Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

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

Abstract……………………………………………………………………………….i

Acknowledgements……………………………………………………………….ii

Table of Contents…………………………………………………………….......iii

List of Figures…………………………………………………………………….v

List of Tables……………………………………………………………………...vi

Chapter 1: Introduction

1.1 Background………………………………………………………………1

1.2 Description of data…………………………………………………….........1

Chapter 2: Literature Review

2.1 Data transformation……………………………………………………….3

2.2 Cluster Analysis…………………………………………………………...4

2.3 Bayesian networks…………………………………………………………...5

Chapter 3: Transformation methodology

3.1 Motivation……………………………………………………………........9

3.2 Selection of constant by simulation………………………………………...11

Chapter 4: Multivariate techniques

4.1 Principal component analysis……………………………………………....16

4.2 Cluster analysis……………………………………………………………..18

Chapter 5: Bayesian networks

5.1 Introduction……………………………………………………………...20

5.2 Learning the Structure of Bayesian networks……………………………...21

5.3 BANJO (Bayesian Network Inference with Java Objects)………………...23

Chapter 6: Results

6.1 Overview……………………………………………………………….. 29

6.2 Heat maps…………………………………………………………………..29

6.3 Graphical representation of Bayesian networks……………………………32

6.4 Comments made on the graphs……………………………………………..38

Chapter 7: Conclusion and Discussion

7.1 Conclusion…………………………………………………………………40

7.2 Discussion ……………..…………………………………………………40

7.3 Future study………………………………………………………………...41

Appendix……………………………………………………………………………43

References…………………………………………………………………………...51

List of Figures

Figure 2.1 An example of a simple Bayesian network structure……………………6

Figure 3.1 Histograms of carvenone and butyl acetate which belong to the group of ‘monoterpene’ and ‘ester’ in the compound database…………………...9

Figure 3.2 Plots of carvenone and isopiperitenone for different delta values……...10

Figure 3.3 Plots of butyl acetate and ethyl pentanoate for different delta values….11

Figure 3.4 Box plot showing gap1 and gap2 for random normal samples and monoterpene group…………………………………………………..13

Figure 3.5 Box plot showing gap1 and gap2 for random normal samples and acid, alcohol & ester group………………………………………………..13

Figure 3.6 Histograms of transformed carvenone and butyl acetate which belong to the group of ‘monoterpene’ and ‘ester’ in the compound database……14

Figure 3.7 Histograms of untransformed and transformed hex-E2-enol which belong to the group of ‘alcohol’ in the compound database…………………...15

Figure 6.1 Heat map representing 123 genotypes and 16 monoterpene compounds…………………………………………………………...30

Figure 6.2 Heat map representing 123 genotypes and 35 compounds of acid, alcohol & esters……………………………………………………………32

Figure 6.3 Compounds pathway known for monoterpene group…………………..33

Figure 6.4 Graphical representation of Bayesian networks for monoterpene group with 45 compounds…………………………………………………..34

Figure 6.5 Graphical representation of Bayesian networks for monoterpene group with 16 compounds chosen by principal component analysis……….....36

Figure 6.6 Graphical representation of Bayesian network for acid, alcohol & ester group with 35 compounds chosen by principal component analysis…..37

List of Tables

Table 5.1 Combination of different initial temperature and cooling factor with maximum time of 15minutes…………………………………………...26

Table 5.2 Combination of different initial temperature and cooling factor with maximum time of 30minutes…………………………………………...27

Table 5.3 Combination of different initial temperature and cooling factor with maximum time of an hour……………………………………………...27

Table 6.1 Division of Bayesian networks graph for monoterpene group with 45 compounds into four levels and compounds present in that level…………………………………………………………………….35