1. Introduction

Non-food biomass crops e.g. switchgrass (*Panicum virgatum* L.), *Miscanthus x giganteus*, and short-rotation coppice poplar (*Populus* spp.) and willow (*Salix* spp.) offer a sustainable source of energy and platform chemicals (Sims et al., 2006). The majority of the energy stored in the crop biomass is in the cell wall which constitutes the largest fraction of lignocellulosic biomass. The three polymers that constitute the bulk of plant cell wall (cellulose, hemicellulose and lignin) rank amongst the most abundant biopolymers on the planet and their proportional concentrations range generally between 40 – 50%, 10 – 40% and 5 – 30% of biomass by weight respectively (McKendry, 2002). The absolute and relative concentrations of the components of the cell wall have a great influence on biomass quality i.e. its suitability for conversion to heat, power and chemical products. However, because biomass can be utilised by a number of conversion routes with differing feedstock demands measures of feedstock quality are often quite specific to how the material is to be used. For example, biomass can be processed thermochemically. These routes include combustion or co-combustion with coal to generate heat and electricity (Allison et al., 2010). Alternatively, biomass can be converted by fast pyrolysis to bio-char, which is receiving much attention as a soil improver and a means of sequestering carbon from the atmosphere into the soil (Laird, 2008; Woolf et al., 2010), and bio-oil, a liquid fuel (Bridgewater, 2003; Mohan et al., 2006). Biomass can also be gasified to produce a combustible gas which has application for the generation of heat and power, and for the chemical synthesis of liquid transport fuels and industrial chemicals (Ptasinski et al., 2007). These thermochemical processes demand feedstocks with low moisture content and high energy density, often equating with high levels of the poly aromatic polymer, lignin.

In contrast, non-thermochemical processes e.g. the production of bioethanol and industrial platform chemicals by biological conversion processes are often inhibited by high levels of lignin. High concentrations of lignin in the feedstock necessitate harsh chemical and heat pre-treatments of the biomass prior to enzymic saccharification. This increases energy inputs and often damages the polysacharide components of the cell wall giving rise to inhibitory products (Carroll and Somerville, 2009; Chang, 2007; Fahmi et al., 2007; Fahmi et al., 2008; Grabber, 2005). There is therefore considerable pressure to optimise feedstock composition and at present the most feasible way to achieve this at a commercial scale is by breeding.
improvement (Clifton-Brown et al., 2008), although agronomic practice may also influence composition (Hodgson et al., 2010). Crop breeding places huge demands on the analyst in terms of large sample numbers and there is a demand for methods that can cope with high rates of sample throughput whilst being of low unit cost.

Biomass is traditionally assessed in terms of cell wall fibre content measured by direct, or as they are often called, gravimetric methods. The different types of fibre are isolated by successively harsh chemical treatments which selectively and sequentially remove the different classes of structural carbohydrates until only lignin remains. The proportion of each fibre fraction in the biomass is quantified by sample weight change. Measurements of this type are not entirely quantitative as treatments are not wholly selective, and it is often impossible to directly compare cell wall parameters measured by different methods (Hatfield et al., 1999). In addition, these procedures are time consuming, costly and of low throughput (Giger-Reverdin, 1995).

The acetyl bromide method for the quantification of lignin is an indirect method that has gained in popularity over recent years. The method was first published by Johnson et al. in 1961 (Johnson et al., 1961) and a modified version of the method has been used to analyse lignin in a wide range of species (Fukushima and Dehority, 2000) from relatively small samples of tissue. Lignin is dissolved from purified cell wall material by extraction into a solution of acetyl bromide in concentrated acetic acid at 50°C, reacted with hydroxylamine and quantified by absorbance change at 280 nm. Quantification requires reference to a standard curve that is set up using known amounts of standard lignin extracted from similar plant material with acetyl bromide (Fukushima and Dehority, 2000) or acidic dioxane (Fukushima and Hatfield, 2001). The concentrations of lignin detected in samples by this method are comparable with those obtained using the widely accepted gravimetric Klason method (Hatfield and Fukushima, 2005) and the method has the advantage that it is more readily worked into a high-throughput scheme (Foster et al., 2010). Adapting direct and indirect methods for high rates of sample throughput is often complex and most likely requires investment in expensive robotised laboratory equipment that may be beyond the resources of many groups (Foster et al., 2010). Consequently many researchers have turned to less expensive technology for new high-throughput methods.

Infrared (IR) spectrometry offers researchers and breeders an alternative approach that is robust and rapid. The majority of the carbon based molecules in animals and plants are highly active in the IR. At its simplest mid-IR spectroscopy is a useful analytical approach in its own right that provides structural information on samples of pure compounds. Generally, the approach taken is to correlate IR spectra with analytical data obtained from the same samples using multivariate regression methods. This approach although seemingly complex overcomes possible nonlinear relationships between absorption and concentration that are encountered in the IR. Such deviations from the Beer-Lambert law may be caused by spectral shifts due to hydrogen bonding, co-variance with other components in the sample and poor design in older instruments (Hsu, 1997). The regression models produced can be then applied to predict the concentrations of the cell wall components in new samples. Properly executed, this approach has been shown to be rapid and robust and therefore reduces the requirement for standard chemical analysis.

Two types of IR spectroscopy have found application for the measurement of chemical composition in biomass: mid-infrared spectroscopy and near infrared reflectance spectroscopy (NIRS). NIRS has a longer history of being used to predict chemical composition in bulk plant samples as until recently the method was more applicable to bulk...
plant samples. Mid-IR spectroscopy in contrast to NIRS informs on fundamental molecular vibrations, rather than on harmonic and overtone absorptions, and for this reason gives a better insight into the molecular bonds present in the sample. Over recent years the technique has been revolutionised by the development of Fourier transform instruments (FTIR) and by improvements in methods for sample presentation and many groups are choosing analysis by FTIR in preference to NIRS. Several reports have shown that the combination of FTIR and multivariate analysis is highly effective. For example, spectra of wood can be used to discriminate between tree species (Huang et al., 2008). Recently, we reported on using FTIR to predict the concentration of lignin and hydroxycinnamic acids (Allison et al., 2009), nitrogen and alkali index (Allison et al., 2009) in samples of grasses by partial least squares regression. In addition to the analysis of bulk samples, FTIR has long had application in the field of microscopy. For example, FTIR microscopy has been used to identify and characterise cell wall mutants and transgenic plants altered in cell wall biosynthetic genes (Chen et al., 1998; McCann et al., 2007; Mouille et al., 2003; Stewart et al., 1997).

Although it is not intended to discuss FTIR microscopy in depth in this chapter because it is not a method of choice for high-throughput screening it deserves some brief discussion and comparison with allied methods. For the study of biological samples FTIR microscopy is often used as a transmission technique where the IR beam passes through the sample. Special objectives are available (see Section 4) which work through contact with the sample but this results in damage to most biological specimens. FTIR microscopy is best suited to the high resolution mapping of chemical composition at the level of the cell. In this type of study high rates of sample throughput and the prediction of composition in bulk samples are not usually experimental objectives. Earlier FTIR microscope instruments took a long time to acquire spectra because the beam passing through the sample filled only part of the detector. New developments in FTIR microscope detector design have decreased spectral acquisition time substantially and many groups have taken advantage of focal plane array (FPA) detectors. These were originally developed for military use and they allow the simultaneous acquisition of many spectra. Over the last decade, FPA detectors have grown from arrays of 8x8 elements to 16 x 16 and now 64 x 64 or even 128 x 128 element detectors are available for groups with sufficient funds. FPA-FTIR microscopes allow very rapid chemical mapping over large areas but the technique is not without fundamental problems. FTIR microscopy is incompatible with water, which prevents the analysis of living samples, and the analyst must be judicious in the use of chemical fixatives to preserve the structure of sectioned materials as the absorption peaks from these substances may mask important features in the spectral data. Furthermore, spatial resolution in the IR is limited to 2 – 5 µm by wavelength and whilst this does not prevent resolution of most plant cells, which range generally between 10 – 50 µm in diameter, cellular components and structures may be substantially smaller than this size and therefore below the limit of resolution. Raman microscopy, which will not be discussed in any depth here, offers a solution to many of these problems. For many years Raman microscopy was considered an expensive and unusual technique despite offering a spatial resolution of better than 1 µm (Schmidt et al., 2010). In addition, Raman microscopy is not affected by water thus making possible the analysis of living samples and samples where cellular structure has been preserved cryogenically. The relatively low signal strength of Raman emission were problematic and often led to poor signal and lengthy analysis times. This matter has been largely overcome by improvements in instrument design and new sensitive Raman techniques e.g. surface enhanced Raman (Knauer et al., 2010) may result in Raman soon becoming the best way to
study and understand the structural organisation of the cell wall at the cellular level. The resolution and flexibility of Raman imaging has been demonstrated recently by several groups studying the ultra-structure and composition of cell wall in tree species and corn stover (Agarwal, 2006; Gierlinger and Schwanninger, 2006; Sun et al., 2010).

In this review we discuss the principles underlying IR spectroscopy, the developments in instrumentation and sample presentation methods that have led to Fourier transform instruments becoming much more compatible with high through-put analysis, and we present data showing how FTIR is being used to assist breeding improvement work in Aberystwyth.

2. The interaction of matter and IR energy

To understand IR spectroscopy it is necessary to have some insight into how energy interacts with the molecules present in living cells. The bonds in molecules are not rigid nor of fixed length and both, bond length and their angle to each other vary about a mean position and vibrate with specific frequencies. Covalent molecules absorb energy in the IR when the frequency of the energy correlates exactly with the vibrational frequency of a chemical bond within the molecule. This absorbance of energy by the molecule causes an increase in the amplitude, but not the frequency of the bond’s vibration about its mean centre. By definition, absorption of IR energy can only occur when it causes a change to the bond’s electric dipole moment, i.e. the bond’s polarity, or the separation of positive and negative electrical charges comprising the covalent bond. Simple homonuclear molecules e.g. N₂ and O₂ cannot absorb in the IR because there is no dipole moment to change. In contrast, it is common for many bonds in covalent heteronuclear molecules to exhibit some degree of polarity e.g. H₂O and CO₂ and most organic compounds absorb strongly in the IR with characteristic absorption frequencies.

The visible, IR and ultraviolet parts of the electromagnetic (EM) spectrum are defined by frequency (ν; the number of wave-cycles passing through a point in one second) measured in Hertz (Hz) and wave length (λ; the length of one complete wave-cycle). The mid-IR region stretches between 3 x 10⁻⁴ and 3 x 10⁻³ cm and is of greatest practical use to the organic chemist because this range corresponds to the fundamental frequencies of molecular vibrations. Frequency and wavelength are related to each other by equation 1, where \( c \) is the velocity of light in a vacuum, and has the value 2.997 x 10⁸ m s⁻¹:

\[
c = \lambda \nu
\]  

Experiments in the early part of the 20th Century demonstrated that when interacting with matter EM radiation is best described as discrete packets of energy called quanta. The energy (\( E \)) of a quantum is directly proportional to its frequency and is calculated by the Bohr equation (equation 2) where \( h \) is the Planck constant (\( h = 6.626 \times 10^{-34} \) Joules-second). Substitution of equation 1 into equation 2 shows that energy decreases with increasing wave length.

\[
E = h\nu
\]  

In IR spectroscopy wavelength is often expressed as its reciprocal, wave number (cm⁻¹), which equates to the number of waves per centimetre. In the mid-IR wave numbers span from 4000 – 400 cm⁻¹; this is a more convenient notation and has the advantage of being proportional with energy.
The infrared spectrum of a sample of a pure compound shows the position of absorptions within the IR spectrum and therefore indicates the functional groups in the molecule. The spectrum can be presented in terms of transmission, where the y axis ranges from 0 – 100% transmission over a range of wave numbers, or in terms of absorption, in which percentage transmission (%T) is converted to absorption (A) by the simple relationship shown in equation 4.

\[ A = \log_{10} \left( \frac{100}{%T} \right) \]  

(3)

For qualitative analysis absorption has the advantage of being proportional to the concentration of the analyte, which can be calculated by equation 4, the Beer-Lambert Law in which \( c \) is the molar concentration of analyte, \( \epsilon \) the molar extinction coefficient (L mol\(^{-1}\) cm\(^{-1}\)) and \( l \) the light path through the sample (cm).

\[ A = \epsilon cl \]  

(4)

Even simple compounds have relatively complex IR spectra showing many absorption bands. This is because there may be many bonds within an individual molecule that are able to absorb in the mid-IR. Bonds within a molecule can vibrate in two ways corresponding to the movements of atoms sharing a bond. Atoms can move relative to each other causing the bond to vary in length, this causes bond stretching, or one atom can move out of plane relative to the other, causing bond bending. For reasons not explained here the maximum number of absorptions in the mid-IR that might be expected from a pure compound comprised of \( N \) atoms can be calculated by equation 5 for linear molecules such as CO\(_2\) and by equation 6 for non-linear molecules such as H\(_2\)O.

\[ \text{Linear} = 3N - 5 \]  

(5)

\[ \text{Non-linear} = 3N - 6 \]  

(6)

Not all possible vibrational modes are active in the IR, for example CO\(_2\) would be expected to have \( 3N - 5 \) absorption bands in its spectrum, in reality it has two as one of the possible stretching modes is symmetrical and not IR active, furthermore the two bending modes are degenerate and show as one combined band. Other reasons why fewer than the theoretical number of IR bands are seen in the spectrum include the absorption not being in the 4000–400 cm\(^{-1}\) range; an absorption being too weak to be observed and absorptions being too close to each other to be resolved on the instrument. The vibration frequency of the absorption in wave numbers (\( \bar{v} \)) is proportional with the force constant or bond stiffness (\( k \), dyne cm\(^{-1}\)) and the masses of the atoms (in grams) sharing the bond (equation 7).

\[ \bar{v} = \frac{1}{2\pi c} \sqrt{\frac{k(m1+m2)}{(m1m2)}} \]  

(7)

3. **The Fourier transform infrared spectrometer**

Modern mid-IR spectrophotometers are almost exclusively Fourier transform instruments as this design offers greater sensitivity and considerably faster scan speeds compared with the older and now largely obsolete dispersive instruments. The main difference between the
two designs is the manner in which the IR beam is produced which is passed through the sample. Dispersive instruments are based on a monochromator; this device splits a polychromatic beam produced by the source into a beam of very narrow wave number. As the instrument scans over a wave number range the monochromator shifts the wave number of the beam through the scan range. This of course means that the time required for each scan is dependant on how quickly the monochromator can be made to work and at any one point in the scan the beam transmitted through the sample is only a small portion of the sources output. In contrast, a beam covering the entire wave number range of the source is passed through the sample in a Fourier transform instrument. Therefore scan time and beam intensity are considerably faster and brighter respectively. The heart of a Fourier transform instrument is an interferometer. Most commonly this is a Michelson interferometer (Figure 1). In this device IR light from a Nernst or Globar source is split at a beam splitter into two beams of equal intensity, one beam is reflected onto a stationary mirror whilst the other is reflected onto a moving mirror. The two beams are recombined to form the transmitted beam at the beam splitter, and this beam is at 90˚to the input beam. The moving mirror produces a varying optical path difference between the two beams resulting in constructive and destructive interference when they are combined. A helium neon laser beam is included in the FTIR spectrophotometer to provide a reference beam of known wave number that can be used to measure precisely the displacement of the moving mirror. This laser is not shown in Figure 1 for the purpose of simplicity. Two types of detector are commonly used in the mid-IR: The deuterium triglycine sulphate (DTGS) detector works at room temperature and has the advantage of great stability and ease of use. For more demanding work increased sensitivity of the mercury cadmium telluride (MCT) may be necessary but this type of detector requires cooling to near liquid nitrogen temperatures in order to work. The resulting interferogram (Figure 2) contains the source’s frequency information modulated in a time domain as a function of the moving mirror’s displacement.

The entire spectrum is therefore measured simultaneously in the interferogram and FTIR instruments offer considerable speed benefits compared to dispersive instruments which mechanically scan from one wave number to another. An interferogram can be obtained in only a few seconds compared to many minutes so allowing many interferograms to be collected in a comparatively short time. These are averaged allowing great improved signal to noise ratios and increased sensitivity. The absorbance spectrum of the sample is produced from the interferogram by Fourier-transformation; the process requires that the analyst also records an interferogram where no sample is present in the beam path. The ratio of the resulting absorbance spectra corresponds to the absorbance spectrum of the sample alone.

Key developments in technology were necessary for Fourier transform based instruments to become practical. Firstly, the development of Fast Fourier transformation (FFT) by Cooley and Tukey (Cooley and Tukey, 1965) offered an algorithm for Fourier transformation that was several orders of magnitude simpler that previous calculation routes whilst only slightly less accurate. Without FFT the lengthy computation would have required unfeasibly powerful and expensive computers. Even the much shorter computation underlying FFT however required the technology to wait for development of affordable desk-top computers to supply the necessary computing power before FTIR based instruments could become practical. In addition, not only does the computer handle the FFT allowing generation of spectra it also allows the analyst the ability to manipulate spectra for integration, base-line correction and averaging, and to analyse the spectra using multivariate approaches such as principal components analysis (PCA) and multivariate regression.
**4. The development of effective sample analysis techniques**

Samples of plant material are generally prepared by oven or freeze-drying to eliminate water which has high absorption in the IR; the dry samples are then finely ground. Generally, the finer that samples can be ground the better but often the analyst finds that for fibrous samples there is trade off between particle size and sample through-put and it is often necessary to determine empirically the maximum size of particle that permits robust analysis. Furthermore, the milled sample must be truly representative of the whole sample and chemical differences between particles of different sizes must be accounted for. It is a good idea to have access to a number of mills differing in design as each offers a combination of through-put and effectiveness that is highly sample specific. The preparation of samples by grinding is often the rate limiting step in IR analysis.
Traditionally, analysis by IR spectroscopy was highly time-consuming. In this approach small amounts of ground sample are ground with a larger quantity of an IR transparent salt such as KBr in a pestle and mortar. The mixture is then pressed hydraulically into transparent discs with between 5 - 10 tonnes of pressure (Kacurakova et al., 2000; Xu et al., 2007). At all stages it is essential that the amount of water in the disk is kept to a minimum and spectra must be taken quickly as the discs are hydroscopic and soon become opaque due to the absorption of water from the atmosphere. This process is time-consuming but allows spectra of a high quality to be obtained. It’s use has been reported by several groups engaged on the structural study of various cell wall components including pectin and hemicellulose; (Kacurakova et al., 2000; Sun and Tomkinson, 2002; Xu et al., 2006; Xu et al., 2007), cellulose (Liu et al., 2006), lignin (Gosselink et al., 2004; Sun et al., 2002; Sun et al., 2002; Xu et al., 2007), pyrolysis char (Hu et al., 2008) and lignin (Faix, 1991; Jung and Himmelsbach, 1989; Monties, 1989; Nakanishi and Kawakami, 1991; Stewart et al., 1997). Alternatively, samples can be ground with an oily mulling agent (typically Nujol) and smeared onto salt disks prior to spectral acquisition.

In recent years a much more rapid method of analysing powdered samples has become widely available and this has made the use of FTIR for larger sample numbers more practicable (Allison et al., 2009; Gosselink et al., 2004). The method relies on crushing a powdered sample against a flattened face of a trapezoid crystal of diamond, germanium or zinc selenide mounted in a rugged plate. The accessory is most often placed in the spectrophotometers sample bay and the IR beam is passed through the crystal by a series of mirrors. Because the IR beam hits the walls of the crystal at a very shallow angle it is reflected within the crystal and exits the far side where it passes to the detector. The technique is known as attenuated total reflectance (ATR). Most ATR accessories are of the single bounce design but for samples of only low absorption it is possible to purchase multi-bounce ATR accessories. Spectra are obtained from the sample only when they are placed in direct contact with the crystal and it is important that the particle size of the sample is sufficiently small and uniform to allow spectral acquisition to be reproducible. Because the IR beam only penetrates a few microns from the crystal surface it is possible to directly analyse aqueous samples. All methods require for spectra to be collected in a two step process; first, a blank spectrum is collected with no sample in place, next a sample spectrum is collected and the spectrum of the sample is obtained by calculating the ratio of the two measurements. Using ATR an operator using a manual spectrophotometer can comfortably collect spectra from approximately 200 samples comfortably in one day.

5. Spectral interpretation and multivariate analysis

The handling and interpretation of the IR spectral data is as much a part of compositional analysis as sample preparation and spectral acquisition and is discussed in this chapter in order for the reader to have a clear insight into the entire process. From the information provided in Section 2 it would appear relatively straight forward for the analyst to assign absorption peaks in the spectrum of a pure sample to particular chemical moieties or molecular structures. In practice, IR spectra are more complex than might be expected. Often additional absorption bands are present in the spectrum due to overtones or combinations of the fundamental vibrations. In addition, the spectra of larger molecules have an additional level complexity caused by the coupling of vibrations over part or the entire molecule. When two oscillating bonds share a common atom, the vibrations of the two bonds are coupled and as one bond contracts, the other bond contracts or expands, as in
asymmetrical and symmetrical stretching. When coupling occurs, the spectral position of bands due to particular bonds become shifted rather than superimposed as might be expected. These shifted vibrations are known as skeletal vibrations and give information on the entire molecule rather than on specific functional groups within the molecule. However, despite these complexities spectral interpretation can generally be achieved without resorting to first principles by remembering the principles underlying equation 7, and by reference to either a textbook on the subject e.g. Stuart (2005) or to one of the collections of IR spectral fingerprints e.g. Movasaghi et al. (2008).

Spectra from samples of complex mixtures such as plant cell wall preparations or ground samples of dried lignocellulosic crops, present an additional level of difficulty to interpret as it is next to impossible to deconvolute the superimposed spectra of the individual components in the sample. Prior knowledge of the likely composition of the sample often allows the identification of spectral features known to be associated with likely components e.g. in plant material the presence of bands at 1590 cm\(^{-1}\) and 1610 cm\(^{-1}\) are correlated with lignin (Monties, 1989), and if no knowledge of sample composition exists IR spectra provides clues as to the kind of compounds which may be present. The analyst can thus make an informed decision on the best methods to use for initial chemical analysis of the sample.

Whilst useful for qualitative analysis, these approaches are not sufficiently robust to allow the prediction of composition in spectra from large numbers of samples as they are too easily influenced by co-variance between chemical components in the samples and by random differences in the spectral data. The goal of using IR spectra to predict chemical composition in complex samples has been made possible by the adoption of multivariate approaches to data analysis and the availability of powerful and affordable desk top computers to handle the data processing and file storage. The simplest and possibly the most commonly used multivariate approach is PCA. This was first described in theoretical terms by Karl Pearson in 1901 (Pearson, 1901) and later made into a practical reality by Hotelling in 1933 (Hotelling, 1933). PCA allows the variance in the spectra data at hundreds of wave numbers to be condensed into a much smaller number of new principal components which explain most of the variance in the original data. Because there are fewer variables the data can be more easily explored graphically or statistically for correlation with experimental, environmental and sample effects. PCA works best with highly correlated data and spectral data is highly correlated. Any single data point in a spectrum is highly influenced by the value of its immediate neighbours. PCA is of great value for the biologist and analytical chemist hoping to understand how differences between samples in geographic location, growth year, species or some experimental treatment such as the addition of fertiliser, have an effect on chemical composition. This is illustrated by Figure 3A which shows a plot of the first and second principal components obtained from 194 mid-IR spectra of reed canary grass and switchgrass samples collected at the end of two consecutive growth years. The two components plotted in this figure represent the majority of the variance in the data set. The data points are coded according to species and it is quite apparent that spectral differences exist between the two grass species and these differences are accounted for primarily by the variance explained by principal component 2. The separation of the data points is however only partial showing that chemical differences between samples from the two species occur in only a proportion of cases. The analyst would usually wish to explore this further and establish whether this was the result of processing and sample preparation e.g. differences in sample moisture content, or whether
these differences reflected or factors influencing the growth and development of the plants during their growth in the field. Re-plotting these data but coding the data points by growth year would result in an identical distribution of data points but show no separation of the data according to growth year by these components (data not shown).

Fig. 3. (A) PCA of 194 mid-infrared attenuated total reflectance absorbance spectrums of samples of reed canary over the spectral range of 500-4000 cm\(^{-1}\) (□ = switchgrass, ● = reed canary grass). (B) Scatter plots showing measured vs. PLS model predicted values for nitrogen content in the training (●) and test (+) spectral data. (Data from Allison et al., 2009)

PCA is obviously a highly useful tool for evaluating differences and similarities between samples. In addition, the analyst may identify the wave numbers where these differences occur. This information is contained in the PCA loadings, a matrix of data produced during the PCA process that shows the influence of each of the original variables i.e. absorbance’s at groups of wave numbers, on each of the new components. Analysis of the loadings for principal component 2 would in this case identify the discriminatory wave numbers between the two species and perhaps make it possible to identify the class of chemical that
differed most between the two species. This very powerful way to look at chemical data will also reveal not just increases or decreases in given compounds but changes in the concentrations of several chemical components simultaneously. PCA and other methods of this kind e.g. canonical correlation, cluster analysis and multidimensional scaling are therefore excellent tools for exploring structure relationships within the spectral data but they cannot be used to determine the concentration of components within the sample mixture. For that it is necessary to employ a multivariate regression approach. Many approaches have been developed over the years e.g. see Otto (2007), and currently many analytical chemists rely on regression by partial least squares (PLS). Other methods such as principal components regression and multivariate linear regression offer subtle differences in capability but the way in which they are employed is similar. Multivariate regression requires the analyst to provide a training set of data from samples that have been also been analysed using standard chemical methods. The variance in these training data is arranged by PLS into new orthogonal components (latent variables) but unlike PCA these new latent variables not only capture variance but also achieve correlation with the analytical data. To ensure that the correlation is genuine and based on the chemical components being measured rather than to some selective fitting of noise the process is controlled by cross validation. The mathematics underlying these calculations is complex and many analytical scientists rely on software from specialist suppliers or instrument manufacturers, although it is also possible to find free-ware packages on the internet. The PLS regression is judged for predictive accuracy using an independent test set of spectral data from chemically analysed samples that were excluded from the PLS regression fitting process. This gives the user confidence that predictions made with the model will be accurate. Confidence in predictions made using the model is obtained by chemically analysing a small percentage of all samples and checking the agreement between the real and predicted values, and by monitoring the variance presented by new samples to ensure that they are adequately explained by the existing model. Figure 3B shows a plot of predicted vs. measured nitrogen content for the 194 samples of reed canary grass and switchgrass. Theoretical perfect fit is displayed as a dotted line in the figure.

6. Example of process

In this example using previously unpublished data we show how FTIR spectra can be used to develop predictive PLS models for several combustion parameters in samples of coal. This may seem to be a curious choice for illustration but the procedure by which the data are analysed and models developed are essentially identical to those employed for developing models to predict aspects of cell wall composition in less ancient biomass. The purpose of this study was to explore the possibility of using FTIR for the process monitoring of coal and biomass destined for co-firing. Two similar studies were published in 2009 which used NIRS (Kim et al., 2009) and FTIR (Geng et al., 2009) and these provide a useful comparison for these findings. Sixty nine samples of powdered coal were provided by E.ON. It was known that the coal samples were from a number of locations but details of their origins were not disclosed. The coal samples had been subjected to proximate analysis and values were supplied of the mass fractions of moisture, volatiles, ash and sulphur, see Table 1. With only minor exceptions the data were normally distributed. Triplicate IR spectra were taken from each sample from 600 cm\(^{-1}\) to 4000 cm\(^{-1}\) using a Golden Gate ATR accessory fitted with a single bounce diamond crystal (Specac, U.K.) placed in the sample compartment of an Equinox 55 FTIR spectrophotometer (Bruker Optik GmbH, Germany).
Each spectrum represented an average of 32 scans at a resolution of 4 cm\(^{-1}\). Spectra were averaged, smoothed and derivatised to the first Savitsky-Golay derivative using a window of 25 wave numbers and a second order polynomial fit before being normalised to unit variance. All spectral processing was performed using MatLab (version 7.10; Mathworks) and the PLS toolbox (version 6.0; Eigenvector Research, Inc.).

<table>
<thead>
<tr>
<th></th>
<th>Xa</th>
<th>Xm</th>
<th>Xs</th>
<th>Xv</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>10.81</td>
<td>11.93</td>
<td>1.124</td>
<td>31.01</td>
</tr>
<tr>
<td>Minimum</td>
<td>2.30</td>
<td>4.00</td>
<td>0.120</td>
<td>24.10</td>
</tr>
<tr>
<td>Maximum</td>
<td>41.10</td>
<td>25.90</td>
<td>2.390</td>
<td>35.60</td>
</tr>
<tr>
<td>SD</td>
<td>4.70</td>
<td>3.72</td>
<td>0.648</td>
<td>2.56</td>
</tr>
</tbody>
</table>

Table 1. Values, range and standard deviation (SD) of Q, calorific value (corrected for ash and moisture content); Xa, ash mass fraction; Xcl, chlorine mass fraction; Xm, moisture mass fraction; Xs, sulphur mass fraction and Xv, volatile mass fraction for 69 samples of powdered coal. Data supplied by E.ON

Figure 4A shows the complete set of 207 spectra taken from the 69 samples before averaging, derivatisation and normalisation. Derivatisation is a commonly employed practice in spectroscopy as it improves the separation of non-resolved peaks. In NIRS it is not unusual to derivatise spectra to the 3\(^{rd}\) or 4\(^{th}\) derivative. In FTIR however, the peaks are better resolved and derivatisation beyond the 1\(^{st}\) or 2\(^{nd}\) derivative typically offers no advantage. The spectra shown in Figure 4B are much more tightly grouped and it is apparent that this regime of pre-processing has improved the noise in the data considerably. Other processes which might have been applied include scatter correction, baseline correction offsets and general least squares regression which serves to decrease noise between chemically similar data points. The 69 spectra were divided into two groups; one group of 59 spectra that were used to develop the partial least squares regression model and an independent test set of 10 spectra which were excluded from the model but which were used to assess model predictive accuracy. The PLS modes were developed using SIMPLS algorithms (de Jong, 1993) and the PLS toolbox software. Models were cross validated using venetian blinds cross validation protocol (7 data splits) and developed to an optimal number of latent variables to ensure that the models were based on variance explaining the parameter of interest rather than on noise in the data. In each case the best number of latent variables to include was indicated by a minimal value of the root mean square error of cross validation (RMSECV). This measure of error can be interpreted as the standard deviation of the unexplained variance in the cross validated regression, and has the useful property of being in the same units as the response variable. This measure of error is a much better indication of model fit than root mean square error of calibration (RMSEC) as the latter does not indicate when the model is over-fitted to the data. The models were tested for predictive accuracy by measuring the root mean square error of prediction (RMSEP) using the independent data test set. For RMSECV, \(n\) relates to the number of spectra in the training data set, \(y_i\) is an observed value obtained by chemical analysis and \(\hat{y}_i\) a value predicted from the cross validated regression model. RMSEC can be calculated using the same equation by substituting the values of \(\hat{y}_i\) with the predicted values from the non-cross validated. Similarly, RMSEP is calculated by using values from the independent data test set.
Application of Fourier Transform Mid-Infrared Spectroscopy (FTIR) for Research into Biomass Feed-Stocks

\[ RMSECV = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \]  

Fig. 4. (A) Triplicate ATR-FTIR spectra (207 spectra) of 69 the samples of powdered coal. (B) First derivative averaged spectra after Savitsky-Golay smoothing and normalisation (69 spectra)

Plots of regression fits for these four models are shown in Figure 5 and the parameters of the four models in Table 2. All of the models are sufficiently good to be used for prediction of these parameters in unknown samples. Ideally, the variance in spectra from the unknown samples would be compared with those used to develop the model in question. This ensures that the model is capable of making an accurate prediction. Some of the samples detected as dissimilar would be analysed chemically and these data incorporated into a new model better able to make predictions. In all circumstances a proportion of the unknown samples, perhaps 5 – 10% of the total number should be analysed chemically to validate model predictions.
Fig. 5. PLS regression fits (red line) for the spectral data from the coal samples for A, moisture; B, volatiles; C, sulphur and D, ash mass fractions. Black and red symbols indicate samples in the training and test data sets respectively. The green line denotes a 1:1 regression fit.

Table 2. Summary of the PLS models developed to measure moisture (Xm), volatiles (Xv), sulphur (Xs) and ash (Xa) mass fractions. Details are shown of the number of latent variables in each model (LV), the percentage variation included from the spectral (% Var x) and non-spectral data (% Var y), predictive bias (Pred Bias) and model goodness of fit is shown for correlation, cross validation and prediction as root mean square error (RMSEC, RMSECV and RMSEP) and as determination coefficients ($R^2$, $R^2_{cv}$ and $R^2_p$).
7. Conclusions

Mid-IR spectrophotometry has become a potentially useful predictive tool for the plant breeder, biologist and engineer as it offers high rates of sample through-put, low unit cost and robust and accurate analysis. Key to this approach has been the development of modern Fourier transform spectrophotometers, the advent of modern personal computers and the availability of chemometrics software. Whilst many still prefer to use near-IR spectral analysis for the prediction of compositional parameters mid-IR gives the analyst additional information on the molecules within the sample which may be more difficult to discern using NIRS.

8. Acknowledgements

This work was funded by the Engineering and Physical Sciences Research council (EPSRC) and carried out as part of the Supergen Consortium in Biomass, Biofuels and Energy Crops (GR/S28204). The author wishes to thank Catherine Morris for making FTIR spectral analysis of the coal samples and Hugh Burnham-Slipper of E.ON for supplying the samples.

9. References


www.intechopen.com


New analytical strategies and techniques are necessary to meet requirements of modern technologies and new materials. In this sense, this book provides a thorough review of current analytical approaches, industrial practices, and strategies in Fourier transform application.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:
