## 1 Background

Vibrational spectroscopy is a method of probing sample molecular vibrations by subjecting them to light irradiation. Most biomolecules present a unique set of vibrations, which consequently produce an identifiable spectroscopic signature. Thus, the technique can be used to detect and quantify changes in sample biomolecular composition. Apart from specificity, vibrational spectroscopy is also very sensitive and can detect minute changes. Additional attributes, such as the noninvasive and nondestructive nature of analysis and amenability to *in vivo* application designs, render it ideal for use in biology and medicine. Applications of this phenomenon are myriad and widespread. It is beyond the scope of this Spotlight to review all biological/medical applications, and readers can refer to excellent reviews on the subject by Hanlon et al.,[1] Tu and Chang,[2] Pence and Mahadevan-Jansen,[3] Petry et al.,[4] Singh et al.,[5] Cialla-May et al.,[6] Zhao et al.,[7] Ahn et al.,[8] Krafft et al.,[9] and many others. One of the major focus areas is the development of diagnostic and treatment tools for the biomedical industry, which has a market with a 6.4% growth rate. Within this industry, vibrational spectroscopy is especially suited for disease screening, early diagnosis, and disease prevention, and caters to a market with a 7.3% growth rate.[8]

What mainly sets this technique apart from conventional screening/diagnostic methodologies is its ability to provide a complete biochemical fingerprint of the sample. Instead of detecting one or a few disease-associated factors, it gives information on the whole metabolome—that is, the overall change in biomolecules such as proteins, lipids, nucleic acids, carbohydrates, and some other specific molecules. This attribute is of particular advantage in complex diseases such as cancer, where malignancy-specific changes vary greatly. The ability to profile an entire sample's biochemistry also makes this a powerful tool for detecting changes from the normal, potentially signaling disease onset. However, this very feature complicates data analysis to a great extent. Multiple spectral signatures need to be compared across samples and give a single definitive output regarding the sample. In light of extreme in-group variations encountered in biological systems, separating healthy tissue from diseased, especially for early onset, can become very challenging.

Over the years, researchers have developed mathematical tools to tackle this problem, giving rise to the field of chemometrics.[10] A themed collection called "Chemometrics: Tutorials" in advanced data analysis methods produced by the *Analytical Methods* journal can be referred to for more insights on preprocessing[11] and multivariate analysis.[12,13] Gautam et al. have reviewed data processing in detail, specifically for vibrational spectroscopy.[14] The primary aim of the analysis is to weigh the importance of each peak in the spectrum with respect to all other spectra, decide which peaks have maximum variation, and reduce the data to take only the chosen peaks into consideration for the final output. It is clear that the analysis relies heavily on spectral variations, and thus it is very important to

remove all confounding variations from the input spectra before subjecting them to analysis. This operation is referred to as *preprocessing*. Common confounding factors corrected for are sample background, optical errors, sample fluorescence, charge-coupled device (CCD) response, and intensity variations. Optical errors and sample background are corrected by acquiring a spectrum without the sample and subtracting it from the sample spectrum. CCD response variation is offset by dividing the sample spectrum by response from a standardized material (such as a National Institute of Standards and Technology material). This step is important for between-machines comparison and need not be applied if all data originate from the same instrument. Fluorescence background may be removed by baseline correction or by spectrum-first derivatization. Baseline correction involves subtracting a polynomial from the sample spectrum and can be subjective. First derivatization is an objective method and reduces the regions of spectra, where any change in the $y$-axis values is gradual with respect to the $x$-axis values to near zero. Since fluorescence signals are broad in nature, fluorescence signals are accordingly reduced to near zero. In the case of sharp vibrational peaks, changes in $y$-axis values are very rapid with respect to the corresponding $x$-axis values, leading to nonzero values, isolating signals from background. Finally, intensity-related variations can be removed by normalization. Normalization helps analysis based only on peak variations—presence/absence/shape changes, rather than peak intensity.

The preprocessed spectra are subjected to multivariate analysis. As mentioned earlier, this selects important spectral variations and provides output based only on the chosen signatures. The most common multivariate analysis tool is principal component analysis (PCA). It is an unsupervised tool that does not take group information into consideration. Simply put, it does not matter whether input spectra are supplied with labels "normal," "abnormal," etc., or are unlabeled. PCA calculates the mean of all input spectra, calculates the variation of each spectrum from the average, and then ranks the varying spectral signatures with respect to prevalence.

Consider that wavenumbers 1200, 1450, and 1680 $cm^{-1}$ vary in every input spectra, 1340 and 1745 $cm^{-1}$ vary in 50% input spectra, and 1300 $cm^{-1}$ is variable in only 2% spectra; then PCA ranks the group 1200, 1450, and 1680 $cm^{-1}$ as most important, 1340 and 1745 $cm^{-1}$ as next, and 1300 $cm^{-1}$ as having little significance. The data are transformed to contain only these important wavenumbers and used to separate spectra on the basis of these wavenumbers. The PCA ranks are called principal components (PCs), and the results are presented as graphs of PC scores, where scores are values assigned to each spectrum depending on the extent of variation with respect to the chosen PC. Thus, the plot of PC1 versus PC2 will group spectra with respect to presence/absence/change in wavenumbers 1200, 1450, and 1680 $cm^{-1}$ and 1340 and 1745 $cm^{-1}$. PCA is usually considered a robust analytical tool, uninfluenced and unbiased. However, it is limited by the number of dimensions that can be plotted and visualized. For example, one cannot

visualize group separation in a plot of PC1 versus PC2 versus PC3 versus PC4, since only three plotting axes are available.

One way to circumvent this problem uses supervised analysis, which requires input in groups with labels. These methods work on principles similar to PCA, but there is a bias toward increasing intergroup separation and decreasing intragroup separation by selecting the best orientation in *n*-dimensions. The results are presented in the form of a confusion matrix; it displays the placement of each individual spectrum in a particular group. The spectrum may be correctly placed in the correct group (the same group as labeled) or in the wrong group. By the unique arrangement of the confusion matrix, all diagonal element spectra are correctly placed, whereas nondiagonal elements are incorrectly grouped. Such discriminant analyses can apply linear, quadratic, partial least square, or other equations to achieve group separations. They are prone to over-fitting, which can lead to erroneous conclusions. To reduce the chances of such errors, the spectra are further subjected to cross-validation. Cross-validation works by dividing the input spectra into training and test groups. The validity of the analysis is then checked based on the power of the subset training spectra group building a model that can correctly predict the test group spectra. The more spectra that are correctly predicted, the better the analysis. A commonly used cross-validation is the leave-one-out (LOO) method, whereby one spectrum is removed and the remaining input spectra are used to train a model, which is then used to predict the group of the one removed spectrum. This is repeated until all spectra have been left out once. The robustness of the model is determined by the number of spectra placed correctly in groups. Finally, for every type of analysis, one can use the built model to predict the group of the spectra whose group is unknown. This process is called *test prediction*.

There are several varied options available for preprocessing and analysis. Every method has its own advantage and disadvantage, and spectroscopists need to adapt the analysis routine that best suits the objectives of their studies. An analysis routine is the group of consecutive steps that are followed to analyze spectroscopy data. This understandably varies from lab to lab as well as experiment to experiment within a lab. There are several types of software available to preprocess spectra and perform multivariate analysis, such as LabSpec, Origin, OPUS, Minitab, Cytospec, and many others. They are excellent tools for the exploration and fixing of a routine. However, once a routine is established, it is time consuming to continue using multiple software for different steps. Programming platforms, such as MATLAB®, R, Python, SciLab, etc., allow scripting that can incorporate all steps and allow an analysis with a single click. Features such as turning on/off certain steps or multivariate analysis can be designed without forgoing the single-click option. Ultimately, the scripts may prove to be quickly exploratory as well as routine analysis tools, with the advantages of rapidity, accuracy, and ease of use. The Spotlight details the simple commands that can achieve this with instructions for using them to yield user-specific codes.