PROCESSING AND ANALYSIS OF CHROMATOGRAPHIC DATA

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Rhiannon Maree Parker
B. Sc. (Hons)

School of Applied Sciences
College of Science, Engineering and Health
RMIT University
March 2012

Declaration

I certify that except where due acknowledgement has been made, the work is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program; any editorial work, paid or unpaid, carried out by a third party is acknowledged; and ethics procedures and guidelines have been followed.

Rhiannon Maree Parker

March 2012

Acknowledgements

First and foremost I would like to express my utmost gratitude to my supervisor, Professor Mike Adams, for his patience, guidance and support throughout my PhD. I appreciate all your advice and expertise; you made my PhD experience interesting and enjoyable. Thanks also for the conference trips, coffee, lunches and dinners. I am indebted to you for giving me the opportunity to work in Switzerland for three months; this experience changed my life for the better and I will never forget it.

I am grateful to the many people at Deakin University, Prof. Neil Barnett, Dr. Xavier Conlan, Dr. Sara Bellomarino and Dr. Tiffany Gunning, who collaborated with the studies. Without them this research would not have been complete.

I express my gratitude to Paul Morrison and Dr. Jacqui Adcock for all their GC×GC assistance.

My sincere thanks to Frederic Begnaud and Alain Chaintreau for all their kindness and guidance during my time in Geneva. Working at Firmenich for three months was invaluable for the development of the software and I greatly appreciate it. I would especially like to thank Frederic for inviting me to spend my birthday with his family on my first day in Geneva; you made me feel at home in a foreign country and for this I am very thankful. I would also like to thank Christian Debonneville for all his help with GC×GC and Lidia Avois-Mateus for her assistance with the flavour samples and for giving me a tour of the quality control laboratories.

Thanks also to all my friends for their support and morning coffee. You made my time at university more enjoyable.

I would like to thank my Uncle Mark for all he has done for me and my Uncle Bruce and Auntie Jill for providing me with a place to live during my time at university, without you this would not have been possible.

To my Nan and Pa, thank you for always being proud of me and supporting me throughout my schooling. Pa, I feel that you are always with me, giving me the strength to do anything.

To my brother, Robert, thank you for just being you. I really appreciate that you were always there to offer encouragement when I needed it most. I could not have wished for anything more.

Finally, to my mum, who has always believed in me and supported me every step of the way. Thank you for listening to my problems and putting up with me when I was stressed. Your love and knowing that you are proud of me has motivated me to do my best and achieve my goals. You have sacrificed so much to give me the best possible start in life, for this and so much more, I am dedicating this thesis to you.

Abstract

Data pre-processing and analysis techniques are investigated for the analysis of one- and two-dimensional chromatographic data. Pre-processing, in particular alignment, is of paramount importance when employing multivariate chemometric methods as these techniques highlight variance, or changes between samples at corresponding variables (i.e. retention times).

Principal components analysis (PCA) was employed to evaluate the effectiveness of alignment. Two methods, correlation optimised warping and *i*coshift were compared for the alignment of high performance liquid chromatography (HPLC) metabolite data. PCA was then employed as an exploratory technique to investigate the influence of phosphite on the secondary metabolites associated with *Lupinus angustifolius* roots inoculated with the pathogen, *Phytophthora cinnamomi*.

In a second application, HPLC with acidic potassium permanganate chemiluminescence detection was evaluated for the analysis of Australian wines from different geographic origins and vintages. Linear discriminant analysis and quadratic discriminant analysis were used to classify red and white wines according to geographic origin. In the analysis of wine vintage, partial least squares and principal components regression were compared for the modelling of sample composition with wine age.

Finally, software was developed for quality control (QC) of flavours and fragrances using comprehensive two-dimensional gas chromatography (GC×GC). The software aims to automatically align and compare a sample chromatogram to a reference chromatogram. A simple method of partitioning the two-dimensional pattern space was employed to select reference control points. Corresponding control points in a sample chromatogram were identified using a triangle-pattern matching algorithm. The reference and sample control points were then used to calculate the translation, scaling and rotation operations for an affine transform, which is applied to the complete sample peak list in order to align reference and sample peaks. Comparison of reference and sample chromatograms was achieved through the use of fuzzy logic.

It is concluded that the pre-processing and chemometric methods investigated here are valuable tools for the analysis of chromatographic data. The developed GC×GC software was successfully employed to analyse real flavour samples for QC purposes.

Publications

During the course of this project, a number of papers based on the work presented in this thesis were accepted for international publication. They are listed for reference:

- S. A. Bellomarino, X. A. Conlan, **R. M. Parker**, N. W. Barnett and M. J. Adams, *Geographical classification of some Australian wines by discriminant analysis using HPLC with UV and chemiluminescence detection*, Talanta, 80 (2009) 833-838.
- S. A. Bellomarino, **R. M. Parker**, X. A. Conlan, N. W. Barnett and M. J. Adams, *Partial least squares and principal components analysis of wine vintage by high performance liquid chromatography with chemiluminescence detection*, Analytica Chimica Acta, 678 (2010) 34-38.

In preparation for submission:

- **R. M. Parker**, J. L. Adcock, F. Begnaud, A. Chaintreau and M. J. Adams, *Registration and alignment of comprehensive two-dimensional gas chromatographic data*, in preparation for submission to Analytica Chimica Acta.
- T. K. Gunning, X. A. Conlan, **R. M. Parker**, G. A. Dyson, M. J. Adams, N. W. Barnett and D. M. Cahill, *Chemometric profiling of plant secondary metabolites from L. angustifolius plant roots inoculated with Phytophthora cinnamomi*, in preparation for submission to Analytical Biochemistry.

Presentations

During the course of this project, a number of public presentations have been made based on the work presented in this thesis. They are listed for reference:

- **R. M. Parker**, M. J. Adams, J. L. Adcock and P. J. Marriott, *Pattern matching in comprehensive two-dimensional gas chromatography*, poster presentation at the 16th annual RACI Environmental and Analytical Division R&D Topics conference, Macquarie University, New South Wales, Australia, 2008.
- **R. M. Parker**, M. J. Adams, P. J. Marriott, F. Begnaud and A. Chaintreau, *Comprehensive two-dimensional gas chromatography analysis software for quality control of flavours and fragrances*, poster presentation at the World Analytical Meeting, Geneva, Switzerland, 2010.
- **R. M. Parker** and M. J. Adams, *GC×GC analysis software for quality control of flavours and fragrances*, oral presentation at Firmenich, Geneva, Switzerland, 2010.
- **R. M. Parker** and M. J. Adams, *Comparison of comprehensive two-dimensional gas chromatograms*, poster presentation at the 18th annual RACI Environmental and Analytical Division R&D Topics conference, University of Tasmania, Tasmania, Australia, 2010.

Table of Contents

Abbreviations

1D	One-dimensional
2D	Two-dimensional
3D	Three-dimensional
ANN	Artificial neural networks
CA	Cluster analysis
COW	Correlation optimised warping
DAD	Diode array detector
Dim 1	First dimension
Dim 2	Second dimension
DTW	Dynamic time warping
EDA	Exploratory data analysis
ESI-QTOF-MS	Electrospray ionisation quadrupole time-of-flight mass spectrometry
FID	Flame ionisation detector
GC	Gas chromatography
GC×GC	Comprehensive two-dimensional gas chromatography
HPLC	High performance liquid chromatography
IRMS	Isotope ratio mass spectrometry
k-NN	*k*- nearest neighbours
LC	Liquid chromatography
LC×LC	Comprehensive two-dimensional liquid chromatography
LDA	Linear discriminant analysis
MIR	Mid-infrared
MS	Mass spectrometry
MSPE	Mean squared prediction error
NIR	Near-infrared
NMR	Nuclear magnetic resonance
PCA	Principal components analysis
PCR	Principal components regression
PLS	Partial least squares
PLS-DA	Partial least squares discriminant analysis
PTW	Parametric time warping
PWA	Piecewise alignment
QC	Quality control
QDA	Quadratic discriminant analysis
RI	Refractive index
SIMCA	Soft independent modelling of class analogy
TOF-MS	Time-of-flight mass spectrometry
UV	Ultraviolet

Chapter 1 - Introduction

1.1 Chromatography

Chromatography represents a range of separation methods in which the components to be separated are distributed between two phases, one of which is stationary (the stationary phase) while the other (the mobile phase) moves in a definite direction [1]. For a gaseous mobile phase, the process is known as gas chromatography (GC) and liquid chromatography (LC) if a liquid is used.

The chromatographic process involves passing the mobile phase over and through the stationary phase. During this process the components of the mixture are distributed between the two phases and the amount of interaction with the sorbent bed results in different migration rates through the system. The sorption-desorption process occurs many times as the molecule moves through the bed, and the time required to do so depends mainly on the proportion of time the molecule is held in the stationary phase. Separation is achieved if the various components emerge from the bed at different times, referred to as retention times [2].

1.1.1 One-dimensional chromatography

1.1.1.1 Gas chromatography

In GC, volatile analytes partition between a stationary phase and a gaseous mobile phase. A traditional GC instrument consists of a carrier gas (mobile phase), an injector, a column (stationary phase) and a detector (Figure 1.1). A volatile liquid or gaseous sample is injected through a septum into a heated port where it evaporates. This vapour is swept through the column by carrier gas and the analytes are separate from one another based on their relative vapour pressures and affinities for the stationary bed. The separated analytes then flow through a detector and the response measured [3, 4].

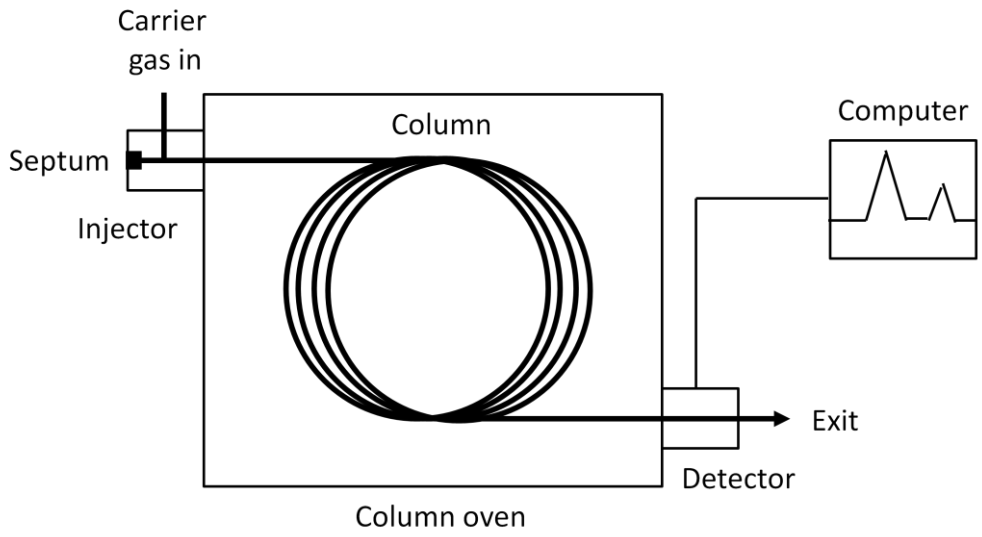


Figure 1.1: Schematic of a GC system. Adapted from [4].

Both open tubular and packed columns are used in GC. Open tubular columns are preferred for general use as they provide higher resolution, shorter analysis times and greater sensitivity. However, they do require higher operating pressure and have less sample capacity. Open tubular columns include wall-coated columns which contain a thin film of stationary liquid phase on the inner wall of the column, support-coated columns which have solid particles coated with liquid stationary phase that are attached to the inner wall and porous-layer columns in which the solid particles are the active stationary phase. Stationary phase selection is based on the rule "like dissolves like". Non-polar columns are best for non-polar solutes, while columns of intermediate polarity are best for intermediate solutes and strongly polar columns are best for strongly polar solutes [3, 4]. The most popular class of liquid stationary phases are silicone polymers, differing in the extent to which they contain polar functional groups. Packed columns contain fine particles of solid support coated with non-volatile liquid stationary phase. Alternatively, the solid particles themselves may be the stationary phase. Despite their inferior resolution, packed columns are used for preparative separations, where a large amount of stationary phase is required, or to separate gases which are poorly retained. Columns are usually made of stainless steel or glass with diatomite solid support that has been silanised to reduce hydrogen bonding to polar solutes [2, 4]. Packed and capillary columns have been reviewed in the literature [5, 6].

GC separations can be carried out at constant temperature (isothermal separation mode) or in the case of complicated analyte mixtures, the column temperature can be increased according to a temperature program. In temperature programming, the temperature of the column is raised during the separation to increase solute vapour pressure and decrease retention times of late eluting compounds. When a constant temperature is used more volatile compounds elute close together and less volatile compounds, if they are even eluted from the column, will elute much later and have broader peak shapes. By increasing the temperature according to a temperature program, all compounds elute and the separation will be fairly uniform. Since analyte retention time depends considerably on the column temperature precise regulation is required to ensure reproducible results [4, 7].

Various injectors have been developed for delivering the sample to the head of the separation column with the smallest possible bandwidth. These include split injection, splitless injection and on-column injection. In a split injection only 0.2-2% of the sample is delivered to the column. This is suited to high resolution work, where the best results are obtained using the smallest amount of sample that can adequately be detected. The sample is rapidly injected through the septum into the evaporation zone of the glass liner. The injector temperature is kept high in order to promote fast evaporation. Carrier gas then sweeps the sample through the mixing chamber, where complete vaporisation and mixing occur. At the split point only a small fraction of vapour enters the chromatography column, while most passes through the needle valve to a waste vent. The portion of sample that does not reach the column is called the split ratio; this typically ranges from 50:1 to 600:1 [4]. Splitless injection is preferred for trace analysis. The same port for split injection is used, however the glass liner is a straight, empty tube with no mixing chamber. A large volume (~ 2 μL) of dilute solution in a low-boiling point solvent is slowly injected into the liner with the split vent closed. The temperature of the injector is lower for splitless injection as the sample spends more time in the port and may decompose at higher temperatures. In splitless injection, ~ 80% of the sample is applied to the column, and little fractionation occurs during injection. The initial column temperature is set 40°C below the boiling point of the solvent, which therefore condenses at the beginning of the column. As the solutes slowly catch up with the condensed plug of solvent, they are trapped in the solvent as a narrow band at the head of the column; this leads to

sharp chromatographic peaks. Chromatography is initiated by raising the column temperature to vaporise the solvent trapped at the head of the column [4]. On-column injection is used for samples that decompose above their boiling point and is often preferred for quantitative analysis. In on-column injection, solution is injected directly onto the column, without going through the hot injector. The initial column temperature is low enough to condense solutes in a narrow zone. Chromatography is initiated by warming the column [4]. Injectors for GC have been discussed in the literature [8-10].

The sample is moved through the column by carrier gas. The carrier gas must be inert and not chemically interact with the sample. The most popular carrier gases are hydrogen, helium and nitrogen and the choice is often dependent on the detector used. Measurement and control of carrier gas flow is essential to ensure a constant and reproducible flow rate and hence reproducible retention times [3].

There are a number of detectors which can be used in GC; these are summarised in Table 1.1. Detectors interact with the eluted compounds and the interaction is converted into an electrical signal which is sent to a recorder. The data is presented by plotting the intensity of the signal versus the time of analysis; this is the so-called chromatogram [7]. The detector signal is proportional to the quantity of each analyte making it possible to perform quantitative analysis. The most commonly employed detector for GC is the flame ionisation detector (FID), as it has high sensitivity, a large linear response range and low noise.

Table 1.1: Summary of commonly used GC detectors [4, 11]

Detector	**Selectivity and principle**	**Approximate detection limit**
Flame-ionization detector	In the FID, eluate is burned in a mixture of hydrogen and nitrogen (in air). Carbon atoms (except carbonyl and carboxyl carbons) produce CH radicals, which are thought to produce CHO^+ ions in the flame: $CH + O \rightarrow CHO^+ + e^-$ Electrons flow from the anode to the cathode, where they neutralise CHO^+ in the flame. This current is the detector signal.	2 pg/s
Thermal conductivity detector	Thermal conductivity detectors are universal detectors for anything providing a difference in thermal conductivity from the carrier gas. Eluate from the column flows over a hot filament. When an analyte emerges from the column, the conductivity of the gas stream decreases, the filament gets hotter, its electrical resistance increases, and the voltage across the filament changes. This change in voltage is measured by the detector.	400 pg/mL (propane)
Electron capture detector	Electron capture detectors are particularly sensitive to halogenated compounds. Gas entering the detector is ionised by high-energy electrons emitted from a radioactive source. The electrons formed are attracted to an anode, producing a baseline current. When analyte molecules with a high electron affinity enter the detector, they capture some of these electrons. The detector responds by varying the frequency of voltage pulses between the anode and the cathode to maintain a constant current. This frequency is converted to a voltage which is proportional to the concentration of the eluting compound.	As low as 5 fg/s

Detector (cont.)	Selectivity and principle (cont.)	Approximate detection limit (cont.)
Mass spectrometry (MS) detector	Mass spectrometers are universal detectors for total or single ion monitoring. To obtain a mass spectrum, gaseous molecules or species desorbed from condensed phases are ionised. These ions are accelerated by an electric field and are separated according to their mass-to-charge ratio, *m/z*.	25 fg to 100 pg
Flame photometric detector	Flame photometric detectors are selective to S- and P-containing compounds. They filter and measure light emitted when a sample is burned in a hydrogen-rich flame.	< 1 pg/s (phosphorous) < 10 pg/s (sulfur)
Sulfur chemiluminescence detector	Sulfur chemiluminescence detectors are selective to sulfur-compounds: S-compounds are oxidised to produce SO, after ozonisation SO gives SO_2^* that decays to ground state producing a signal.	100 fg/s (sulfur)

GC applications have been frequently reviewed in the literature and include the analysis of essential oils [12], pesticides [13], petroleum [14], oil [15], fatty acids [16], steroids [17] and biological fluids [18].

1.1.1.2 High performance liquid chromatography

High performance liquid chromatography (HPLC) is one of the most extensively employed liquid chromatographic methods; the two most important forms of HPLC are normal phase and reversed phase. In normal phase chromatography the retention order is based on increasing hydrophilicity, while in reversed-phase chromatography it is based on increasing hydrophobicity [19].

HPLC uses high pressure to force solvent through closed columns containing very fine particles which give high-resolution separations. A typical HPLC instrument consists of solvent (mobile phase), a pump, an injector, a column (stationary phase) and a detector (Figure 1.2).

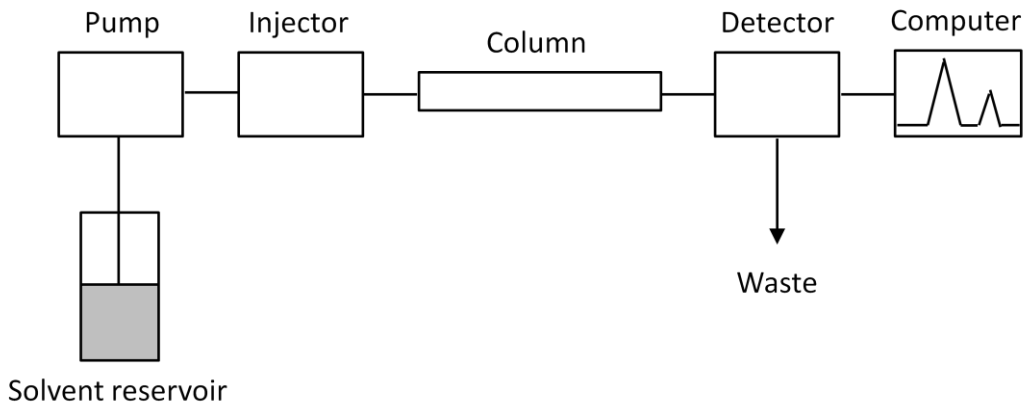


Figure 1.2: Schematic of a HPLC system. Adapted from [2].

In HPLC, column efficiency is increased by increasing the rate at which solute equilibrates between the stationary and mobile phases. For GC, with an open tubular column, this is achieved by decreasing the thickness of the stationary phase and reducing the diameter of the column so that molecules can quickly diffuse between the channel and the stationary phase. However, since diffusion in liquids is 100 times slower than in gases, it is generally not feasible to use open tubular columns as the diameter of the solvent channel is too great for a solute molecule to cross in a short time. Hence, packed columns are often used in liquid chromatography. The efficiency of packed columns increases as the size of the stationary-phase particles decreases. This is due to the fact that they provide more uniform flow through the column and the distance the solute must diffuse in the mobile phase is less. However, smaller particle size results in resistance to solvent flow. The most common stationary phase for HPLC is highly pure, spherical, microporous particles of silica that are permeable to solvent and have a surface area of several hundred square meters per gram [4]. Monolithic columns can also be used in HPLC. In some instances monoliths offer an advantage over particle packed columns, namely, they can be operated at high flow rates allowing fast separation of complex mixtures. Monolithic columns consist of one piece of continuous, porous material and can be divided into two main classes based on whether they employ organic or inorganic precursors [20]. Both packed and monolithic have been reviewed by in the literature [21-24].

In adsorption chromatography, solvent molecules compete with solute molecules for sites on the stationary phase and elution occurs when solvent displaces solute from the

stationary phase. The more polar the solvent, the greater its eluent strength for adsorption chromatography with bare silica. The greater the eluent strength, the more rapidly solutes will be eluted from the column. Adsorption chromatography on bare silica is an example of normal phase chromatography, where a polar stationary phase and a less polar solvent are used. In reversed-phase chromatography, the stationary phase is non-polar or weakly polar and the solvent is less polar. This means a less polar solvent has a higher eluent strength. Peak tailing is eliminated in reversed-phase chromatography as the stationary phase has few sites that can strongly absorb solute to cause tailing [4].

When elution is performed with a single solvent (or constant solvent mixture), it is referred to as isocratic elution. If one solvent does not provide sufficiently rapid elution of all components, then gradient elution can be used. Gradient elution in HPLC is analogous to temperature programming in GC. It involves changing the eluent strength of the mobile phase, which gradually changes the composition of the eluent entering the column; this accelerates the elution of peaks which would otherwise elute late or not at all. Gradient elution is widely employed in reversed-phase chromatography [25].

Selection of the appropriate mobile phase is based on a wide range of criteria, including viscosity, ultraviolet (UV) transparency, refractive index (RI), boiling point, purity and it must be inert with respect to sample compounds. As a general rule the mobile phase should not be detector-active, otherwise unwanted baseline effects and extra peaks may appear in the chromatogram [26].

The mobile phase is delivered at a constant flow rate through the column by a pump. The pump must be capable of generating high pressures with high flow accuracy and precision at the chosen flow rate. Moreover, the flow should be pulse-free and the internal volume must be low in order to enable a quick solvent change. Most HPLC pumps use a reciprocating piston design [26].

A sample injector introduces the sample into the chromatograph as a sharp plug to minimise dispersion and peak broadening. Valve injectors are generally used in commercial instrumentation as the sample is introduced with minimal flow interruption [27].

There are many detectors which can be used in HPLC; these are summarised in Table 1.2. The most commonly employed detector for HPLC is the UV-visible absorbance detector as many compounds of interest absorb in the UV (or visible) range [28].

Table 1.2: Summary of commonly used HPLC detectors [4, 28-30]

Detector	**Selectivity and principle**	**Approximate limit of detection (ng)**
UV-Vis absorbance detector	UV-Vis absorbance detectors are near universal as many solutes absorb UV light. They measure the amount of light transmitted through a solution given by Beer's law: $A = \in cl$	0.1-1
Refractive index detector	RI detectors are sensitive to all analytes that have a RI different than the mobile phase. Light passes through a cell with pure solvent and is directed to a photocell by a deflection plate. When solute with a different RI enters the cell, the beam is deflected and the photocell output changes. The deflection of light is measured by Snell's law: $n_1 \sin\theta_1 = n_2 \sin\theta_2$	100-1000
Fluorescence detector	Fluorescence detectors are very sensitive, however they only respond to the few analytes that fluoresce. They excite the eluate with a laser and measure fluorescence.	0.001-0.01
Conductivity detector	Conductivity detectors are very selective and can be used whenever the sample bands have different conductivity from the running buffer.	0.5-1
Electrochemical detector	Electrochemical detectors are very selective and sensitive; they measure the electron flow generated at electrode surfaces during oxidation or reduction reactions.	

Detector (cont.)	Selectivity and principle (cont.)	Approximate limit of detection (cont.)
Evaporative light scattering detector	Evaporative light scattering detectors are considered universal as they respond to any analyte that is significantly less volatile than the mobile phase. Eluate is formed into a uniform dispersion of droplets. Solvent is evaporated from the droplets, leaving a fine mist of solid particles. These particles are then detected by the light they scatter.	0.1-1

Chemiluminescence detection can also be employed in HPLC. Chemiluminescence is simply the emission of light from a chemical reaction. There are two types of chemiluminescence reactions, direct and indirect. In direct chemiluminescence, the reaction between compounds A and B forms a product or intermediate in an electronically excited state (C^*) that returns to its ground state by the ejection of a photon (Equation 1.1). In indirect chemiluminescence, instead of C^* returning to the ground state by photon ejection, it can undergo energy transfer with a suitable fluorophore, which in turn may then exhibit its characteristic fluorescence emission (Equation 1.2).

$$A + B \rightarrow C^* + D \qquad \text{Equation 1.1}$$
$$C^* \rightarrow C + light$$

$$C^* + \text{fluorophore} \rightarrow [\text{fluorophore}]^* + products \qquad \text{Equation 1.2}$$
$$[\text{fluorophore}]^* \rightarrow [\text{fluorophore}]^* + light$$

Chemiluminescence detection is offers a simple, low cost and sensitive means to quantify a wide variety of compounds. Many chemiluminescence reagents have been used for detection in HPLC, including luminol, tris(2,2'-bipyridyl)ruthenium(III) and potassium permanganate [31].

A number of HPLC applications have been reviewed in the literature, these include pharmaceuticals [32], clinical analysis [33], polymers [34], food [35] and phenolic compounds [36].

1.1.1.3 Comparison of gas and liquid chromatography

When comparing gas and liquid chromatography it is necessary to first look at the analytes to be separated and their matrix. Next, the parameters that are most important to each technique can be examined. In GC, the two most important parameters are the nature of the stationary phase, which can be changed to adapt to the separation problem and the temperature, which is critical as GC is limited to volatile samples that are thermally stable. The two most important parameters in HPLC are the nature of the stationary phase and the nature of the mobile phase. The applicability of HPLC is wider than GC as both the stationary and mobile phases can be changed to adapt to the separation problem. However, HPLC can be restricted by insolubility [2, 26].

GC is typically used for the analysis of non-polar and semi-polar, volatile and semi-volatile chemicals. However, chemical derivatisation and pyrolysis can be performed on polar and non-volatile compounds, respectively, to permit their analysis by GC. HPLC is used for separating all types of organic chemicals independent of polarity or volatility [37].

Kivilopolo et al. [38] compared GC and LC methods for the analysis of phenolic acids in herb extracts. The results were compared in terms of their linearity, speed of analysis, selectivity, sensitivity and repeatability. Both methods proved suitable for the determination of phenolic acids. The sensitivity of both methods was good, but the linear range in LC was relatively small in comparison to GC. LC provided shorter analysis times and more straight-forward sample preparation than GC, which required liquid extraction, evaporation and derivatisation. On the other hand, the method development for GC was simpler than LC and GC provided better repeatability and easier identification of unknown species. LC proved a good choice for semi-quantitative analysis of phenolic acids, while GC was more useful for the accurate quantification of low molecular weight phenolic acids. The major drawback of GC in the analysis of phenolic acids, is the need for volatility, which limits the range of compounds that can be analysed.

1.1.1.4 One-dimensional chromatographic data

Chromatographic signal

The chromatographic data considered in this thesis is generated using a univariate detector, e.g. an FID, thus only data from such a detector will be discussed. The overall signal of a chromatogram can be considered as comprising three major components [39]:

1. The analytical signal, which contains the signal of any analyte present; it generally depends on the detector sensitivity and the capability of the chromatographic system.
2. The background signal, which is any signal that is not related to the analyte signal and shows some sort of systematic behaviour. The background often depends on the chromatographic conditions. The terms background and baseline are often referred to as the same phenomena, however background is more appropriate when talking about the contribution to the analytical signal and baseline is more often used in the literature when dealing with the correction of offset.
3. The noise, which is any unsystematic (random) variation in the signal; it essentially depends on the detector sensitivity.

The components of a chromatographic signal are shown in Figure 1.3.

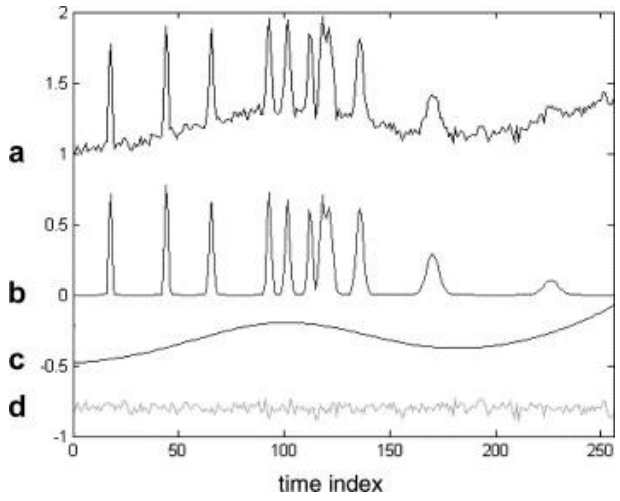


Figure 1.3: Components of a chromatographic signal: (a) overall signal, (b) relevant signal, (c) background, and (d) noise. From [40].

Structure of chromatographic data

The data obtained from a one-dimensional (1D) chromatographic experiment is in the form of a vector. A vector is a single ordered, time series, row or column of numbers written as:

$$\boldsymbol{x_j} = x_1, x_2 \ldots x_J \qquad \text{Equation 1.3}$$

Where $\mathbf{x_j}$ represents the specific, recorded detector response at time j in the range j = 1...J.

When performing a chromatographic study with more than one sample, the data is presented in the form of a matrix consisting of I rows (i = 1...I) of samples and J columns (j = 1...J) of variables and is written as:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1J} \\ \vdots & \ddots & \vdots \\ x_{I1} & \cdots & x_{IJ} \end{bmatrix} \qquad \text{Equation 1.4}$$

1.1.2 Comprehensive two-dimensional chromatography

The conditions defining a comprehensive two-dimensional separation were proposed by Schoenmakers et al. [41], based on the definitions formulated by Giddings [42], and can be summarised as [43]:

1. Every part of the sample is subjected to two different separations.
2. Equal percentages (100% or lower) of all sample components pass through both columns and eventually reach the detector.
3. The separation (resolution) obtained in the first dimension is essentially maintained.

1.1.2.1 Comprehensive two-dimensional gas chromatography

In comprehensive two-dimensional gas chromatography (GC×GC), two GC separations based on fundamentally different separation mechanisms are applied to the entire sample. An interface known as the modulator separates the first column

eluate into a large number of adjacent small fractions. Each individual fraction is then refocused and injected into the second GC column. The second column is often shorter and narrower than the first column, allowing fast separation in the second dimension to produce very narrow peaks. These narrow peaks require fast detectors in order to properly reconstruct the second dimension chromatograms. The outcome of a GC×GC run is a large series of high-speed, second-dimension chromatograms, which can be stacked side-by-side to form a two-dimensional chromatogram with one dimension representing the retention time on the first column and the other, the retention time on the second column. Visualisation is achieved through the use of contour plots, image plots and occasionally three-dimensional (3D) plots [44]. This process is depicted in Figure 1.4. GC×GC has been extensively reviewed in the literature [44-51]. These reviews discuss the principles of GC×GC, columns, modulation, detectors and applications.

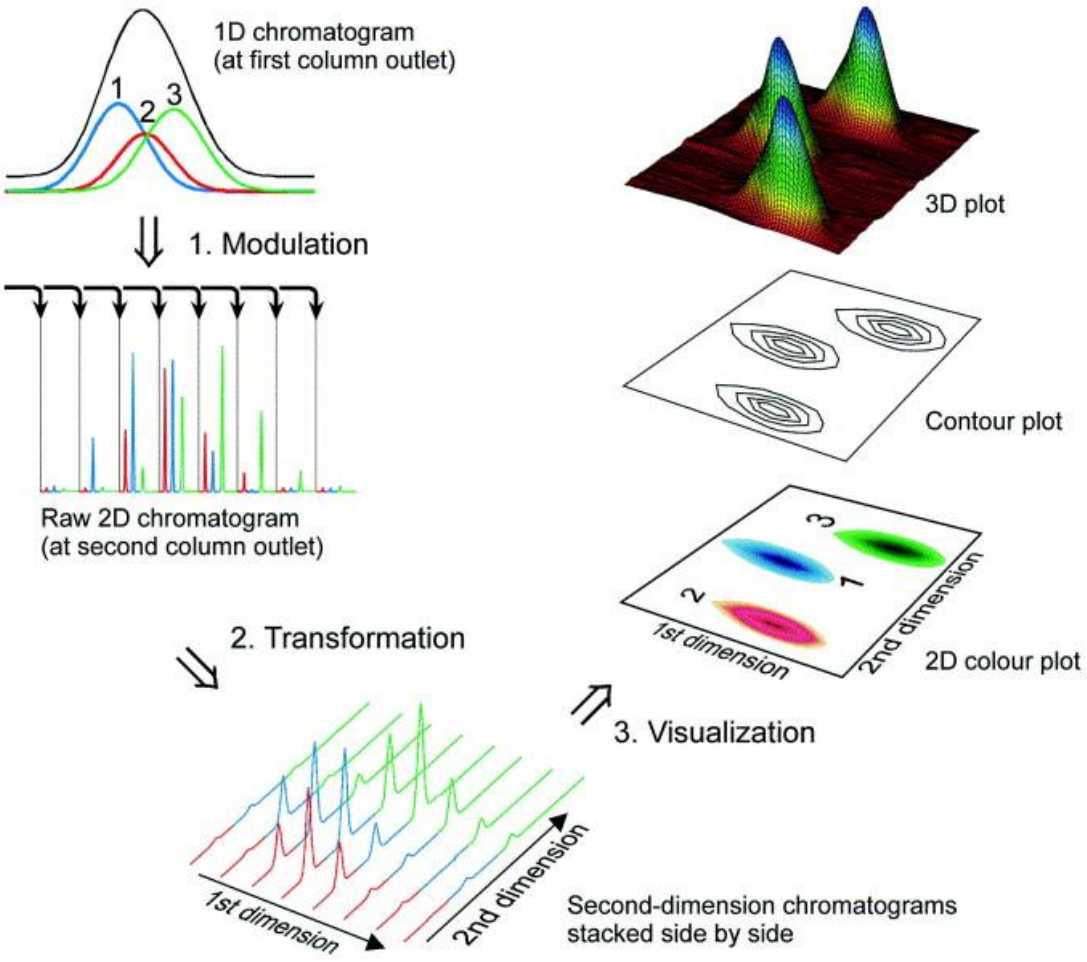


Figure 1.4: Generation and visualisation of a GC×GC chromatogram, from [44]

There is general agreement regarding the main advantages of GC×GC over conventional 1D GC in the literature [49, 52, 53]. Most notably, the peak capacity is much higher, which improves separation. Secondly, due to the refocusing process in the modulator, as well as the improved analyte separation, detectability is improved. Thirdly, chemically related compounds appear as ordered structures; this facilitates group-type analysis and the provisional classification of unknowns [44].

A typical GC×GC system consist of an injector, first dimension column, a modulator, second dimension column and a detector. A schematic of a GC×GC system is shown in Figure 1.5.

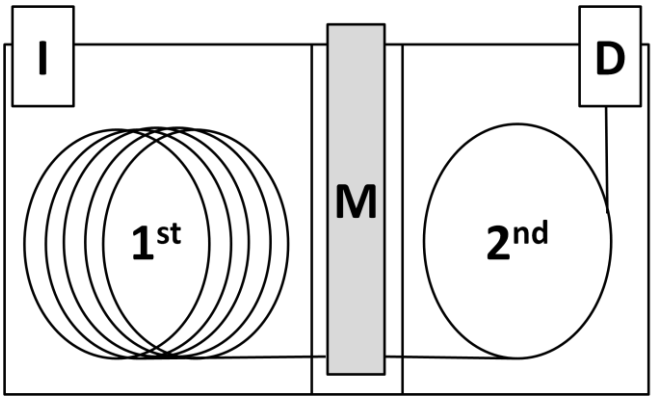


Figure 1.5: Schematic of a GC×GC system. I = injector, M = modulator, D = detector, 1st GC oven with first-dimension column and 2nd (separate) GC oven with the second-dimension column. Adapted from [44].

In GC×GC, two different and independent separation mechanisms are used in the two GC columns. In most cases, a non-polar stationary phase is used in the first dimension and a more polar stationary phase in the second dimension. Using a non-polar first dimension, analytes are separated according to boiling point and volatility. As the second dimension separation is fast, it is carried out under essentially isothermal conditions with little influence from sample volatility, so the separation is governed by the specific analyte interactions with the stationary phase [45, 47, 48].

The modulator can be considered the "heart" of a GC×GC system. It must serve three functions: (i) to continuously trap small adjacent fractions of the effluent from the first column whilst the first-dimension separation proceeds; (ii) to refocus the trapped

fractions either in time or space; (iii) to inject the refocused fractions as narrow pulses into the second-dimension column [44]. It is preferred that the separation in the second-dimension is finished before the injection of the next fraction or the second dimension retention time will exceed the modulation period, causing the second dimension peaks to appear in a later modulation than they were injected. This phenomenon is known as wrap-around [45, 46]. There are two main categories of commercially available modulators, thermal modulators and valve-based modulators. Thermal modulators are based on a temperature increase such as those using a "slotted heater" [54] or inversely, cryogenic modulation. Today, cryogenic modulation is used almost exclusively. The first cryogenic modulator was the longitudinally modulated cryogenic system [55, 56], which is based on a moving cold trap. Later, cryogenic jet modulators, with either carbon dioxide or liquid nitrogen, were developed [57-60]. These jet modulators have the advantage of no moving parts. There are fewer valve-based modulators employed in GC×GC, these include the diaphragm modulator [61] and the differential flow modulator [62]. A review of developments in GC×GC modulation is provided by Adahchour et al. [46].

A GC×GC detector must offer higher sampling rates, as the refocusing of peaks from the first column to the second, produces very narrow peaks. These narrow peaks require detectors with small internal volumes and fast acquisition rates in order to ensure a faithful representation of the chromatographic peak shape. Several detectors are suitable for GC×GC peak characterisation, including the FID which has been the detector of choice because of its small internal volume, fast slew rate, and corresponding high sampling rate. MS detectors are also important in GC×GC analyses as they provide structural information, which brings an additional dimension to the system. Combining GC×GC with time-of-flight mass spectrometry (TOF-MS) allows high separation power based on the combined use of chromatographic resolution and mass spectral resolution [63]. A review of detector technologies for GC×GC is provided by von Muhlen et al. [64].

1.1.2.2 Comprehensive two-dimensional liquid chromatography

The operating principles of comprehensive two-dimensional liquid chromatography (LC×LC) are similar to that of GC×GC. A typical LC×LC system consists of two

pumps, two columns, an injector, an interface (modulator) and a detector. An example of a typical LC×LC set-up is shown in Figure 1.6. The two columns are connected by the interface (usually a high pressure switching valve), which ensures the collection of the entire first dimension effluent in aliquots of predefined volumes and enables automatic re-injection of these fractions onto the second dimension column [65].

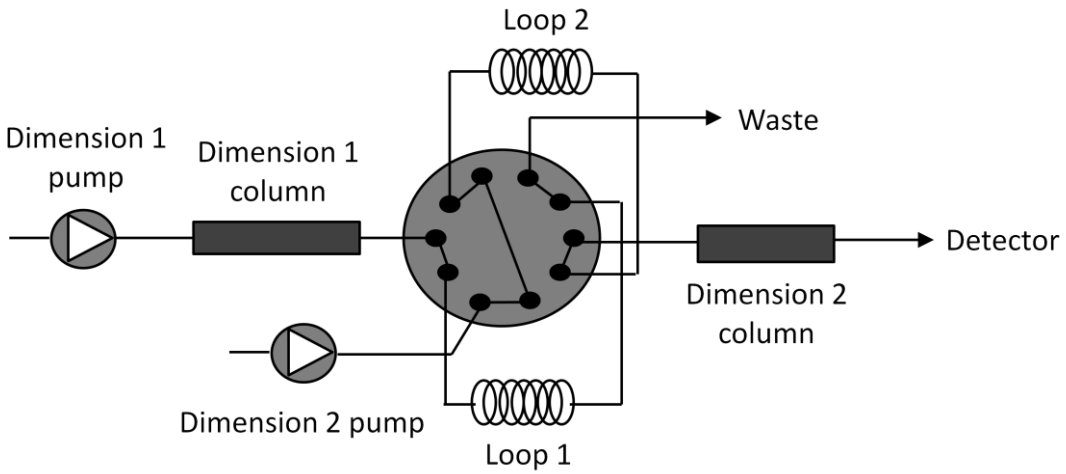


Figure 1.6: Schematic of an LC×LC system, adapted from [65]

A number of LC×LC reviews appear in the literature [43, 65-70]. These reviews discuss the principles of LC×LC, columns, interfaces, mobile phase compatibility, detection and applications.

1.1.2.3 Two-dimensional chromatographic data

The primary data generated by a GC×GC or LC×LC system is similar to the time response data generated in 1D chromatography. A two-column matrix is obtained, with the first column representing the time and the corresponding detector response signal in the second. Since the modulation period and the sampling frequency are known the raw data matrix can be converted and re-shaped to form a two-dimensional matrix (Equation 1.5).

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1J} \\ \vdots & \ddots & \vdots \\ x_{I1} & \cdots & x_{IJ} \end{bmatrix} \qquad \text{Equation 1.5}$$

Where i = 1 to I columns represent the second dimension chromatogram and j = 1 to J rows are the first dimension chromatogram.

The matrix resulting from this "data folding" can be visualised by appropriate software as a contour diagram.

1.1.2.4 Two-dimensional chromatographic applications

The separation provided by 1D chromatography can be significantly enhanced using 2D chromatography. 1D chromatography generally does not provide sufficient separation of complex mixtures, such as petroleum. Petroleum samples are complex as they contain a very large number of saturated and unsaturated alkanes, cyclic alkanes, aromatics and heteroatom-containing compounds. The number of compounds in petroleum samples increases exponentially with boiling point, and 1D GC can only fully separate constituents in the low boiling range, i.e. up to C_9 for straight-run hydrocarbon fractions, and even less for olefin-containing fractions [44].

Capillary 1D GC has routinely been used to analyse the volatile constituents of flavours and fragrances. However, the complex nature of these samples results in extended run times and extensive peak co-elutions which present a challenge for complete qualitative analysis [71, 72]. Shellie et al. [73] compared the separation of French lavender and tea tree essential oils by 1D GC and GC×GC. The 1D GC and GC×GC separations of tea tree oil are shown in Figure 1.7 (a and b), respectively. This study highlighted the increased separation of GC×GC and its ability to produce more baseline resolved peaks than traditional 1D GC. It was suggested that the peak capacity of GC×GC is up to 10 times higher than that of 1D GC and while the separation was no faster, within a similar analysis time GC×GC obtained higher sensitivity, greater resolution of peaks and a fingerprint pattern [73].

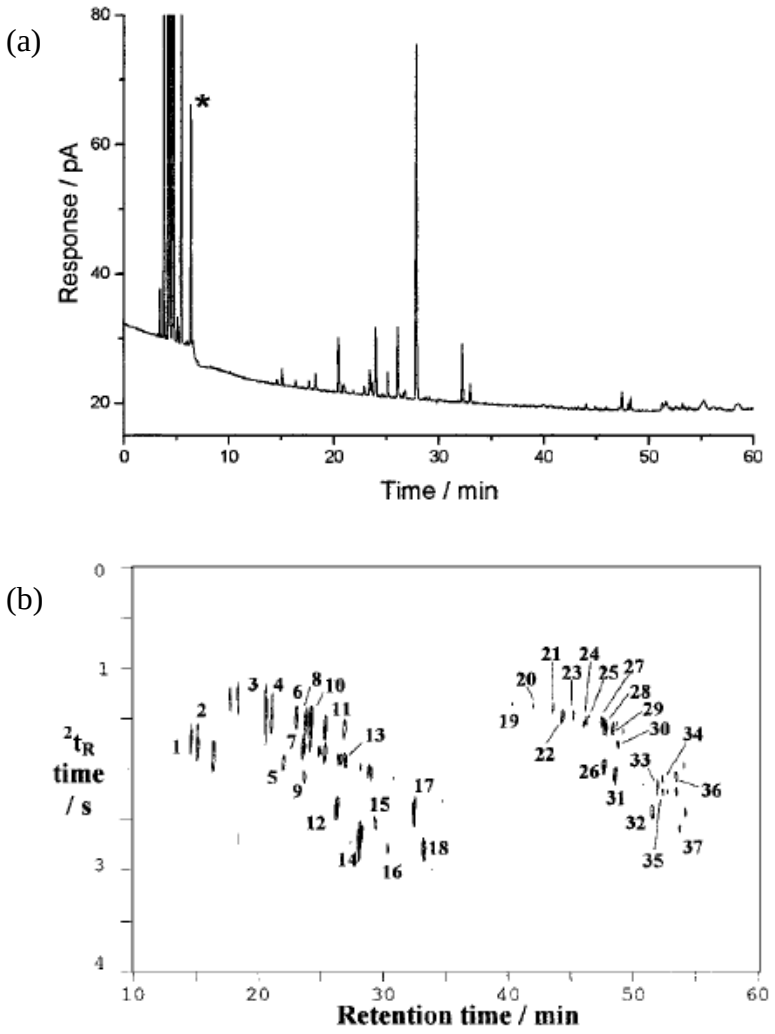


Figure 1.7: (a) 1D GC tea tree oil separation (b) GC×GC tea tree oil separation, from [73]. Peak assignments are provided in [73].

The application of GC×GC has been reviewed in the literature for the analysis of petroleum [74], drugs [75], flavours and fragrances [76], food and beverages [77], environmental monitoring [78] and metabolomics [79].

Since GC×GC has reached a higher degree of maturity than LC×LC, there are fewer applications of LC×LC in the literature; however some recent applications employing LC×LC in the areas of food and beverages, polymers and complex biological samples can be found in references [80-82].

1.2 Data pre-processing

Pre-processing of chromatographic data, to correct for artifacts in the signal, is important to enhance the quality of the individual chromatograms as well as improve the interpretation of results obtained using multivariate data analysis techniques. Standard pre-processing of chromatograms involves baseline correction to help eliminate artifacts such as column bleed as well as normalisation and scaling methods to reduce instrument and sample variations. Alignment of chromatographic data is also important as retention time shifts can occur due to variations in the flow rate and temperature, mobile phase composition and stationary phase decomposition. This can cause problems for multivariate data analysis as these techniques assume that compounds elute at the same time in all chromatograms. When this assumption is not met, variation modelled by the chemometric methods does not simply correspond to the chemical variations but rather a combination of chemical sources and retention time variations [83].

The application of chemometric tools to chromatographic data was traditionally performed using data that were processed to provide a list of detected, integrated peak areas or heights. However, in recent years there is an increasing interest in the direct chemometric interpretation of raw chromatographic signals. Integrated peak tables reduce the number of variables, remove baseline noise, and can remove the signal from irrelevant compounds if exact peaks are known. Problems can arise using integrated peaks since the analysis is restricted to identified compounds and as a result important information may be excluded. There are also many errors that can occur during the integration of raw signals due to poorly-resolved or missing peaks, which may skew the results and subsequent analysis. By applying chemometric tools directly to the raw data, all the information is preserved; however other issues become more important, most notably retention time shifts and the population of available variables. When comparing raw data, alignment is crucial to ensure that the peak for a given component is always registered in the exact same position in the data matrix. When using raw data the number of variables measured for each sample will outnumber the number of samples available in the data set. These overdetermined systems can defeat many chemometric techniques due, for example, to collinear variables. However, this may be overcome by the use of cross-validation (section 1.3.3) [84].

The choice of pre-processing methods not only depends on the type of data used (raw or processed) and the sample information required, but also on the subsequent data analysis method since different data analysis methods focus on different aspects of the data. For example, principal components analysis (section 1.3.1) attempts to explain as much variation as possible in as few components as possible, whereas a clustering method (section 1.3.2) focuses on the analysis of (dis)similarities. Changing data properties using data pre-treatment may therefore enhance the results a PCA analysis, while obscuring the results of a clustering method [85].

A number of pre-processing strategies for chromatographic data appear in the literature [86-90]. These strategies are employed prior to multivariate data analysis and typically involve baseline correction, alignment, normalisation and scaling.

1.2.1 Baseline correction

Baseline drift is mainly caused by continuous variations in experimental conditions, such as temperature, solvent programming in liquid chromatography, or temperature programming in gas chromatography. This makes baseline correction a common requirement in chromatographic studies and of great importance for peak detection and comparison of different chromatographic signals [91]. Baseline correction methods are commonly employed to eliminate interferences due to drift, column bleed and overlap of broad or poorly defined peaks [40].

Baseline correction methods for chromatographic data frequently involve fitting a low order polynomial curve and subtracting it from the overall signal. These methods are based on the construction of either a local or a global baseline, an example of this is shown in Figure 1.8. A global fit using a second order polynomial is shown in (a) and an accurate correction is not achieved across the whole chromatogram. When using a local method with several second order polynomials (b), the curve is not smooth, but a more accurate fit of the baseline is obtained. Local methods work well if it is possible to find points in the baseline where there are no peaks, if peaks are not coeluting, and if the signal-to-noise ratio is high. When this is not the case, the baseline may be better described using a global polynomial fit with higher order polynomials to account for a more complex or curved baseline [39].

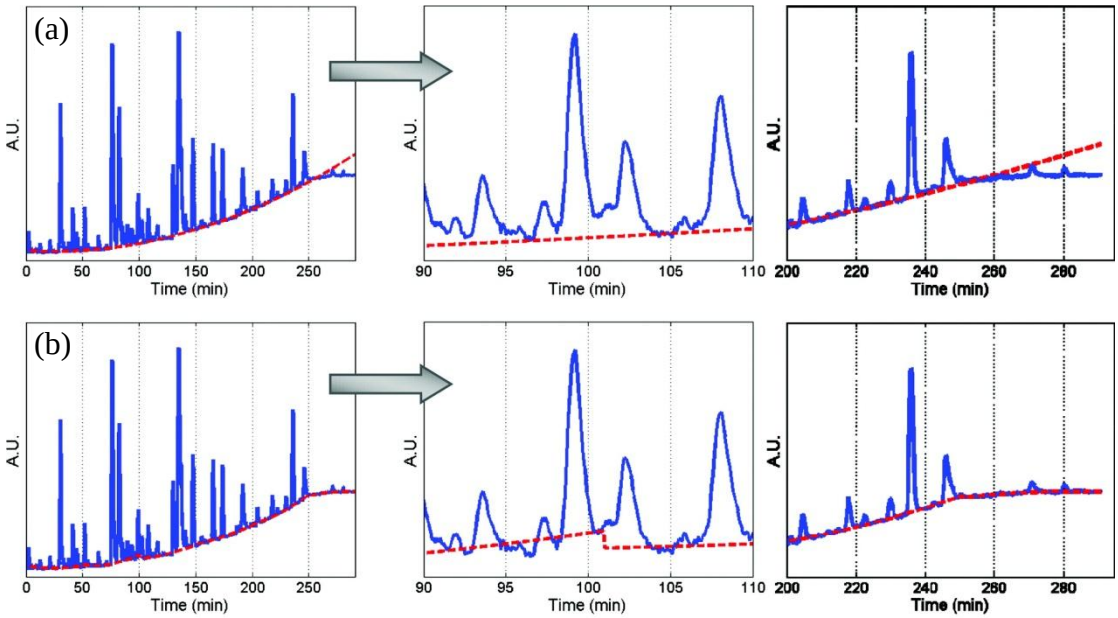


Figure 1.8: Curve-fitting baseline methods (a) global fit using a second order polynomial and (b) local fit using several second order polynomials. From [39].

A global baseline correction technique using asymmetric least squares smoothing was proposed by Eilers [92]. This method works by fitting an initial polynomial of a specified order to all data points in the chromatogram. By iteratively weighting positive deviations from the polynomial more than negative deviations, the polynomial will at some point approximate the baseline (within a predefined limit), which is then subtracted from the original signal [93]. This process is illustrated in Figure 1.9. After some modifications, this approach has been extended to 2D data for use in 2D gel electrophoresis [94, 95].

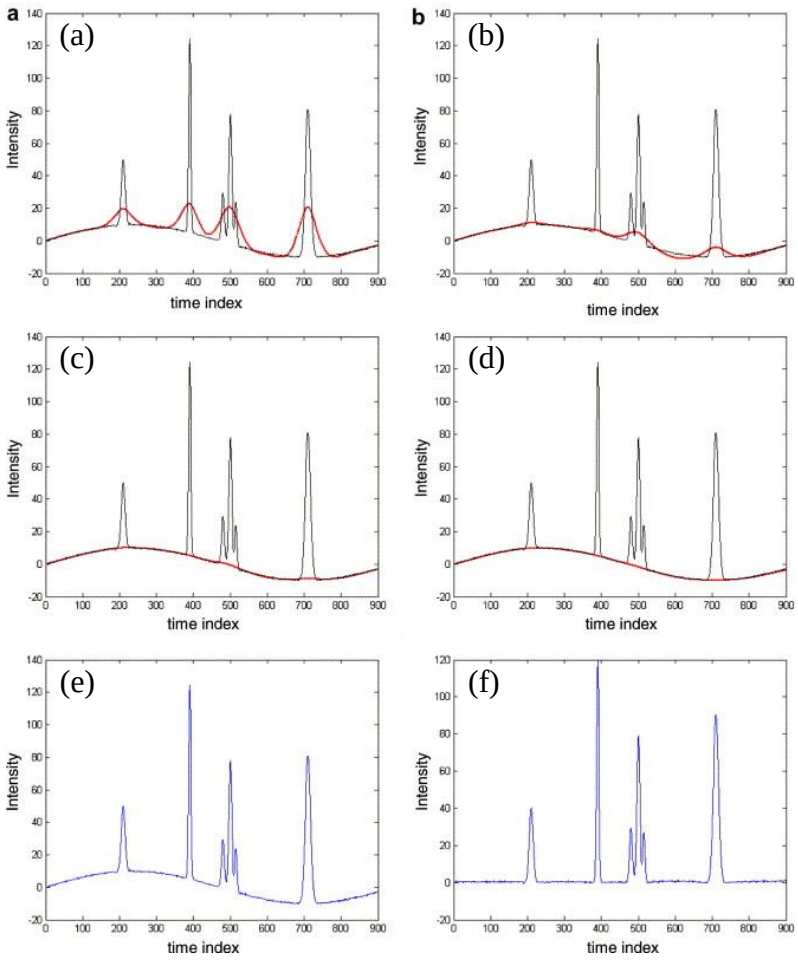


Figure 1.9: Global baseline correction using asymmetric least squares (a-d) consecutive baseline estimates, (e) original signal and (f) baseline corrected signal. From [40].

Gan et al. [96] also fitted a polynomial to estimate the baseline of a chemical signal. This process is shown in Figure 1.10; an initial polynomial is fitted, which is set as the automatic threshold, and parts of the signal that are above this threshold are cut out. The new signal replaces the original one for use in the next iteration and this iterative process continues until there is no change in the modelled baseline.

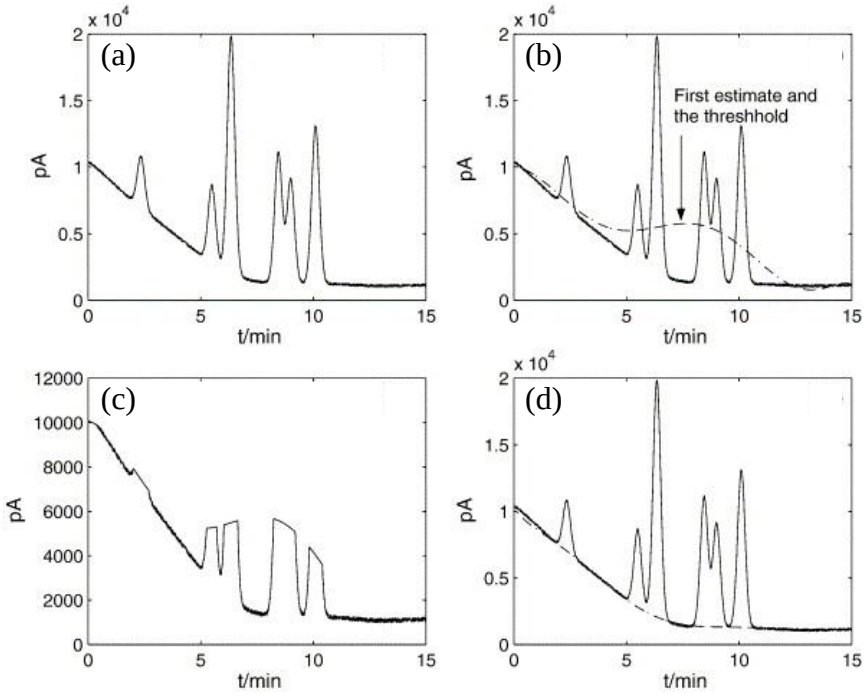


Figure 1.10: Fitted polynomial baseline estimate (a) simulated signal, (b) initial polynomial fit (threshold), (c) parts of signal above threshold removed and (d) final estimated baseline. From [96].

Numerical differentiation, usually first or second order, provides a straightforward method to remove baselines in chromatographic data. This method relies on the assumption that the slope of the baseline is relatively small compared to that of a chromatographic peak. Since numerical differentiation leads to a considerable decrease in the signal-to-noise ratio, the Savitzky-Golay algorithm [97] can be used to obtain the derivatives and smooth the data [86, 88, 89].

1.2.2 Alignment

Variation in the elution time of an analyte is frequently observed in chromatographic analyses; this can obscure chemical variations and makes alignment of chromatographic signals necessary [90, 98-102]. Retention time variations can be due to subtle, random, and often unavoidable changes in instrument parameters. Pressure, temperature and flow rate fluctuations may cause an analyte to elute at a different retention time in replicate runs. Matrix effects and stationary phase decomposition may also cause retention time shifting [103]. Alignment is particularly important when multivariate chemometric analysis techniques are to be performed as these

techniques are generally employed to interpret chemical variation as changes from sample-to-sample at corresponding variables (i.e. retention times) [104]. Therefore in order to effectively interpret the between-sample differences, retention time variations need to be eliminated or reduced as the chemometric model may describe retention time shifts and conceal significant information regarding the samples [105].

For an alignment method to be successful, it must synchronise the signals and preserve the peak information (i.e. peak area and shape). The more flexible the alignment method, the more adjustment is possible, but there is increased risk of introducing artifacts. On the other hand, less flexible methods imply only smaller peak shifts can be corrected but with reduced risk of changing the chromatographic profile. Using a method with low flexibility that is able to correct for shifts present in the data is the goal for all alignment procedures [39, 104].

One issue for alignment is the selection of an appropriate reference chromatogram. Ideally the reference chromatogram should be representative, contain as many common peaks as possible, be reproducible, clean (contain no artifacts) and a real chromatogram (not generated). A number of reference chromatogram selections have been proposed in the literature, including using the first chromatogram [90, 92], the middle chromatogram [89, 106, 107], the chromatogram containing the highest number of common constituents [99, 101, 108], the chromatogram with the highest correlation coefficient to all the other chromatograms [100, 109], or a generated chromatogram such as the mean [110]. Daszykowski et al. [83] conducted a study on the use of different reference chromatograms and found the chromatogram with the highest correlation coefficient gave the most satisfactory results.

Alignment of data may only require a linear shift of the whole chromatogram, i.e. a linear translation of the data vector. However, if the column is changed between runs or if samples are measured over a long period of time, non-linear shifts may occur. These shifts are characterised by a different degree of shifting for multiple peaks across samples and can be seen as peaks shifting independently from one another in the same chromatogram. Thus, more complex shift correction may be needed [87].

The most commonly employed method for aligning chromatographic signals is correlation optimised warping (COW) [88, 89, 106, 111-115]. COW was originally developed by Nielsen et al. [116] and is based on aligning a sample chromatogram to

a reference by piecewise linear stretching and compression. The algorithm requires two user input parameters, the segment length and slack size, which are selected in order to optimise the overall correlation between the reference and sample. The reference and sample chromatograms are divided into segments of user defined segment length. Each segment in the sample is stretched or compressed by shifting the position of its end point by the slack value. The resulting segment is then linearly interpolated to the corresponding segment in the reference. For each possible end point of the segment, the correlation coefficient between the interpolated segment and the corresponding reference segments is computed. This is performed on all segments. A warping solution is then constructed as a cumulative sum of the correlation coefficients of the previous segments. After examining all possible end points of all segments, the optimal warping path is constructed [99, 100]. A 2D COW algorithm that extends the 1D COW algorithm for 2D chromatographic profiles was proposed by Zhang et al. for aligning GC×GC TOF-MS data [117].

Dynamic time warping (DTW) has also been employed for the alignment of chromatographic data [101, 102]. DTW nonlinearly warps two signals in such a way that similar events are aligned and a minimum distance between them is obtained. The algorithm uses dynamic programming to find the optimal path of warping and employs the squared Euclidean distance metric as the optimisation criterion. Dynamic programming solves combinatorial optimisation problems in order to find the optimal warping path by examining all the possible combinations of data points on the time axis [118].

Piecewise alignment (PWA) is related to COW in that it divides the sample and reference chromatograms into windows of a user-specified length, but instead of applying stretching and shrinking interpolation just prior to calculating the correlation, each window in the sample chromatogram is iteratively shifted, point-by-point, within a specified limit along the retention time axis, thus saving computation time. The Pearson correlation coefficient between the sample and reference is calculated at each shift and the shift that gives the maximum correlation coefficient is used to correct that window of the sample chromatogram. The desired retention time corrections are assigned to the centre point of the windows and the shifts to be applied in the regions between the window centres are calculated by linear interpolation [103, 110, 119]. Pierce et al. [120] further modified this algorithm to develop a comprehensive 2D

retention time alignment algorithm which aims to correct the entire chromatogram and preserve the separation information in both dimensions.

Instead of dividing the chromatograms into segments that can subsequently be stretched or compressed, parametric time warping (PTW) [92] calculates a global second order polynomial (the warping function) by minimising the sum of the squares between the reference and sample chromatograms. The sample chromatogram is then interpolated to the points in the warping function to obtain a chromatogram that is aligned to the reference chromatogram [93].

A recently developed nuclear magnetic resonance (NMR) alignment method, *i*coshift [121], has also been applied to the alignment of chromatographic data [122-124]. The *i*coshift algorithm divides the chromatograms into segments and aligns these to the corresponding segments in the reference chromatogram. Each chromatogram is independently aligned to the reference by shifting the segments sideways and maximising the cross-correlation between the segments. The *i*coshift algorithm requires two user input parameters, the segment length and maximum shift. The algorithm uses a fast Fourier transform engine to boost the simultaneous alignment of all chromatograms in the data set. Interpolation is avoided by filling in the missing parts on the edges of the segments with either missing values or by repeating the value of the boundary point [121, 125]. *i*coshift differs from COW and PWA as there is no interpolation step; it can also be further differentiated from COW as alignment is achieved through sideways shifting rather than stretching and shrinking.

A summary of the different alignment techniques is given in Table 1.3.

Table 1.3: Summary of alignment techniques

Method	**Input parameters**	**Optimisation criteria**	**Aligns by**	**Advantages**	**Disadvantages**	**Applications**
COW	Reference, segment length, slack size	Correlation coefficient	Stretching/shrinking	Correct complex shifts. Option to automatically optimise parameters. Provides good and robust alignment.	Small segments and large slack can change peak shape. Time consuming due to many interpolation steps.	Tested on many different types of shifted data. Able to correct most shifts. [98, 99-102, 118]
DTW	Reference, local continuity constraints, band constraints	Squared Euclidean distance	Elementary transitions	Correct complex shifts	Distance not best measure of similarity. Can be too flexible. Time consuming.	Tested on simple and severely shifted data. Too flexible for simple shifts. [101, 102, 118]
PWA	Reference, segment length, sideways shifting	Correlation coefficient	Shifting	Correct complex shifts. Peak shape preserving.	Only linear shifts are permitted within segments. Potential problem if peaks split between segments.	Tested on severely shifted data, with peaks shifted across neighbour peaks [110]
PTW	Reference, warping function coefficients	Sum of squared residuals	Stretching/shrinking	Fast and simple. No risk of peak splitting or artifacts as a global warping function is used.	Low flexibility due to the quadratic warping function. Only correct non-complex shifts.	Tested for many types of shifts, but really only suitable for small systematic shifts [98, 118]

Method (cont.)	**Input parameters (cont.)**	**Optimisation criteria (cont.)**	**Aligns by (cont.)**	**Advantages (cont.)**	**Disadvantages (cont.)**	**Applications (cont.)**
*i*coshift	Reference, interval length, shift	Correlation coefficient	Shifting	Fast and simple. Provides good alignment.	Peaks can be split between segments. Visual inspection required to identify presence of artifacts.	Tested on a broad range of shifts [121, 125]

1.2.3 Normalisation and scaling

The recorded signal intensity is affected by instrumental variations and chromatographic conditions as well as sampling and sample preparation. Variations in the data due to these factors can mask the compositional information and distort any subsequent data analysis. Hence, a normalisation step is a prerequisite for effective comparison of chromatographic data [88].

Normalisation is performed on the rows of the 1D data matrix and comprises methods to make the data from all samples directly comparable with each other. One common method involves normalising each chromatogram to have unit total detector response; this is referred to as normalisation to a constant sum [126]. This type of normalisation is useful for characterising different groups of samples, but not for establishing a calibration model as some quantitative information may be lost. For quantitative analysis, it is often recommended to normalise according to an internal standard peak as each peak still contains a quantitative measure of the analyte concentration. This relies on the assumption that all peaks behave in a similar manner and can be corrected by the same internal standard. If this is not the case, multiple internal standards may need to be used [39].

In multivariate data analysis and when dealing with models that focus on variability in data, it is usual practice to scale the data. Scaling is performed on the columns of the 1D data matrix (i.e. on each chromatographic intensity across all samples). A review of the various scaling methods can be found in an article by van den Berg et al. [85]. The review describes how each of the scaling methods emphasise different aspects of the data and notes the advantages and disadvantages of each method. A summary of the methods described in this review are provided in Table 1.4.

Although normalisation and scaling operations serve different purposes, it is usual practice to use both to aid in comparing chromatograms and prior to performing multivariate data analysis [126].

Table 1.4: Overview of column scaling pre-treatments (from [85]). The mean is estimated as: $\bar{x}_i = \frac{1}{J}\sum_{j=1}^{J} x_{ij}$ and the standard deviation is estimated as $s_i = \sqrt{\frac{\sum_{j=1}^{J}(x_{ij} - \bar{x}_i)^2}{J-1}}$. $\tilde{x}$ and $\hat{x}$ represents the data after different pre-treatment steps.

Method	**Formula**	**Goal**	**Advantages**	**Disadvantages**
Centering	$\tilde{x}_{ij} = x_{ij} - \bar{x}_i$	Focus on the differences and not the similarities in the data	Remove the offset from the data	When the data is heteroscedastic, the effect of this pre-treatment method is not always sufficient
Autoscaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$	Compare variables based on correlations	All variables become equally important	Inflation of the measurement errors
Range scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\left(x_{i_{max}} - x_{i_{min}}\right)}$	Compare variables relative to the response range	All variables become equally important. Scaling is related to response.	Inflation of the measurement errors and sensitive to outliers
Pareto scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$	Reduce the relative importance of large values, but keep data structure particularly intact	Stays closer to the original measurement than autoscaling	Sensitive to large fold changes
Vast scaling	$\tilde{x}_{ij} = \frac{\left(x_{ij} - \bar{x}_i\right)}{s_i} \cdot \frac{\bar{x}_i}{s_i}$	Focus on variables that show small fluctuations	Aims for robustness, can use prior group knowledge	Not suited for large induced variation without group structure
Level scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\bar{x}_i}$	Focus on relative response	Suited for identification	Inflation of the measurement errors

Method (cont.)	**Formula (cont.)**	**Goal (cont.)**	**Advantages (cont.)**	**Disadvantages (cont.)**
Log transformation	$\tilde{x}_{ij} = log_{10}(x_{ij})$ $\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	Correct for heterscedasticity, pseudo scaling. Make multiplicative models additive.	Reduce heterscedasticity, multiplicative effects become additive	Difficulties with values with large relative standard deviation and zeros
Power transformation	$\tilde{x}_{ij} = \sqrt{(x_{ij})}$ $\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	Correct for heterscedasticity, pseudo scaling	Reduce heterscedasticity, no problems with small values	Choice of square root is arbitrary

1.2.4 Smoothing

Smoothing is applied to a chromatographic signal in order to increase the signal-to-noise ratio for the peaks of interest.

Noise in chemical signals is generally defined as the instantaneously irreproducible signals caused by interfering physical or chemical processes, imperfections in the experimental apparatus, and other irregularities, which often complicate the results. A number of methods such as moving average and the Savitzky-Golay smoothing method have been developed to compensate for these irregularities [91].

Moving average is a simple smoothing method, which involves averaging an odd number of points and assigning this value to the central point. The window then shifts by one point and the procedure is repeated, thus preserving data density. Another similar method is box-car averaging. Box-car averaging involves dividing the data into a series discrete, equally spaced bands and replacing each band by a centriod average value. It is then shifted the entire box-car length along the vector, thereby decreasing data density [127].

The most widely used smoothing technique in analytical science is the polynomial filter suggested by Savitzky and Golay [97]. The principle of the Savitzky-Golay smoother is simple; a low-order polynomial is fitted to a window in the chromatogram and the fitted value in the middle of the window is used as the smoothed data. The window is then shifted to the right and this procedure is repeated until all desired smoothed values have been computed [128]. The Savitzky-Golay smoother requires two input parameters; the window width and the order of the polynomial. Vivo-Truyols and Schoenmakers [129] developed an algorithm for selecting the window size. The method is based on a comparison of the fitting residuals (i.e. differences between the input signal and the smoothed signal) with the noise of the instrument; the window size that yields the autocorrelation of the residuals closest to the autocorrelation of the noise of the instrument is considered optimal.

A smoother based on penalized least squares, which extends a method proposed by Whittaker [130] has been develpoed by Eilers [128] as an alternative to the Savitzky-Golay smoother. The smoother is reported to be extremely fast, provide continuous control over smoothness and interpolate automatically.

1.3 Data analysis

1.3.1 Principal components analysis

Principal components analysis (PCA) is the most widely used multivariate technique for exploratory analysis in chromatography [131-140]. A set of correlated variables are transformed into a set of uncorrelated latent variables (principal components) such that the first few components explain most of the variation in the data.

PCA involves rotating and transforming the original axes, each representing an original variable, into new axes (so-called latent variables). This transformation is performed in such a way that the new axes lie along the direction of maximum variance of the data with the constraint that the new axes are orthogonal (uncorrelated). PCA can reveal the variables, or combinations of variables, that describe some inherent structure in the data. The first principal component (PC) is the linear combination with the largest variance that best summarise the distribution of the data. The second PC is uncorrelated with the first and accounts for the largest remaining variance; this process is continued until the total variance is accounted for. Each PC is characterised by two pieces of information, the scores, and the loadings. Loadings are the coefficients of the linear combinations of the original variables and scores are the coordinates of the original data in the new coordinate system [127].

PCA is a mathematical transformation of the original data matrix, which takes the form:

$$\boldsymbol{X} = \boldsymbol{T} \cdot \boldsymbol{P} + \boldsymbol{E} \qquad \text{Equation 1.6}$$

Where $\boldsymbol{X}$ is the original data matrix of I rows and J columns; $\boldsymbol{T}$ is the scores matrix of I rows and A columns; $\boldsymbol{P}$ is the loadings matrix of A rows and J columns; $\boldsymbol{E}$ is an error matrix. This is shown graphically in Figure 1.11.

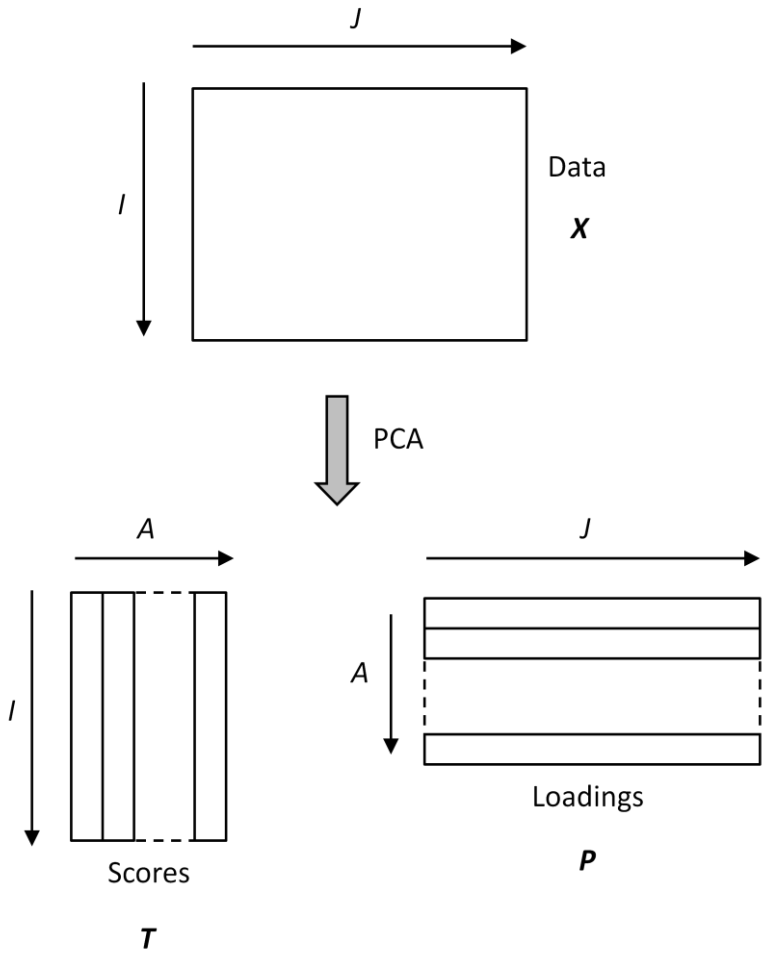


Figure 1.11: Principles of PCA, adapted from [141]

The variance explained by each PC is referred to as an eigenvalue. The earlier (and more significant) the components, the larger their variance.

Since a large fraction of the variance is usually described by few PCs, the data can be visualised by plotting the scores and loadings against each other [142]. Plotting scores against each other can help elucidate relationships between samples and how they are connected and grouped. Loading vectors plotted as chromatograms can show the components where variation was observed and thereby which variables affected the relationships in the score vectors. Plotting loading vectors against each other also reveals information about the connection between variables. The combination of score and loading plots, in the form of a biplot, will often reveal which samples are connected to which variables. However, this is more complicated when using raw chromatographic data [143].

In chromatography, visualisation of scores and loadings has recently been employed to discriminate food and beverages [144, 145], analyse polycyclic aromatic hydrocarbons in environmental samples [146, 147], classify steroid metabolites [148] and differentiate gasoline for forensic purposes [149, 150]. PCA has also been employed to identify outliers in chromatographic data [151].

PCA and visual representation of the scores plots associated with the first few components is often the end to multivariate data analysis, however the scores can be used as inputs to other multivariate techniques. For example, scores may be used for both unsupervised and supervised pattern recognition techniques that require the number of samples be larger than the number of variables or when variables are highly correlated [152]. PCA scores have been used as inputs to discriminant analysis for classifying wines [153-155].

Data evaluation in chromatography by PCA was reviewed by Cserhati [156]. The review included the application of PCA to gas and liquid chromatographic data for health care, food and environmental applications, as well as miscellaneous applications such as the selection of chromatographic columns.

1.3.2 Unsupervised pattern recognition

In the case of unsupervised pattern recognition, often referred to as cluster analysis (CA), no class knowledge is required and no assumptions need be made regarding the class to which a sample may belong.

CA involves converting the data into some corresponding set of similarity, or dissimilarity, measures between each sample with the subsequent aim of dividing a set of objects into several groups or clusters so that objects within the same group are more similar to each other than objects in different groups. CA has been extensively discussed in a tutorial by Bratchell [157]. The tutorial discusses distance and similarity measures, hierarchical clustering techniques, such as nearest neighbour, furthest neighbour or average linkage, which are used to link objects as well as visualisation and interpretation of results using a dendrogram.

CA is often used as the first step in chemometric data analysis as it can provide a simple visual representation of the results in the form of a dendrogram, this aids in identifying underlying patterns and structure in the data. In chromatography, CA has recently been employed in medical analyses to differentiate cancer patients and healthy controls [158, 159] as well as distinguishing between Alzheimer's disease patients and healthy volunteers [160]. CA has also been used to classify chromatographic columns [161, 162].

1.3.3 Supervised pattern recognition

Supervised pattern recognition techniques use known information about the class membership of samples to a certain group (or class) in order to classify new, unknown, samples to one of the known classes based on measurement patterns. There are several types of supervised pattern recognition methods which essentially differ in the way they achieve classification. There are those focused on discriminating among classes, such as linear discriminant analysis (LDA) and *k*-nearest neighbours (*k*-NN) and those oriented towards modelling classes, such as soft independent modelling of class analogy (SIMCA) [163].

Whatever the method used for classification, supervised pattern recognition techniques essentially consist of the following steps [163, 164]:

1. Selection of a training set, which consists of objects of known class membership for which variables are measured. The training set is used for the optimisation of parameters characteristic of each technique.
2. Variable selection. The variables that are meaningful for the classification are kept, while variables that are noisy or that have no discriminating power are eliminated.
3. Building a model using the training set. A mathematical model is derived between the variables measured on the training set and their known classes.
4. Validation of the model using an independent test set or cross-validation, in order to evaluate the reliability of the classification achieved.

Validation is one of the most important aspects of supervised pattern recognition. Model validation evaluates the number of significant variables needed to characterise

the data set, the model prediction ability for unknown samples and the representative character of the data used to produce the model. Model validation ensures that the supervised pattern recognition technique is good enough to perform the classification of unknown samples. This is achieved by examining how successful the model is at classifying unknown objects, i.e. by evaluating the recognition and prediction abilities of the model. The recognition ability is defined as the percentage of the samples in the training set correctly classified during the modelling step. The prediction ability is the percentage of the samples in the test set correctly classified by using the models developed in the training step. In cross-validation, the prediction ability of the model is determined by developing a model using the training set and using the test set to test the model. Both training and test sets contain samples representative of each class. The model development and testing is repeated several times increase the probability that a sample will be used in the training and test sets. This can be done using K-fold cross-validation or leave-one-out cross validation [163].

Alonso-Salces et al. [165] compared LDA, *k*-NN and SIMCA for the classification of Galician and French ciders. 100% recognition ability for both classes was achieved using LDA, while 100% prediction ability was achieved using *k*-NN. SIMCA was able to achieve 100% recognition and prediction abilities, but only for the Galician ciders. The results of the supervised pattern recognition techniques were found to be complementary.

Supervised pattern recognition in food analysis has been reviewed by Berrueta et al. [163]. The review discusses the various supervised pattern recognition techniques, including LDA, partial least squares discriminant analysis (PLS-DA), *k*-NN, SIMCA and artificial neural networks (ANN). The review also discusses validation models, such as K-fold and leave-one-out cross-validation and provides a detailed literature review of supervised pattern recognition techniques to classify wines, edible oils, honey, dairy foods, meat and fruit.

1.3.3.1 *k*-nearest neighbours

k-NN methods are simple to understand and implement. In *k*-NN, an unknown sample is classified according to the majority vote of its *k*-nearest neighbours, where *k* is an odd number. The *k*-NN method requires three steps [166, 167]:

1. Calculating the distance between an unknown object and all the members in the training set. The most common distance metric is Euclidean distance.
2. Selecting the *k* nearest neighbours to the unknown.
3. Predicting class membership of the unknown by applying the majority rule to the *k* nearest samples.

In chromatography, *k*-NN along with LDA and support vector machines was recently employed discriminate HPLC fingerprints of raw and processed rhubarb samples [168]. All methods provided satisfactory results, however the best classification was achieved using *k*-NN.

Despite being a simple method that provides good classification results, *k*-NN is less frequently employed than LDA for supervised pattern recognition of chromatographic data.

1.3.3.2 Discriminant analysis

Discriminant analysis is one of the most powerful and commonly used supervised pattern recognition techniques. The object of discriminant analysis is to fit a line or surface through a set of data points that provides a maximum discrimination between groups present in the data [127]. Objects lying on the same side of the line are considered as belonging to the same group.

LDA is one variant of discriminant analysis, in which the discrimination boundaries are linear. LDA is based on the determination of a linear discriminant function, which maximises the ratio of between-class variance and minimise the ratio of within-class variance. LDA selects a direction that achieves maximum separation among the given classes. LDA requires that the variance-covariance matrices of the established classes can be pooled, which is only possible when these matrices can be considered equal. A quadratic discriminant function (QDA) is another function used for discrimination. QDA establishes parabolic boundaries and is subject to fewer constraints in the distribution of the objects in space than LDA. In QDA the distance to each class is calculated using the sample variance-covariance matrix of each class rather than the overall pooled matrix. Both LDA and QDA require that the number of samples is greater than that of the variables [163, 169].

The results of discriminant analysis can be displayed in the form of a contingency table, referred to as a confusion matrix, of known parent groups against classified groups [127]. An example confusion matrix is shown in Table 1.5.

Table 1.5: Example confusion matrix. X_x is the number of objects in X correctly classified as X, X_y is the number of objects in Y incorrectly classified as X, Y_x is the number of objects in X incorrectly classified as Y and Y_y is the number of objects in Y correctly classified as Y. T_{AX} and T_{AY} are the total number of objects actually in X and Y, respectively. T_{PX} and T_{PY} are the total number of objects predicted in X and Y, respectively.

	Predicted X	**Predicted Y**	**Total**
Actual X	X_x	Y_x	T_{AX}
Actual Y	X_y	Y_y	T_{AY}
Total	T_{PX}	T_{PY}	

In chromatography, LDA is more frequently employed for classification than QDA [170-177]. It is most commonly employed for classification of food and beverages according origin and variety.

LDA and QDA, among other methods, were compared by Dixon and Brereton [169] for the classification of six synthetic data sets, with varying degrees of distribution between two classes. The LDA and QDA models were calculated using the first two PCs. For the classification of a training set, QDA achieved a greater number of correct classifications in three of the data sets; LDA achieved greater classification in just one of the data sets and on two occasions LDA was more accurate at classifying class 1, while QDA was more accurate at classifying class 2.

1.3.3.3 SIMCA

A SIMCA model consists of a collection of PCA models, one for each class in the data set. This is shown graphically in Figure 1.12. Each class can have a different number of PCs, which are defined by the user or cross-validation. This ensures a sufficient number of PCs are retained to account for most of the variation within each class, while ensuring a high signal-to-noise ratio by not including noisy PCs in the class model. SIMCA determines the class distance as well as the modelling and discriminatory powers. The class distance can be calculated as the geometric distance from the PC models or by using a confidence level (usually 95%), which assumes each class is bound by a region of space. The discriminatory power measures how well a variable discriminates between two classes. This differs from the modelling

power in the sense that a variable may able to model one class well, but this does not necessarily imply that it is able to discriminate two groups effectively [163].

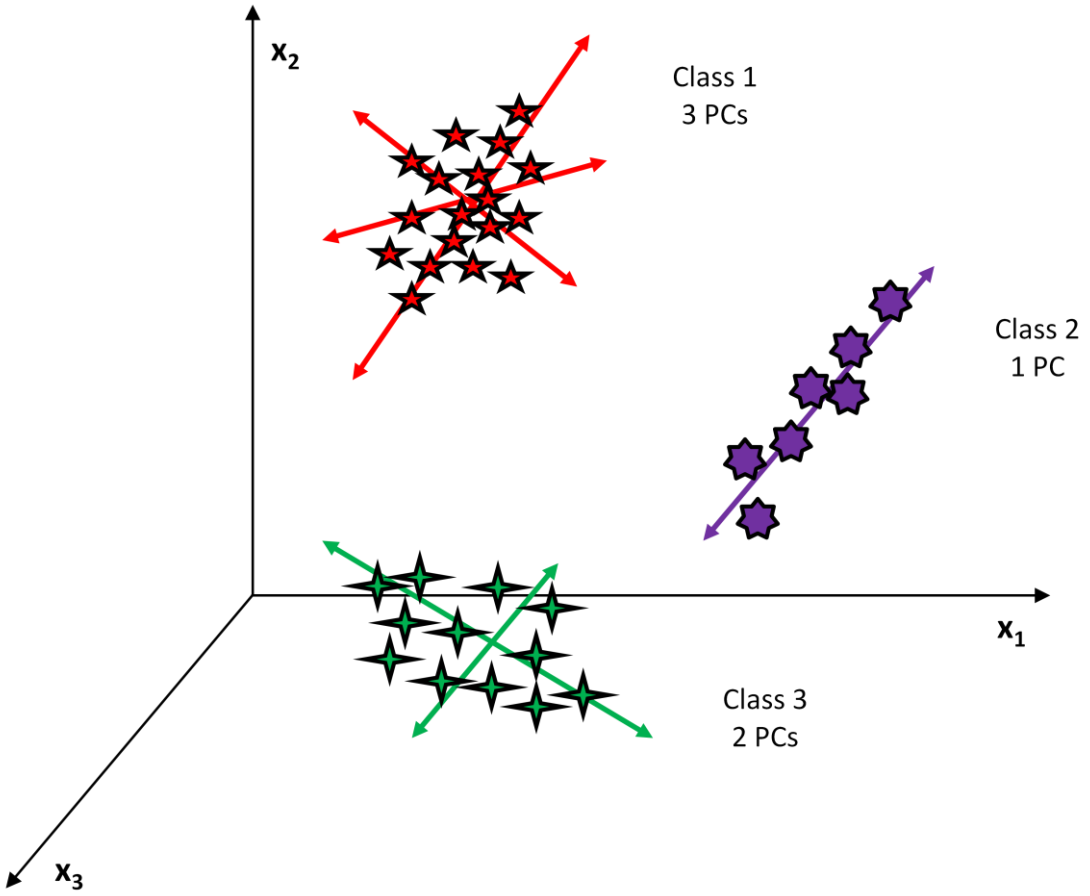


Figure 1.12: Graphical representation of a SIMCA model, adapted from [178]

Recently, SIMCA has been employed in chromatographic analysis to identify extra virgin olive oil adulteration [179], screen the quality of commercial Brazilian gasoline [180], discriminate herbal extracts [181] and evaluate the quality control of herbal medicines [182]. However, despite providing good classification results, SIMCA is still less frequently employed than LDA for supervised pattern recognition in chromatography.

1.3.4 Regression analysis

Regression analysis is a class of techniques that study how measured independent or response variables vary as a function of a single so-called dependent variable. The principal aim in undertaking regression analysis is to develop a suitable mathematical model for descriptive or predictive purposes. The model can be used to confirm some idea or theory regarding the relationship between variables or it can be used to predict some general, continuous response function [127].

Multivariate regression techniques were reviewed by Brereton [183]. The article discusses and compares basic regression methods using case studies.

1.3.4.1 Partial least squares regression

Partial least squares (PLS) is probably the most commonly employed regression method in chromatography [184-191]. PLS is a multivariate projection method for modelling a relationship between dependent variables (Y) and independent variables (X). PLS aims to find the components in the input matrix (X) that describe as much of the relevant variations in the input variables as possible, while having the maximal correlation with the target value in Y, giving less weight to the variations that are irrelevant or noisy. Hence, PLS models both X and Y simultaneously to find the latent variables in X that will predict the latent variables in Y [163]. Two models are obtained in PLS:

$$\boldsymbol{X} = \boldsymbol{T} \cdot \boldsymbol{P} + \boldsymbol{E} \qquad \text{Equation 1.7}$$

$$\boldsymbol{c} = \boldsymbol{T} \cdot \boldsymbol{q} + \boldsymbol{f} \qquad \text{Equation 1.8}$$

Where $\boldsymbol{X}$ represents the original data matrix and $\boldsymbol{c}$ is the concentrations. The first equation is similar to that of PCA, however the scores matrix also models the concentrations. The matrix $\boldsymbol{T}$ is common to both equations and the error matrices in $\boldsymbol{X}$ and $\boldsymbol{c}$ blocks are given by $\boldsymbol{E}$ and $\boldsymbol{f}$, respectively [192]. These matrices are represented graphically in Figure 1.13.

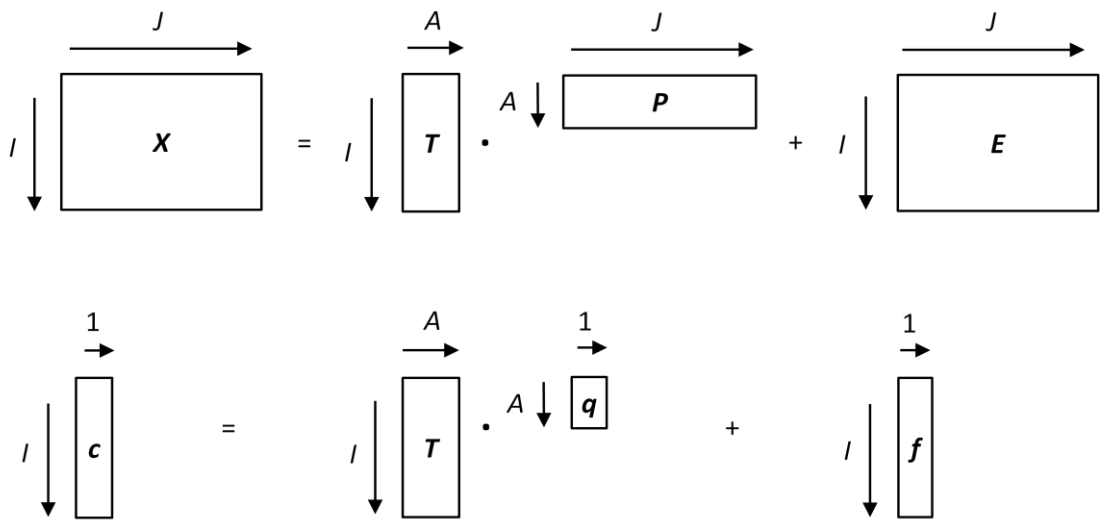


Figure 1.13: Principles of PLS, adapted from [141]

The theory of PLS has been extensively described by Wold et al. in a number of publications [193-202].

1.3.4.2 Principal components regression

As with PLS, principal components regression (PCR) models the dependent and independent variables to construct new components, however PCR differs from PLS in the way it constructs the components. PCR creates the components to explain the observed variability in the independent variables, without considering the dependent variables.

In chromatography, PLS and PCR have been compared by Wentzell and Vega Montoto [203] using simulations of complex mixtures. They found no significant differences in the prediction errors reported by PCR and PLS; however PLS usually required fewer latent variables, but this did not influence the predictive ability. This article also provided a review of PLS and PCR comparisons in the literature, with the common consensus being that PCR and PLS were similar, but generally PLS required less latent variables.

1.4 Chemometric applications in chromatography

Some select examples of data pre-processing and analysis methods in gas and liquid chromatography are provided in Table 1.6.

Table 1.6: Select examples of chemometric application in chromatography

Description	**Method**	**Data pre-processing**	**Data analysis**	**Reference**
Chemical fingerprinting of petroleum biomarkers	GC	Baseline correction, normalisation, alignment	PCA	[89]
Classification of gasoline	GC	Baseline correction, alignment, normalisation	PCA, LDA	[103]
Chemometric analysis of diesel fuel for forensic and environmental applications	GC	Baseline correction, alignment	PCA	[109]
Assessment of oil weathering	GC	Baseline correction, alignment, normalisation	PCA	[112]
Association and discrimination of diesel fuels	GC	Baseline correction, alignment, normalisation	PCA	[150]
Analysis of petroleum compositional similarity	GC×GC	Scaling	PCA	[204]
Multivariate pattern recognition of petroleum-based accelerants	GC	Normalisation, scaling	PCA, CA, SIMCA	[205]
Identification of adulteration of gasoline	GC	Normalisation	CA, *k*-NN	[206]
Quantification of naphthalenes in jet fuel	GC×GC	Alignment	PLS	[207]
Chemical fingerprinting of unevaporated automotive gasoline samples	GC	Scaling	PCA, LDA	[208]
Prediction of total green tea antioxidant capacity	HPLC	Alignment	PCA, PLS	[111]

Description (cont.)	**Method (cont.)**	**Data pre-processing (cont.)**	**Data analysis (cont.)**	**Reference (cont.)**
Profiling water soluble compounds in tea extracts	HPLC	Baseline correction, alignment, normalisation	PCA	[113]
Comparative analysis of the chromatographic fingerprints of sage samples	HPLC, GC	Baseline correction, alignment, smoothing,	PCA	[114]
Effectiveness of the use of triacylglycerols for the quantification of olive oil in vegetable oil blends	HPLC	Baseline correction, alignment, scaling	PLS	[123]
Prediction of sensory properties of Brazilian Arabica roasted coffees	GC	Normalisation, smoothing	PLS	[209]
Identification of volatile aroma-active compounds in oregano	GC	Scaling	CA, LDA, PLS	[210]
Classification of Chilean wines from different years, valleys and vineyards	HPLC	Normalisation	LDA, QDA, *k*-NN	[211]
Classification of vegetable oils characterised by the content of fatty acids	GC	Scaling	PCA, LDA	[212]
Analysis of tocopherols and triglycerides in coffee and their use as authentication parameters	HPLC	Normalisation	PCA, LDA	[213]
Development of a data mining system for metabolite identification	GC	Scaling	PCA, SIMCA	[214]
Exploring metabolic fingerprints in urine samples from prostate and bladder cancer patients	LC	Baseline correction, alignment, normalisation, scaling	PCA	[215]
Separation, characterisation and semi-quantitation of phospholipids from extracts of complex biological samples	HPLC	Normalisation, scaling	PCA	[216]

Description (cont.)	**Method (cont.)**	**Data pre-processing (cont.)**	**Data analysis (cont.)**	**Reference (cont.)**
Arsenic speciation patterns in freshwater fish	HPLC	Normalisation	CA	[217]
Determination of pesticides in vegetables	GC	Scaling	PLS, PCR	[218]
Characterisation of herbal extracts and importance of pre-processing	HPLC	Alignment, normalisation	PCA	[90]
Monitoring and detection of unknown impurities in an industrial insulin intermediate	LC	Baseline correction, alignment, normalisation, scaling	PCA	[124]
Classification of pharmaceutical substances	HPLC	Normalisation, scaling	PCA, CA	[131]
Chromatographic fingerprints for quality control of herbal medicines	GC	Alignment	PCA	[219]
Qualitative and quantitative analysis of chemical components from herbal medicines	HPLC	Alignment	PCA	[220]
Chemometric detection of thermally degraded samples in the analysis of drugs of abuse	GC	Baseline correction, normalisation, scaling	PCA, SIMCA	[221]
Pattern recognition techniques for screening drugs of abuse	GC	Baseline correction, normalisation, scaling	PCA, SIMCA	[222]
Forensic classification of ballpoint pen inks	HPLC	Scaling	PCA, LDA	[223]
Characterising stationary phases	LC	Scaling	PCA, CA	[224]
Determination of orthogonal chromatographic systems to characterise impurities in drug substances	HPLC	Scaling	PCA, CA	[225]
Characterisation of reversed-phase liquid chromatographic columns by chromatographic tests	LC	Scaling	PCA	[226]

1.5 Scope

This thesis investigates the use of pre-processing techniques and chemometric analysis of chromatographic data. The overall objectives of this research are:

1. To compare COW and *i*coshift for the alignment of HPLC data using PCA (**Chapter 2**).
2. To employ PCA as an exploratory technique to investigate metabolic changes in plant roots in response to a pathogen using HPLC (**Chapter 2**).
3. To evaluate HPLC with acidic potassium permanganate chemiluminescence detection for the classification of Australian wines according to geographic origin and vintage (**Chapter 3**).
4. To classify wines according to geographic origin using discriminant analysis (LDA and QDA) (**Chapter 3**).
5. To construct a regression model to correlate chromatographic peaks with wine age using PLS and PCR (**Chapter 3**).
6. To develop software for quality control of flavours and fragrances (**Chapter 4** and **Chapter 5**).
7. To develop an algorithm for the automated alignment of GC×GC chromatograms using affine transformation (**Chapter 4**).
8. To develop an algorithm for the automated comparison of GC×GC chromatograms using fuzzy logic (**Chapter 5**).

Chapter 2 - Exploratory Data Analysis of Plant Metabolite Profiles Using HPLC

2.1 Introduction

Exploratory data analysis (EDA) is commonly used to gain visual insight and simplify the interpretation of data. EDA is applied in order to remove redundancy and noise while retaining as much of the information present in the original data as possible. The most frequently employed EDA technique is PCA. PCA reduces the dimensionality of the data by transforming the original variables into PCs, linear combinations of the original variables that are uncorrelated so there is no redundancy in the data. PCs are characterised by scores and loadings, which can be plotted in order to achieve visualisation [163].

In chromatography, PCA has recently been employed in the exploratory analysis of food and beverages [190, 191], pharmaceuticals [87, 124], environmental [147, 227], biological [148, 215] and forensic samples [149, 228].

When PCA is performed on multivariate data sets, a number of sources of variation may be encountered. EDA techniques such as PCA describe variation between samples, however chromatographic variation induced by the instrument or sampling procedure may serve to complicate this. Since PCA finds the directions of maximum variance, the chromatographic variations may be combined with actual sample variations, making interpretation of the results difficult. In general, the main source of chromatographic variation is retention time shifting, which is caused by variations in flow rate and temperature, mobile phase composition and stationary phase decomposition. These shifts can be systematic for all peaks as well as random for individual peaks [105]. Retention time shifts can cause problems as PCA interprets changes between samples at corresponding retention times; this requires uniform presentation of data, i.e. all signals should be of equal length and, when placed as rows in a data matrix, corresponding variables (peaks) should be in the same column of the matrix [98, 104]. In order to effectively interpret the between-sample

differences, chromatographic variations have to be eliminated or reduced. If not, the PCA model may describe retention time shifts and conceal significant information regarding the samples as the true sample variations may be small in comparison and as a result difficult to distinguish in the PCA results [105].

Several techniques have been proposed in the literature for aligning chromatographic data, including COW [100, 101, 116], DTW [101, 102], PWA [103, 110, 119] and PTW [92]. Of these techniques COW is the most commonly employed [88, 89, 106, 111-115]; it is based on aligning a sample chromatogram to a reference by piecewise linear stretching or compression in combination with interpolation. COW is easy to implement and is considered to be of low flexibility (i.e. more peak shape preserving) [100].

The NMR alignment method, *i*coshift [121], has also been successfully applied to align chromatographic data [122-124, 229]. The *i*coshift algorithm has been demonstrated to be orders of magnitude to be faster than COW as there is no interpolation step and alignment is achieved by shifting the segments sideways, rather than shrinking or stretching the segments as in COW.

There is still no generally accepted standard measure for assessing alignment quality. A number of methods have been proposed in the literature, including PCA [83, 89, 90, 99, 101, 108, 113, 118], correlation coefficients [98, 102, 125] and visual inspection [98, 102].

Szymansaka et al. [118] evaluated DTW, COW and PTW for the analysis of urinary nucleosides of cancer patients and healthy subjects. The alignment algorithms were compared in terms of shift correction, computation time and ease of parameter optimisation. PCA was employed to evaluate the quality of alignment by looking at the separation achieved in the scores plot and the percentage of variance explained in the PCs. PCA demonstrated the significant advantages of COW and DTW over PTW or unaligned data. The separation between healthy and cancer patients was improved in the scores plot and more variance was explained in the first PC as the retention time shifts were removed as a source of variation [105]. The most effective algorithm for shift correction was found to be COW, which was also faster than DWT. PTW was the fastest and simplest to be employed, however it was not as precise as COW or DTW [118].

In this chapter PCA is employed to compare and evaluate two alignment methods, COW and *i*coshift. As an EDA technique, PCA is subsequently applied to aligned HPLC metabolite profiling data in order to investigate the influence of phosphite on the secondary metabolites associated with *Lupinus angustifolius* roots inoculated with *Phytophthora cinnamomi*.

The soil-borne plant pathogen, *P. cinnamomi* is the causal agent of widespread disease in vegetation and agriculture. It interrupts the physiological and chemical processes of a plant through the invasion of its root system [230].

There is no method currently available to eradicate *P. cinnamomi* in native forests without destroying the native plants themselves. Current management strategies aim to reduce spread of the pathogen and help build up the defences of existing host plants. The exogenous application of phosphite has shown some success in reducing the symptoms of disease associated with this pathogen in both agriculture and native forest incursions [231, 232].

The mode of action of phosphite has not been established and a clear understanding of the mechanism is required to optimise processes involved in the application to the plant of interest. This results in the need to develop methodology to monitor plant metabolites influenced by the application of phosphite. GC has primarily been used as the separation method for profiling primary plant metabolites, however the derivatisation of extracts required for GC analysis is known to degrade the glycosides and esters of secondary metabolites. Hence, HPLC has become the method of choice for secondary metabolite profiling [233].

2.2 Experimental

2.2.1 Samples

Plant growth

Certified *L. angustifolius* seeds (Department of Primary Industries, VIC, Australia) were surface sterilised in ethanol for 30 seconds prior to a triple rinse in distilled water. The seeds were set in a soil-free plant growth system and root tips were removed six days after germination to encourage the growth of lateral roots. Development of adequate foliage for the application of phosphite was achieved after 14 days [234, 235].

Application of phosphite treatment

The commercially available product 'Throw Down' systemic fungicide (Nipro Products Pty Ltd, QLD, Australia), with the active constituent of 400 g L^{-1} phosphorous acid (H_3PO_3), present as mono (KH_2PO_3) and di-potassium phosphite (K_2HPO_3) was used as the source of phosphite for this study. To aid chemical adhesion of the phosphite solution to plant foliage, an adjuvant (Biotrol Oil, Gulf A G Pty Ltd, Clayton, VIC, Australia) was added.

The seedlings were removed from the controlled growth medium 14 days after germination for foliage treatment. The treatments were applied as a fine mist from a hand held spray bottle (Canyon Corp Pty Ltd, Melbourne, VIC, Australia) until the treatment began to run off the leaf surface. After treatment, the seedlings were left to dry for 15 minutes and placed back into the controlled growth conditions.

Treatments were applied to the foliage of 14-day-old seedlings as 5 g L^{-1} phosphite in deionised water with the adjuvant added at 2.5 mL L^{-1}. Deionised water was used as the control. After allowing 48 hours for translocation, root tips were inoculated with a few strands of clarified *P. cinnamomi* hyphae.

At time points 0, 12 and 72 hours post inoculation, the lower 20 mm of roots were excised and external mycelia removed with forceps. Roots were immediately flash frozen in liquid nitrogen to quench metabolic activity and were stored at -80°C. The

metabolites were extracted using water, acetonitrile and isopropanol (2:3:3) and separated via HPLC [234, 235].

2.2.2 Chromatographic analysis

Chromatographic separations were performed by employing a Hewlett Packard 1100 series high performance liquid chromatograph (Agilent Technologies, Blackburn, VIC, Australia). Root extracts were separated through a reversed phase C_{18} column (Waters, 250 mm x 4.6 mm, particle size 5 µm, Alltima, Alltech Ass. Inc., Deerfield, IL, US). The mobile phase was comprised of two solvents; A: water / formic acid 0.1% v/v and B: acetonitrile / formic acid 0.1% v/v. The pH of the mobile phase at 0.1% was 2.5. Formic acid was added to the mobile phases to aid phenolic and alkaloid elution from the samples. The flow rate was 1.0 mL^{-1} min and the volume injected was 20 µL. The column was kept at room temperature. The HPLC system was equipped with a Diode Array Detector (DAD) (1200 DAD, Agilent Technologies). A wavelength of 280 nm was chosen for detection of phenylpropanoid metabolites in this protocol [234, 235].

2.2.3 Mass spectrometry

The samples were analysed via electrospray ionisation quadrupole time-of-flight mass spectrometry (ESI-QTOF-MS) (6210 MSD TOF-MS, Agilent Technologies). The MS was calibrated using a standard tuning mix (G2421A, Agilent Technologies). The mass spectra were recorded in negative ion mode based on the following ESI source settings; drying gas (N_2) flow rate and temperature (7 L min^{-1}, 350°C), nebuliser gas (N_2) pressure (30 psi), capillary voltage 3.0 kV, vaporiser temperature 350°C, and cone voltage 60 V. MS data acquisition was carried out using Agilent MassHunter Workstation Acquisition for TOF/Q-TOF (B.02.00 (B1128)) and data analysis carried out using Agilent MassHunter Qualitative Analysis (version B.03.01) [234, 235].

2.2.4 Data pre-processing and analysis

All data manipulation and analysis algorithms were implemented using Matlab (V7.10 (R2010a), MathWorks Inc, MA, USA), the PLS Toolbox (V4.0.2, Eigenvector Research Inc., WA, USA) and in-house developed algorithms.

Chemometric methods were applied to raw chromatographic data. The data was first smoothed using the Savitzky-Golay algorithm with an 11 point quadratic filter, then

aligned and normalised using an internal standard (vanillic acid). Two alignment methods, COW and *i*coshift, were compared. The algorithms are both publically available and were sourced from [236].

PCA was employed as an EDA technique as well as a measure of assessing the effectiveness of alignment by examining the structure of the data in the scores plot and the percentage of variance explained by the PCs. PCA was applied to mean-centred data.

2.3 Results and discussion

The data set consists of four classes; *L. angustifolius* treated with water, *L. angustifolius* treated with phosphite, *L. angustifolius* inoculated with the plant pathogen *P. cinnamomi* and treated with water and *L. angustifolius* inoculated with the plant pathogen *P. cinnamomi* and treated with phosphite. The non-inoculated samples serve as controls to monitor the plants response to the water and phosphite treatments.

The plant roots were analysed at 0, 12 and 72 hours post inoculation. The average chromatograms for each time point are shown in Figure 2.1. There are no metabolites in the 0 or 12 hour data that are not present in the 72 hour data and the metabolites in the 72 hour data are also present in higher concentrations. As a result, all comparative investigations are carried out using the 72 hour data.

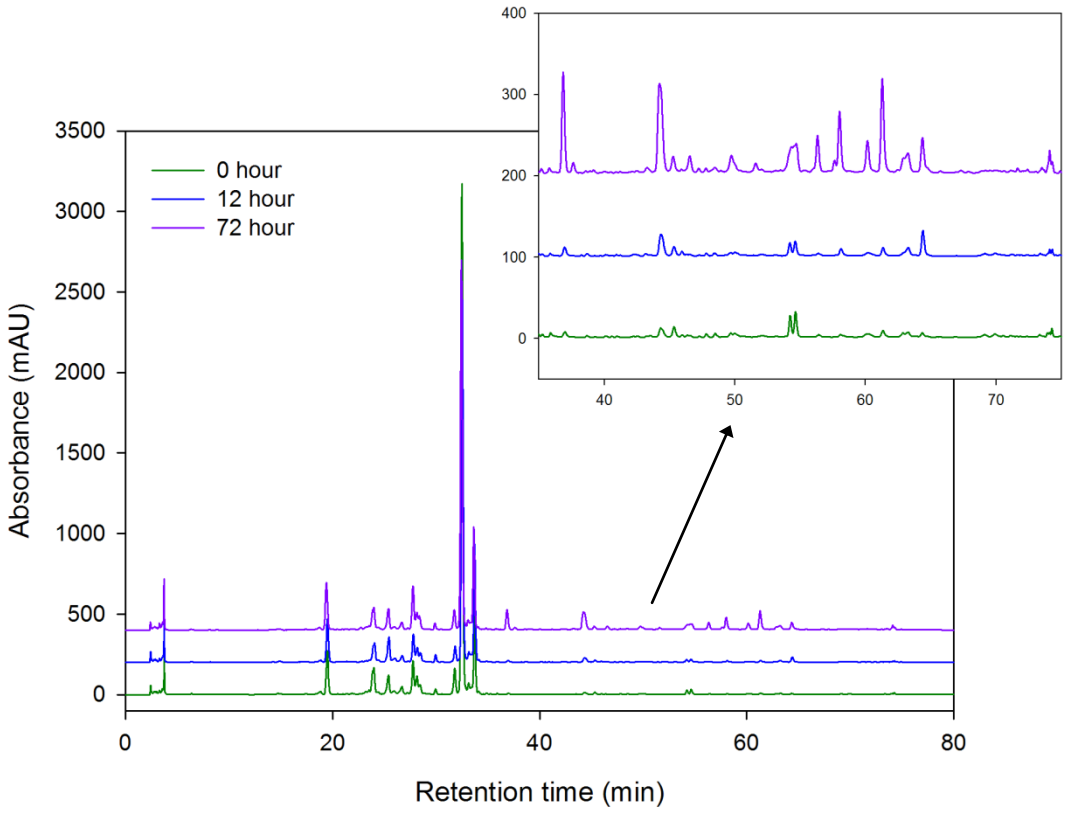


Figure 2.1: Average aligned chromatograms for the 0, 12 and 72 hour time points. In order to aid visualisation, the chromatograms were offset by 100 points.

The average chromatograms of the 72 hour data for the non-inoculated and *P. cinnamomi* inoculated, water-treated and phosphite-treated *L. angustifolius* roots are shown in Figure 2.2. Tentative peak assignments are provided in Table 2.1. Metabolites were identified by first determining their molecular weight from the molecular ions obtained using ESI-QTOF-MS. This information was compared to an open access database [237] to compile a list of possible metabolite identities. The MS fragmentation patterns of the metabolites were then compared to MS fragmentation patterns of plant based metabolites published on an open access database [237] and in the literature [238-242] to compile the final list of metabolites. A full discussion/interpretation of MS results is provided in [234].

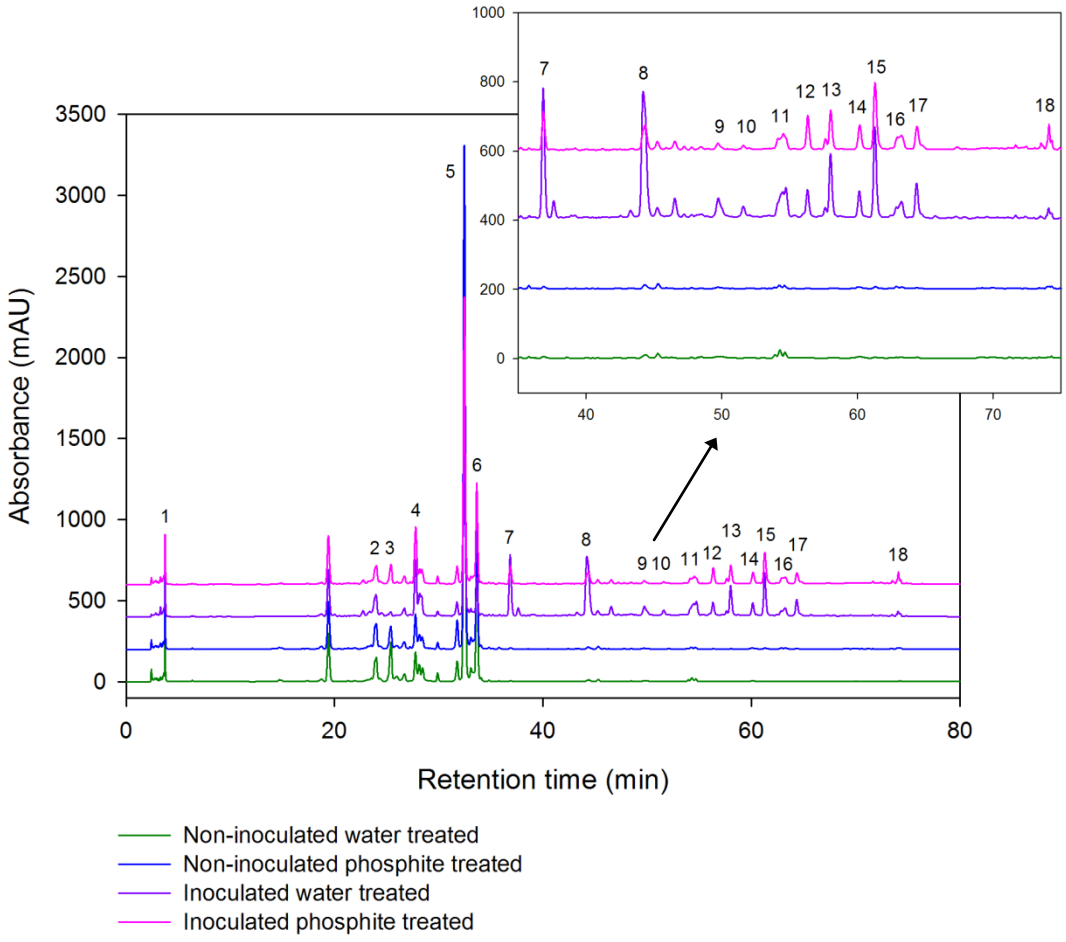


Figure 2.2: Average aligned chromatograms of the 72 hour data with peaks identified (the peak at 19 minutes is the internal standard (vanillic acid)). Tentative peak assignments are provided in Table 2.1. In order to aid visualisation, the chromatograms were offset by 200 points.

Table 2.1: Tentative peak assignments

Peak	Compound
1	3,4,5-trihydroxybenzoic acid (gallic acid)
2	Genistein 4’,7-O-diglucoside malonylated I
3	Genistein 4',7-di-O-glucoside malonylated II
4	2'-Hydroxygenistein 7-O-glucoside malonylated I
5	Genistein 7-O-glucoside malonylated I
6	Genistein 7-O-glucoside malonylated II
7	5,7,2',4'-Tretrahydroxyisoflavone (2'-Hydroxygenistein)
8	4',5,7-trihydroxyisoflavone (genistein)
9	3-Rha-Gal-Glc-soyasapogenol B
10	Rha-Gal-GlcA-Soyasapogenol E
11	Flavonoid glucoside
12	Luteone or Licoisoflavone A or Lupinisoflavone C
13	Luteone or Licoisoflavone A or Lupinisoflavone C
14	Lupinisoflavone A
15	Luteone or Licoisoflavone A or Lupinisoflavone C
16	Wighteone, Isowighteone and or Lupiwighteone
17	Wighteone, Isowighteone and or Lupiwighteone
18	Angustone A (2'-Hydroxyisolupalbigenin)

From this it can be seen that there is an increase in the concentration of the aglycones of genistein (peak 8) and 2’hyrdoxygenistein (peak 7), the saponins (peaks 9 and 10) and prenylated isoflavones (peaks 12-18) post inoculation. This is accompanied by a decrease in the concentration of malonylated genistein glucosides (peaks 2, 3 and 5).

Aglycones are often stored in high concentrations as inactive glycosides for downstream biosynthesis in response to stress [243]. In lupin species, a well recognised defence response is the prenylation of aglycones such as genistein and 2’hyrdoxygenistein resulting in the production of prenylated isoflavones which are known to be highly anti-fungal metabolites [244, 245]. As well as prenylated isoflavones, *L. angustifolius* accumulated saponin-based metabolites in response to *P. cinnamomi*. Saponins exhibit a range of antimicrobial and antifungal activity [246]. The increase in peaks 9 and 10 post pathogen inoculation suggests that these saponins have a defence function in *L. angustifolius*.

The observed increase in most of the defence metabolites appears to be lower in the phosphite-treated samples compared with the water-treated. This may be due to down-

regulation of the plant's immune response after the pathogen is arrested. In contrast, water-treated plants would continue to produce defence metabolites as the pathogen progressed.

The increase in angustone A (peak 18) in the phosphite-treated *L. angustifolius* post inoculation is above that of the plants normal defence response (water-treated). Hence, it is possible that phosphite enhances the production of this metabolite, which may play a role in the plant's defence response.

It has been previously suggested that high concentrations of phosphite in plant roots work directly on *P. cinnamomi*, preventing the pathogen from forming an association with the plant [247]. However, the results presented here demonstrate that this was not the case. If the pathogen was arrested prior to making an association with the plant, the metabolite profiles of the non-inoculated and inoculated phosphite-treated *L. angustifolius* would be equivalent as there would be no change to plant metabolism.

The water-treated and phosphite-treated *L. angustifolius* produce comparable metabolite profiles post inoculation with *P. cinnamomi*. This confirms that the pathogen is able to make an association with the plant despite the high concentration of phosphite in the root tissue and trigger the plants secondary metabolic defence response. These results will be discussed further when PCA is employed as an EDA technique (section 2.3.2).

2.3.1 PCA comparison of alignment methods

The first step in performing alignment is the selection of an appropriate reference chromatogram. The COW program provides several options for the choice of reference; namely the mean, median, maximum and highest correlation signals. The mean will contain peaks from all of the chromatograms, however using a generated chromatogram may cause problems as averaging a series of poorly aligned chromatograms would tend to produce a heavily distorted chromatogram and introduce artifacts [83]. Daszykowski et al. [83] conducted a study on the use of different reference chromatograms. They found that optimal results were obtained using the chromatogram with the highest mean correlation to the other chromatograms.

However, correlation can suffer if the data set contains two or more classes and as a result the alignment may be affected by the class the reference belongs to [100, 125].

Figure 2.3 (a) shows the potential reference chromatograms available in the COW program. There are significant differences between the references from approximately 50-70 minutes, hence this section of the chromatograms is expanded in Figure 2.3 (b) and examined further. From this it can be seen that the peaks in the chromatogram with the maximum signal are distorted; this may cause problems as the interpolation of the data points in COW is guided by the shape of the peaks in the reference [100]. The peak at around 54.5 minutes is supposed to be two separate peaks, but the only reference that shows this is the chromatogram with the highest correlation. In order to see what effect this will have on alignment, all four potential reference chromatograms were examined to align the data.

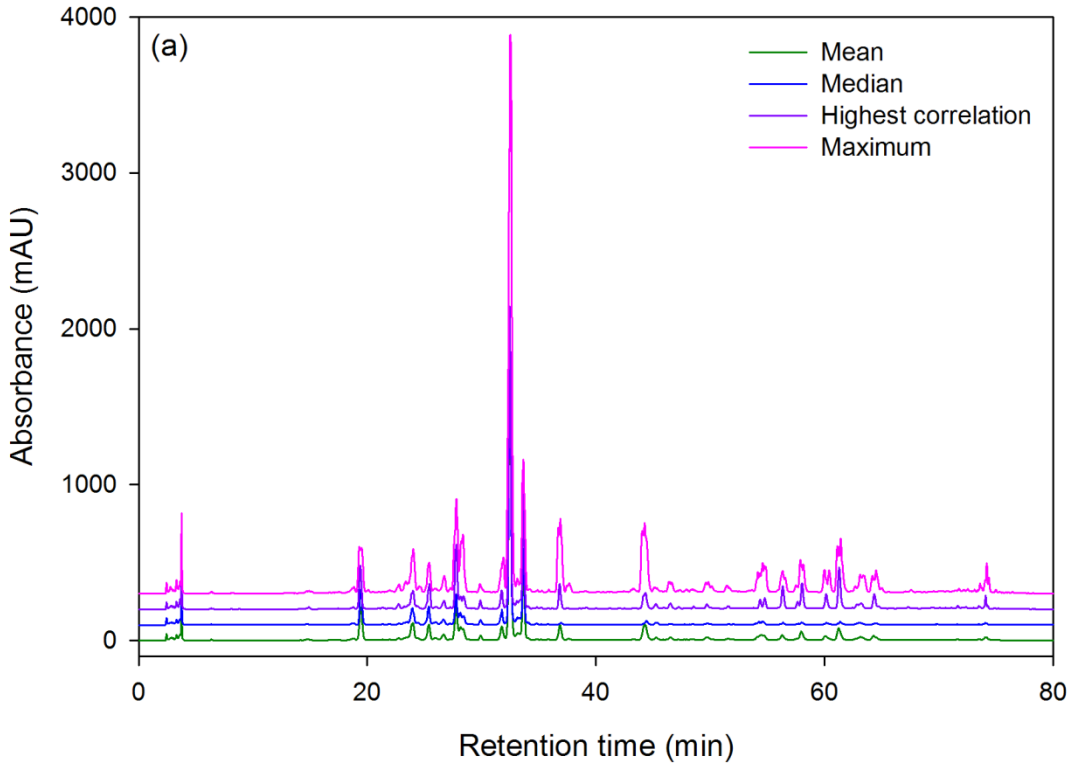


Figure 2.3: (a) potential reference chromatograms available in the COW program. In order to aid visualisation, the chromatograms were offset by 100 points.

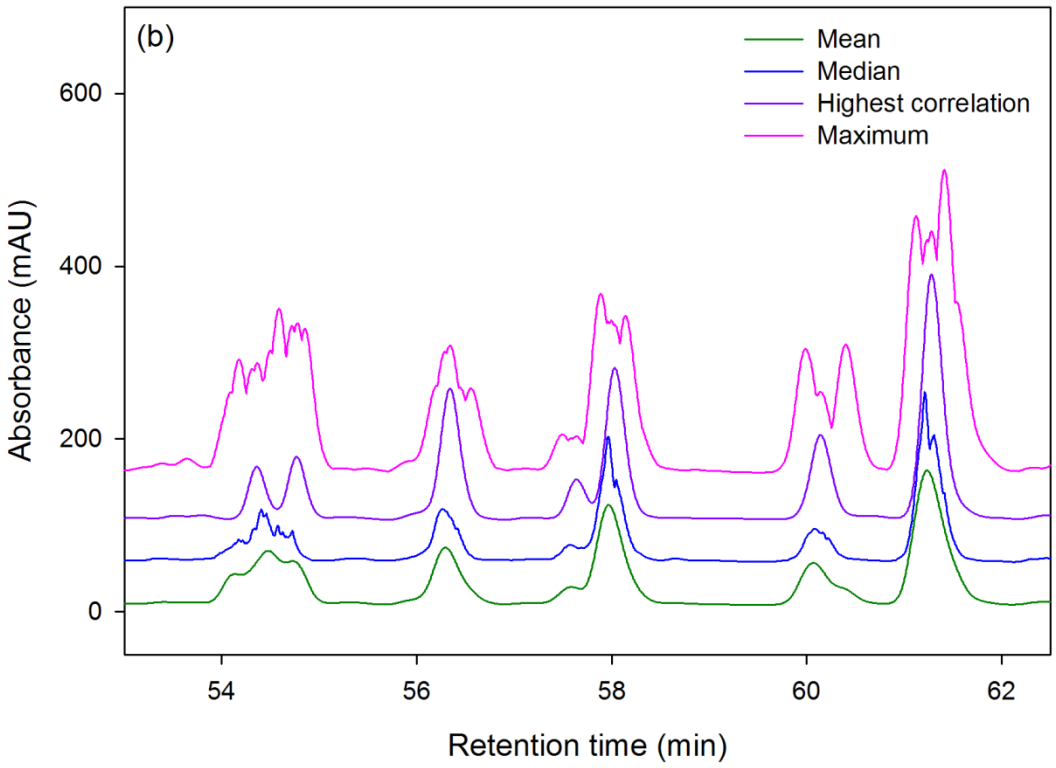


Figure 2.3: (b) expanded section of (a). In order to aid visualisation, the chromatograms were offset by 50 points.

Figure 2.4 shows the data before alignment. The data is then aligned to each of the different references using the COW algorithm with an arbitrarily selected segment length of 20 points and slack size of 5.

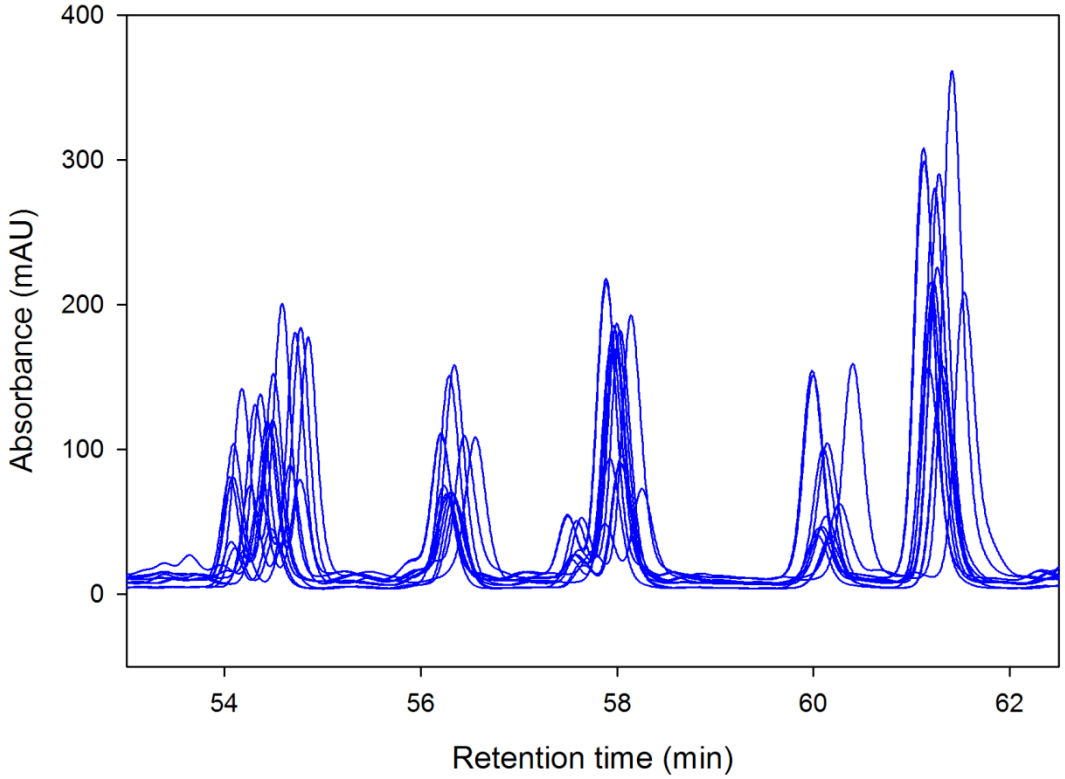


Figure 2.4: Expanded section of chromatograms before alignment

The results of alignment are shown in Figure 2.5. The alignment obtained for the peaks after approximately 55 minutes was successful for all references, however problems arose for the peaks at approximately 54.5 minutes. This was due to the poor shape of these peaks in the mean, median and maximum references. The most accurate alignment was achieved using the reference chromatogram with the highest correlation and as a result this chromatogram was chosen as the reference. The fact that there are four classes present in the data set does not present a problem as all the major peaks were found in the chromatogram with the highest correlation coefficient. The same reference was used for both COW and *i*coshift alignment.

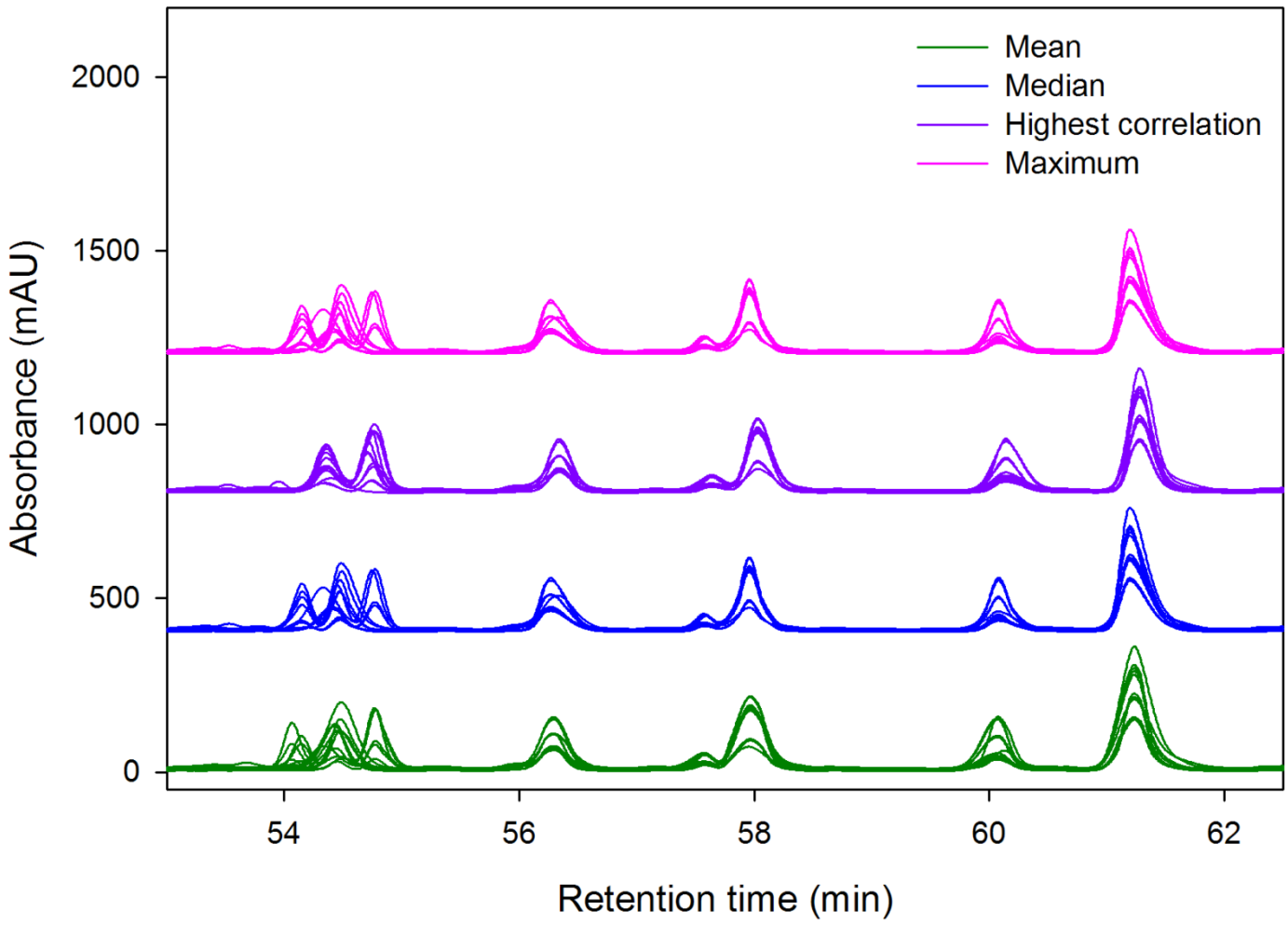


Figure 2.5: Results of COW alignment using mean, median, highest correlation and maximum reference chromatograms. In order to aid visualisation, the chromatograms were offset by 400 points.

COW requires the optimisation of two parameters, the segment length and slack size. Combining a small segment length and large slack size (i.e. high flexibility) will result in interpolation steps over many data points and thus the possibility of aligning peaks effectively, but this also increases the risk of changing the peak shapes and areas [100]. As a result, Skov et al. [100] developed an algorithm (built into the COW program) to optimise the selection of the segment length and slack size, whilst still preserving the peak area and shape. The optimisation method requires the user to define an optimisation space in which the selection of the segment length and slack size is to be based. This optimisation space is derived from the average peak widths in the reference chromatogram and the general observed shift in the peaks. As a rule of thumb, the segment length optimisation space, L, is:

$$L = PW_A \pm \frac{PW_A}{2} \qquad \text{Equation 2.1}$$

Where PW_A is the approximate average peak width measured at the base of all peaks in the reference chromatogram. In the data presented here, the average peak width is 110 points; therefore a segment length optimisation space of 55 to 165 was chosen. The correct slack size search space is more difficult to define as features such as different local peak shifts and increased flexibility of the COW algorithm in the middle of the chromatogram will have an effect on the outcome of the alignment procedure. According to Skov et al. [100] a rule of thumb is that if the number of data points before the first peak and after the last peak are approximately the same as the peak widths, then a slack size search space ranging from 1 to 15 is appropriate; this range was employed here. Based on the optimisation space, the algorithm selected a segment length of 105 and a slack size of 1.

The *i*coshift program does not have a method for optimising the segment length and slack, so these must be selected by the user via trial and error. Since the optimum segment length and slack size were already chosen for COW, they were also used for the *i*coshift algorithm to allow direct comparison between the alignment methods. Visual inspection of the *i*coshift aligned results is required because when segments of constant length are used there is the possibility that the segment edges may be located in a peak, leading to artifacts in the peak shapes. Furthermore, due to the lack of an interpolation step and the use of sideways shifting instead of stretching and shrinking, missing parts of the segment edges are replaced by repeating the value of the

boundary point; this can also result in artifacts [121, 125]. Visual inspection revealed the presence of artifacts using a segment length of 105 and a slack size of 1, an example of this is shown in Figure 2.6.

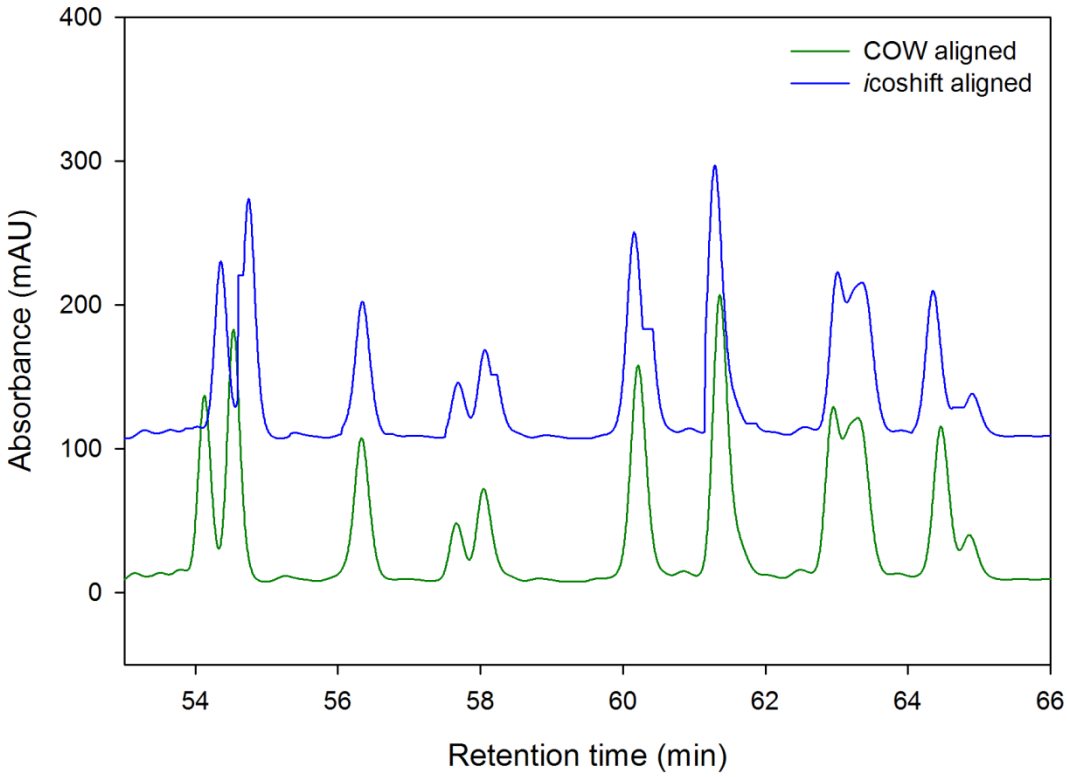


Figure 2.6: Artifacts observed in the *i*coshift aligned data (segment length = 105 and slack size = 1). In order to aid visualisation, the chromatograms were offset by 100 points.

In an attempt to remove the artifacts, different combinations of segment lengths and slack sizes were tested; the results are summarised in Table 2.2.

Table 2.2: Results using different segment length and slack size combinations for *i*coshift alignment

Segment length	Slack size	Artifacts	Quality of alignment
50	1	yes	poor
100	1	yes	poor
150	1	yes	poor
200	1	yes	poor
50	5	yes	average
100	5	no	average
150	5	yes	average
200	5	no	average
50	10	yes	average
100	10	yes	average
150	10	yes	average
200	10	yes	average
50	“best”	yes	good
100	“best”	no	good
150	“best”	yes	good
200	“best”	no	good

The “best” slack allows the algorithm to search for the shift in each segment that maximises the correlation between the reference and sample. This results in a different slack value for each segment and provides the algorithm with more shifting freedom.

The results from Table 2.2 indicate that both segment lengths of 100 and 200 with the “best” slack resulted in good alignment with no artifacts. However, 100 is closest to the segment length used in COW alignment (segment length = 105) and as a result will allow more accurate comparison between the alignment methods.

A small selected region of the chromatograms is used as an example to illustrate the quality of alignment achieved using a segment length of 100 with the various slack sizes. The results are presented in Figure 2.7.

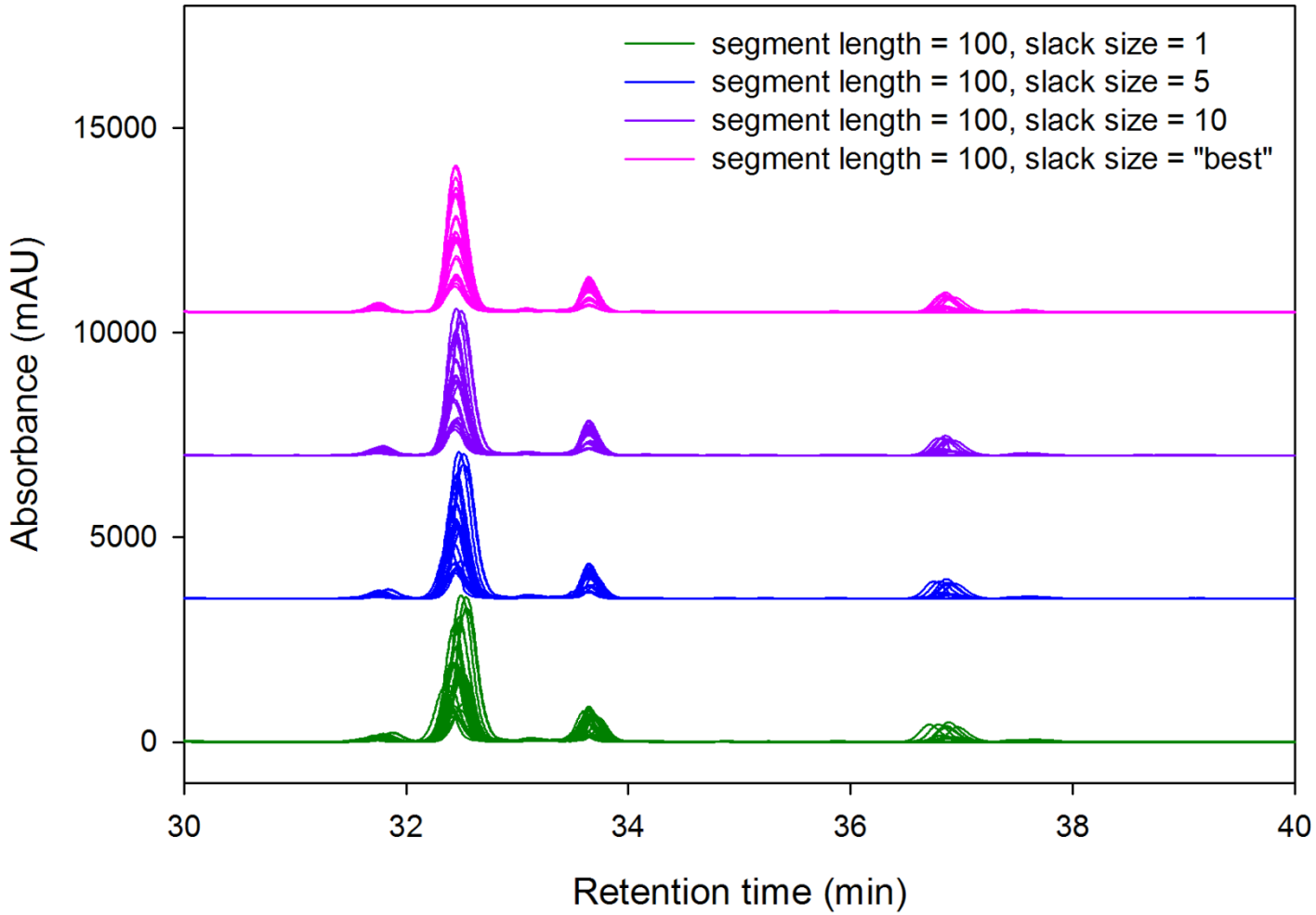


Figure 2.7: Results of *i*coshift alignment with a segment length of 100 and slack sizes of 1, 5, 10 and "best". In order to aid visualisation, the chromatograms were offset by 3500 points

Figure 2.7 indicates that a segment length of 100 and the "best" slack provide the most accurate alignment. Hence, this data was used for further analysis.

PCA was performed on the unaligned as well as the COW and *i*coshift aligned data to evaluate the effectiveness of alignment. Since the data consists of four classes (non-inoculated and inoculated, water-treated and phosphite-treated), it is expected that PCA will be able to discriminate between the classes. The unaligned and aligned data were smoothed, normalised and mean-centred prior to performing PCA, so the only difference in pre-processing was alignment. Table 2.3 gives the eigenvalues for the first ten PCs.

Table 2.3: PCA eigenvalues for the unaligned and COW and *i*coshift aligned data

	Unaligned		**COW aligned**		***i*coshift aligned**	
PC number	**% variance**	**% cumulative variance**	**% variance**	**% cumulative variance**	**% variance**	**% cumulative variance**
1	63.45	63.45	86.75	86.75	88.10	88.10
2	24.65	88.10	9.10	95.85	8.35	96.45
3	4.83	92.93	1.75	97.60	1.38	97.83
4	3.22	96.15	1.00	98.60	0.88	98.71
5	1.18	97.33	0.34	98.94	0.35	99.06
6	0.95	98.28	0.28	99.22	0.33	99.39
7	0.42	98.70	0.22	99.44	0.15	99.54
8	0.32	99.02	0.20	99.64	0.15	99.69
9	0.26	99.28	0.13	99.77	0.09	99.78
10	0.18	99.46	0.06	99.83	0.06	99.84

The aligned data sets contain more variance in the first PC than the unaligned data, this is in accordance with other published studies [83, 89, 101, 108, 113, 118]. More variance is explained in the first PC after alignment as the retention time shifts are removed as a source of variation [105]. Before alignment, variations between the samples and the peak retention times are described by the model; this means that more PC's are required to effectively describe the added variation. However, after alignment, only the sample variations remain and as a result can be effectively explained by the first PC.

Since most of the variance is accounted for in the first two PCs, they were subsequently examined and interpreted. Figure 2.8 (a to c) shows the scores plot of the first two PCs for the data before alignment, after alignment with COW and after alignment with icoshift, respectively. Before alignment (Figure 2.8 (a)) no separation between the classes is evident. After alignment with COW (Figure 2.8 (b)) and *i*coshift (Figure 2.8 (c)) the four classes are separated and easily identified.

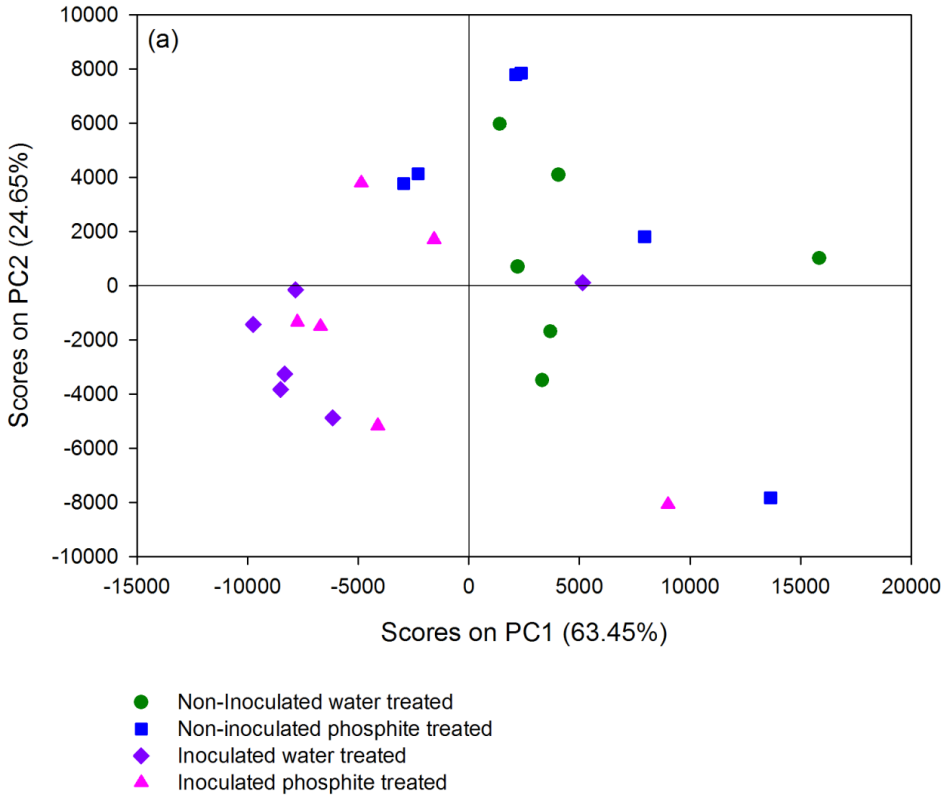
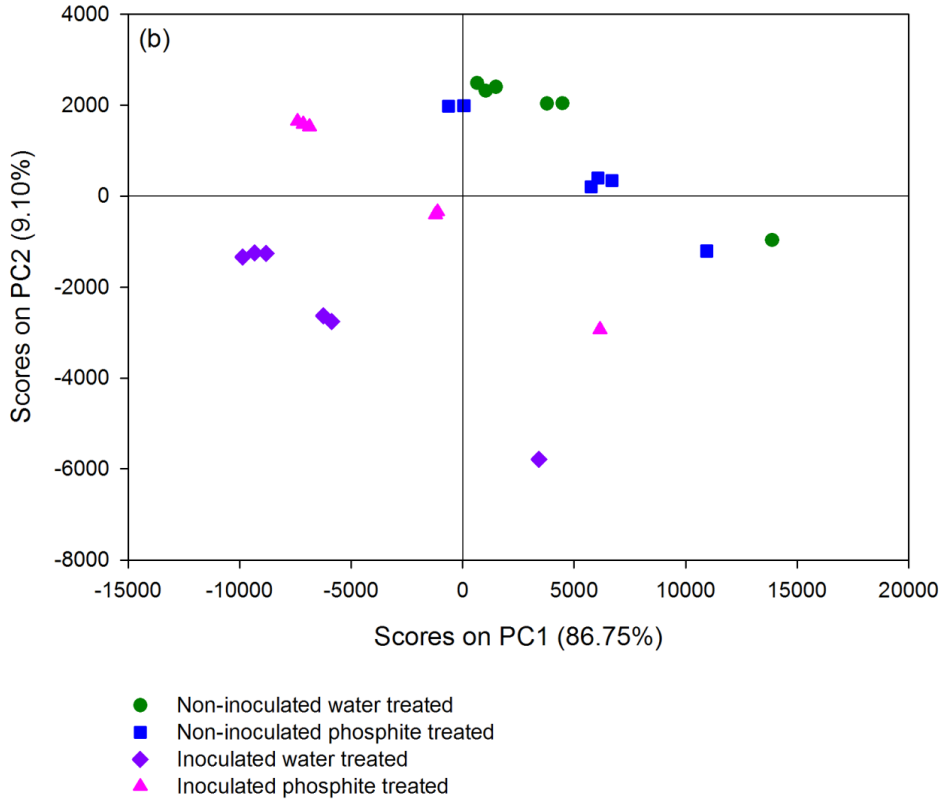


Figure 2.8: Scores plots for the first two PCs (a) before alignment



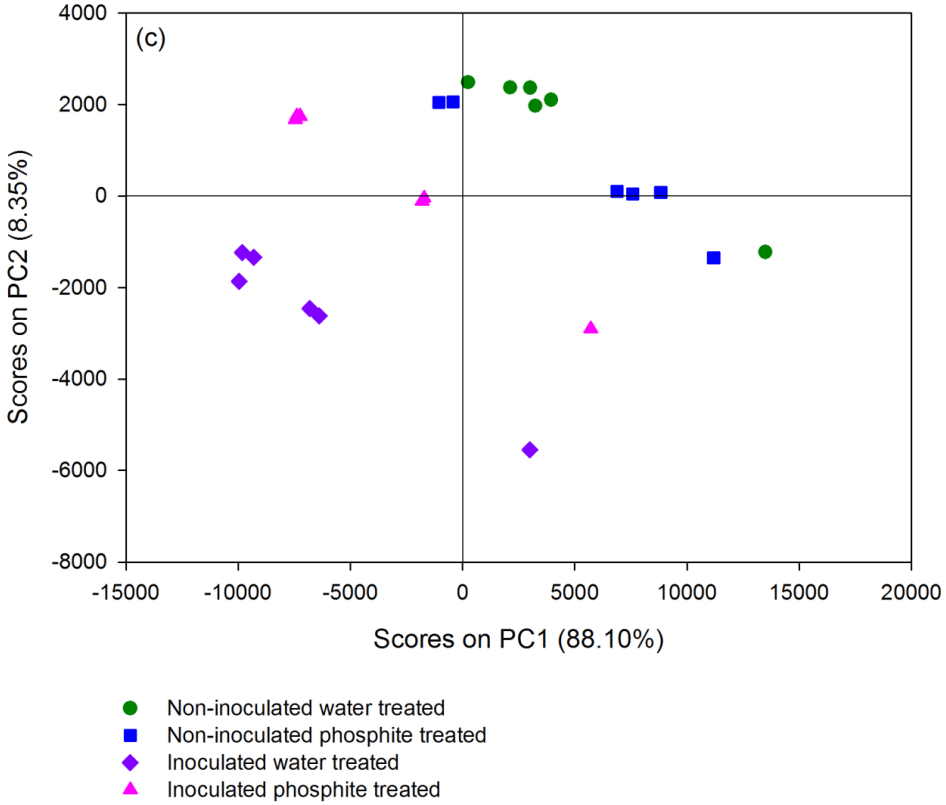


Figure 2.8: Scores plots for the first two PCs (b) COW aligned and (c) icoshift aligned

Before alignment, the loadings plot (Figure 2.9 (a)) shows broad variation with no peaks defined due to the peak shifts. However, after alignment with COW (Figure 2.9 (b)) and *i*coshift (Figure 2.9 (c)) the peaks responsible for separation between the different classes in the scores plots (Figure 2.8) are more evident. The *i*coshift aligned data produces slightly sharper peaks, which may be due to the alignment being more accurate.

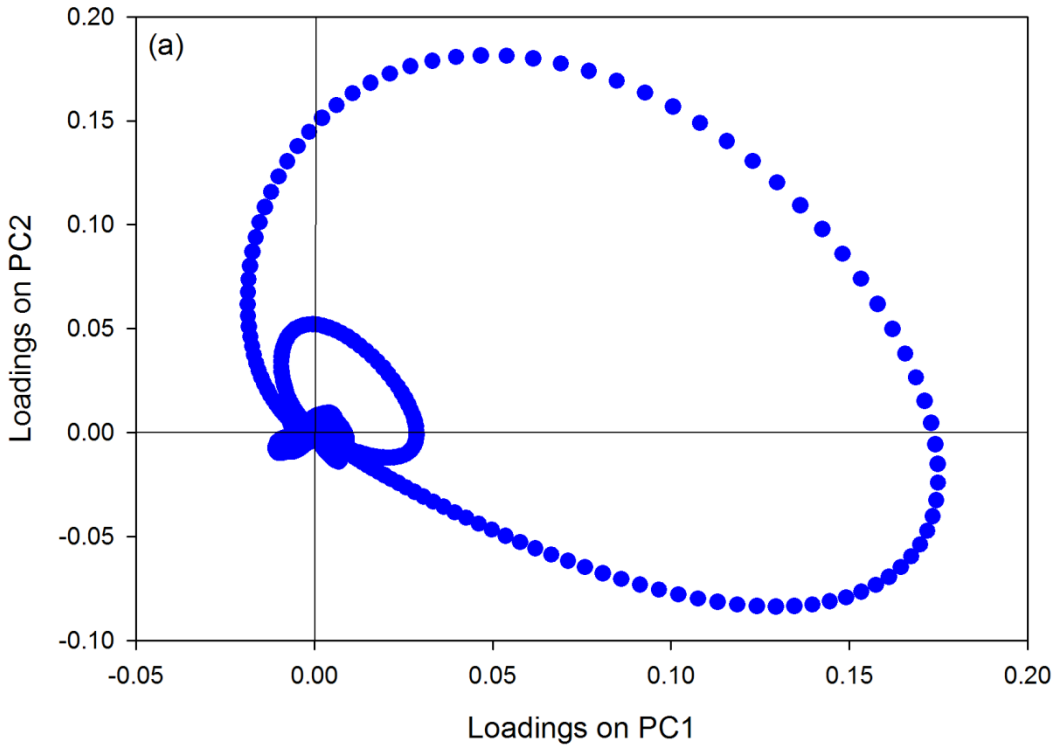


Figure 2.9: Loadings plots for the first two PCs (a) before alignment

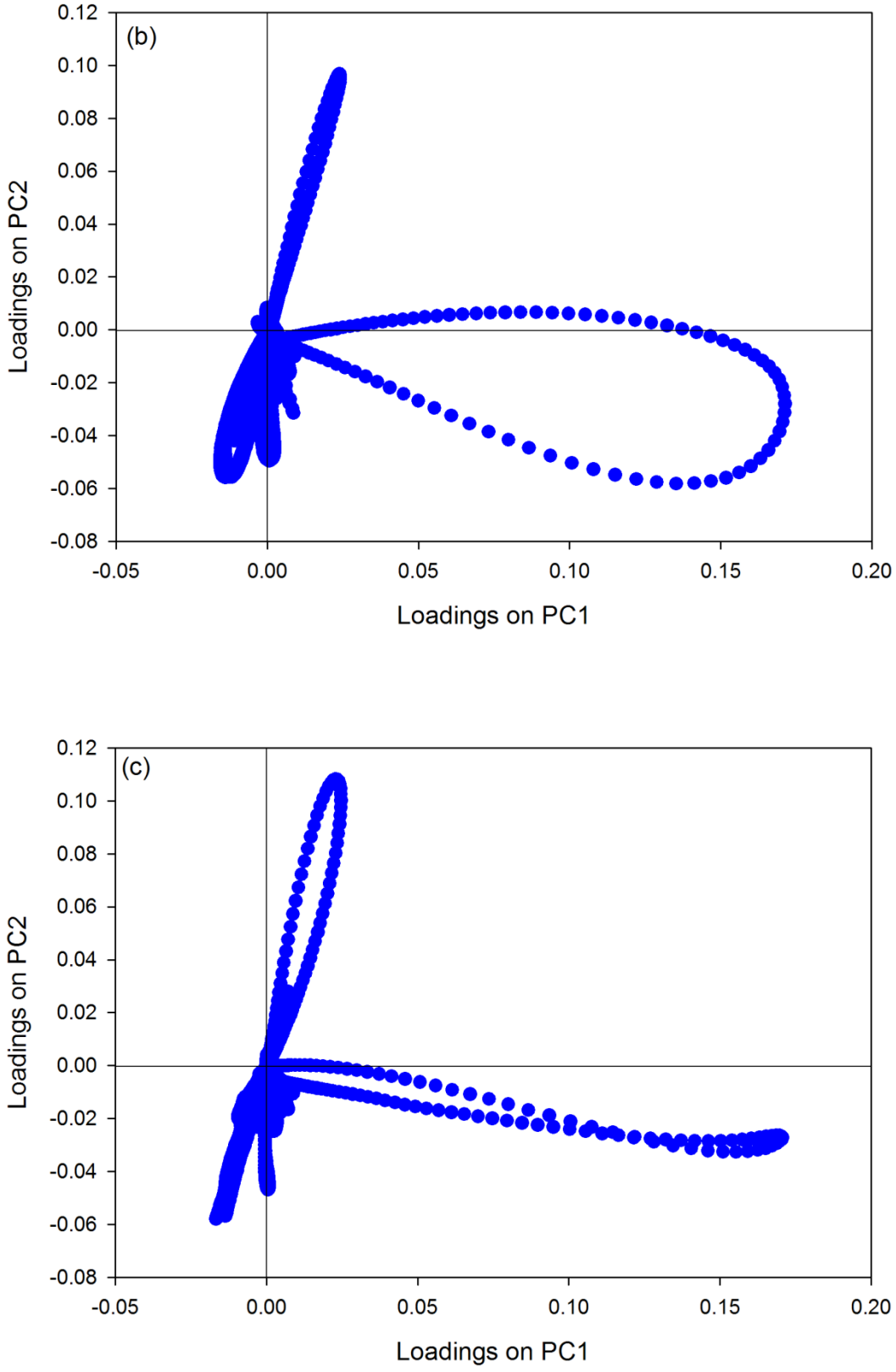


Figure 2.9: Loadings plots for the first two PCs (b) COW aligned and (c) *i*coshift aligned

Comparison of the alignment methods by PCA revealed that more variance was explained in the first PC and between class separation in the scores plot became evident after alignment. Both COW and *i*coshift were successful in aligning the data, however the *i*coshift aligned data explained slightly more variance and produced sharper peaks in the loadings plot. As a result, the *i*coshift aligned data is used in the exploratory analysis by PCA.

2.3.2 Exploratory analysis by PCA

From the scores plot (Figure 2.8 (c)) it can be seen that the non-inoculated (water and phosphite treated) samples clustered close together, which would suggest that the plant has a similar response to both treatments and as a result changes in the metabolites post inoculation can be attributed to the plants response to the pathogen.

The loadings on PC1 and PC2 are plotted against time in order to interpret the separation observed in the associated scores plot. The loadings are shown in Figure 2.10 with peaks identified.

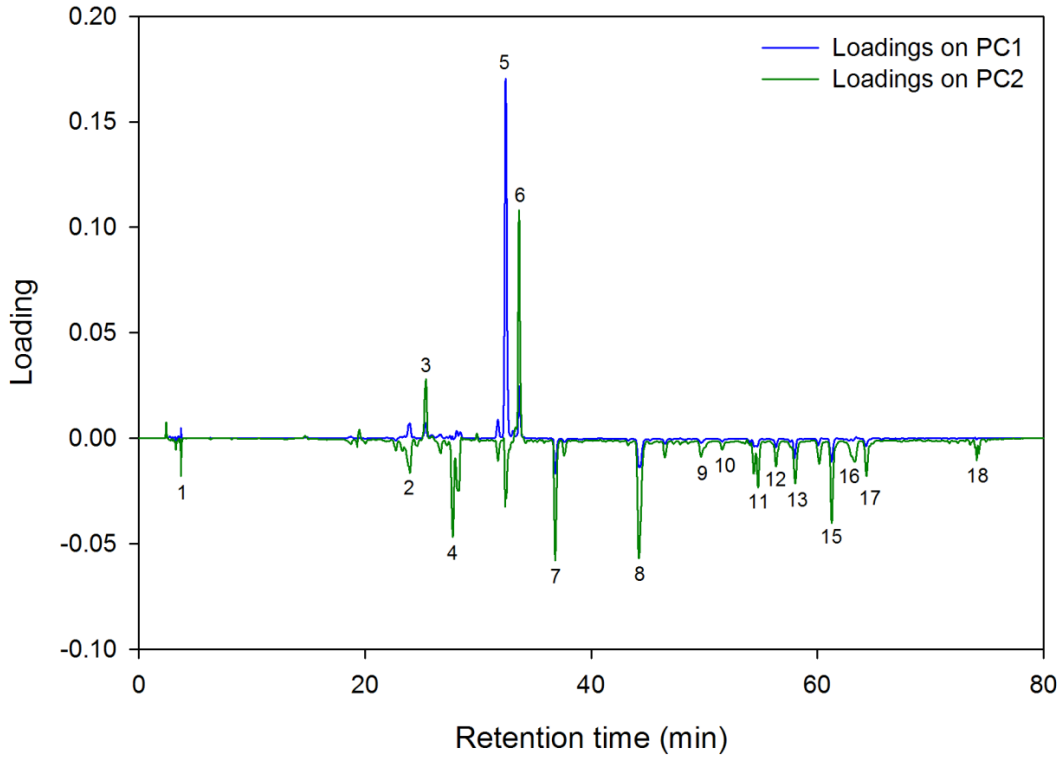


Figure 2.10: Loadings plot (PC1 and PC2) for icoshift aligned data with peaks identified. Tentative peak assignments are provided in Table 2.1.

The loadings on PC1 indicate the non-inoculated samples are separated based on the malonylated genistein glucosides (peaks 5 and 6), while the inoculated samples are separated according to the aglycones of genistein (peak 8) and 2'hyrodxygenistein (peak 7). On PC2, the loadings again highlight separation of the non-inoculated samples based on malonylated genistein glucosides (peaks 3 and 6) and separation of the inoculated samples based on the aglycones of genistein (peak 8) and 2'hyrodxygenistein (peak 7) as well as the prenylated isoflavones (peaks 12-18). These results concur with those obtained by the comparison of the average chromatograms (Figure 2.2).

From these results it is proposed that in response to the pathogen, *P. cinnamomi*, *L. angustifolius* up-regulates a defence beginning with the cleavage of glycosides from stored genistein and 2'hyrodxygenistein (peaks 2,3,5 and 6), which results in accumulation of aglycones of genistein (peak 8) and 2'hyrodxygenistein (peak 7). The prenylation of the genistein and 2'hyrodxygenistein aglycones then results in the production of a range of secondary metabolites (peaks 12-18), which are likely to be either a response to stress or a failed defence response. However, since many of the isoflavones tentatively assigned as secondary metabolites have previously been associated with the defence response of lupin species to pathogenic fungi [248], it is probable that the metabolic response observed here is an example of a failed defence response by *L. angustifolius*. The proposed metabolism of the 2'hyrodxygenistein and genistein glucosides is shown in Figure 2.11.

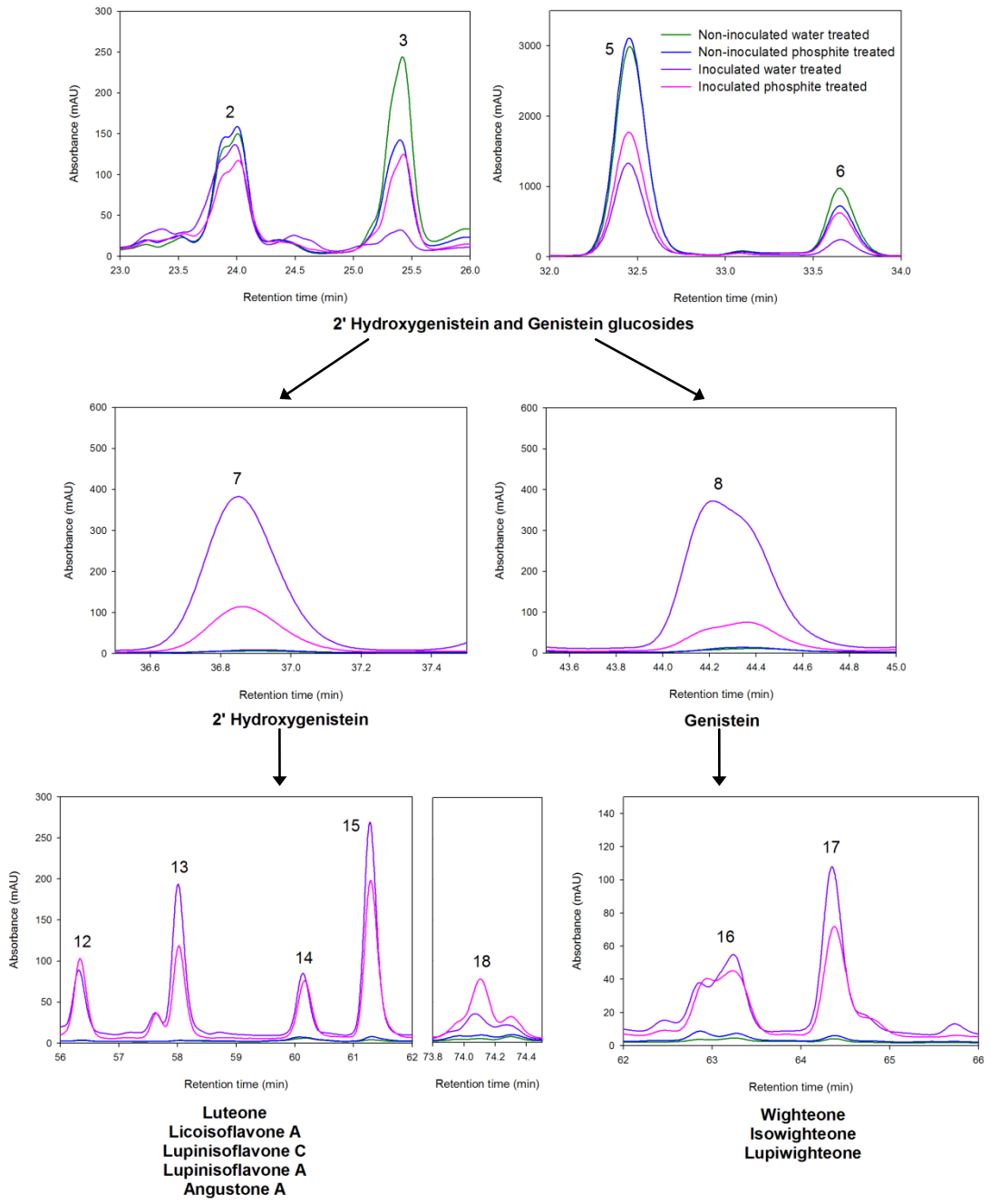


Figure 2.11: Proposed metabolism of 2'hydroxygenistein and genistein glucosides [234, 235]. Tentative peak assignments are provided in Table 2.1.

2.4 Conclusion

PCA was employed to compare and evaluate the effectiveness of COW and *i*coshift for the alignment of HPLC data. By comparing the explained variance and the separation in the scores plot before and after alignment, it was observed that after alignment more variance was explained in the first PC as the retention time shifts were removed as a source of variation. This also meant that the separation in the scores plot was enhanced as it was based on sample differences rather than retention time differences.

Both COW and *i*coshift were successful in aligning the data, with *i*coshift being slightly better as it explained more variance in fewer components. Alignment using *i*coshift was also orders of magnitude faster than COW. However, *i*coshift required multiple combinations of segment lengths and slack sizes in order to remove artifacts introduced by the algorithm.

PCA was also employed as an exploratory method to profile metabolites in *L. angustifolius* inoculated with the pathogen, *P. cinnamomi*, and treated with both water and phosphite. The results demonstrated that the pathogen stimulates a similar response using both treatments, differing only in concentration. This response was proposed to be a component of the plants defence against the pathogen. The triggering of the defence response suggests that the pathogen is able to make an association with the plant despite the presence of phosphite in the root tissue.

Chapter 3 - Classification of Wines by HPLC with Chemiluminescence Detection

3.1 Introduction

Classification of wines is important in the food industry [249, 250]. Wine has become a commodity of significant commercial value, with consumer expectations depending on many factors, including grape variety, maturity and geographic origin [251]. In Europe, most wine producing countries associate wine quality and value with both climate and soil characteristics, in particular defined by geographical classification or denomination of origin systems [252-254]. Recently, the determination of food authenticity and the detection of adulteration have become major issues in the food industry and are attracting an increasing amount of attention from wine producers, consumers and researcher as the quality of wine has an obvious commercial value [251, 252]. Hence, a number of multivariate data analysis techniques, including PCA, CA, LDA, PLS-DA, SIMCA, ANN and PLS have been employed classify wines (Table 3.1).

Classification of wines according to geographic origin, variety and vintage have been successfully achieved by measuring chemical compounds present in the wine matrix, such as phenolic compounds [255-262], minerals [263-265], volatile compounds [266-270] and amino acids and amines [271, 272].

Phenolic compounds are particularly important components and their composition in wine depends on the grape variety, vineyard location, and ageing, among many other factors. They contribute to their sensorial properties, being responsible for red wine colour, flavour, astringency and bitterness, both directly and indirectly through interaction with proteins, polysaccharides or other phenolic compounds. In addition to contributing to the olfactory profile of the wine, phenolic acids are precursors of volatile phenols, which enrich wines with varying aromas. They are also responsible for browning reactions of the wine and are considered to be essential elements during

preservation and ageing [273]. Phenolic compounds and commonly determined by HPLC with UV detection [257, 262, 274-277].

Chemiluminescence has been established as a valuable detection technique with advantages including low limits of detection, wide linear dynamic ranges and speed of response [278]. Chemiluminescent detection methods have recently been employed to determine phenolic compounds in wine [279-282]. Phenolic compounds have been shown to be particularly sensitive towards acidic potassium permanganate chemiluminescence reactions, with polyphenol's generally producing a greater response than simple phenols [283]. Costin et al. [278] successfully employed acidic potassium permanganate chemiluminescence detection to monitor the total phenolic/antioxidant levels in wine. It was found that acidic potassium permanganate chemiluminescence detection was selective with minimal interferences observed from non-phenolic components in the wine. The inherent selectivity for phenolic compounds makes acidic potassium permanganate an excellent chemiluminescent reagent for the study of polyphenol variations between wines.

In this chapter HPLC with acidic potassium permanganate chemiluminescence detection is evaluated for the analysis of Australian wines from different geographic origins and vintages. LDA and QDA are compared for the classification of red and white wines according to geographic origin. PLS and PCR are also examined for the modelling of sample composition with wine age.

Table 3.1: Some recent examples of chemometric classification of wines according to geographic origin, variety and vintage

Study	**Method**	**Chemometric technique**	**Reference**
Geographic discrimination of wines	GC-MS	PCA, LDA, PLS-DA	[153]
Classification of Riesling wines from different countries	Visible and NIR spectroscopy	PCA, LDA, PLS-DA	[154]
Geographic classification of Spanish and Australian Tempranillo red wines	Visible and NIR spectroscopy	PCA, LDA, PLS-DA	[251]
Geographic classification of Italian wines	Flame atomic absorption and emission spectrophotometry	PCA, CA, LDA, SIMCA	[264]
Geographic classification of young red wines from the Canary Islands	HPLC	PCA, LDA	[273]
Geographic classification of Australian and New Zealand wines	UV, visible, near-infrared (NIR) and mid-infrared (MIR) spectroscopy	PCA, PLS-DA, SIMCA	[284]
Discrimination between Shiraz wines from different Australian regions	UV-visible, NIR and MIR spectroscopy	PCA, LDA, SIMCA	[285]
Differentiation of certified brands of origins of Spanish wines	Headspace solid-phase microextraction gas chromatography	PCA, LDA, ANN	[286]
Authentication of Italian CDO wines	Chemical analyses	SIMCA, unequal class modelling	[287]
Characterisation of the geographic origin of Italian red wines	Chemical analyses, chromatography, emission spectroscopy and NMR	PCA, CA, DA	[288]
Differentiation of Slovenian wines according to geographic origin	NMR and isotope ratio mass spectrometry (IRMS)	PCA, CA, ANN	[289]
Classification of Hungarian wines according to geographic origin, wine-making technology, grape variety and year of vintage	Ion-exchange chromatography	PCA, LDA	[272]

Study (cont.)	**Method (cont.)**	**Chemometric technique (cont.)**	**Reference (cont.)**
Determination of origin and vintage of Slovenian wines	NMR and IRMS	PCA, LDA	[290]
Differentiation and classification of wines according to origin, grape variety and ageing process	UV-vis spectroscopy	PCA, SIMCA	[291]
Varietal discrimination of red and white wines	MIR spectroscopy	PCA, LDA	[155]
Classification of Australian white wines according to varietal origin	MS-based electronic nose	PCA, LDA, PLS-DA	[292]
Discrimination of Australian white wines according to varietal origin	Visible and NIR spectroscopy	PCA, LDA, PLS-DA	[293]
Classification of rice wines according to ageing time	HPLC	PCA, PLS-DA	[294]
Determination of the age of sherry wines	Gas and liquid chromatography	PLS, multiple linear regression	[295]
Classification of wines according to vintage year	Chemical analyses	PCA, LDA	[296]

3.2 Experimental

3.2.1 Samples

Geographic

Finished high-quality Cabernet Sauvignon (*n* = 34) and Chardonnay (*n* = 22) wines were collected from the Geelong (Cabernet Sauvignon, n = 10, Chardonnay, n = 11) and Coonawarra (Cabernet Sauvignon, n = 24, Chardonnay, n = 11) wine regions of Australia with a total of 21 different wineries (Punters Corner Wines, Wynns Coonawarra Estate, Brand's of Coonawarra, Balnaves of Coonawarra, Wingara Wine Group, Hollick Wines Pty Ltd, Rymill Winery, Majella Wines, Di Giorgio Family Wines Pty Ltd, Flint's of Coonawarra, Redman Wines, Leconfield, Moorabool Estate, Provenance Wines, Scotchmans Hill, Bannockburn Vineyards, Pettavel Pty Ltd, Eagles Rise, Clyde Park Vineyard, and Lethbridge Wines) involved in the study [297, 298].

Vintage

Finished high-quality Cabernet Sauvignon (*n* = 17) wines were collected from the Pirramimma Winery in the McLaren Vale wine region of Australia from the 1971 to 2003 vintages [297, 299].

Both the geographic and vintage wine samples were stored in centrifuge tubes at -18^{o}C until required. For analysis the samples were thawed, equilibrated at room temperature and mixed thoroughly, using a vortex mixer, prior to filtering through a 0.45 μm nylon membrane filter (Acrodisc PSF syringe filters; Pall Australia, VIC, Australia). All wines were sampled in duplicate [297-299].

3.2.2 Chromatographic analysis

Chromatographic separations were performed by employing a Hewlett Packard 1100 series high performance liquid chromatograph equipped with a quaternary pump, solvent degasser, autosampler (Agilent Technologies, Forest Hill, VIC, Australia) and a DAD (1200 DAD, Agilent Technologies). The HPLC was fitted with a Chromolith Performance RP-18e 100 × 4.6 mm column and a 5 mm monolithic guard column

(Merck, Kilsyth, VIC, Australia). The eluent from the DAD (254 nm) was merged post-column with a chemiluminescent reagent (section 3.2.4) prior to the chemiluminescence detector. Thus the column eluent was propelled sequentially through each detector. The HPLC pump, DAD and data acquisition from the chemiluminescence detector were controlled using Hewlett Packard Chemstation Software (Agilent Technologies). Wines were analysed by injecting 20 μL aliquots of the samples and separated at a flow rate of 3 mL min^{-1}. Solvent composition of 3% methanol in an aqueous solution of trifluoroacetic acid (0.1% v/v) was increased to 30% methanol over 12 minutes, which was then raised to 70% methanol for a further 10 minutes [297-299].

3.2.3 Mass spectrometry

Characterisation of the detected, prominent wine constituents was gained with the aid of high-resolution mass spectrometry. In order to achieve an optimum negative ion signal the chromatography was performed without trifluoroacetic acid. The sample stream was split post chromatographic separation with a portion (50%) directed to the mass spectrometer in order to reduce overloading the mass spectrometer. A 6210 MSDTOF mass spectrometer (Agilent Technologies) was used with the following conditions: drying gas, nitrogen (7mL min^{-1}, 350°C); nebulizer gas, nitrogen (16 psi); capillary voltage, 4.4 kV; vaporizer temperature, 350°C; and cone voltage, 60 V. The MS was calibrated using a standard tuning mix (G2421A, Agilent Technologies) [297-299].

3.2.4 Chemicals

HPLC grade methanol was obtained from BDH (Poole, UK). All mobile phases were filtered through a 0.45 μm nylon membrane filter. The permanganate chemiluminescence reagent was made by dissolving potassium permanganate (5×10^{-4} M; Ajax, Auburn, NSW, Australia) in a 1% (w/v) sodium polyphosphate (Sigma-Aldrich, Castle Hill, NSW, Australia) solution and the pH adjusted to 2.0 using sulfuric acid (Rhone-Poulenc, Melbourne, VIC, Australia) [297-299].

3.2.5 Data pre-processing and analysis

All data manipulation and analysis algorithms were implemented using Matlab (V7.10 (R2010a), MathWorks Inc, MA, USA), the PLS Toolbox (V4.0.2, Eigenvector Research Inc., WA, USA) and in-house developed algorithms.

Chemometric methods were applied to raw chromatographic data to preserve all the important information. The chromatograms were smoothed using an 11 point quadratic filter, aligned using COW [236] and normalised to a constant total area (constant sum). The COW parameters selected by the optimisation algorithm for the geographic and vintage data are given in Table 3.2. Despite the fact that *i*coshift was shown to be the more accurate alignment method in Chapter 2, COW is used in this chapter as the chemiluminescence detector results in non-systematic peak shifts and baseline drift due to the solvent gradient.

Table 3.2: Optimised segment lengths and slack sizes for COW alignment

	Segment length	**Slack size**
Geographic (Cabernet Sauvignon)	153	1
Geographic (Chardonnay)	106	1
Vintage	154	4

PCA was employed as both an EDA technique and a pre-processing method in order to reduce the dimensionality of the chromatographic data. PCA was performed on mean-centred data.

LDA and QDA were applied to the PCs to classify the Cabernet Sauvignon and Chardonnay wines according to geographic origin. The classification models were developed and validated using leave-one-out cross-validation and the results displayed in the form of a confusion matrix.

PLS and PCR were applied to data from the Pirramimma wines in order to construct a regression model to correlate the identified chromatographic peaks with wine age.

3.3 Results and discussion

3.3.1 Classification of Cabernet Sauvignon wines according to geographic origin

Average chromatograms of Cabernet Sauvignon from the two wine growing regions are shown in Figure 3.1. The average chromatograms were generated by averaging the aligned chemiluminescence traces. Peak assignments are provided in Table 3.3. A full discussion/interpretation of MS results is provided in [297].

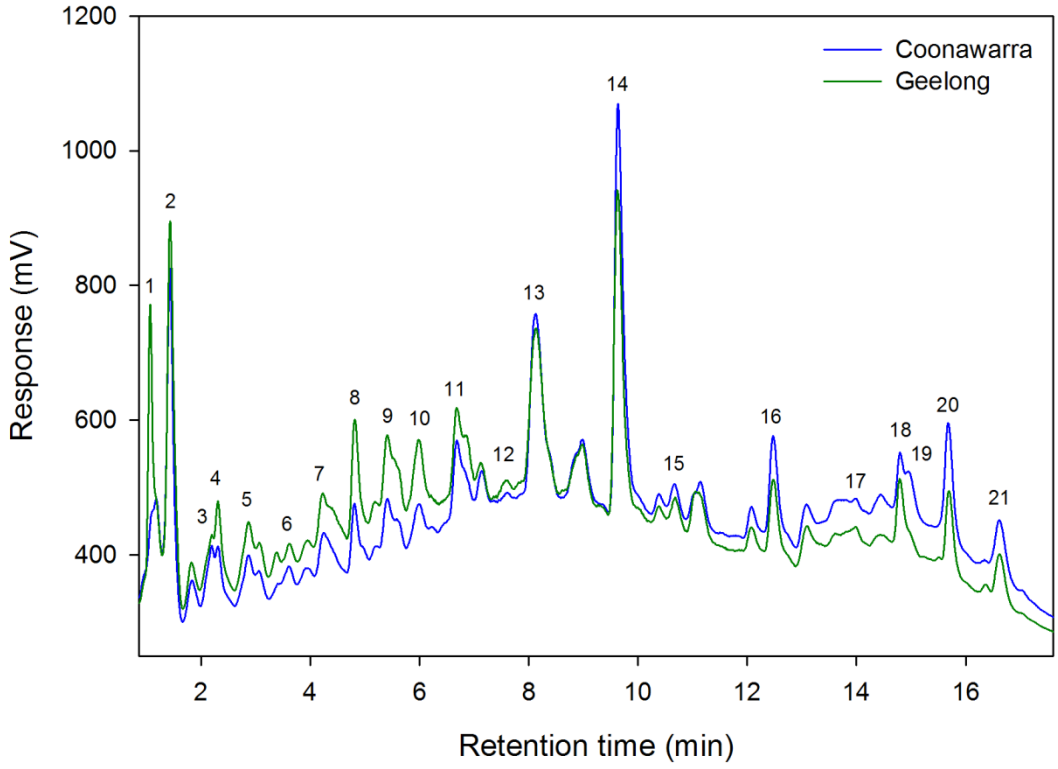


Figure 3.1: Average aligned chromatograms of Cabernet Sauvignon wines from the Coonawarra and Geelong regions. Peak assignments are provided in Table 3.3.

Variation can be observed between the wine varieties from the two regions. In general the intensity of the chromatographic peaks is higher between 0 and 8 minutes for the Geelong region, with the Coonawarra wines presenting a greater intensity for the peaks after 8 minutes. Based on the reversed-phase liquid chromatography, this may indicate that the Geelong wines contain higher concentrations of smaller water-soluble phenols and fewer less soluble polyphenolic tannins than those from Coonawarra.

Table 3.3: Peak assignments

Peak	Compound	Peak	Compound
1	Cinnamic acid	12	Epicatechin
2	Tartaric acid	13	Ethyl gallate
3	Gallic acid	14	Myricetin
4	Vanillic acid	15	Syringic acid
5	Gallocatechin	16	Procyanidin B
6	Quercetin hexoside-gallate	17	Procyanidin A
7	Catechin	18	Resveratrol
8	Epigallocatechin	19	Picied
9	Coumaric acid	20	Morin
10	Caffeic acid	21	Malvidin
11	Sinapic acid		

PCA was performed as an EDA technique and as a method for reducing the dimensionality of the data by using the scores as inputs to discriminant analysis. Table 3.4 gives the eigenvalues for the first ten PCs.

Table 3.4: PCA eigenvalues for Cabernet Sauvignon

PC number	% variance	% cumulative variance
1	47.03	47.03
2	21.89	68.92
3	16.49	85.41
4	3.16	88.57
5	2.29	90.86
6	1.57	92.43
7	1.47	93.90
8	1.24	95.14
9	0.78	95.92
10	0.64	96.56

The scores plot for the first two PCs, which account for 68.92% of the variance, is plotted in Figure 3.2 (a). Clustering of the wines according to production region is evident, however there is still not complete separation. Therefore the third PC is added (16.49% of the variance) in attempt to gain complete separation between the Coonawarra and Geelong wines. The 3D scores plot of the first three PCs is shown Figure 3.2 (b).

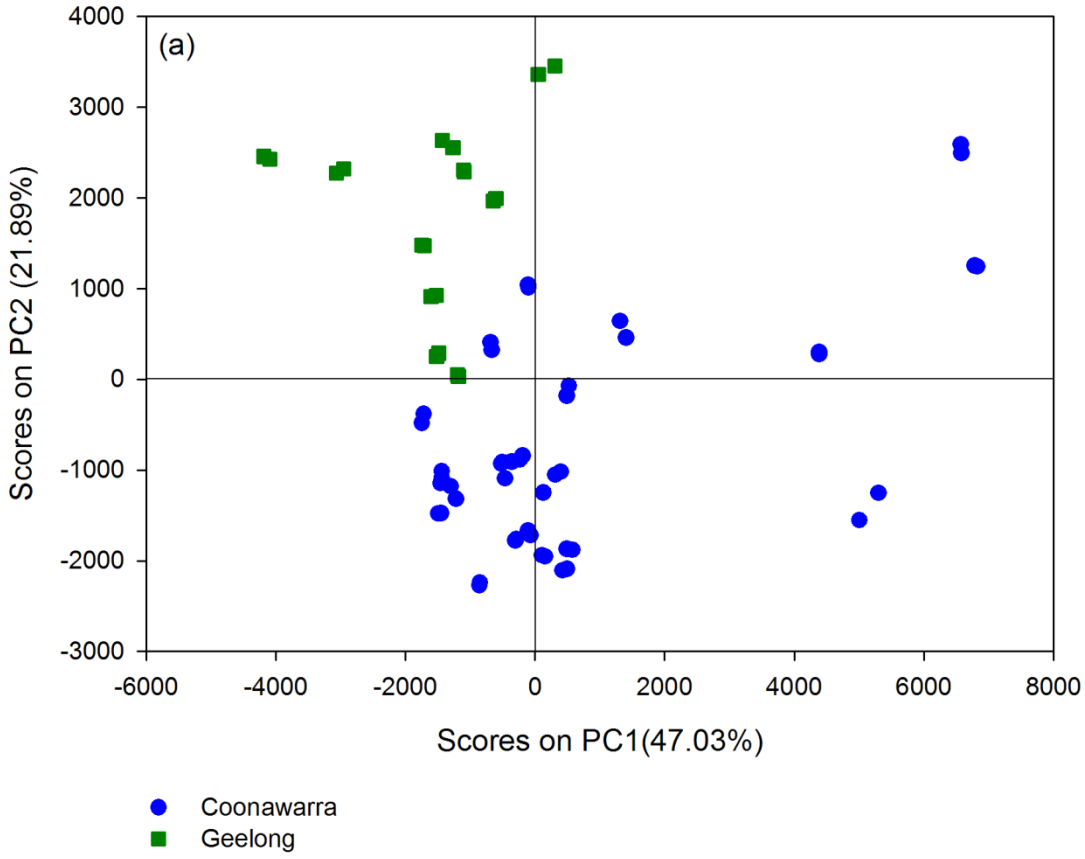


Figure 3.2: Cabernet Sauvignon scores plots (a) first two PCs

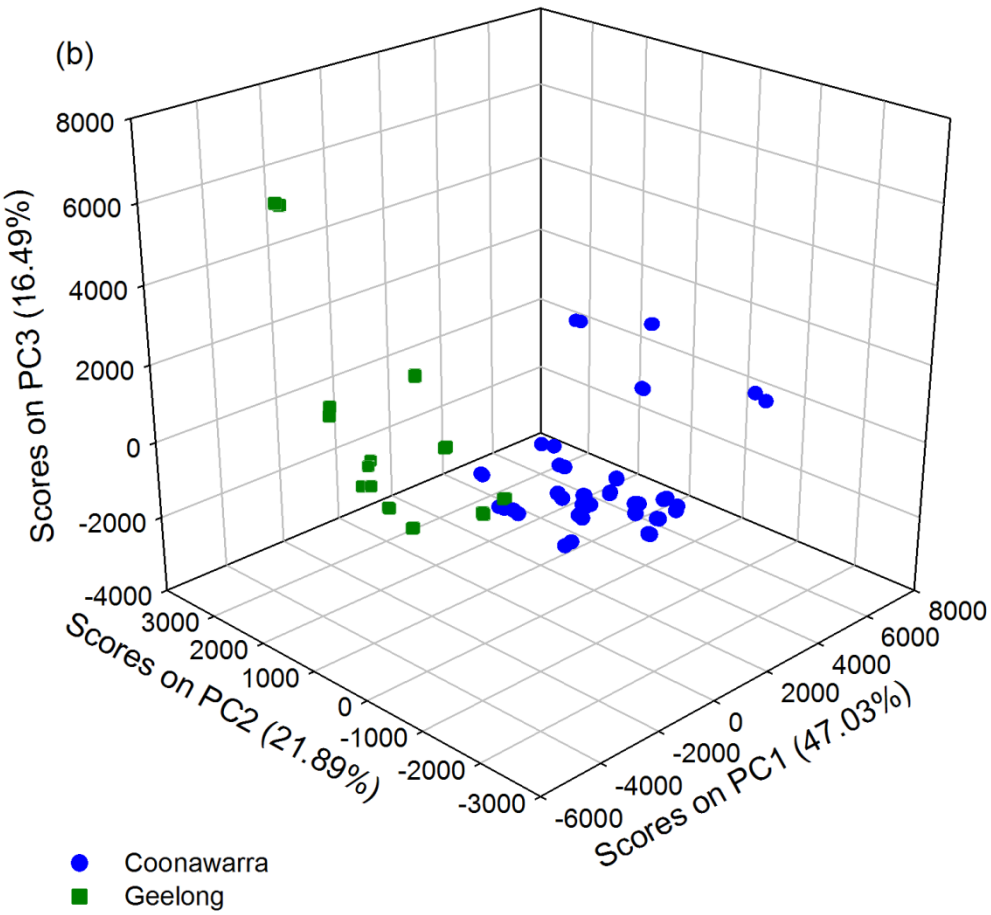


Figure 3.2: Cabernet Sauvignon scores plots (b) first three PCs

There is no noticeable improvement in the separation between the Coonawarra and Geelong wines when the third PC is added and since visualisation and interpretation is simpler using two PCs, only the first two PCs are used for the discussion. The loadings associated with the first two PCs are shown in Figure 3.3.

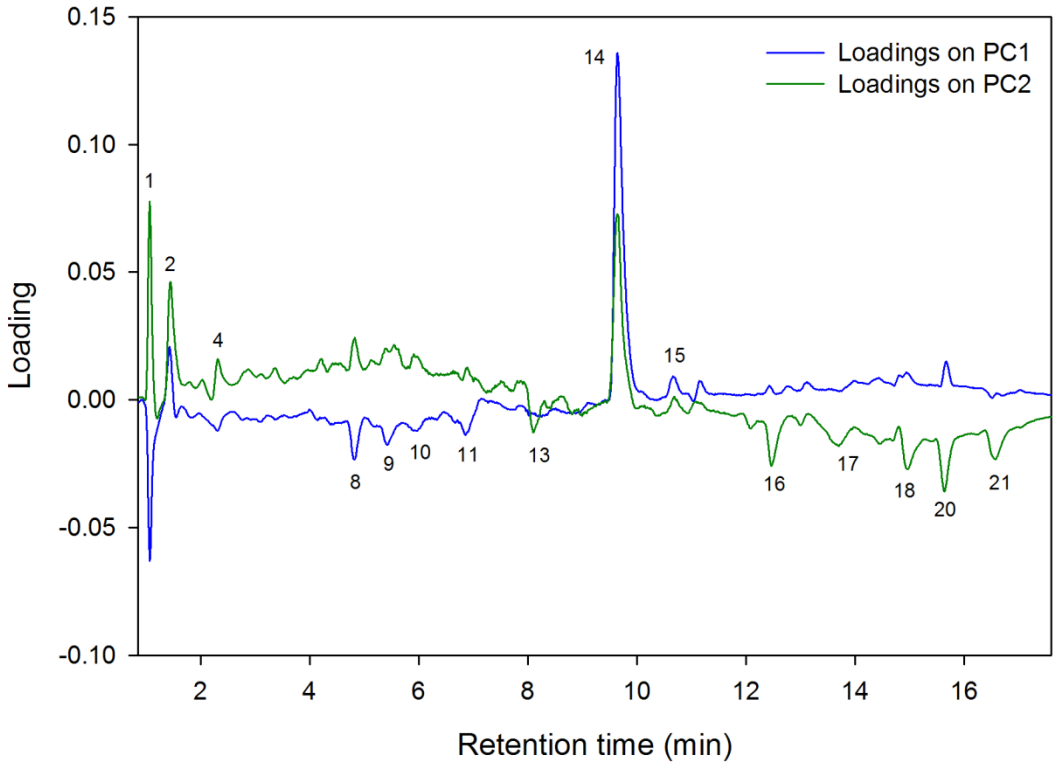


Figure 3.3: Cabernet Sauvignon loadings for the first two PCs. Peak assignments are provided in Table 3.3.

The loading plots highlight cinnamic acid (peak 1), tartaric acid (peak 2) and myricetin (peak 14) as significant compounds in the classification of these red wines. The levels of myricetin have been found to differ according to geographic origin [258, 300] and since tartaric acid is one of the most common acids found in red wine, it will play a role in the classification.

Discrimination between the Coonawarra and Geelong regions was obtained using LDA and QDA. Discriminant analysis requires that the number of variables is less than the number of samples and that the variables are not correlated. By reducing the data with PCA, these two important criteria for robust discriminant analysis are satisfied [155]. Since two PCs adequately separated the Coonawarra and Geelong wines in the scores plot (Figure 3.2 (a)), LDA and QDA were performed using the first two PCs and the results are summarised as a confusion matrix (Table 3.5).

Interpretation of the LDA confusion matrix (Table 3.5 (a)) is as follows; for the total 48 Coonawarra wines, 46 of them were correctly classified as Coonawarra, while 2 of them were incorrectly classified as Geelong wines giving an accuracy of 96%. For the 20 Geelong wines, 18 of them were correctly classified as Geelong and 2 of them were incorrectly classified as Coonawarra, to give an accuracy of 90%. The overall accuracy of 94% for the LDA classification was calculated according to Equation 3.1.

$$A_{overall} = \frac{\sum_{i=1}^{k} n_{ii}}{n} \qquad \text{Equation 3.1}$$

Where n_{ii} is number of samples correctly classified in each class (diagonal elements of the matrix) and n is the total number of samples in the matrix.

QDA (Table 3.5 (b)) provided the same overall accuracy (94%) as LDA. However, QDA correctly classified all 20 of the Geelong wines, but could only correctly classify 44 of the 48 Coonawarra wines.

Despite having the same overall accuracy, the classification results of LDA and QDA differ. This is due to LDA employing a linear discrimination boundary, while QDA uses a parabola. The results suggest that the parabolic boundary employed by QDA improves the classification of the Geelong wines, but in doing so, the correct classification of the Coonawarra wines is reduced.

Table 3.5: Cabernet Sauvignon discriminant analysis results using the first two PCs (a) LDA and (b) QDA

(a)

	Coonawarra	**Geelong**	**Total**	**Accuracy**
Coonawarra	46	2	48	96%
Geelong	2	18	20	90%
Overall accuracy				94%

(b)

	Coonawarra	**Geelong**	**Total**	**Accuracy**
Coonawarra	44	4	48	92%
Geelong	0	20	20	100%
Overall accuracy				94%

3.3.2 Classification of Chardonnay wines according the geographic origin

Average chromatograms of Chardonnay from the two wine growing regions are shown in Figure 3.4. The average chromatograms were generated by averaging the aligned chemiluminescence traces. Peak assignments are provided in Table 3.3.

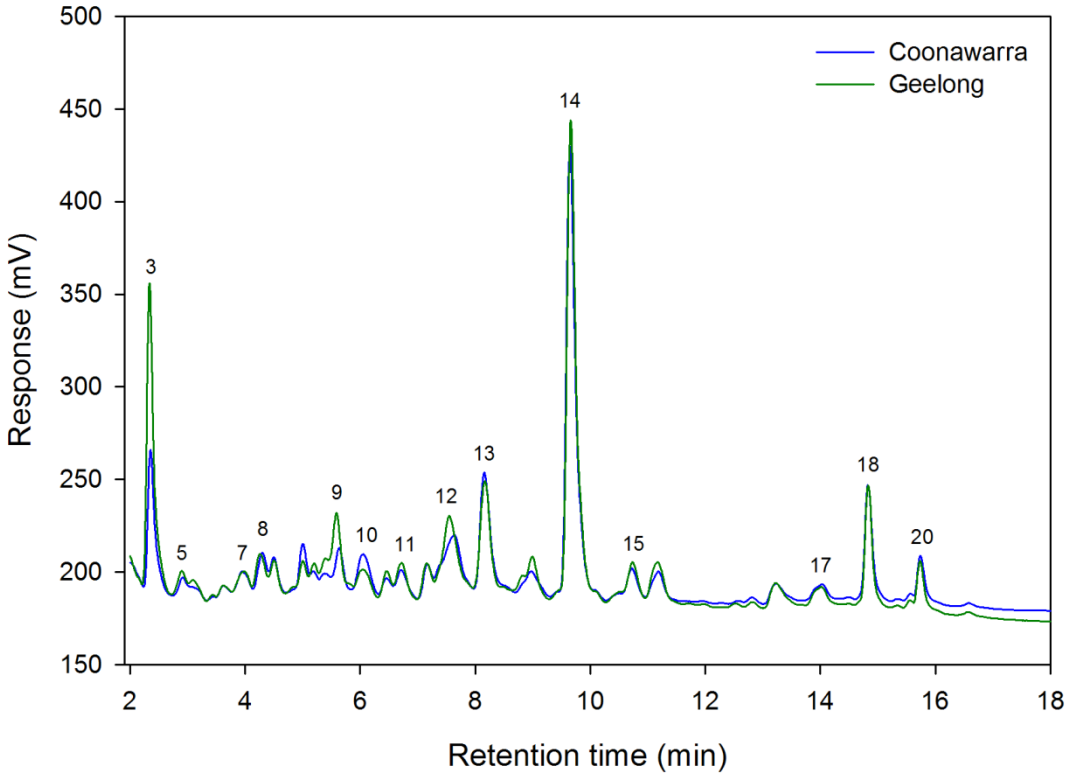


Figure 3.4: Average aligned chromatograms of Chardonnay wines from the Coonawarra and Geelong regions. Peak assignments are provided in Table 3.3.

Variation between Chardonnay from the Coonawarra and Geelong wine growing regions is not as evident as with the Cabernet Sauvignon, however changes in concentration of the simple phenolic acids, gallic acid (peak 3), coumaric acid (peak 9) and epicatechin (peak 12) can be observed, all of which are higher in the Geelong wines. The concentrations of these species have been shown to vary with geographic origin [301].

As with the Cabernet Sauvignon, PCA was employed as an EDA technique and to reduce the dimensionality of the data. Table 3.6 gives the eigenvalues for the first ten

PCs. Since most of the variance is explained in the first two PCs (52.74% and 29.39% respectively), they are used for PCA.

Table 3.6: PCA eigenvalues for Chardonnay

PC number	**% variance**	**% cumulative variance**
1	52.74	52.74
2	29.39	82.13
3	4.33	86.46
4	3.35	89.81
5	2.50	92.31
6	1.76	94.07
7	1.53	95.60
8	1.03	96.63
9	0.70	97.33
10	0.55	97.88

The scores plot for the first two PCs is shown in Figure 3.5 (a). Clustering of the wines according to geographic origin is evident, however there is some overlap between the two classes. Therefore the third PC is added (4.33% of the variance) in attempt to gain complete separation between the Coonawarra and Geelong wines. The 3D scores plot of the first three PCs is shown Figure 3.5 (b).

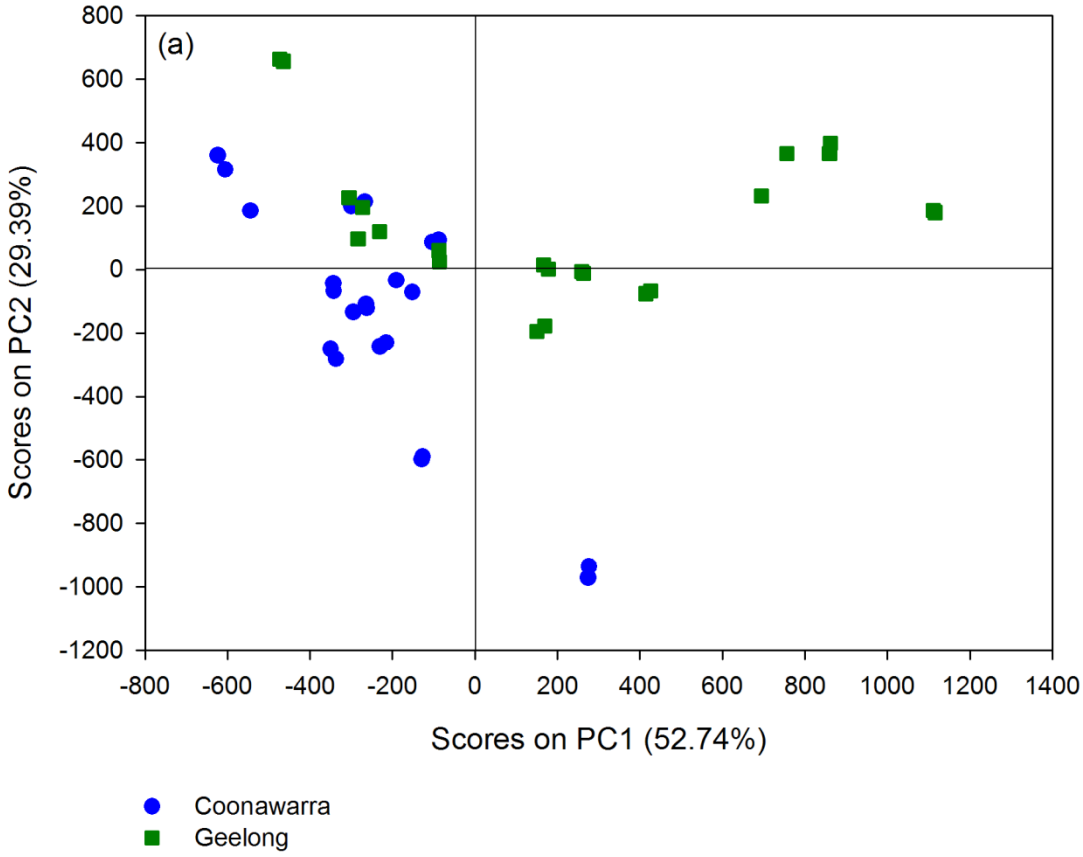


Figure 3.5: Chardonnay scores plots (a) first two PCs

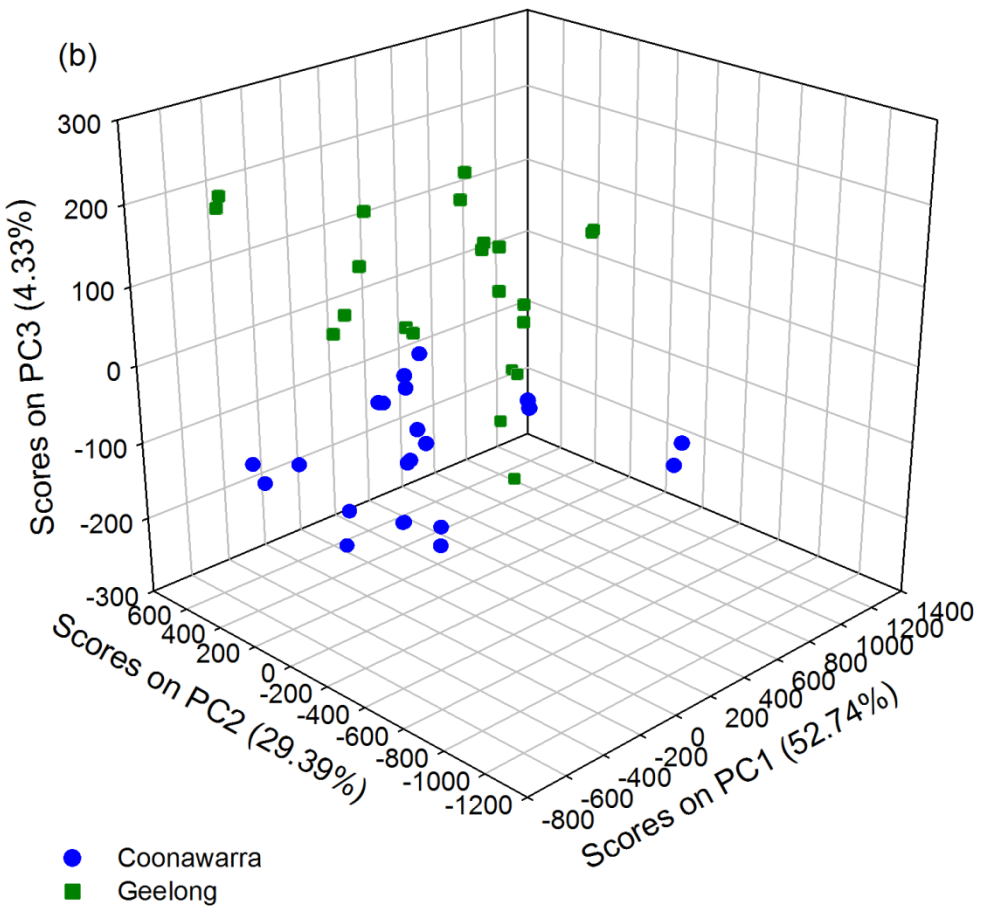


Figure 3.5: Chardonnay scores plots (b) first three PCs

As with the Cabernet Sauvignon wines, separation between the Coonawarra and Geelong wines is not really improved by adding in the third PC, thus only the first two PCs are used for the discussion. The loadings associated with the first two PCs are plotted in Figure 3.6.

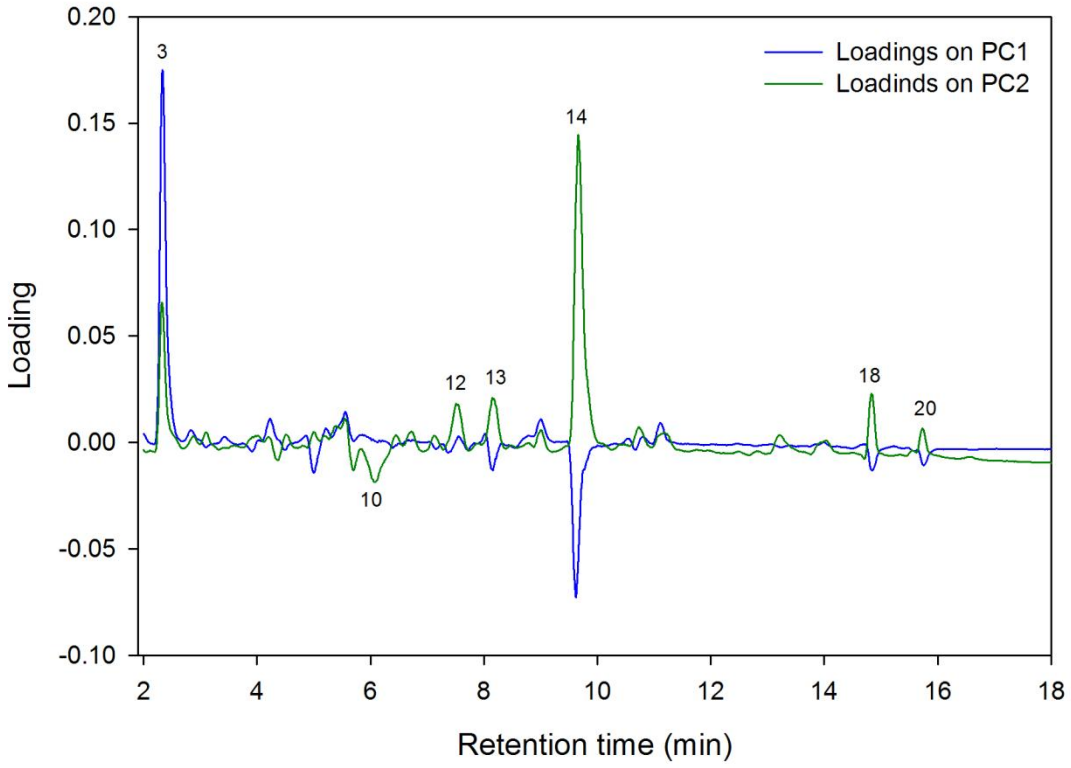


Figure 3.6: Chardonnay loadings for the first two PCs. Peak assignments are provided in Table 3.3.

The loading plots highlight gallic acid (peak 3) and myricetin (peak 14) as significant compounds for the geographic classification of these white wines. Gallic acid has been shown to vary according to geographic origin [258]. Although myricetin is more commonly associated with red wines [302], it contains many hydroxyl moieties making it an ideal candidate for reactivity with acidified potassium permanganate. The chemical structure of myricetin is similar to that of quercetin (myricetin has an extra hydroxyl moiety), which has been shown to have an exceptionally high response to permanganate chemiluminescence [278]. This high sensitivity to chemiluminescence detection enables myricetin to play a significant role here.

LDA and QDA were used to discriminate the Chardonnay wines according to production region. As with Cabernet Sauvignon, PCs were used as inputs to discriminant analysis. Since the Geelong and Coonawarra wines were adequately separated by two PCs in the scores plot (Figure 3.5 (a)), LDA and QDA were performed using the first two PCs and the results are summarised in Table 3.7.

The overall accuracy of QDA was higher, 82% compared with 77% for LDA. LDA correctly classified 20 of the 22 Coonawarra wines, however only 14 of the 22 Geelong wines were able to be correctly classified. On the other hand, QDA was able to correctly classify 16 of the 22 Geelong wines and 20 of the 22 Coonawarra wines. These results suggest that parabolic discriminating boundary in QDA was more accurate at classifying the Geelong wines, while maintaining the same accuracy as the linear boundary for classifying the Coonawarra wines.

Table 3.7: Chardonnay discriminant analysis results using the first two PCs (a) LDA and (b) QDA

(a)

	Coonawarra	**Geelong**	**Total**	**Accuracy**
Coonawarra	20	2	22	91%
Geelong	8	14	22	64%
Overall accuracy				77%

(b)

	Coonawarra	**Geelong**	**Total**	**Accuracy**
Coonawarra	20	2	22	91%
Geelong	6	16	22	73%
Overall accuracy				82%

3.3.3 Regression analysis of wine vintage

The average chromatogram from the 33 chromatograms (representing the 17 Cabernet Sauvignon wines) is shown in Figure 3.7. The average chromatograms were generated by averaging the aligned chemiluminescence traces. Peak assignments are provided in Table 3.3.

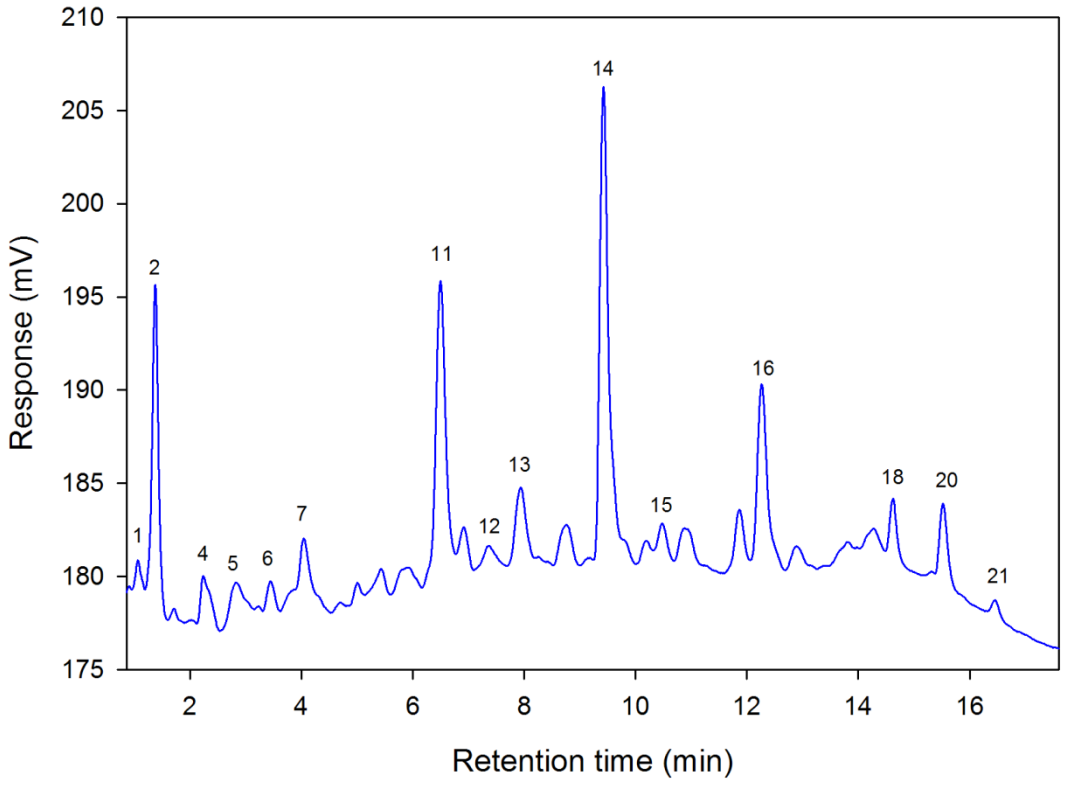


Figure 3.7: Average aligned chromatogram of the 33 chromatograms (representing the 17 Cabernet Sauvignon wines). Peak assignments are provided in Table 3.3.

PCA was employed as an EDA technique, Figure 3.8 (a) shows the scores plot for the first two PCs and separation between wines produced in the 1970's, 1980's, 1990's and 2000's is evident. The loading plots for the first two PCs are provided in Figure 3.8 (b), tartaric acid (peak 2), catechin (peak 7), sinapic acid (peak 11) and myricetin (peak 14) were identified as significant compounds for the separation of wines produced in the 1970's, 1980's, 1990's and 2000's.

Wine vintage is strongly influenced by tartaric acid [303]. Goldberg et al. [304] found catechin levels to vary according to vintage and geographic location, with Pinot Noir

varying by 30% in the USA, while remaining stable between vintages in the Beaujolais and Burgundy regions of France. Myricetin content was found to vary between vintages grown in several countries by McDonald et al. [300]. Variations included 7% in Australian Cabernet Sauvignon, 40% in USA and Bulgarian Cabernet Sauvignon, 50% in Chilean Merlot, 34% in USA Merlot, and 17% in French Pinot Noir. Similar results were found for a range of Bulgarian [305] and Italian [306] wines.

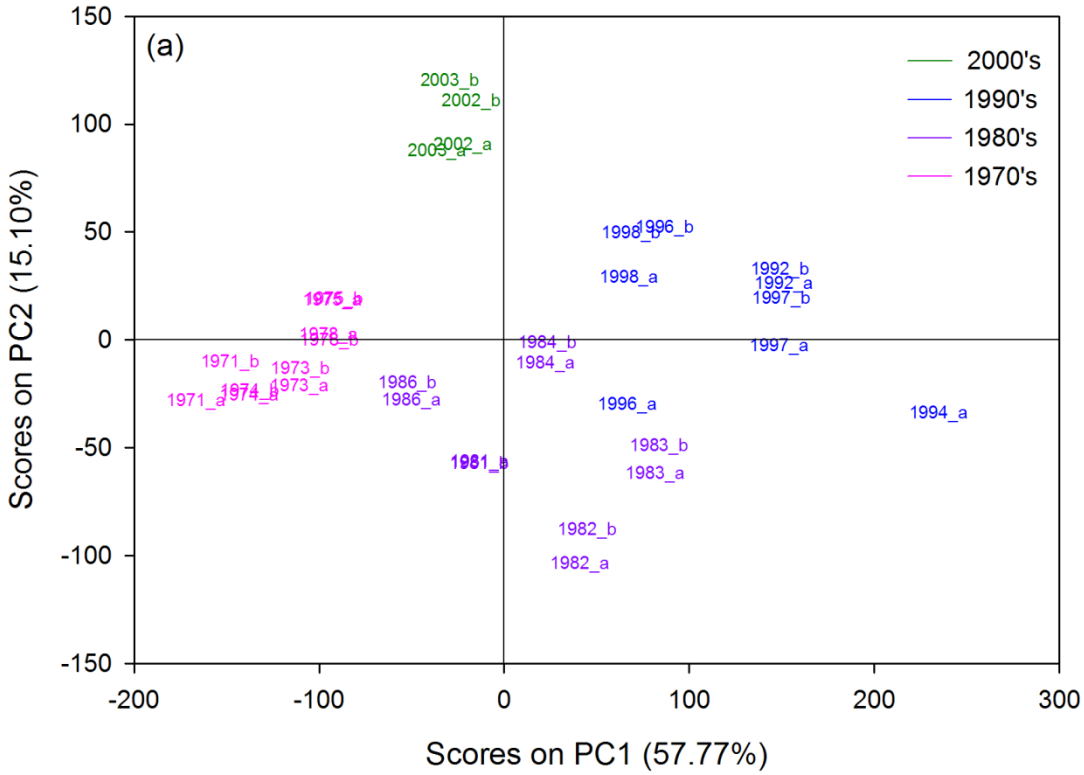


Figure 3.8: Vintage data PCA results for the first two PCs (a) scores plot, where “a” and “b” refer to the duplicate analyses

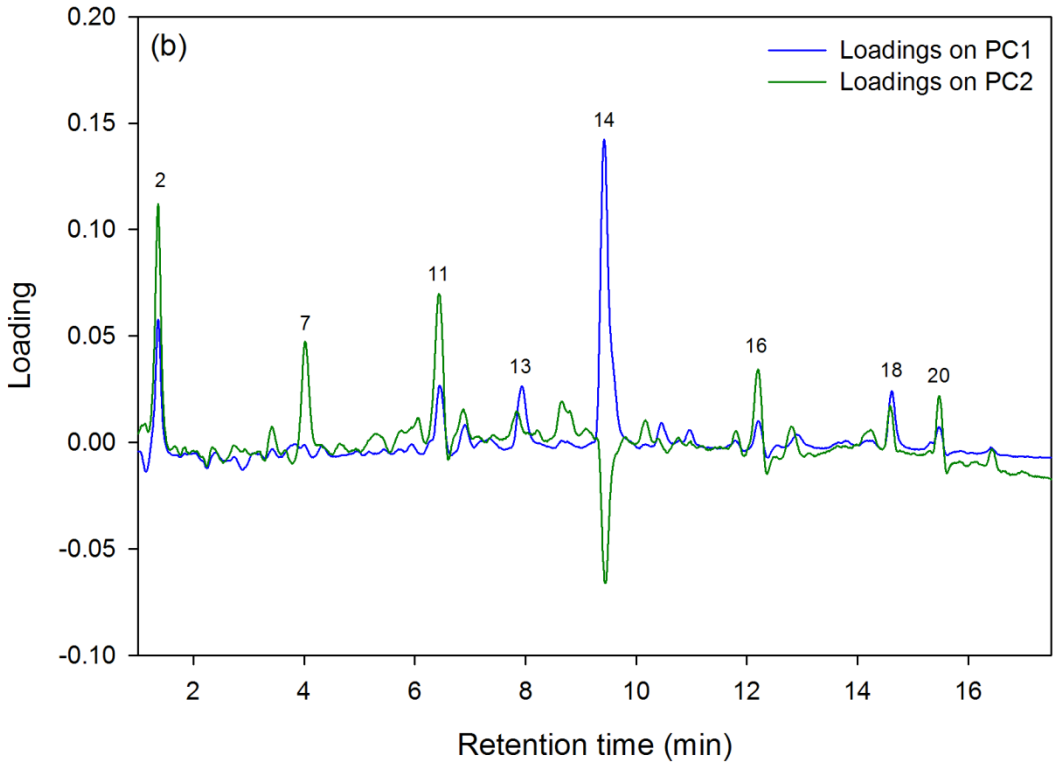


Figure 3.8: Vintage data PCA results for the first two PCs (b) loadings plot. Peak assignments are provided in Table 3.3.

A model of the relationships between sample composition and age was developed by PLS and PCR. Prior to developing the model, the number of components in the model must be selected. A simple way of doing this is by looking at the variance explained by the model, given in Table 3.8. Figure 3.9 shows the percentage of cumulative variance explained in the dependent variable (Y) for both the PLS and PCR models.

Table 3.8: PLS and PCR latent variables (X-block is the variance explained in the independent variable and Y-block is the variance explained in the dependent variable)

	PLS				**PCR**			
	X-block		**Y-block**		**X-block**		**Y-block**	
Component	**% variance**	**% cumulative variance**	**% variance**	**% cumulative variance**	**% variance**	**% cumulative variance**	**% variance**	**% cumulative variance**
1	55.07	55.07	61.58	61.58	57.77	57.77	40.65	40.65
2	17.52	72.59	24.43	86.01	15.10	72.87	39.32	79.97
3	4.56	77.15	7.95	93.96	8.99	81.86	0.81	80.78
4	5.57	82.72	2.83	96.79	6.48	88.34	1.65	82.43
5	5.93	88.65	0.90	97.69	3.41	91.75	0.93	83.36
6	3.24	91.89	0.89	98.58	2.21	93.96	6.42	89.78
7	2.52	94.41	0.33	98.91	1.77	95.73	1.35	91.13
8	1.27	95.68	0.28	99.19	1.16	96.89	3.75	94.88
9	1.14	96.82	0.21	99.40	0.64	97.53	2.62	97.50
10	0.84	97.66	0.19	99.59	0.53	98.06	0.13	97.63

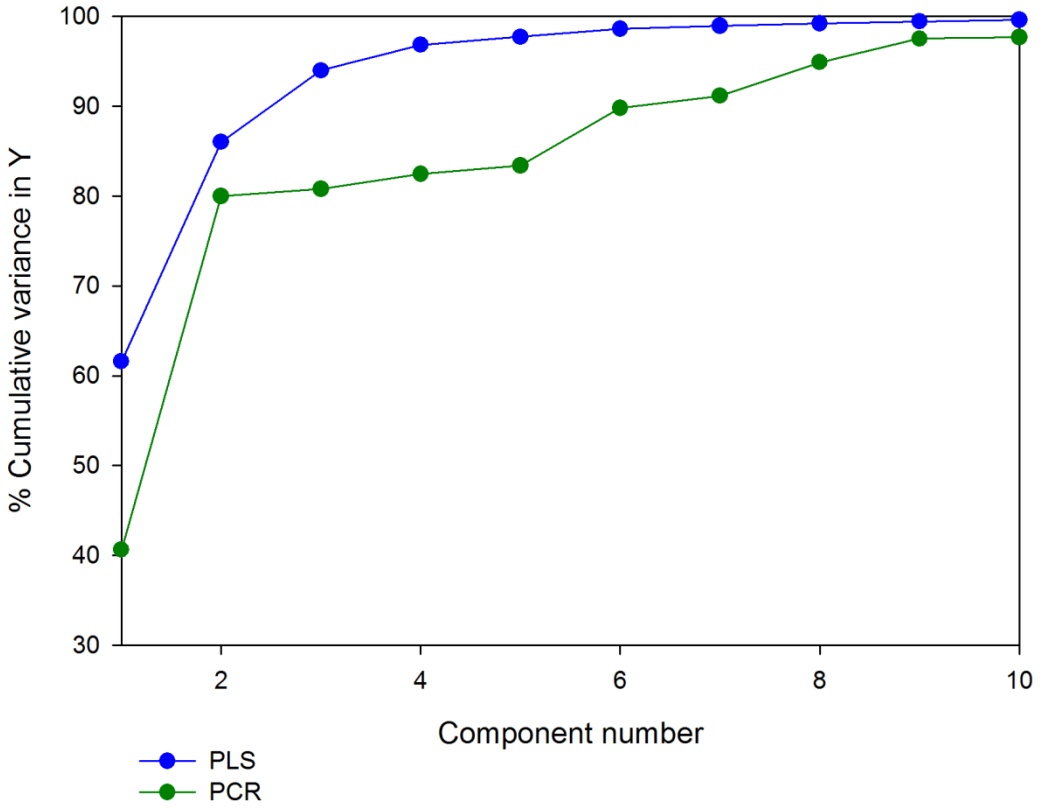


Figure 3.9: Percent cumulative variance explained in the dependent variable for PLS and PCR

Figure 3.9 suggests that 6 components explain enough of the variance as this is where it starts to plateau. The variance explained in Y is lower for PCR as the model is constructed to explain the independent variable (X), rather than Y; this can be seen in Table 3.8 as more variance is explained in X by PCR compared with PLS. Figure 3.10 (a and b) shows the predicted age versus actual age for the 6 component PLS and PCR models, respectively. From this it can be seen that PLS is more accurate at fitting the dependent variable (age) than PCR. This is also confirmed by considering the R^2 values, 0.9858 and 0.8978 for PLS and PCR, respectively.

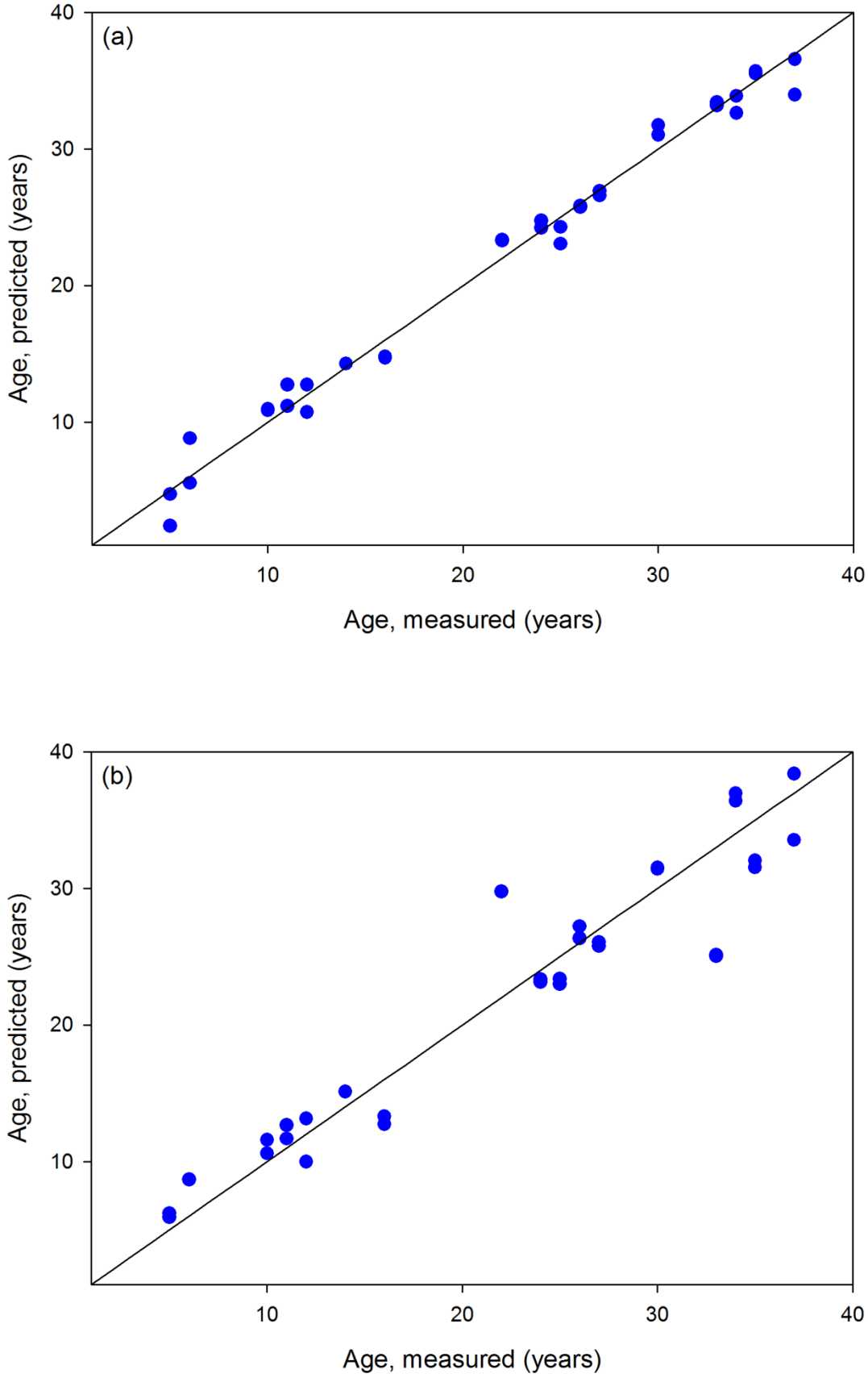


Figure 3.10: Predicted age versus measured age for a 6 component model (a) PLS and (b) PCR

A more sophisticated method for selecting the number of components in a model is cross-validation as it avoids over-fitting by not reusing the same data to fit a model and to estimate prediction error. The mean squared prediction error (MSPE) is estimated from a training set and is frequently used for assessing the performance of a regression model and selecting the optimal number of components in PLS and PCR. Commonly employed internal estimators include K-fold cross-validation and leave-one-out cross-validation [307]. The K-fold cross-validation estimate is given by:

$$MSPE_{cv.K} = \frac{1}{n_T} \sum_{k=1}^{K} \sum_{i \epsilon T_k} (f_k(x_i) - y_i)^2 \qquad \text{Equation 3.2}$$

Where the training set (T) is randomly divided in K segments (T_k, $k = 1,\ldots,K$) of roughly equal size and f_k is the predictor trained on $T \backslash T_k$ [307].

The MSPE for both PLS and PCR was estimated using 10-fold cross-validation and the results are shown in Figure 3.11.

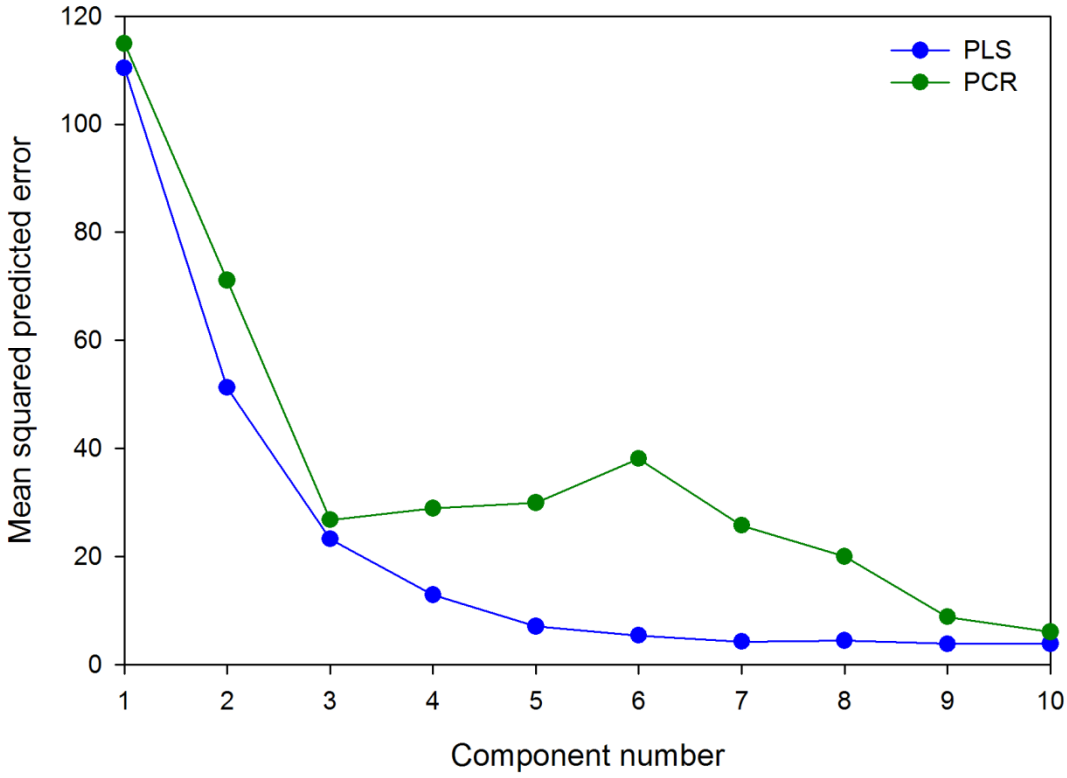


Figure 3.11: PLS and PCR MSPE curves estimated using 10-fold cross-validation

The MSPE curve for PLS regression indicates that 7 components provide sufficient prediction accuracy, as this is where the curve starts to plateau. On the other hand, PCR needs all 10 components to get the same prediction accuracy. In fact, the sixth component in PCR increases the prediction error of the model, suggesting that the combination of predictor variables contained in that component are not strongly correlated with Y, which can again be attributed to the fact that PCR constructs components to explain X, not Y. As a result the final regression models were developed using 7 and 10 components for PLS and PCR, respectively.

The PLS model using 7 components explained 98.91% of the variance in age. Figure 3.12 (a) shows the predicted age versus actual age, and good agreement is evident with an R^2 value of 0.9891. The regression vector, which represents the 7 components, is shown in Figure 3.12 (b).

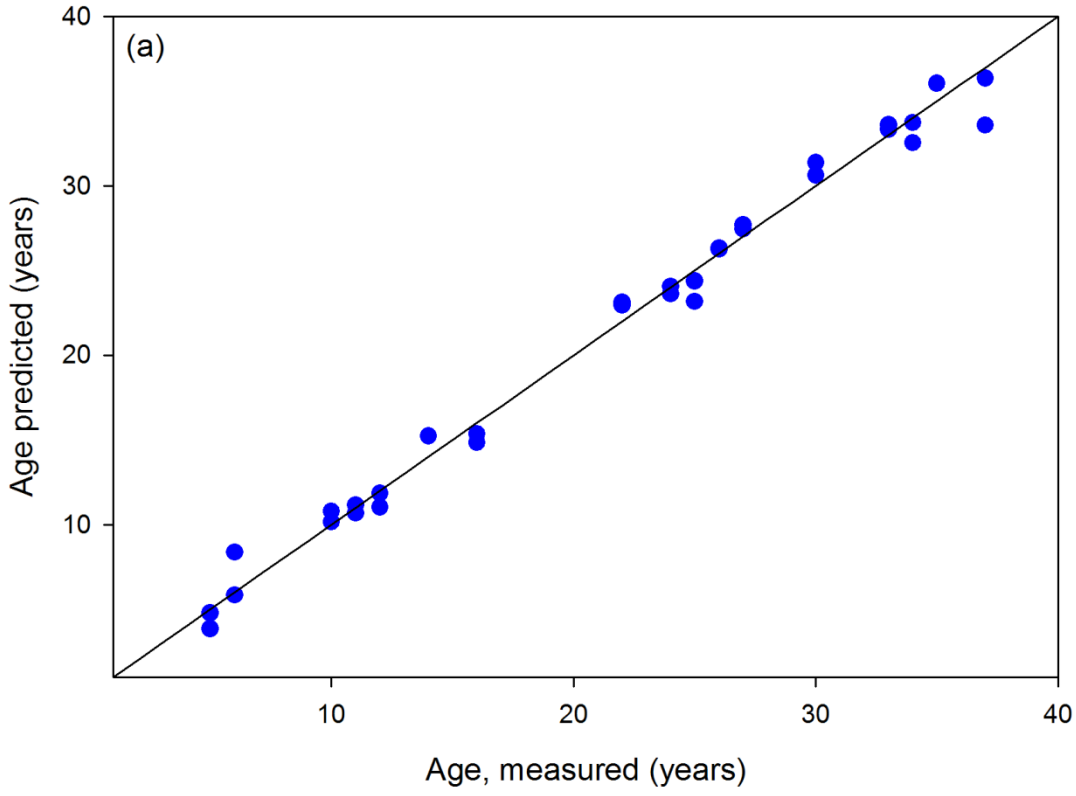


Figure 3.12: 7 component PLS model (a) predicted versus measured age

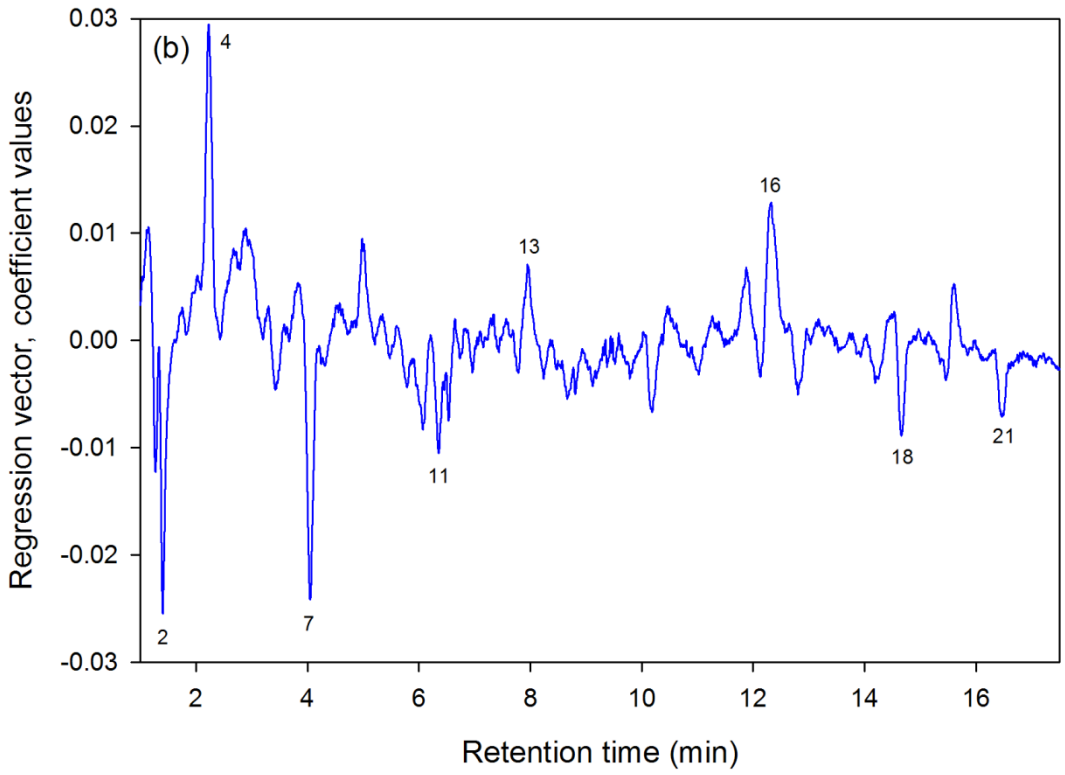


Figure 3.12: 7 component PLS model (b) regression vector. Peak assignments are provided in Table 3.3.

The PCR model using 10 components explained 97.63% of the variance in age, which is very close to the variance explained in the PLS model. Figure 3.13 (a) shows the predicted age versus actual age, and good agreement is evident with an R^2 value of 0.9763. The regression vector, which represents the 10 components, is shown in Figure 3.13 (b). Both the PLS and PCR models effectively fit the data, however PLS requires fewer components and has a slightly higher R^2 value.

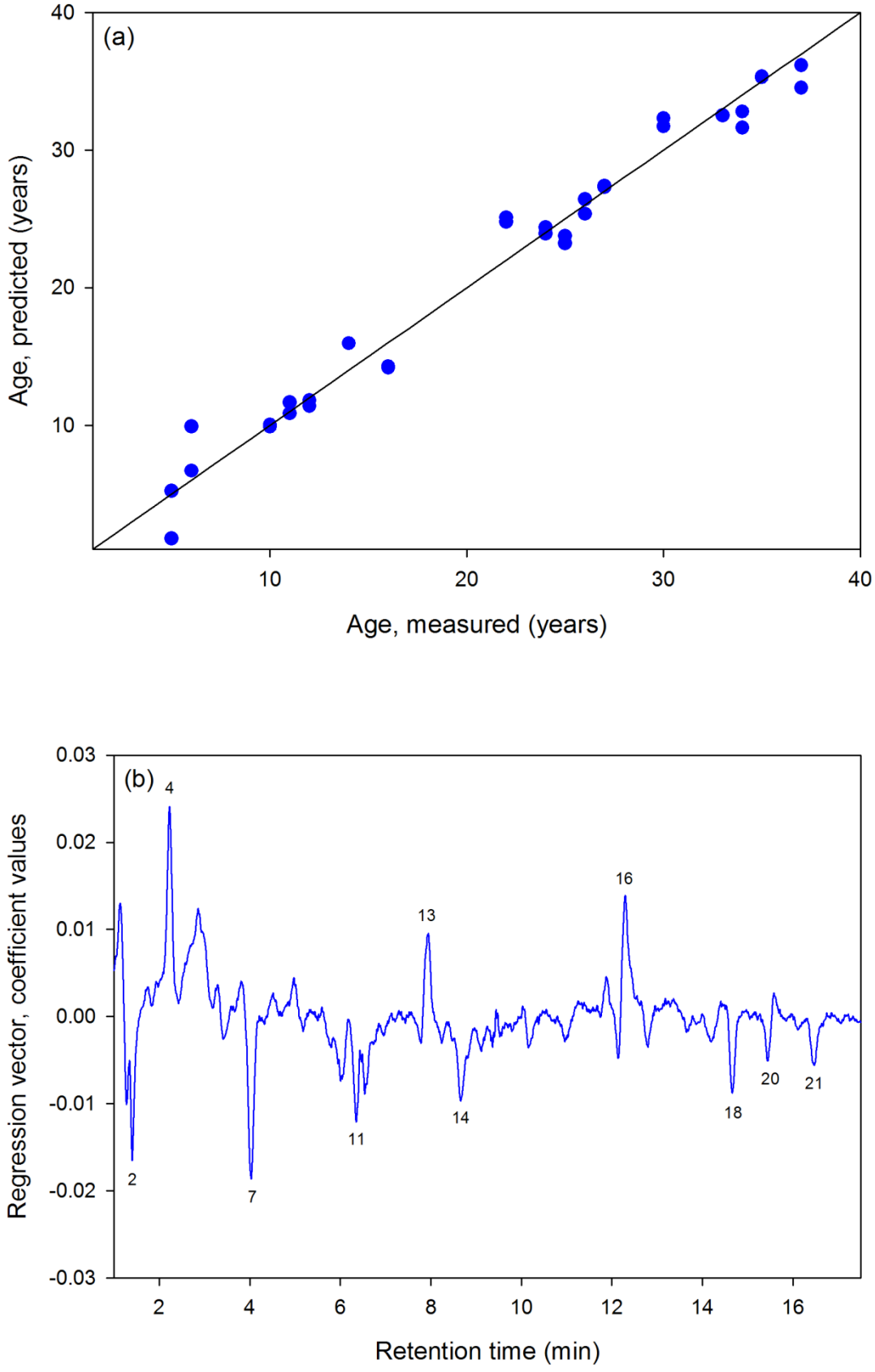


Figure 3.13: 10 component PCR model (a) predicted versus measured age and (b) regression vector. Peak assignments are provided in Table 3.3.

The regression vectors for both the 7 component PLS model and the 10 component PCR model appear similar and highlight a negative trend in tartaric acid (peak 2), catechin (peak 7), sinapic acid (peak 11), resveratrol (peak 18) and malvidin (peak 21) indicating that the concentrations of these analytes is higher in the younger wines and decrease during ageing. Vanillic acid (peak 4), ethyl gallate (peak 13) and procyanadin B (peak 16) show a positive trend and therefore increase over the ageing period and are higher in the older wines.

Kelebek at al. [308] found the content of malvidin to decrease, while procyanadin B and vanillic acid increased during ageing. Malvidin was also found to decrease by Schwartz et al. [309]. Significant decreases in catechin were observed by Chamkha et al. [310] in Pinot Noir and Gris et al. [311] in Cabernet Franc and Merlot wine varieties. Although these studies were not conducted over a long period of time such as that in this work, it could be expected that these trends would continue during ageing as found in our study.

In a study conducted over a wider range of vintages, Chira et al. [312] correlated wine age with the phenolic content of Bordeaux Cabernet Sauvignon (1978 to 2005) and Merlot (1979 to 2003) wines. It was found that phenolic compounds could discriminate both wine varieties according to vintage.

PCR also highlights a negative trend in myricetin (peak 14) and morin (peak 20). The fact that myricetin was found to be important in PCR was not unexpected as the components were calculated by PCA and myricetin was found to be significant in the PCA loadings (Figure 3.8 (b)). This highlights a significant difference between PCA and PLS. PCA is only concerned with variance and any relationship between the data and the dependent variable (age) is irrelevant. On the other hand, the relationship between X and Y is important in PLS. Thus, myricetin appears in PCA as its levels change a lot, however these changes are not correlated to age and are therefore of low importance in PLS. The reverse of this is evident for vanillic acid, where it is not even discernible in PCA, but plays an important role in PLS as it is related to age.

3.4 Conclusion

HPLC with acidic potassium permanganate chemiluminescence detection was found to be suitable for the geographic classification of Australian red and white wines. For the red wines, PCA adequately separated the Coonawarra and Geelong regions using two PCs and highlighted cinnamic acid, tartaric acid and myricetin as significant maker compounds for identification of geographic origin. LDA and QDA were employed to discriminate the wines according to geographic origin. Using two PCs, LDA and QDA had the same overall accuracy of 94%, however the number of Coonawarra and Geelong wines correctly classified by each technique were different. This was due to the discrimination boundaries employed in each technique; the parabola employed by QDA improved the classification of the Geelong wines, but in doing so, reduced the correct classification of the Coonawarra wines. For the white wines, adequate separation according to production region was achieved using PCA with two PCs; gallic acid and myricetin were identified as important compounds in terms of geographic origin. QDA was slightly better at discriminating the wines with an overall accuracy of 82% compared with 77% for LDA. This was due to the parabolic discriminating boundary in QDA being more accurate at classifying the Geelong wines, while still maintaining the same accuracy as the linear boundary for classifying the Coonawarra wines.

In the analysis of wine vintage, HPLC with acidic potassium permanganate chemiluminescence detection was again found to be suitable. PLS and PCR were compared for the modelling of sample composition and wine age. PLS required 7 components, while PCR required 10 components to achieve similar predictive ability. The PCR model required more components as it was constructed to explain the independent variable and not the dependent variable and as a result more components were need to effectively explain the dependent variable. Both methods highlighted tartaric acid, vanillic acid, catechin, sinapic acid, ethyl gallate, procyanadin B, resveratrol and malvidin as analytes that vary throughout the ageing process. PCR also found myricetin to be important; this was not unexpected as the components in PCR were calculated by PCA and myricetin was found to be significant in the PCA loadings. This highlighted a significant difference between PCA and PLS. Since PCA is only concerned with variance, any relationship between the data and the dependent variable (age) is irrelevant. On the other hand, PLS models are developed to describe

the relationship between the dependent and independent variables. Hence, myricetin appears in PCA as its levels change a lot, however these changes are not correlated to age and are therefore of low importance in PLS.

The phenolic compounds identified with acidic potassium permanganate chemiluminescence detection were found to be valuable for the analysis of wine vintage as well as discriminating the geographic origin of red and white wines.

Chapter 4 - GC×GC Quality Control Software: Data Alignment

4.1 Introduction

Quality control (QC) involves monitoring a process and eliminating causes leading to unsatisfactory performance. The scope of QC varies considerably depending on the context in which it is to be used, however all QC processes involve analysis, collection of information and the interpretation and presentation of results. In the analytical laboratory, a quality system seeks to assure the analytical results are accurate and representative of the test sample being analysed [313].

QC differs from "normal" analysis in that the sample is well characterised; its composition is known and the analysis is being performed to ensure that it is within some defined tolerance values and, hence, "fit for purpose".

The aim of the work described here is to develop software for the final data analysis phase of the QC of flavours and fragrances. The software is required to compare a new sample to a known reference material in order to accept or reject the new sample.

Flavours and fragrances are made up of numerous constituents, many of which are complex mixtures such as natural extracts or essential oils [314]. They are often characterised by the presence of many volatile components, belonging to several classes of compounds in a wide range of concentrations [315]. Since most flavour and fragrance compounds are volatile, 1D GC methods are routinely employed for QC of flavours and fragrances [316-319]. However, even when combined with identification/confirmation techniques such as MS, 1D GC generally does not provide sufficient separation power for complex qualitative or quantitative analysis. This results in the need for a greater degree of separation, hence the use of GC×GC for QC purposes has been investigated [314, 320, 321]. Since flavours and fragrances are composed of a number of chemical classes with different polarities, there is an opportunity to exploit the polarity differences of closely eluting compounds through the two separation mechanisms employed in GC×GC. Provided there is a separation

mechanism which permits their resolution on the second column, co-eluting compounds from the first column will subsequently be resolved on the second [12].

GC×GC provides advantages such as enhanced resolution and sensitivity [50] and increased peak capacity [322, 323] compared to 1D GC. The greater separation capacity afforded by GC×GC provides data sets that are typically 3 to 10 times larger than 1D GC [324]. This makes the technique an information rich source of chemical data that requires sophisticated computerised data processing methodologies in order to extract the maximum relevant and useful information. Commonly employed GC×GC data processing methods include visualisation, background correction, peak detection and quantification [325-329].

Comparison of chromatograms is required for QC in order to compare a manufactured product with a standard. Shellie et al. [330] developed a number of methods for comparing GC×GC chromatograms, including direct chromatogram comparison, chromatogram subtraction and averaging routines, as well as a method for generating relative weighted peak surface difference chromatograms and a more conventional Students *t*-test statistical approach.

Comparison of chromatograms depends on accurate alignment of features to ensure that the same analytes are being compared [120, 326, 331, 332]. Several methods have been proposed for the alignment of GC×GC data and most are based on procedures originally developed for 1D GC. It should be noted, however, that in 2D separations the alignment is more critical due to the inherently higher variability of the retention times in the short second dimension time window [333]. Due to the limited number of modulation cycles per one-dimensional run, the number of data points available in the first dimension is limited (i.e. data density is low). In the second dimension, the data acquisition rate is much higher than in the first dimension (2 orders of magnitude or more), hence more information is obtained in this direction [328]. Fraga et al. [334] and van Mispelaar et al. [335] proposed algorithms to correct retention time variations in comprehensive 2D separations, however both methods can only be applied to small, local regions of interest in the chromatogram. In order to correct retention time variations over the entire chromatogram in both separation dimensions, Zhang et al. [117] and Pierce et al. [120] extended the 1D alignment methods of correlation optimised warping and piecewise alignment, respectively.

A GC×GC chromatogram can be considered similar to a digital image, where each resolved chemical species produces a cluster of pixels at a pair of characteristic retention values defined by each column. Image based comparison methods for GC×GC data sets are described by Hollingsworth et al. [336].

This chapter describes the development of novel QC software that involves automated alignment of GC×GC chromatograms obtained with a univariate detector such as a FID. Each chromatogram, reference and sample, is reduced to a list of component peaks. Suitable reference peaks, termed control points, are then identified in the reference chromatogram. An astronomical pattern matching algorithm [337] is used to match these control points to features in the sample chromatogram. Subsequent alignment of the sample data is performed by affine transformation. A model fragrance is used to illustrate implementation of the software.

4.2 Experimental

4.2.1 Samples

The sample examined to evaluate and test the software is a model fragrance provided by Firmenich (Firmenich SA, Meyrin, Switzerland). It contained a representative range of perfumery ingredients, including esters, aldehydes, ketones and amines.

4.2.2 Instrumentation

The GC×GC system was based on an Agilent 6890 GC (Agilent Technologies, Wilmington, DE, USA) equipped with a split-splitless injector and a FID detector. A two-stage double loop modulator (ZX1; Zoex Corporation, TX, USA) was installed in the GC oven. This modulator consists of a cold jet (nitrogen gas cooled by liquid nitrogen) and a hot jet (heated air, at the temperature of the GC oven temperature + 150°C, duration = 350 ms) positioned orthogonal to each other. A double trapping loop was positioned in the flow of both jets. The cold jet operated constantly to trap compounds within the double loop assembly, whilst the hot jet pulses periodically, acting to both divert the flow of the cold jet to release trapped compounds, and to heat the cold spot to actively remobilise the trapped compounds more quickly [57, 338]. The modulation period was 1.5 s. The first dimension column (HP-FFAP 15 m × 0.25 mm × 0.25 µm; Agilent J&W) was linked to the second column (DB-1 1 m × 0.1 mm × 0.1 µm; Agilent J&W) via a deactivated silica column (called the transfer line, 0.1 mm i.d., 1.75 m). The transfer line was installed in the modulator in a double loop configuration. Press fits were used between the first column and the transfer line and between the transfer line and the second column. The trapping of the compounds was performed in the transfer line. The GC inlet was heated to 250°C, and a 0.1 µL injection volume (using a 1 µL syringe, Hamilton 7001N, ga 0.47/70mm/pst 2, P/N 80135/01) was used with a split ratio of 30:1. For the reference data helium was used as carrier gas at a flow of 1 mL min^{-1}, in constant flow mode. The oven temperature program started from 40°C then increased at 15°C min^{-1} up to 230°C with a 7 min hold. A flame ionisation detector was used at 250°C, with nitrogen makeup gas, and a data acquisition rate of 100 Hz.

Further sample chromatograms were obtained by varying the flow rate and temperature ramp in order to generate chromatograms with induced peak shifts. The details are summarised in Table 4.1.

Table 4.1: Details of the conditions used to obtain the varied sample chromatograms. (-) refers to no change

Sample	**Δ Flow rate (mL/min)**	**Δ Temperature ramp (°C/min)**
1	- 0.1	(-)
2	+ 0.1	(-)
3	(-)	- 0.5
4	(-)	+ 0.5

4.2.3 Data processing

All data manipulation and analysis algorithms were developed and implemented in-house using Matlab (V7.10 (R2010a), MathWorks Inc, MA, USA). The raw data from the GC×GC system were acquired using Agilent ChemStation vE01.01.335 (Agilent Technologies). It was then exported in comma-separated values format (.csv) using an in-house macro and imported into Matlab for processing.

4.3 Program development

The developed software is employed to align reference and sample chromatograms in order to allow accurate comparison for QC purposes. The software involves peak detection of the reference and sample chromatograms, reference control point selection, identification of corresponding sample control points, and alignment of the sample to the reference chromatogram. A model fragrance was used to illustrate development of the algorithm. The fragrance was analysed under standard conditions (section 4.2.2) and the GC×GC chromatogram was used as the reference. The fragrance was then analysed after increasing the temperature ramp (sample 4 in Table 4.1) to provide a test sample chromatogram with induced peak shifts in order to illustrate the alignment process.

4.3.1 GC×GC data

GC×GC data is similar to the time-response data generated in 1D GC (Figure 4.1).

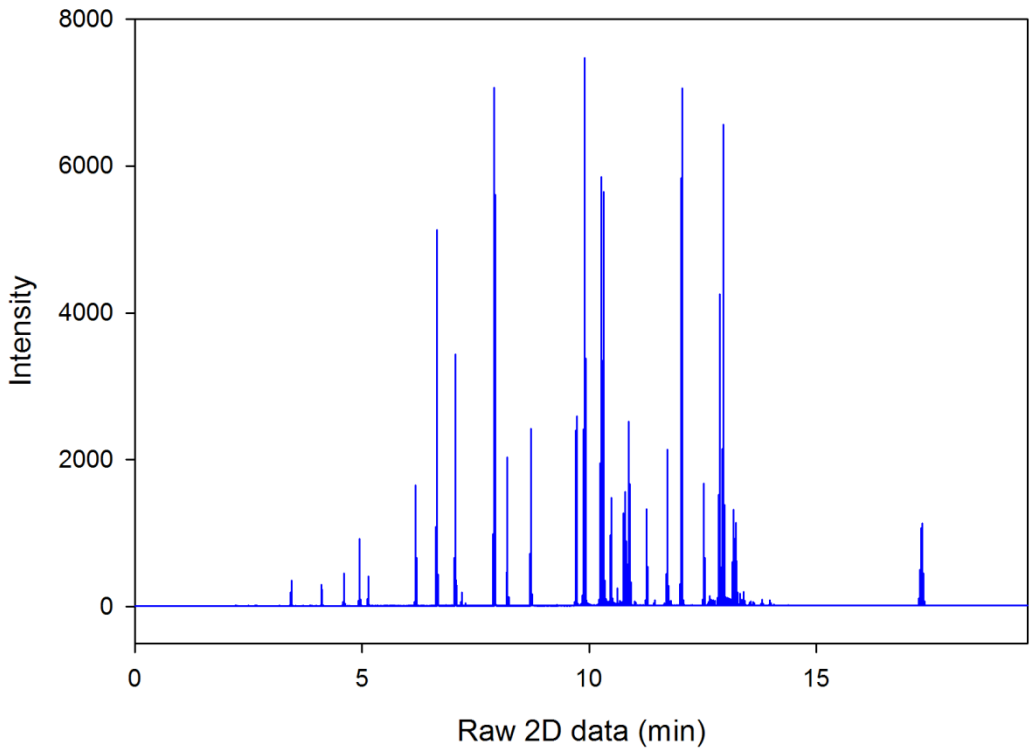


Figure 4.1: Raw GC×GC data of the model fragrance

Since the modulation period and sampling frequency are known accurately the recorded data can be folded every cycle (Equation 4.1) to form a 2D matrix or array as shown in Figure 4.2.

$$cycle = mod * freq \qquad \text{Equation 4.1}$$

Where *mod* is the modulation period (s) and *freq* is the sampling frequency (Hz). In this work a modulation period of 1.5s and a sampling frequency of 100Hz was used, therefore every 150 points the data is folded to form a new column in the 2D matrix.

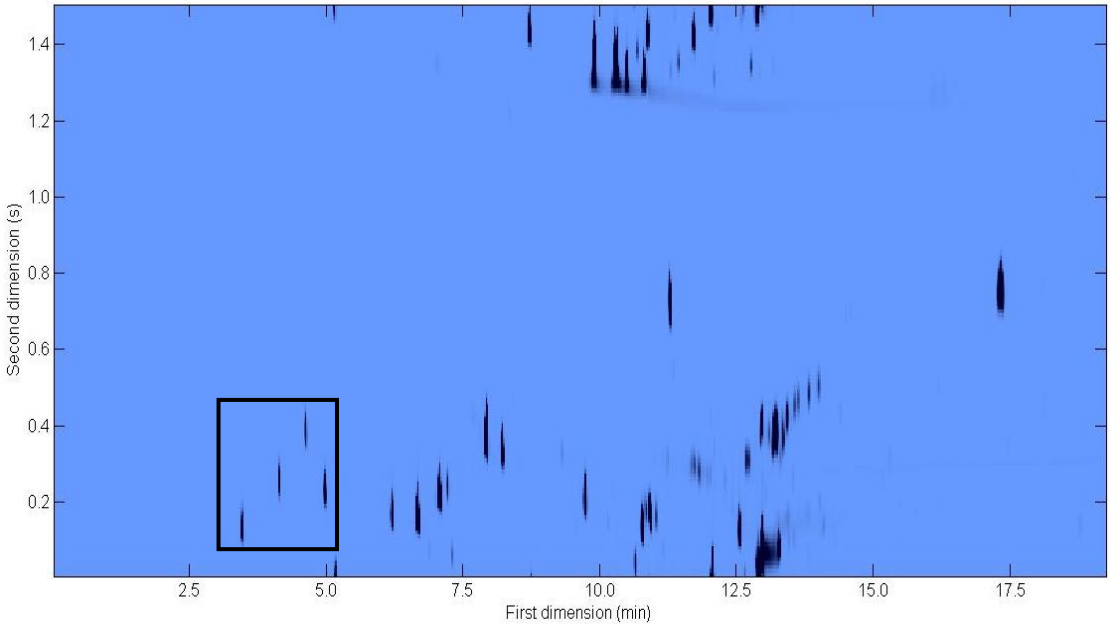


Figure 4.2: 2D image of the model fragrance GC×GC data matrix

To simplify interpretation, only a small selected region of the chromatogram (highlighted in Figure 4.2) is used to demonstrate development of the software.

4.3.2 Peak detection

Peak detection is performed using an algorithm based on that described by Peters et al. [328]. The algorithm involves 1D peak detection, 2D peak merging and quantification of the 2D peaks.

It is assumed that each column of the 2D array represents a 1D chromatogram in the second dimension and as a result a 1D peak detection algorithm can be applied to each of these 1D chromatograms. Figure 4.3 highlights columns 197-201 of the 2D data matrix in Figure 4.2, from this it can be seen that each 1D chromatogram (columns in the data array) contains one peak that should be found by a 1D peak detection algorithm.

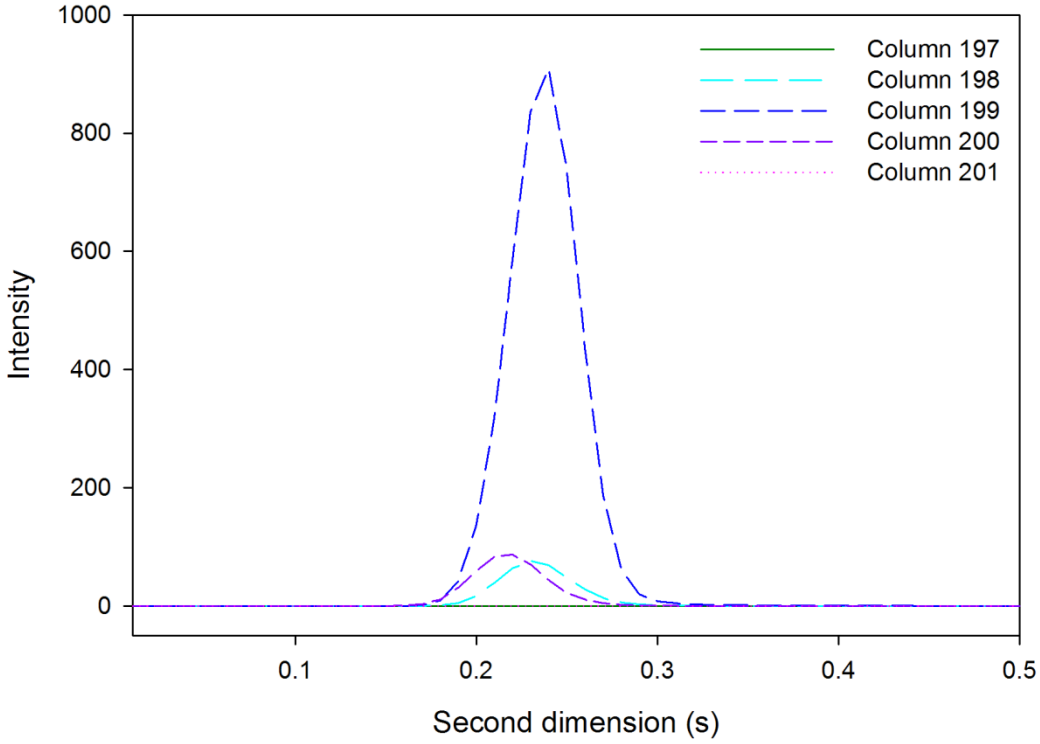


Figure 4.3: Columns 197-201 from the GC×GC data matrix in Figure 4.2

The 1D peak detection algorithm employed in this work was based on that described by Vivo-Truyols et al. [339] and is undertaken on the data as a single dimension vector, as displayed in Figure 4.1. Baseline correction is achieved by subtracting from each segment, corresponding to a second dimension column, the median value of the segment. The complete 1D data vector ($d0_i$) is then smoothed and peak locations identified from the first ($d1_i$) and second ($d2_i$) derivatives obtained using the Savitzky-Golay algorithm [97] with a 7-point quadratic polynomial. The peak location is given by the second derivative and the range (width) of each peak is derived from the first

derivative data. Figure 4.4 shows an example of the original, first and second order derivatives with the defined peak characteristics.

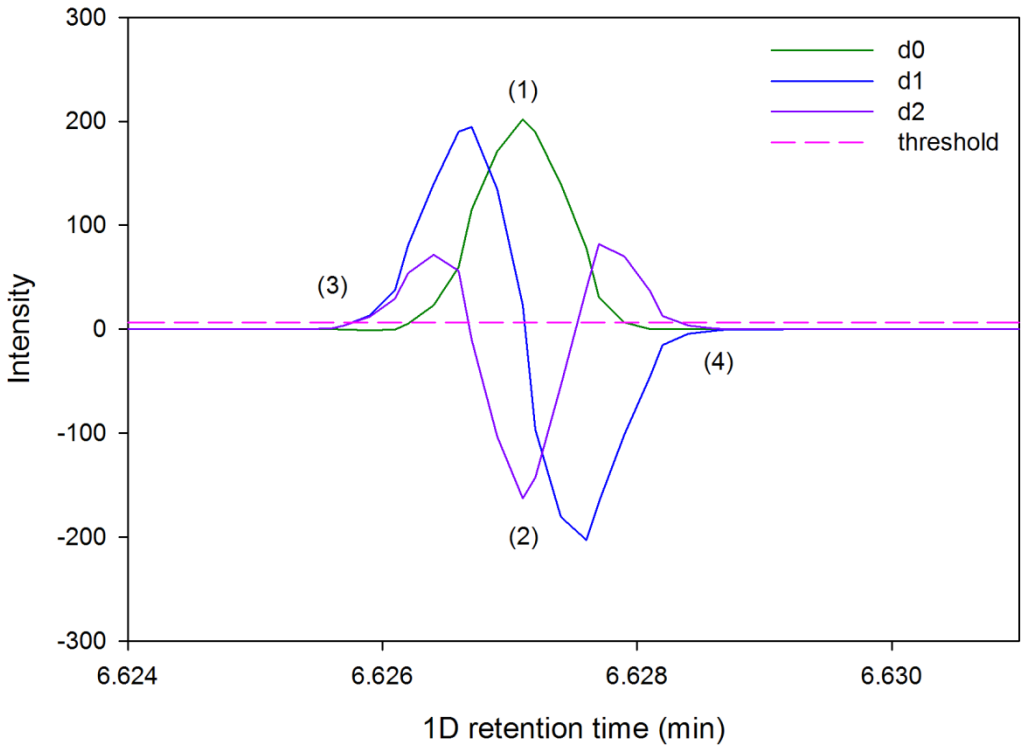


Figure 4.4: Example of the original (d0), first (d1) and second (d2) order derivatives. Peak characteristics are labelled. (1) maximum peak height (maximum in d0), (2) peak location (minimum in d2), (3) peak start point (first point in d1 above the threshold) and (4) peak end point (last point in d1 above the threshold). To allow comparison of these three plots on the same vertical axis, d0 was divided by 5.

To define peak characteristics, start and end points, and to establish a minimum peak intensity, a threshold value is set with only peaks exceeding this threshold considered. The threshold employed in this work in defined according to Equation 4.2.

$$threshold = 10 \cdot \frac{1}{n} \sum_{i=L}^{n-1} \left| d2_i - \left(\frac{d2_{i+1} - d2_{i-1}}{2} \right) \right| \qquad \text{Equation 4.2}$$

The threshold value is based on the second derivative data vector and is ten times the average difference of adjacent values. After detecting all peaks exceeding the threshold value in the 1D chromatogram, the output is a list of peaks comprising their first and second dimension indices, the start and end points and the area under the peak.

The algorithm then merges appropriate peaks into 2D clusters. A 2D cluster is a collection of 1D peaks in consecutive 1D chromatograms that are considered to belong to the same peak, and therefore should be merged to form a single object. When a 2D cluster is unable to be extended with more 1D peaks, a 2D peak is considered complete and defined. The first 1D peak in the first second dimension chromatogram is considered as the first 2D cluster. Next, all 1D peaks found in the adjacent second dimension chromatogram are considered as candidates for merging. Overlap and unimodality criteria are applied in order to accept or reject the merging of these 1D peaks with the first 2D cluster. The overlap criterion examines the degree of overlap between the 1D peaks in consecutive second dimension chromatograms. Depending on the adjacent peak regions considered, five different situations can be distinguished. These are outlined below and illustrated in Figure 4.5 where 1D peak A is defined as the last 1D peak of the existing 2D cluster and 1D peak B is the candidate peak for merging.

(a) Both peaks start at the same location in the second dimension
(b) Peak A starts and ends later than peak B
(c) Peak B starts and ends later than peak A
(d) Peak B starts later than peak A, but ends earlier
(e) Peak A starts later than peak B, but ends earlier

The percentage of overlap is calculated according to Equation 4.3.

$$OV = (b/a) \cdot 100 \qquad \text{Equation 4.3}$$

Where b is the region of the candidate peak (1D peak B) that is overlapped with the peak region of 1D peak A and a is the region of peak A. A threshold is then selected; in this work a threshold of 40% was used. If OV is greater than the threshold, the 1D candidate peak is accepted; if not, this candidate peak is rejected and the algorithm proceeds to the next candidate peak. In cases (d) and (e), one of the peaks is incorporated in the peak region of the other peak and as a result the candidate peak is always accepted.

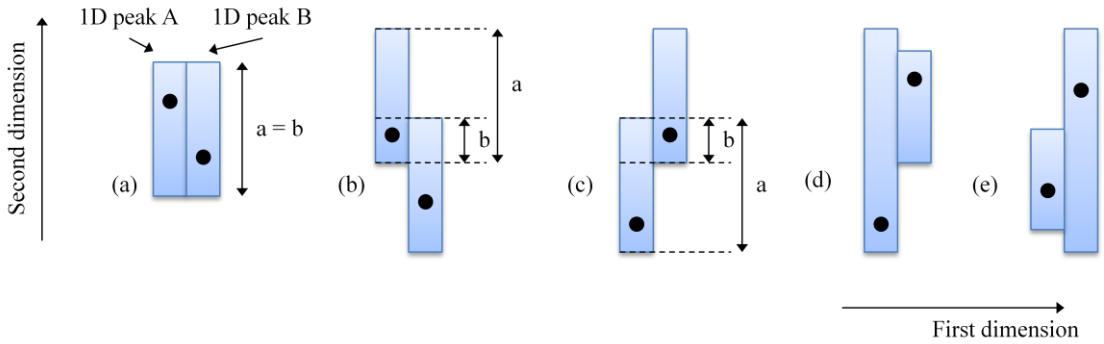


Figure 4.5: Schematic representation of peak regions of adjacent 1D peaks for the five different cases. Adapted from [328].

All 1D peaks that meet the overlap criterion are then subjected to a unimodality criterion. The unimodality criterion examines the peak maxima profile, which represents the chromatographic peak profile in the first dimension and therefore it should only show one maximum (i.e. unimodal). With increasing first dimension locations, if a maximum has already been detected, only candidate peaks with decreasing intensities are accepted. However, if the intensity of the candidate peak is greater than the previous maximum, the 2D cluster is considered complete and a new cluster is started. The 1D peak maxima profile of the peaks in the selected region of Figure 4.2 is depicted in Figure 4.6 (a). From this it can be seen that there are four separate 2D clusters, each made up of multiple 1D peaks. The 1D peaks in each cluster then need to be merged to form four 2D peaks. In order to identify which 2D peaks the 1D peaks in Figure 4.6 (a) correspond to, they are plotted as 2D peaks in Figure 4.6 (b).

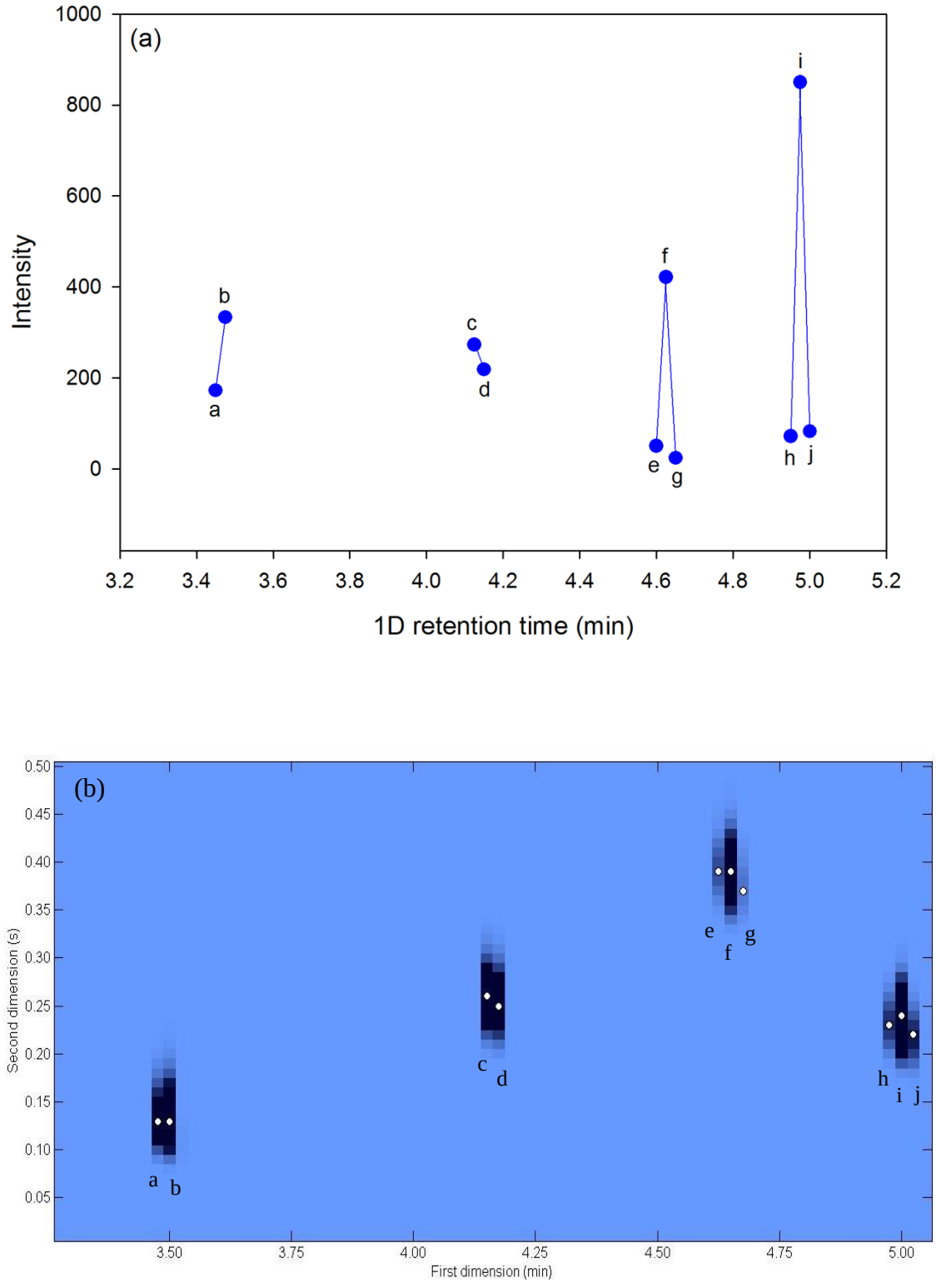


Figure 4.6: (a) 1D peak maxima profile of the peaks in the selected region of Figure 4.2 and (b) 2D image plot identifying the 2D peaks that correspond to the 1D peaks in (a). a-j identify the corresponding peaks

If more than one peak meets the overlap and unimodality criteria, the candidate peak with the second dimension location closest to the last 1D peak in the 2D cluster is selected for merging. Once this process is complete, peaks identified as belonging to a single cluster are merged to form a 2D peak and its underlying volume is given by summing the areas of the contributing 1D peaks. This peak merging process is illustrated in Figure 4.7.

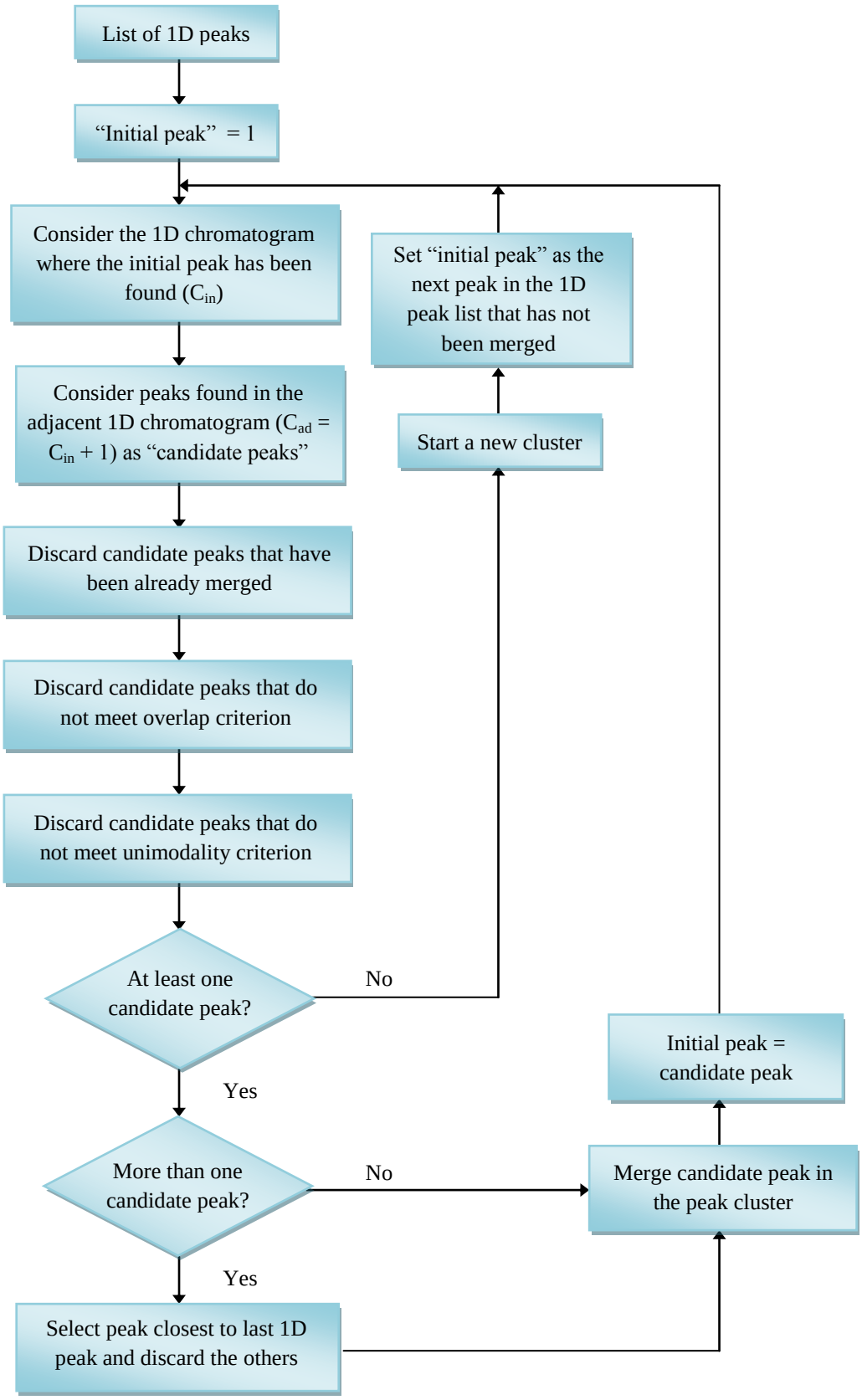


Figure 4.7: General flow chart of procedure for merging 1D peaks to form 2D clusters. Adapted from [328].

At this point in the software, the output for both the reference and sample data is the 2D data matrix displayed as an image (Figure 4.8) and a list of the identified peaks with their corresponding first and second dimension locations and the relative volume enclosed by each peak (Table 4.2).

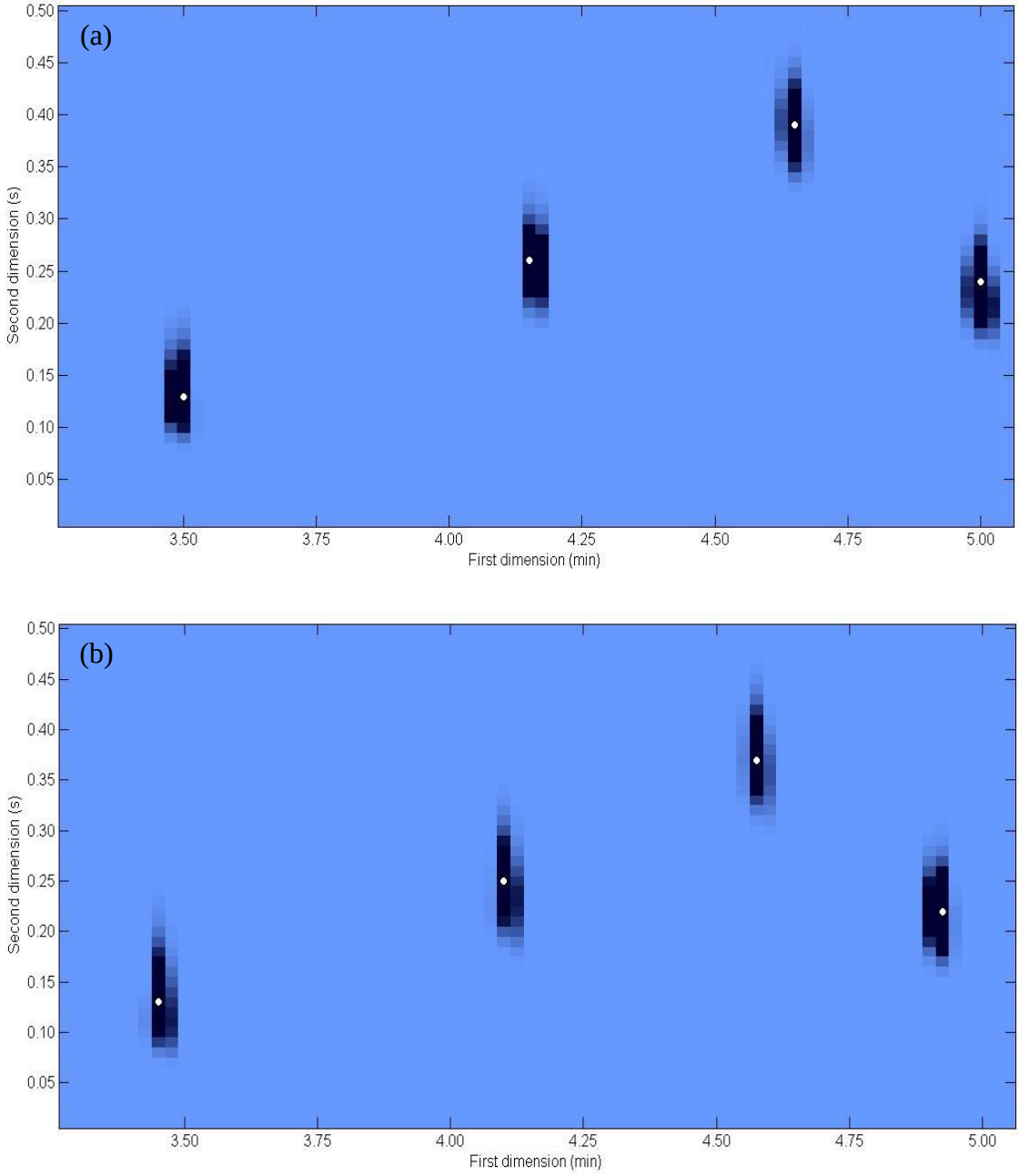


Figure 4.8: 2D image generated by software with peaks identified (a) reference and (b) sample

Table 4.2: List of peaks identified by software with their first and second dimension locations and relative volume enclosed by each peak

Reference			Sample		
Dim 1 (min)	**Dim 2 (s)**	**Volume (% total)**	**Dim 1 (min)**	**Dim 2 (s)**	**Volume (% total)**
3.48	0.13	20.99	3.43	0.13	21.08
4.13	0.26	20.81	4.08	0.25	20.65
4.63	0.39	20.22	4.55	0.37	20.28
4.98	0.24	37.98	4.90	0.22	37.99

4.3.3 Aligning chromatograms

In order to match and compare peaks arising from a reference material and a sample material, the generated peaks, as x- and y- locations and z-intensity data, are considered as points in an image and it is necessary to identify and select peaks that are common to both sets of data. These peaks are referred to as index points or control points.

The process of overlaying two or more maps or images of similar data is referred to as registration. Registration geometrically aligns a reference and some sample image. Zitova and Flusser [340] reviewed modern and traditional methods of image registration and in the case of aligning 2D chromatograms we can identify the following four steps. (1) feature detection, by which the locations of distinctive and characteristic objects in the reference image are noted. In a 2D chromatogram these features can be represented by the location of peak maxima, and are termed index points or control points. (2) feature matching, where the correspondence between the control points in the sample and the reference chromatograms is established. (3) determine a mapping function that aligns the sample control points with those from the reference. (4) transform the complete sample chromatogram by the mapping function to achieve correspondence between the two chromatograms. All sample peak locations are moved using this transformation function, allowing subsequent comparison of all peaks between chromatograms.

Note that it is not necessary that the complete sample chromatogram be transformed to match that of the reference. We only need to match corresponding peaks, i.e. chemical components, between the chromatograms, thus saving computation time.

4.3.3.1 Selection of control points

Selection of suitable control points is required to derive the translation, scaling and rotation parameters for data transformation. Although manual selection of these common peaks is generally straightforward, the process is subjective and can be time consuming. Instead an efficient and effective automated method to carry out this task is proposed. Here, selection of control points in the reference chromatogram is undertaken by partitioning the chromatogram into a set of equal sized segments, the number of which can be user selected according to the distribution of peaks within the GC×GC pattern space. In this case 24 sectors were employed, 12 along the first dimension and 2 along the second dimension. The mean peak volume for all peaks across the entire chromatogram is calculated and the peak in each sector with the volume closest to the mean is selected as a control point, thus providing up to 24 reference control points covering the pattern space. This procedure reduces the likelihood that neither very intense peaks nor minor peaks are likely to be selected as control points. Intense peaks can arise due to overloading the GC system and are often characterized by broad, asymmetric, and irregular profiles with unclear peak maxima locations between similar samples. Similarly, minor or weak peaks can be problematic since their retention times are more likely to be influenced by environmental factors, instrument noise, or the presence of intense peaks, and may not be present in all samples to be compared.

However, as the simple illustrative example chromatograms contain only four peaks and the affine transform requires a minimum of three peaks, all four peaks will be used as control points.

Once control points are selected in the reference chromatogram, identification of corresponding points in the sample chromatogram is undertaken. The algorithm employed here is derived from a star recognition algorithm originally developed by Groth for the comparison of star maps [337]. The Groth pattern-matching algorithm compares two lists of star coordinate positions and identifies individual points from one list (a reference) and their likely counterparts in the other, sample list. The 2D coordinate lists are matched according to the similarity of triangles formed between every combination of three points within each list. Geometrically similar pairs of triangles, one from each list, are identified and a voting process provisionally

highlights points that appear in multiple triangle pairs as being common to both lists. The algorithm is insensitive to translation, rotation, magnification or inversion; it can tolerate minor random errors or distortions and does not require the two lists to be of equal length. Modifications to the original algorithm have been proposed for matching photographic images of whale sharks [341]. Here, Groth's algorithm has been adapted for GC×GC analysis.

From the coordinates of the reference control points, triangles are formed between every combination of three points with a peak position defining each of a triangle's vertices (v_1, v_2, and v_3 defined by the locations (x_1,y_1), (x_2,y_2), and (x_3,y_3) respectively). For each list of *n* points, the total number of triangles, *N*, is given by Equation 4.4. Thus for the 4 control points in the reference chromatogram there are 4 triangles describing the peak distribution.

$$N = n(n-1)(n-2)/6 \qquad \text{Equation 4.4}$$

The triangle vertices are arranged to make the shortest side, r_2, lie between vertices 1 and 2, the intermediate side, r_1, between vertices 2 and 3, and the longest side, r_3, between vertices 1 and 3. The triangles are then defined by the following six geometric properties as illustrated in Figure 4.9.

i) The location of the centroid of the triangle, $\bar{x}, \bar{y}$,

$$\bar{x} = (x_1 + x_2 + x_3)/3 \qquad \text{Equation 4.5}$$
$$\bar{y} = (y_1 + y_2 + y_3)/3$$

ii) The ratio of the longest to the shortest sides, R.

$$R = r_3/r_2 \qquad \text{Equation 4.6}$$

iii) The cosine, C, of the angle, α, at vertex 1

iv) The rotation angle, θ, defined by the angle between vertex 1, the centroid and the x-axis.

$$\theta = tan^{-1}\left[\frac{(y_1 - \bar{y})}{x_1 - \bar{x}}\right] \qquad \text{Equation 4.7}$$

v) The perimeter of the triangle, P.

$$P = r_1 + r_2 + r_3 \qquad \text{Equation 4.8}$$

vi) The orientation of the triangle (whether vertices 1, 2 and 3 are in a clockwise or counter-clockwise direction), O. In terms of vertex coordinates,

$$(y_2 - y_1, x_2 - x_1) \cdot (x_3 - x_1, y_3 - y_1) > 0 \qquad \text{Equation 4.9}$$

describes a clockwise oriented triangle [342].

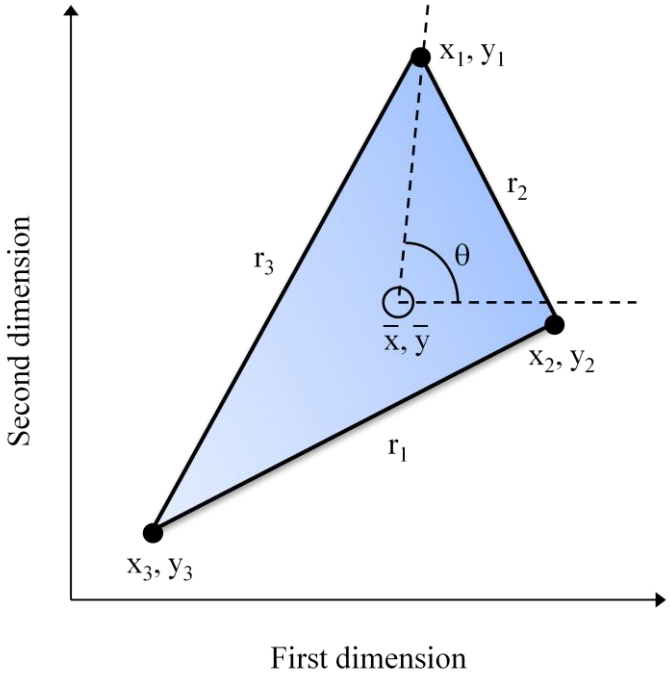


Figure 4.9: The characteristic features of triangles between any three peak locations

These geometric properties are required to prevent incorrect matching of the triangles. The orientation of the triangle, whether the sides are specified in a clockwise or counter-clockwise manner, provides a useful discriminating tool to ensure only triangles facing the same direction are matched. The rotation angle, θ, prevents severely rotated triangles from being incorrectly matched. The rotation angle was an addition to Groth's algorithm proposed by Arzoumanian et al. [341] for matching shark patterns and was employed as a 'local' measure of rotation that provides some insensitivity to distortions. Matching triangles with similar values of C ensures that only triangles with a similar angle at vertex 1 are considered. Triangles of significantly different sizes are not matched due to R and P and by looking at the

location of triangle centres, triangles which are similar in all other properties but are located in significantly different regions of the 2D chromatogram are not matched. This feature of examining the location of the triangle centres is a new measure employed in this work. Unlike a star map or photograph of a shark, the peaks in a GC×GC chromatogram will not be severely shifted between chromatograms; therefore triangles located at significantly different regions of the chromatogram will not be matched based on the location of the triangle centre.

The reference control points and the characteristics of their triangles are employed to match corresponding points in sample chromatograms and subsequently align the peaks.

4.3.3.2 Matching reference and sample control points

Once the triangle properties of the reference control points have been defined, the entire sample peak coordinate list is examined. The six geometric properties are calculated for all possible triads from the sample peak list. The reference and sample triangles are illustrated in Figure 4.10 and their corresponding geometric properties are provided in Table 4.3. From this it can be seen that despite the peak locations being shifted in the chromatogram (Table 4.2) the triangle characteristics remain similar.

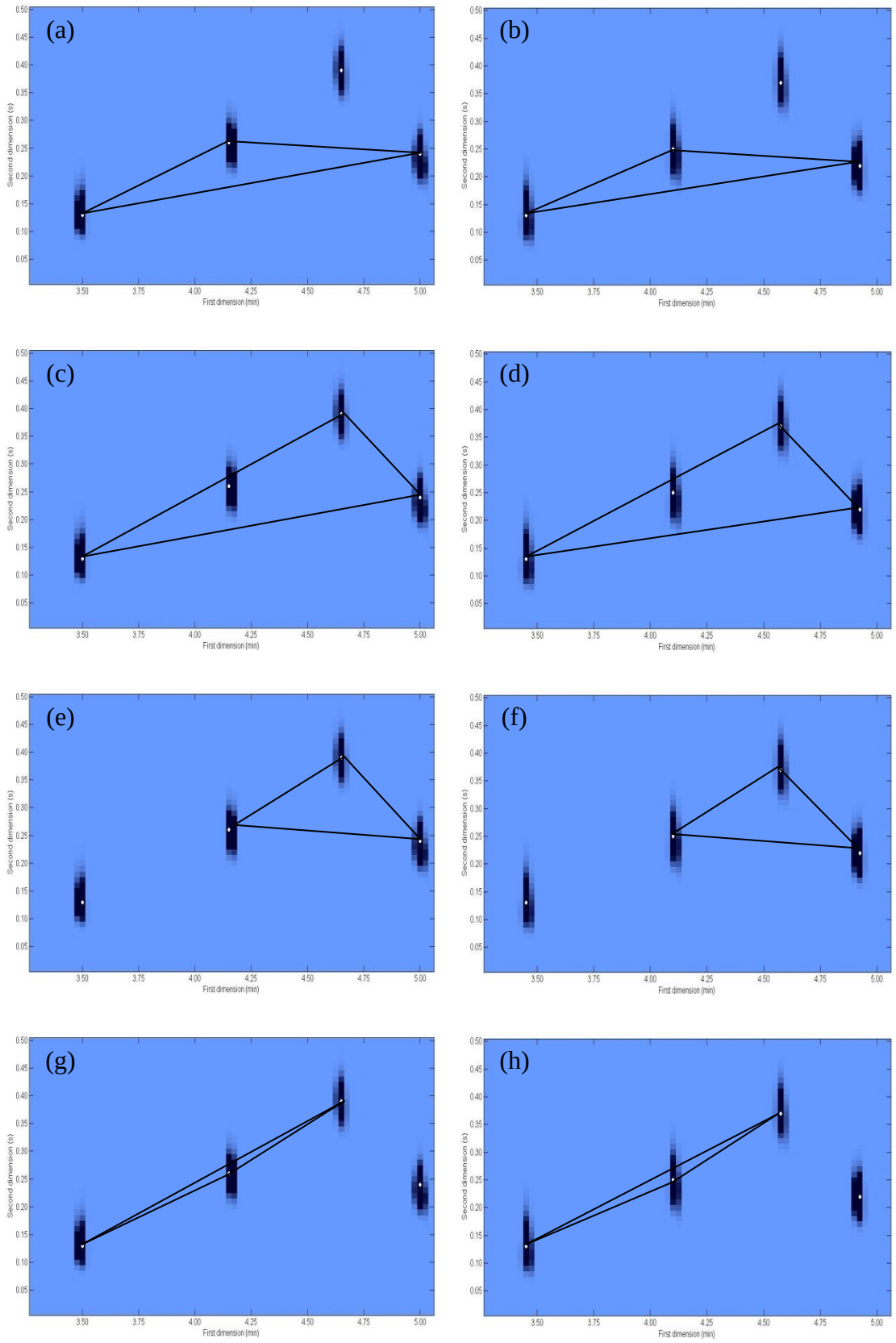


Figure 4.10: Reference and sample triangles (a) reference triangle 1, (b) sample triangle 1, (c) reference triangle 2, (d) sample triangle 2, (e) reference triangle 3, (f) sample triangle 3, (g) reference triangle 4 and (h) sample triangle 4

Table 4.3: Geometric properties for the four reference and sample triangles in Figure 4.10

	Reference						**Sample**					
Triangle	**Centroid (x,y)**	**R**	**C**	**θ**	**P**	**O**	**Centroid (x,y)**	**R**	**C**	**θ**	**P**	**O**
1	38.67,21.00	2.10	0.96	0.27	124.13	-1.00	36.33, 20.00	2.08	0.96	0.24	121.45	-1.00
2	45.33, 25.33	2.97	0.54	-0.05	134.36	1.00	42.67, 24.00	2.91	0.56	-0.08	131.20	1.00
3	54.00, 29.67	1.66	0.72	-0.34	78.43	1.00	51.33, 28.00	1.61	0.75	-0.37	76.12	1.00
4	34.00, 26.00	2.22	1.00	0.53	105.76	-1.00	31.67, 25.00	2.27	1.00	0.51	102.11	-1.00

The computed properties for every triangle from the sample chromatogram are compared with each of those from the reference chromatogram in order to find matches. The triangles are first compared according to orientation, then closeness of centroids, followed by perimeter, rotation angle, ratio of side lengths, and cosine of the vertex angle. A successful match is noted if all comparisons are within predefined tolerance values (selected empirically according to typical uncertainty of coordinate measurements). From this a list of matched triangle pairs is compiled, each of which involves three pairs of matched vertex points. Table 4.4 shows the tolerances employed in this work compared with those employed by Groth [337] and Arzoumanian et al [341].

To improve the efficiency of the matching algorithm, it is necessary to filter the triangles. Triangles with large length ratios and cosine values close to 1 are removed as such elongated and flattened triangles have a high probability of causing false matches and as a result weaken the algorithms discriminating ability. This process removes triangle 4 (Table 4.3) as it was a flattened triangle with a cosine value of 1.

Table 4.4: Tolerance values for triangle matching

Parameter	**Groth**	**Arzoumanian**	**This work**	**Description**
$\mathbf{R_{max}}$	10	8	8	Maximum triangle side length ratio
$\mathbf{C_{max}}$	NA	0.99	0.99	Maximum cosine of angle at vertex 1
$\mathbf{\theta_{max}}$	NA	10°	20%	Maximum relative triangle rotation
$\mathbf{P_{max}}$	NA	NA	10%	Maximum difference between triangle perimeter
Centriod	NA	NA	20	Maximum difference in position of triangle centre

In order to determine which peaks are truly common to both reference and sample, it is assumed that matching peaks will appear in more matching triangles. Therefore the number of times a sample peak appears in a matching triangle is calculated, and this value is referred to as the vote. The sample peak having the highest vote for each reference peak is selected, providing a list of peaks that are assumed to be the same in

both the reference and sample chromatograms. This list of matching peaks identifies corresponding analytes in the chromatograms that can be used for subsequent derivation of the transformation function in order to achieve alignment of the chromatograms. The matching reference and sample peaks for the example data are provided in Table 4.5. The results indicate that all the peaks were successfully matched.

Table 4.5: Results of triangle matching. List of matched reference and sample peaks (first and second dimension locations) with the vote

Reference		**Sample**		
Dim 1 (min)	**Dim 2 (s)**	**Dim 1 (min)**	**Dim 2 (s)**	**Vote**
3.48	0.13	3.43	0.13	2
4.13	0.26	4.08	0.25	2
4.63	0.39	4.55	0.37	1
4.98	0.24	4.90	0.22	1

The maximum number of votes, V_{max}, for a list of K points, is given by:

$$V_{max} = \sum_{i=1}^{K-2} i \qquad \text{Equation 4.10}$$

For the 4 sample peaks, the maximum number of votes that a peak can receive according to Equation 4.10 is 3. Thus the 2 votes obtained for two of the peaks in Table 4.5 represents good matching, while the peaks that got only 1 vote belonged to triangles that varied more than the tolerances defined in Table 4.4 and hence were not found to be matching triangles.

4.3.3.3 Alignment

In order to align the sample to the reference it is necessary to derive and apply some suitable transformation function. Transformation or mapping functions can be divided into two broad categories [339]. Global models use all the control points to estimate a single transformation function for the whole image, while local mapping functions treat the image as comprising patches of data each of which will have its own separate transformation model.

A wide variety of transforms are discussed in the literature and of the global models, low degree polynomials are most frequently encountered. Transformation to align a sample to a reference consists of translation, scaling and rotation, with the simplest and most robust of the linear models being the affine transform. We assume the reference control points, **R**, have been shifted in the corresponding sample data, **S**, by a linear combination of translation, scaling and rotation operations. Then, using homogeneous coordinates,

$$\mathbf{R} = \begin{bmatrix} x_{R1} & x_{R2} & \dots & x_{Rn} \\ y_{R1} & y_{R2} & \dots & y_{Rn} \\ 1 & 1 & \dots & 1 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} x_{S1} & x_{S2} & \dots & x_{Sn} \\ y_{S1} & y_{S2} & \dots & y_{Sn} \\ 1 & 1 & \dots & 1 \end{bmatrix}, \qquad \text{Equation 4.11}$$

The transformation matrix, **A,** is defined as,

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ 0 & 0 & 1 \end{bmatrix} \qquad \text{Equation 4.12}$$

Thus,

$$\boldsymbol{R} = \boldsymbol{A} \cdot \boldsymbol{S} \qquad \text{Equation 4.13}$$

Where $a_{1,1}$ and $a_{2,2}$ are the scaling coefficients for x and y respectively, $a_{1,2}$ and $a_{2,1}$ are the rotation coefficients and $a_{1,3}$ and $a_{2,3}$ describe translation along the x- and y-axes.

The affine transform requires a minimum of three non-collinear control points, but generally many more than this are available. The parameters of the fitting function are then calculated by means of a least-squares fit, minimising the sum of squared errors at the control points.

$$\boldsymbol{A} = \boldsymbol{R} \cdot \boldsymbol{S}^{\boldsymbol{T}} (\boldsymbol{S} \cdot \boldsymbol{S}^{\boldsymbol{T}})^{-\mathbf{1}} \qquad \text{Equation 4.14}$$

A is determined from coordinates of the control points in the sample chromatogram matched, using the triangles algorithm, to the control points in the reference chromatogram. Once calculated, it is applied to the complete list of sample peak coordinates (Equation 4.13).

The affine transform is known to be a quick and simple method that is effective in reducing retention time variations associated with chromatographic data [336, 343].

The results of performing an affine alignment are shown in Figure 4.11. The reference chromatogram is plotted and the sample peaks before and after alignment are shown. From this it can be seen that the sample peaks are closer to the reference peaks after alignment via affine transformation. This allows the reference and sample peaks to be accurately compared.

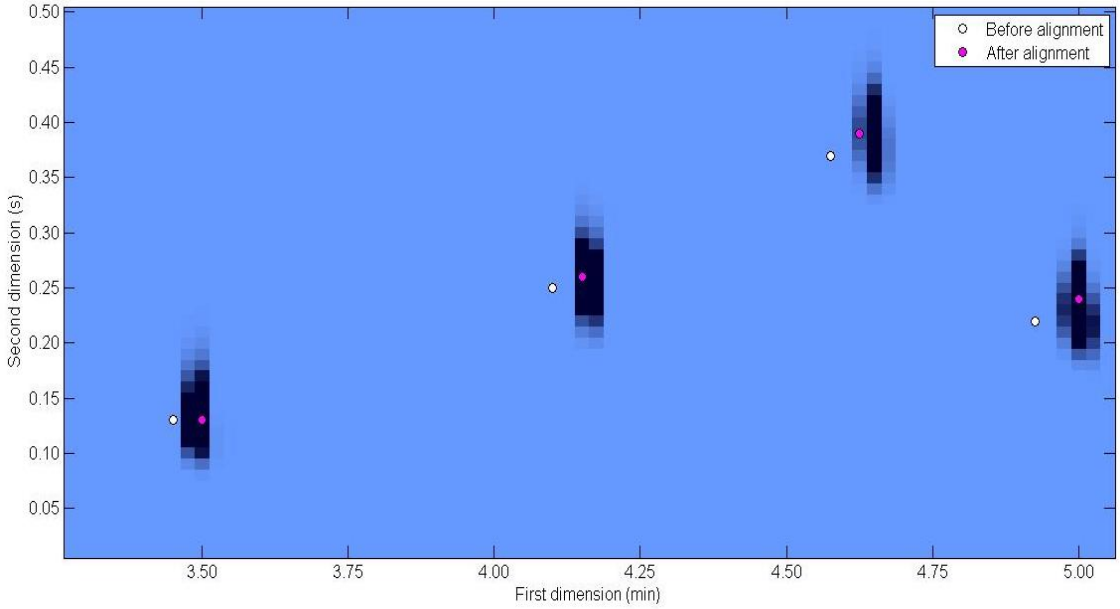


Figure 4.11: Results of alignment. Reference chromatogram with sample peaks plotted before and after alignment via affine transformation

4.4 Results and discussion

The software was employed to align peak lists from entire reference and sample chromatograms. A model fragrance was used to illustrate implementation of the algorithm. As described above, the fragrance was analysed under standard conditions (section 4.2.2) and the GC×GC chromatogram obtained was used as the reference image. The fragrance was then analysed under modified conditions to provide a series of test chromatograms with induced peak shifts. The flow rate was decreased (sample 1) and increased (sample 2) by 0.1 mL/min and the temperature ramp was decreased (sample 3) and increased (sample 4) by 0.5°C/min to produce four sample data sets (Table 4.1).

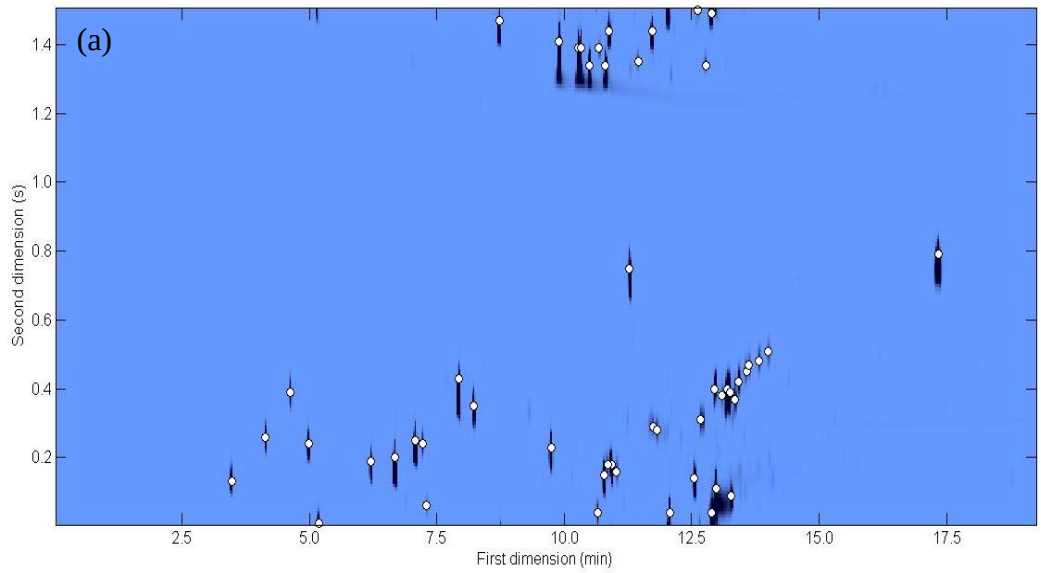
Peak detection was performed on the reference and the four sample chromatograms. The results of peak detection are provided in Table 4.6. For each chromatogram the data is sorted according to the first dimension index and second dimension index. The reference and four sample chromatograms, with peaks marked, are illustrated in Figure 4.12.

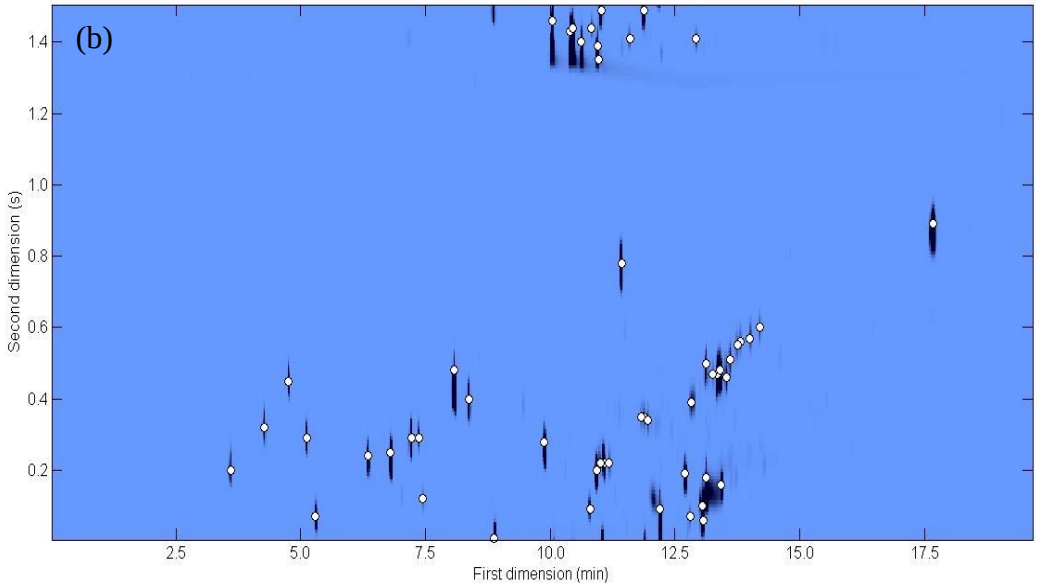
Table 4.6: List of peaks identified in the reference and four sample chromatograms (first and second dimension locations and volume)

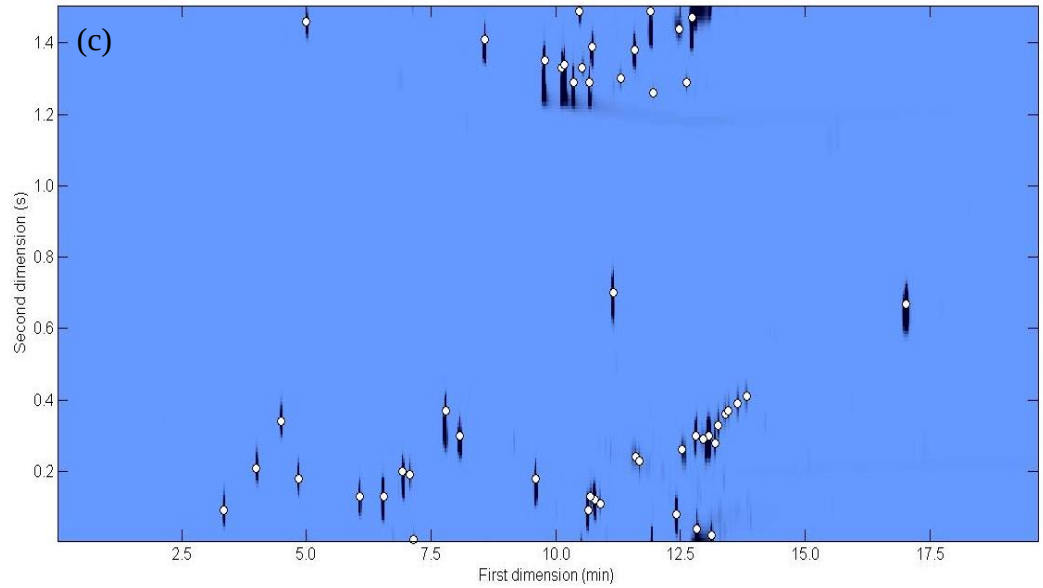
Reference			Sample 1			Sample 2			Sample 3			Sample 4		
Dim 1 (min)	**Dim 2 (s)**	**Volume (%)**	**Dim 1 (min)**	**Dim 2 (s)**	**Volume (%)**	**Dim 1 (min)**	**Dim 2 (s)**	**Volume (%)**	**Dim 1 (min)**	**Dim 2 (s)**	**Volume (%)**	**Dim 1 (min)**	**Dim 2 (s)**	**Volume (%)**
3.48	0.13	0.37	3.60	0.20	0.38	3.35	0.09	0.36	3.50	0.16	0.39	3.43	0.13	0.36
4.13	0.26	0.36	4.28	0.32	0.36	4.00	0.21	0.36	4.20	0.28	0.38	4.08	0.25	0.36
4.63	0.39	0.35	4.75	0.45	0.37	4.50	0.34	0.35	4.70	0.41	0.37	4.55	0.37	0.34
4.98	0.24	0.66	5.13	0.29	0.68	4.85	0.18	0.66	5.05	0.25	0.69	4.90	0.22	0.66
5.18	1.00	0.30	5.30	0.07	0.32	5.00	1.46	0.31	5.25	0.02	0.32	5.05	1.50	0.31
6.20	0.19	1.45	6.35	0.24	1.49	6.08	0.13	1.45	6.33	0.19	1.51	6.08	0.17	1.45
6.68	0.20	4.69	6.80	0.25	4.76	6.55	0.13	4.66	6.80	0.21	4.80	6.53	0.18	4.67
7.08	0.25	3.08	7.23	0.29	3.13	6.93	0.20	3.08	7.23	0.26	3.15	6.93	0.24	3.08
7.23	0.24	0.16	7.38	0.29	0.17	7.08	0.19	0.15	7.38	0.25	0.17	7.08	0.23	0.16
7.30	0.06	0.02	7.45	0.12	0.02	7.15	0.01	0.02	7.45	0.07	0.02	7.15	0.05	0.02
7.93	0.43	11.24	8.08	0.48	11.35	7.80	0.37	11.23	8.10	0.45	11.38	7.75	0.40	11.20
8.23	0.35	1.83	8.38	0.40	1.84	8.08	0.30	1.83	8.40	0.37	1.85	8.05	0.32	1.81
8.73	1.47	2.12	8.88	0.01	2.14	8.58	1.41	2.11	8.93	1.46	2.15	8.53	1.46	2.12
9.75	0.23	3.42	9.88	0.28	3.43	9.60	0.18	3.43	9.98	0.24	3.43	9.10	0.32	0.01
9.90	1.41	10.61	10.05	1.46	10.72	9.78	1.35	10.57	10.15	1.39	10.71	9.50	0.22	3.43
10.28	1.39	8.59	10.40	1.43	6.95	10.13	1.33	7.14	10.53	1.38	8.59	9.68	1.39	10.60
10.33	1.39	4.83	10.45	1.44	6.59	10.18	1.34	6.25	10.58	1.38	4.96	10.03	1.38	9.30
10.50	1.34	1.78	10.63	1.40	1.79	10.35	1.29	1.77	10.75	1.33	1.77	10.08	1.37	4.13
10.65	0.04	0.17	10.80	0.09	0.18	10.48	1.49	0.18	10.90	0.04	0.18	10.23	1.34	1.77
10.68	1.39	0.07	10.83	1.44	0.07	10.53	1.33	0.07	10.95	1.38	0.07	10.40	0.04	0.18

Reference (cont.)			Sample 1 (cont.)			Sample 2 (cont.)			Sample 3 (cont.)			Sample 4 (cont.)		
Dim 1 (min)	**Dim 2 (s)**	**Volume (%)**	**Dim 1 (min)**	**Dim 2 (s)**	**Volume (%)**	**Dim 1 (min)**	**Dim 2 (s)**	**Volume (%)**	**Dim 1 (min)**	**Dim 2 (s)**	**Volume (%)**	**Dim 1 (min)**	**Dim 2 (s)**	**Volume (%)**
10.78	0.15	1.43	10.93	0.20	1.43	10.65	0.09	1.40	11.05	0.15	1.48	10.43	1.39	0.06
10.80	1.34	1.74	10.95	1.39	1.61	10.68	1.29	1.72	11.08	1.34	1.75	10.53	0.13	1.49
10.85	0.18	0.06	10.98	1.35	0.15	10.70	0.13	0.07	11.15	1.44	2.01	10.55	1.33	1.75
10.88	1.44	2.03	11.00	0.22	0.05	10.73	1.39	2.02	11.18	0.19	1.60	10.60	1.44	2.02
10.93	0.18	1.60	11.03	1.49	2.00	10.78	0.12	1.61	11.30	0.18	0.06	10.65	0.17	1.60
11.03	0.16	0.08	11.08	0.22	1.59	10.90	0.11	0.08	11.58	0.77	1.67	10.78	0.15	0.08
11.28	0.75	1.73	11.18	0.22	0.07	11.15	0.70	1.75	11.73	1.35	0.05	11.00	0.71	1.75
11.45	1.35	0.07	11.43	0.78	1.70	11.30	1.30	0.06	12.03	1.43	1.67	11.15	1.36	0.07
11.73	1.44	1.69	11.60	1.41	0.07	11.58	1.38	1.67	12.05	0.31	0.13	11.43	1.43	1.68
11.75	0.29	0.14	11.83	0.35	0.03	11.60	0.24	0.14	12.10	0.30	0.09	11.45	0.28	0.13
11.83	0.28	0.09	11.88	1.49	1.67	11.68	0.23	0.08	12.38	0.04	9.43	11.53	0.27	0.08
12.08	0.04	9.55	11.88	0.35	0.10	11.90	1.49	9.58	12.40	1.31	0.02	11.75	0.03	9.59
12.55	0.14	1.61	11.95	0.34	0.08	11.95	1.26	0.02	12.88	0.14	1.58	12.08	0.01	0.30
12.63	1.50	0.03	12.20	0.09	9.79	12.43	0.08	1.60	12.98	0.01	0.03	12.25	0.13	1.60
12.68	0.31	0.29	12.70	0.19	1.59	12.48	1.44	0.37	13.00	0.32	0.28	12.33	0.01	0.03
12.78	1.34	0.04	12.80	0.07	0.03	12.53	0.26	0.28	13.10	1.34	0.04	12.35	0.31	0.28
12.90	0.04	4.25	12.83	0.39	0.29	12.63	1.29	0.04	13.23	0.02	4.40	12.45	1.35	0.04
12.90	1.49	0.35	12.93	1.41	0.04	12.73	1.47	4.51	13.28	0.37	0.67	12.58	0.03	4.35
12.95	0.40	0.69	13.05	0.10	4.19	12.80	0.30	0.70	13.30	0.10	8.76	12.58	1.50	0.26
12.98	0.11	8.88	13.08	0.06	0.30	12.83	0.04	8.71	13.33	1.48	0.12	12.63	0.40	0.69
13.10	0.38	0.04	13.13	0.18	8.57	12.95	0.29	0.05	13.43	0.37	0.04	12.65	0.11	8.71

Reference (cont.)			Sample 1 (cont.)			Sample 2 (cont.)			Sample 3 (cont.)			Sample 4 (cont.)		
Dim 1 (min)	**Dim 2 (s)**	**Volume (%)**	**Dim 1 (min)**	**Dim 2 (s)**	**Volume (%)**	**Dim 1 (min)**	**Dim 2 (s)**	**Volume (%)**	**Dim 1 (min)**	**Dim 2 (s)**	**Volume (%)**	**Dim 1 (min)**	**Dim 2 (s)**	**Volume (%)**
13.20	0.40	2.33	13.13	0.50	0.68	13.08	0.30	3.24	13.53	0.38	1.98	12.78	0.38	0.04
13.25	0.39	0.92	13.25	0.47	0.05	13.13	0.02	0.64	13.58	0.38	1.18	12.88	0.39	3.25
13.28	0.09	0.64	13.35	0.47	1.81	13.20	0.28	0.23	13.60	0.08	0.57	12.95	0.09	0.62
13.35	0.37	0.24	13.40	0.48	1.34	13.25	0.33	0.21	13.68	0.37	0.22	13.03	0.37	0.27
13.43	0.42	0.19	13.43	0.16	0.55	13.40	0.36	0.08	13.75	0.41	0.20	13.10	0.43	0.18
13.58	0.45	0.08	13.53	0.46	0.27	13.45	0.37	0.07	13.90	0.44	0.06	13.23	0.45	0.07
13.63	0.47	0.07	13.60	0.51	0.18	13.65	0.39	0.09	13.98	0.46	0.06	13.30	0.47	0.05
13.83	0.48	0.09	13.75	0.55	0.07	13.83	0.41	0.09	14.15	0.48	0.09	13.50	0.49	0.09
14.00	0.51	0.08	13.80	0.56	0.07	17.03	0.67	2.88	14.35	0.50	0.09	13.68	0.51	0.08
17.35	0.79	2.85	14.00	0.57	0.08				17.68	0.76	2.77	16.98	0.77	2.85
			14.20	0.60	0.09									
			17.68	0.89	2.75									


(a)
Second dimension (s)
First dimension (min)


(b)
Second dimension (s)
First dimension (min)


(c)
Second dimension (s)
First dimension (min)

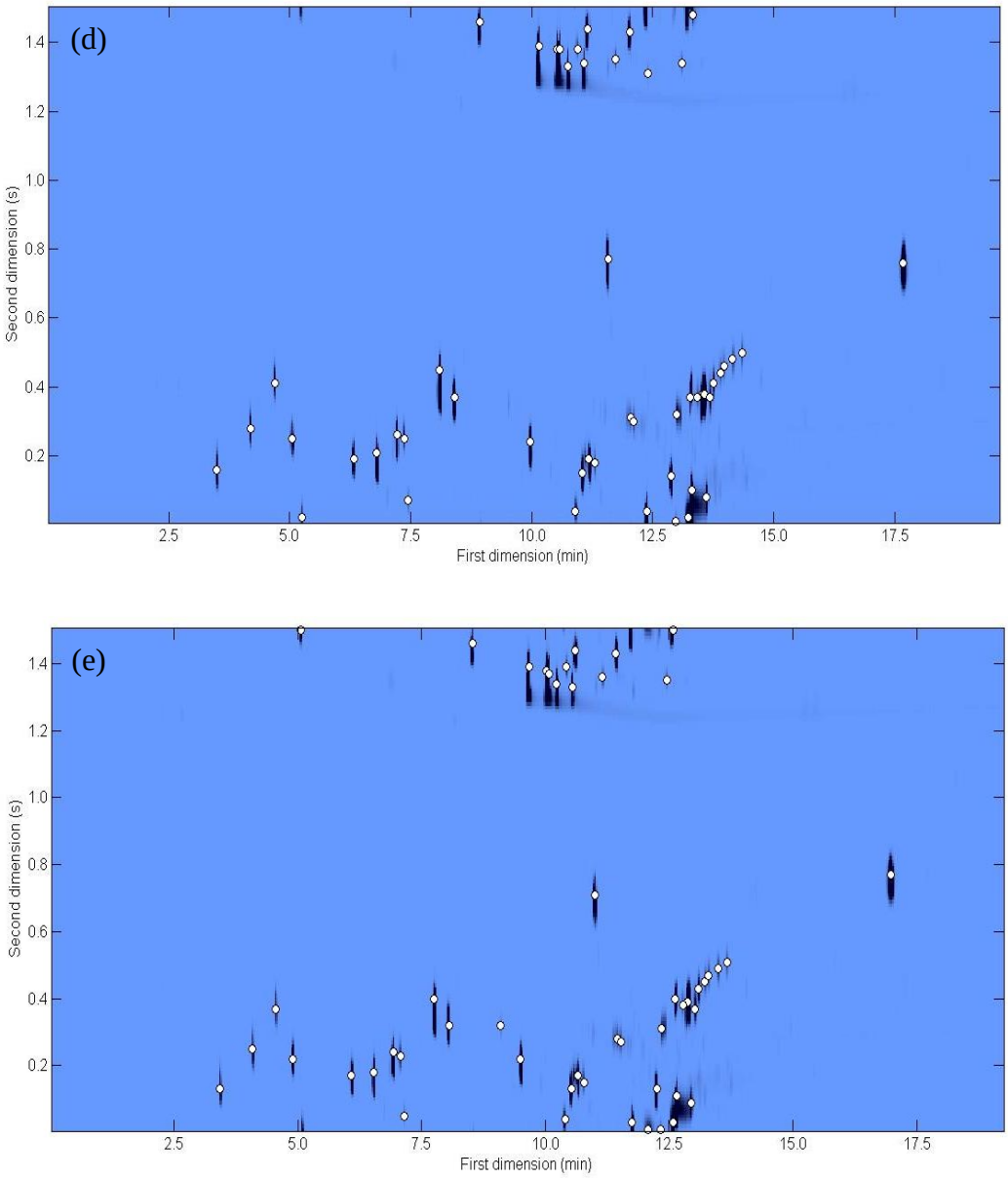


Figure 4.12: 2D images with peaks identified (a) reference, (b) sample 1, (c) sample 2, (d) sample 3 and (e) sample 4

Following peak detection, the reference and sample control points are selected. The reference control points are selected as described above and shown below in Figure 4.13. The 24 segments of the chromatogram are marked for illustration purposes and the 11 highlighted peaks represent the selected reference control points.

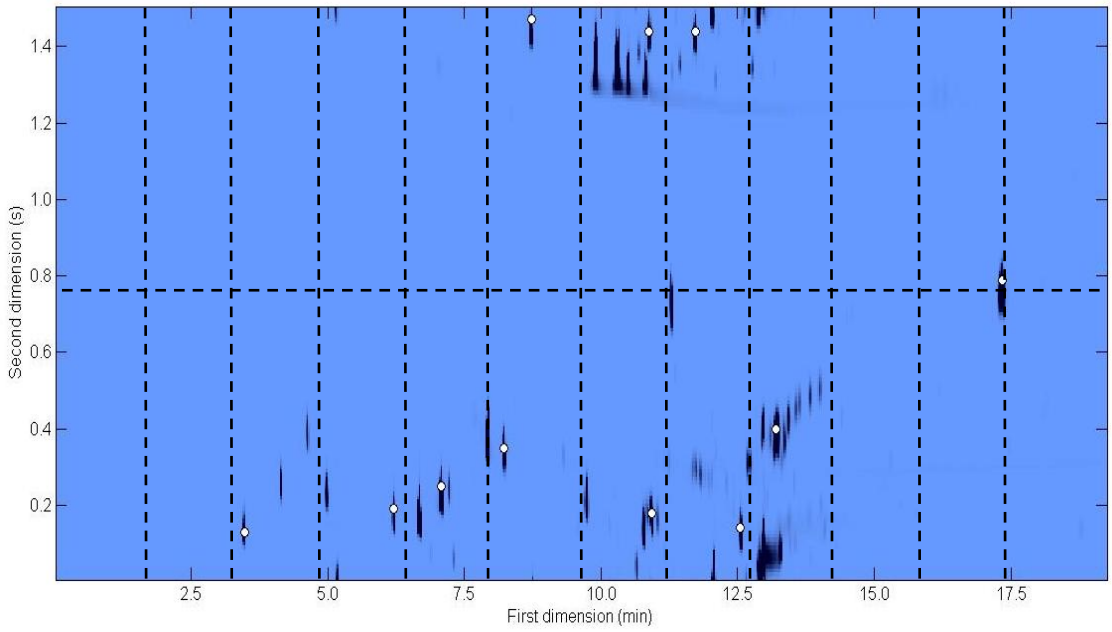


Figure 4.13: 11 selected reference control points, the 24 segments are marked for illustration purposes

The coordinates corresponding to the 11 control points provide 165 triangles against which a sample chromatogram can be compared.

Consider sample 4 with the increased temperature ramp. For the complete set of 51 peaks identified in the sample chromatogram, 20,825 discrete triangles can be formed. However, to minimise problems associated with small peaks, triangles are only formed using sample peaks with intensities greater than 20% of the minimum reference control point. This produces a list of 11,480 sample triangles.

Some triangles are poorly suited for comparison. Triangles with large length ratios, R>8, and cosine, C, values greater than 0.99 are discarded from both the reference and sample triangle lists. Such elongated and flattened triangles can be falsely matched, which weakens the algorithms discriminating ability [341]. Figure 4.14 shows the distribution of R and C values for the triangles derived from the sample peak coordinates, the filtering criteria, R>8 and C>0.99 are highlighted for illustration purposes. The results of filtering provide 131 reference triangles and 8,522 sample triangles for matching, requiring a total of 1,116,382 comparisons.

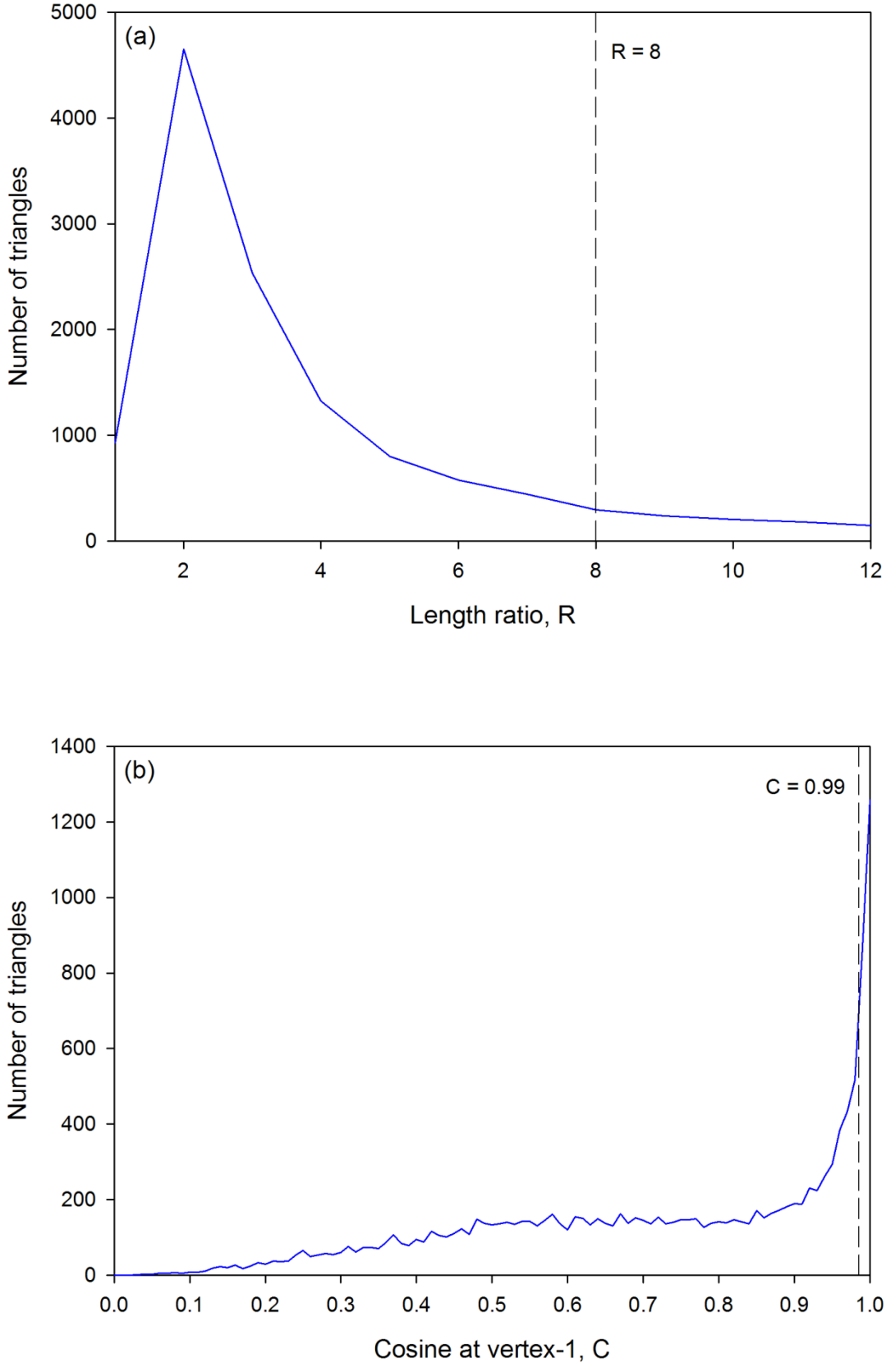


Figure 4.14: Distribution of (a) length ratios and (b) cosines at vertex-1, for the triangles derived from the coordinates of the sample 4 chromatogram

The time taken to construct the list of matching triangles can be dramatically reduced by filtering the sample triangles during the triangle comparison algorithm; this also increases the efficiency and effectiveness of pattern matching. Examining triangle orientation eliminates approximately half of the comparisons. Comparing closeness of the triangle centres as being within an acceptable level (+/- 20 in dimension 1) and eliminating all sample triangles with perimeters different to the reference triangles by more than 10% reduces subsequent comparisons to around 20,000. The list of acceptable matches is further reduced by comparison of the rotation angle, side-length ratio and vertex-1 cosine. The list of matching triangles obtained contains the coordinates of every sample triangle that is matched to each reference triangle. The number of times each coordinate appears in a match is referred to as the vote for that coordinate. The votes are then examined and those matches having the coordinates with the highest number of votes are retained. This results in a final list of coordinates that is used to define the affine transform between this list and the corresponding reference control points. Results of triangle matching are provided in Table 4.7.

Table 4.7: Results of triangle matching. List of matched reference and sample 4 peaks (first and second dimension locations) with the vote

Reference		**Sample 4**		
Dim 1 (min)	**Dim 2 (s)**	**Dim 1 (min)**	**Dim 2 (s)**	**Vote**
3.48	0.13	3.43	0.13	62
6.20	0.19	6.08	0.17	60
7.08	0.25	6.93	0.24	55
8.23	0.35	8.05	0.32	47
8.73	1.47	8.53	1.46	55
10.88	1.44	10.60	1.44	43
10.93	0.18	10.65	0.17	49
11.73	1.44	11.43	1.43	65
12.55	0.14	12.25	0.13	27
13.20	0.40	12.88	0.39	31
17.35	0.79	16.98	0.77	84

Of the 11 control points identified in the original reference data, all are correctly matched in the sample data. For the 51 sample peaks, the maximum number of votes

that a peak can receive is 1225 (Equation 4.10), however the highest vote obtained was 84. This suggests that many of the triangles were filtered out by the tolerances defined in Table 4.4, which improves the efficiency and accuracy of the algorithm.

The list of matching reference and sample control points are then used to calculate the affine transformation matrix of translation, scaling and rotation parameters, which is then applied to all 51 peaks in the sample chromatogram. This transformation could be applied to the complete chromatogram to generate an array or image for visual comparison. Figure 4.15 shows the reference chromatogram with the sample peaks plotted before and after alignment with the affine transform. From this it can be seen that the sample peaks are much closer after alignment than before, which indicates that the selected control points and the affine transformation were successful for aligning the reference and sample data.

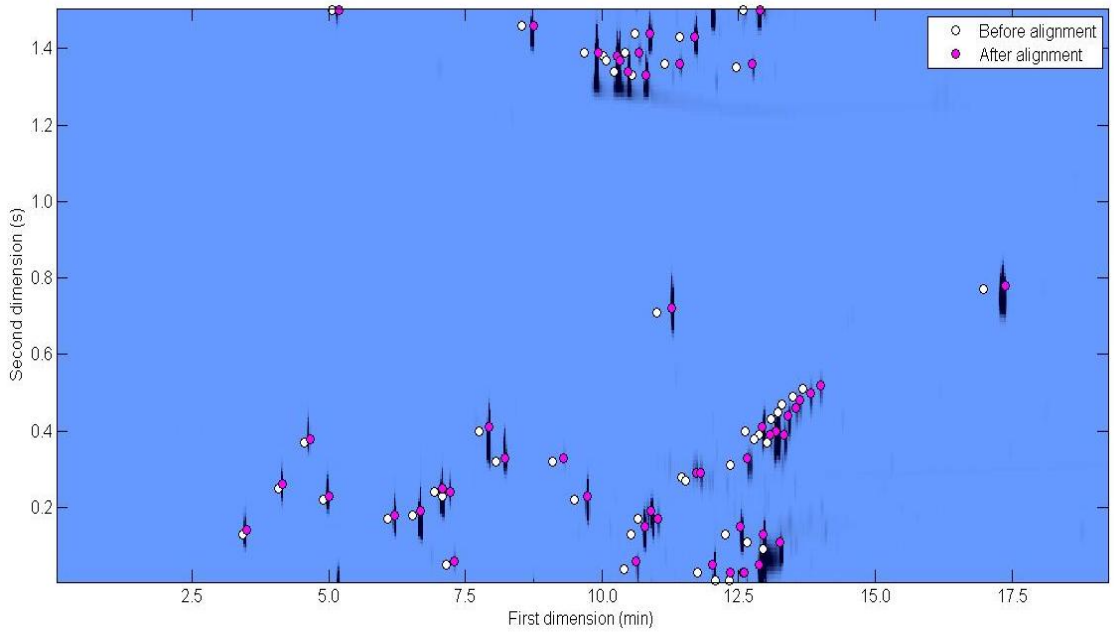


Figure 4.15: Results of alignment. Reference chromatogram with sample 4 peaks plotted before and after alignment via affine transformation

The remaining sample chromatograms (samples 1-3) were compared to the reference chromatogram in a similar manner using the same 11 control points. The results of triangle matching are provided in Table 4.8 and the subsequent alignment results are shown in Figure 4.16.

Table 4.8: Results of triangle matching. List of matched reference and sample (1-3) peaks (first and second dimension locations) with the vote

Reference		Sample 1			Sample 2			Sample 3		
Dim 1 (min)	**Dim 2 (s)**	**Dim 1 (min)**	**Dim 2 (s)**	**Vote**	**Dim 1 (min)**	**Dim 2 (s)**	**Vote**	**Dim 1 (min)**	**Dim 2 (s)**	**Vote**
3.48	0.13	3.60	0.20	56	3.35	0.09	70	3.50	0.16	58
6.20	0.19	6.35	0.24	42	6.08	0.13	65	6.33	0.19	56
7.08	0.25	7.23	0.29	46	6.93	0.20	66	7.23	0.26	62
8.23	0.35	8.38	0.40	47	8.08	0.30	57	8.40	0.37	50
8.73	1.47	9.88	0.28	4	8.58	1.41	60	8.93	1.46	62
10.88	1.44	11.03	1.49	37	10.73	1.39	50	11.15	1.44	46
10.93	0.18	11.08	0.22	51	10.90	0.11	49	11.05	0.15	48
11.73	1.44	11.88	1.49	44	11.58	1.38	61	12.03	1.43	54
12.55	0.14	12.70	0.19	26	12.43	0.08	21	12.88	0.14	22
13.20	0.40	13.35	0.47	20	13.25	0.33	21	13.28	0.37	20
17.35	0.79	17.68	0.89	66	17.03	0.67	67	17.68	0.76	58

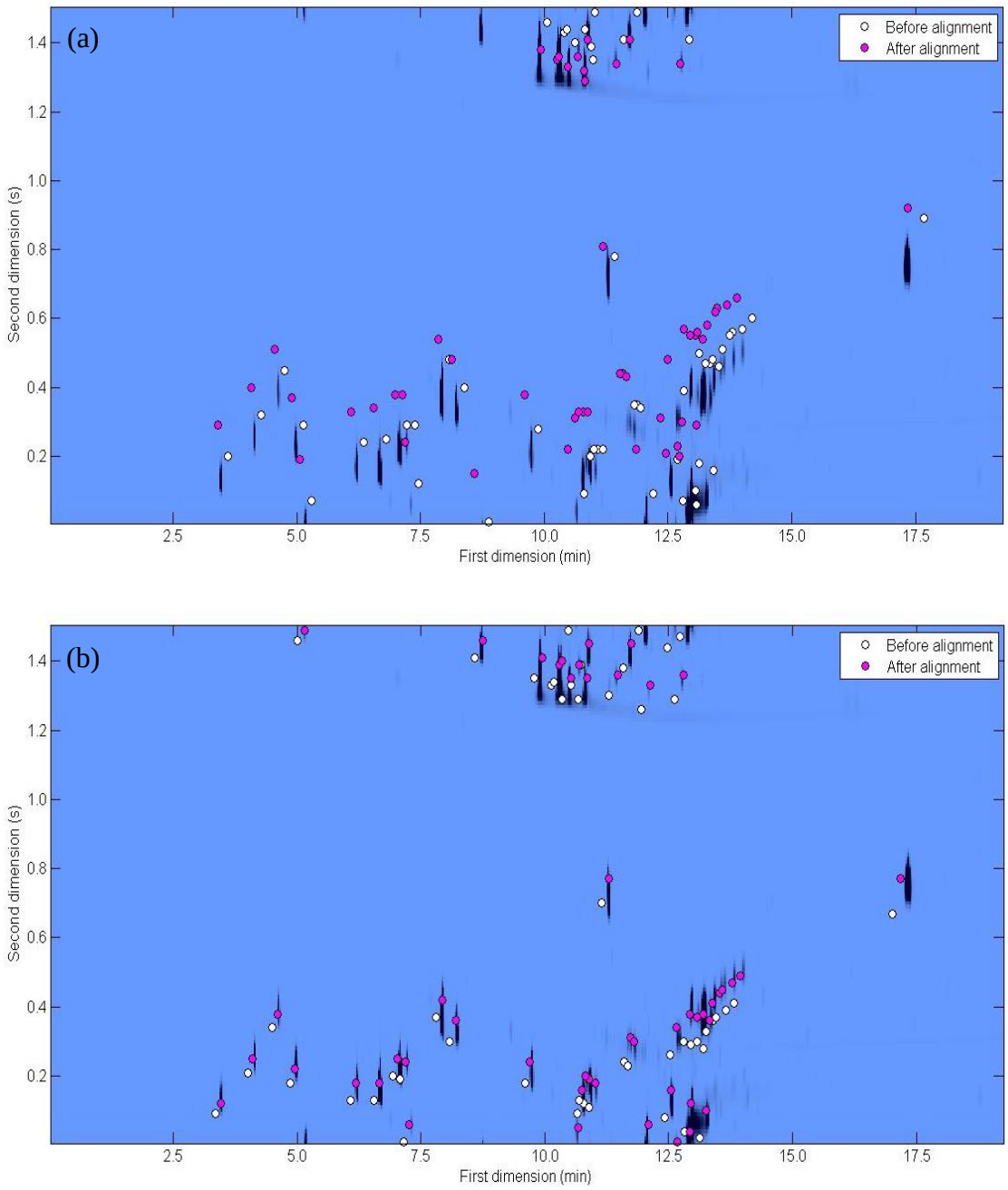


Figure 4.16: Results of alignment. Reference chromatogram with sample peaks plotted before and after alignment via affine transformation (a) sample 1, (b) sample 2

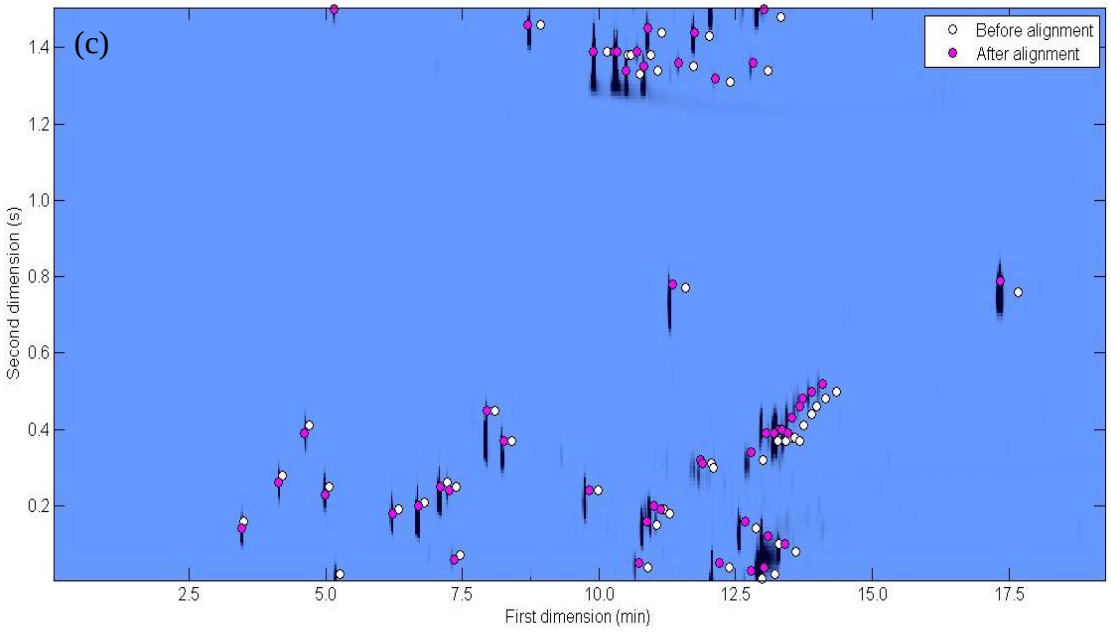


Figure 4.16: Results of alignment. Reference chromatogram with sample peaks plotted before and after alignment via affine transformation (c) sample 3

All of the sample peaks in Figure 4.16 appear to be closer to the reference peaks after alignment, except for the sample chromatogram obtained with the reduced flow rate (sample 1). This may be due to an incorrect control point match and as a result the reference control points and matched sample 1 control points are plotted for inspection (Figure 4.17). The circled peaks are incorrectly matched due to wrap-around. As the circled reference control point is wrapped around in the sample 1 chromatogram, the triangle patterns are different and can no longer be correctly matched according to their geometric properties. It can also be seen in Table 4.8 that the vote for this peak (9.88,0.28) is very low, which can indicate an incorrect match.

Figure 4.16 also indicates that although affine transformation is able to sufficiently align the reference and sample peaks, the alignment results may be inadequate. This is due to the fact that the affine transform is a global solution and is therefore not exact for individual peaks.

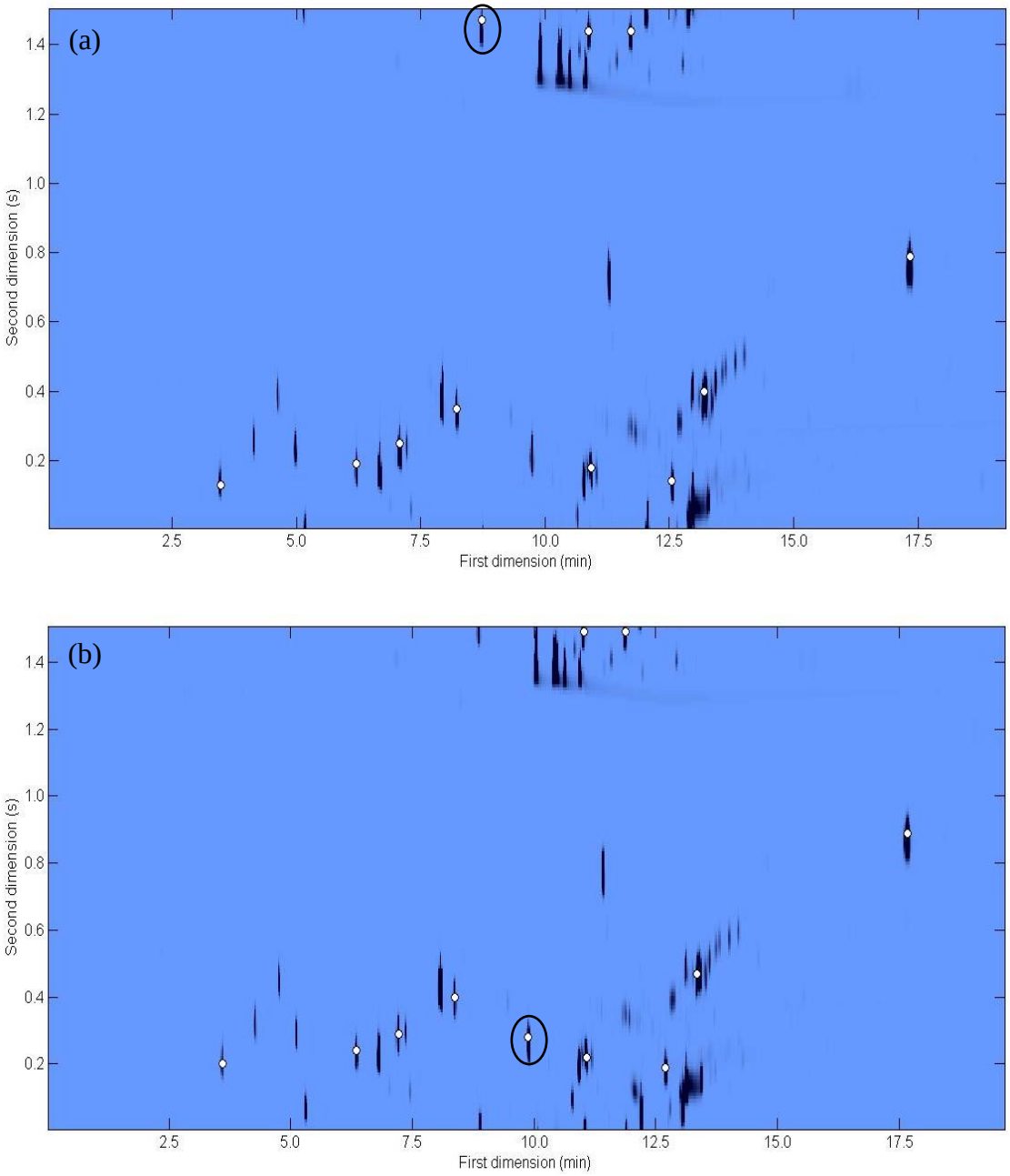


Figure 4.17:(a) reference control points and (b) sample 1 control points. The circled peaks are the incorrectly matched control points

The final output produced from the algorithm is a list of aligned peaks between the reference and sample chromatograms allowing direct comparison of analytes within samples.

4.5 Conclusion

The comparison of GC×GC chromatograms requires the analytical data to be aligned, it has been demonstrated in this chapter that this can be achieved automatically using a pattern-matching algorithm originally developed for comparing two-dimensional star maps. A major aspect of this algorithm is its ability to compare coordinate lists of different size. Adopted for GC×GC data, this feature is exploited to identify the correspondence between a small number of control points taken from a reference chromatogram and a larger list of coordinates from a sample chromatogram; this dramatically reduces computation time. Once corresponding reference and sample control points are identified, they are used to calculate the translation, scaling and rotation operations for the global affine transform. The affine transformation is then applied to the sample data in order to align reference and sample peaks.

Problems can arise for GC×GC data when peak wrap-around occurs. Wrap-around changes the triangle patterns and as a result they can no longer be correctly matched according to their geometric properties and false matches can occur. Despite the affine transform being a simple method for alignment, it is a global solution and is therefore not exact for individual peaks. These issues will be looked at in the next chapter.

Chapter 5 - GC×GC Quality Control Software: Data Comparison

5.1 Introduction

In the previous chapter software was described for the automated alignment of GC×GC chromatograms. A simple method of partitioning the reference chromatogram was employed to aid selection of reference control points that cover the pattern space. These control points were then compared with the entire sample peak coordinate list using a triangle pattern matching algorithm, originally developed for the comparison of star maps, to identify corresponding points in the sample chromatogram. Once reference and sample control points were identified, they were used to calculate the translation, scaling and rotation operations for an affine transform. The affine transformation was then applied to the complete sample peak list in an attempt to align reference and sample peaks.

A few problems were identified with use of the software, these were generally associated with peak wrap-around and the use of a global affine transform for alignment. These issues are addressed in this chapter as well as further extending the software to perform sample comparisons for the purpose of QC. Comparison of reference and sample chromatograms is achieved through the use of fuzzy logic with a trapezoidal membership function. Fuzzy comparisons are ideally suited for this type of analysis, they are flexible and, as a result, can compensate for the lack of accuracy obtained with an affine transform. The developed software is employed to analyse a number of real flavour samples for QC purposes.

5.2 Experimental

5.2.1 Samples

The samples examined to evaluate and test the software are a model fragrance and real flavour samples provided by Firmenich (Firmenich SA, Meyrin, Switzerland).

5.2.2 Instrumentation

The GC×GC system was based on an Agilent 6890 GC (Agilent Technologies, Wilmington, DE, USA) equipped with a split-splitless injector and a FID detector. A two-stage double loop modulator (ZX1; Zoex Corporation, TX, USA) was installed in the GC oven. This modulator consists of a cold jet (nitrogen gas cooled by liquid nitrogen) and a hot jet (heated air, at the temperature of the GC oven temperature + 150°C, duration = 350 ms) positioned orthogonal to each other. A double trapping loop was positioned in the flow of both jets. The cold jet operated constantly to trap compounds within the double loop assembly, whilst the hot jet pulses periodically, acting to both divert the flow of the cold jet to release trapped compounds, and to heat the cold spot to actively remobilise the trapped compounds more quickly [57, 338]. The modulation period was 1.5 s. The first dimension column (HP-FFAP 15 m × 0.25 mm × 0.25 µm; Agilent J&W) was linked to the second column (DB-1 1 m × 0.1 mm × 0.1 µm; Agilent J&W) via a deactivated silica column (called the transfer line, 0.1 mm i.d., 1.75 m). The transfer line was installed in the modulator in a double loop configuration. Press fits were used between the first column and the transfer line and between the transfer line and the second column. The trapping of the compounds was performed in the transfer line. The GC inlet was heated to 250°C, and a 0.1 µL injection volume (using a 1 µL syringe, Hamilton 7001N, ga 0.47/70mm/pst 2, P/N 80135/01) was used with a split ratio of 30:1. For the reference data helium was used as carrier gas at a flow of 1 mL min^{-1}, in constant flow mode. The oven temperature program started from 40°C then increased at 15°C min^{-1} up to 230°C with a 7 min hold. A flame ionisation detector was used at 250°C, with nitrogen makeup gas, and a data acquisition rate of 100 Hz.

Further sample fragrance chromatograms were obtained by varying the flow rate and temperature ramp in order to generate chromatograms with induced peak shifts (Table 4.1).

5.2.3 Data processing

All data manipulation and analysis algorithms were developed and implemented in-house using Matlab (V7.10 (R2010a), MathWorks Inc, MA, USA). The raw data from the GC×GC system were acquired using Agilent ChemStation vE01.01.335 (Agilent Technologies). It was then exported in comma-separated values format (.csv) using an in-house macro and imported into Matlab for processing.

5.3 Program development

5.3.1 Improved control point selection

The alignment process using an affine transformation is made more stable by increasing the number of control points employed. Since it is a least squares fit, a larger number of control points reduces the effects outliers and provides for a more accurate equation of fit. In order to ensure the analytical data is more evenly spread across the pattern space, and hopefully increase the number of selected control points, the reference chromatogram is rotated to occupy the centre of the pattern space.

As in the previous chapter, the reference chromatogram is that obtained from a model fragrance analysed under standard conditions (section 5.2.2). To rotate the chromatogram to the centre of the pattern space, the reference chromatogram is projected in the second dimension by summing along all columns in the data matrix (Equation 5.1).

$$ref\_proj_i = \sum_{j=1}^{J} X_{i,j} \qquad \text{Equation 5.1}$$

This provides a profile of the reference which is used to identify the spread of the peaks. A parabola is then constructed and the correlation between the parabola and the reference profile is calculated. Figure 5.1 shows the parabola with the reference profile, it can be seen that the spread of the peaks lies on the edges of the parabola and the objective is to shift these peaks to the centre of the parabola and hence the centre of the chromatogram, which allows for selection of more, better placed reference control points.

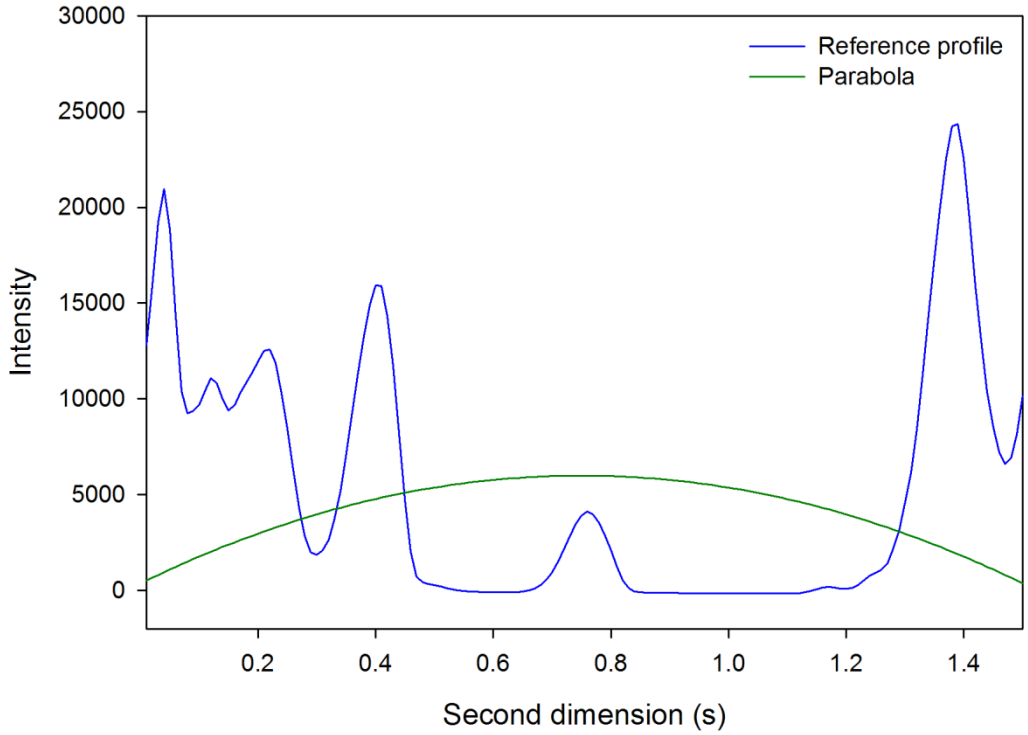


Figure 5.1: Parabola and reference profile

To centre the chromatogram, the reference profile vector is incrementally shifted and correlated with the parabola. The shift required to achieve the highest correlation, i.e. the position that centres the chromatogram, is recorded. This shift is referred to as the offset. The complete reference chromatogram as a 1D vector is then circularly shifted by the offset, resulting in a chromatogram where the peaks occupy the centre of the pattern space. Figure 5.2 shows the parabola with the shifted reference profile and the reference chromatogram before and after rotation is shown in Figure 5.3.

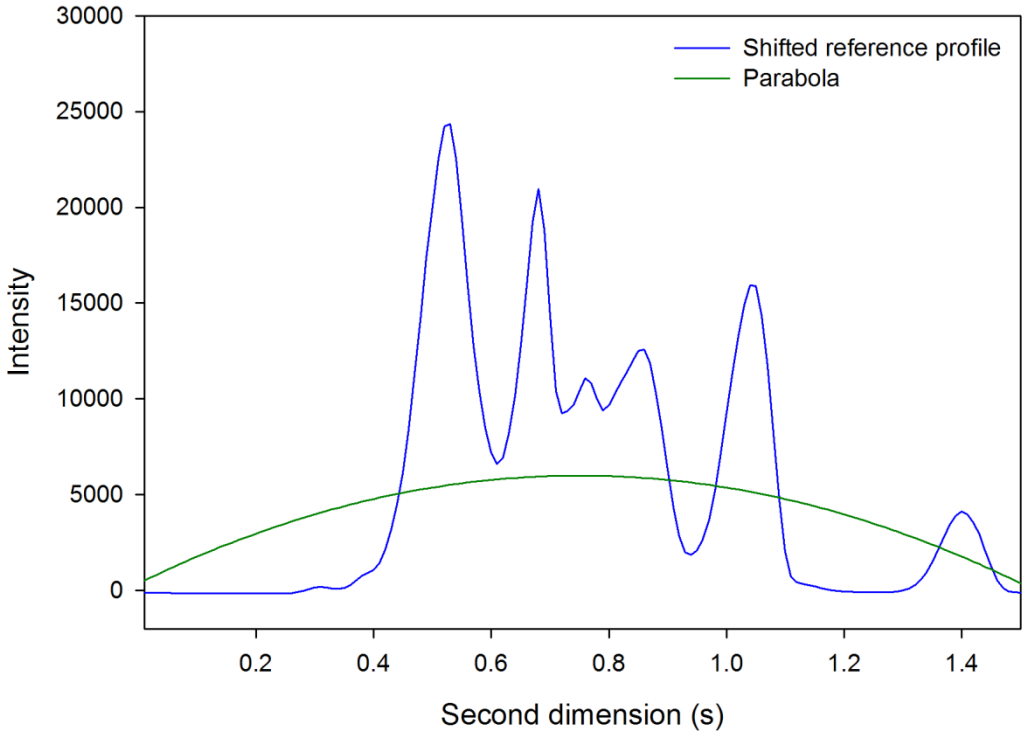


Figure 5.2: Parabola and shifted reference profile

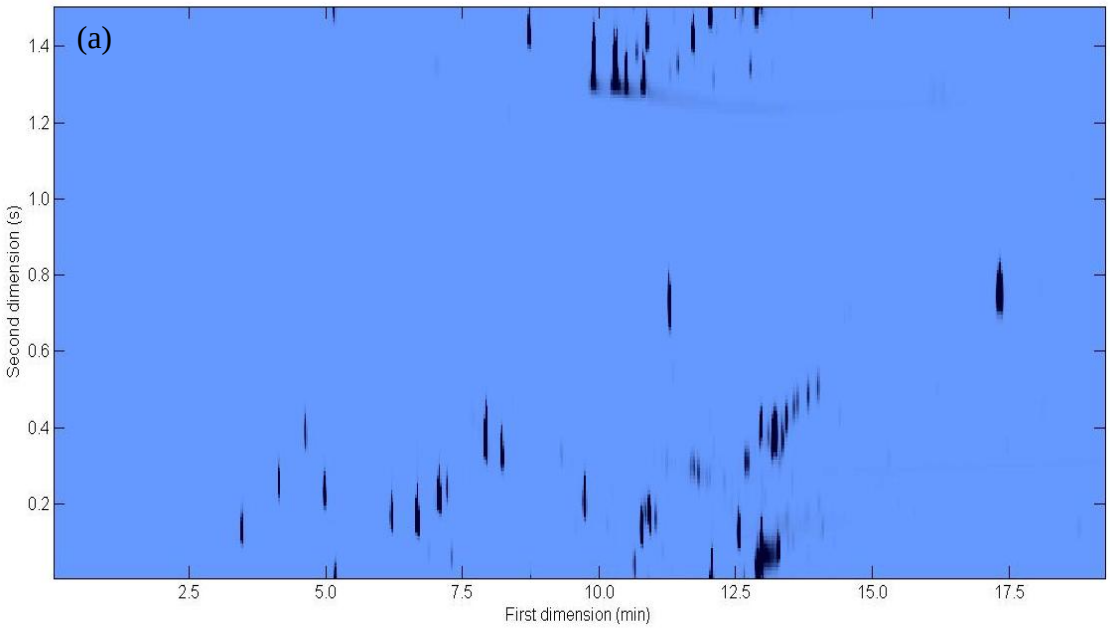


Figure 5.3: (a) original reference chromatogram

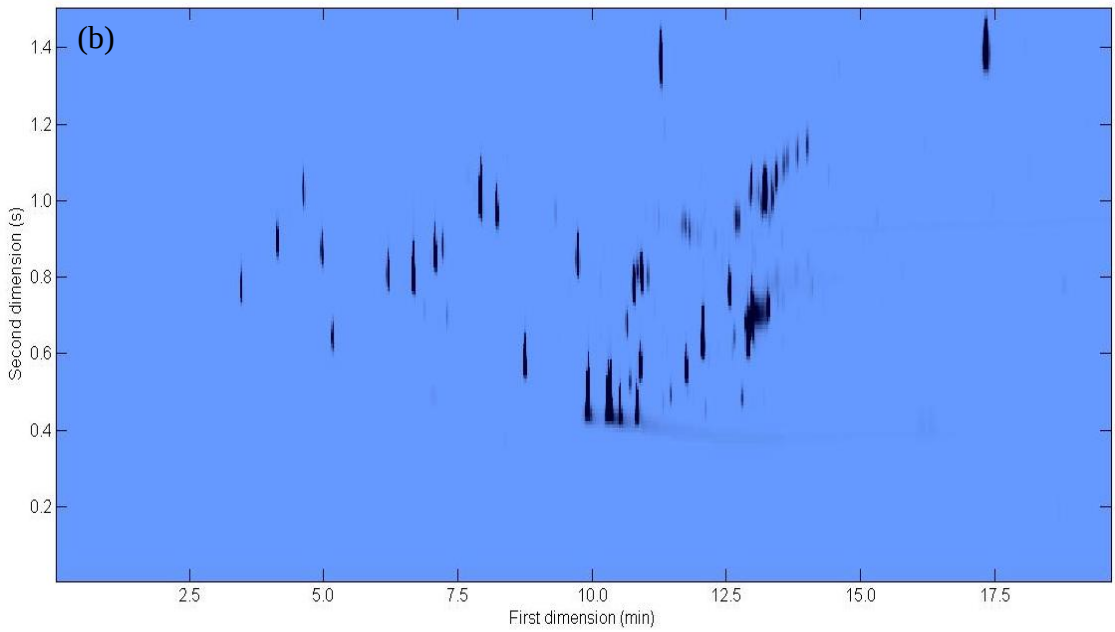


Figure 5.3: (b) rotated reference chromatogram

The control points are then selected as described in Chapter 4 and shown in Figure 5.4. From this it can be seen that there are now 14 control points, compared with the 11 selected using the original version of the software (Chapter 4). A higher number of control points is better for the affine transform as it gives it more points across the pattern space to calculate the transformation. The affine transformation is improved by a higher number of control points as it is a least squares fit and more points results in a better equation approximation.

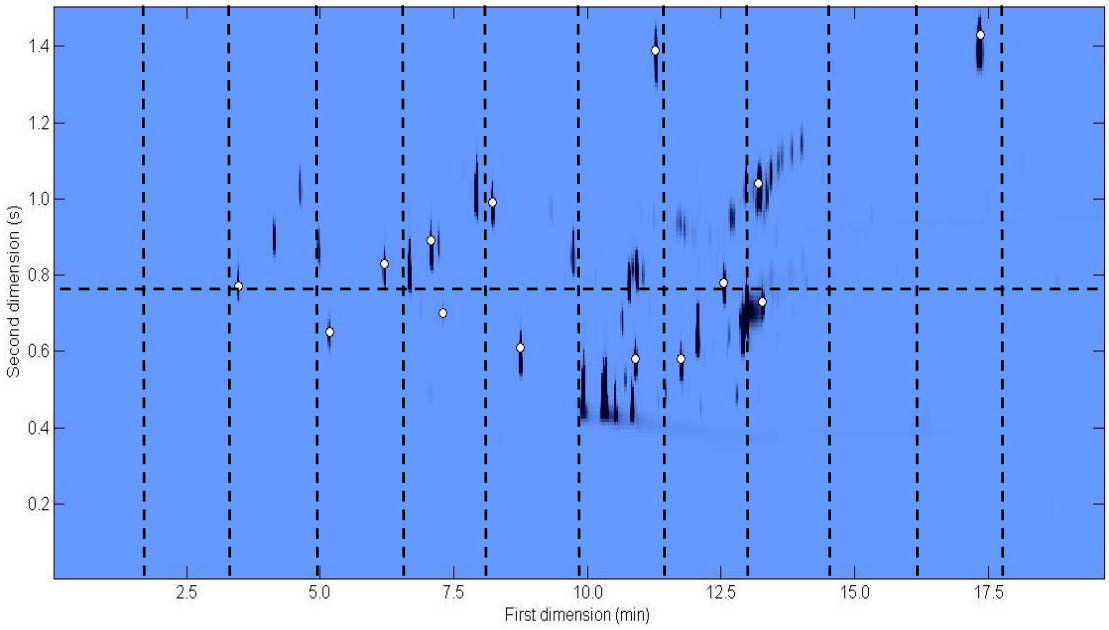


Figure 5.4: 14 selected reference control points, the 24 segments are marked for illustration purposes

5.3.2 Wrap-around

Wrap-around occurs when the separation in the second-dimension is not finished before the next fraction is injected. Thus, the second dimension retention time exceeds the modulation period and part or all of a peak appears in a later modulation than that in which it was injected. This is illustrated in Figure 5.5 and it can be seen that the same highlighted peak has been wrapped around between the two chromatograms. Wrap-around causes problems in pattern matching as the second dimension retention times are significantly different, in the example below the peak in the first chromatogram is located at 8.73,1.47 and the second chromatogram it is at 8.88,0.01. This results in different patterns and hence different triangle properties, which will cause incorrect peak matching.

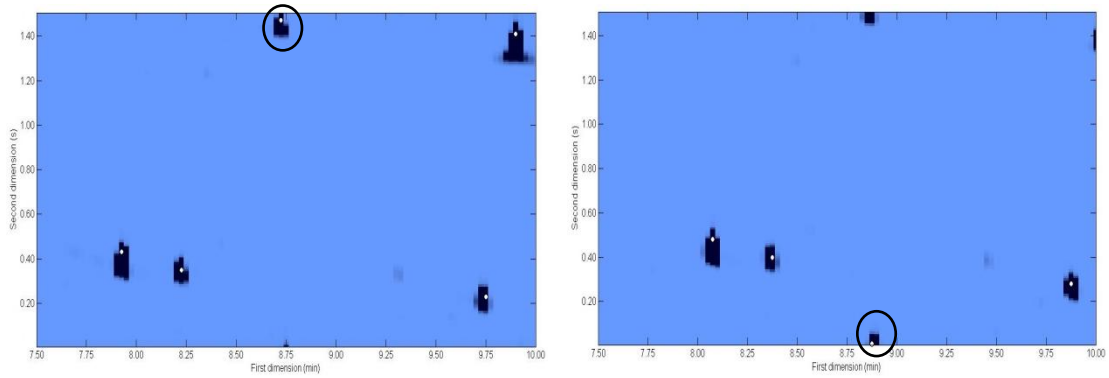


Figure 5.5: Example of wrap-around. The circled peak is wrapped around between the two chromatograms

In order to correct the effects of wrap-around, the sample chromatogram is aligned with the reference in the centre of the pattern space. Since the reference and sample are aligned in the second dimension, the selected reference control points will not be wrapped around in the sample chromatogram and as a result can be accurately matched. The sample used to illustrate this part of the algorithm is the sample with the decreased flow rate (sample 1 in Table 4.1), as wrap-around issues were observed for this sample in Chapter 4.

To align the sample chromatogram with the reference in the centre of the pattern space, the sample profile is found in the same manner as the reference, by summing the chromatogram in the second dimension (Equation 5.1). The correlation between the new, shifted, reference profile and the sample profile is calculated and the offset between the reference and sample is defined. Figure 5.6 shows the sample profile and shifted reference profile.

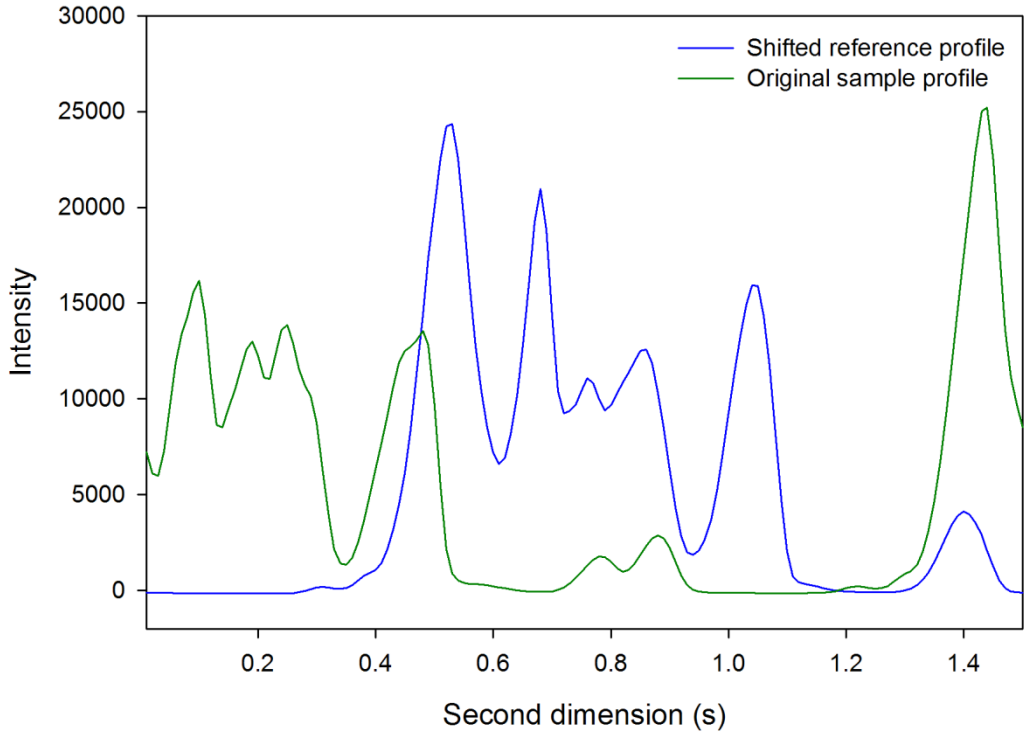


Figure 5.6: Original sample profile and shifted reference profile

The sample as a 1D vector is then circularly shifted by the offset, to provide a sample chromatogram that is aligned with the reference in the centre of the pattern space (Figure 5.7). The rotated sample chromatogram is shown in Figure 5.8.

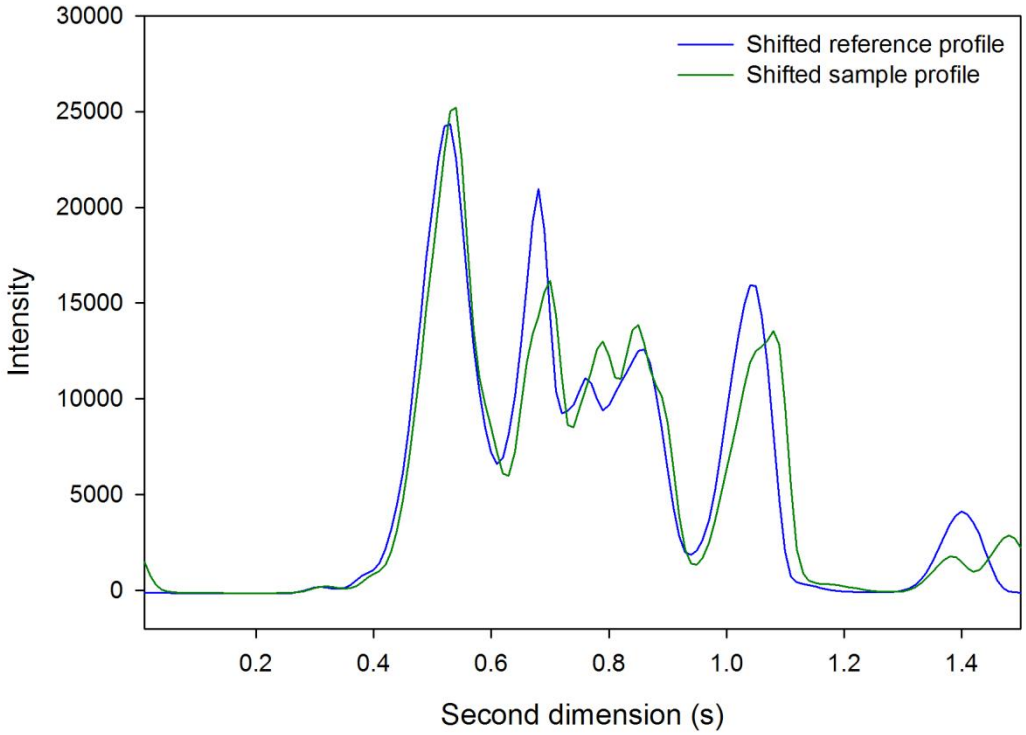


Figure 5.7: Shifted reference and sample profiles

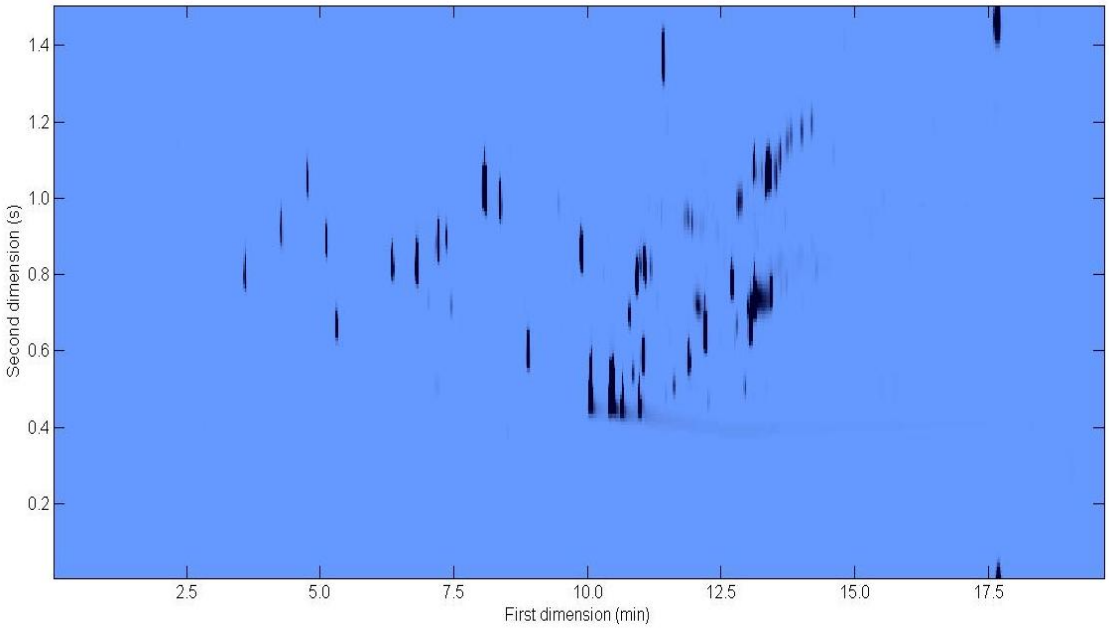


Figure 5.8: Rotated sample chromatogram

5.3.3 Matching reference and sample control points

The 14 control points selected from the new, rotated, reference chromatogram as described above are matched to the entire peak list from the sample. The sample peaks are detected as described in Chapter 4.

The triangles are then formed for the reference control points and the entire sample list and the triangles are filtered to remove triangles poorly suited for matching as previously described. Finally, the votes are examined and the matches with the highest number of votes are assumed to be the matching sample control points. The results of triangle matching are provided in Table 5.1.

Table 5.1: Results of triangle matching. List of matched reference and sample peaks (first and second dimension locations) with the vote

Reference		**Sample**		
Dim 1 (min)	**Dim 2 (s)**	**Dim 1 (min)**	**Dim 2 (s)**	**Vote**
3.48	0.77	3.60	0.80	63
5.18	0.65	5.30	0.67	123
6.20	0.83	6.35	0.84	57
7.08	0.89	7.23	0.89	52
7.30	0.70	7.45	0.72	80
8.23	0.99	8.38	1.00	50
8.75	0.61	8.88	0.61	70
10.90	0.58	11.05	0.59	115
11.28	1.39	11.43	1.38	40
11.75	0.58	11.90	0.59	83
12.55	0.78	12.70	0.79	59
13.20	1.04	13.35	1.07	74
13.28	0.73	13.43	0.76	41
17.35	1.43	17.68	1.49	215

Of the 14 reference control points, all are correctly matched in the sample. This would suggest that rotating the reference and sample chromatograms to occupy the centre of the pattern space has removed the effect of wrap-around as well as improving the affine transformation by increasing the number of control points. Many of the votes

are also higher than that of the un-rotated chromatograms (Tables 4.7 and 4.8), which indicates that more triangles were able to be accurately matched.

5.3.4 Reducing false matches

Another issue observed in the previous chapter was that of incorrect matching, these matches were identified as having a low vote and significantly different coordinates. In order to correct for this a vote cut off and/or coordinate tolerance can be set. First, a vote cut-off is evaluated. A vote cut off was applied by both Groth [337] and Arzoumanian et al. [341], where matches were removed if the vote dropped by a factor of two. In this work, a vote cut-off was set at half of the maximum vote; anything lower than half of the maximum vote was removed. This vote cut-off was applied to all four sample chromatograms and the results are provided in Table 5.2. From this it can be seen that too many matches are removed and in the case of sample 2 only two matches remain. Two control points are not sufficient for an affine transform, which requires a minimum of three points.

Applying a tolerance on the coordinates was also evaluated. If the coordinates are greater than twice the average tolerance (Equation 5.2) apart, they are removed. The results for the four samples are provided in Table 5.3. The results indicate that the last peak at 17.35,1.43 (reference) is removed from sample 1 and 2, despite being a correct match. However, peak coordinates that are too far apart are not recommended for an affine transform as it will adversely affect the transformation. Since the affine transformation is a least squares fit, peaks that deviate significantly result in a poor approximation as the method is not robust. For this reason, the coordinate tolerance was subsequently used in the algorithm as a means of reducing false matches.

$$T_{coord} = \sqrt{[cycle\ (M_{1x} - M_{2x})]^2 + \left(M_{1y} - M_{2y}\right)^2} \qquad \text{Equation 5.2}$$

Where *cycle* is defined in Equation 4.1, *x* and *y* are the first and second dimension locations and 1 and 2 refer to the reference and sample, respectively.

Table 5.2: Results of vote cut off for the reference and four sample chromatograms

Reference		Sample 1			Sample 2			Sample 3			Sample 4		
Dim 1 (min)	**Dim 2 (s)**	**Dim 1 (min)**	**Dim 2 (s)**	**Vote**	**Dim 1 (min)**	**Dim 2 (s)**	**Vote**	**Dim 1 (min)**	**Dim 2 (s)**	**Vote**	**Dim 1 (min)**	**Dim 2 (s)**	**Vote**
5.18	0.65	5.30	0.67	123	5.03	0.66	124	5.25	0.68	125	5.08	0.66	111
7.30	0.70										7.15	0.71	70
8.75	0.61										8.55	0.62	73
10.90	0.58	11.05	0.59	115				11.18	0.60	109	10.63	0.60	99
13.20	1.04										12.78	1.04	79
17.35	1.43	17.68	1.49	215	17.03	1.37	200	17.68	1.42	210	16.98	1.43	137

Table 5.3: Results of coordinate tolerance for the reference and four sample chromatograms

Reference		Sample 1			Sample 2			Sample 3			Sample 4		
Dim 1 (min)	**Dim 2 (s)**	**Dim 1 (min)**	**Dim 2 (s)**	**Vote**	**Dim 1 (min)**	**Dim 2 (s)**	**Vote**	**Dim 1 (min)**	**Dim 2 (s)**	**Vote**	**Dim 1 (min)**	**Dim 2 (s)**	**Vote**
3.48	0.77	3.60	0.80	63	3.35	0.79	52	3.50	0.82	57	3.43	0.79	60
5.18	0.65	5.30	0.67	123	5.03	0.66	124	5.25	0.68	125	5.08	0.66	122
6.20	0.83	6.35	0.84	57	6.08	0.83	53	6.33	0.85	61	6.08	0.83	62
7.08	0.89	7.23	0.89	52	6.93	0.90	52	7.23	0.92	63	6.93	0.90	56
7.30	0.70	7.45	0.72	80	7.15	0.71	67	7.45	0.73	78	7.15	0.71	73
8.23	0.99	8.38	1.00	50	8.08	1.00	59	8.40	1.03	68	8.05	0.98	45
8.75	0.61	8.88	0.61	70	8.60	0.61	54	8.95	0.62	64	8.55	0.62	64
10.90	0.58	11.05	0.59	115	10.75	0.59	96	11.18	0.60	109	10.63	0.60	106
11.23	1.39	11.43	1.38	40	11.15	1.40	41	11.58	1.43	43	11.00	1.37	44
11.75	0.58	11.90	0.59	83	11.60	0.58	66	12.05	0.59	82	11.45	0.59	71
12.55	0.78	12.70	0.79	59	12.43	0.78	39	12.88	0.80	55	12.25	0.79	57
13.20	1.04	13.35	1.07	74	13.08	1.00	50	13.53	1.04	57	12.88	1.05	69
13.28	0.73	13.43	0.76	41	13.13	0.72	33	13.60	0.74	40	12.95	0.75	40
17.35	1.43							17.68	1.42	210	16.98	1.43	215

5.3.5 Alignment

As in Chapter 4, the matched, corresponding reference and sample control points are used to calculate the affine transformation matrix of translation, scaling and rotation parameters, which is then applied to the sample peaks. The results are summarised in Table 5.4. The average Euclidean distance between the coordinates of the reference and sample control points was calculated both before and after pattern matching and alignment. In every case the pattern matching and subsequent affine transform considerably reduced the differences between the reference and sample control points.

Table 5.4: Average Euclidean distance between the reference and sample control points before (uncorrected) and after (corrected) pattern matching and affine transformation

Sample	Before	After
1	6.03	1.02
2	5.80	0.83
3	9.45	1.11
4	9.06	0.87

Once the reference and sample chromatograms are aligned, they can be returned to the original, un-rotated, coordinate system by simply subtracting the offset values.

5.3.6 Chromatogram comparison

After alignment, the final comparison of the reference and sample chromatograms is achieved through the use of fuzzy logic. Fuzzy logic is a convenient way of mapping an input space to an output space by introducing vagueness to eliminate sharp boundaries dividing members of a class from non-members. Rather than having "hard" yes-no answers, fuzzy reasoning allows for "not-quite" yes or no answers to class membership. This type of reasoning is common in human language, but can be difficult for a computer. Reasoning in fuzzy logic is simply a matter of generalising the familiar yes-no (Boolean) logic by assigning "true" the numerical value 1 and "false" 0, however this is also extended to allow in-between values of, for example, 0.6 etc [344-346]. This feature of fuzzy logic is exploited in the comparison of

GC×GC chromatograms to compare a reference peak to the entire sample peak list in order to identify its matching peak. Instead of "yes this is the matching sample peak" or "no it isn't", the peaks are assigned a degree of matching which makes the algorithm flexible enough to tolerate the imprecise alignment resulting from the affine transform. The degree of matching is given by the membership function which is a curve that defines how each point in the input space is mapped to a membership value between 0 and 1 [345]. Membership functions include triangular, trapezoidal and Gaussian functions; these are illustrated below in Figure 5.9. In this work a symmetrical trapezoidal membership function was employed.

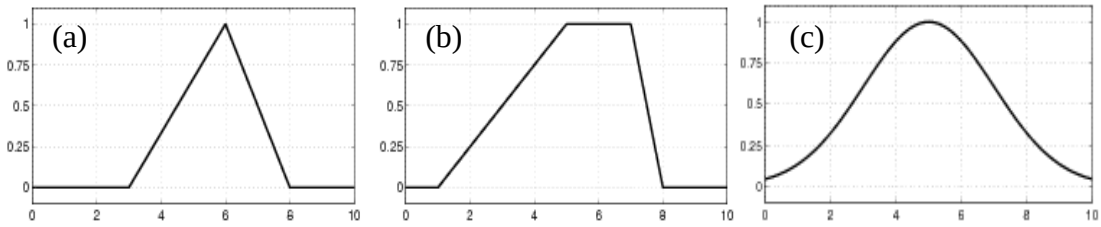


Figure 5.9: Membership functions (a) triangular, (b) trapezoidal and (c) Gaussian. From [345].

The comparison begins by first converting the 2D reference and sample peaks lists into 1D peak positions according to Equation 5.3.

$$X_{1D} = (x - 1) \cdot cycle + y \qquad \text{Equation 5.3}$$

Where x and y are the first dimension and second dimension peak locations, respectively and *cycle* is defined in Equation 4.1.

In this example, the reference is the standard model perfume, as used previously, and the sample is that with the reduced temperature ramp (sample 3). The 1D peak positions for the reference and sample are shown in Figure 5.10.

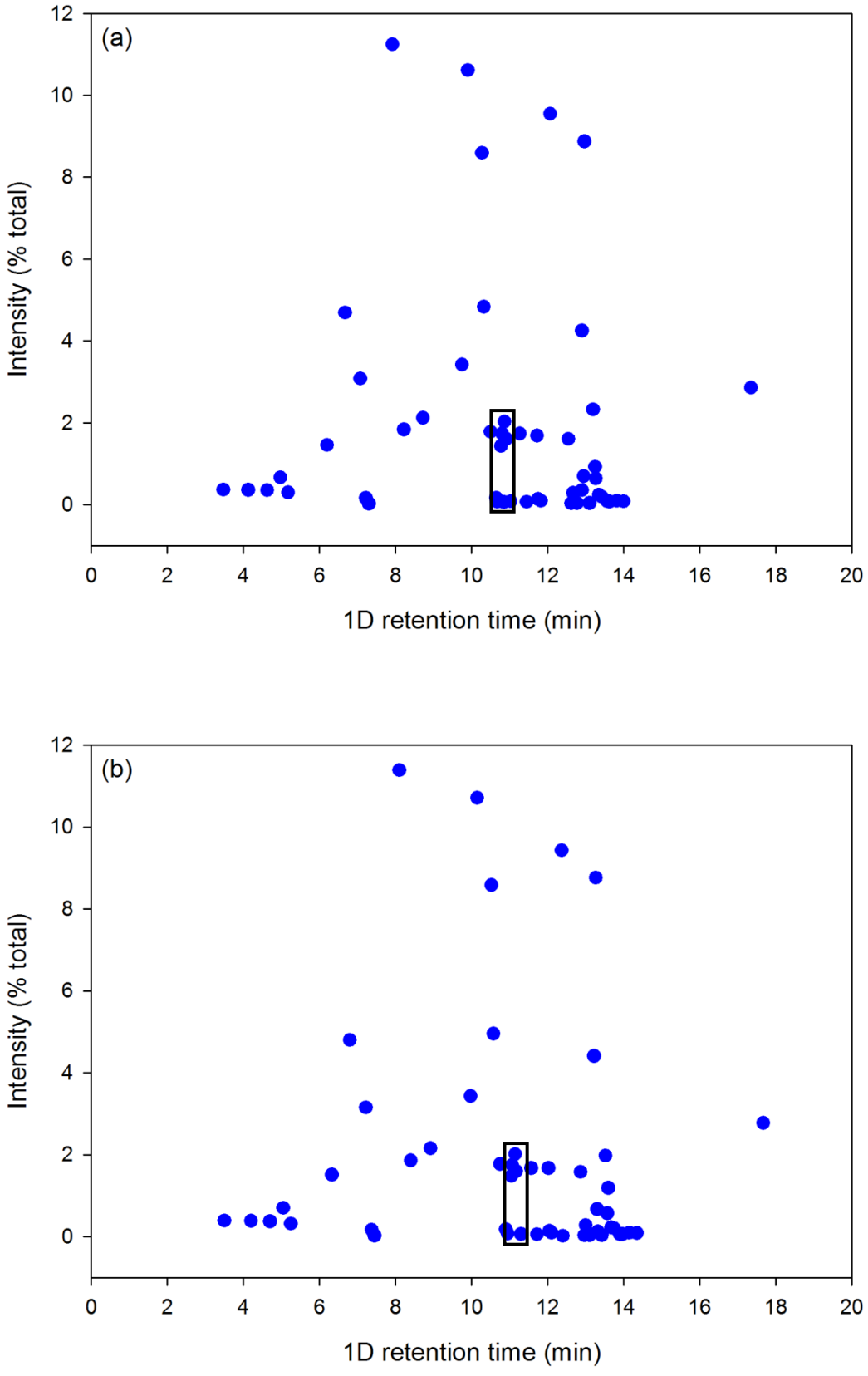


Figure 5.10: 1D peak locations (a) reference and (b) sample

The highlighted section in Figure 5.10 is expanded in Figure 5.11 and as a simple example the highlighted reference peak is compared to the sample peaks (peaks 1-6) in order to find its matching peak. It is simple to identify the matching peaks by eye, however the problem is getting a computer to automatically identify matching peaks and this is done using the fuzzy membership function.

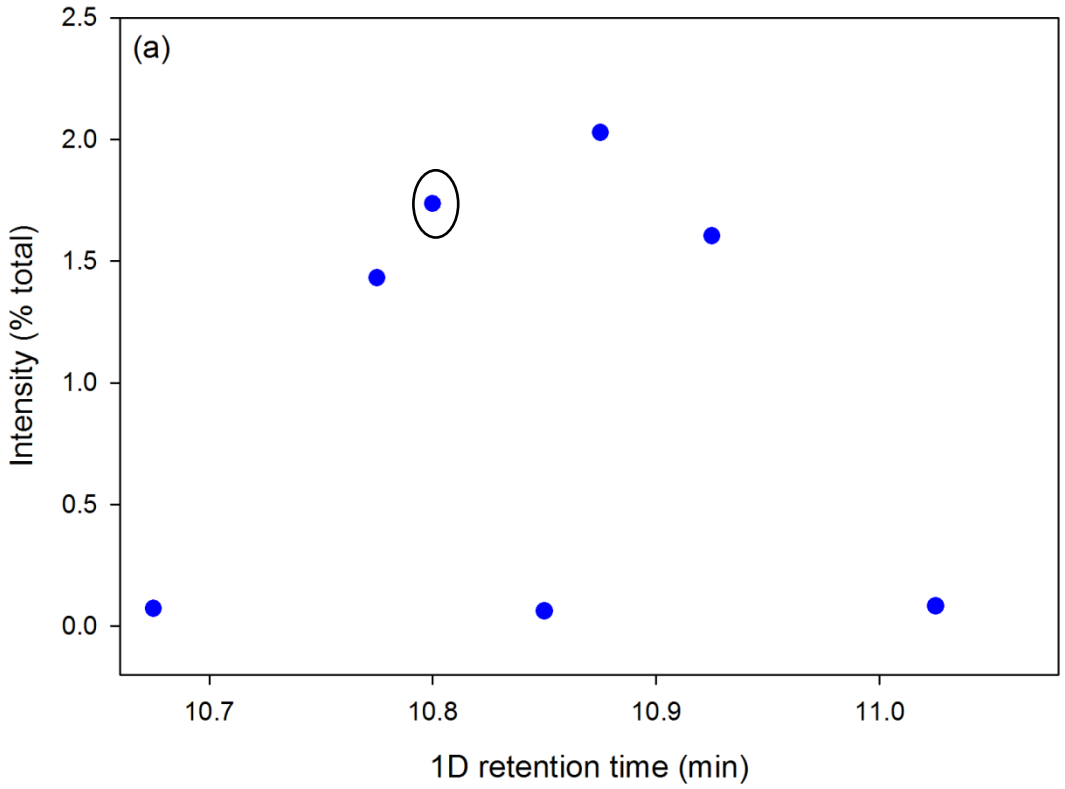


Figure 5.11: Expanded section of Figure 5.10 (a) reference with selected peak for matching highlighted

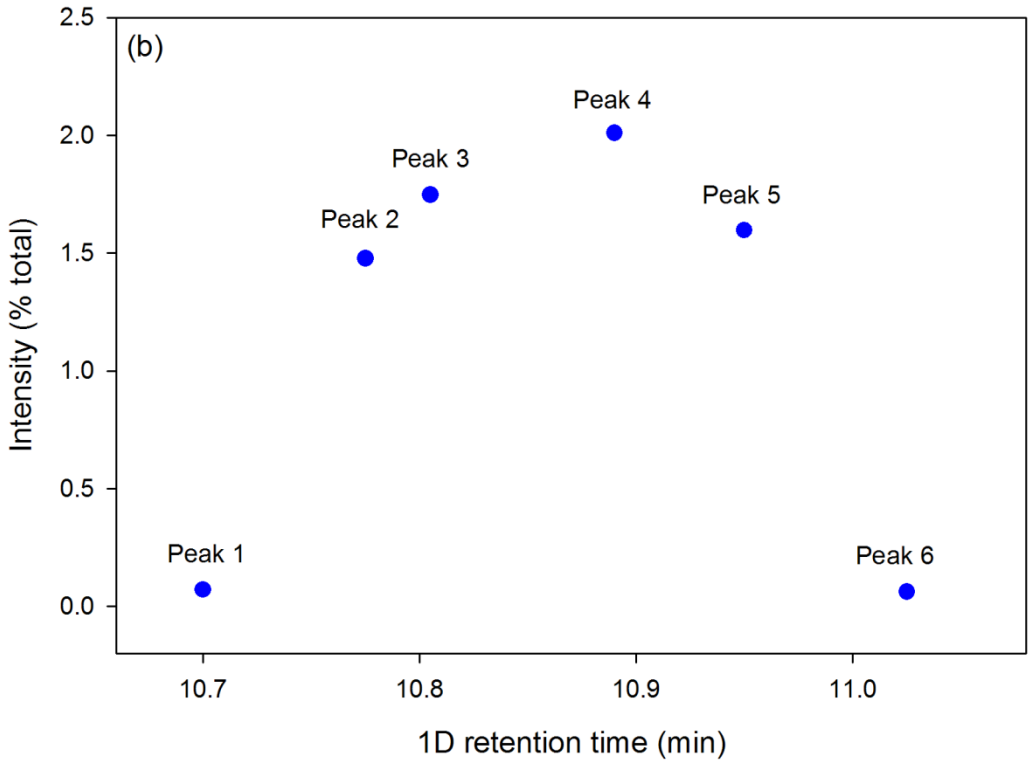


Figure 5.11: Expanded section of Figure 5.10 (b) sample with peaks 1-6 highlighted

Membership values between 0 and 1 for the sample peaks to each of the reference peaks are then determined from the membership function. The closer the membership value is to 1, the more likely it is that the peaks are a match. A 3D representation of the trapezoidal membership function is shown in Figure 5.12 (a); this function essentially fits over a reference peak and the membership values of the sample peaks are determined by their closeness to the reference peak. However, as the algorithm operates in 1D space, the actual membership values are determined as in Figure 5.12 (b).

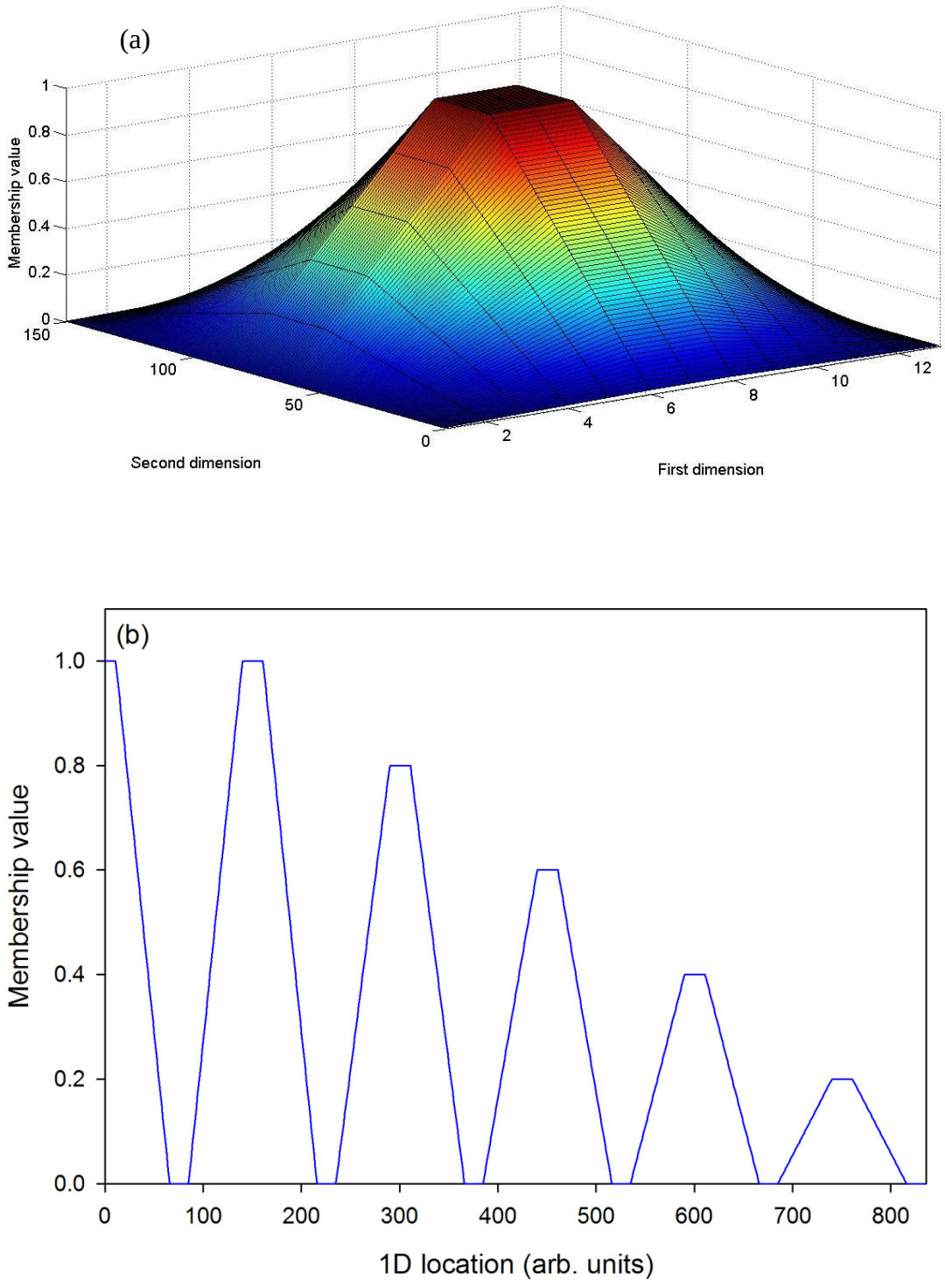


Figure 5.12: Trapezoidal membership function in (a) 3D and (b) 1D forms

The membership values for the sample peaks (peaks 1-6) to the highlighted reference peak are provided in Table 5.5. The membership values indicate that sample peak 3 is the closest match to the reference peak, however in order to make sure that an

incorrect peak is not matched simply as it is the closest, the membership values are scaled according to the peak intensity (Equation 5.4). This ensures that the peaks are relatively the same size. The scaled membership values are provided in Table 5.5. The results confirm that peak 3 is the most likely match to the selected reference peak. It can also be seen that the membership values of some peaks drop significantly after scaling. For example, peak 4 originally had a membership value of 0.58, but after scaling the value dropped to just 0.02. This indicates that the intensity of the peak was significantly different from the reference and as a result is not the correct match.

Table 5.5: Original membership values and scaled membership values for sample peaks (peaks 1-6) to the selected reference peak

Peak number	**Membership value**	**Scaled membership values**
1	0.20	0.00
2	0.52	0.43
3	1.00	0.99
4	0.58	0.02
5	0.60	0.52
6	0.23	0.21

$$M_{scale} = M \cdot \frac{\min(Z_i)}{\max(Z_i)} \qquad \text{Equation 5.4}$$

Where M is the original membership value and Z is the peak intensity.

After the membership values are scaled, the algorithm finds the maximum membership value of every sample peak to each reference peak. This results in a list of matching reference and sample peaks that can be compared.

5.4 Results and discussion

The software developed in this chapter and the previous chapter is employed to analyse a series of real flavour samples. The data contains two batches of each flavour, one that passed QC and the other that failed. The failed samples were failed in the QC laboratory by either 1D GC or taste testing and the software is employed to see if GC×GC is able to identify why it was failed.

When comparing a new sample to a reference for QC purposes, it is necessary to identify if any new peaks appear in the sample that are not in the reference ("extra peaks") or peaks that may be present in the reference but are missing from the sample ("missing peaks"). Tolerance values also need to be set in order to pass or fail matches, a peak may be the same in both the reference and sample but the concentrations may vary significantly, as a result the product should not be passed. The tolerance values employed in this work were originally developed for 1D GC QC and are provided in Table 5.6.

Table 5.6: Tolerance values employed to pass or fail matches

	Fragrances	**Flavours**
Tolerance (%)	**Peak area (%)**	**Peak area (%)**
2	10-100	20-100
5	3-10	6-20
10	1-3	0.9-6
50	0.2-1	0.15-0.9
100	0.02-0.2	0.03-0.15

The final comparison output to be used in a QC laboratory contains a list of the reference and sample peaks, their corresponding percentage volume and a comment as to whether the sample peak passed, failed, etc. as well as a graphical representation of the results. The comparison requires the user to input the file names (reference and sample), modulation period, sampling frequency and tolerance values. No manual intervention is required. The comparison output for one of the flavour samples is provided in Figure 5.13. The results indicate that there are two extra peaks and a missing peak as well as a number of fails based on component concentrations; this

would suggest a possible contamination in the production. The comparison outputs for the remaining flavour samples are provided in Appendix 1.

The time taken to perform the entire comparison, including peak detection, control point selection, alignment and comparison, varies depending on the number of peaks and increases with the number of peaks. For the 10 flavour samples, the average computation time was 23.2 seconds (range 2.0 to 120.5 seconds). This computation was performed on an Asus laptop equipped with an Intel Core i7 processor running at 1.60 GHz and 4.00 GB of RAM. The operating system was Microsoft Windows 7 (64 bit).

(a)

Date: 13-Dec-2011 18:08:12

Reference: FLAVOUR 001 PASS
Sample: FALVOUR 001 FAIL
Modulation period (s): 1.5
Sampling frequency (Hz): 100

Tolerance 1: 20.00% - 100.00% = 2.0%
Tolerance 2: 6.00% - 20.00% = 5.0%
Tolerance 3: 0.90% - 6.00% = 10.0%
Tolerance 4: 0.15% - 0.90% = 50.0%
Tolerance 5: 0.03% - 0.15% = 100.0%

Ref 1	Ref 2	Vol%	Smpl 1	Smpl 2	Vol%	Diff	Comment
102	137	4.61	102	137	4.35	-5.61	Pass
141	12	1.28	141	12	1.22	-4.53	Pass
151	148	1.25	151	147	0.97	-22.00	Fail
165	19	19.47	165	19	18.90	-2.94	Pass
167	135	1.49	167	135	1.38	-7.27	Pass
189	137	1.52	189	137	1.54	1.50	Pass
192	140	0.06	192	140	0.08	43.92	Pass
199	149	10.65	199	149	8.45	-20.70	Fail
202	21	1.32	202	21	1.28	-2.91	Pass
216	21	6.60	216	22	6.46	-2.06	Pass
224	44	2.09	224	44	2.05	-1.74	Pass
251	21	0.66	252	24	2.44	268.23	Fail
276	7	0.46	277	10	1.52	230.69	Fail
280	5	0.13	281	8	0.56	334.66	Fail
327	141	26.12	327	140	24.15	-7.55	Fail
376	10	0.09	376	10	0.05	-43.06	Pass
399	126	0.08	399	127	0.25	199.48	Fail
433	127	0.17	433	128	0.46	177.73	Fail
466	1	21.20	465	150	20.71	-2.33	Fail
515	149	0.69	517	3	2.97	329.47	Fail
0	0	0.00	234	20	0.07	100.00	Extra Peak
0	0	0.00	451	133	0.03	100.00	Extra Peak

(b)

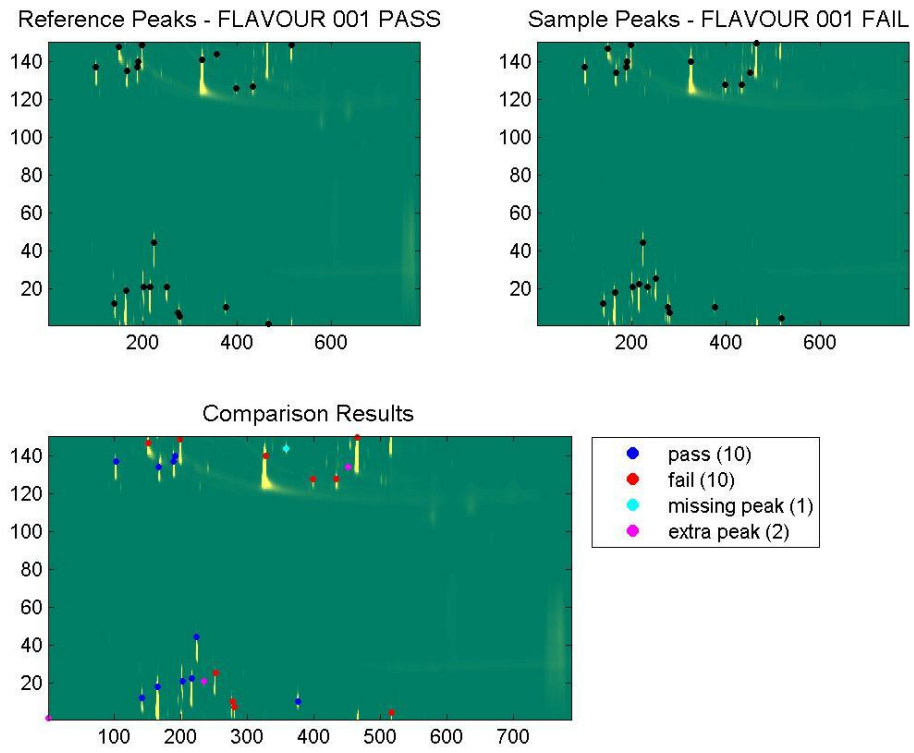


Figure 5.13: FLAVOUR_001comparison output (a) text output and (b) graphic output

5.5 Conclusion

The software developed in Chapter 4 was further extended in this chapter to improve the selection of reference control points by rotating the reference chromatogram to occupy the centre of the pattern space. This resulted in an increased number of control points, which aids alignment via affine transformation. In order to account for peak wrap-around, the sample chromatogram was aligned with the reference in the centre of the pattern space. By aligning the reference and sample in the second dimension, the selected reference control points were not wrapped around in the sample chromatogram and as a result could be accurately matched. A tolerance on the coordinates was also added to help reduce false matches.

A method of comparing reference and sample peaks was developed using fuzzy logic with a trapezoidal membership function. The flexibility of fuzzy comparisons allows peaks to be assigned a degree of matching rather than a “hard” 0 or 1; this was able to compensate for the imprecise alignment of the global affine transformation. The membership values were also scaled according to the peak intensity to aid correct matching as it ensured that the peaks were relatively the same size.

The final version of the software was then successfully employed for analysing real flavour samples for QC purposes.

Chapter 6 - Conclusions and Further Work

Data pre-processing and chemometric techniques proved to be essential for the analysis of chromatographic data.

PCA was employed for both exploratory analysis and to evaluate the effectiveness of alignment on HPLC data. Two alignment methods, COW and *i*coshift were compared using PCA; while both methods were successful in aligning the data, *i*coshift was slightly better as it explained more variance in fewer components. Alignment using *i*coshift was also orders of magnitude faster than COW. However, *i*coshift required multiple combinations of segment lengths and slack sizes in order to remove artifacts introduced by the algorithm. As an exploratory technique, PCA was applied to profile metabolites in *L. angustifolius* inoculated with the pathogen, *P. cinnamomi*, and treated with both water and phosphite. Since the pathogen stimulated a similar response using both treatments, the response was proposed a component of the plants defence against the pathogen. This demonstrated that the pathogen was able to make an association with the plant despite the presence of phosphite in the root tissue.

HPLC with acidic potassium permanganate chemiluminescence detection was evaluated for the analysis of Australian wines from different origins and vintages. PCA was again employed for exploratory analysis as well as a pre-processing technique to reduce the dimensionality of the data. PCs were used as inputs to LDA and QDA in order to discriminate red and white wines from the Coonawarra and Geelong wine growing regions. For the red wines, LDA and QDA had the same overall accuracy; however the number of Coonawarra and Geelong wines correctly classified by each technique was different. This was due to the difference between the linear and parabolic discrimination boundaries employed by LDA and QDA, respectively. For the white wines, QDA was found to be slightly more accurate than LDA. In the analysis of wine vintage, PLS and PCR were compared for the modelling of sample composition with wine age. PCR required more components than PLS to achieve similar predictive ability. This was due to the PCR model being constructed to explain the independent variable rather than the dependent variable and as a result more components were needed to effectively explain the dependent variable.

Software was developed for quality control of flavours and fragrances using GC×GC. The software automatically aligns and compares a sample chromatogram to a reference chromatogram. To ensure a sufficient number of reference control points were selected, the reference chromatogram was rotated to occupy the centre of the pattern space and a simple method of partitioning the 2D pattern space was employed to select the reference control points. In order to prevent peak wrap-around, the sample chromatogram was aligned with the reference in the centre of the pattern space. The selected reference control points were then compared to the entire sample peak list using a triangle pattern matching algorithm to identify the corresponding control points in the sample chromatogram. Once reference and sample control points were identified, they were used to calculate the translation, scaling and rotation operations for an affine transform. The affine transformation was then applied to the complete sample peak list in order to align reference and sample peaks. Comparison of reference and sample chromatograms was achieved through the use of fuzzy logic with a trapezoidal membership function. Finally, the developed software was successfully employed to analyse a number of real flavour samples for QC purposes.

Further work

The software could be further extended to replace any "hard" decisions with fuzzy decisions. This would make the software more flexible and robust. For example, the tolerance values could be made fuzzy to prevent peaks not being matched simply because their volumes were just outside the tolerance cut-off points.

As the software was developed in collaboration with a flavour and fragrance manufacturer, it was only tested on flavour and fragrance data. Hence, the software could be applied to different sample types such as biological samples in order to identify abnormalities in blood or urine. This would considerably extend the ease at which GC×GC could be employed for routine monitoring and analysis.

The software could also be modified for the analysis of LC×LC data. Some preliminary results indicate that the peak detection algorithm may need to be altered to account for the difference in LC×LC peak shapes compared with GC×GC.

Finally, since the software is still in Matlab code, it could be complied to be a standalone program or turned into a graphic user interface.

References

1. S. Ahuja, in *Separation Science and Technology, Vol. 4* (Ed.: S. Ahuja), Academic Press, California, **2003**.
2. J. M. Miller, *Chromatography Concepts and Contrasts*, 2nd ed., John Wiley & Sons Inc., New Jersey, **2005**.
3. H. M. McNair, J. M. Miller, *Basic Gas Chromatography*, 2nd ed., John Wiley & Sons Inc., New Jersey, **2009**.
4. D. C. Harris, *Quantitative Chemical Analysis*, 6th ed., W. H. Freeman and Company, New York, **2003**.
5. C. A. Cramers, P. A. Leclercq, *Journal of Chromatography A*, **1999**, *842*, 3-13.
6. C. F. Poole, S. K. Poole, Journal of Chromatography A, **2008**, *1184*, 254-280.
7. T. Cserhati, *Multivariate Methods in Chromatography: A Practical Guide*, John Wiley & Sons Inc., Chichester, **2008**.
8. R. Bailey, *Journal of Environmental Monitoring*, **2005**, *7*, 1054-1058.
9. M. Minones Vazquez, M. E. Vazquez Blanco, S. Muniztegui Lorenzo, P. Lopez Mahia, E. Fernandez Fernandez, D. Prada Rodriguez, *Journal of Chromatography A*, **2001**, *919*, 363-371.
10. W. Engewald, J. Teske, J. Efer, *Journal of Chromatography A*, **1999**, *856*, 259-278.
11. D. G Wesmoreland, G. R. Rhodes, *Pure & Appl. Chem*., **1989**, 61, 1147-1160.
12. P. J. Marriott, R. Shellie, C. Cornwell, *Journal of Chromatography A,* **2001**, *936,* 1-22.
13. G. R. van der Hoff, P. van Zoonen, *Journal of Chromatography A*, **1999**, *843*, 301-32.
14. Z. Wang, M. Fingas, *Journal of Chromatography A*, **1997**, *774*, 51-78.
15. J. Blomberg, P. J. Schoenmakers, U. A. Th. Brinkman, *Journal of Chromatography A*, **2002**, *972*, 137-173.
16. T. Seppanen-Laakso, I. Laakso, R. Hiltunen, *Analytica Chimica Acta*, **2002**, *465*, 39-62.
17. B. G. Wolthers, G. P. B. Kraan, *Journal of Chromatography A*, **1999**, *843*, 247-274.

18. K. K. Pasikanti, P. C. Ho, E. C. Y. Chan, *Journal of Chromatography B*, **2008**, *871*, 202-211.
19. G. Vidya Sagar, Instrumental Methods of Drug Analysis, PharmaMed Press, 2009.
20. G. Guiochon, *Journal of Chromatography A*, **2007**, *1168*, 101-168.
21. K. K. Unger, R. Skudas, M. M. Schulte, *Journal of Chromatography A*, **2008**, *1184*, 393-415.
22. A. Ghanem, T. Ikegami, *Journal of Separation Science*, **2011**, *34*, 1945-1957.
23. K. Cabrera, *Journal of Separation Science*, **2004**, *27*, 843-852.
24. F. Svec, *Journal of Separation Science*, **2004**, *27*, 1419-1430.
25. M. H. Chen, C. Horvath, *Journal of Chromatography A*, **1997**, *788*, 51-61.
26. V. R. Meyer, *Practical High-Performance Liquid Chromatography*, 5th ed., John Wiley & Sons Inc., Chichester, **2010**.
27. A. Weston, P. R. Brown, *HPLC and CE Principles and Practice*, Academic Press, California, **1997**.
28. M. Swartz, *Journal of Liquid Chromatography & Related Technologies*, **2010**, *33*, 1130-1150.
29. D. Corradini, *Handbook of HPLC*, 2nd ed., CRC Press, Boca Raton, **2011**.
30. E. S. Yeung, R. E. Synovec, *Analytical Chemistry*, **1986**, 58,1237-1265.
31. N. W. Barnett, P. S. Francis, in *Encyclopaedia of Analytical Science, Vol. 5*, Elsevier, Oxford, **2005**.
32. A. V. Kostarnoi, G. B. Golubitskii, E. M. Basova, E. V. Budko, V. M. Ivanov, *Journal of Analytical Chemistry*, **2008**, *63*, 516-529.
33. D. J. Anderson, *Analytical Chemistry*, **1999**, *71*, 314R-327R.
34. C. H. Lochmuller, C. Jiang, Q. Liu, V. Antonucci, M. Elomaa, *Critical Reviews in Analytical Chemistry*, **1996**, *26*, 29-59.
35. H. M. Merken, G. R. Beecher, *Journal of Agricultural and Food Chemistry*, **2000**, *48*, 577-599.
36. K. M. Kalili, A. de Villiers, *Journal of Separation Science*, **2011**, *34*, 854-876.
37. S. J. Lehotay, J. Hajslova, *Trends in Analyitcal Chemistry*, **2002**, *21*, 686-697.
38. M. Kivilompolo, V. Oburka, T. Hyotylainen, *Analytical and Bioanalytical Chemistry*, **2007**, *388*, 881-887.
39. J. M. Amigo, T. Skov, R. Bro, *Chemical Reviews*, **2010**, *110*, 4582-4605.

40. M. Daszykowski, B. Walczak, *Trends in Analytical Chemistry*, **2006**, *25*, 1081-1096.
41. P. Schoenmakers, P. Marriott, J. Beens, *LC-GC Europe*, **2003**, *June*, 335-339.
42. J. C. Giddings, in *Multidimensional Chromatography: Techniques and Aplications*, (Ed.: H. J. Cortes), Marcel Dekker, New York, **1990**.
43. R. A. Shellie, P. R. Haddad, *Analytical and Bioanalytical Chemistry*, **2006**, *386*, 405-415.
44. J. Dalluge, J. Beens, U. A. Th. Brinkman, *Journal of Chromatography A*, **2003**, *1000*, 69-108.
45. M. Adahchour, J. Beens, R. J. J. Vreuls, U. A. Th. Brinkman, *Trends in Analytical Chemistry*, **2006**, *25*, 438-454.
46. M. Adahchour, J. Beens, R. J. J. Vreuls, U. A. Th. Brinkman, *Trends in Analytical Chemistry*, **2006**, *25*, 540-553.
47. M. Adahchour, J. Beens, U. A. Th. Brinkman, *Journal of Chromatography A*, **2008**, *1186*, 67-108.
48. H. Cortes, B. Winniford, J. Luong, M. Pursch, *Journal of Separation Science*, **2009**, *32*, 883-904.
49. P. Marriott, R. Shellie, *Trends in Analytical Chemistry*, **2002**, *21*, 573-583.
50. R. C. Y. Ong, P. J. Marriott, *Journal of Chromatographic Science*, **2002**, *40*, 276-291.
51. T. Gorecki, O. Panic, N. Oldridge, *Journal of Liquid Chromatography & Related Technologies*, **2006**, *29*, 1077-1104.
52. J. B. Phillips, J. Xu, *Journal of Chromatography A*, **1995**, *703*, 327-334.
53. J. B. Phillips, J. Beens, *Journal of Chromatography A*, **1999**, *856*, 331-347.
54. J. B. Phillips, E. B. Ledford, *Field Analytical Chemistry and Technology*, **1996**, *1*, 23-29.
55. R. M. Kinghorn, P. J. Marriott, *Journal of High Resolution Chromatography*, **1998**, *21*, 620-622.
56. P. J. Marriott, R. M. Kinghorn, *Analytical Sciences*, **1998**, *14*, 651-659.
57. E. B. Ledford, C. Billesbach, *Journal of High Resolution Chromatography*, **2000**, *23*, 202-204.
58. J. Beens, M. Adahchour, R. J. J. Vreuls, K. van Altene, U. A. Th. Brinkman, *Journal of Chromatography A*, **2001**, *919*, 127-132.

59. M. Pursch, P. Eckerle, J. Biel, R. Streck, H. Cortes, K. Sun, B. Winniford, *Journal of Chromatography A*, **2003**, *1019*, 43-51.
60. J. Harynuk, T. Gorecki, *Journal of Chromatography A*, **2003**, *1019*, 53-63.
61. C. A. Bruckner, B. J. Prazen, R. E. Synovec, *Analytical Chemistry*, **1998**, *70*, 2796-2804.
62. J. V. Seeley, F. Kramp, C. J. Hicks, *Analytical Chemistry*, **2000**, *72*, 4346-4352.
63. L. Ramos, in *Comprehensive Analytical Chemistry, Vol. 55* (Ed.: D. Barcelo), Elsevier B.V., Amsterdam, **2009**.
64. C. von Muhlen, W. Khummueng, C. Alcaraz Zini, E. Bastos Caramao, P. J. Marriott, *Journal of Separation Science*, **2009**, *29*, 1909-1921.
65. I. Francois, K. Sandra, P. Sandra, *Analytica Chimica Acta*, **2009**, *641*, 14-31.
66. D. R. Stoll, X. Li, X. Wang, P. W. Carr, S. E. G. Porter, S. C. Rutan, *Journal of Chromatography A*, **2007**, *1168*, 3-43.
67. M. Kivilompolo, J. Pol, T. Hyotylainen, *LC-GC Europe*, **2011**, *May*, 232-243.
68. J. Pol, T. Hyotylainen, *Analytical and Bioanalytical Chemistry*, **2008**, *391*, 21-31.
69. P. Schoenmakers, *LC-GC North America*, **2008**, *26*, 600-608.
70. P. Jandera, *LC-GC Europe*, **2007**, *October*, 510-525.
71. J. D. Dimendja, S. B. Stanfill, J. Grainger, D. G. Patterson, *Journal of High Resolution Chromatography*, **2000**, *23*, 208-214.
72. M. Adahchour, J. Beens, R. J. J. Vreuls, A. M. Batenburg, E. A. Rosing, U. A. Th. Brinkman, *Chromatographia*, **2002**, *55*, 361-367.
73. R. Shellie, P. Marriott, C. Cornwell, *Journal of High Resolution Chromatography*, **2000**, *23*, 554-560.
74. C. von Muhlen, C. Alcaraz Zini, E. Bastos Caramao, P. J. Marriott, *Journal of Chromatography A*, **2006**, *1105*, 39-50.
75. B. Mitrevski, P. Wynne, P. J. Marriott, *Analytical and Bioanalytical Chemistry*, **2011**, *401*, 2361-2371.
76. R. Shellie, P. Marriott, *Flavour and Fragrance Journal*, **2003**, *18*, 179-191.
77. J. E. Welke, C. A. Zini, *Journal of the Brazilian Chemical Society*, **2011**, *22*, 609-622.
78. O. Panic, T. Gorecki, *Analytical and Bioanalytical Chemistry*, **2006**, *386*, 1013-1023.

79. M. F. Almstetter, P. J. Oefner, K. Dettmer, *Analytical and Bioanalytical Chemistry*, **2012**, *402*, 1993-2013.
80. P. Dugo, F. Cacciola, P. Donato, D. Airando-Rodriguez, M. Herrero, L. Mondello, *Journal of Chromatography A*, **2009**, *1216*, 7483-7487.
81. A. Greiderer, L. Steeneken, T. Aalbers, G. Vivo-Truyols, P. Schoenmakers, *Journal of Chromatography A*, **2011**, *1218*, 5787-5793.
82. M. Eggink, W. Romero, R. J. Vreuls, H. Lingeman, W. M. A. Niessen, H. Irth, *Journal of Chromatography A*, **2008**, *1188*, 216-226.
83. M. Daszykowski, B. Walczak, *Journal of Chromatography A*, **2007**, *1176*, 1-11.
84. J. J. Harynuk, A. P. de la Mata, N. A. Sinkov, in *Chemometrics in Practical Applications*, (Ed.: K. Varmuza), InTech, **2012**.
85. R. A. van den Berg, H. C. J. Hoefsloot, J. A. Westerhuis, A. K. Smilde, M. J. van der Werf, *BMC Genomics*, **2006**, *7*, 1-15.
86. N. J. Nielsen, G. Tomasi, R. J. N. Frandsen, M. B. Kristensen, J. Nielsen, H. Giese, J. H. Christensen, *Metabolomics*, **2010**, *6*, 341-352.
87. K. Laursen, S. Sondergaard Frederiksen, C. Leuenhagen, R. Bro, *Journal of Chromatography A*, **2010**, *1217*, 6503-6510.
88. J. H. Christensen, G. Tomasi, *Journal of Chromatography A*, **2007**, *1169*, 1-22.
89. J. H. Christensen, G. Tomasi, A. B. Hansen, *Environmental Science & Technology*, **2005**, *39*, 255-260.
90. M. M. W. B. Hendriks, L. Cruz-Juarez, D. De Bont, R. D. Hall, *Analytica Chimica Acta*, **2005**, *545*, 53-64.
91. F. Chau, Y. Liang, J. Gao, X. Shao, in *Chemical Analysis; A Series of Monographs on Analytical Chemistry and its Applications*, *Vol. 164* (Ed.: J. D. Winefordner), John Wiley & Sons Inc., New Jersey, **2004**.
92. P. H. C. Eilers, *Analytical Chemistry*, **2004**, *76*, 404-411.
93. T. Skov, Doctoral thesis, University of Copenhagen, 2008.
94. K. Kaczmarek, B. Walczak, S. de Jong, B. G. M. Vandeginste, *Acta Chromatographica*, **2005**, *15*, 82-96.
95. M. Daszykowski, I. Stanimirova, A. Bodzon-Kulakowska, J. Silberring, G. Lubec, B. Walczak, *Journal of Chromatography A*, **2007**, *1158*, 306-317.
96. F. Gan, G. Ruan, J. Mo, *Chemometrics and Intelligent Laboratory Systems*, **2006**, *82*, 59-65.

97. A. Savitzky, M. J. E. Golay, *Analytical Chemistry*, **1964**, *36*, 1627-1639.
98. A. M. van Nederkassel, D. Daszykowski, P. H. C. Eilers, Y. Vander Heyden, *Journal of Chromatography A*, **2006**, *1118*, 199-210.
99. A. M. van Nederkassel, C. J. Xu, P. Lancelin, M. Sarraf, D. A. MacKenzie, N. J. Walton, F. Bensaid, M. Lees, G.J. Martin, J. R. Desmurs, D. L. Massart, J. Smeyers-verbeke, Y. Vander Heyden, *Journal of Chromatography A*, **2006**, *1120*, 291-298.
100. T. Skov, F. van den Berg, G. Tomasi, R. Bro, *Journal of Chemometrics*, **2006**, *20*, 484-497.
101. G. Tomasi, F. van den Berg, C. Andersson, *Journal of Chemometrics*, **2004**, *18*, 231-241.
102. V. Pravdova, B. Walczak, D. L. Massart, *Analytica Chimica Acta*, **2002**, *456*, 77-92.
103. K. M. Pierce, J. L. Hope, K. J. Johnson, B. W. Wright, R. E. Synovec, *Journal of Chromatography A*, **2005**, *1096*, 101-110.
104. K. J. Johnson, B. W. Wright, K. H. Jarman, R. E. Synovec, *Journal of Chromatography A*, **2003**, *996*, 141-155.
105. G. Malmquist, R. Danielsson, *Journal of Chromatography A*, **1994**, *687*, 71-88.
106. J. H. Christensen, A. B. Hansen, U. Karlson, J. Mortensen, O. Andersen, *Journal of Chromatography A*, **2005**, *1090*, 133-145.
107. D. Bylund, R. Danielsson, G. Malmquist, K. E. Markides, *Journal of Chromatography A*, **2002**, *961*, 237-244.
108. E. Szymanska, M. J. Markuszewski, X. Capron, A. M. van Nederkassel, Y. Vander Heyden, M. Markuszewski, K. Krajka, R. Kaliszan, *Journal of Pharmaceutical and Biomedical Analysis*, **2007**, *43*, 413-420.
109. A. M. Hupp, L. J. Marshall, D. I. Campbell, R. Waddell Smith, V. L. McGuffin, *Analytica Chimica Acta*, **2008**, *606*, 159-171.
110. K. M. Pierce, B. W. Wright, R. E. Synovec, *Journal of Chromatography A*, **2007**, *1141*, 106-116.
111. A. M. van Nederkassel, M. Dazykowski, D. L. Massart, Y. Vander Heyden, *Journal of Chromatography A*, **2005**, *1096*, 177-186.
112. L. M. Malmquist, R. R. Olsen, A. B. Hansen, O. Andersen, J. H. Christensen, *Journal of Chromatography A*, **2007**, *1164*, 262-270.

113. L. Zheng, D. G. Watson, B. F. Johnston, R. L. Clark, R. Edrada-Ebel, W. Elseheri, *Analytica Chimica Acta*, **2009**, *642*, 257-265.
114. M. Daszykowski, M. Sajewicz, J. Rzepa, M. Hajnos, D. Staszek, L. Wojtal, T. Kowalska, M. Waksmundzka-Hajnos, B. Walczak, *Acta Chromatographica*, **2009**, *21*, 513-530.
115. K. Ropkins, D. C. Carlsaw, P. S. Goodman, J. E. Tate, *Trends in Analyitcal Chemistry*, **2009**, *28*, 373-391.
116. N. P. V. Nielsen, J. M. Carstensen, J. Smedsgaard, *Journal of Chromatography A*, **1998**, *805*, 17-35.
117. D. Zhang, X. Huang, F. E. Regnier, M. Zhang, *Analytical Chemistry*, **2008**, *80*, 2664-2671.
118. E. Szymanska, M. J. Markuszewski, X. Capron, A. M. van Nederkassel, Y. Vander Heyden, M. Markuszewski, K. Krajka, R. Kaliszan, *Electrophoresis*, **2007**, *28*, 2861-2873.
119. N. E. Watson, M. W. VanWingerden, K. M. Pierce, B. W. Wright, R. E. Synovec, *Journal of Chromatography A*, **2006**, *1129*, 111-118.
120. K. M. Pierce, L. F. Wood, B. W. Wright, R. E. Synovec, *Analytical Chemistry*, **2005**, *77*, 7735-7743.
121. F. Savorani, G. Tomasi, S. B. Engelsen, *Journal of Magnetic Resonance*, **2010**, *202*, 190-202.
122. P. de la Mata-Espinosa, J.M. Bosque-Sendra, R. Bro, L. Cuadros-Rodriguez, *Analytical and Bioanalytical Chemistry*, **2011**, *399*, 2083-2092.
123. P. de la Mata-Espinosa, J.M. Bosque-Sendra, R. Bro, L. Cuadros-Rodriguez, *Talanta*, **2011**, *85*, 177-182.
124. K. Laursen, U. Justesen, M. A. Rasmussen, *Journal of Chromatography A*, **2011**, *1218*, 4340-4348.
125. G. F. Giskeodegard, T. G. Bloemberg, G. Postma, B. Sitter, M. B. Tessem, I. S. Gribbestad, T. F. Bathen, L. M. C. Buydens, *Analytica Chimica Acta*, **2010**, *683*, 1-11.
126. A. Craig, O. Cloarec, E. Holmes, J. K. Nicholson, J. C. Lindon, *Analytical Chemistry*, **2006**, *78*, 2262-2267.
127. M. J. Adams, *Chemometrics in Analytical Spectroscopy*, 2nd ed., (Ed.: N. W. Barnett), The Royal Society of Chemistry, Cambridge, **2004**.
128. P. H. C. Eilers, *Analytical Chemistry*, **2003**, *75*, 3631-3636.

129. G. Vivo-Truyols, P. J. Schoenmakers, *Analytical Chemistry*, **2006**, *78*, 4598-4608.
130. E. T. Whittaker, *Proceedings of the Edinburgh Mathematical Society*, **1923**, *41*, 63.
131. A. Detroyer, V. Schoonjans, F. Questiers, Y. Vander Heyden, A. P. Borosy, Q. Guo, D. L. Massart, *Journal of Chromatography A*, **2000**, *897*, 23-36.
132. R. J. M. Vervoort, A. J. J. Debets, H. A. Claessens, C. A. Cramers, G. J. de Jong, *Journal of Chromatography A*, **2000**, *897*, 1-22.
133. S. Rocha, L. Maeztu, A. Barros, C. Cid, M. A. Coimbra, *Journal of the Science of Food and Agriculture*, **2003**, *84*, 43-51.
134. J. F. Cotte, H. Casabianca, S. Chardon, J. Lheritier, M. F. Grenier-Loustalot, *Analytical and Bioanalytical Chemistry*, **2004**, *380*, 698-705.
135. M. R. Euerby, P. Petersson, *Journal of Chromatography A*, **2005**, *1088*, 1-15.
136. J. S. Camara, M. Arminda Alves, J. C. Marques, *Talanta*, **2006**, *68*, 1512-1521.
137. S. Risticevic, E. Carasek, J. Pawliszyn, *Analytica Chimica Acta*, **2008**, *617*, 72-84.
138. J. Chen, Y. Lu, D. Wei, X. Zhou, *Chromatographia*, **2009**, *70*, 981-985.
139. F. A. Jabalpurwala, J. M. Smoot, R. L. Rouseff, *Phytochemistry*, **2009**, *70*, 1428-1434.
140. G. Xiang, H. Yang, L, Yang, X. Zhang, Q. Cao, M. Miao, *Microchemical Journal*, **2010**, *95*, 198-206.
141. R.G. Brereton, Chemometrics*; Data Analysis for the Laboratory and Chemical Plant*, John Wiley & Sons Inc., Chichester, **2003**.
142. M. Otto, *Chemometrics; Statistics and Computer Application in Analytical Chemistry*, 2nd ed., Wiley-VCH, Germany, 2007.
143. K. H. Liland, *Trends in Analyitcal Chemistry*, **2011**, *30*, 827-841.
144. L. Vaclavik, O. Lacina, J. Hajslova, J. Zweigenbaum, *Analytica Chimica Acta*, **2011**, *685*, 45-51.
145. A. Kende, D. Portwood, A. Senior, M. Earll, E. Bolygo, M. Seymour, *Journal of Chromatography A*, **2010**, *1217*, 6718-6723.
146. N. Ratola, J. M. Amigo, M. S. N. Oliveira, R. Araujo, J. A. Silva, A. Alves, *Environmental and Experimental Botany*, **2011**, *72*, 339-347.
147. S. J. de Andrade, J. Cristale, F. S. Silva, G. J. Zocolo, M. R. R. Marchi, *Atmospheric Environment*, **2010**, *44*, 2913-2919.

148. A. Fragkaki, Y. S. Angelis, A. Tsantili-Kakoulidou, M. Koupparis, C. Georgakopoulos, *Journal of Steriod Biochemistry & Molecular Biology*, **2009**, *115*, 44-61.
149. M. Monfreda, A. Gregori, *Journal of Forensic Sciences*, **2011**, *56*, 372-380.
150. L. J. Marshall, J. W. McIlroy, V. L. McGuffin, R. Waddel Smith, *Analytical and Bioanalytical Chemistry*, **2009**, *394*, 2049-2059.
151. A. W. Michell, D. Mosedale, D. J. Grainger, R. A. Barker, *Metabolomics*, **2008**, *4*, 191-201.
152. A. Rencher, Multivariate Methods of Analyis, 2nd ed., John Wiley & Sons Inc., Canada, **2002**.
153. D. Ballabio, T. Skov, R. Leardi, R. Bro, *Journal of Chemometrics*, **2008**, *22*, 457-463.
154. L. Liu, D. Cozzolino, W. U. Cynkar, R. G. Dambergs, L. Janik, B. K. O'Niell, C. B. Colby, M. Gishen, *Food Chemistry*, **2008**, *106*, 781-786.
155. C. J. Bevin, R. G. Dambergs, A. J. Fergusson, D. Cozzolino, *Analytica Chimica Acta*, **2008**, *621*, 19-23.
156. T. Cserhati, *Biomedical Chromatography*, **2010**, *24*, 20-28.
157. N. Bratchell, *Chemometrics and Intelligent Laboratory Systems*, **1989**, *6*, 105-125.
158. R. Liu, Y. Jia, W. Cheng, J. Ling, L. Liu, K. Bi, Q. Li, *Talanta*, **2011**, *83*, 751-756.
159. K. Kim, P. Aronvo, S. O. Zakherkin, D. Anderson, B. Perround, I. M. Thompson, R. H. Weiss, *Molecular & Cellular Proteomics 8.3*, **2009**, 558-570.
160. L. Liu, Q. Li, N. Li, J. Ling, R. Liu, Y. Wang, L. Sun, X. Hui Chen, K. Bi, *Journal of Separation Science*, **2011**, *34*, 1198-1204.
161. P. Jandera, K. Vynuchalova, T. Hajek, P. Cesla, G. Vohralik, *Journal of Chemometrics*, **2008**, *22*, 203-217.
162. M. H. Abraham, C. F. Poole, S. K. Poole, *Journal of Chromatography A*, **1999**, *842*, 79-114.
163. L. A. Berrueta, R. M. Alonso-Salces, K. Heberger, *Journal of Chromatography A*, **2007**, *1158*, 196-214.
164. B. M. Vandeginste, *Handbook of Chemometrics and Qualimetrics: Part B, Vol. 20B*, Elsevier Science B.V., Amsterdam, **1998**.

165. R. M. Alonso-Salces, C. Herrero, A. Barranco, D. M. Lopez-Marquez, L. A. Berrueta, B. Gallo, F. Vicente, *Food Chemistry*, **2006**, *97*, 438-446.
166. B. K. Lavine, *Critical Reviews in Analytical Chemistry*, **2006**, *36*, 153-161.
167. J. Broseus, M. Vallat, P. Esseiva, *Chemometrics and Intelligent Laboratory Systems*, **2011**, *107*, 343-350.
168. Y. Ni, R. Song, S. Kokot, *Analytical Methods*, **2012**, *4*, 171-176.
169. S. J. Dixon, R. G. Brereton, *Chemometrics and Intelligent Laboratory Systems*, **2009**, *95*, 1-17.
170. H. H. Jelen, A. Ziolkowska, A. Kaczmarek, *Journal of Agricultural and Food Chemistry*, **2010**, *58*, 12585-12591.
171. M. J. Lerma-Garcia, E. F. Simo-Alfonso, A. Mendez, L. J. Lliberia, J. M. Herrero-Martinez, *Food Research International*, **2011**, *44*, 103-108.
172. E. Pouliarekou, A. Badeka, M. Tasioula-Margari, S. Kontakos, F. Longobardi, M. G. Kontominas, *Journal of Chromatography A*, **2011**, *1218*, 7354-7542.
173. A. B. Cerezo, W. Tesfaye, M. E. Soria-Diaz, M. Jesus Torija, E. Mateo, M. C. Garcia-Parrilla, A. M. Troncoso, *Journal of Food Composition and Analysis*, **2010**, *23*, 175-184.
174. J. S. Camara, P. Herbert, J. C. Marques, M. A. Alves, *Analytica Chimica Acta*, **2004**, *513*, 203-207.
175. C. Pizarro, I. Esteban-Diez, C. Saenz-Gonzalez, J. M. Gonzalez-Saiz, *Analytica Chimica Acta*, **2008**, *608*, 38-47.
176. A. Bentabol Manzanares, Z. Hernandez Garcia, B. Rodriguez Galdon, E. Rodriguez Rodriguez , C. Diaz, Romero, *Food Chemistry*, **2011**, *126*, 664-672.
177. K. L. Tritt, C. J. O'Bara, M. J. M. Wells, *Journal of Agricultural and Food Chemistry*, **2005**, *53*, 5304-5312.
178. B. M. Wise, *PLS_Toolbox Version 4.0; User Manual*, Eigenvector Research, **2006**.
179. F. Priego Capote, J. Ruiz Jimenez, M. D. Luque de Castro, *Analytical and Bioanalytical Chemistry*, **2007**, *388*, 1859-1865.
180. R. R Hatanaka, D. L. Flumignan, J. E. de Oliveira, *Chromatographia*, **2009**, *70*, 1135-1142.
181. J. R. Lucio-Gutierrez, J. Coello, S. Maspoch, *Analytica Chimica Acta*, **2012**, *710*, 40-49.

182. S. P. Boyle, P. J. Doolan, C. E. Andrews, R. G. Reid, *Journal of Pharmaceutical and Biomedical Analysis*, **2011**, *54*, 951-957.
183. R. G. Brereton, *Analyst*, **2000**, *125*, 2125-2154.
184. M. Anzar, R. Lopez, J. Cacho, V. Ferreira, *Journal of Agricultural and Food Chemistry*, **2003**, *51*, 2700-2707.
185. Y. Niu, X. Zhang, Z. Xiao, S. Song, K. Eric, C. Jia, H. Yu, J. Zhu, *Journal of Chromatography B*, **2011**, *879*, 2287-2293.
186. M. Vilanova, Z. Genisheva, A. Masa, J. M. Oliveira, *Microchemical Journal*, **2010**, *95*, 240-246.
187. M. Andjelkovic, J. Van Camp, M. Pedra, K. Renders, C. Socaciu, R. Verhe, *Journal of Agricultural and Food Chemistry*, **2008**, *56*, 5181-5185.
188. K. M. Pierce, S. P. Schale, *Talanta*, **2011**, *83*, 1254-1259.
189. V. A. Watts, C. E. Butzke, R. B. Boulton, *Journal of Agricultural and Food Chemistry*, **2003**, *51*, 7738-7742.
190. N. Rodriguez, M. C. Ortiz, L. Sarabia, E. Gredilla, *Talanta*, **2010**, *81*, 255-264.
191. M. Gonzalez Alvarez, C. Gonzalez-Barreiro, B. Canch-Grande, J. Simal-Gamdara, *Food Chemistry*, **2011**, *129*, 890-898.
192. R. G. Brereton, *Applied Chemometrics for Scientists*, John Wiley & Sons Inc., Chichester, **2007**.
193. S. Wold, M. Sjostrom, L. Eriksson, *Chemometrics and Intelligent Laboratory Systems*, **2001**, *58*, 109-130.
194. S. Wold, N. Kettaneh-Wold, B. Skagerberg, *Chemometrics and Intelligent Laboratory Systems*, **1989**, *7*, 53-65.
195. S. Wold, *Technometrics*, **1993**, *35*, 136-139.
196. F. Lindgren, P. Geladi, S. Rannar, S. Wold, *Journal of Chemometrics*, **1994**, *8*, 349-363.
197. F. Lindgren, P. Geladi, A. Berglund, M. Sjostrom , S. Wold, *Journal of Chemometrics*, **1995**, *9*, 331-342.
198. J. Trygg, S. Wold, *Chemometrics and Intelligent Laboratory Systems*, **1998**, *42*, 209-220.
199. S. Wold, J. Trygg, A. Berglund , H. Antti, *Chemometrics and Intelligent Laboratory Systems*, **2001**, *58*, 131-150.
200. S. Wold, *Chemometrics and Intelligent Laboratory Systems*, **2001**, *58*, 83-84.

201. S. Wold, M. Josefson, J. Gottfries, A. Linusson, *Journal of Chemometrics*, **2004**, 1*8*, 156-165.
202. S. Wold, M. Hoy, H. Martens, J. Trygg, F. Westad, J. MacGregor, B. M. Wise, *Journal of Chemometrics*, **2009**, 23, 67-68.
203. R. D. Wentzell, L. Vega Montoto, *Chemometrics and Intelligent Laboratory Systems*, **2003**, *65*, 257-279.
204. G. T. Ventura, G. J. Hall, R. K. Nelson, G. S. Frysinger, B. Raghuraman, A. E. Pomerantz, O. C. Mullins, C. M. Reddy, *Journal of Chromatography A*, **2011**, *1218*, 2584-2592.
205. E. S. Bodle, J. K. Hardy, *Analytica Chimica Acta*, **2007**, *589*, 247-254.
206. V. L. Skrobot, E. V. R. Castro, R. C. C. Pereira, V. M. D. Pasa, I. C. P. Fortes, *Energy & Fuels*, **2005**, *19*, 2350-2356.
207. K. J. Johnson, B. J. Prazen, D. C. Young, R. E. Synovec, *Journal of Separation Science*, **2004**, *27*, 410-416.
208. P. M. L. Sandercock, E. Du Pasquier, *Forensic Science International*, **2003**, *134*, 1-10.
209. J. S. Ribeiro, F. Augusto, T. J. G. Salva, R. A. Thomaziello, M. M. C. Ferreira, *Analytica Chimica Acta*, **2009**, *634*, 172-179.
210. A. Bansleben, I. Schellenberg, J. W. Einax, K. Schaefer, D. Ulrich, D. Bansleben, *Analytical and Bioanalytical Chemistry*, **2009**, *395*, 1503-1512.
211. N. H. Beltran, M. A. Duarte-Mermoud, M. A. Bustos, S. A. Salah, E. A. Loyola, A. I. Pena-Neira, J. W. Jalocha, *Journal of Food Engineering*, **2006**, *75*, 1-10.
212. D. Brodnjak-Voncina, Z. Cencic Kodba, M. Novic, *Chemometrics and Intelligent Laboratory Systems*, **2005**, *75*, 31-43.
213. A. G. Gonzalez, F. Pablos, M. J. Martin, M. Leon-Camacho, M. S. Valdenebro, *Food Chemistry*, **2001**, *73*, 93-101.
214. H. Tsugawa, Y. Tsujimoto, M. Arita, T. Bamba, E. Fukusaki, *BMC Genomics*, **2011**, *12*, 1-13.
215. R. Danielsson, E. Allard, P. J. R. Sjoberg, J. Berquist, *Chemometrics and Intelligent Laboratory Systems*, **2011**, *108*, 33-48.
216. G. B. Mortuza, W. A. Neville, J. Delaney, C. J. Waterfield, P. Camilleri, *Biochimica et Biophysica Acta*, **2003**, *1631*, 136-146.
217. Z. Slejkovec, Z. Bajc, D. Z. Doganoc, *Talanta*, **2004**, *62*, 931-936.

218. A. Garrido Frenich, F. J. Arrebola Liebanas, M. Mateu-Sanchez, J. L. Martinez Vidal, *Talanta*, **2003**, *60*, 765-774.
219. C. Xu, Y. Liang, F. Chau, Y. Vander Heyden, *Journal of Chromatography A*, **2006**, *1134*, 253-259.
220. F. Gong, Y. Liang, Y. Fung, F. Chau *Journal of Chromatography A*, **2004**, *1029*, 173-183.
221. M. Praisler, J. Van Bocxlaer, A. De Leenheer, D. L. Massart, *Journal of Chromatography A*, **2002**, *962*, 161-173.
222. M. Praisler, I. Dirinck, J. Van Bocxlaer, A. De Leenheer, D. L. Massart, *Talanta*, **2000**, *53*, 177-193.
223. A. Kher, M. Mulholland, E. Green, B. Reedy, *Vibrational Spectroscopy*, **2006**, *40*, 270-277.
224. K. Le Mapinhan, J. Vial, A. Jardy, *Journal of Chromatography A*, **2004**, *1030*, 135-147.
225. E. Van Gyseghem, S. Van Hemelryck, M. Daszykowski, F. Questier, D. L. Massart, Y. Vander Heyden, *Journal of Chromatography A*, **2003**, *988*, 77-93.
226. D. Visky, Y. Vander Heyden , T. Ivanyi, P. Baten, J. De Beer, Z. Kovacs, B. Noszal, P. Dehouck, E. Roets, D. L. Massart , J. Hoogmartens, *Journal of Chromatography A*, **2003**, *1012*, 11-29.
227. J. Liang, G. Ma, H. Fang, L. Chen, P. Christie, *Environmental Earth Sciences*, **2011**, *62*, 33-42.
228. J. M. Baerncopf, V. L. McGuffin, R. W. Smith, *Journal of Forensic Sciences*, **2011**, *56*, 70-81.
229. G. Tomasi, F. Savorani, S. B. Engelsen, *Journal of Chromatography A*, **2011**, *1218*, 7832-7840.
230. D. M. Cahill, G. M. Weste, B. R. Grant, *Plant Physiology*, **1986**, *81*, 1103-1109.
231. M. J. Aberton, B. A. Wilson, D. M. Cahill, *Australasian Plant Pathology*, **1999**, *28*, 225-234.
232. R. Daniel, B. A. Wilson, D. M. Cahill, *Australasian Plant Pathology*, **2005**, *34*, 541-548.
233. J. W. Allwood, D. I. Ellis, R. Goodacre, *Physiologia Plantarum*, **2008**, *132*, 117-135.
234. T. K. Gunning, Doctoral thesis, Deakin University, 2010.

235. T. K, Gunning, X. A. Conlan, R. M. Parker, G. A. Dyson, M. J. Adams, N. W. Barnett, D. M. Cahill, in preparation for submission to *Analytical Biochemistry*.
236. Quality & Technology; Plant Food Science group & Spectroscopy and Chemometrics group, 07/06/11, <http://www.models.kvl.dk/algorithms>.
237. Mass Bank; High Quality Mass Spectral Database, 20/05/11, <http://www.massbank.jp>.
238. P. Bednarek, L. Kerhoas, J. Einhorn, R. Franski, P. Wojtaszek, M. Rybus-Zajac, M. Stobiecki, *Journal of Chemical Ecology*, **2003**, *29*, 1127-1142.
239. P. Kachlicki, J. Einhorn, D. Muth, L. Kerhoas, M. Stobiecki, *Journal of Mass Spectrometry*, **2008**, *43*, 572-586.
240. P. Kachlicki, L. Marczak, L. Kerhoas, J. Einhorn, M. Stobiecki, *Journal of Mass Spectrometry*, **2005**, *40*, 1088-51103.
241. D. Muth, P. Kachlicki, P. Krajewski, M. Przystalski, M. Stobiecki, *Metabolomics*, **2009**, *5*, 354-362.
242. D. Muth, E. Marsden-Edwards, P. Kachlicki, M. Stobiecki, *Phytochemical Analysis*, **2008**, *19*, 444-452.
243. M. Hsieh, T. L. Graham, *Phytochemistry*, **2001**, *58*, 995-1005.
244. S. Tahara, R. K. Ibrahim, *Phytochemistry*, **1995**, *38*, 1073-1094.
245. S. Tahara, Y. Katagiri, J. L. Ingham, J. Mizutani, *Phytochemistry*, **1994**, *36*, 1261-1271.
246. S. G. Sparg, M. E. Light. J. van Staden, *Journal of Ethnopharmacology*, **2004**, *94*, 219-243.
247. T. J. Jackson, T. Burgess, I. Colquhoun, G. E. StJ. Hardy, *Plant Pathology*, **2000**, *49*, 147-154.
248. S. Tahara, *Journal of the Agricultural Chemical Society of Japan*, **1984**, *58*, 1247-1257.
249. I. S. Arvanitoyannis, M. N. Katsota, E. P. Psarra, E. H. Soufleros, S. Kallithraka, *Trends in Food Science & Technology*, **1999**, *10*, 321-336.
250. S. Kelly, K. Heaton, J. Hoogewerf, *Trends in Food Science & Technology*, **2005**, *16*, 555-567.
251. L. Liu, D. Cozzolino, W. U. Cynkar, M. Gishen, C. B. Colby, *Journal of Agricultural and Food Chemistry*, **2006**, *54*, 6754-6759.

252. L. M. Reid, C. P. O'Donnell, G. Downey, *Trends in Food Science & Technology*, **2006**, *17*, 344-353.
253. C. Cordella, I. Moussa, A. Martel, N. Sbirrazzuoli, L. Lizzani-Cuvelier, *Journal of Agricultural and Food Chemistry*, **2002**, *50*, 1751-1764.
254. U. Fischer, D. Roth, M. Christmann, *Food Quality and Preference*, **1999**, *10*, 281-288.
255. M. Garcia-Marino, J. M. Hernandez-Hierro, C. Santos-Buelga, J. C. Rivas-Gonzalo, M. T. Escribano-Bailon, *Talanta*, **2011**, *85*, 2060-2066.
256. L. Jaitz, K. Siegl, R. Eder, G. Rak, L. Abranko, G. Koellensperger, S. Hann, *Food Chemistry*, **2010**, *122*, 366-372.
257. S. Ledda, G. Sanna, G. Manca, M. A. Franco, A. Porcu, *Journal of Food Composition and Analysis*, **2010**, *23*, 580-585.
258. V. Rastija, G. Srecnik, M. M. Saric, *Food Chemistry*, **2009**, *115*, 54-60.
259. M. S. Dopico-Garcia, A. Fique, L. Guerra, J. M. Afonso, O. Pereira, P. Valentao, P. B. Andrade, R. M. Seabra, *Talanta*, **2008**, *75*, 1190-1202.
260. A. de Villiers, P. Majek, F. Lynen, A. Crouch, H. Lauer, P. Sandra, *European Food Research Technology*, **2005**, *221*, 520-528.
261. L. Gambelli, G. P. Santaroni, *Journal of Food Composition and Analysis*, **2004**, *17*, 613-618.
262. N. Landrault, P. Poucheret, P. Ravel, F. Gasc, G. Cros, P. Teissedre, *Journal of Agricultural and Food Chemistry*, **2001**, *49*, 3341-348.
263. M. del Mar Castineira Gomez, I. Feldmann, N. Jakubowski, J. T. Andersson, *Journal of Agricultural and Food Chemistry*, **2004**, *52*, 2962-2974.
264. S. Frias, J. E. Conde, J. J. Rodriguez-Bencomo, F. Garcia-Montelongo, J. P. Perez-Trujillo, *Talanta*, **2003**, *59*, 335-344.
265. M. J. Baxter, H. M. Crews, M. J. Dennis, I. Goodall, D. Anderson, *Food Chemistry*, **1997**, *60*, 443-450.
266. S. Cabredo-Pinillos, T. Cedron-Fernandez, C. Saenz-Barrio, *European Food Research Technology*, **2008**, *226*, 1317-1323.
267. E. Falque, P. Darriet, E. Fernandez, D. Dubourdieu, *International Journal of Food Science and Technology*, **2008**, *43*, 4640475.
268. B. Fedrizzi, F. Magno, D. Badocco, G. Nicolini, G. Versini, *Journal of Agricultural and Food Chemistry*, **2007**, *55*, 10880-10887.

269. I. Alvarez, J. L. Aleixandre, M. J. Garcia, A. Casp, L. Zunica, *European Food Research Technology*, **2003**, *217*, 173-179.
270. J. L. Aleixandre, V. Lizama, I. Alvarez, M. J. Garcia, *Journal of Agricultural and Food Chemistry*, **2002**, *50*, 751-755.
271. E. H. Soufleros, E. Bouloumpasi, C. Tsarchopoulos, C. G. Biliaderis, *Food Chemistry*, **2003**, *80*, 261-273.
272. K. Heberger, E. Csomos, L. Simon-Sakadi, *Journal of Agricultural and Food Chemistry*, **2003**, *51*, 8055-8060.
273. M. A. Rodriguez-Delgado, G. Gonzalez-Hernandez, J. E. Conde-Gonzalez, J. P. Perez-Trujillo, *Food Chemistry*, **2002**, *78*, 523-532.
274. A. Rodriguez-Bernaldo de Quiros, M. A. Lage-Yusty, J. Lopez-Hernandez, *Food Research International*, **2009**, *42*, 1018-1022.
275. F. Fang, J. Li, Q. Pan, W. Huang, *Food Chemistry*, **2007**, *101*, 428-433.
276. S. Kallithraka, A. Mamalos, D. Makris, *Journal of Agricultural and Food Chemistry*, **2007**, *55*, 3233-3239.
277. M. A. Rodriguez-Delgado, G. Gonzalez, J. P. Perez-Trujillo, F. J. Garcia-Montelongo, *Food Chemistry*, **2002**, *76*, 371-375.
278. J. W. Costin, N. W. Barnett, S. W. Lewis, D. J. McGillivery, *Analytica Chimica Acta*, **2003**, *499*, 47-56.
279. E. Nalewajko-Sieliwoniuk, I. Tarasewicz, A. Kojlo, *Analytica Chimica Acta*, **2010**, *668*, 19-25.
280. S. Girotti, F. Fini, L. Bolelli, L. Savini, E. Sartini, G. Arfelli, *Luminescence*, **2006**, *21*, 233-238.
281. Q. Zhang, H. Cui, A. Myint, M. Lian, L. Liu, *Journal of Chromatography A*, **2005**, *1095*, 94-101.
282. J. Zhou, H. Ciu, C. Wan, H. Xu, Y. Pang, C. Duan, *Food Chemistry*, **2004**, *88*, 613-620.
283. J. L. Adcock, P. S. Francis, N. W. Barnett, *Analytica Chimica Acta*, **2007**, *601*, 36-67.
284. D. Cozzolino, W. U. Cynkar, N. Shah, P. A. Smith, *Food Chemistry*, **2011**, *126*, 673-678.
285. R. Riovanto, W. U. Cynkar. P. Berzaghi, D. Cozzolino, *Journal of Agricultural and Food Chemistry*, **2011**, *59*, 10356-10360.

286. J. M. Jurado, O. Ballesteros, A. Alcazar, F. Pablos, M. J. Martin, J. L. Vilchez, A. Navalon, *Analytical and Bioanalytical Chemistry*, **2008**, *390*, 961-970.
287. F. Marini, R. Bucci, A. L. Magri, A. D. Magri, *Chemometrics and Intelligent Laboratory Systems*, **2006**, *84*, 164-171.
288. M. Brescia, V. Caldarola, A. De Giglio, D. Benedetti, F. P. Fanizzi, A. Sacco, *Analytica Chimica Acta*, **2002**, *458*, 177-186.
289. I. J. Kosir, M. Kocjancic, N. Ogrinc, J. Kidric, *Analytica Chimica Acta*, **2001**, *429*, 195-206.
290. N. Ogrinc, I. J. Kosir, M. Kocjancic, J. Kidric, *Journal of Agricultural and Food Chemistry*, **2001**, *49*, 1432-1440.
291. M. Urbano, M. D. Luque de Castro, P. M. Perez, J. Garcia-Olmo, M. A. Gomez-Nieto, *Food Chemistry*, **2006**, *97*, 166-175.
292. D. Cozzolino, H. E. Smyth, W. Cynkar, R. G. Dambergs, M. Gishen, *Talanta*, **2005**, *68*, 382-387.
293. D. Cozzolino, H. E. Smyth, M. Gishen, *Journal of Agricultural and Food Chemistry*, **2003**, *51*, 7703-7708.
294. F. Shen, Y. Ying, B. Li, Y. Zheng, Q. Zhuge, *Food Chemistry*, **2011**, *129*, 565-569.
295. D. A. Guillen, M. Palma, R. Natera, R. Romero, C. G. Barroso, *Journal of Agricultural and Food Chemistry*, **2005**, *53*, 2412-2417.
296. M. Giaccio, A. Del Signore, *Journal of the Science of Food and Agriculture*, **2004**, *84*, 164-172.
297. S. A. Bellomarino, Doctoral thesis, Deakin University, 2009.
298. S. A. Bellomarino, X. A. Conlan, R. M. Parker, N. W. Barnett, M. J. Adams, *Talanta*, **2009**, *80*, 833-838.
299. S. A. Bellomarino, R. M. Parker, X. A. Conlan, N. W. Barnett, M. J. Adams, *Analytica Chimica Acta*, **2010**, *678*, 34-38.
300. M. S. McDonald, M. Hughes, J. Burns, M. E. J. Lean, D. Matthews, A. Crozier, *Journal of Agricultural and Food Chemistry*, **1998**, *46*, 368-375.
301. D. M. Goldberg, E. Tsang, A. Karumanchiri, G. J. Soleas, *American Journal of Enology and Viticulture*, **1998**, *49*, 142-151.
302. H. Vuorinen, K. Maatta, R. Torronen, *Journal of Agricultural and Food Chemistry*, **2000**, *48*, 2675-2680.

303. G. Gremaud, S. Quaile, U. Piantini, E. Pfammatter, C. Corvi, *European Food Research Technology*, **2004**, *219*, 97-104.
304. D. M. Goldberg, A. Karumanchiri, E. Tsang, G. J. Soleas, *American Journal of Enology and Viticulture*, **1998**, *49*, 23-34.
305. S. Tsanova-Savova, F. Ribarova, *Journal of Food Composition and Analysis*, **2002**, *15*, 639-645.
306. A. Ferrandino, S. Guidoni, *European Food Research Technology*, **2010**, *230*, 417-427.
307. B. Mevik, H. R. Cederkvist, *Journal of Chemometrics*, **2004**, *18*, 422-429.
308. H. Kelebek, A. Canbas, M. Jourdes, P. Teissedre, *Analytical Letters*, **2011**, *44*, 991-1008.
309. M. Schwarz, G. Hofman, P. Winterhalter, *Journal of Agricultural and Food Chemistry*, **2004**, *52*, 498-504.
310. M. Chamkha, B. Cathala, V. Cheynier, R. Douillard, *Journal of Agricultural and Food Chemistry*, **2003**, *51*, 3179-3184.
311. E. F. Griss, F. Mattivi, E. A. Ferreira, U. Vrhovsek, R. C. Pedrosa, M. T. Bordignon-Luiz, *Food Chemistry*, **2011**, *126*, 213-220.
312. K. Chira, N. Pacella, M. Jourdes, P. Teissedre, *Food Chemistry*, **2011**, *126*, 1971-1977.
313. B. E. Broderick, W. P. Cofino, R. Cornelis, K. Heydorn, W. Horwitz, D. T. E. Hunt, R. C. Hutton, H. M. Kingston, H. Huntau, R. Baudo, D. Rossi, J. G. van Raaphorst, T. T. Lub, P. Schramel, F. T. Smyth, D. E. Wells, A. G. Kelly, *Mikrochimica Acta*, **1991**, *2*, 523-542.
314. R. Shellie, P. Marriott, A. Chaintreau, *Flavour and Fragrance Journal*, **2004**, *19*, 91-98.
315. L. Mondello, A. Casilli, P. Q. Casilli, P. Dugo, R. Costa, S. Festa, G. Dugo, *Journal of Separation Science*, **2004**, *27*, 442-450.
316. J. Jiao, N. Ding, T. Shi, X. Chai, P. Cong, Z. Zhu, *Analytical Letters*, **2011**, *44*, 648-655.
317. M. Scandinaro, P. Q. Tranchida, R. Costa, P. Dugo, G. Dugo, L. Mondello, *LC-GC Europe*, **2010**, *October*, 456-464.
318. S. A. Savchuk, G. M. Kolesov, *Journal of Analytical Chemistry*, **2005**, *60*, 752-771.
319. A. van Asten, *Trends in Analyitcal Chemistry*, **2002**, *21*, 698-708.

320. Z. L. Cardeal, P. P. de Souza, M. D. R. Gomes de Silva, P. J. Marriott, *Talanta*, **2008**, *74*, 793-799.
321. E. M. Humston, J. D. Knowles, A. McShea, R. E. Synovec, *Journal of Chromatography A*, **2010**, *1217*, 1963-1970.
322. J. C. Giddings, *Analytical Chemistry*, **1984**, *56*, 1258-1270.
323. G. Guiochon, L. A. Beaver, M. F. Gonnord, A. M. Siouffi, M. Zakari, *Journal of Chromatography*, **1983**, *255*, 415-437.
324. L. M. Blumberg, *Journal of Chromatography A*, **2003**, *985*, 29-38.
325. S. E. Reichenbach, M. Ni, V. Kottapalli, A. Visvanathan, *Chemometrics and Intelligent Laboratory Systems*, **2004**, *71*, 107-120.
326. M. Kallio, M. Kivilompolo, S. Varjo, M. Jussila, T. Hyotylainen, *Journal of Chromatography A*, **2009**, *1216*, 2923-2927.
327. S. E. Reichenbach, M. Ni, D. Zhang, E. B. Ledford, *Journal of Chromatography A*, **2003**, *985*, 47-56.
328. S. Peters, G. Vivo-Truyols, P. J. Marriott, P. J. Schoenmakers, *Journal of Chromatography A*, **2007**, *1156*, 14-524.
329. J. Beens, H. Boelens, R. Tijssen, J. Blomberg, *Journal of High Resolution Chromatography*, **1998**, *21*, 47-54.
330. R. A. Shellie, W. Welthagen, J. Zrostlikova, J. Spranger, M. Ristow, O. Fiehn, R. Zimmermann, *Journal of Chromatography A*, **2005**, *1086*, 83-90.
331. Th. Groger, M. Schaffer, M. Putz, B. Ahrens, K. Drew, M. Eschnre, R. Zimmermann, *Journal of Chromatography A*, **2008**, *1200*, 8-16.
332. J. Vial, H. Nocairi, P. Sassiat, S. Mallipatu, G. Cognon, D. Thiebaut, B. Teillet, D. N. Rutledge, *Journal of Chromatography A*, **2009**, *1216*, 2866-2872.
333. S. Castillo, I. Mattila, J. Miettinen, M. Oresic, T. Hyotoylainen, *Analytical Chemistry*, **2011**, *83*, 3058-3067.
334. C. G. Fraga, B. J. Prazen, R. E. Synovec, *Analytical Chemistry*, **2001**, *73*, 5833-5840.
335. V. G. van Mispelaar, A. C. Tas, A. K. Smilde, P. J. Schoenmakers, A. C. van Asten, *Journal of Chromatography A*, **2003**, *1019*, 15-29.
336. B. V. Hollingsworth, S. E. Reichenbach, Q. Tao, A. Visvanathan, *Journal of Chromatography A*, **2006**, *1105*, 51-58.
337. E. J. Groth, *The Astronomical Journal*, **1986**, *91*, 1244-1248.
338. E. B. Ledford, *Zoex Corporation*, **2003**.

339. V. Vivo-Truyols, J. R. Torres-Lapasio, A. M. van Nederkassel, Y. Vander Heyden, D. L. Massart, *Journal of Chromatography A*, **2005**, *1096*, 133-145.
340. B. Zitova, J. Flusser, *Image and Visual Computing*, **2003**, *21*, 977-1000.
341. Z. Arzoumanian, J. Holmberg, B. Norman, *Journal of Applied Ecology*, **2005**, *42*, 999-1011.
342. C. Narayanaswami, *IBM Technical Disclosure Bulletin*, **1995**, *38*, 53
343. M. Ni, S. E. Reichenbach, A. Visvanathan, J. TerMaat, E. B. Ledford, *Journal of Chromatography A*, **2005**, *1086*, 165-170.
344. G. J. Klir, T. A. Folger, *Fuzzy Sets, Uncertainty, and Information*, Prentice Hall, New Jersey, **1988**.
345. *Fuzzy Logic Toolbox; Users Guide*, V. 2, The MathWorks Inc., USA, **2000**
346. H. Bandmer, S. Gottwald, *Fuzzy Sets Fuzzy Logic Fuzzy Methods with Applications*, John Wiley & Sons Inc., Chichester, **1995**.

Appendices

Appendix 1: Flavour sample comparison outputs

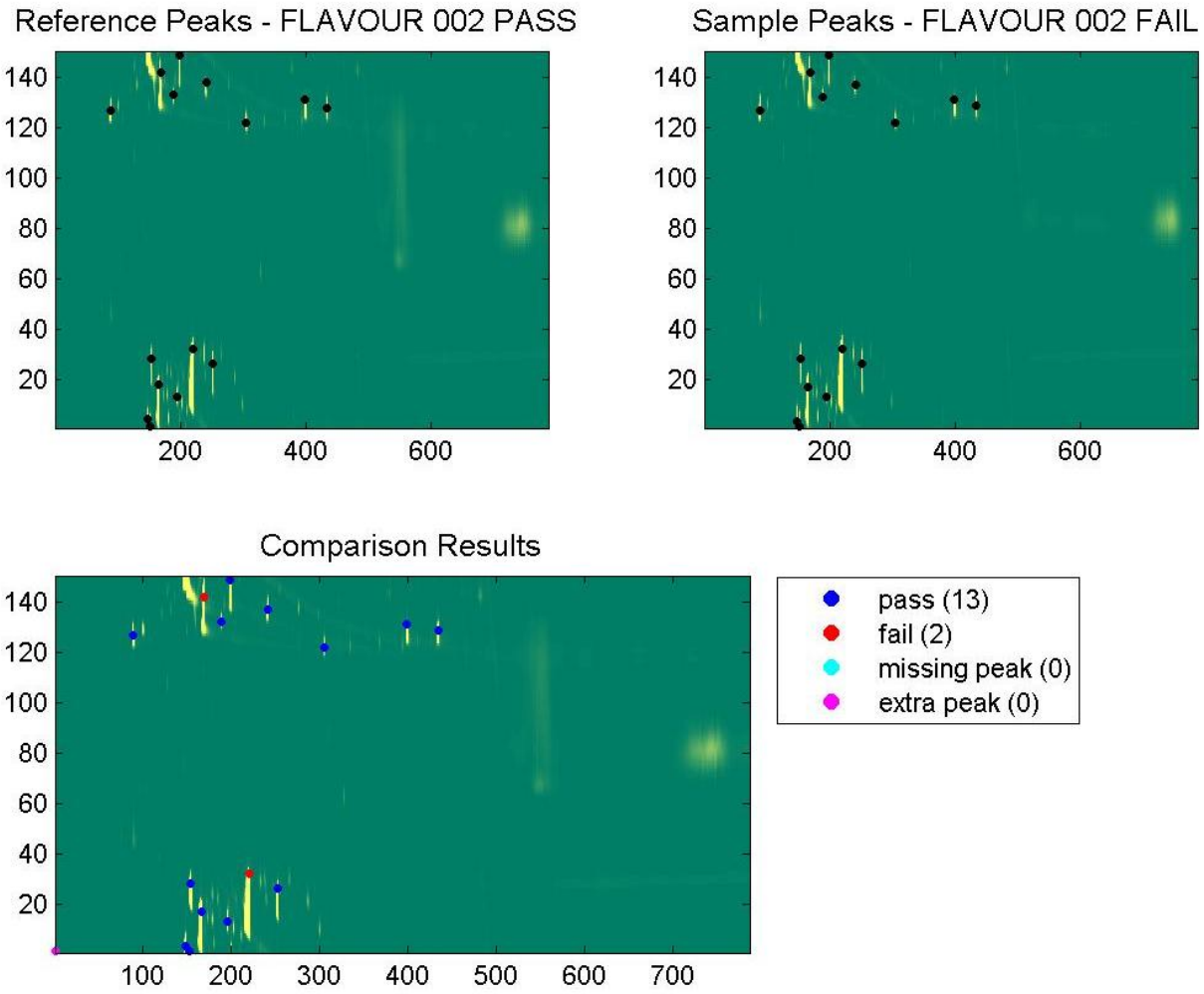
FLAVOUR_002

```
Date: 25-Jan-2012 22:05:09

Reference: FLAVOUR 002 PASS
Sample: FLAVOUR 002 FAIL
Modulation period (s): 1.5
Sampling frequency (Hz): 100

Tolerance 1: 20.00% - 100.00% = 2.0%
Tolerance 2: 6.00% - 20.00% = 5.0%
Tolerance 3: 0.90% - 6.00% = 10.0%
Tolerance 4: 0.15% - 0.90% = 50.0%
Tolerance 5: 0.03% - 0.15% = 100.0%
```

Ref 1	Ref 2	Vol%	Smpl 1	Smpl 2	Vol%	Diff	Comment
89	127	0.39	89	127	0.56	43.71	Pass
148	4	2.00	148	3	1.88	-6.28	Pass
152	1	2.74	152	1	2.87	4.61	Pass
154	28	2.30	154	28	2.23	-2.87	Pass
165	18	18.63	166	17	18.29	-1.79	Pass
169	142	17.91	169	142	16.54	-7.61	Fail
189	133	0.06	189	132	0.04	-42.26	Pass
196	13	0.14	196	13	0.12	-11.26	Pass
199	149	4.41	199	149	4.43	0.32	Pass
220	32	46.33	220	32	48.18	3.99	Fail
241	138	0.21	241	137	0.13	-35.59	Pass
252	26	3.71	252	26	3.89	4.87	Pass
305	122	0.10	305	122	0.07	-35.87	Pass
399	131	0.66	399	131	0.47	-29.19	Pass
434	128	0.27	434	129	0.21	-21.22	Pass


Reference Peaks - FLAVOUR 002 PASS
Sample Peaks - FLAVOUR 002 FAIL
Comparison Results
140
120
100
80
60
40
20
200
400
600
100
300
500
700
pass (13)
fail (2)
missing peak (0)
extra peak (0)

FLAVOUR_003

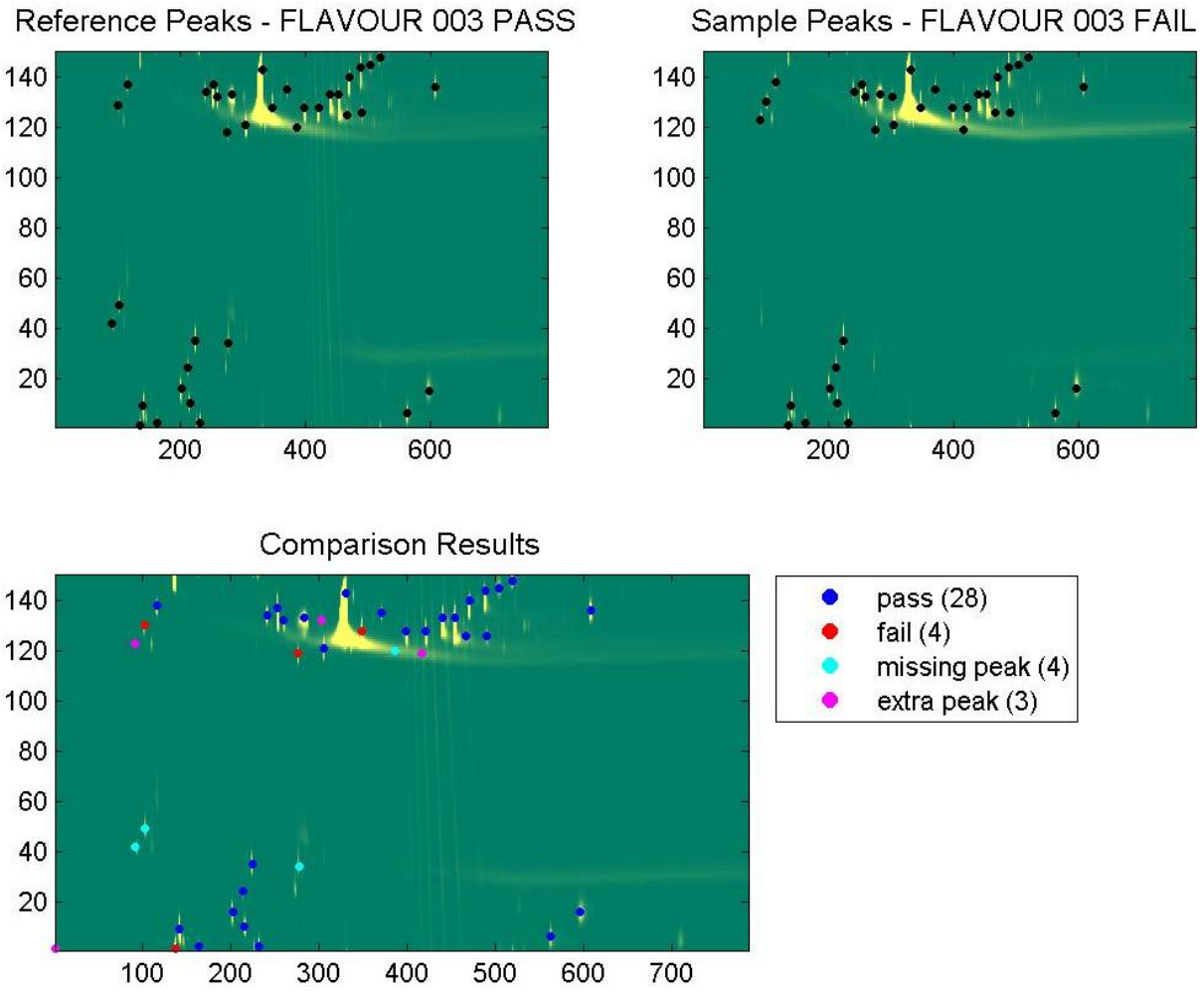
Date: 25-Jan-2012 22:05:33

Reference: FLAVOUR 003 PASS
Sample: FLAVOUR 003 FAIL
Modulation period (s): 1.5
Sampling frequency (Hz): 100

Tolerance 1: 20.00% - 100.00% = 2.0%
Tolerance 2: 6.00% - 20.00% = 5.0%
Tolerance 3: 0.90% - 6.00% = 10.0%
Tolerance 4: 0.15% - 0.90% = 50.0%
Tolerance 5: 0.03% - 0.15% = 100.0%

Ref 1	Ref 2	Vol%	Smpl 1	Smpl 2	Vol%	Diff	Comment
92	42	0.03	0	0	0.00	0.00	Missing Peak
101	129	0.03	102	130	0.10	256.10	Fail
103	49	0.06	0	0	0.00	0.00	Missing Peak
116	137	0.05	116	138	0.07	41.38	Pass
137	1	1.19	137	1	1.03	-13.79	Fail
141	9	0.72	141	9	0.64	-11.33	Pass
163	2	0.04	163	2	0.04	-10.02	Pass
202	16	0.37	202	16	0.33	-8.84	Pass
213	24	0.03	213	24	0.03	-7.26	Pass
216	10	0.07	215	10	0.06	-9.32	Pass
224	35	0.17	225	35	0.15	-8.30	Pass
232	2	0.10	232	2	0.09	-6.44	Pass
241	134	0.11	241	134	0.10	-3.66	Pass
253	137	1.70	253	137	1.61	-5.28	Pass
260	132	0.12	260	132	0.13	8.90	Pass
276	118	0.04	276	119	0.12	175.47	Fail
277	34	0.02	0	0	0.00	0.00	Missing Peak
284	133	0.44	283	133	0.45	2.85	Pass
305	121	0.19	305	121	0.19	0.48	Pass
331	143	85.97	331	143	85.87	-0.13	'Pass'
348	128	2.04	348	128	2.68	31.25	Fail
371	135	0.02	371	135	0.02	-6.41	Pass
386	120	0.03	0	0	0.00	0.00	Missing Peak
399	128	0.45	399	128	0.52	15.28	Pass
421	128	0.40	421	128	0.38	-6.54	Pass
440	133	0.94	440	133	0.89	-5.30	Pass
454	133	2.35	454	133	2.13	-9.38	Pass
467	125	0.04	467	126	0.03	-2.65	Pass
470	140	0.78	470	140	0.73	-6.62	Pass
488	144	0.96	488	144	0.89	-7.34	Pass
490	126	0.02	490	126	0.02	-8.14	Pass

504	145	0.04	504	145	0.02	-47.61	Pass
519	148	0.06	519	148	0.06	-6.61	Pass
562	6	0.10	562	6	0.08	-13.81	Pass
597	15	0.10	596	16	0.16	55.84	Pass
607	136	0.14	608	136	0.13	-6.39	Pass
0	0	0.00	92	123	0.05	100.00	Extra Peak
0	0	0.00	302	132	0.03	100.00	Extra Peak
0	0	0.00	416	119	0.04	100.00	Extra Peak


Reference Peaks - FLAVOUR 003 PASS
Sample Peaks - FLAVOUR 003 FAIL
Comparison Results
140
120
100
80
60
40
20
200
400
600
100
300
500
700
pass (28)
fail (4)
missing peak (4)
extra peak (3)

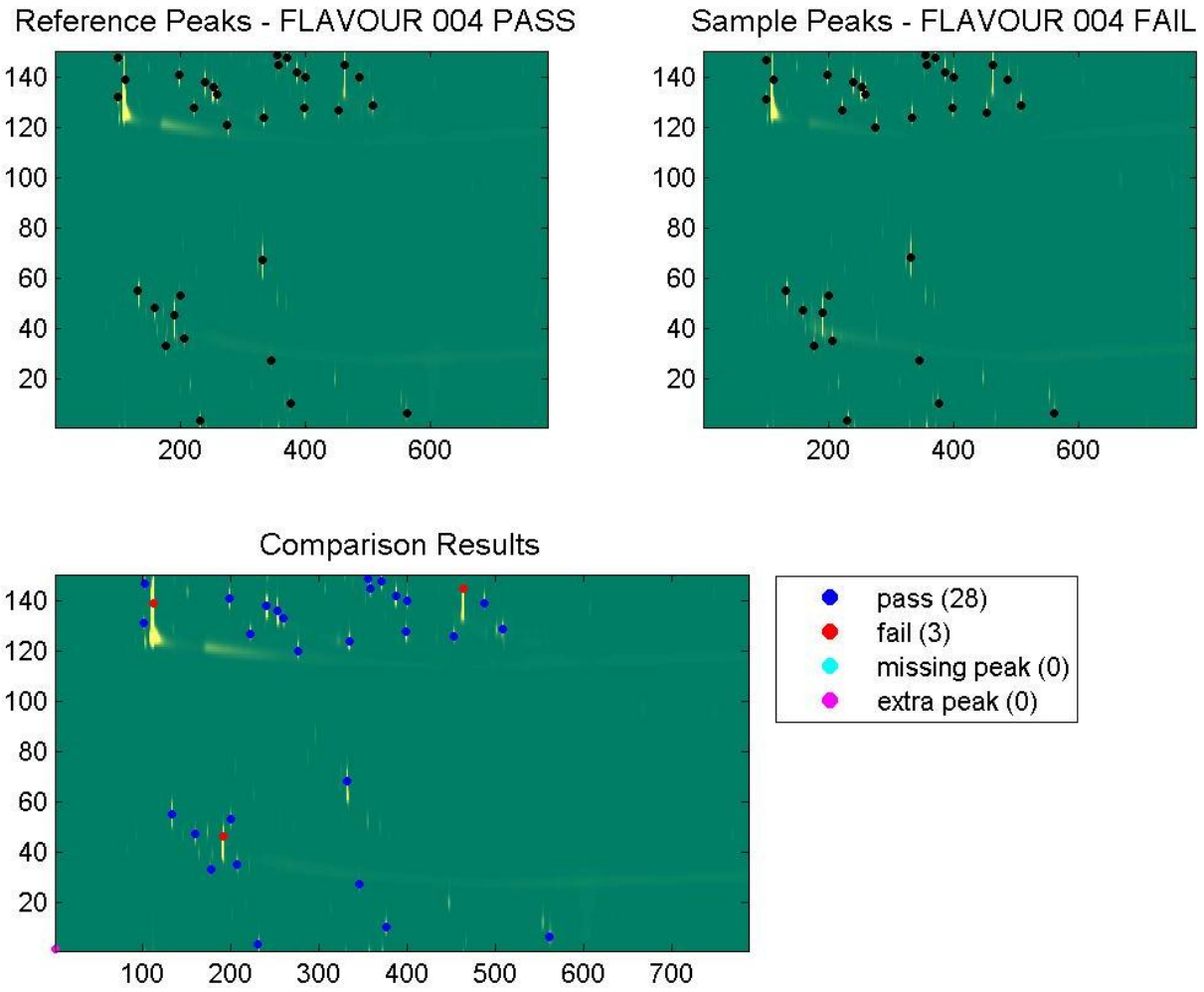
FLAVOUR_004

Date: 25-Jan-2012 22:05:56

Reference: FLAVOUR 004 PASS
Sample: FLAVOUR 004 FAIL
Modulation period (s): 1.5
Sampling frequency (Hz): 100

Tolerance 1: 20.00% - 100.00% = 2.0%
Tolerance 2: 6.00% - 20.00% = 5.0%
Tolerance 3: 0.90% - 6.00% = 10.0%
Tolerance 4: 0.15% - 0.90% = 50.0%
Tolerance 5: 0.03% - 0.15% = 100.0%

Ref 1	Ref 2	Vol%	Smpl 1	Smpl 2	Vol%	Diff	Comment
102	132	0.07	101	131	0.05	-32.45	Pass
102	148	0.06	102	147	0.06	-7.02	Pass
112	139	90.08	112	139	86.91	-3.51	Fail
133	55	0.39	133	55	0.41	3.50	Pass
159	48	0.07	159	47	0.08	12.82	Pass
177	33	0.02	177	33	0.04	83.82	Pass
191	45	1.56	191	46	1.82	16.68	Fail
198	141	0.19	198	141	0.22	16.86	Pass
200	53	0.03	200	53	0.04	38.21	Pass
207	36	0.03	207	35	0.04	36.71	Pass
222	128	0.04	222	127	0.05	18.94	Pass
232	3	0.08	231	3	0.10	29.70	Pass
240	138	0.43	240	138	0.54	24.85	Pass
253	136	0.76	253	136	0.95	23.58	Pass
260	133	0.11	260	133	0.14	24.19	Pass
276	121	0.11	276	120	0.11	-3.98	Pass
332	67	0.53	332	68	0.75	40.43	Pass
334	124	0.11	334	124	0.13	15.72	Pass
345	27	0.02	345	27	0.03	45.33	Pass
356	149	0.06	356	149	0.08	47.67	Pass
358	145	0.03	358	145	0.04	27.99	Pass
371	148	0.09	371	148	0.11	30.31	Pass
376	10	0.04	376	10	0.06	44.88	Pass
387	142	0.14	387	142	0.19	31.86	Pass
399	128	0.23	399	128	0.26	13.60	Pass
400	140	0.03	400	140	0.05	71.79	Pass
453	127	0.04	453	126	0.04	15.80	Pass
463	145	4.41	463	145	6.37	44.48	Fail
487	140	0.05	487	139	0.06	35.71	Pass
508	129	0.05	508	129	0.06	22.82	Pass
562	6	0.04	561	6	0.03	-23.81	Pass


Reference Peaks - FLAVOUR 004 PASS
Sample Peaks - FLAVOUR 004 FAIL
Comparison Results
140
120
100
80
60
40
20
200
400
600
100
200
300
400
500
600
700
pass (28)
fail (3)
missing peak (0)
extra peak (0)

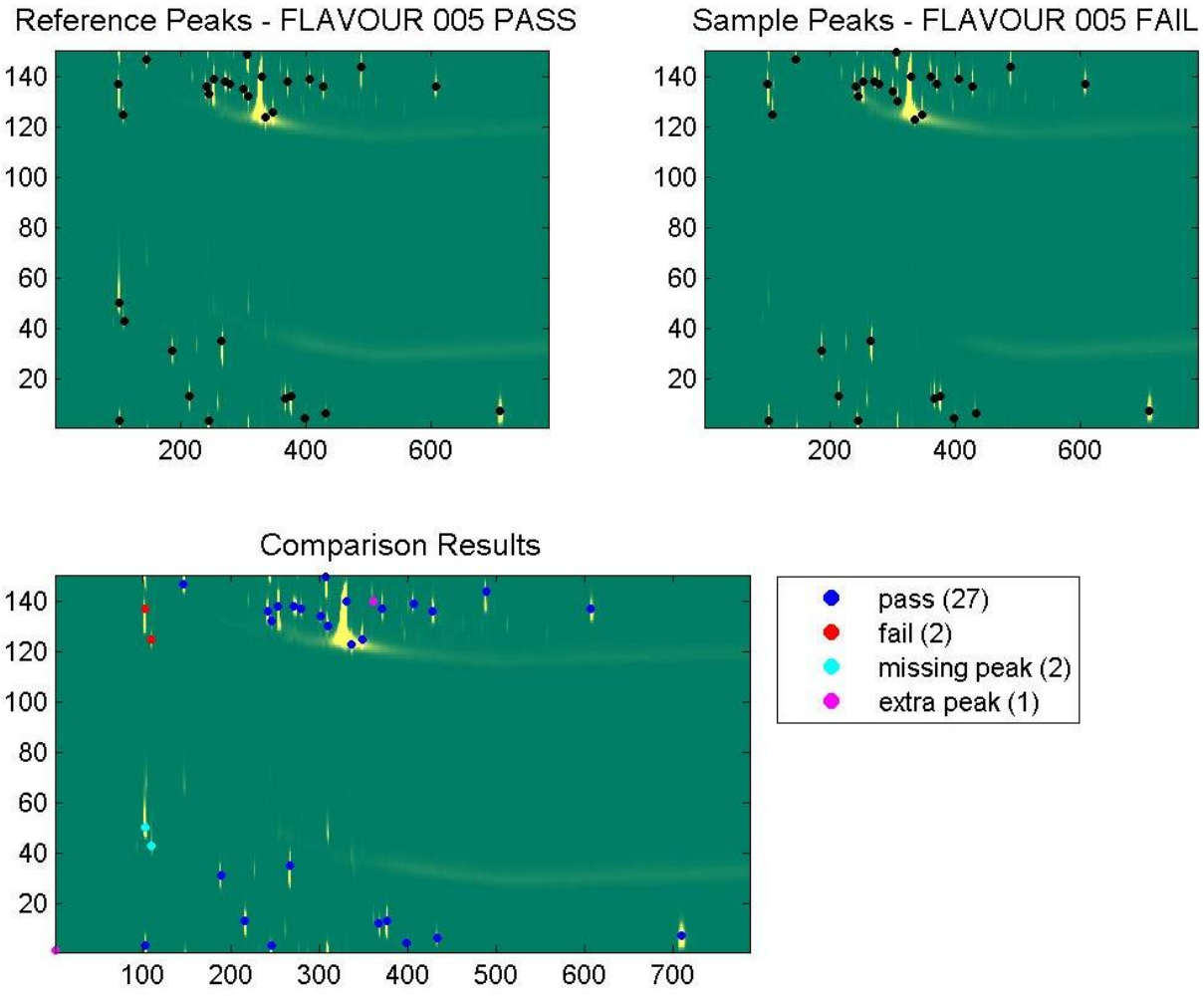
FLAVOUR_005

Date: 25-Jan-2012 22:06:11

Reference: FLAVOUR 005 PASS
Sample: FLAVOUR 005 FAIL
Modulation period (s): 1.5
Sampling frequency (Hz): 100

Tolerance 1: 20.00% - 100.00% = 2.0%
Tolerance 2: 6.00% - 20.00% = 5.0%
Tolerance 3: 0.90% - 6.00% = 10.0%
Tolerance 4: 0.15% - 0.90% = 50.0%
Tolerance 5: 0.03% - 0.15% = 100.0%

Ref 1	Ref 2	Vol%	Smpl 1	Smpl 2	Vol%	Diff	Comment
102	137	6.12	102	137	5.59	-8.61	Fail
103	3	0.78	103	3	0.66	-15.51	Pass
103	50	0.38	0	0	0.00	0.00	Missing Peak
109	125	0.08	109	125	0.18	142.11	Fail
110	43	0.02	0	0	0.00	0.00	Missing Peak
147	147	0.18	146	147	0.19	10.18	Pass
188	31	0.35	188	31	0.36	2.34	Pass
215	13	0.35	215	13	0.34	-0.90	Pass
241	136	0.36	241	136	0.38	4.89	Pass
245	3	0.35	245	3	0.35	1.30	Pass
245	133	0.05	245	132	0.05	-3.24	Pass
253	139	6.12	253	138	6.16	0.61	Pass
266	35	1.44	266	35	1.46	1.04	Pass
271	138	0.20	271	138	0.28	43.74	Pass
279	137	0.04	279	137	0.05	29.73	Pass
301	135	0.03	301	134	0.04	26.92	Pass
307	149	1.98	307	150	2.01	1.56	Pass
308	132	0.07	309	130	0.09	27.45	Pass
330	140	76.37	330	140	76.78	0.53	Pass
336	124	0.54	336	123	0.56	2.79	Pass
348	126	0.71	348	125	0.77	8.26	Pass
367	12	0.19	367	12	0.20	3.27	Pass
371	138	0.31	371	137	0.31	1.67	Pass
376	13	0.62	376	13	0.64	2.26	Pass
399	4	0.02	399	4	0.02	3.86	Pass
406	139	0.09	406	139	0.12	35.67	Pass
428	136	0.28	428	136	0.28	0.65	Pass
432	6	0.03	433	6	0.03	9.99	Pass
488	144	1.14	488	144	1.18	3.54	Pass
607	136	0.18	607	137	0.22	24.15	Pass
709	7	0.61	710	7	0.61	0.00	Pass
0	0	0.00	361	140	0.08	100.00	Extra Peak


Reference Peaks - FLAVOUR 005 PASS
Sample Peaks - FLAVOUR 005 FAIL
Comparison Results
140
120
100
80
60
40
20
200
400
600
100
300
500
700
pass (27)
fail (2)
missing peak (2)
extra peak (1)

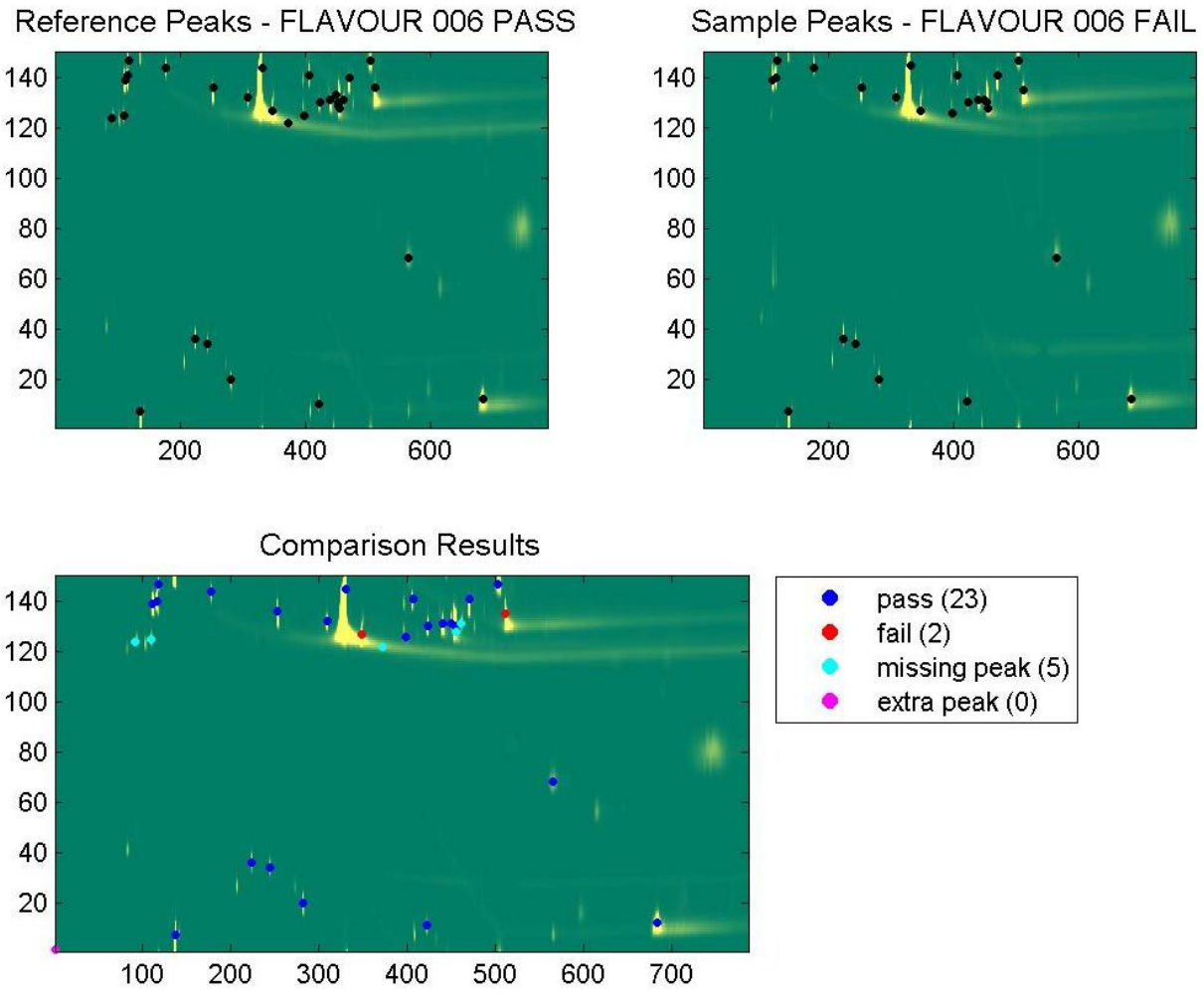
FLAVOUR_006

Date: 25-Jan-2012 22:06:26

Reference: FLAVOUR 006 PASS
Sample: FLAVOUR 006 FAIL
Modulation period (s): 1.5
Sampling frequency (Hz): 100

Tolerance 1: 20.00% - 100.00% = 2.0%
Tolerance 2: 6.00% - 20.00% = 5.0%
Tolerance 3: 0.90% - 6.00% = 10.0%
Tolerance 4: 0.15% - 0.90% = 50.0%
Tolerance 5: 0.03% - 0.15% = 100.0%

Ref 1	Ref 2	Vol%	Smpl 1	Smpl 2	Vol%	Diff	Comment
92	124	0.04	0	0	0.00	0.00	Missing Peak
110	125	0.05	0	0	0.00	0.00	Missing Peak
112	139	0.63	111	139	0.85	34.79	Pass
116	141	0.40	116	140	0.43	6.87	Pass
118	147	0.06	118	147	0.06	-2.97	Pass
137	7	5.28	137	7	5.27	-0.25	Pass
177	144	0.15	177	144	0.15	0.68	Pass
224	36	0.11	224	36	0.11	1.19	Pass
244	34	0.03	244	34	0.03	-1.20	Pass
253	136	0.67	253	136	0.68	1.85	Pass
282	20	0.21	282	20	0.21	0.67	Pass
309	132	0.11	309	132	0.08	-25.33	Pass
331	144	86.89	331	145	87.54	0.75	Pass
348	127	0.92	348	127	0.81	-11.22	Fail
372	122	0.10	0	0	0.00	0.00	Missing Peak
399	125	0.03	399	126	0.02	-8.16	Pass
406	141	0.26	406	141	0.19	-25.67	Pass
422	10	0.02	422	11	0.02	-1.80	Pass
423	130	0.13	423	130	0.13	0.69	Pass
440	131	0.21	440	131	0.22	2.42	Pass
450	133	0.08	450	131	0.02	-69.58	Pass
453	130	0.21	455	128	0.19	-11.14	Pass
454	129	0.37	454	130	0.48	29.52	Pass
455	128	0.08	0	0	0.00	0.00	Missing Peak
461	131	0.05	0	0	0.00	0.00	Missing Peak
470	140	0.45	470	141	0.51	12.69	Pass
504	147	0.27	503	147	0.25	-6.80	Pass
511	136	1.37	511	135	1.16	-14.98	Fail
565	68	0.04	565	68	0.06	46.20	Pass
683	12	0.65	683	12	0.45	-31.08	Pass


Reference Peaks - FLAVOUR 006 PASS
Sample Peaks - FLAVOUR 006 FAIL
Comparison Results
140
120
100
80
60
40
20
200
400
600
100
200
300
400
500
600
700
pass (23)
fail (2)
missing peak (5)
extra peak (0)

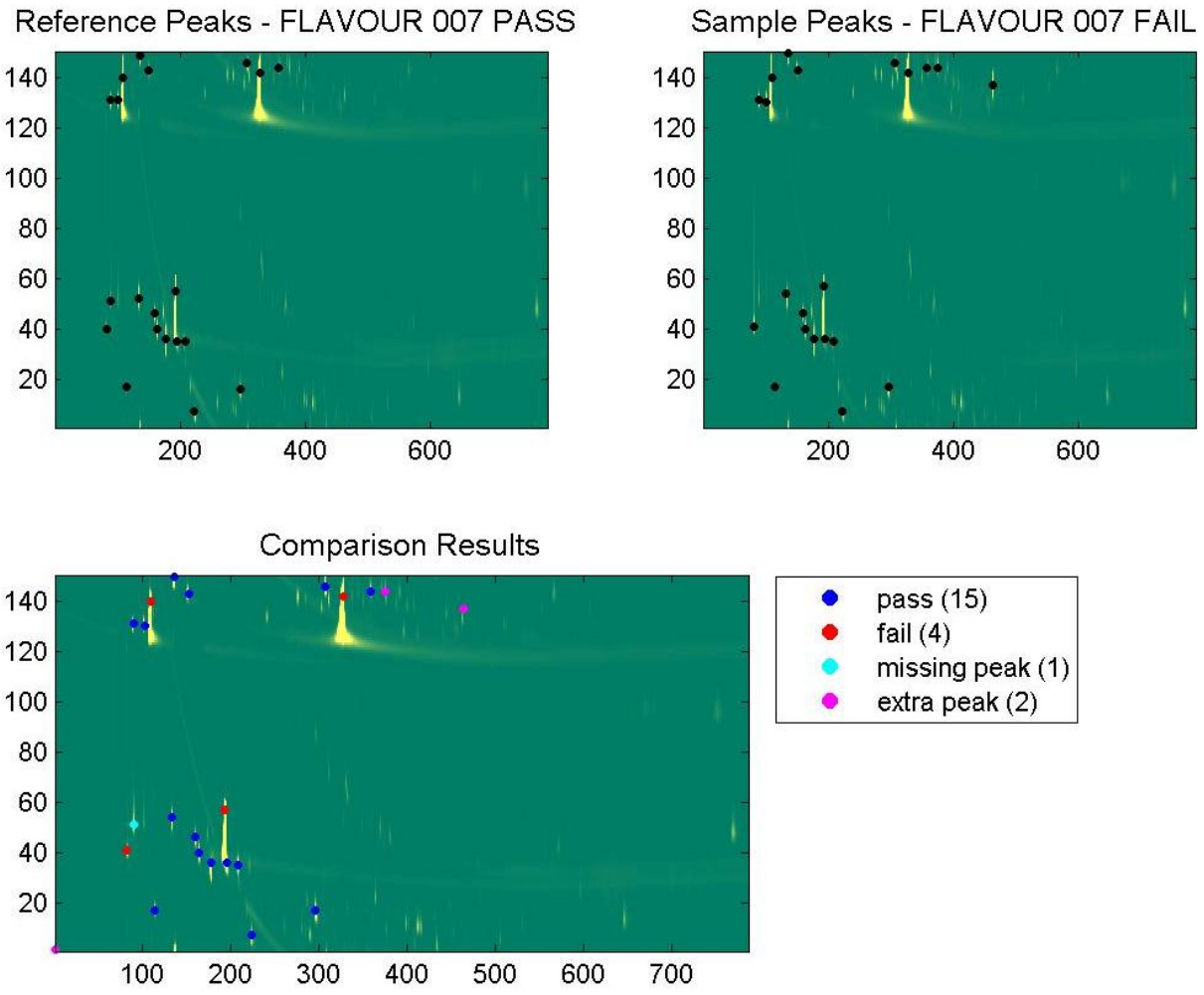
FLAVOUR_007

Date: 25-Jan-2012 22:06:42

Reference: FLAVOUR 007 PASS
Sample: FLAVOUR 007 FAIL
Modulation period (s): 1.5
Sampling frequency (Hz): 100

Tolerance 1: 20.00% - 100.00% = 2.0%
Tolerance 2: 6.00% - 20.00% = 5.0%
Tolerance 3: 0.90% - 6.00% = 10.0%
Tolerance 4: 0.15% - 0.90% = 50.0%
Tolerance 5: 0.03% - 0.15% = 100.0%

Ref 1	Ref 2	Vol%	Smpl 1	Smpl 2	Vol%	Diff	Comment
83	40	0.02	82	41	0.06	210.64	Fail
89	131	0.05	90	131	0.04	-7.49	Pass
90	51	0.03	0	0	0.00	0.00	Missing Peak
102	131	0.12	102	130	0.09	-23.24	Pass
109	140	44.11	110	140	39.89	-9.57	Fail
114	17	0.02	114	17	0.02	5.26	Pass
134	52	0.15	133	54	0.13	-13.50	Pass
136	149	0.25	136	150	0.22	-12.18	Pass
151	143	0.06	152	143	0.06	11.31	Pass
160	46	0.03	160	46	0.03	-2.29	Pass
164	40	0.05	164	40	0.05	-6.49	Pass
177	36	0.42	177	36	0.37	-10.34	Pass
194	55	23.43	193	57	22.22	-5.15	Fail
196	35	0.05	196	36	0.05	-10.55	Pass
208	35	0.04	208	35	0.03	-8.43	Pass
223	7	0.07	223	7	0.07	3.03	Pass
296	16	0.09	296	17	0.10	11.77	Pass
307	146	0.14	307	146	0.16	8.53	Pass
328	142	30.76	328	142	35.86	16.60	Fail
358	144	0.03	358	144	0.04	17.35	Pass
0	0	0.00	375	144	0.02	100.00	Extra Peak
0	0	0.00	463	137	0.30	100.00	Extra Peak


Reference Peaks - FLAVOUR 007 PASS
Sample Peaks - FLAVOUR 007 FAIL
Comparison Results
140
120
100
80
60
40
20
200
400
600
100
300
500
700
pass (15)
fail (4)
missing peak (1)
extra peak (2)

FLAVOUR_008

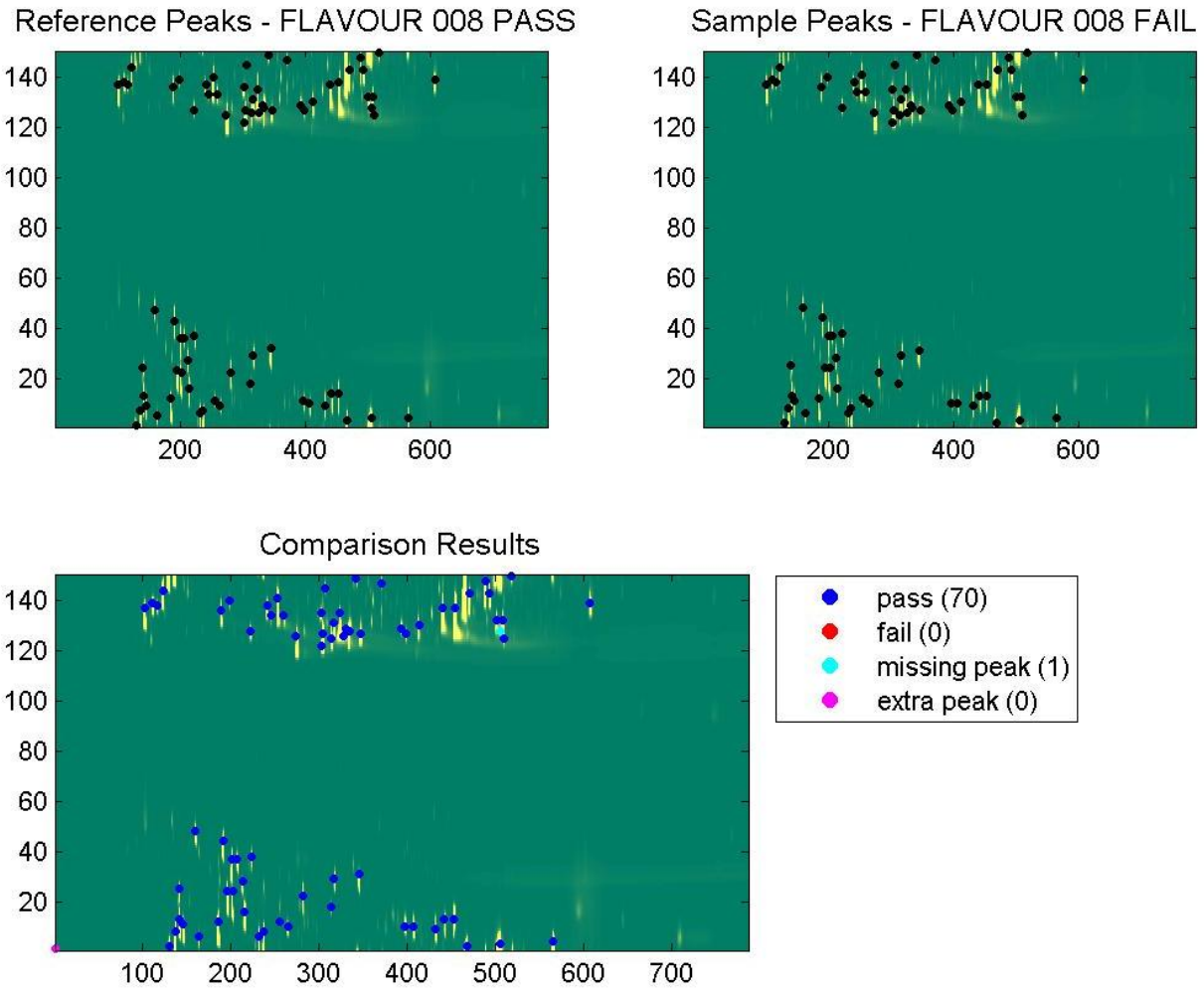
Date: 25-Jan-2012 22:07:35

Reference: FLAVOUR 008 PASS
Sample: FLAVOUR 008 FAIL
Modulation period (s): 1.5
Sampling frequency (Hz): 100

Tolerance 1: 20.00% - 100.00% = 2.0%
Tolerance 2: 6.00% - 20.00% = 5.0%
Tolerance 3: 0.90% - 6.00% = 10.0%
Tolerance 4: 0.15% - 0.90% = 50.0%
Tolerance 5: 0.03% - 0.15% = 100.0%

Ref 1	Ref 2	Vol%	Smpl 1	Smpl 2	Vol%	Diff	Comment
102	137	2.90	102	137	2.95	1.89	Pass
111	138	0.21	111	139	0.21	-1.00	Pass
117	137	0.06	116	138	0.06	7.06	Pass
123	144	1.91	123	144	1.93	1.08	Pass
130	1	1.73	130	2	1.75	0.68	Pass
137	7	5.17	137	8	5.23	1.12	Pass
141	24	0.14	141	25	0.14	-1.09	Pass
142	13	1.63	142	13	1.64	0.38	Pass
146	9	0.55	146	11	0.57	3.43	Pass
160	47	0.14	160	48	0.13	-6.51	Pass
163	5	0.07	163	6	0.07	-7.96	Pass
186	12	2.19	186	12	2.21	1.18	Pass
189	136	0.53	189	136	0.53	0.63	Pass
191	43	0.82	191	44	0.83	1.03	Pass
196	23	1.17	196	24	1.18	0.92	Pass
199	139	0.03	199	140	0.03	-9.76	Pass
201	36	1.22	201	37	1.22	0.64	Pass
202	22	1.71	202	24	1.73	0.94	Pass
207	36	0.09	207	37	0.09	0.82	Pass
213	27	0.20	213	28	0.20	0.62	Pass
215	16	0.90	215	16	0.91	0.89	Pass
222	127	0.07	222	128	0.06	-18.08	Pass
223	37	0.12	223	38	0.12	0.23	Pass
232	6	0.47	232	6	0.47	0.81	Pass
237	7	1.15	237	8	1.16	0.90	Pass
241	137	0.20	241	138	0.20	1.37	Pass
245	133	0.03	245	134	0.03	2.08	Pass
253	140	4.40	253	141	4.47	1.63	Pass
256	11	0.02	256	12	0.02	4.47	Pass
260	133	0.10	260	134	0.10	1.21	Pass
264	9	0.05	265	10	0.05	0.75	Pass

274	125	2.51	274	126	2.57	2.52	Pass
282	22	0.91	282	22	0.91	0.14	Pass
302	136	0.13	302	135	0.10	-25.26	Pass
303	122	0.05	303	122	0.05	18.84	Pass
304	127	2.27	304	127	2.31	1.69	Pass
307	145	0.02	307	145	0.02	-0.60	Pass
313	18	0.07	313	18	0.05	-17.92	Pass
314	126	0.50	314	125	0.51	0.08	Pass
317	29	0.26	317	29	0.26	-0.32	Pass
317	131	0.26	317	131	0.26	-0.43	Pass
324	135	2.63	324	135	2.62	-0.55	Pass
327	126	0.08	327	126	0.06	-30.67	Pass
331	129	0.13	331	129	0.12	-10.68	Pass
334	128	1.48	334	128	1.50	1.37	Pass
342	149	0.07	342	149	0.07	-0.53	Pass
345	32	0.49	345	31	0.49	-0.91	Pass
347	127	0.31	347	127	0.31	1.14	Pass
371	147	0.02	371	147	0.02	-1.87	Pass
393	129	0.06	393	129	0.06	1.32	Pass
397	11	0.10	397	10	0.09	-0.33	Pass
399	127	0.09	399	127	0.09	-0.25	Pass
407	10	0.18	407	10	0.18	-0.20	Pass
413	130	0.56	413	130	0.56	0.66	Pass
432	9	0.25	432	9	0.25	-0.64	Pass
440	137	1.93	440	137	1.93	-0.14	Pass
442	14	0.55	442	13	0.54	-0.69	Pass
453	14	0.83	453	13	0.82	-0.92	Pass
454	138	7.03	454	137	6.98	-0.74	Pass
467	3	38.24	468	2	38.05	-0.49	Pass
470	143	1.30	470	143	1.29	-0.42	Pass
488	148	1.80	488	148	1.79	-0.47	Pass
493	143	0.30	493	143	0.30	-0.74	Pass
501	132	0.19	501	132	0.17	-10.34	Pass
505	4	3.26	505	3	3.22	-1.11	Pass
505	128	0.03	0	0	0.00	0.00	Missing Peak
508	132	0.27	508	132	0.30	13.67	Pass
509	125	0.03	509	125	0.03	16.21	Pass
518	150	0.05	518	150	0.05	1.76	Pass
565	4	0.37	565	4	0.37	-1.67	Pass
607	139	0.28	607	139	0.28	-0.09	Pass


Reference Peaks - FLAVOUR 008 PASS
Sample Peaks - FLAVOUR 008 FAIL
Comparison Results
140
120
100
80
60
40
20
200
400
600
100
200
300
400
500
600
700
pass (70)
fail (0)
missing peak (1)
extra peak (0)

FLAVOUR_009

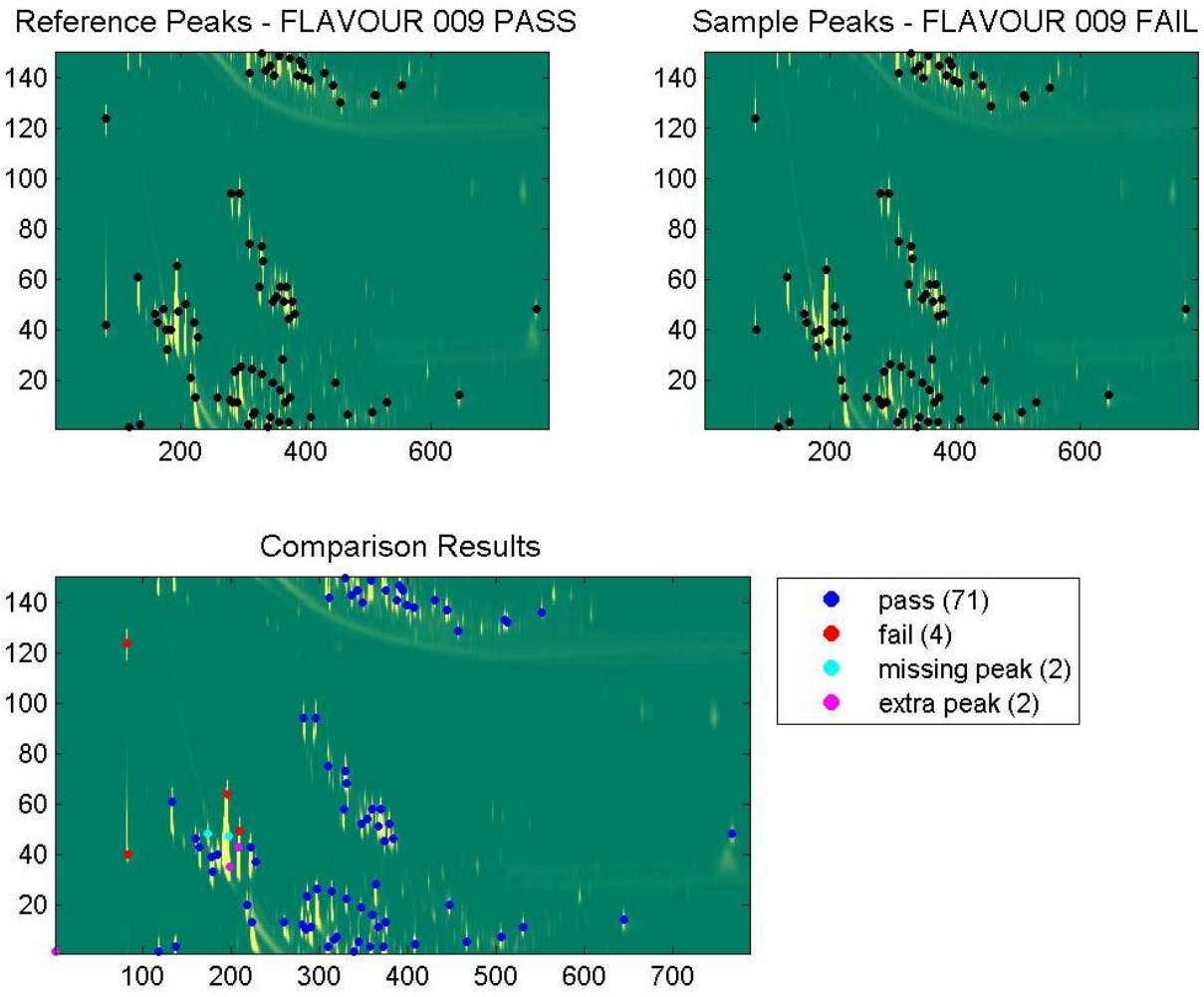
Date: 25-Jan-2012 22:09:51

Reference: FLAVOUR 009 PASS
Sample: FLAVOUR 009 FAIL
Modulation period (s): 1.5
Sampling frequency (Hz): 100

Tolerance 1: 20.00% - 100.00% = 2.0%
Tolerance 2: 6.00% - 20.00% = 5.0%
Tolerance 3: 0.90% - 6.00% = 10.0%
Tolerance 4: 0.15% - 0.90% = 50.0%
Tolerance 5: 0.03% - 0.15% = 100.0%

Ref 1	Ref 2	Vol%	Smpl 1	Smpl 2	Vol%	Diff	Comment
82	42	0.17	83	40	0.05	-69.06	Fail
82	124	1.00	82	124	1.14	13.91	Fail
118	1	1.38	118	1	1.27	-7.55	Pass
133	61	2.30	133	61	2.16	-6.13	Pass
137	2	1.43	137	3	1.32	-7.58	Pass
160	46	0.15	160	46	0.14	-6.39	Pass
164	43	0.82	164	43	0.77	-5.77	Pass
173	48	0.12	0	0	0.00	0.00	Missing Peak
177	40	2.12	177	39	2.07	-2.27	Pass
179	32	0.10	179	33	0.17	68.66	Pass
185	40	0.59	185	40	0.56	-5.86	Pass
196	65	51.18	196	64	59.33	15.94	Fail
197	47	3.92	0	0	0.00	0.00	Missing Peak
209	50	8.28	209	49	3.99	-51.82	Fail
217	21	0.85	218	20	0.78	-7.90	Pass
222	43	2.16	222	43	2.05	-5.15	Pass
224	13	2.43	224	13	2.30	-5.34	Pass
228	37	0.03	228	37	0.03	-2.63	Pass
260	13	0.23	260	13	0.22	-4.20	Pass
280	12	0.14	280	12	0.14	6.05	Pass
282	94	0.16	282	94	0.17	1.19	Pass
284	11	0.16	284	10	0.15	-4.16	Pass
286	23	0.18	286	23	0.16	-6.51	Pass
290	11	0.61	290	11	0.59	-3.13	Pass
295	94	0.44	295	94	0.42	-3.27	Pass
297	25	3.66	297	26	3.58	-2.40	Pass
309	2	4.03	309	3	3.89	-3.49	Pass
310	74	0.14	310	75	0.15	5.20	Pass
311	142	0.15	311	142	0.15	-1.71	Pass
314	24	1.59	314	25	1.58	-0.69	Pass
317	6	0.05	317	6	0.05	-6.64	Pass
319	7	0.11	319	7	0.11	-5.22	Pass
327	57	0.23	327	58	0.23	-2.08	Pass

329	73	0.38	329	73	0.36	-4.59	Pass
329	150	0.04	329	150	0.03	-5.92	Pass
330	22	0.16	330	22	0.16	-0.19	Pass
331	67	0.31	331	68	0.31	-2.66	Pass
336	143	0.13	336	143	0.11	-13.58	Pass
339	1	0.10	339	1	0.09	-6.42	Pass
343	145	0.06	343	145	0.06	-6.34	Pass
344	5	0.04	344	5	0.04	-3.76	Pass
347	19	0.03	347	19	0.03	-0.54	Pass
347	51	0.07	347	52	0.07	-0.33	Pass
349	141	0.10	349	140	0.09	-5.72	Pass
354	53	0.06	354	54	0.05	-5.54	Pass
357	3	0.95	357	3	0.89	-6.24	Pass
358	149	0.84	358	149	0.81	-2.84	Pass
359	57	0.08	359	58	0.08	-3.92	Pass
360	16	0.04	360	16	0.04	-2.64	Pass
363	28	0.54	363	28	0.53	-1.97	Pass
366	51	0.10	366	51	0.11	16.98	Pass
367	11	0.06	367	11	0.06	-2.09	Pass
369	57	0.91	369	58	0.86	-5.77	Pass
372	3	1.12	372	3	1.07	-4.41	Pass
373	44	0.08	373	45	0.08	1.38	Pass
374	148	0.49	375	145	0.45	-7.40	Pass
375	13	0.07	375	13	0.07	-1.49	Pass
379	51	0.37	379	52	0.34	-8.36	Pass
383	46	0.03	383	46	0.03	-5.78	Pass
387	141	0.15	387	141	0.14	-4.72	Pass
390	147	0.33	390	147	0.31	-5.90	Pass
394	145	0.04	394	145	0.04	-3.79	Pass
399	140	0.22	399	139	0.21	-5.51	Pass
407	139	0.09	407	138	0.09	-6.28	Pass
408	5	0.09	408	4	0.09	-3.45	Pass
430	142	0.04	430	141	0.04	-9.16	Pass
444	137	0.19	444	137	0.17	-6.16	Pass
447	19	0.04	447	20	0.04	-5.64	Pass
456	130	0.09	457	129	0.08	-7.32	Pass
466	6	0.10	466	5	0.09	-4.32	Pass
505	7	0.08	505	7	0.07	-5.67	Pass
509	133	0.09	509	133	0.08	-9.22	Pass
512	133	0.06	512	132	0.06	-8.67	Pass
530	11	0.04	530	11	0.04	-9.50	Pass
552	137	0.05	551	136	0.04	-13.96	Pass
644	14	0.04	644	14	0.05	36.24	Pass
767	48	0.03	767	48	0.04	7.57	Pass
0	0	0.00	198	35	0.16	100.00	Extra Peak
0	0	0.00	208	43	1.45	100.00	Extra Peak


Reference Peaks - FLAVOUR 009 PASS
Sample Peaks - FLAVOUR 009 FAIL
Comparison Results
pass (71)
fail (4)
missing peak (2)
extra peak (2)

FLAVOUR_010

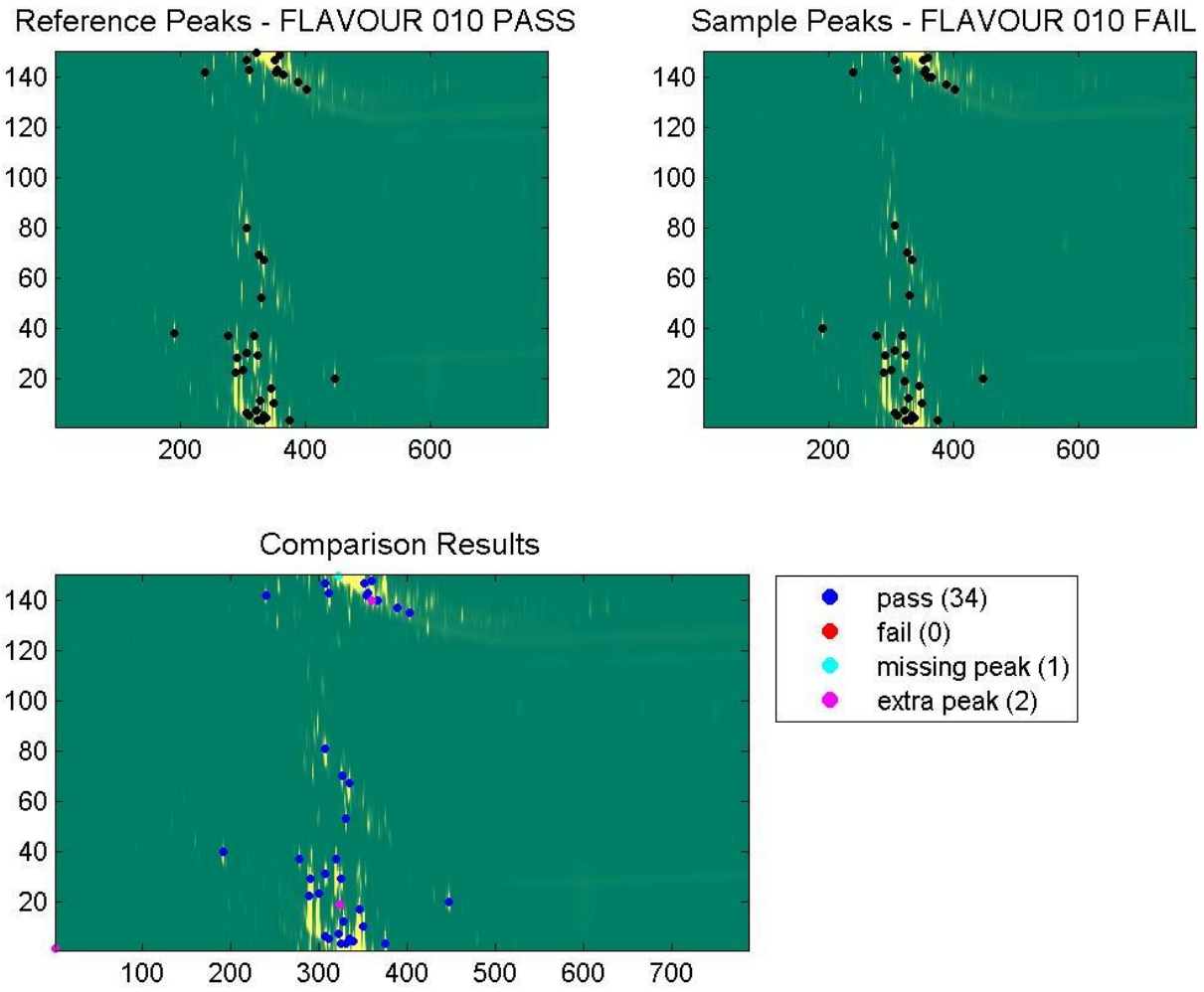
Date: 25-Jan-2012 22:10:21

Reference: FLAVOUR 010 PASS
Sample: FLAVOUR 010 FAIL
Modulation period (s): 1.5
Sampling frequency (Hz): 100

Tolerance 1: 20.00% - 100.00% = 2.0%
Tolerance 2: 6.00% - 20.00% = 5.0%
Tolerance 3: 0.90% - 6.00% = 10.0%
Tolerance 4: 0.15% - 0.90% = 50.0%
Tolerance 5: 0.03% - 0.15% = 100.0%

Ref 1	Ref 2	Vol%	Smpl 1	Smpl 2	Vol%	Diff	Comment
191	38	0.03	191	40	0.03	1.05	Pass
240	142	0.02	240	142	0.02	0.18	Pass
278	37	0.03	278	37	0.02	-3.75	Pass
288	22	2.31	288	22	2.31	0.16	Pass
290	28	15.48	290	29	15.48	0.02	Pass
300	23	8.78	300	23	8.77	-0.01	Pass
307	6	0.06	307	6	0.06	-0.40	Pass
307	30	0.09	307	31	0.08	-1.82	Pass
307	80	0.16	307	81	0.16	-1.51	Pass
307	147	0.12	307	147	0.12	-0.65	Pass
311	5	0.04	311	5	0.03	-17.22	Pass
311	143	0.15	311	143	0.15	-0.01	Pass
319	37	3.59	319	37	3.55	-0.92	Pass
322	7	1.77	322	7	1.83	3.66	Pass
322	150	0.06	0	0	0.00	0.00	Missing Peak
325	3	0.21	325	3	0.21	1.10	Pass
325	29	0.08	325	29	0.08	-2.04	Pass
326	69	0.03	326	70	0.03	-1.63	Pass
328	11	5.99	328	12	5.99	0.00	Pass
330	52	0.04	330	53	0.03	-3.40	Pass
331	3	0.20	331	3	0.20	-0.43	Pass
334	5	0.81	334	5	0.82	0.33	Pass
334	67	0.21	334	67	0.21	-1.10	Pass
338	4	1.01	338	4	1.04	2.59	Pass
346	16	53.06	346	17	53.09	0.06	Pass
350	10	3.38	350	10	3.39	0.31	Pass
351	147	0.30	351	147	0.29	-4.74	Pass
354	142	0.03	354	142	0.03	0.60	Pass
356	143	0.03	356	143	0.03	-1.11	Pass
359	149	0.61	359	148	0.49	-19.61	Pass
366	141	0.04	366	140	0.03	-20.43	Pass

375	3	1.05	375	3	1.05	-0.20	Pass
388	138	0.04	388	137	0.03	-16.64	Pass
403	135	0.02	403	135	0.02	1.54	Pass
447	20	0.03	447	20	0.03	-4.40	Pass
0	0	0.00	323	19	0.02	100.00	Extra Peak
0	0	0.00	360	140	0.10	100.00	Extra Peak


Reference Peaks - FLAVOUR 010 PASS
Sample Peaks - FLAVOUR 010 FAIL
Comparison Results
140
120
100
80
60
40
20
200
400
600
100
200
300
400
500
600
700
pass (34)
fail (0)
missing peak (1)
extra peak (2)