CHAPTER 1

Introduction

The subject matter of this book is a synthesis between chemometrics and metabolomics, both relatively recent scientific disciplines. This chapter describes the background to these disciplines and then introduces the background to the case studies which are used to illustrate the chemometric methods and describes some software packages that can be used to obtain results described in this text.

1.1 CHEMOMETRICS

The name chemometrics was first proposed by Svante Wold in 1972 in the context of spline fitting [1]. Together with Bruce Kowalski, they founded the International Chemometrics Society and the term slowly took off in the 1970s. However, the pioneers did not widely use this term for some years, but a major event that catalysed it was a workshop in Cosenza, Italy, in 1983 [2] where many of the early pioneers met. After this time several initiatives took off, including the main niche journals, *Journal of Chemometrics* (Wiley) [3] and *Chemometrics and Intelligent Laboratory Systems* (Elsevier) [4], together with courses and the first textbooks [5, 6] with regular reviews and ACS (American Chemical Society) symposia starting a few years earlier [7].

However, these events primarily concern name recognition and organisation, and the main seeds for the subject were sown many years earlier.

Applied statistics was one of the main influences on chemometrics, although the two approaches have diverged in recent years. The modern framework for applied statistics was developed in the early 20th century and we still use terminology first defined during these decades. Before that, early academic statistics was mainly mathematical and theoretical, often linked to probability theory, game theory, statistical mechanics, distributions etc. and viewed as a subdiscipline of mathematics. Although many early pioneers had already used approaches previously that we would now regard as the

Data Analysis and Chemometrics for Metabolomics, First Edition. Richard G. Brereton.

Companion website: www.wiley.com/go/Brereton/ChemometricsforMetabolomics

forerunners of modern applied statistics, their ideas were not well incorporated into mainstream thinking until the early 20th century.

A problem in the 19th century was partly the division of academic disciplines. Would a mathematician talk to a biologist? They worked in separate institutes and had separate libraries and training. For applied statistics to develop, less insular thinking was required. There also needed to be some level of non-academic contribution as many of the catalysts were at the time linked to industrial, agricultural and medical problems. With core academic disciplines, the application of statistical methods in physics and chemistry, which would eventually progress to quantum mechanics and statistical mechanics, fell outside mainstream applied statistics and has led to specialist statistically-based methods that are largely unrelated to chemometrics.

However, in the first three decades of the 20th century, there was a revolution in thinking. Such changes primarily involved formalising ideas that had been less well established over the previous decades and even centuries. Karl Pearson [8] and William Gossett publishing under the pseudonym 'Student' [9] are recognised as two of the early pioneers. Pearson set up the first statistics department in the world, based in London, and his 1900 paper first introduced the idea of a *p* value, although historic predecessors can be traced several centuries back [10–12].

It was not until after the First World War that applied statistical methods were properly formalised in their modern incarnation. Ronald Fisher was possibly the most important figure in developing a modern framework for statistical methodology that many people still use today. In 1925 he published *Statistical Methods for Research Workers* [13] and established the concepts of *p* values, significance tests and ANOVA (analysis of variance). Ten years later he wrote a book that described the basis of almost all statistical experimental designs [14] used even now, and his paper on classification of irises (the plants) [15] is an essential introduction to multivariate classification techniques, with this dataset used even now for demonstrating and comparing new approaches. Other important workers over that period, included Harold Hotelling, who among others was attributed with progressing the widespread use and recognition of PCA (principal components analysis) [16, 17] and Jerzy Neymar and Ergon Pearson who developed alternative approaches to hypothesis tests to those proposed by Fisher [18].

During the interwar period, many of the cornerstones of modern applied statistics were developed, and we continue to use methods first introduced during this era; many approaches used in chemometrics have a hundred-year vintage. However, there were some significant differences from modern practice. There was no capacity to perform intensive computations or generate large quantities of analytical data, so applications were more limited. Agriculture was at the forefront. During this era, the old landowning classes had to modernise to survive: many farm labourers left for the cities and agriculture became more automated. The relationship between landowners and tenants weakened and larger farms were viewed more as an industry rather than the birthright of aristocratic classes. This required a significant change in production, and agricultural statistics was very important, especially to improve the economies of Western Nations. Other important driving forces came from the use of psychology to interpret test scores, and from economics. Common to all these types of data is that experiments involved considerable investment in time, so it was reasonable to spend substantial effort analysing the results, some required weeks of manual calculations, as

data was expensive and precious. In modern days, spectra, in contrast, can be obtained relatively rapidly and quickly, so spending days or weeks performing statistical calculations would be an unbalanced use of resources.

Furthermore, without the aid of computers many of the multivariate methods we now take for granted would involve a large amount of time. Salsburg [19] claims as follows: 'To get some idea of the physical effort involved, consider Table VII that appears on page 123 of *Studies in Crop Variation. I.* [20]. If it took about one minute to complete a single large-digit multiplication, I estimate that Fisher needed about 185 hours of work to generate that table. There are fifteen tables of similar complexity and four large complicated graphs in the article. In terms of physical labor alone, it must have taken at least eight months of 12-hour days to prepare the tables for this article.' Of course, Fisher would have had many assistants to perform calculations, and he would have been very well resourced compared to most workers of the time. Hence, only quite limited statistical studies could be performed routinely. Some algorithms and designs such as Yates' algorithm [21] were developed with simplicity of calculation in mind as the data had special mathematical properties and although still reported in some textbooks even now are not so crucial to know about with the advent of modern computing power. Computers can invert large matrices very quickly, whereas a similar calculation might take days or longer using manual methods. In areas such as quantum chemistry, a calculation that may take up an entire PhD via manual calculations can now be done in seconds or less using modern computing.

The statistician of the first half of the 20th century would be armed with logarithm tables, calculators, slide rules and special types of graph paper, and in many cases would tackle less data-rich problems than nowadays. However, there was a gap between the mathematical literature where quite sophisticated methods could be described, often in intensely theoretical language, and the practical applications of much more limited and in most cases simpler approaches. Many of the more elaborate methods of those early days would not have much widespread practical use, but modern-day multivariate statistics can now take advantage of them. The chemometrician can routinely use methods that on very large spectroscopic or chromatographic datasets that were inconceivable prior to the widespread availability of modern computers.

In the post-war years, chemical manufacturing was of increased importance and multivariate methods were applied by industrial chemical engineers [22]. G.E.P. Box worked with a group in the chemical company ICI in the UK for some years, before moving to the US. His text [23] written together with two co-authors, is considered a classic in modern statistical thinking for applied scientists emphasising experimental design and regression modelling and brings the work of the early 20th century into the modern era.

In the 1970s, mainstream applied statistics started to diverge from chemometrics. In chemometrics, we often come across short fat datasets, where the number of variables may far exceed the number of samples. For example, we may record thousands of mass spectral or NMR or chromatographic data points for each of perhaps 20–100 samples. These sorts of problems were not conceivable to the original statistical pioneers, measurements were expensive, so variables were scarce. Fisher's classic iris data [15] consisted of 150 samples but only four variables. Once sample sizes are less than the number of variables, some classic approaches for multivariate data analysis are no longer directly applicable, an important one is the Mahalanobis distance [24]

and the corresponding method of LDA (linear discriminant analysis) [15] which have to be adapted and cannot be directly applied to data whose variable to sample ratio exceeds 1. More recent methods such as PLS (partial least squares) [25, 26] and SIMCA (Soft Independent Modelling of Class Analogy) [27] were advocated in the 1970s and 1980s very much with the needs of chemometricians in mind and coped well with such data. Chemometricians are often very interested in variables, for example, which are most significant out of possibly hundreds of candidates, whereas statisticians emphasise the significance of factors, in many cases using univariate tests. Although both types of thinking may come to similar types of conclusions, some traditional statistical approaches are invalid, as an example if there are more variables to samples, we cannot use LDA to provide us information as to which variables are the most significant, whereas PLS might provide an answer. As such problems were inconceivable before the 1970s, the impetus to providing niche solutions for the chemometrician is of 50-year vintage. Metabolomics is often a rich source of information where the number of variables far exceeds the number of samples so chemometrics is needed for statistical interpretation.

Although there are still a few workers in the chemometrics field who identify themselves as statisticians, their influence became quite limited after the 1980s, judging by attendance at chemometrics meetings, development of texts and courses and so on. In a way, this divergence is similar to those in areas such as quantum mechanics or statistical mechanics, which are also founded on statistical principles but are primarily led by numerate scientists.

However, the strong statistical parentage is an important cornerstone of chemometrics. As many applications are moving away from the original ones of quantitative analytical and physical chemistry into more hypothesis-based science such as metabolomics, the original aims of measuring more precisely or predicting more accurately are gradually being supplemented by more statistical aims to generate and test hypotheses. In the latter case, chemometrics, whilst once routed in the physical sciences, is being adapted to tackle problems from those in biology, geology, psychology and so on, and requires a return to core statistical thinking. It is questioned for the future whether this will attract more applied statisticians back into the field, or whether niche chemometrics experts will return to the mainstream statistical literature and then incorporate more statistical thinking into their publications and software without the need to bring mainstream statisticians directly into their collaborations.

Another parent of chemometrics was quantitative chemistry. Historically, it is important to understand that interdisciplinary research was not very widespread during most of the formative years in the 20th century. Academia was highly compartmentalised, with students opting for specialist courses in, for example, chemistry, or mathematics or biology. Most university teachings except at the basic levels would have been by staff from a single department. Students might be introduced to statistical concepts at an early stage but if by mathematicians in a somewhat abstract manner. If carried forward in courses such as physical and analytical chemistry, statistical methods would have been strongly oriented towards univariate measurement and estimation or in specialist areas such as quantum chemistry.

At a research level, departments would have their own libraries and staff rooms. In many countries, to obtain academic positions staff would have to be very focused.

An analytical chemist might have to pass a committee for tenure or habilitation, which is narrowly defined in part because subsequent teaching duties will also be highly compartmentalised. Libraries, journals, books, staff rooms, even sports teams and social events would often be organised by departments. A chemist, if working in the right subdiscipline, could snatch some snippets of statistical thinking where needed, but it would mainly be developed independently within their own environment. An analytical chemist would be unlikely to visit a mathematics library or read a mainstream statistics journal, but would gather their statistical knowledge from analytical journals and books; the internet was not yet available so the ability to search for papers from a wide selection of journals was mainly restricted to visits to other departments' libraries. Analytical chemists would be introduced to niche statistics, for example, learning about precision and accuracy of measurements, but most would have limited exposure to other statistical texts except at a basic level.

Industrial cross-over with mainstream academics was also rather limited until the fourth quarter of the 20th century. Thus, the work, for example, G.E.P. Box and colleagues were doing in ICI in the 1950s would be unknown to chemists in many prestigious universities. Formal statistical design of experiments, so important for improved industrial productivity, would not be adopted by mainstream synthetic chemists in an academic environment until the last decade of the 20th century.

Hence, the development of statistical thinking within mainstream chemistry, such as multivariate analysis and statistical experimental design, developed in quite unique ways. A few brave workers did, however try to introduce statistical and computational ideas to the chemistry community. Statistical methods such as univariate linear regression, determining accuracy and precision of measurements and so on, have been part of the analytical and physical chemists' toolbox for more than a century. However, many techniques now recognised as part of chemometrics were shown to be recognised within the mainstream chemistry community. In 1949, Mandel draws together a number of statistical techniques, including design of experiments and ANOVA [28], in a paper which has been cited just seven times at time of writing (December 2023), an almost forgotten paper. W.J. Youden, a pioneering statistician, wrote 37 articles for *Industrial and Engineering Chemistry* between 1950 and 1957, many on experimental designs and very relevant to analytical chemists and chemical engineers, which in turn have been cited only 60 times in total. In 1952, he published a review in the journal *Analytical Chemistry* [29] citing 154 references, which has, in turn, only received seven citations. Yet a paper in the journal *Cancer* by the same author published in 1950 [30] has received over 7000 citations at the time of writing. It would be a value judgement as to whether these different publications contained more in-depth or original information, and the difference in reception would primarily relate to the difference in readership.

Hence, although analytical chemists were aware of traditional statistical methods, for example, how to fit a straight line or how to determine the 95% confidence in a mean, they were relatively uninterested at the time in the statistical revolution of Fisher and colleagues, and a traditional course in chemistry would not cover systematic experimental design, ANOVA, multilinear regression, multivariate pattern recognition, etc. It took until the 1970s before the importance of these approaches, well established 50 years ago in the mainstream statistical literature, started to slowly become recognised

within the chemistry community. Chemometrics by name was fundamentally born within the chemical sciences, where the first applications of for example chromatography and spectroscopy for the analysis of complex mixtures were developed. As time has progressed, although the first uses of these instruments were within the analytical and physical community, their application to data-rich sciences such as in the biomedical sciences where large mixtures of chemical compounds have a metabolic significance, has led to a much wider applicability of these instrumental techniques, and so the concepts from chemometrics. The original applications of NMR and MS in chemistry were primarily for the structural elucidation of individual molecules and it took several decades before they became established for the studies of mixtures.

It was not until the 1970s that analytical chemists started widely recognising the potential of statistical principles for experimental design. Stan Deming, whose first paper was published in 1967, published a well-cited paper in analytical chemistry on simplex optimisation in 1973 [31] followed up by a more comprehensive statistically based and well-regarded book in 1987 [32].

A parallel development happened within the physical chemistry community. Spectroscopists studied problems involving overlapping peaks of mixtures. They probably did not have much access to the statistical literature at the time, but did read the physical literature for example about eigenanalysis, which is related to PCA. There were several papers in the 1960s of which two are cited [33, 34]. Physical/analytical chemists of the time developed their methods in isolation to statisticians and, due to the limited availability of computers, applied their approaches to what we now regard as quite simple problems.

The prolific pioneer Ed Malinowski put together many of the first approaches to multivariate resolution of mixtures in the 1970s with statistical and algorithmic descriptions, culminating in his classic book published in 1980 [35]. Many of the methods now recognised as multivariate curve resolution that have a high profile in chemometrics, emerge from Malinowski's early work. Unlike most chemometrics methods, for example for classification or exploratory data analysis or regression, factor analysis (as defined by Malinowski) or multivariate curve resolution have a very specific role in chemometrics, so play a unique role; even now many approaches for resolving peaks in coupled chromatography incorporated in elaborate software have their origin in concepts first advocated by Malinowski.

By the 1980s, ideas from quantitative analytical and physical chemistry had started to become formalised both in the area of multivariate analysis and experimental design, and were slowly recognised initially by a rather small group primarily of analytical chemists. In the 2000s, basic chemometric concepts were starting to be introduced in general analytical chemistry courses and books and rapidly developed into software that was essential for burgeoning applications of instrumental analysis such as metabolomics as discussed in Section 1.4.

The third catalyst was scientific computing. Many of the approaches described in the first half of the 20th century remained theoretical without the availability of good computer power, and prior to the 1970s, only really quite simple problems could be solved by chemometric methods.

In the 1950s, very few scientists and engineers had access to computers for daily calculations. Computers were large and expensive institutional machines often developed

for defence (as the initiative originally came out of the Second World War). Individual researchers rarely had direct access, programs had to be sent in using punch cards or paper tape to trained operators, who would then feed the instructions to a mainframe and output would usually be in the form of printer paper sent back to the user. If there was a mistake in a program, this was very costly.

Early programs were in assembly language or machine code, which would be very hard for non-experts and take time to develop simple sets of instructions. In the 1950s, IBM developed a high-level language, called Fortran (Formula Translation) [36], which was the basis of most scientific software for the next two decades. The language continues to be developed with regular updates. Most classical scientific software was written using Fortran. The vast majority of quantum chemistry packages are still written in Fortran, and some build on subroutines over 50 years old. The 1960s was a particularly fruitful time for quantum mechanics computing. Prior to this era, it could take a graduate student almost an entire PhD just to calculate a few quantum mechanical integrals, so the use of scientific computers revolutionised this field.

However, access to mainframes capable of running large scientific programs was very limited until the late 1960s and early 1970s. Scientists who were not viewed as hard-core physical scientists had limited possibilities, and it took some years before access to significant scientific computer power was broadened.

In the very late 1960s, a few chemists did get access to significant computing power. Peter Jurs and his co-worker Bruce Kowalski were some of the first writing a series of papers in the analytical literature [37]. The impetus had been from Djerassi and coworkers who had pioneered the use of computers in structural elucidation [38] leading to the field of artificial intelligence. The Arthur program [39], developed for both mainframes and VAX minicomputers was one of the first widespread chemometrics packages but still one had to be very expert to install and use it.

In the 1980s, two important developments happened in computing that would move chemometrics from a specialist discipline to one that had the potential to be more widespread. The first and most important was the development of microcomputers in the late 1970s to early 1980s, the most successful scientific models of the type based on the original IBM PC [40]. Once micros became widespread and relatively user-friendly, chemometrics software could extend to a far wider user base and allow laboratory-based scientists rather than primarily specialists with access to expensive communal mainframes, to access chemometrics software.

Another important development of the time was MATLAB, originally developed by Cleve Moler [41] in 1981. Most chemometricians like to think in terms of matrices, and languages such as Fortran and BASIC were not naturally oriented towards matrix operations at the time. MATLAB, however, allows the programmer to develop code using matrix and vector algebra directly and contains fast algorithms for operations such as inverting large matrices. Operations such as PCA are also built-in. Since its early days, the software has been substantially expanded with GUIs (graphic user interfaces) and extensive graphics. MATLAB had a significant role in the development of chemometrics because workers could quickly develop matrix-oriented algorithms and often swapped code. Although many developing chemometrics methods now prefer R or Python, there still is an important group of MATLAB users, and this environment strongly catalysed the fundamental developments from the 1980s onwards.

The three catalysts, namely applied statistics, analytical/physical chemistry and scientific computing, converged in the 1980s to provide a fertile environment for the establishment of the discipline. Several events were responsible to create a specific launching pad for what is now well regarded as a coherent body of knowledge.

The NATO-sponsored meeting in Cosenza, Italy, in 1983 [2] brought together many of the experts working in this field at the time, not all would have regarded themselves as fundamental chemometricians until then. The journal *Analytical Chemistry* started publishing biennial fundamental reviews under the name 'chemometrics' as from 1980 [7] bringing together the last two years of papers in the field. Several regular workshops were established. The first comprehensive texts were published [5, 6] although some more specialist books had been published covering specific areas, but without the name 'chemometrics' in the title, in the years before. Two journals were established, published by Wiley [3] and Elsevier [4]. Several series of conferences were started at this period. These attracted mainly niche workers, most of whom had a back in computing within a chemistry environment and were not yet attracting biologists. There was a separate and very well-established area of biological statistics, mainly oriented towards univariate data.

The applications at this phase were fairly simple, with NIR (near infrared) calibration and HPLC (high-performance liquid chromatography) deconvolution predominating, and relatively small sample sizes. Some approaches such as PLS [26, 27] and SIMCA (self independent modelling of class analogy) [27] as well as Malinowski's extensive methods which he called factor analysis [35] were very oriented towards and widely reported within the chemometrics community rather than using a general statistics toolbox, although PLS (originating in economics) has found a home more generally since. The emphasis at the time was primarily to be able to measure and estimate accurately, often in areas such as pharmaceutical and food science, where the concentration of an ingredient or of a reactant had to be estimated by spectroscopic or chromatographic means accurately and quickly. Pattern recognition that has a high profile in modern chemometrics and metabolomics was not a very large part of the original literature. Many dedicated packages were developed over this period, most of which are still in existence now, these will be described in the section on software (1.4) below.

Hence, most of the tools now recognised as chemometrics were available in the 1980s. However, the early promise did not materialise at the time, unlike many other data-rich areas such as bioinformatics and QSAR (quantitative structure–activity relationship) which had a similar vintage. As from the mid-1990s, there was a tremendous interest in the application of chemometrics to fundamental analytical chemistry, for example the resolution of overlapping peaks in HPLC, which generated a large number of technically sophisticated but in most cases not particularly widespread papers and a very introspective number of groups and conferences. There was less scope for huge innovation compared to the 1970s when scientific computing was rapidly evolving and multivariate statistical methods were quite new to numerate chemists. In 2008, Paul Geladi and Phil Hopke ask 'Is there a future for chemometrics?' [42]. Many viewed the subject as rather a technical niche area without much general applicability. The number of specialist chemometrics groups in the world had hardly changed over the decades, just with a few new faces replacing those that had retired or left for other fields. It was not a particularly attractive area for researchers, without much funds, and appeared to flatten off.

In contrast, during the last 15 years or so, there has been a substantial renaissance. This is because chemometrics has moved out of its original comfort zone of quantitative analytical and physical chemistry and been embraced in areas such as metabolomics, among others. Big datasets are now readily available using instruments, such as LCMS (liquid chromatography mass spectrometry), NMR (nuclear magnetic resonance), GCMS (gas chromatography mass spectrometry) and so on. The capability of instruments to analyse many samples, each in turn containing large quantities of information, means a new and urgent need to correctly interpret these immense datasets and also to understand how to correctly design the experiments and perform the sampling. Thus, over more than a decade, there has been a renewed urgent need for chemometrics expertise. The sort of problems tackled nowadays often differs from the traditional analytical chemistry problems. The latter often involved measuring more accurately. For example, can a spectroscopic technique estimate the concentration of an analyte in a mixture without chromatography, and if so how well? We might create some reference standards or perform some independent and slower method of analysis, such as HPLC, and use multivariate analysis of a series of mixtures to create a calibration model. However, in many metabolomics problems, we are not certain of the answer in advance. For example, we may obtain the LCMS of donors' serum with and without a disease. How certain there is enough information in the serum to distinguish each group? It may depend on whether the disease was correctly diagnosed and how far it has progressed. There will be confounding factors such as age or diet or genetics – how representative were the samples? And then if we think we can separate groups, which metabolites are most likely to be markers for the disease? How confident are we? In such a situation we do not know the answer in advance and are generating and testing hypotheses. In fact, most of science outside the core physical sciences is primarily based on hypothesis testing. Much of early chemometrics was developed by programmers good at algorithms and matrices, whereas the needs of biological scientists are more hypothesis testing. This text aligns chemometrics methods primarily from the point of view of hypothesis formulation, which communicates more closely with the language of clinicians and biologists.

Chemometrics has had a renaissance because its methods are being applied to many scientific problems outside core quantitative chemistry. Nevertheless, most of the original methods, first pioneered over 50 years ago, are still relevant, and techniques such as molecular spectroscopy or chromatography, once the domain of chemists, are now widely used in many scientific fields. Understanding the fundamental statistical basis of chemometrics is an essential aid to safely and usefully employing these techniques, which have become widely available due to a plethora of software and datasets.

1.2 METABOLOMICS

The central dogma of biology is illustrated in Figure 1.1. In its simplest form DNA makes RNA, which makes proteins which make metabolites. The metabolic profile influences phenotype and so the characteristics of all organisms.

The study of systems biology is to connect these steps.

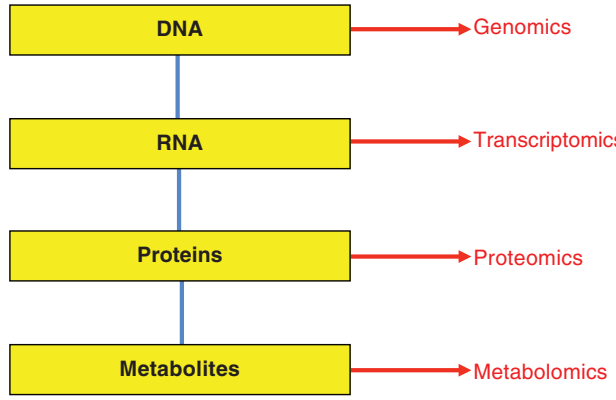

FIGURE 1.1 Central dogma of biology.

The concept of a genotype is the oldest and was first defined early in the 20th century by a number of papers by Johannsen [43]. The concept of a genome was first described by Winkler in 1920 [44]. There is in fact no universally agreed definition of these two terms, despite widespread usage; however, a common distinction is as follows. A genome is an organism's complete DNA, which includes all of its genes (coding/non-coding) and intergenic regions. The genotype refers to the genetic information for a particular trait. In humans, the genome, for example, includes the portion of human DNA that codes for hair colour and decides the genotype for that trait.

Genomics therefore studies the entirety of an organism's DNA. With improvements in high-throughput sequences, the first whole genomes were sequenced in 1980s and 1990s. The first complete genome sequence of a eukaryotic organelle, the human mitochondrion, was reported in 1981 [45] and the first chloroplast genomes followed in 1986 [46]. In 1992, the first eukaryotic chromosome, chromosome III of brewer's yeast *Saccharomyces cerevisiae,* was sequenced [47]. The first free-living organism to be sequenced was that of *Haemophilus influenzae* in 1995 [48]. It was however the human genome project that catalysed this scientific discipline with the complete sequencing announced in 2003 [49], although a small number of sequences still remained.

Whereas the concept of a genome was well developed, the name genomics and the concept of this as a discipline is rumoured to have been proposed in 1986, with a birthplace in the journal *Genomics* in 1987 [50]. This was the first recognised omics discipline and great-grandparent of metabolomics. With genomics arrived a large amount of data and this catalysed the arrival of computationally intense disciplines such as bioinformatics to mine these large datasets, placing computing at the centre of modern biological research.

Next up was transcriptomics, this time concerned with RNA. The earliest known use of the noun transcriptome is in the 1990s. The earliest known use of this term in the scientific literature is from 1997 [51], which was the first key work to report the transcriptome of an organism describing 60,633 transcripts expressed in *S. cerevisiae* using serial analysis of gene expression. With the rise of high-throughput technologies and bioinformatics and the subsequent increased computational

power, it became increasingly efficient and easy to characterise and analyse large amounts of data.

Third up, but with an earlier start, is proteomics. The first studies of proteins that might now be regarded as proteomics began in 1975, after the introduction of the two-dimensional gel and mapping of the proteins from the bacterium *Escherichia coli*. However, the first formal mention of the idea of a proteome was in 1995 [52]. The concept of proteomics followed soon after [53]. There was a rapid increase in publications in this area in the late 1990s culminating in the journal *Proteomics*, founded in 2001 [54], which published 156 papers in its first year, an unusual volume for a new journal. Several other journals were founded and named with the word proteomics in the title subsequently.

With the growth in proteomics came a growth in instrumental methods to characterise proteins. Important to readers of this book is the development of mass spectrometric methods which are the most widespread currently used. Most studies now involve some form of LCMS/MS. There are a large number of approaches for the collection and interpretation of such data for proteomics, but the result is usually a table in which the identified proteins and their relative concentrations or intensities are listed. This table can then be subjected to statistical methods, in a very similar way to metabolomics, although the methods for dealing with the raw instrumental data can be quite complex. But there is a lot of similarity in the protocol for interpreting and analysing proteomic data as for metabolomic data, except that the raw datasets tend to be more complex. In proteomics, good instrumental analysis, data analysis, statistical interpretation and computing are integrated and have been from the early days.

Stepping into this stable is metabolomics. The chain DNA to RNA to protein has the main outcome of creating metabolites that result in so-called phenotypic expression, so metabolites are the end product of this sequence of events. For many years, metabolism was the province of biochemistry and bio-organic chemistry and somewhat divorced from mainstream biology. The original omics techniques were developed with nucleic acids and proteins in mind. Chemists focused on small molecules by isolating and synthesising individual metabolites rather than studying entire metabolic profiles.

With the advent of coupled chromatography; however, it now became possible to detect and characterise whole metabolic profiles by GCMS [55] and LCMS [56]. NMR was also demonstrated as a technique to analyse whole metabolic profiles [57] although it took some years to improve sensitivity to develop this as a widespread technique in metabolomic studies. However, with advances in instrumentation, all three techniques are now essential workhorses for metabolomics.

The first published use of the concept of a metabolome was in 1998 [58, 59]. The metabolome refers to the complete set of small molecule (<1.5 kDa) metabolites found within a biological sample. The first recorded use of the derivative term metabolomics in the formal scientific literature is from 2003 [60]. The journal *Metabolomics* was established in 2005 [61]. Many of the early pioneers from the 1990s are still in leading positions within the discipline, unlike genomics or proteomics where the leaders have mainly passed the torch onto future generations. A large number of metabolome databases have since been developed, of which the human metabolome database is the most prominent [62]. These are distinct from metabolite databases, which are much

larger, but less specific, some of which are described in Chapter 2. The aim of the databases is primarily the identification and characterisation of metabolites, which then can be used for statistical analysis.

Public domain datasets consisting primarily of analytical LCMS, GCMS and NMR data have been uploaded in various formats to a number of large and well-maintained databases, such as the Metabolomics Workbench [63] and Metabolights [64]. These are quite distinct from the metabolite databases described above that characterise analytical data such as spectra of individual metabolites rather than whole profiles.

It is common the define the metabolomics workflow as illustrated in Figure 1.2. There are variants in the literature, but this term started gaining currency in the late 2000s and is now widely employed to illustrate the various steps in a typical analysis.

Steps 1 and 2 involve sample collection and preparation (usually by extraction). In some descriptions, these are combined into a single stage. The chemometrics expert is sometimes involved in step 1, using experimental design to choose experimental conditions or sampling designs to choose which samples to analyse. In large clinical sampling projects that may take several years; however, the chemometrics expert may have less initial input although he or she may be able to select samples for pattern

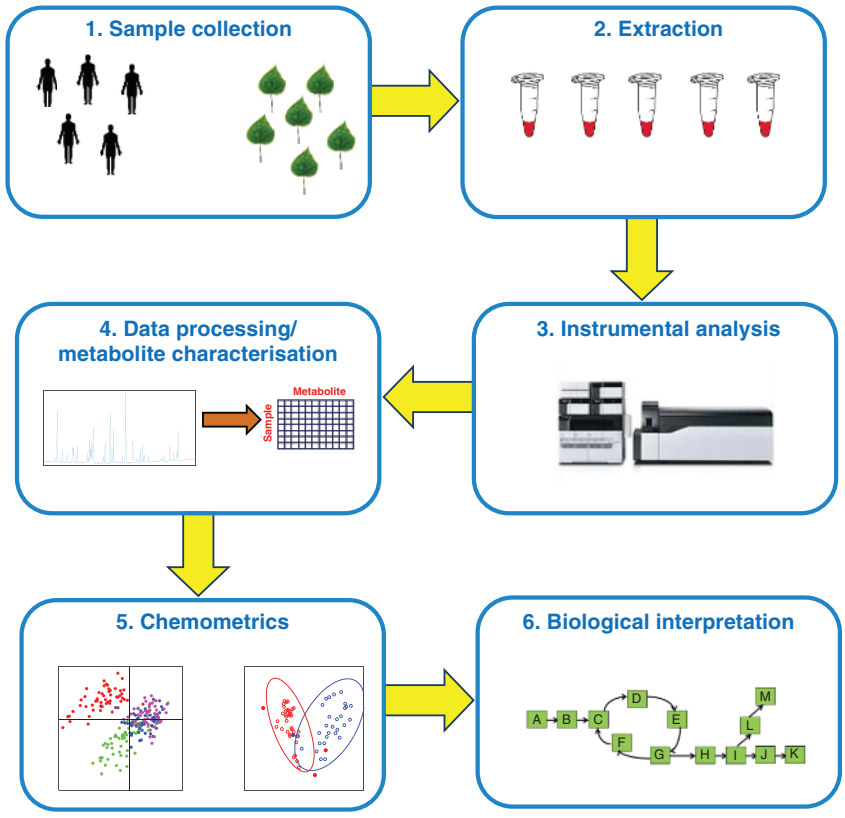

FIGURE 1.2 Metabolomics work flow.

recognition often several years after initial collection, from a larger set of samples as discussed in Chapter 6. In smaller laboratory-based studies, it is usually good practice that the expert who analyses the data also advises on the experimental design where feasible and is therefore involved in planning the work from the beginning, rather than after the data becomes available.

Step 3 is normally the experimental domain, often laboratory based involving analytical chemists. However, chemometrics experts can have a role, for example, in advising on instrumental stability, quality control, quantification, number of runs and so on.

Step 4 is sometimes combined either with step 3 and viewed as an essential part of processing raw spectroscopic or chromatographic data, or as part of step 5, the chemometrics. This book will not cover this stage in detail. As an example, there are an ever-increasing and evolving number of software-based approaches for processing and annotating (or identifying metabolites in) LCMS datasets; these are often very technical and bespoke sometimes with only some steps in the public domain and will change over the period of the lifetime of this text. A book that covered these would be as long as this text and would become rapidly out of date. However, the main principles and, where appropriate, webpages for some of the more widespread packages at time of writing, are described in Chapter 2. Some laboratories develop their own in-house approaches, others use free public domain packages, and others use commercial products often oriented towards specific instruments. The outcome is usually a matrix, often a peak table, which can then be used for subsequent statistical analysis. In some cases, raw spectroscopic data such as NMR frequency domain intensities can alternatively be used without initially identifying metabolites, so stage 4 is rather straightforward. However, most users of both coupled chromatography and NMR would prepare a peak table using a variety of software. In the case of NMR the columns of the data matrix might consist of the concentrations of identified metabolites corresponding to several spectroscopic peaks in different regions of the spectrum.

Step 5 involves chemometrics and statistical data analysis or pattern recognition. This key step is the main focus of this text, although we place it in context in Chapters 1–3. Most metabolomics data is very complex and often cannot be successfully interpreted without statistical and computational approaches and hence chemometrics is integrated as an essential step in almost all metabolomics studies. This book is restricted to describing statistical techniques although there is some trend towards machine learning. Although the latter is very powerful and takes advantage of the ready availability of modern computing, most metabolomics data, whereas it is substantially more complex compared to information obtained from single metabolite studies, its complexity is much lower than encountered in problems that require typical bioinformatics data mining, as example in human genomics. Hence, currently and probably for the foreseeable future, chemometric and statistical techniques are likely to be of the correct sophistication for metabolomic studies. The end product of this step often results in proposed biomarkers, which can be investigated further.

Step 6 involves biological interpretation, usually in the form of pathways, in favourable cases leading to clinical or toxicological explanations. Some metabolomic studies might stop at step 5, for example, just to identify potential biomarkers or classify samples. In this book, we will not extend our discussions to the interpretation of

our results, which inevitably must be done on a case-by-case basis according to the individual aim of the experimentation.

Although the most widespread applications of metabolomics involve the use of NMR or coupled chromatography for biomarker exploration, not everyone who would describe their work as metabolomics uses these techniques, and it is anticipated that there will be some readers with other viewpoints. Single wavelength HPLC, Raman and FTIR can be used to study whole metabolic profiles either from extracts or intact. Some studies do not involve biomarker discovery but may involve just classification using metabolic profiles or calibration to a biological property. Whereas the focus of this book will be on biomarker discovery, we will also discuss other applications, which some will view as part of metabolomics. The overriding theme is studying whole metabolic profiles using spectroscopic or chromatographic techniques in order to relate these to biological properties.

The term metabonomics is sometimes used as an alternate [65]; in practice, there is no well-accepted difference with the term metabolomics and we will not distinguish between these in this book. Metabonomics is more commonly used in NMR in a clinical context but there is no exclusive use of this term. Lipidomics, is also another term some readers may have come across and can be considered a subfield of metabolomics [66]. It is the study of the lipidome, that is, the total lipid content within a cell, organ or biological system. Lipids are often simply defined as hydrophobic biological substances generally soluble in organic solvents. We will not distinguish these fields in this text, as similar statistical methods are required.

There have been several general articles about the application of chemometrics to metabolomics of which we cite some early well-known ones [67–71]. Review articles within specific application areas are widespread also. A recent edited book contains a number of articles by a variety of authors mainly related to chemometrics and metabolomics [72].

This book is distinct in that it provides in-depth numerical and graphical descriptions of many of the common chemometrics methods used in metabolomics, illustrated by applications.

1.3 CASE STUDIES

The methods described in this text are illustrated by a mixture of case studies and simulations.

The case studies are all public domain and the datasets are available on the publisher's website, to allow readers to reproduce results using their own favourite software. Sufficient details are available in this text to allow the reproduction of the numbers and graphs in this book, which are not tied to any specific software package.

The experimental case studies are described in greater detail in Chapter 3, so we only summarise in brief below, providing readers with references to descriptions of the source experiments. It is not the purpose of this book to reproduce a full step-by-step analysis of each dataset, as this is already available in excellent published articles, but to use datasets where appropriate to illustrate the application of different methods in appropriate chapters. Some datasets may be more useful for illustrating variable

selection, classification performance, calibration or exploratory methods. Some datasets are only used a few times in this book, others are used in several chapters. We do not always employ the same approaches as in the original papers, and it is not the purpose of this book to copy existing literature, and for reasons of brevity, this book only describes some of the most common approaches that in this author's opinion are widespread. However, some readers will be interested in the full chemometric analysis and biological interpretation, and we list source papers below for readers wishing a more comprehensive background to the data.

Case Study 1: Presymptomatic study of humans with rheumatoid arthritis using blood plasma and LCMS was analysed in Umeå, Sweden, from samples obtained from the Biobank of Northern Sweden [73].

Case Study 2: Diagnosis of malaria in human blood plasma of children using GCMS, was analysed in Umeå, Sweden, from samples obtained in Rwanda [74].

Case Study 3: Measurement of Triglycerides in children's blood serum using NMR was obtained from a large study of school children in Norway, and the data were analysed in the University of Bergen, Norway [75].

Case Study 4: Glucose intolerance and diabetes in humans as assessed by blood serum using NMR involved clinical data collected from Beijing, China, and analysed in Wuhan, China [76].

Case Study 5: Metabolic changes in maize due to cold as assessed by NMR involved plant growth experiments performed and analysed in INRAE, France's National Research Institute for Agriculture, Food and Environment [77].

Case Study 6: Effect of nitrates on different parts of wheat leaves as analysed by FTIR was supplied by the University of Manchester, UK [78].

Case Study 7: Rapid discrimination of Enterococcal Bacteria in faecal isolates by Raman Spectroscopy was analysed in the University of Manchester, UK, with samples obtained from Belfast, UK [79].

Case Study 8: Effects of salinity, temperature and hypoxia on *Daphnia magna* metabolism as studied by GCMS, involving environmental toxicology, was obtained and analysed in IDAEA-CSIC, Barcelona, Spain [80].

Case Study 9: Bioactivity in a Chinese herbal medicine studied using HPLC was obtained in Hong Kong, in collaboration with the University of Bergen, Norway [81].

Case Study 10: Diabetes in mice studied by LCMS was obtained from UC Davis, California, USA [82], and is also described in case study ST000075 in the Metabolomics Workbench [63].

1.4 SOFTWARE

Crucial to chemometrics or all forms of statistical data analysis is software.

There are a large number of approaches, and which is chosen will depend on the background and expertise of the user, also how much time they have to develop code or analyse data, how important user friendliness and good graphics are, and of course

cost and technical support. Some packages are also well supported via dedicated courses and books. Often, different people have very strong views as to which package or environment is most suited. It may depend on whether you are in a large organisation with limited time to learn the basics of software development, whether you have a background in programming, whether you are intending to develop your own methods or just trying to learn the basics of different approaches before you use them.

This author wrote all programs for the methods described in this book, created all graphs and performed all calculations, using MATLAB, a large suite of software maintained by MathWorks [83]. It was originally invented by Cleve Moler in the 1960s [41] but was first released in a compiled version in the 1980s. Many early chemometricians used this as it is a very matrix-oriented environment. Over time the basic package has been substantially enhanced especially with very sophisticated graphics and very many toolboxes. The company runs many courses and there are a large number of MATLAB-based books available. There are several chemometrics-based software libraries that can be downloaded for free. It became very popular in the early days of the subject as its initial expansion took place at about the same time as the first events and organisation of chemometrics, so was a useful way chemometricians could talk together and exchange code and data, and was a matrix-oriented language with an extensive scripting language. Today there are around four million users, many in engineering. Scripts can be compiled so users without access to the package can use software written in MATLAB.

There are of course many other approaches.

Fortran is the oldest scientific programming language, first developed by IBM in the 1950s [36]. Not many people use Fortran in the chemometrics community currently, but it is considered a much lower level programming environment than MATLAB, in that there is much more flexibility in accessing the computer and it can be compiled into fast code. There are numerous numerical and graphical libraries, and matrix operations are easy. For development of fast code, it is preferred but it is more difficult to program than MATLAB. Since its original release in 1954, Fortran has undergone numerous updates to become a modern-day widespread numerical programming environment and still remains a well established programming language, and it is entirely possible to program in the methods from the text using Fortran. Some higher-level packages have been originally developed using Fortran so is still a choice for developing packages or applications in chemometrics and very good for numerical computations.

BASIC (Beginners' All-purpose Symbolic Instruction Code) was originally developed in the 1960s as an instructional language for college students [84], but was more widely used and refined in the next few decades, especially for use on microcomputers which started to emerge in the 1980s; Fortran was at the time more oriented towards large mainframes. Early versions of the SIMCA package (see below) running on micros were written in BASIC. As time moved on there were an increasing number of modifications for use in the early micros, and VBA (Visual Basic for Applications) [85] was developed by Microsoft with a particularly popular application to produce Add-ins for Excel: VBA was first launched with MS Excel 5.0 in 1993. There are several add-ins to perform common chemometrics methods for chemometrics in Excel [86, 87], which can be used for many basic procedures such as PLS and PCA. These together with

Excel's data analysis tools such as ANOVA or regression and numerous downloadable tools means that most chemometrics calculations can be performed in Excel although graphics are limited. BASIC though is an interpreted language and so slow for larger calculations, but can be sped up using DLLs (Dynamic Link Libraries) for functions often using compiled code in C (see below) and Visual Basic still remains a popular programming language.

C and C++ can be used for chemometrics but do require a good knowledge of programming and how computers work. C was originally developed at Bell Labs 1972 and 1973 to construct utilities running on Unix [88]. The likelihood many readers of this book will use C is quite low unless they are involved in developing a package from scratch, although it could be useful for compiled DLLs to link to VBA as above. C++ is particularly flexible and often used for programming operating systems but requires a good grasp of object-oriented programming.

At a higher level, the R programming language [89] is gaining substantial popularity within the chemometrics and statistics community, and has a comparable chemometrics user base to MATLAB. It is a free software originating from the University of Aukland, New Zealand, based on the precursor statistical language S. It has been widely developed subsequently, with thousands of contributed routines. The Comprehensive R Network (CRAN) [90] at time of writing consists of around 20,000 contributed routines and is the largest depository. The R journal published by the R foundation [91] was founded in 2009. A great deal of chemometrics and multivariate statistical software has now been implemented in this environment and is freely downloadable. Users with a background in bioinformatics are more likely to favour R than MATLAB, although users from an engineering background are more likely to prefer MATLAB. Despite R's substantial popularity in the statistical community, a survey of popularity of programming languages at the time of writing [92] places R lower than some of the other languages described elsewhere (Python, MATLAB, Visual Basic, FORTRAN) but free availability and a large depository of statistical routines makes R an increasingly widespread environment for the type of analyses discussed in this book, and worth learning for newcomers.

Python is another high-level programming environment that is freely available, created first in the late 1980s [93]. Interestingly, at the time of writing, Python, which is freely available, was listed as the most popular general programming language beating even C [92]. It has undergone a large number of releases, with version 3.8 released in October 2019 being the oldest still supported. There is some chemometrics software available in Python, which is open source and free to download [94].

There will be many readers of this book that do not wish to program methods. Some may be satisfied for example with Excel Add-ins or able to use MATLAB or R packages with limited programming knowledge, but others will want all-inclusive packages. In any organisation there may be a variety of users all with different computer skills and experiences. Some of the developers of packaged software produce excellent courses on how to handle data using software that has been developed for many years or decades and is oriented towards the laboratory-based user without much time to learn to program. This of course depends on how a unit is structured, some may have data analysis specialists who can develop their own procedures, but others will have laboratory-based scientists who want to immediately analyse the results of their data

as it becomes available without much programming but perhaps after attending one or more hands-on courses. Most of the methods in this book are possible to reproduce using niche packaged software, although aspects such as data pre-processing, variable selection and so on might be under default control of the software, and some software is more flexible than others. The less flexible the software, in general, the more automated but the more the decisions have been made for you in advance. The more flexible, the more you can control the outcome but may have to understand some of the procedures in processing the data. In all the environments described above the numerical results in this book should be obtained if the various details are followed carefully, although graphical output may differ. For the packaged software described below, in some cases especially using default methods, it may not always be possible to obtain the same results as in this book; however, it should be possible to understand why there are differences: in some cases, this can be corrected by changing parameters in the processing steps or by macro languages/scripting extensions to the basic package where available. Being able to compare numerical results between packages is a very important step in the understanding and safe use of statistical approaches for data analysis. In this text sufficient numerical and graphical output is presented to allow users to check against their own software and if necessary develop modifications for themselves where there are differences. If in doubt about a method it is recommended to use more than one package and compare results.

PLS Toolbox [95] was developed within the MATLAB environment. It was first released in 1989 at a time chemometrics was becoming established, with a demand for MATLAB-based scripts. Eigenvector Research, based in Seattle, which distributes the software, and supports this by courses and consultancy, was founded in 1995. A version called Solo which does not use the MATLAB command line but consists of compiled MATLAB code with a GUI is also available. This software is continually evolving and contains a wide range of chemometrics methods; it also allows users to flexibly develop their own approaches without programming from scratch. It would be suitable for both users without much or any programming experience or for users who are more comfortable doing some programming and want flexibility via macro commands. Eigenvector provides extensive training and consultancy based on their software.

The SIMCA package (to be distinguished from the method) was released and first developed at about the same time [96], based on the group in Umeå, Sweden. It was first released by Umetri, and at time of writing is now marketed via Sartorius [97]. It is a well-known software, based very much on the philosophy of Svante Wold and his colleagues in the early days of the subject. Unlike PLS Toolbox, it is very much a packaged software, allowing less flexibility, but very much oriented towards the laboratory-based user without requiring much software expertise. Many of the approaches such as SIMCA, OPLS and Hotelling's T^2etc are embedded into the software, which will guide the user along a certain pathway of data analysis, and many of the original developers were leaders in the early days who developed protocols for analysis of multivariate analytical data. There are extensive books and courses associated with using this package.

The Arthur program [37], written in Fortran developed for both mainframes and VAX minicomputers, from the University of Washington in Seattle, was one of the

first widespread chemometrics packages but still one had to be very expert to install it. From this seed emerged the company Infometrix in 1978 [98] which was probably the first software-based company primarily focussed on chemometrics, who developed the Pirouette package which continues to be updated and available today.

The UNSCRAMBLER software was originally developed in 1986 by Harald Martens from Norway and later by CAMO Software [99]. It was released at about the same time as SIMCA from Sweden. The late 1980s were a pioneering time for chemometrics software, with the change from mainframes and minicomputers to micros as the workhorse, and the first organised literature, meetings and courses in chemometrics, reflecting the need of the time. This software is still available and has been updated over time.

The Sirius package was also developed at around this time [100] from Bergen, Norway. It is distributed from Pattern Recognition Systems (PRS) and has facilities for many multivariate approaches of interest to readers of this book, including PLSDA, SIMCA and many variable selection approaches.

Some users of metabolomics software will have a prior background in applied statistics. Although many will be able to use R, there is a large community that prefers packages originally developed primarily for statisticians.

SAS (Statistical Analysis System) is the largest and most broadly based package oriented towards statisticians. It originated at North Carolina State University from 1966 until 1976, after which SAS Institute was incorporated [101]. SAS is the largest privately owned software company in the world. It has a very extensive suite of statistical and graphical software. SAS software includes a Base SAS component that performs analytical functions and more than 200 other modules that add graphics, spreadsheets or other features. Even some statistical concepts like Type I, Type II and Type III tests in ANOVA were first defined by SAS institute which has had a strong influence on applied statistics for many decades. The software is very comprehensive in areas such as ANOVA and regression. Most methods described in this book can be performed using SAS. There are extensive books, highly informative manuals, dedicated conferences and courses. There is a strong user base and certification [102]. However, the software is regarded as harder to use than packaged software such as SIMCA, PLS Toolbox, UNSCRAMBLER, Pirouette or Sirius, and not oriented towards the hands-on laboratory-based user, but the statistically trained modeller; some general programming skills may be desired, but this is not a low-level programming language. For the applied statistician moving into chemometrics though, most of the tools will be available in SAS, although it is not very commonly used in this area, probably due to the background of most users. SAS has a scripting language that allows users to develop their own approaches in addition to the add-on packages.

SPSS (Statistical Package for the Social Sciences) is one of the earliest general statistical packages, now maintained by IBM [103]. It was first released in 1968 for batch processing on mainframes, but in the 1980s was also implemented for micros. Although originally developed for social sciences, its applicability is wider and there are many multivariate procedures available. Scripts can also be developed using a macro language. Some biologists use SPSS for pattern recognition. It is regarded by many as somewhat easier to use compared to SAS, so would be suited as an entry-level statistical package for those without much statistical training, but is less comprehensive.

Although there are several other statistical packages readily available, the use of niche statistical software is not very common in metabolomics, so we will not produce a comprehensive list of statistical software and list only two of the most common and widespread above. Nevertheless, if using scripts or a macro language these statistical packages are likely to be suitable for most of the methods described in this book, and for readers working in large institutes such as universities it is quite common for there to be institutional licences and often in-house support for general statistical software.

REFERENCES

1. S. Wold (1972). Spline functions, a new tool in data-analysis. *Kemisk Tidskrift*, 3, 34–37.
2. B.R. Kowalski (ed) (1984). *Chemometrics: Mathematics and Statistics in Chemistry*, Reidel, Dordrecht.
3. *Journal of Chemometrics* (1987). Wiley, Chichester.
4. *Chemometrics and Intelligent Laboratory Systems* (1986). Elsevier, Amsterdam.
5. M.A. Sharaf, D.L. Illman and B.R. Kowalski (1986). *Chemometrics*, Wiley, New York.
6. D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte and L. Kaufman (1988). *Chemometrics: A Textbook*, Elsevier, Amsterdam.
7. B.R. Kowalski (1980). Chemometrics. *Analytical Chemistry*, 52, R112–R122.
8. K. Pearson (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science 5* (50), 157–175.
9. Student (aka W S Gossett) (1908). The probable error of a mean. *Biometrika*, 6, 1–25.
10. J. Arbuthnot (1710). An argument for Divine Providence, taken from the constant regularity observed in the births of both sexes. *Philosophical Transactions of the Royal Society of London*, 27, 186–190.
11. D. Bernoulli (1735). Recherches physiques et astronomiques. *Pieces qui ont Remporte le Prix Double de l'Academie Royale des Sciences en,* 1734, 93–122.
12. P. Laplace (1778). Mémoire sur les probabilités. *Mémoires de l'Académie Royale des Sciences de Paris*, 9, 227–332.
13. R.A. Fisher (1925). *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh.
14. R.A. Fisher (1935). *The Design of Experiments*, Oliver and Boyd, New York.
15. R.A. Fisher (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, 179–188.
16. H. Hotelling (1933). Analysis of a complex of statistical variables into principal components. *Journal of Education & Psychology,* 24 (417–441), 498–520.
17. H. Hotelling (1936). Simplified calculation of principal components. *Psychometrika,* 1, 27–35.
18. J. Neyman and E.S. Pearson (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: part I. *Biometrika*, 20A, 175–240.
19. D. Salsburg (2001). *The Lady Tasting Tea*, W. H. Freeman and Co., New York.

20. R.A. Fisher (1921). Studies in crop variation. I. An examination of the yield of dressed grain from Broadbalk. *The Journal of Agricultural Science,* 11, 107–135.
21. F. Yates (1937). The design and analysis of factorial experiments, Technical Communication of the Commonwealth Bureau of Soils 35 *Commonwealth Agricultural Bureaux,* Farnham Royal.
22. O.L. Davies (ed) (1956). *Statistical Methods in Research and Production,* Longman, London.
23. G.E.P. Box, W.G. Hunter and J.S. Hunter (1978). *Statistics for Experimenters*, Wiley, New York.
24. P.C. Mahalanobis (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2, 49–55.
25. S. Wold, M. Sjöström and L. Eriksson (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58, 109–130.
26. P. Geladi and B.R. Kowalski (1986). Partial least squares: a tutorial. *Analytica Chimica Acta*, 185, 1–19
27. S. Wold (1976). Pattern-recognition by means of disjoint principal components models. *Pattern Recognition*, 8, 127–139.
28. J. Mandel (1949). Statistical methods in analytical chemistry. *Journal of Chemical Education*, 26, 534–539
29. H.J. Hader and W.J. Youden (1952). Experimental statistics. *Analytical Chemistry*, 24, 120–124.
30. W.J. Youden (1950). Index for rating diagnostic tests. *Cancer*, 3, 32–25
31. S.N. Deming and S.L. Morgan (1973). Simplex optimization of variables in analytical-chemistry. *Analytical Chemistry*, 45, A278–279.
32. S.N. Deming and S.L. Morgan (1987) *Experimental Design: A Chemometric Approach*, Elsevier, Amsterdam.
33. R.M. Wallace and S.M. Katz (1964). A method for determination of rank in analysis of absorption spectra of multicomponent systems. *The Journal of Physical Chemistry*, 68, 3890–3892.
34. D. Katakis (1965). Matrix rank analysis of spectral data. *Analytical Chemistry*, 37, 876–878.
35. E.R. Malinowski and D.G. Howery (1980). *Factor Analysis in Chemistry*, Wiley, New York.
36. J.W. Backus, H. Herrick and I. Ziller, (1954). Preliminary Report: Specifications for the IBM Mathematical FORmula TRANSlating System, FORTRAN. Programming Research Group, Applied Science Division*, International Business Machines Corporation*.
37. B.R. Kowalski, P.C. Jurs, T.L. Isenhour and C.N. Reilly (1969). An investigation of combined patterns from diverse analytical data using computerized learning machines. *Analytical Chemistry*, 41, 1949–1953
38. R.K. Lindsay, B.G. Buchanan, E.A. Feigenbaum and J. Lederberg (1980). *Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project*, McGraw-Hill, New York.
39. A.M. Harper, D.L. Duewer, B.R. Kowalski and J.L. Fasching (1977). ARTHUR an experimental data analysis: the heuristic use of a polyalgorithm, in B.R. Kowalski

(ed), *Chemometrics: Theory and Applications, ACS Symp. Ser. 52*, American Chemical Society, Washington, DC.

40. A. Pollock (1981). Big IBM's Little Computer. *New York Times*, Section D, page 1.
41. C. Moler (1981). *MATLAB Users' Guide*, University of New Mexico, Albuquerque, New Mexico.
42. P. Geladi and P.K. Hopke (2008). Is there a future for chemometrics? Are we still needed?. *Journal of Chemometrics*, 22, 289–290.
43. W. Johannsen (1911). The genotype conception of heredity. *American Naturalist*, 45, 129–159.
44. H.L. Winkler (1920). *Verbreitung und Ursache der Parthenogenesis im Pflanzen- und Tierreiche*, Verlag Fischer, Jena.
45. S. Anderson, A.T. Bankier, B.G. Barrell, M.H. de Bruijn, A.R. Coulson, J. Drouin, et al. (1981). Sequence and organization of the human mitochondrial genome. *Nature*, 290 (5806), 457–465.
46. K. Ohyama, H. Fukuzawa, T. Kohchi, H. Shirai, T. Sano, S. Sano, et al. (1986). Chloroplast gene organization deduced from complete sequence of liverwort Marchantia polymorpha chloroplast DNA. *Nature*, 322 (6079), 572–574.
47. S.G. Oliver, O.J. van der Aart, M.L. Agostoni-Carbone, M. Aigle, L. Alberghina, D. Alexandraki, et al. (1992). The complete DNA sequence of yeast chromosome III. *Nature*, 357 (6373), 38–46.
48. R.D. Fleischmann, M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269 (5223), 496–512.
49. I. Noble (2003). Human genome finally complete, http://news.bbc.co.uk/1/hi/sci/tech/2940601.stm.
50. V.A. McKusick and F.H. Ruddle (1987). A new discipline, a new name, a new journal. *Genomics*, 1, 1–2.
51. V.E. Velculescu, L. Zhang, W. Zhou, J. Vogelstein, M.A. Basrai, D.E. Bassett, et al. (1997). Characterization of the yeast transcriptome. *Cell*, 88, 243–251.
52. V.C. Wasinger, S.J. Cordwell, A. Cerpa-Poljak, J.X. Yan, A.A. Gooley, M.R. Wilkins, et al. (1995). Progress with gene-product mapping of the Mollicutes: *Mycoplasma genitalium*. *Electrophoresis*, 16, 1090–1094.
53. P. James (1997). Protein identification in the post-genome era: the rapid rise of proteomics. *Quarterly Reviews of Biophysics*, 30, 279–331.
54. *Proteomics* (2001). Wiley, Chichester.
55. S.C. Gates and C.C. Sweeley (1978). Quantitative metabolic profiling based on gas chromatography. *Clinical Chemistry*, 24, 1663–1673.
56. B.F. Cravatt, O. Prospero-Garcia, G. Siuzdak, N.B. Gilula, S.J. Henriksen, D.L. Boger, et al. (1995). Chemical characterization of a family of brain lipids that induce sleep. *Science*, 268, 1506–1509.
57. D.I. Hoult, S.J. Busby, D.G. Gadian, G.K. Radda, R.E. Richards and P.J. Seeley (1974). Observation of tissue metabolites using ^{31}P nuclear magnetic resonance. *Nature*, 252, 285–287.

58. S.G. Oliver, M.K. Winson, D.B. Kell and F. Baganz (1998). Systematic functional analysis of the yeast genome. *Trends in Biotechnology*, 16, 373–378.
59. H. Tweeddale, L. Notley-McRobb and T. Ferenci (1998). Effect of slow growth on metabolism of *Escherichia coli*, as revealed by global metabolite pool ("Metabolome") analysis. *Journal of Bacteriology*, 180, 5109–5116.
60. M. Satake, B. Dmochowska, Y. Nishikawa, J. Madaj, J. Xue, Z.W. Guo et al. (2003). Vitamin C metabolomic mapping in the lens with 6-deoxy-6-fluoro-ascorbic acid and high-resolution ^{19}F-NMR spectroscopy. *Investigative Ophthalmology and Visual Science*, 44, 2047–2058
61. *Metabolomics* (2001). Springer Nature, Heidelberg.
62. D.S. Wishart, D. Tzur, C. Knox, R. Eisner, A.C. Guo, N. Young, et al. (2007). HMDB: the Human Metabolome Database. *Nucleic Acids Research*, 35 (Database issue), D521–D526.
63. Metabolomics Workbench, https://www.metabolomicsworkbench.org/.
64. Metabolights, https://www.ebi.ac.uk/metabolights/.
65. J.K. Nicholson, L.C. Lindon and E. Holmes (1999). 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, 29, 1181–1189.
66. M.R. Wenk (2005). The emerging field of lipidomics. *Nature Reviews Drug Discovery*, 4, 594–610.
67. J. van der Greef and A.K Smilde (2005). Symbiosis of chemometrics and metabolomics: past, present, and future. *Journal of Chemometrics*, 19, 376–386.
68. J. Trygg, E. Holmes and T. Lundstedt (2007). Chemometrics in metabonomics. *Journal of Proteome Research*, 6, 469–479.
69. D.S. Waterman, F.W. Bonner and J.C. Lindon (2009). Spectroscopic and statistical methods in metabonomics. *Bioanalysis*, 1, 1559–1578.
70. J. Trygg, J. Gullberg, A.I. Johansson, P. Jonsson and T. Moritz (2006). Chemometrics in metabolomics — an introduction. In: K. Saito, R.A. Dixon, L. Willmitzer (ed) *Plant Metabolomics. Biotechnology in Agriculture and Forestry*, vol. 57, pp 117–128 Springer, Berlin, Heidelberg.
71. R. Madsen, T. Lundstedt and J. Trygg (2010). Chemometrics in metabolomics—a review in human disease diagnosis. *Analytica Chimica Acta*, 659, 23–33.
72. J. Jaumot, C. Bedia and R. Tauler (ed) (2018). *Data analysis for Omic sciences: methods and applications*, Comprehensive Analytical Chemistry vol. 82, Elsevier, Amsterdam.
73. I. Surowiec, L. Ärlestig, S. Rantapää-Dahlqvist and J. Trygg (2016). Metabolite and lipid profiling of biobank plasma samples collected prior to onset of Rheumatoid arthritis. *PLoS One*, 11, e0164196.
74. I. Surowiec, J. Orikiiriza, E. Karlsson, M. Nelson, M. Bonde, P. Kyamanwa, et al. (2015). Metabolic signature profiling as a diagnostic and prognostic tool in pediatric *Plasmodium falciparum* malaria. *Open Forum Infectious Diseases*, 2, 2.
75. E. Aadland, O.M. Kvalheim, S.A. Anderssen, G.K. Resaland and L.B. Andersen (2018). The multivariate physical activity signature associated with metabolic health in children. *International Journal of Behavioral Nutrition and Physical Activity*, 15, 77.

76. X.Y. Zhang, Y.L. Wang, F.H. Hao, X.H. Zhou, X.Y. Han, H.R. Tang, et al. (2009). Human serum metabonomic analysis reveals progression axes for glucose intolerance and insulin resistance statuses. *Journal of Proteome Research*, 8, 5188–5195.
77. M. Urrutia, M. Blein-Nicolas, S. Prigent, S. Bernillon, C. Deborde, T. Balliau, et al. (2021). Maize metabolome and proteome responses to controlled cold stress partly mimic early-sowing effects in the field and differ from those of Arabidopsis. *Plant, Cell and Environment*, 44, 1504–1521.
78. J.W. Allwood, S. Chandra, Y. Xu, W.B. Dunn, E. Correa, L. Hopkins, et al. (2015). Profiling of spatial metabolite distributions in wheat leaves under normal and nitrate limiting conditions. *Phytochemistry*, 115, 99–111.
79. N. AlMasoud, Y. Xu, D.I. Ellis, P. Rooney, J.F. Turton and R. Goodacre (2016). *Analytical Methods*, 8, 7603–7613.
80. E. Garreta-Lara, B. Campos, C. Barata, S. Lacorte and R. Tauler (2018). Combined effects of salinity, temperature and hypoxia on *Daphnia magna* metabolism. *Science of the Total Environment*, 610, 602–612.
81. F.T. Chau, H.Y. Chan, C.Y. Cheung, C.J. Xu, Y. Liang and O.M. Kvalheim (2009). Recipe for uncovering the bioactive components in herbal medicine *Analytical Chemistry*, 81, 7217–7225.
82. J. Fahrmann, D. Grapov, J. Yang, B. Hammock, O. Fiehn, G.I. Bell, et al. (2015). Systemic alterations in the metabolome of diabetic NOD mice delineate increased oxidative stress accompanied by reduced inflammation and hypertriglyceremia. *The American Journal of Physiology – Endocrinology and Metabolism*, 308, E978–E989.
83. Mathworks, www.mathworks.com.
84. J.G. Kemeny, T.E. Kurtz and E. Thomas (1964). *Basic: A Manual for BASIC, the Elementary Algebraic Language Designed for use with the Dartmouth Time Sharing System*, Dartmouth College Computation Center, Hanover, N.H.
85. *Getting Started with VBA in Office,* https://learn.microsoft.com/en-us/office/vba/library-reference/concepts/getting-started-with-vba-in-office.
86. R.G. Brereton (2003). *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*, Wiley, Chichester.
87. A.L. Pomerantsev (2014). *Chemometrics in Excel*, Wiley, Chichester.
88. D.M. Ritchie (1993). The development of the C language. *ACM SIGPLAN Notices*, 28, 201–208.
89. F.M. Giorgi, C. Ceraolo and D. Mercatelli (2022). The R language: an engine for bioinformatics and data science. *Life*, 12, 648.
90. *The Comprehensive R Archive Network*, https://cran.r-project.org/.
91. V. Carey (2009). Special section: the Future of R. *R Journal*, 1, 3.
92. *TIOBE Index*, https://www.tiobe.com/tiobe-index/.
93. *General Python FAQ*, https://docs.python.org/3/faq/general.html#why-was-python-created-in-the-first-place.
94. https://pypi.org/project/chemometrics/.
95. *PLS_Toolbox*, https://eigenvector.com/software/pls-toolbox/.

96. W. Dunn (1987). SIMCA-3B, a pattern-recognition program. *Chemometrics and Intelligent Laboratory Systems,* 2, 126-127.
97. SIMCA https://www.sartorius.com/en/products/process-analytical-technology/data-analytics-software/mvda-software/simca.
98. *History of Infometrix,* https://infometrix.com/company/history-of-infometrix/.
99. V. Tysso, K. Esbensen and H. Martens (1987). UNSCRAMBLER – an interactive program for multivariate calibration and prediction. *Chemometrics and Intelligent Laboratory Systems*, 2, 239–243.
100. O.M. Kvalheim and T.V. Karstang (1997). A general purpose program for multivariate data analysis. *Chemometrics and Intelligent Laboratory Systems*, 2, 235–237.
101. E.S. Nourse, B.G. Greenberg, G.M. Cox, D.D. Mason, J.E. Grizzle, N.L. Johnson, et al. (1978). Statistical training and research: the University of North Carolina System. *International Statistical Review*, 46, 171–206.
102. G. Balfour (1999). Certification program validates SAS users *Computerworld Canada*.
103. N.H. Nie, D.H. Bent and H.C. Hadlai (1970). *SPSS: Statistical Package for the Social Sciences,* McGraw-Hill, New York.