technology* (September 19 - 22, 2005). IAT. IEEE Computer Society, Washington, DC, 759-762. doi: http://dx.doi.org/10.1109/IAT.2005.81

Young, H., Gilbert, J.M., Murray, S.H. and Ball, R.D. (1995). Factors contributing to kiwifruit flavour. Hort. Res. Client Report 95/112.