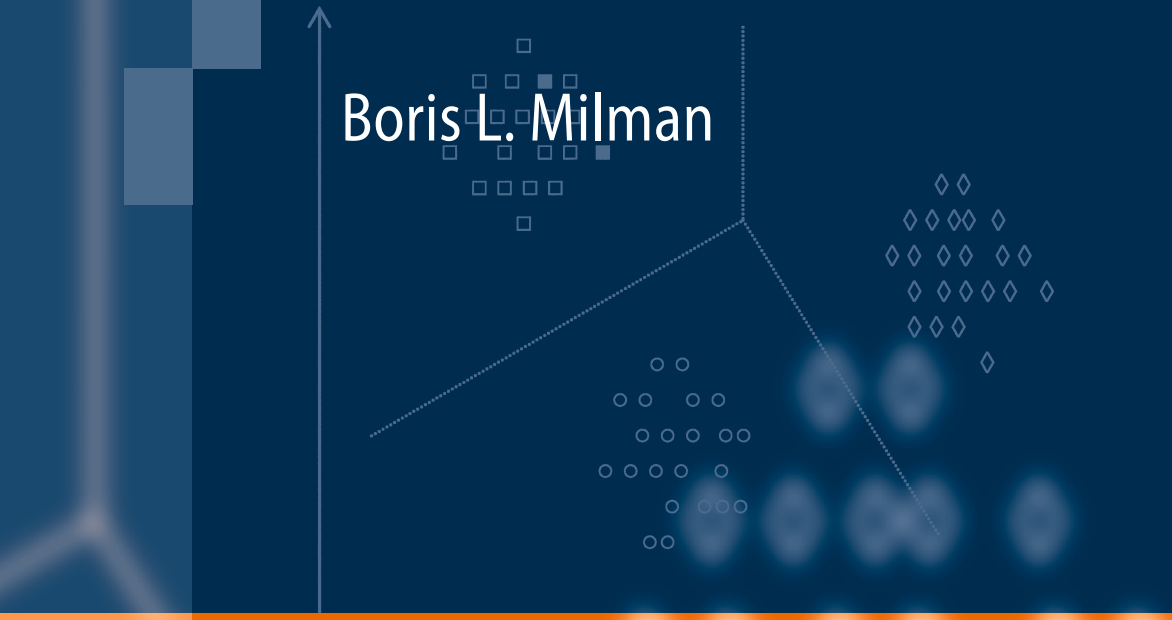


Boris L. Milman

# Chemical Identification and its Quality Assurance

Springer

# Chemical Identification and its Quality Assurance

Boris L. Milman

# Chemical Identification and its Quality Assurance

Springer

Boris L. Milman
D.I. Mendeleyev Inst. for Metrology
(VNIIM) and Cent. for Ecol. Saf. of Russ. Acad. of Sciences
65, 9 Morskaya nab
199226 St. Petersburg
Russia
bmilman@mail.rcom.ru

ISBN 978-3-642-15360-0 e-ISBN 978-3-642-15361-7
DOI 10.1007/978-3-642-15361-7
Springer Heidelberg Dordrecht London New York



*Cover design:* WMX Design GmbH, Heidelberg, Germany

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

*to Inna*

# Preface

Analytical chemistry plays a very important role in most fields of science, industrial and environmental control, healthcare, and many other areas of academic, ecological, economic, and social values. Many innovations in analytics ultimately result in the discovery of new complex chemical compounds, clear and thorough understanding of living nature, improvement of quality and safety of goods, reduction in pollutant levels in the environment, and so on. Also, progress in analytical chemistry, as well as in any basic science, is certainly important as such.

In its structure, this science is the holistic unity of qualitative and quantitative analysis, which can be considered separately in the fields of research, innovation introduction, learning the discipline in universities, training and education programmes, etc. For the last two decades, a succession of scientists specialized in general analytical methodology, chemical metrology, and analytical fields where detection and identification of chemical compounds is of particular importance, perceived and expressed an opinion that modern qualitative chemical analysis had been insufficiently described in general treatises and guidances on analytics, metrology, standardization, quality assurance, and so on. Unlike analytical techniques for qualitative and quantitative determinations, well-presented in books and reviews, theoretical principles of identification and general experimental approaches to its implementation have not received comprehensive treatment in the literature. This prevents progress in the development and consistent validation of particular qualitative procedures, quality assurance of the proper analytical data, and expressing and reporting identification errors analogously to errors/uncertainties in quantitative analysis.

This book entirely devoted to chemical identification has been written especially to

- Remove "skewness" of presentation of two principal parts of chemical analytics in the literature on an analytical methodology
- Generalize approaches to identification of both various chemical compounds and samples containing these compounds
- Summarize methods of estimating trueness of identification results

- Draw the special attention of practical analysts to non-target qualitative analysis seldom or never considered in the general literature on analytical chemistry
- Spotlight issues of quality assurance and control in identification/qualitative analysis

The book is intended for anyone engaged in analytical and bioanalytical chemistry: professionals in reference, test, and control laboratories; scientists in research laboratories of universities and chemical, pharmaceutical, and biotechnology companies; graduate students of analytical chemistry, chemometrics, chromatography, spectroscopy, and quality assurance and control. In style, the book is both monograph and also laboratory guidance/manual. I hope that these two aspects complement each other.

The book begins with the consideration of basic principles of chemical identification, including main concepts and terminology (Chap. 1). Following are chapters covering analytical techniques (Chap. 2) and statistical/calculation methods (Chap. 3) required for identification purposes. Only brief information is given here, with references to comprehensive literature sources. Chapter 4 deals with different quantities and indices expressing trueness of results of qualitative analysis, detection and identification, and rates of their errors. In the book, procedures for qualitative analysis are divided into target identification by methods (Chap. 5) and unknown/non-target analysis (Chap. 7). For the latter, prior data extracted from chemical databases are very essential (Chap. 6). Identification/classification of objects such as foodstuffs, pollutions, microorganisms, materials, and so on is described in Chap. 8. Finally, issues of quality assurance and control in relation to qualitative analytical procedures are explored in Chap. 9.

Three remarks are necessary. First, because of my professional interest, low molecular compounds are covered to a greater extent than high molecular ones. Nevertheless, the latter are also of concern, in the respect that progress of analysis of high molecules, first of all in proteomics, affects the development of general analytical methodology. Second, general issues of chemical analysis are discussed only if related to identification problems. Third, qualitative procedures related to identification, such as detection, are also considered.

My view on the subject of chemical identification was formed not only by me alone but also as the result of cooperation with other persons. I would like to name them here.

Prof. Miguel Valcárcel (University of Córdoba) invited me to participate in the MEQUALAN project. Dr. Willie May (NIST) and Dr. Reenie Parris (NIST) introduced me to the analytical laboratory responsible for the development of reference methods and materials. Dr. Stephen Stein (NIST) was my supervisor in the project of building a library of tandem mass spectra. Mr. W.A. Hardcastle sent me the LGC document on qualitative analysis cited in the book. Dr. Steven Lehotay (USDA Agricultural Research Service) sent his recent articles on identification methodology. Dr. Valeri Babushok (NIST) introduced me to details of the database on retention indices. Dr. Inna Zhurkovich (Institute of Plant Protection) provided me with considerable advice about general issues of chemical analysis, and explained

many details concerning liquid chromatography and pesticide analysis. Dr. Yana Russkikh and Mrs. Lyuba Tselikova (Centre for Ecological Safety) recorded some mass spectra presented in the book. I wish to thank them all for their cooperation, discussions, introductions, or assistance.

I also thank and formally acknowledge the following publisher and persons for permission to use their materials as figures in this book: Elsevier Publishers, Dr. John Cottrell (Matrix Science), Dr. Per Daling (SINTEF), and Dr. S. Stein.

My special thanks to the editorial staff of Springer Chemistry, and personally Dr. Steffen Pauly, for all they did for publishing this book.

Saint Petersburg
November 2010

Boris L. Milman

# Contents

# Abbreviations and Symbols

| | |
|---|---|
| 1D | One-dimensional |
| 2D | Two-dimensional |
| 3D | Three-dimensional |
| AAFS | American Academy of Forensic Sciences, USA |
| ACS | American Chemical Society, USA |
| AIST | National Institute of Advanced Industrial Science and Technology, Japan |
| ANN | Artificial neural network |
| AOAC | Association of Official Analytical Chemists |
| AORC | Association of Official Racing Chemists |
| APCI | Atmospheric pressure chemical ionization |
| ASTM | American Society for Testing and Materials, USA |
| BAM | Federal Institute for Materials Research and Testing, Germany |
| CA | Chemical Abstracts |
| CAC/GL | Guideline(s) of Codex Alimentarius Commission |
| CAS | Chemical Abstracts Service |
| CE | Capillary electrophoresis |
| CI | Chemical ionization |
| CID | Collision-induced dissociation |
| CITAC | Cooperation on International Traceability in Analytical Chemistry |
| CRM | Certified Reference Material(s) |
| CSI | Chemical Substance Index to CA |
| DA | Discriminant analysis |
| DAD | Photo diode array detector |
| DNA | Deoxyribonucleic acid(s) |
| DP | Discriminating power |
| EC | European Commission |
| ECD | Electron capture detector |
| EI | Electron ionization |
| EPA | Environmental Protection Agency, USA |

| | |
|---|---|
| EPO | European Patent Office |
| ESI | Electrospray ionization |
| EU | European Union |
| EURACHEM | European network of organizations of analytical chemists and chemical metrologists |
| FA | Factor analysis |
| FAO | Food and Agriculture Organization |
| FDA | Food and Drug Administration, USA |
| FID | Flame ionization detector |
| FN | False negative (result) |
| FP | False positive (result) |
| FPD | Flame photometric detector |
| FT | Fourier transform |
| GC | Gas chromatography |
| GMD | Golm Metabolome Database |
| HMDB | Human Metabolome Database |
| HOSE | Hierarchical organization of spherical environments |
| HPLC | High-performance liquid chromatography |
| HPTLC | High-performance thin layer chromatography |
| HRMS | High-resolution mass spectrometry |
| $HRMS^n$ | High-resolution multistage/multiple mass spectrometry |
| ICR | Ion cyclotron resonance |
| IEC | International Electrotechnical Commission |
| InChI | IUPAC international chemical identifier |
| INFOODS | International Network of Food Data Systems |
| IP | Identification point(s) |
| IR | Infrared, infrared spectroscopy |
| IRMM | European Institute for Reference Materials and Measurements |
| ISO | International Organization for Standardization |
| IT | Ion trap |
| IUPAC | International Union of Pure and Applied Chemistry |
| KI | Kovats retention index(ices) |
| *k*-NN | *k*-nearest neighbors |
| LC | Liquid chromatography |
| LDA | Linear discriminant analysis |
| LIT | Linear ion trap |
| LRI | Linear retention index(ices) |
| M | Molecule |
| MALDI | Matrix-assisted laser desorption/ionization |
| MEKC | Micellar electrokinetic chromatography |
| MF | Match factor(s) |
| MLL | Mean list length |
| MMD | Manchester Metabolomics Database |
| MRPL | Minimum required performance limit(s) |
| MS | Mass spectrometry |

| | |
|---|---|
| $MS^1$ | Single mass spectrometry |
| $MS^2$, MS/MS | Tandem mass spectrometry |
| $MS^n$ | Multistage/multiple mass spectrometry |
| MSSJ | Mass Spectrometry Society of Japan |
| N | Negative |
| NCI | Negative chemical ionization |
| NIH | National Institutes of Health, USA |
| NIR | Near-infrared, near-infrared spectroscopy |
| NIST | National Institute of Standards and Technology, USA |
| NMIJ | National Metrology Institute of Japan, Japan |
| NMR | Nuclear magnetic resonance |
| NPD | Nitrogen phosphorous detector |
| OECD | Organization for Economic Cooperation and Development |
| OPCW | Organization for the Prohibition of Chemical Weapons |
| P | Positive |
| PAH | Polycyclic aromatic hydrocarbons |
| PBM | Probability-based matching |
| PCA | Principal component analysis |
| PCB | Polychlorinated biphenyls |
| PCDD/F | Polychlorinated dibenzodioxins and dibenzofurans |
| PFPD | Pulsed flame photometric detector |
| PLS | Partial least squares |
| Q | Quadrupole |
| QA | Quality assurance |
| QC | Quality control |
| QSRRs | Quantitative structure-(chromatographic) retention relationships |
| RI | Retention index(ices) |
| RM | Reference material(s) |
| RN | Registration number(s) |
| ROC | Receiver operating characteristic(s) |
| RRT | Relative retention time(s) |
| RT | Retention time(s) |
| SIM | Selected/single ion monitoring |
| SIMCA | Soft independent modeling of class analogy |
| SIMS | Secondary ion mass spectrometry |
| SMILES | Simplified molecular input line entry specification |
| SOFT | Society of Forensic Toxicologists, USA |
| sPRG | Proteomics standards research group |
| SRM | Selected/single reaction monitoring |
| TaMaSA | Tandem mass spectral abstracts |
| TLC | Thin layer chromatography |
| TN | True negative (result) |
| ToF | Time-of-flight |
| TP | True positive (result) |
| TQ | Triple quadrupole |

| | |
|---|---|
| UNIIM | Ural Research Institute for Metrology, Russia |
| UPLC | Ultra performance liquid chromatography |
| USDA | United States Department of Agriculture |
| USPTO | United States Patent and Trademark Office |
| UV | Ultraviolet |
| UV–Vis | Ultraviolet and visible (spectroscopy) |
| VNIIM | D.I. Mendeleyev Institute for Metrology, Russia |
| WADA | World Anti-Doping Agency |
| WHO | World Health Organization |
| WIPO | World Intellectual Property Organization |
| $c$ | Concentration |
| $CC\alpha$ | Critical concentration as decision limit |
| $CC\beta$ | Critical concentration as detection limit |
| $CPPV$ | Cumulative positive predictive value |
| $FDR$ | False discovery rate |
| $FNR$ | False negative rate |
| $FOR$ | Direct MF as dot-product |
| $FPR$ | False positive rate |
| $FR$ | False rate |
| $H_0$ | Null hypothesis |
| $\overline{H}_0$ | Alternative hypothesis |
| $H_1$ | First hypothesis |
| $H_2$ | Second hypothesis |
| $I$ | Spectral peak intensity, analytical signal intensity |
| $m/z$ | Mass-to-charge ratio |
| $NPV$ | Negative predictive value |
| $PPV$ | Positive predictive value |
| $p$ | Probability |
| $Prob$ | Probability of identification in the NIST MS Search program |
| $Pv$ | Prevalence |
| $Rev$ | Reverse MF as dot-product |
| $s$ | Sample standard deviation |
| S/N | Signal-to-noise ratio |
| $Sp$ | Statistical specificity |
| $St$ | Statistical sensitivity |
| $t$ | Parameter of Student's $t$ distribution |
| $TNR$ | True negative rate |
| $TPR$ | True positive rate |
| $x$ | Variable, quantity |
| $z$ | Quantile of normal distribution |
| $\alpha$ | Type I error |
| $\beta$ | Type II error |
| $\Delta_c$ | Confidence interval |
| $\sigma$ | Population standard deviation |
| $\upsilon$ | Degrees of freedom |

# Chapter 1
# Principles of Identification

**Abstract** In this initial chapter, concepts and terms related to qualitative chemical analysis are outlined and discussed. Chemical identification is defined as assigning an analyte to one from known chemical compounds or a group/class of compounds. General principles for identification through the use of chemical tests and instrumental measurements are formulated. Qualitative analytical procedures and approaches to implement them are classified. Components of identification procedures are further described. Objects for identification such as compounds, substances, and analyzed samples are discussed in great detail, including identifiers of the objects. Known chemical substances, which amount to more than 110 million entities, are statistically reviewed. Finally, two key metrological issues, traceability in identification operations and qualitative scale of measurements, are discussed.

## 1.1 Introduction

The value of identification procedures is hard to overestimate.

> Qualitative analysis may take place without quantitative analysis, but quantitative analysis requires the identification (qualification) of the analytes for which numerical estimates are given [1].

Another obvious statement is that any science or scientific field should be based on the proper theory. Therefore one could discuss "theory for qualitative analysis" or "theory of chemical identification". Nevertheless, a theory of such a type seems to be a rather loose (miscellaneous) structure consisting of (a) pieces of theories from chemistry, physics, statistics, and so on, (b) empirical regularities derived from experiments in analytical chemistry, and (c) elements of theory of decision-making (see Chap. 3). To date, it is hard to consider this theoretical basis as a holistic, rigorous, and logically consistent theory. *Principles of chemical identification*, treated in this chapter, is a more adequate term for defining the theoretical foundation of this part of chemical analysis.

One or other of the general principles of qualitative analysis (identification) have been developed over the last few decades, possibly starting from the book [2]

B.L. Milman, *Chemical Identification and its Quality Assurance*,
DOI 10.1007/978-3-642-15361-7_1, 

in which detection errors at low analyte concentrations were considered and estimated. The reviews and articles [3–9], the special issue of Trends in Analytical Chemistry [10], and two European Commission documents [11, 12] seem to be the most valuable recent contributions in theoretical principles for chemical identification. Those were issued before, at the same time, or after the author's work devoted to this topic [13–19].

## 1.2 The Concept of Identification

To identify, from Latin *identificare*, means to

> . . . recognize as being, establish the identity of someone or something [20].

In chemistry, we consider identification of chemical entities such as chemical elements, their compounds, and group/class/mixture of compounds. In chemical analysis, those are components of systems/samples chemically analyzed, i.e., analytes. Analyzed samples themselves can also be identified, i.e., classified by means of techniques and methods of chemical analysis. In the book, any identification related to both chemical entities and chemical analysis is considered as a chemical one.

A number of definitions for chemical identification have been proposed (see [19]). Combining some of them and the above consideration, the following definition can be proposed:

*Chemical identification is assigning an analyte (analytical signal) to one of the set of known individual chemical compounds or to a group/class of compounds.*

The definition should be supplemented by some remarks and explanations.

1. An analyst often is not able to see the substance he/she determines because an analyte is present in a sample in a very low amount. All an analyst sees is an analytical signal, e.g., chromatographic or spectral peak. In such cases, identification relates to different kinds of signal processing rather than direct manipulation of the substance.
2. According to our treatment, identification is both an analytical procedure and its result. It is also a "bridge" between a procedure and a result, which is the analyst's idea/decision on identity between an analyte and one of the known compounds. The identity is concluded on the basis of identification criteria established in advance or of *ad hoc* criteria.
3. A differentiation should be made between *chemical substances* and *chemical compounds*. An (individual) compound is formed from different elements and has a definite molecular structure. A substance may be (a) formed by a single element, or (b) composed from one or more different individual compounds. *Substances*, *chemicals*, and *materials* considered as products may be synonyms: e.g., "benzene" may be both *compound* and *substance*. "Benzene 99%" manufactured by some chemical company is *substance*, *chemical*, and so on. It is also clear that *compound* and *substance* are not absolutely different concepts. *Compound* and *pure complex substance* are terms certainly related to each other.

*Compound(s)* will be preferentially used in this book when discussing identification operations (different from qualitative analysis II; see below).

4. In biochemical analysis, e.g., in analytical proteomics, known compounds may be virtual (possible, predicted; see Chap. 7).
5. Identification of individual compounds may be required to identify/classify an analyzed sample itself as the definite kind/type/grade/brand of products, materials, compositions, formulations, and so on. This type of chemical analysis was named *qualitative analysis II* [12, 19, 21] (see Chap. 8), and can be called *identification II*. A sample can be also identified using methods of fingerprinting, without individual recognition of its components. As a rule, such methods are also referred to as chemical analysis. Therefore, the definition of identification can be extended:

*Chemical identification is also assigning an analyzed sample to one of the classification groups for specimens, materials, products, foodstuffs, pollutions, living organisms, and so on, using the techniques and methods of chemical analysis.*

For qualitative analysis II, notions of *identification* and *classification* are obviously similar. Here the terms of *authenticity* and *authentication* are also used; authentication is a confirmation of identity. *Classification* occurs in "usual" identification when an analyte is assigned to one of the classes of chemical compounds, e.g., organic nitrogen ones. For other terms and concepts, see Chap. 8.

*Identification* and *qualitative analysis* is another pair of related terms (Fig. 1.1). They are similar in their general sense, and used in different fields of chemical analysis. *Identification* is mainly used as the name for the corresponding procedure in instrumental analysis of organic and bioorganic substances. *Qualitative analysis* refers to determination of elements or compound classes by relatively simple methods/techniques such as qualitative/spot reactions and chemical test kits.

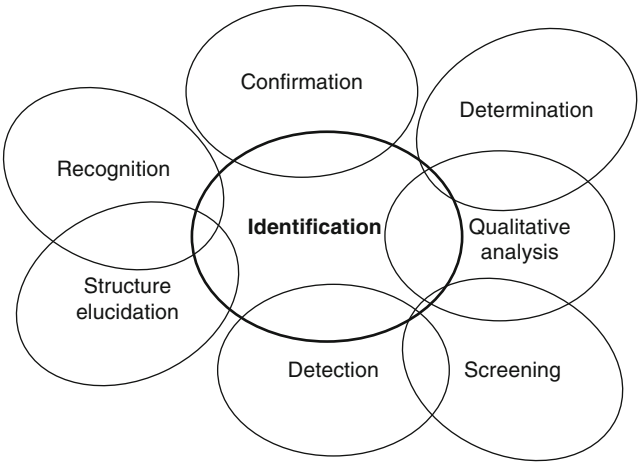



**Fig. 1.1** Representation of the concepts and terminology related to identification. The *overlapping circle* shows very approximately a similarity degree between different terms

There are also a number of other terms and concepts (Fig. 1.1) resembling *identification* and *qualitative analysis* in one characteristic or another.

- *Determination* covers both quantitative and qualitative aspects of chemical analysis.
- *Screening* is fast chemical analysis (often a multianalyte one) of a lot of samples, with only preliminary conclusions about quantitative and qualitative results.
- Screening results should be verified and proved by a *confirmatory* (*confirmation)* method with a higher reliability.
- *Detection* is that the analytical signal was received (and can be tentatively identified).
- *Structure elucidation* relates to new compounds synthesized or isolated from natural samples.
- *Recognition* mainly refers to results of the use of a computer algorithm for *pattern recognition*.

## 1.3 General Principles for Identification

Based on the collective experience of many generations of analytical scientists which is expressed in the literature, general principles for chemical identifications can be formulated as follows.

*An analyte is considered to be unambiguously identified as the compound A when*

- *Physical, chemical, and biological properties of an analyte and A are identical*
- *Those of an analyte and all other compounds are different*[1]

*If the properties of an analyte and compounds A, B, C and so on are not differentiated, there is a case of ambiguous/group identification.*

*If a sample contains the compound A specific for the kind of the sample or the group of compounds* $A_1$, $A_2$, $A_3$, …, $A_n$ *in specific ratios, a sample can be identified/classified as referring to the particular specimen, material, product, food, pollution, living organism, and so on.*

There are two types of properties required for identification. First, there are qualitative features/characteristics which are the chemical properties, such as color/spot reactions, gas evolution, and precipitation. This is the field of classical qualitative analysis (e.g., see [22, 23]). In modern laboratories, qualitative reactions are

[1]Different properties of chemical substances are usually correlated. So mismatch in one property for a pair of compounds will lead to a difference in plenty of other properties. On the contrary, the match in a few properties (but not the only one) between an analyte and the compound A will probably result in (a) matching all others, and (b) difference from those of other compounds, followed by (c) reliable identification of an analyte as A.

For negative identification of a target, a mismatch rather than a match in properties should be proved.

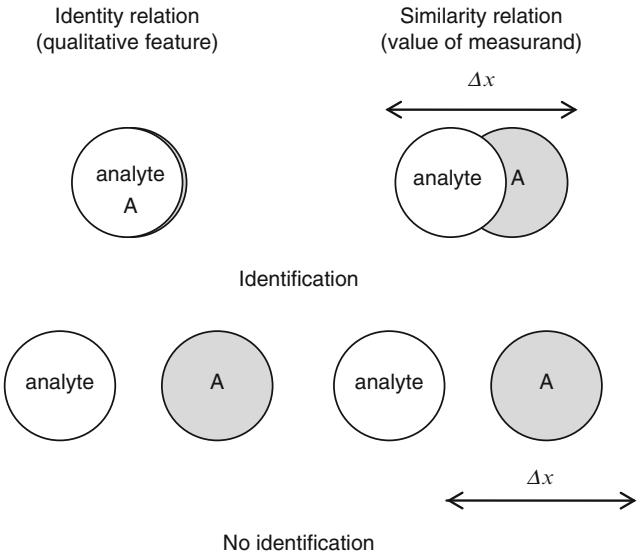



**Fig. 1.2** Chemical identification as comparing chemical and physical properties which are features and quantities respectively. Identity of features and similarity in values of the quantity observed/measured by the analytical chemist lead to identification of an analyte as the compound A (*circles*). Mismatching features or a significant difference in values imply that the identification is not achieved

often performed using paper strips, indicator papers, powders and tubes, their kits, and so on [24].

Classical qualitative analysis is related to determination of chemical elements. As for organic analysis, qualitative reactions are insufficiently specific for determination of most individual organic substances, many of which possess rather similar properties. Therefore, chemical test methods are mainly effective for recognizing classes of compounds related to various heteroatomic groups [25]. This is true in many respects for biological test methods for determining chemical compounds [26].

The principle of identification using these methods is determining an identity of qualitative/discrete features (Fig. 1.2). If such features are indistinguishable, it is appropriate to change to methods and techniques for identification based on comparing physical properties which are values of measured continuous[2] quantities (Fig. 1.2)

Before the appearance of spectral analysis and chromatography, chemists isolated pure substances and obtained quantitative measures such as density, boiling and melting point, refractive index, and so on to identify those substances [25]. In such experiments, an amount of substance of at least 1–10 mg was required. With

[2]Consideration of physical quantities as continuous ones is an approximation ignoring the discrete structure of matter and quantum effects. So they can be more properly named "quasi-continuous quantities".

the progress of analytical techniques, this amount had been steadily decreased. Reliable identification of at most a few nanograms of a complex organic/bioorganic compound is a routine procedure when ion masses, corresponding peak intensities, and retention parameters are measured in chromatography mass spectrometry (Chaps. 5 and 7).

When discussing the role of measurements in performing chemical identification, it should be noted that any value of measurand has some error/uncertainty [27]. In other words, replicated values of measured quantities diverge from each other. Therefore, the statement is true that similarity of values rather than their identity is essential for identification based on comparing physical properties (Fig. 1.2).

However, the concept of identity with regard to a measurand can be also applicable.

1. Rounded values can be taken into account: for example, $m/z$ values in low-resolution mass spectrometry are integral/discrete ones (Fig. 1.3). Their identity is one of the criteria for identification (Chap. 5).
2. A set of any quantities is divided into ranges which can be considered as different discrete features. If the value of measurand for an analyte falls within the particular reference value range $\Delta x$ specific for the compound A, it means (a) identity/matching of features of an analyte and the compound A, and (b) the possibility for identifying an analyte as A (Fig. 1.2). Here, different value ranges of measurands correspond to criteria for identification of various compounds (see below).

The main physical quantities in identification procedures based on spectrometry and chromatography are wavelengths, frequencies, masses, and times. The second

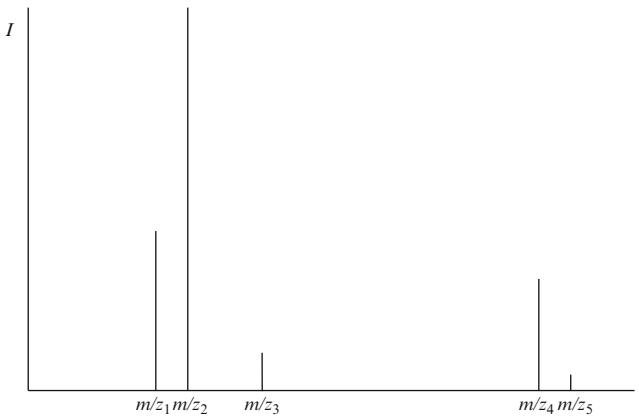



**Fig. 1.3** Typical mass spectrum. The quantities measured for identification and quantitative determination are mass-to-charge ratios ($m/z$) of ions and ion currents/counts/abundances. For most ions of low molecules, $z = 1$. The ion masses $m$ are measured in Da. Integral masses are called mass numbers. Ion abundances are commonly expressed as relative intensities of mass peaks ($I$, %)

dimension of corresponding spectra or chromatograms is the intensity of signals/peaks (Fig. 1.3). Here, intensity ratios of spectral peaks may be sufficiently specific to characterize analytes.

Recently, the concept of *identification point* was introduced [11]. One point is one property, i.e., one feature or one value of a measurand (one value ratio) to characterize an analyte. There should be several identification points selected for reliable identification (see Chap. 5).

## 1.4 Components of Identification

In the procedures under consideration, there are something that is identified (analyte), somebody who identifies (analyst), and many things (techniques and methods, reference data and expert systems, etc.) needed for identification itself. Identification as the system consists of those elements (Fig. 1.4). Correspondingly, quality assurance (Chap. 9) should be provided at the level of both the system and its elements.

Analytes as targets for identification are core components. The nature and origin of analyte and matrix/sample predetermine the choice of analytical techniques and approaches to implement qualitative determination (Chap. 2). It also obvious that there is no chemical analysis without suitable instruments (Chap. 2) and particular methods, whether the latter are standard (Chap. 5) or *ad hoc* (Chap. 7) ones. Methods are used as guidance for carrying out analytical experiments.

A growing role is played by data/information, software, computers, data systems, and global networks (Chaps. 6 and 7). Computers with appropriate software control laboratory instruments, and are used for searching and processing various data. Chemical databases containing records on features, structures, and properties of compounds, especially spectral libraries, are essential in unknown analysis, as defined in Sect. 1.5.1 below. This book emphasizes the value of special/prior data that describe the origin and use of different substances and their occurrence (Chap. 6). Rare compounds, as compared with abundant ones, can be excluded from consideration when solving most analytical problems.

Data obtained in analytical experiments and prior information are processed using statistical and expert programs. Statistical methods (Chap. 3) are valuable for qualitative as well as quantitative analysis, because they are used not only in standard operations but also intended for expression and estimation of identification reliability/error (Chap. 4). Knowledge expressed in different forms, ideally as a component of computer expert systems (Chap. 7), is also one of the attributes of identification procedures.

The main role in identification belongs to analysts themselves who choose analytical techniques/methods, set up identification hypotheses (Chap. 3), establish criteria for identification, and make the eventual decision on the nature of analytes based on those criteria. Corresponding reference materials are aimed at confirmation of identification results (Chaps. 5, 7–9). It is good practice when the analyst's

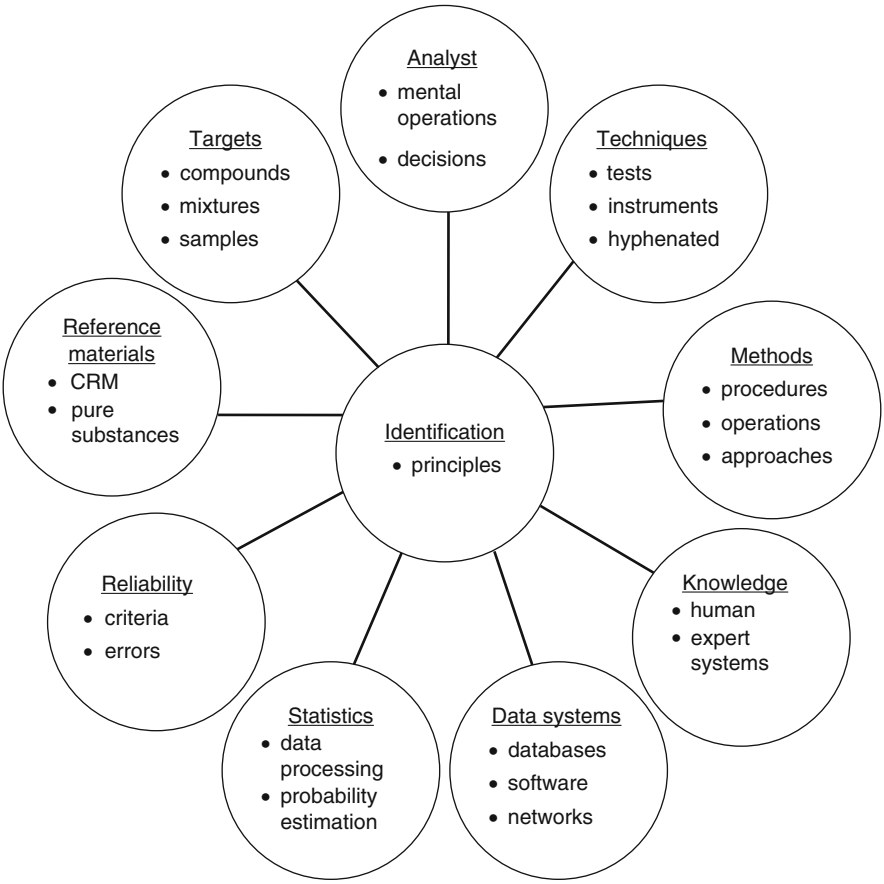



**Fig. 1.4** Individual components of identification procedures

decision is supported by estimates of identification reliability. As a result of (a) personal attitudes and (b) insufficient skill, experience, and responsibility of an analyst, a result of identification including its reliability may (a) be personally biased, and (b) contain human mistakes (e.g., see [9]).

## 1.5 Types and Objects of Identification

### *1.5.1 Main Classification*

It may be clear or not clear what is to be analyzed, and correspondingly there are two types of chemical analysis:

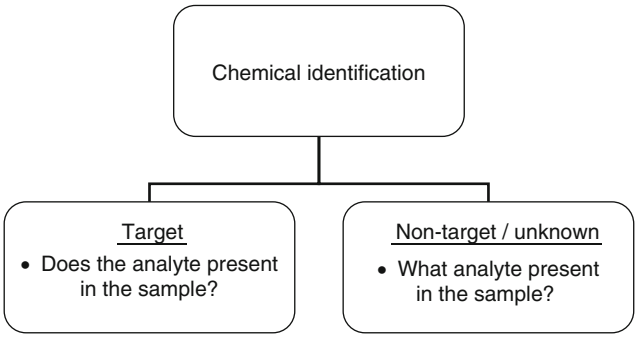



**Fig. 1.5** Main classification of identification procedures

- *Target* analysis is a determination of analytes specified before performing analytical procedures.
- *Non-target*/*unknown* analysis is a determination of analytes unknown to a chemist (but not necessarily absolutely new compounds) before analyzing corresponding samples.

Correspondingly, there are two types of chemical identification. The difference between them is shown in Fig. 1.5.

In the two cases, approaches to identification are not the same. Target analytes, which are often regulated chemicals, are known to an analyst. So he or she can use standard/validated methods for their determination, including standard operations for identification. There are two possible results of target identification, which are yes or no responses. Either one may be true or false (Chap. 4). Confirmation of identification results may be required (Chap. 5).

Identification procedures of the second type are far more challenging to implement. General protocols for this kind of analysis are not only absent but also very difficult to develop due to a profusion of possible analytes and approaches for analyzing them. Unknown identification is initially based both on several different analytical techniques and on various data (Chaps. 6 and 7). The skill and experience of analysts engaged in non-target analyses are crucial for obtaining unambiguous true results.

For discussion of some terminological differences between *unknown* and *non-target*, see Sect. 7.1.

### *1.5.2 Subtypes of Identification*

In this subsection, classification of sorts of identification is further considered.

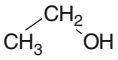
The widespread identification subtype is identification of *individual* compounds. Below, this sort of identification is supposed by default unless otherwise specified. However, it is only one of the possible types of identification (see Table 1.1).

**Table 1.1** Types of identification exemplified by ethanol **1.1**

| Classification unit | Example | Type of identification |
|---|---|---|
| Element | C, H, O | Element determination |
| Atomic group | $CH_3$, $CH_2$, OH | Structure elucidation |
| Individual compound | $CH_3CH_2OH$ | Individual identification |
| Group of compounds | Alcohols $C_1$-$C_4$ | Group identification |
| Narrow class of compounds | Aliphatic alcohols | Group identification |
| Wide class of compounds | Organic oxygen compounds | Group identification |
| Substance, chemical, reagent, material | Ethanol standard, 10% aqueous solution | Qualitative analysis II[a] |
| Product, commodity | Ethanol–gasoline mixture as biofuel | Qualitative analysis II[a] |

[a]In this or similar cases, the qualitative procedure may be equivalent to common tests for authenticity which include (a) composition analysis, (b) contamination/impurity detection and identification, (c) fragrances and odor identification, and so on

Considering that chemical identification is essentially chemical classification, the identification subtype is concluded to be predetermined by the type of the classification unit (Table 1.1).



**1.1**

If the classification unit is a *group*/*class* of compounds, there is one or other subtype of group identification. Examples are aliphatic alcohols or oxygen compounds as a whole (see Table 1.1).

It is also appropriate to differentiate between *unambiguous* and *ambiguous* identification. The first is the exact assignment of the analyte to the corresponding classification unit, e.g., identification of ethanol as "ethanol". Ambiguous identification means that only a wider class is determined, involving also other classified groups at a necessary level. For the case of ethanol (Table 1.1), recognition of this compound as "aliphatic alcohol" is certainly ambiguous, because this classification unit covers also methanol, *n*-propanol, and the plethora of other alcohols. So group identification of individual chemical compounds is always ambiguous. As such, identification is rather "not fully accurate" than false; it may be valid when solving some analytical problems.

Group identification of compound mixtures may be also unambiguous and ambiguous. The first case can be observed when there are no other mixtures within this class (with this name). In the second one, there are two or more objects of the same name.

In the following situations, only group identification is possible or reasonable:

- The properties of compounds are very similar. In analytical conditions, they can not be separated and characterized by their individual features. Enantiomer pairs in achiral media are very good examples of this.


**Fig. 1.6** Objects for individual, group, and sample identification. The vitamin group can be identified in the middle sample. The latter is also classified as a brand of multivitamin product produces by some pharmaceutical company

- Identification of many individual components of mixtures is possible but laborious and expensive. Furthermore, the mixtures are used as a whole. Here, natural mixtures of organic compounds such as petroleum fractions are good examples.

In qualitative analysis II (Chap. 8), identification of samples of compounds and their groups is carried out. This is essential in industrial, custom, environmental, and other controls performed using techniques of chemical analysis. Identification II includes also chemotaxonomy (chemosystematics) where living organisms, e.g., bacteria or plants, are classified/identified according to a similarity in their biochemical compositions. The objects mentioned above are differentiated by

- Specific/characteristic compounds and relationships between their amounts or
- Fingerprinting analytical signals in corresponding spectra or chromatograms

Different subtypes of chemical identification are illustrated in Fig. 1.6.

### *1.5.3 Identifiers*

An analytical chemist recognizes chemical compounds not only in samples but also in labels, documents, databases, and so on. For this purpose, identifiers are used.

An unambiguous chemical identifier is a unique set of symbols such as letters, numbers, lines, characters, and so on attributed to a chemical element, compound, or substance for their unambiguous recognition in records, texts, and data systems.

Among these are:

- Systematic names
- Registration numbers (RN) in data systems
- Formulas, first of all structural ones

- Line identifiers to represent two- or three-dimensional (2D or 3D) structural formulas

Such identifiers are shown in Table 1.2.

There are two basic systems of chemical nomenclature, which are IUPAC [30, 31] and CAS [30, 32] naming, both being widespread. For many chemical species, there is also a profusion of other names: (a) traditional, trivial, or semisystematic, and (b) trade/brand ones. The latter mainly specify substances which are formulations produced by chemical companies rather than absolutely pure individual compounds as something abstract. Examples of trade and related names of Carbendazim (Table 1.2) are: Carbendazole, Mecarzole, Carbendazime, Carbendazol, Bavistin, Thicoper, Derosal, Funaben, and so on [28, 33]. These are name synonyms that can be found in many chemical databases. A trade name may be an ambiguous identifier.

The most known registration numbers (the numeric identifiers) are connected to the Chemical Abstract Service data system [34]. A CAS RN contains up to 10 digits, divided by two hyphens into three parts (see Tables 1.2 and 1.3). This number designates only one substance, not necessarily the individual compound. The mixture or group of isomers (see Table 1.3) or even non-isomeric compounds may have the unique RN. This reflects a possibility of ambiguousness of identification results; see Example 1.1.

**Table 1.2** Unambiguous identifier for the pesticide Carbendazim [28, 29]

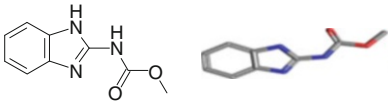
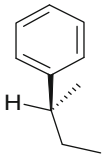
| Type | Identifier |
|---|---|
| IUPAC Name | Methyl N-(1H-benzimidazol-2-yl)carbamate |
| CAS Name | Methyl 1*H*-benzimidazol-2-ylcarbamate |
| CAS RN | 10605-21-7 |
| Structural formulas |  |
| SMILES | COC(=O)NC1=NC2=CC=CC=C2N1 |
| InChI | InChI=1S/C9H9N3O2/c1-14-9(13)12-8-10-6-4-2-3-5-7(6)11-8/h2-5H,1H3,(H2,10,11,12,13) |

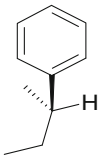
**Table 1.3** Identifiers for *sec*-butylbenzenes $C_{10}H_{14}$

| # | CAS RN | Semisystematic name | IUPAC name | Structure |
|---|---|---|---|---|
| 1 | 5787-29-1 | (*R*)-*sec*-Butylbenzene | [(2*R*-Butan-2-yl]benzene | **1.2** |
| 2 | 5787-28-0 | (*S*)-*sec*-Butylbenzene | [(2*S*-Butan-2-yl]benzene | **1.3** |
| 3 | 36383-15-0 | (±)-*sec*-Butylbenzene | [(2*RS*-Butan-2-yl]benzene | **1.2** + **1.3** |
| 4 | 135-98-8 | *sec*-Butylbenzene | Butan-2-ylbenzene | **1.2** or **1.3** or **1.2** + **1.3** |

**Example 1.1.** Identification of each enantiomer, **1.2** or **1.3**, as (*R*)- or (*S*)-sec-butylbenzene respectively is the unambiguous conclusion. The proper recognition of the isomeric mixture of **1.2** and **1.3** as the racemic pair, (±)-sec-butylbenzene, is the unambiguous group identification. Conversely, assigning the general name of *sec*-butylbenzene to either enantiomer is an ambiguous group identification, because this classification unit corresponds to three outcomes of determination: the individual compound (a) **1.2** or (b) **1.3**, or (c) the mixture **1.2** + **1.3** (Table 1.3).




**1.2**




**1.3**

Many CAS RN refer to patented formulations, e.g., pharmaceutical and agrochemical compositions. In general, RN of this data system are assigned to chemical substances in a wide interpretation of this concept [34] (Fig. 1.7).

Structural formulas as identifiers are 2D or 3D dimensional ones (e.g., see Table 1.2). These can be created and modified by numerous molecular drawing programs. It should be noted that common molecular formulas (brutto formulas), e.g., $C_2H_6O$, are not unambiguous identifiers. That formula belongs to ethanol $CH_3CH_2OH$ and dimethyl ether $CH_3OCH_3$ as well.

For computer input, storage, and processing of chemical information, structural formulas are supplemented or substituted by line symbol identifiers such as SMILES (Simplified Molecular Input Line Entry Specification) [35, 36] and InChI (IUPAC International Chemical Identifier) [37] (e.g., see Table 1.2). Both identifiers express chemical structures in standard machine-readable formats. There are computer programs, e.g., the same molecule editors, for transformation of structures to line identifiers and for the reverse operation (see [36, 37]).

In contrast to CAS RN, (a) the line notations are freely usable, non-proprietary, and not assigned by the only organization, and (b) the corresponding structural information can also be human-readable. Generally, all the above identifiers have been entered in modern chemical databases and spectral libraries (Chaps. 6 and 7). Certainly, modern chemical data systems also have their own identifiers, codes, and notions. The latter refer to not only chemical entities but also different information useful for identification, e.g., spectral bands [38].

### *1.5.4 Known Chemical Substances*

Several universal chemical spaces, i.e., large sets of possible or available chemical compounds, may be discerned:

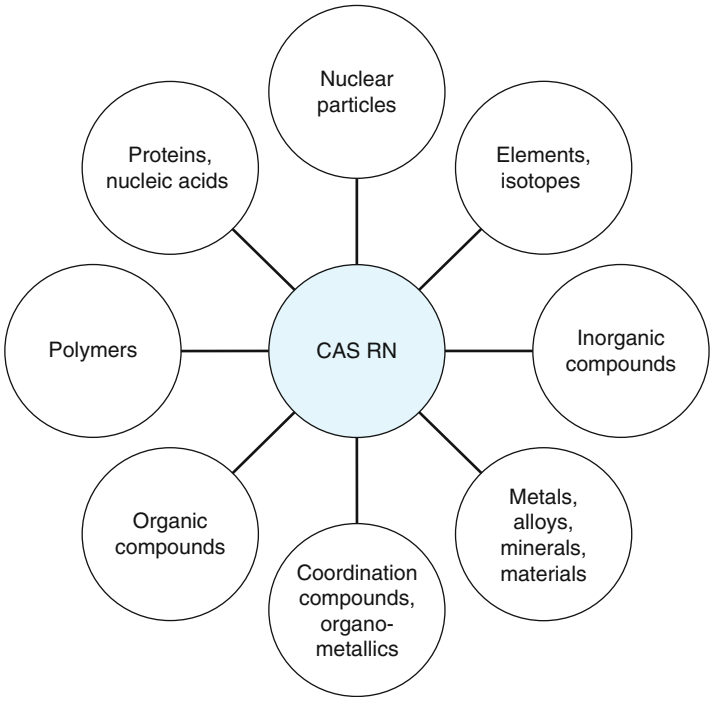



**Fig. 1.7** Groups of chemical entities according to CAS classification.

1. All possible stable compounds
2. Compounds which can be synthesized by chemists with the use of known reactions
3. Known compounds synthesized or isolated from natural sources
4. Widely occurring (abundant) chemical compounds

Here, each preceding set is far more numerous than the following one. Really, the dimensionality (the number of set elements, i.e., compounds) of set 1 is estimated as $10^{20}-10^{200}$ [39] or $10^{60}$ [40] compounds (drug-like ones with molecular masses $\leq$ 500 Da). The dimensionality of set 2 ("virtual organic chemistry space" [41]) is supposed to be between $10^{10}$ [42] and $10^{20}-10^{24}$ [41] compounds.

By definition, the number of possible answers obtained in identification does not exceed the number of known chemical substances, i.e., the dimensionality of set 3. By October 2009, more than 110 million substances had been recorded [43]. According to the simple binary classification, this set was composed of more than 50 million organic and inorganic substances, i.e., low-molecule ones, and more than 61 million of biosequences [43]. The overall number of CAS RN exceeds that of individual chemical compounds. At the same time, the author' observation is that most RN of low-molecule substances are connected to individual compounds.

The growth of known substances is demonstrated in Fig. 1.8. So far, most of them are proteins, nuclear acids, and other biosequences. They emerged in a great number at the end of the previous century and at the start of this one, due to the progress in genomics and proteomics. The rate of their growth slowed down by 2004–2005 (Fig. 1.8).

In recent years, the set of low-molecule substances has been rapidly increasing (Fig. 1.8). This is possibly due both to the growing synthetic/analytical capacities of chemists, and to the patenting of new formulations of "older" chemical compounds. So, a new intersection of the trend lines for low and high molecules can be predicted for the future (see Fig. 1.8).

Availability of the vast plethora of known/registered compounds does not imply that any of them could be detected in analysis of materials, foodstuffs, environmental and biological samples, and so on. All known compounds can be divided into abundant (popular, widespread, commonly occurring, i.e., belonging to the chemical space 4, see above) and rare ones. Rare compounds, which are about 99% of all the substances (see Sect. 6.3), can hardly be determined in the above matrices. Regular analytes are abundant compounds which are:

- Industrial chemicals and impurities in them, their metabolites and different conversion/decomposition products
- Solvents
- Components of fossil fuels
- Toxins and toxic compounds of a different origin
- Biocompounds, both low-molecular ones such as metabolites and high molecules, e.g., main nuclear acids, proteins, and carbohydrates
- Substances formed on storage and processing of wastes, and some others

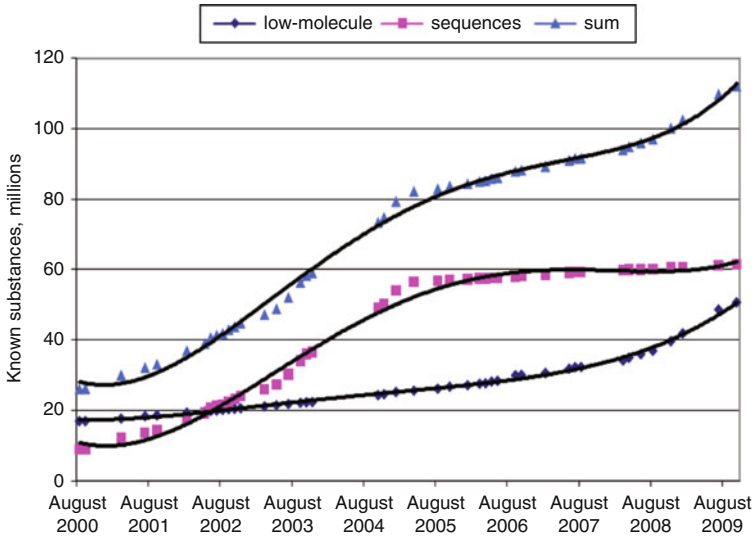



**Fig. 1.8** The overall number of substances registered in CAS

Most abundant compounds are regulated by international, national, and local organizations. Regulated compounds/substances have been entered in the CHEMLIST special database [44].

## 1.6 Principal Approaches to Identification

General procedures which are specially performed for identification or in which experimental analytical data are particularly used for that purpose can be named general approaches for identification. Four such approaches are given and commented on in Table 1.4.

As a whole, the identification reliability increases in series: 4<3<2<1 (for the numbers, see Table 1.4). Only the most reliable means [first, *comparison with reference data* obtained in conditions very similar to experimental ones (the version of the approach 2) and second, *co-analysis with authentic analytical standards* (the approach 1)] are used in target determinations by validated confirmatory methods (Chap. 5). In the literature, identification of high reliability may be referred to as a *definitive* one.

In non-target screening, all possible means are required, with the enforced use of only approaches 3 (*comparison with non-experimental reference data*) and 4 (*spectral interpretation*) in cases of very rare and new analytes where reference data/materials are not available. Identification of the intermediate reliability typical for screening procedures is called a *tentative*, *preliminary*, or *putative* one.

## 1.7 Metrological Issues

The foregoing shows that identification is based on observations and measurements. For organic and bioorganic compounds, only measurements provide reliable results of individual identification, with measurands being wavelengths, frequencies, ion masses, retention times, and so on. Metrology is connected to identification (in general, qualitative analysis) as well as to quantitative analysis [12, 19, 45, 46]. Metrological aspects of qualitative analytical procedures will be considered partly in this section and further in Chap. 9.

**Traceability.** This is among the basic concepts connecting metrology and chemical analysis. By definition, *metrological traceability* is

> property of a measurement result whereby the result can be related to a reference through a documented unbroken chain of calibrations, each contributing to the measurement uncertainty [47].

The concepts of *traceability chain* and *traceability to a measurement unit* are also used in this context [47].

**Table 1.4** Means of identification

| # | Means | Essence | Remarks |
|---|---|---|---|
| 1 | Co-analysis | Comparison of properties of an analyte and a authentic reference material in the same experiment (by simultaneous analysis) | Spiking the sample with the reference material of the analyte does not lead the significant distortion of the analytical signal. It is the strongest evidence for the presence of the analyte. Classical mixed probe in melting point measurements, co-chromatography and co-spectrometry (spectral mixing experiment) are the best examples |
| 2 | Comparison to experimental reference data | Reference data originate from the experiments; experimental conditions may be somewhat different in compared cases | The widespread cases are spectral libraries and databases on chromatography RI, where reference data were recorded on the same instruments in not the same conditions or on other instruments. The most reliable reference data originate from the same laboratory and analytical instrument and replicate/successive experiments (successive spectral scans, chromatographic runs) |
| 3 | Comparison to theoretical/ predicted reference data | Reference data are calculated/predicted based on theories, empirical regularities, correlations, models, and so on | The popular instances are estimating of GC RI (Sect. 7.2), simulation of NMR spectra (Sect. 7.6), and prediction of mass spectra of peptides (Sects. 4.4.2.3, 7.4.1.4, and 7.7.2). In the absence of valid experimental values, the data, if accurately estimated, can be used in screening-type procedures. The use of data originated from an experiment and corrected by one or other calculation method can be referred to both the above and this approach |
| 4 | Data interpretation | Conclusion about molecular composition and structure made from data (spectra) using special rules and algorithms | Individual spectral lines/peaks, more properly, their features including intensities and intensity ratios, and also complex spectral patterns are assigned to some atoms or atomic/functional groups. Software for interpreting spectral data is the core of computer expert systems. This approach is the principal one in structure elucidation of new compounds |

In chemistry, (a) a reference is an analytical standard (reference material) and (b) a measurand is an *amount of substance*, with the *mole* as the SI base unit. Also, it has been emphasized that

> . . . identity and amount. . . together constitute an "amount of substance"[48]

and

> the mole is, by definition, the amount of a specified substance. . .[45].

Indeed,

> when the mole is used, the elementary entities must be specified and may be atoms, molecules, ions, electrons, other particles, or specified groups of such particles [49].

It follows that the concept of traceability [45, 48, 50, 51] should be treated in the wide version of both qualitative and quantitative determinations. In other words, traceability claims are to be proposed to demonstrate "unbroken chains" covering both identity and amount of substance (see [48]).

The consequence of such "metrologization" of identification processes is the focus on a comparison between analytes and references (i.e., analytical standards) in traceability chains; this comparison can be considered as a variant of special "calibration for identity". Analytical chemists have traditionally considered matching properties of an analyte and a reference material in the same experimental conditions as the perfect evidence for identification (Sect. 1.6). In line with that, such procedures have also the highest metrological quality because analytes are "related to a reference" in the direct, shortest, and most certain way (Fig. 1.9).

Other means of identification shown in Fig. 1.9 are of worse metrological quality, in accordance with conclusions of analysts on their reliability (see above). The RM values obtained in different experimental conditions, including an analytical instrument from another manufacturer and/or an impure reference material, may be significantly biased to a basic value. This increases uncertainties related to that value, and therefore probabilities of identification errors. In the subsequent approach, theoretical values may be inaccurate, i.e., the theory or prediction model may be a weak link in "unbroken chain of comparisons," though such data as NMR spectra or GC retention indices are well-predictable. Lastly, spectral interpretation commonly based on rules, regularities, pattern recognition, etc. is as yet hardly traceable to any references in a direct way.

Thus, the traceability concept holds a central position in the metrology of qualitative analysis (see also [7, 12]). Demonstration of traceability calls for availability of reference materials (authentic pure compounds, matrix standards and so on; see Chap. 9). The concept (in versions of "sample traceability", "material traceability") is also applicable to qualitative analysis II where it is important, e.g., to find out an origin of food products under their characterization (Chap. 8).

**Nominal scale.** In metrology of qualitative analysis, not only quantitative but also other scales [52], ordinal and nominal, are of value (see Table 1.5). Identification itself can be represented as a measurement on a nominal/classification scale. Conditional points in the scale are specified by corresponding identifiers (Fig. 1.10). There may not only be CAS RN but also other numerical codes. Given that these are just codes, the only arithmetic operations allowed for this scale are equality and inequality. The operations are equivalent to considering the null and contrary hypotheses (Sect. 3.6).

The scale shown in Fig. 1.10 is advantageous for demonstrating some results and errors of qualitative analysis. As an identification result, a point on the nominal scale assigned to an analyte may be true or incorrect/false. In the case of assigning the name of the individual compound (e.g., (*R*)- or (*S*)-sec-butyl benzene; see Fig. 1.10 and also Table 1.3) to the compound group (butyl benzenes), it is a

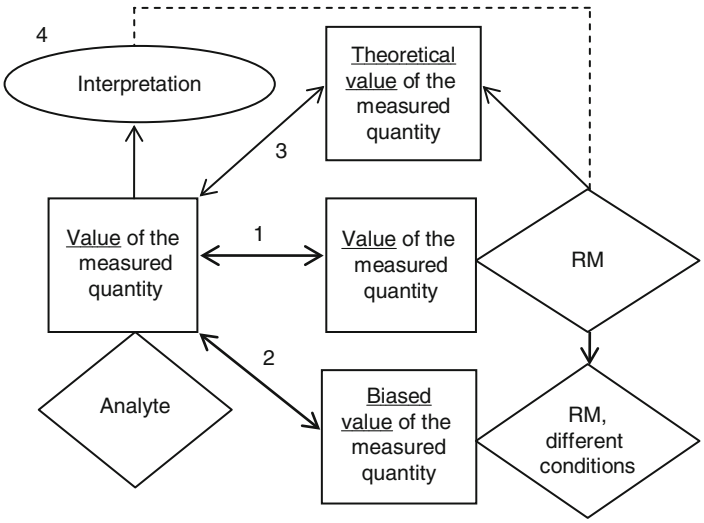



**Fig. 1.9** Traceability chains in identification of an analyte using general approaches. 1. Comparison of the value of the measurand between the analyte and the authentic RM in the same experiment ("calibration for identity"). 2. Comparison with the value of RM obtained in not the same experimental conditions or, in other words, with data of worse/unknown quality. 3. Comparison with the theoretical/predicted value of the RM quantity. 4. Interpretation of experimental data (spectra) obtained for analyte. Numbering coincides with that in Table 1.4

**Table 1.5** Different scales related to identification procedures

| Scale | Example |
|---|---|
| Quantitative (ratio, interval) | Time (RT, RRT), mass ($m/z$), wavelength, frequency, and so on |
| Ordinal | Rank of match factors in a hit list |
| Nominal | Chemical identifier |

false identification. The opposite outcome (i.e., attributing the group identifier to the individual substance) is group identification or individual "underidentification" ("not fully accurate" identification), which is a not very significant error.

A special sort of false results may be due to inaccurate records in chemical databases, including electronic spectral libraries:

- Incorrect chemical names
- Multiple registration of unique chemicals
- Confusion of identifiers (e.g., CAS RN) in the case of optic isomers
- Attribution of bases/acids to RN of corresponding salts and vice versa

These can be named "identifier errors".

So a nominal scale is another aspect of metrology of qualitative chemical analysis which emphasizes an importance of correct chemical identifiers in computer-assisted identification. A somewhat analogous nominal scale is used in medical data systems [53]. In qualitative analysis II, the concept of nominal scale

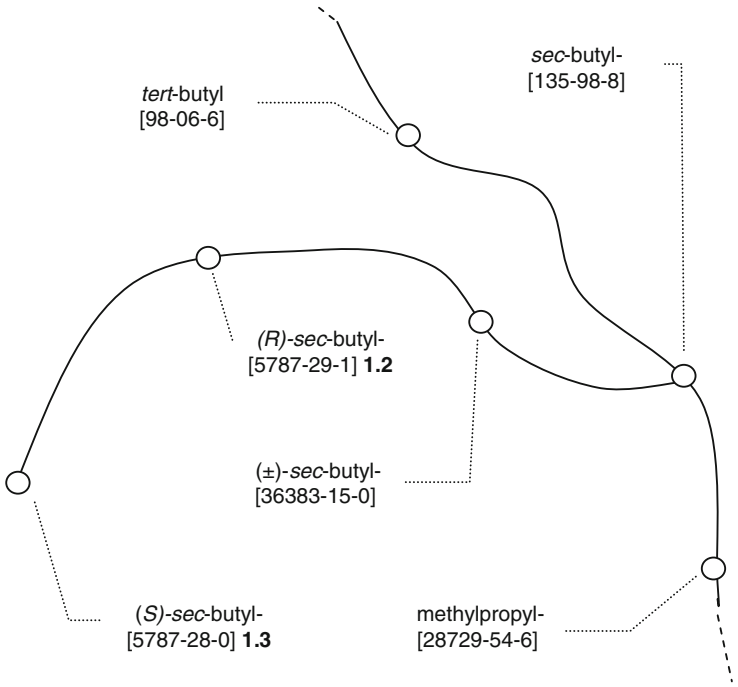



**Fig. 1.10** Representation of nominal scale for some isomeric alkyl benzenes $C_{10}H_{14}$. The scale itself and points in it are conditional. Points are specified by identifiers (here names and CAS RN) of individual compounds or groups of isomers (e.g., sec-butylbenzene). Identification is measurement on such a scale, i.e., assigning one of the identifiers to the analyte. The version of the nominal scale for all the isomers is given in the review [17]

is less applicable. The reason is that classification units such as "wine brand" or "pollution type" are not accurately defined as constant compositions. It is not clear how many points and what identifiers should be in such classification scales. Taking into account that it is also hard to make authentic RM for foodstuffs, pollutions, etc. (Sect. 8.1.4), analysis II seems not to be reliably supported by metrology.

## References

1. Currie LA (1995) Nomenclature in evaluation of analytical methods, including detection and quantification capabilities (IUPAC Recommendations 1995). Pure Appl Chem 67:1699–1723
2. Komar' NP (1955) Basics of qualitative chemical analysis. Book 1: Ionic equilibria (In Russian). Kharkov University Publisher, Kharkov
3. Ellison SLR, Gregory S, Hardcastle WA (1998) Quantifying uncertainty in qualitative analysis. Analyst 123:1155–1161
4. Hartstra J, Franke JP, De Zeeuw RA (2000) How to approach substance identification in qualitative bioanalysis. J Chromatogr B 739:125–137

5. Valcárcel M, Cárdenas S, Gallego M (2000) Qualitative analysis revisited. Crit Rev Anal Chem 30:345–361
6. Bethem R, Boison J, Gale J, Heller D, Lehotay S, Loo J, Musser S, Price P, Stein S (2003) Establishing the fitness for purpose of mass spectrometric methods. J Am Soc Mass Spectrom 14:528–541
7. Ríos A, Barceló D, Buydens L, Cárdenas S, Heydorn K, Karlberg B, Klemm K, Lendl B, Milman B, Neidhart B, Stephany R, Townshend A, Valcárcel M, Zschunke A (2003) Quality assurance of qualitative analysis in the framework of 'MEQUALAN' European project. Accred Qual Assur 8:68–77
8. De Zeeuw RA (2004) Substance identification: the weak link in analytical toxicology. J Chromatogr B 811:3–12
9. Lehotay SJ, Mastovska K, Amirav A, Fialkov AB, Martos PA, de Kok A, Fernández-Alba AR (2008) Identification and confirmation of chemical residues in food by chromatography-mass spectrometry and other techniques. Trends Anal Chem 27:1070–1090
10. Modern qualitative analysis (2005) Trends Anal Chem 24:461–555
11. Commission Decision 2002/657/EC, August 12, 2002, implementing Council Directive 96/23/EC concerning the performance of analytical methods and interpretation of results (2002) Off J Eur Commun L 221:8–36
12. Valcárcel M, Cárdenas S, Barceló D, Buydens L, Heydorn K, Karlberg B, Klemm K, Lendl B, Milman B, Neidhart B, Ríos A, Stephany R, Townshend A, Zschunke A (2002) Metrology of qualitative chemical analysis. Report EUR 20605. EC, Luxembourg
13. Milman BL, Konopelko LA (2000) Identification of chemical substances by testing and screening of hypotheses. I. General. Fresenius J Anal Chem 367:621–628
14. Milman BL, Kovrizhnych MA (2000) Identification of chemical substances by testing and screening of hypotheses. II. Determination of impurities in n-hexane and naphthalene. Fresenius J Anal Chem 367:629–634
15. Milman BL (2002) A Procedure for decreasing uncertainty in the identification of chemical compounds based on their literature citation and cocitation. Two case studies. Anal Chem 74:1484–1492
16. Mil'man BL, Konopel'ko LA (2004) Uncertainty of qualitative chemical analysis: general methodology and binary test methods. J Anal Chem 59:1128–1141
17. Milman BL (2005) Identification of chemical compounds. Trends Anal Chem 24:493–508
18. Milman BL (2005) Literature-based generation of hypotheses on chemical composition using database co-occurrence of chemical compounds. J Chem Inf Model 45:1153–1158
19. Milman BL (2008) Introduction to chemical identification (In Russian). VVM, Saint Petersburg
20. Princeton University WordNet. http://wordnetweb.princeton.edu/perl/webwn?s=identify. Accessed 11 Oct 2009
21. Valcárcel M, Cárdenas S, Simonet BM, Carrillo-Carrión C (2007) Principles of qualitative analysis in the chromatographic context. J Chromatogr A 1158:234–240
22. Kunze UR, Schwedt G (1996) Grundlagen der qualitativen und quantitativen Analyse (In German). Georg Thieme, Stuttgart
23. Otto M (2000) Analytische Chemie. Wiley-VCH, Weinheim
24. Zolotov YA, Ivanov VM, Amelin VG (2002) Chemical test methods of analysis. Elsevier, Amsterdam
25. Bentley KW (1963) Elucidation of organic structures by physical and chemical methods. Wiley, New York
26. Eggins BR (2002) Chemical sensors and biosensors. Wiley, Chichester
27. Guide to the expression of uncertainty in measurement (1993) ISO, Geneva
28. NIH PubChem. http://pubchem.ncbi.nlm.nih.gov. Accessed 11 October 2009
29. Compendium of Pesticide Common Names. http://www.alanwood.net/pesticides/index.html. Accessed 11 Oct 2009
30. Banks JE (1976) Naming organic compounds: a programmed introduction to organic chemistry. Saunders, Philadelphia PA
31. IUPAC Recommendations on organic and biochemical nomenclature, symbols and terminology etc. http://www.chem.qmul.ac.uk/iupac. Accessed 11 Oct 2009

32. Roeges NPG, De Moor MO. A simple guide to the nomenclature in organic chemistry. www.kahosl.be/site/index.php?p=/nl/downloads/1615/orgnompdf. Accessed 21 Oct 2010
33. ChemIndustry.com. http://www.chemindustry.com. Accessed 11 Oct 2009
34. CAS Registry. http://www.cas.org/expertise/cascontent/registry/regsys.html. Accessed 12 Oct 2009
35. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28:31–36
36. OpenSMILES http://www.opensmiles.org. Accessed 12 Oct 2009
37. The IUPAC international chemical identifier (InChI). http://www.iupac.org/inchi. Accessed 12 Oct 2009
38. Standard ASTM E204 – 98(2007) Standard practices for identification of material by infrared absorption spectroscopy, using the ASTM coded band and chemical classification index. http://www.astm.org/Standards/E204.htm. Accessed 25 April 2010
39. Van Deursen R, Reymond JJ (2007) Chemical space travel. ChemMedChem 2:636–640. doi:10.1002/cmdc.200700021
40. Dobson CM (2004) Chemical space and biology. Nature 432:824–828
41. Ertl P (2003) Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. J Chem Inf Comput Sci 43:374–380
42. BioSolveIT http://www.biosolveit.de/datasets. Accessed 11 Oct 2009
43. CAS Registry Number and Substance Counts. http://www.cas.org/cgi-bin/cas/regreport.pl. Accessed 11 Oct 2009
44. CAS Chemlist. http://www.cas.org/expertise/cascontent/regulated/index.html. Accessed 12 Oct 2009
45. King B (1997) Metrology and analytical chemistry: bridging the cultural gap. Metrologia 34:41–47
46. Kipphardt H, Matschat R, Panne U (2008) Metrology in chemistry – a rocky road. Microchim Acta 162:35–41
47. International vocabulary of metrology. Basic and general concepts and associated terms (VIM) (2008). Joint Committee for Guides in Metrology. http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2008.pdf. Accessed 25 April 2010
48. King B (2000) The practical realization of the traceability of chemical measurements standards. Accred Qual Assur 5:429–436
49. Unit of amount of substance (mole). http://www.bipm.org/en/si/base_units/mole.html. Accessed 25 April 2010
50. King B (2001) Meeting the measurement uncertainty and traceability requirements of ISO/IEC standard 17025 in chemical analysis. Fresenius J Anal Chem 371:714–720
51. EURACHEM/CITAC Guide: Traceability in Chemical Measurement (2003). http://www.measurementuncertainty.org/mu/EC_Trace_2003_print.pdf. Accessed 12 Oct 2009
52. Pfanzagl J (1971) Theory of Measurement. Physical-Verlag, Wursburg-Wien
53. Forsum U, Hallander HO, Kallner A, Karlsson D (2005) The impact of qualitative analysis in laboratory medicine. Trends Anal Chem 24:546–555

# Chapter 2
# Techniques and Methods of Identification

**Abstract** In this chapter, techniques and method of chemical analysis are discussed, with the focus on their potential for use in identification procedures. It is demonstrated that analytical techniques providing more information, in particular molecular spectrometry, are preferred for identification. Other techniques are just briefly considered, with for the exception of chromatography, whose combination with spectrometric techniques sharply increases possibilities and trueness of identification. As a whole, mass spectrometry is superior to other spectral techniques in such features as sensitivity, selectivity, generation possibility of molecular mass/formula, and combinability with chromatography. Different types of mass spectrometric instruments are outlined, with many performances tabulated. Experimental conditions for identification of volatile, non-volatile, and high-molecule compounds are discussed. Next, classification of chemical methodologies is given where screening and confirmatory methods are noted. Related procedures, sample treatment, and quantitative determination are also considered as ones affecting qualitative analysis.

## 2.1 General

Any analytical techniques can be used for the purpose of identification, though their potentialities are not the same.

An analytical process can be considered as a generation of information [1, 2]. In turn, unambiguous true identification, especially that of unknown compound (Sect. 1.5.1), needs a large amount of information. The reason is that the results of the procedure are very often complex chemical compounds. Their molecules differ between each other in elements and the number of their atoms, types of chemical bonds, configurations and conformations. The molecule complexity increases with the number and diversity of atoms, bonds, molecular configurations/conformations. Correspondingly, the amount of information required for the full description of complex molecules and differentiation between them is also increased. This is expressed, for example, in a length of the line notation (see Table 1.2). Thus analytical techniques providing more information (Table 2.1), such as those of

B.L. Milman, *Chemical Identification and its Quality Assurance*,
DOI 10.1007/978-3-642-15361-7_2, 

**Table 2.1** Amount of information generated by different techniques [1, 3]

| Technique | Potential information, bits |
|---|---|
| Spot test | 1 |
| Titrimetry | 100 |
| Emission spectral analysis | ≤2,000,000 |
| X-ray spectroscopy | ≤50,000 |
| Polarography | 800 |
| Gas chromatography | 8,000 |
| UV–Vis spectrometry | ≤1,000 |
| IR spectrometry | ~10,000 |
| Mass spectrometry | ~2,000,000 |

molecular spectrometry, are preferred for identification, other factors being equal. Using proper methods, higher selectivity is achieved, which also expressed in a larger number of identification points (Chap. 5). At the same time, some techniques generating a lot of information such as emission spectral analysis (Table 2.1) are not applicable in molecular analysis with its numerous identification problems.

Statistical data on articles related to identification prove the above conclusion (Fig. 2.1). Spectrometry (-scopy) first, then mass spectrometry and chromatography are the top techniques used for the purpose.

Another three techniques of molecular analysis are also prominent (Fig. 2.1). Electrophoresis is a very important separation technique somewhat analogous to chromatography. Fluorescence techniques are very widespread in biochemical analysis, and often do not provide unambiguous identification; rather, they are techniques for selective detection of certain compounds. The third common technique in this series, X-ray diffraction, is used in structure elucidation of new compounds and qualitative analysis II, e.g., for identification of minerals.

## 2.2 Elemental Analysis

Qualitative determination of elements/metals/ions is rarely named "identification of elements", though this is what it means. In contrast to molecular qualitative analysis, with a lot of organic compounds having very similar properties, elemental identification is relatively simple in implementation, because elements are not numerous and differ notably in their properties. Elemental analysis is well-described in literature (e.g., see [4, 5]. Here, related techniques are only listed:

- Qualitative reactions: spot/tube tests, other chemical test systems
- Flame test
- Polarography and related methods
- Photometry and spectrophotometry
- Atomic emission/absorption spectroscopy
- X-ray fluorescence analysis

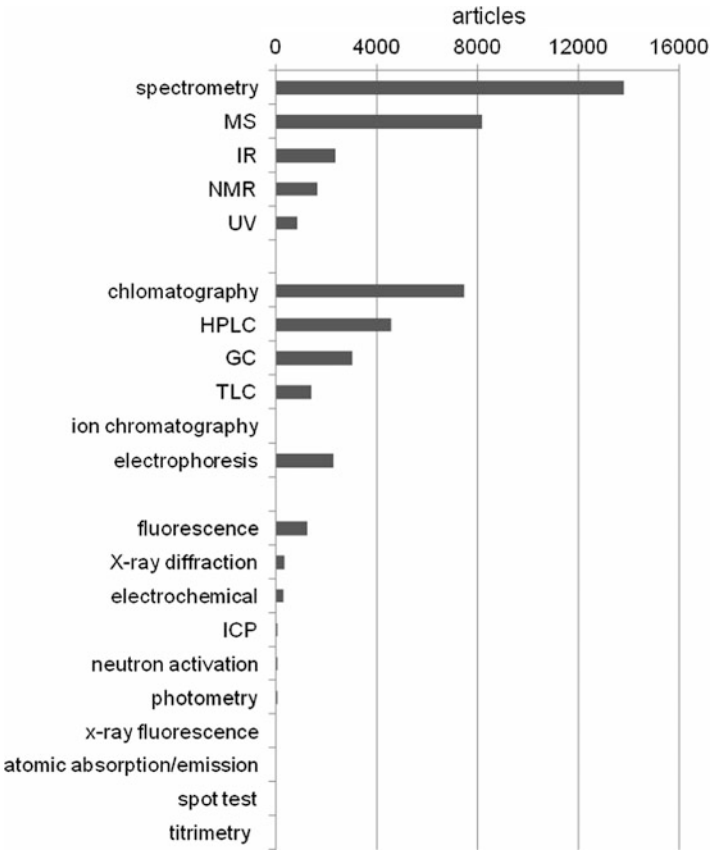



**Fig. 2.1** The number of scientific articles on identification performed by different techniques. The Google scholar engine was used for the search between 23 and 25 August 2009. In searches, articles with a combination of (**a**) the "identification" key word and (**b**) the corresponding technique name in the titles were retrieved

- Neutron activation analysis
- Ion chromatography
- Inductively coupled plasma mass spectrometry
- NMR and some others

These techniques mainly refer to inorganic analysis and can be applied to organic one as well. There are also special techniques of organic elemental analysis which use elemental analyzers (e.g., [6]).

Elemental analysis has been advanced in the version of speciation analysis, which may be a combination of the former with a molecular one [7]. Speciation is a determination of the particular chemical form, e.g., a charge/valence of a metal ion or a molecular/complex compound in which an element occurs in a sample. Analytical problems of the second kind are solved using techniques of molecular analysis.

Data obtained from elemental analysis can be required for identification/classification of samples themselves, i.e., in qualitative analysis II (Chap. 8).

## 2.3 Electrochemistry

Determination of inorganic and organic compounds by electroanalytical techniques includes identification of analytes. For this purpose, polarographic and voltammetric techniques [8] seems to be the most popular ones. The techniques are applicable for identification of electrochemically reducible (e.g., nitro, nitroso, and azo) compounds and oxidizable (aromatic amines, phenols) compounds (e.g., see [9–11]). Voltammetric peak potentials are quantities measured for identification. Two types of electrochemical devices, sensors [11, 12] and selective HPLC detectors [13], are of value for advanced chemical analysis.

## 2.4 X-ray Diffraction

X-ray diffraction is used for structure determination of inorganic and organic solids and identification of crystalline phases [14, 15]. In these types of analysis, diffraction theory and/or the comparison of the positions and intensities of the diffraction peaks to libraries of known crystalline materials are exploited. Multiple phases in a sample can be recognized. Identification of minerals in geological samples is the best known example of the use of the technique in qualitative analysis II [14].

## 2.5 Microanalytical Systems

One of the trends in analytical chemistry, miniaturization of techniques [16], is expressed in the appearance of, for example, numerous chemical test systems [17] and sensors [12]. They are very suitable for:

- Purposes of detection and screening
- Field and industrial analysis
- Beginning analysts
- Qualitative analysis II

and applicable in both elemental and molecular analysis, but not sufficiently selective to unambiguously identify most complex molecular species.

## 2.6 Biological Techniques for Chemical Analysis

Developing biosensor techniques combined with electrochemical devices can be used for screening of some chemicals [12, 18]. Methods based on bioassays, e.g., ones using enzymes, are specific to certain substances and sensitive, i.e., suitable for qualitative confirmation [19, 20]

## 2.7 Chromatography and Related Techniques

The main chromatography techniques and capillary electrophoresis (CE) are briefly described in Table 2.2. They are not only separation techniques but also complete analytical ones, because the instruments include detectors. The chromatographic signal is at least a two-dimensional one, as are most other analytical signals. One measurand is a retention/migration parameter. The second measurand is an overall intensity of a signal. For detection, just the fact of the presence of a signal itself, i.e., a yes response, is adequate (Chap. 4). In chromatography combined with spectrometry, signals are of complicated structure consisting of individual spectral peaks.

All the quantities are used for identification. The retention/migration parameters are purely chromatographic quantities for qualitative analysis. The range criteria for them are included in both analytical methods (retention times, Chap. 5) and non-target analysis (indices, Chap. 7). Co-chromatography is of prime value in confirmatory analysis (Sects. 5.2 and 5.4). Changes in polarity of stationary or mobile (LC, TLC) phase provides additional evidence for confirmation.

In identification procedures as well as in quantitative analysis, chromatographic resolution is the definitive parameter. As it increases (capillary columns > packed columns in GC, UPLC > HPLC > column chromatography), selectivity of determination also rises, and probability of false and inconclusive results diminishes.

Other identification capabilities depend on the detector type. First, in the case of a specific detector, a chromatographic signal itself may be diagnostic in terms of identification. Examples are nitrogen phosphorous and electron capture detectors in GC (Table 2.2), which indicate the presence of N and P and halogens respectively in an analyte molecule. Second, signals of the basic universal detector, mass spectrometer, and some other spectral tools are multiline spectra unambiguously characterizing many analytes (see below). Chromatographs in such hyphenated instruments can be rather considered as suitable inlet devices.

## 2.8 Molecular Spectrometry

Main spectrometric techniques usable in identification procedures are outlined in Table 2.3. Mass spectrometry provides more useful information (Table 2.1), and has more analytical applications and less limitations than other methods (see also Sect. 7.8). However, many laboratories use one or more other spectrometric techniques if possible for more reliable qualitative determinations.

**Table 2.2** Chromatographic and related techniques

| Technique | Principle | Measurand | Detector | Application |
|---|---|---|---|---|
| GC [21] | In the modern form, separation of a gaseous mixture into individual components (compounds) on passing a gas flow through a thin glass column, the inner walls of which are coated with a special nonvolatile liquid | Time, relative time, and index of retention | Universal (MS, flame ionization, katharometer), specific (nitrogen phosphorus, electron capture), multi-element (atomic emission) | Gas, volatile, and semi-volatile compounds[a] |
| LC (HPLC, UPLC[b]) [22, 23] | Separation of a liquid mixture into individual components on passing a liquid through a relatively thin steel column packed with particles or a porous layer of stationary phase | Time, relative time, and index of retention | UV–Vis and MS and also electrochemical, refractive index, fluorescence | Non-volatile compounds including almost all biochemicals |
| TLC [24] | Separation technique for a liquid mixture somewhat resembling LC, where the stationary phase is a layer of solid particles spread on a flat plate | Retention factor | Visual, detection reactions, fluorescence, UV–Vis, densitometers | Non-volatile compounds |
| Capillary electrophoresis [25, 26] | In the simplest form, separation of dissolved ionizable compounds in silica capillary, due to migration of their ions on application of high electric field | Migration time | UV and also fluorescence, electrochemical, MS | Nucleic acids and nucleotides, pharmaceuticals, proteins, various ionizable compounds |

[a]Compounds with boiling points up to approximately 350–400°C at atmospheric pressure. In these cases, vapor pressure of analytes is sufficient to analyze them by GC or GC–MS on sample heating up to not higher than about 300°C

[b]The new version of the HPLC instrument (Waters Corporation, USA), with higher performances for resolution and sensitivity [27]

**Table 2.3** Techniques of molecular spectrometry (-scopy) for identification

| Technique | Principle | Main applications | Limitations |
|---|---|---|---|
| UV–Vis [28, 29] | Measurements of light absorption at different wavelengths in ultraviolet (wavelengths 190–400 nm ) and visible (wavelengths 400–780 nm) part of the spectrum due to electronic excitation | Detection for HPLC | Spectra characterize chromophore types rather than individual compounds |
| IR[a] [28, 30] | Absorption measurement of IR radiation (wavenumbers from 13,000 to 10 $cm^{-1}$, wavelengths from 0.78 to 1,000 μm) due to vibration excitation | Structure elucidation (determination of functional groups), qualitative analysis II (polymers, plastics, resins, food, and so on) | Relatively low sensitivity (≥1–10 μg is commonly needed for spectral recording[b]); low compatibility of IR detector with GC and especially LC |
| NMR [28, 31] | Absorption of radiation in the radiofrequency range of the electromagnetic spectrum (hundreds of MHz) due to changes in the spin states of the atom nucleus | Structure elucidation of pure compounds, metabolomics, qualitative analysis II | Relatively low sensitivity (≥100 μg is commonly needed for $^{1}H$ spectral recording, with lesser amounts in a few hours acquisition time[c]); slow progress in LC–NMR |
| MS [28, 32, 33] | Measurement of mass (up to $10^{6}$ Da) and amount of ions (down to a few counts) generated from atoms/molecules of a substance | All kinds of chemical analysis | Lower applicability in direct analysis of unpolar high-molecular compounds |

[a]Raman spectroscopy [34], together with IR called vibrational spectroscopy, provides complementary information for the particular functional groups
[b]It was noted that from 5 to 20 ng was sufficient for recording spectra by GC–FTIR [35]
[c]Measuring limits and analysis times have been sharply reduced with the progress in NMR technique [36]

### *2.8.1 UV–Vis Spectroscopy*

Spectra of this kind rarely lead to unambiguous identification of individual compounds, and rather characterize classes of unsaturated organic compounds [28]. Numerous brand names of UV–Vis spectrometers are manufactured. Such "simple" spectrometers are not easily applicable to identification of substances in mixtures.

However, a liquid chromatograph is easily combined with a photo-diode array detector (DAD), which is the widespread analytical instrument for analyzing complex mixtures [29]. For the purpose of identification

- Reference value tables of $\lambda_{max}$, wavelengths at absorption maximums, and $\varepsilon$, molar absorptivities
- Full spectra entered in spectral collections, databases, and e-libraries (Sect. 7.3.2)

are commonly used

### 2.8.2 IR Spectroscopy

In classical IR spectroscopy, pure organic compounds were elucidated/identified by means of spectral interpretation using reference tables containing (a) specific wavenumber of absorption bands of different functional groups and (b) specific band absorption (strong, medium, or weak) [28, 30, 35]; see also references in Sect. 7.5. The modern state of the technique, typically using FT-IR instruments, is characterized by widespread application of:

- Electronic libraries of IR spectra (Sect. 7.5)
- NIR (13,000–4,000 $cm^{-1}$), with minimal or no sample preparation, fast determination, and reduced costs, for analysis of foodstuff, pharmaceuticals, chemicals, polymers, and so on, e.g., in qualitative analysis II (Chap. 8)

Substantial limitations of the technique become apparent when mixtures of compounds and their traces are analyzed (see Table 2.3 and also Sect. 7.8).

### 2.8.3 NMR Spectroscopy

The technique of high-resolution NMR is indispensable for structure elucidation of pure chemical compounds [28, 31]. Depending on the nucleus, the main types are $^{1}H$ and $^{13}C$ and also $^{15}N$, $^{17}O$, $^{19}F$, $^{29}Si$, and $^{31}P$ NMR. There may be 1D or 2D versions of the spectra; the role of 2D $^{1}H$-$^{13}C$ spectra used for identification has been growing.

In the classical approach to structure elucidation by spectral interpretation, reference tables of corresponding measurand values, chemical shifts and spin–spin coupling constants, accounting for signal multiplicity, are used. The spectral values are very sensitive to changes in molecular configurations and conformations. So NMR techniques are of the first value in the solution of stereochemical problems.

Now, there are two advanced approaches to identification/structure elucidation (Sect. 7.6). First, a computer spectral simulation is applicable. NMR spectra are easily predictable for hypothetical structures, and can be used for comparison with spectra recorded for analytes. Second, such comparisons can be performed if reference databases comprising of experimental NMR spectra are available. Both

approaches are rapidly developed, and well deserve more attention from the analytical chemist. Nevertheless, relatively low (a) sensitivity and (b) identification power in relation to individual components of complicated mixtures (see Table 2.3) still limit NMR applicability in qualitative analysis I. In contrast, NMR applications in qualitative analysis II seem to be in progress (Chap. 8).

### *2.8.4 Mass Spectrometry and Chromatography Mass Spectrometry*

As a whole, this technique is superior to other spectral ones in the combination of features such as sensitivity, selectivity, generation possibility of molecular mass/formula, and combinability with chromatography. Gas or liquid chromatographs as inlet devices to mass spectrometers separate complex mixtures of chemical compounds for their subsequent detection and recognition, with increased selectivity of combined techniques.

Thus, mass spectrometry and chromatography mass spectrometry have the highest potential for qualitative determination of complex organic compounds in complex mixtures/matrices. This advantage, together with the perfect capabilities of quantitative analysis (methods of isotope dilution), results in a rapid development and widespread application of the two techniques. It is clearly proved by the statistical data. Mass spectrometers held a 42% share of the global market for instruments for molecular analysis [37]. The number of mass spectrometers in the world grew from 34,000 in 1999 to more than 200,000 in 2005 [37].

Figure 2.2 shows a typical schematic diagram of mass spectrometers. The type of mass analyzer determines the main features of mass spectrometer and its "generic name." The most popular mass analyzers are specially specified in Table 2.4. The most popular combinations of mass analyzers with different ion sources and chromatographs as commercially manufactured mass spectrometers and chromatograph-mass spectrometers are placed in Table 2.5. The choice of the instrument for identification depends on properties of analytes and also data types necessary for identification. The latter are:

- Masses of the most important ions, i.e., molecular and analogous ones and intensities of the mass peaks
- Masses of individual fragment ions and intensities of their mass peaks
- Accurate ion masses and corresponding molecular formulas
- Full mass spectra, including tandem and high-resolution mass spectra

In mass spectrometry, the most important compound properties are volatility, polarity of molecule, and molecular mass. Based on these properties, all compounds are divided into three groups.

***Gases, volatile, and semi-volatile compounds.*** These compounds are less numerous than non-volatile ones, but until recent years have more often been analyzed by mass spectrometry.

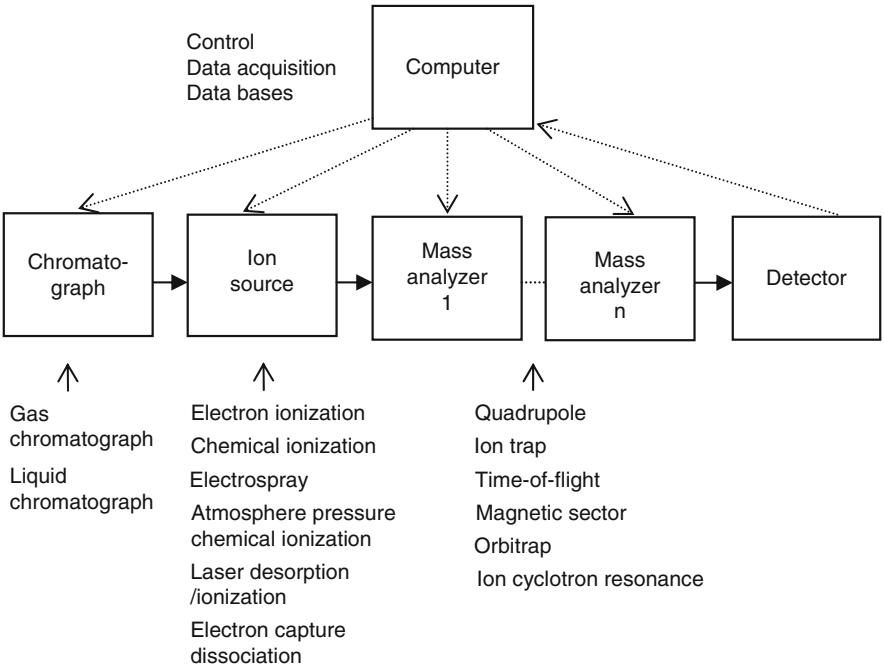



**Fig. 2.2** Schematic of mass spectrometer. Main methods of ionization and types of mass analyzer are specified. Tandem mass spectrometers (tandem-in-space instruments) are made of several analyzers. Chromatograph as separation instrument is sometimes substituted by electrophoresis device. Mass spectrometers can be certainly used without chromatographs, e.g., instruments with laser desorption/ionization (MALDI) and also secondary ion emission (SIMS)

Here, the combination GC–MS including EI (CI is less frequently applied) and quadrupole mass analyzer is the standard working instrument. In most cases, identification of volatile compounds is reliable. There are several reasons to form this conclusion. First, the number of the volatile substances is theoretically limited. A difference in properties between them is larger than among a huge number of non-volatile compounds having higher (and indefinitely high) molecular masses. Therefore, gases and volatile compounds are more easily differentiated for forthcoming identification.

A second group of reasons partly related to the first one are:

- The properties of the compounds under discussion are also well-studied, and corresponding values of measurand are well-reproduced
- Databases containing EI mass spectra and GC retention indices are commercially produced (Chap. 7)
- Many efficient analytical methods of quantitative and qualitative determination of this group of compounds have been developed and validated

**Table 2.4** Modern mass analyzers and their combinations[a] [38–40]

| Mass analyzer | Mass range | Mass accuracy | Fragments for identification | Price[b] | $MS^n$ |
|---|---|---|---|---|---|
| Quadrupole | + | + | EI: from ++ to +++ ESI: + | $ | no |
| Triple quadrupole | + | + | ESI: ++ | from $$ to $$$ | $MS^2$ |
| Ion trap[c] | + | + | ESI: to ++ | from $ to $$ | $MS^n$ |
| Time-of-flight[d] | +++ | up to +++ | MALDI: +[e] | $$ | no |
| Quadrupole-time-of-flight | from + to +++ | up to +++ | ESI and MALDI: ++ | $$$ | $MS^2$ |
| Orbitrap | + | +++ | from + to ++[f] | from $$$ to $$$$ | [g] $MS^n$ |
| Ion cyclotron resonance | + | +++[h] | from + to ++[f] | $$$$ | [g] $MS^n$ |

[a]General interpretation of symbols unless otherwise stated: +++ high, ++ medium, + low. For mass range and accuracy, + corresponds to a few thousand Da and from a few hundredth through several tenths of Da respectively. In the case of fragmentation: + a few fragments, ++ not very characteristic/reproducible fragmentation, +++ reproducible fragment spectra providing reliable identification

[b]Price grows up from $ (up to about $ 100,000) through $$$$ (not less than about one million dollars)

[c]The common ion trap is a quadrupole one. Now analyzers of the newer type, linear ion traps, with better performances, are also manufactured

[d]Unlimited mass range, high-speed scans, identification based on accurate molecular mass (accuracy about a few ppm)

[e]For recording mass spectrum of fragments in MALDI, the Q-ToF and ToF-ToF instruments have entered into practice

[f]Depends on the ionization technique and the tandem combination with different analyzers providing $MS^n$ capabilities

[g]Combinations with ion traps

[h]The highest mass accuracy

Reference comparisons to full (MS libraries) or partial (a few peak) spectra and GC retention parameters and co-chromatography (co-spectrometry) are general means for identification of these compounds (Chaps. 5 and 7).

***Non-volatile low-molecular compounds.*** In comparison to the above group, they have higher molecular mass and/or are more polar. Some polar compounds can be derivatized into corresponding volatile ones for further identification by GC–MS. If derivatization is impossible, inefficient or not achieved, various mass spectrometers and liquid chromatograph mass spectrometers (Tables 2.4 and 2.5) are used for the purpose.

ESI is the main ionization method. However, corresponding mass spectra are not rich in peaks of fragment ions. Tandem mass spectrometry ($MS^n$) is required where fragmentation is enhanced, due to collisions (collision activation) of analyte ions with the gas target within the special chamber. Integer-valued molecular masses may be insufficient for differentiating between heavier molecules of many

**Table 2.5** The most common mass spectrometers and chromatograph mass spectrometers for organic and bioorganic analysis

| Instrument | Application | Comment |
|---|---|---|
| Gas chromatograph–mass spectrometers | Volatile and semi-volatile organic compounds | • Rather simple, bench top, and inexpensive instruments<br>• Commonly EI and single quadrupole mass analyzers |
| Liquid chromatograph low-resolution mass spectrometers | Non-volatile low-molecular organic compounds | • Increasing role of tandem instruments, i.e. ion traps and triple quadrupoles<br>• ESI is the most popular ionization |
| Liquid chromatograph–high-resolution mass spectrometers | Non-volatile organic compounds including high-molecular bio compounds, proteomics | • Expensive instruments: time-of-flight, Orbitrap, and ion cyclotron resonance ones combined with other analyzers<br>• ESI and some other ionization techniques |
| Mass spectrometers for non-volatile compounds without chromatography | Bio compounds, proteomics, polymers | • Sample as a thin surface layer of organic compound in matrix<br>• MALDI and also SIMS to a lesser degree |

compounds. So HRMS leading to accurate molecular mass is required more than in the case of volatile substances. The combination of $HRMS^n$ is especially advantageous (Chap. 7).

Techniques and methods of identification using HPLC–MC are in progress, and approaches for many analytes have not yet been advanced. General challenges in identification are due to the following factors:

- ESI–$MS^n$ spectra and also HPLC retention parameters are not very reproducible
- Libraries of $MS^n$ spectra are far from being complete

That is why identification means directly based on the use of reference materials, starting with co-chromatography and co-mass spectrometry, are more important than in the case of volatile analytes.

***High-molecular compounds.*** These compounds, both bio substances and synthetic polymers, are non-volatile "by definition." High mass, up to $10^6$–$10^7$ Da, is the challenge because the corresponding mass range is only covered by ToF instruments. Therefore, mass spectrometric analysis of molecular fragments formed in the process of proteolysis (proteins) or pyrolysis (synthetic high molecules) is typical for the field. However, in the case of polar N-containing bio polymers (peptides, proteins and other), ESI produces multicharged ions $[M + nH]^{n+}$ that shorten the $m/z$ range necessary for comprehensive MS analysis of these compounds with the use of non-ToF tools. Therefore, this ionization technique became the leading one in bio mass spectrometry.

Determination of unpolar polymer compounds is a challenge for MS, because they are hardly ionized. Here, some analytical approaches are proposed in which MALDI, for example, is engaged [41].

Another concern is the production of reference materials providing the strongest evidence for identification (Chaps. 1, 8, 9). For example, there are no available standards for most proteins. So identification of many of them (Chap. 7) seems to be tentative.

## 2.9 Methods

Different techniques form the basis of various analytical methods. In turn, methods are classified according to techniques used for a proper determination of chemical compounds.

There is also another classification of chemical methodologies consisting of a hierarchy of analytical techniques, methods, procedures, and protocols [42]. In its description, the top level placed by *technique* lacks numerous details with regard to chemical operations. As lower hierarchy levels are reached, techniques become more specific. A *method* document consists of descriptions of individual *procedures*. At the bottom level, *protocol*, a complete description of all operations included to perform chemical analyses is represented [42]. Also, *standard operating procedure* occurs in the literature as a sort of synonym for a protocol.

Analytical methods are also divided on the basis of their reliability, i.e., a level of erroneous results. Now, many analysts emphasize that there are *screening* and *confirmatory* methods, with the latter being more reliable than the former [43] (see Table 2.6).

> Screening method means methods that are used to detect the presence of a substance or class of substances at the level of interest. These methods have the capability for a high sample throughput and are used to sift large numbers of samples for potential non-compliant results. They are specifically designed to avoid false compliant results [FN – Author]. Confirmatory method means methods that provide full or complementary information enabling the substance to be unequivocally identified and if necessary quantified at the level of interest [43].

Confirmatory methods are mainly based on mass spectrometry. Earlier, the similar pair of methods was named *routine* and *reference methods* (see [43]). The latter definition often occurs in the modern literature for specifying a method of a

**Table 2.6** Identification errors permitted in analytical methods

| Method | Probability | |
|---|---|---|
| | FN | FP |
| Screening | $<1:10^4$ | $<1:5$ |
| Confirmatory | $<1:10^4$ | $<1:10^4$ |

Proposals for residue analysis [44]

high (the highest) metrological quality. The synonyms for a reference method are a *definitive, absolute, or primary* method. However, these four terms are not often used only as characteristics of qualitative analysis (identification).

Identification using screening and confirmatory methods will be thoroughly treated in Chap. 5.

## 2.10 Preceding and Related Procedures

If full chemical analysis is performed, operations of qualitative determination are combined with those of quantitative analysis or those preceding/subsequent to the latter. Another preceding procedure or set of procedures is a sample treatment. Identification can be considered as being independent from the two different ones. However, the ways in which sample treatments and quantitative determinations are made have an influence on the result of identification. This will be very briefly outlined.

### *2.10.1 Sample Treatment*

In organic analysis, e.g., carried out by chromatography and mass spectrometry, most samples cannot be directly analyzed because

- The sample phase or chemical form of the analyte is not compatible with the analytical technique
- Non-target sample components and matrices themselves interfere with determination of target compounds and
- Targets present in sample in too low/high amounts

Therefore samples should be treated before analysis. Tens of different procedures for preparing samples for analysis are described [45]. The choice among them depends on whether a sample is gas, liquid, or solid (Fig. 2.3).

Gas samples are often analyzed directly by being injected into chromatographs and spectrometers. Analytes contained in liquids and solids should be isolated, concentrated/diluted, and possibly chemically transformed. Different methods of extraction, separation, clean-up, derivatization, and so on, consisting of many simpler operations, are required (Fig. 2.3). In these procedures, analytes may be lost or not separated from interfering substances; this is one of the sources of false results of detection and identification. Therefore, the presence of the analytes in the samples, rather than only in extracts, should be confirmed (see Sect. 5.3).

Some sample components and the sample itself can be identified, by IR spectroscopy for example, without numerous preparation operations.

In unknown analysis (Chap. 7), many operations of sample treatment, e.g., dissolution in various solvents, different methods of extraction, and even digestion

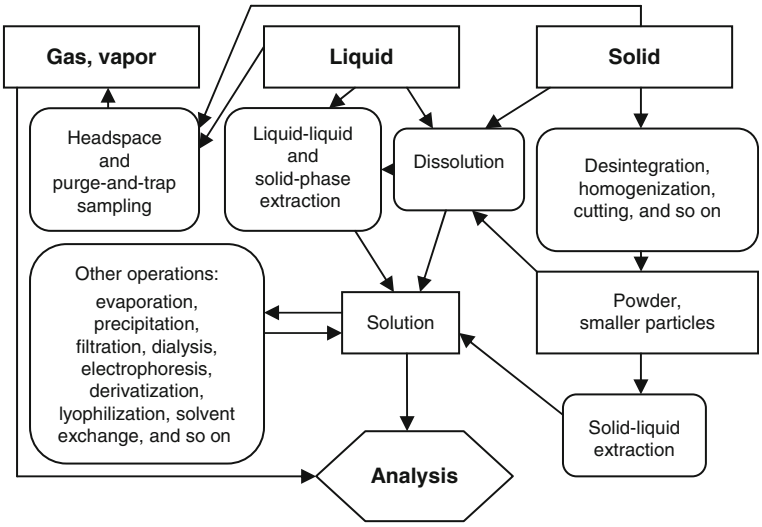



**Fig. 2.3** Flow chart for treating gas, liquid, and solid samples in chromatographic and mass spectrometric analysis. Only the most popular procedures and operations without multistep sequences for their implementation are shown

of a matrix itself, are often required to study a qualitative composition in great detail. The information obtained is also useful for qualitative analysis II, although fingerprinting of intact samples may be sufficient to differentiate between them (Chap. 9).

## *2.10.2 Quantitative Analysis*

Measurements for quantitative and qualitative analysis are or may be done simultaneously, but using not the same measurands; e.g., (a) the intensity of the basic spectral peak and (b) the relationship between intensities of several lines in the same spectrum are destined for (a) quantitative determination and (b) identification respectively.

Often, but not always, sensitivity and selectivity/specificity (and also a spectral resolution correlated with selectivity) are inversely dependent. This should be taken into account when it is necessary to choose optimal conditions for implementing identification combined with quantitative analysis within the same experiment. This can be exemplified by MS, where data can be recorded in the mode of SIM/SRM [46] or full scans which are typical for quadrupole instruments. In the first case, a few peaks are recorded, the maximum sensitivity is achieved, and quantitative determination is made. Nevertheless, the amount of information (the number of IP) may be insufficient for unambiguous identification. In the second situation, full spectra are obtained, which make it possible to reliably identify an analyte

(more selective qualitative determination), and this demands a larger amount of substance (analysis at not the best sensitivity).

Quantitative measurements are essential for estimating some limit performances which also are required in qualitative analysis (see Sect. 4.3). Also, quantitative determination may be part of procedures of qualitative analysis II, where a relationship between amounts of sample components is one of the quantities for characterization of a sample itself (Chap. 8).

## References


1. Eckschlager K, Danzer K (1994) Information theory in analytical chemistry. Wiley, New York
2. Thieme D, Müller RK (1997) Information theory and systematic toxicological analysis in "general unknown" poisoning cases. Fresenius J Anal Chem 358:785–792
3. Eckschlager K, Danzer K, Matherny M (1989) Informationstheorie in der Analytik. II: Mehrkomponentenanalyse. Fresenius Z Anal Chem 334:1–8
4. Kunze UR, Schwedt G (1996) Grundlagen der qualitativen und quantitativen Analyse. Georg Thieme, Stuttgart
5. Otto M (2000) Analytische Chemie. Wiley-VCH, Weinheim
6. Organic elemental analysis. PerkinElmer 2400 series II CHNS/O elemental analyzer. http://las.perkinelmer.com/content/RelatedMaterials/Brochures/BRO_2400SeriesIICHNSOelementalAnalyzer.pdf. Accessed 20 March 2010
7. Lund W (1990) Speciation analysis – why and how? Fresenius J Anal Chem 337:557–564
8. Monk PMS (2001) Fundamentals of electroanalytical chemistry. Wiley, Chichester
9. Smyth WF, Smyth MR (1987) Electrochemical analysis of organic pollutants. Pure Appl Chem 59:245–256
10. Filipiak M (2001) Electrochemical analysis of polyphenolic compounds. Anal Sci Suppl 17: i1667–i1670
11. Barek J, Moreira JC, Zima J (2005) Modern electrochemical methods for monitoring of chemical carcinogens. Sensors 5:148–158
12. Eggins BR (2002) Chemical sensors and biosensors. Wiley, Chichester
13. Flanagan RJ, Perrett D, Whelpton R (2005) Electrochemical detection in HPLC. Analysis of drugs and poisons. Royal Society of Chemistry, Cambridge
14. Brindley GW, Brown G (1984) Crystal structures of clay minerals and their X-ray identification. Mineralogical Society, London
15. Ladd MFC, Palmer RA (2003) Structure determination by X-ray crystallography. Kluwer, New York
16. Ríos A, Escarpa A, González MA, Crevillén AG (2006) Challenges of analytical microsystems. Trends Anal Chem 25:467–479
17. Zolotov YA, Ivanov VM, Amelin VG (2002) Chemical test methods of analysis. Elsevier, Amsterdam
18. Yogeswaran U, Chen SM (2008) A review on the electrochemical sensors and biosensors composed of as Sensing Material. Sensors 8:290–313
19. Farré M, Brix R, Barceló D (2005) Screening water for pollutants using biological techniques under European Union funding during the last 10 years. Trends Anal Chem 24:532–545
20. Guidelines on the Use of Mass Spectrometry (MS) for identification, confirmation and quantative determination of residues (2005) CAC/GL 56-2005. http://www.codexalimentarius.net/download/standards/10185/cxg_056e.pdf. Accessed 02 April 2010

21. Scott RPW (2003) Gas chromatography. http://www.library4science.com. Accessed 26 April 2010
22. Scott RPW (2003) Liquid chromatography. http://www.library4science.com. Accessed 26 April 2010
23. Snyder LR, Kirkland JJ, Dolan JW (2010) Introduction to modern liquid chromatography. Wiley, Hoboken
24. Sherma J, Fried B (2003) Handbook of thin-layer chromatography. Marcel Dekker, New York
25. Introduction to capillary electrophoresis. Beckman Coulter. http://www.beckmancoulter.com/literature/Bioresearch/360643-CEPrimer1.pdf. Accessed 26 April 2010
26. Weinberger R (2000) Practical capillary electrophoresis. Academic, San Diego CA
27. Swartz ME (2005) Ultra performance liquid chromatography (UPLC): An introduction. LC-GC 23: 8-14. http://chromatographyonline.findanalytichem.com/lcgc/data/articlestandard/lcgc/242005/164646/article.pdf. Accessed 26 April 2010
28. Silverstein RM, Bassler GC, Morill TC (1974) Spectrometric identification of organic compounds. Wiley, New York
29. Huber L (1989) Application of diode-array detection in high performance liquid chromatography. Hewlett-Packard Co. Publication number 12-5953-2330
30. Coates J (2000) Interpretation of infrared spectra, a practical approach. In: Meyers RA (ed) Encyclopedia of analytical chemistry. Wiley, Chichester
31. Hornak JP (2002) The basics of NMR. http://www.cis.rit.edu/htbooks/nmr. Accessed 29 April 2010.
32. McLafferty FW, Tureĉek F (1993) Interpretation of mass spectra. University Science Books, Sausalito CA
33. Pramanik BN, Ganguly AK, Gross ML (2002) Applied electrospray mass spectrometry. Marcel Dekker, New York
34. Ferraro JR, Nakamoto K, Brown CW (2003) Introductory Raman spectroscopy. Academic Press, San Diego CA
35. Hsu CPS (1997) Infrared spectroscopy. In: Settle FA (ed) Handbook of instrumental techniques for analytical chemistry. Prentice Hall. http://www.prenhall.com/settle/chapters/ch15.pdf. Accessed 29 April 2010
36. Raftery D (2004) High-throughput NMR spectroscopy. Anal Bioanal Chem 378:1403–1404
37. Mil'man BL, Konopel'ko LA (2006) Modern mass spectrometry: proportions of developments. Mass Spectrom (Moscow, in Russian) 3:271–276
38. Milman BL (2008) Introduction to chemical identification (In Russian). VVM, Saint Petersburg
39. Schuhmacher R, Sulyok M, Krska R (2008) Recent developments in the application of liquid chromatography–tandem mass spectrometry for the determination of organic residues and contaminants. Anal Bioanal Chem 390:253–256
40. Schaeffer-Reiss C (2009) A brief summary of the different types of mass spectrometers used in proteomics. In: Thompson JD, Schaeffer-Reiss C, Ueffing M (eds) Functional proteomics: methods and protocols. Humana Press, Totowa, NJ, http://www.springerlink.com/content/v43522nl40137640. Accessed 29 April 2010
41. Pasch H, Schrepp W (2003) MALDI-TOF mass spectrometry of synthetic polymers. Springer, Berlin
42. Currie LA (1995) Nomenclature in evaluation of analytical methods, including detection and quantification capabilities (IUPAC Recommendations 1995). Pure Appl Chem 67:1699–1723
43. Commission Decision 2002/657/EC, August 12, 2002, implementing Council Directive 96/23/EC concerning the performance of analytical methods and interpretation of results (2002) Off J Eur Commun L 221:8–36
44. Stephany RW (1997) How to assess the reliability of your residue identification? Establishment of Reference Methods: EU CRL workshop. Berlin, December 10–12, 1997
45. Mitra S (2003) Sample preparation techniques in analytical chemistry. Wiley, Hoboken NJ
46. Introduction to MS quantitation and modes of LC/MS monitoring. http://www.ionsource.com/tutorial/msquan/intro.htm. Accessed 29 April 2010.

# Chapter 3
# Probability, Statistics, and Related Methods

**Abstract** The probability/statistical methods used for identification purposes are briefly considered. The basic statement is that many phenomena and procedures included in qualitative analysis are of a probabilistic nature. The probability of yes/no responses in target detection is described by binomial distribution. Values of quantities required for identification, such as retention times in chromatography, wavelengths and frequencies in optical spectroscopy, masses in mass spectrometry, intensities (heights, areas) of any analytical signals, are considered as normally distributed (including *t*-distributed) ones over probabilities. Parameters of the distributions are used in calculations incorporated into procedures of detection and identification. Multivariate statistics connected with chemometrics is essential for classification/authentication of samples, i.e., qualitative analysis II. Bayesian statistics takes into account a prior probability that an analyte is present in a sample.

In the second part of this chapter, operations of setting up, testing, and screening of hypotheses as the core processes of qualitative analysis, are considered. The simplest are hypotheses for a detection operation, e.g., '$H_0$: an analyte is absent in the sample'. In identification, analogous hypotheses: '$H_0$: the analyte is compound A', and '$\overline{H}_0$: the analyte is not compound A' are set up and tested. Identification hypotheses are transformed into experimental and statistical ones to be accepted or rejected on the basis of corresponding criteria, both range/tolerance and statistical criteria. False acceptance or rejection of hypotheses leads to false positive/negative results of identification or detection, the probability of which can be estimated.

## 3.1 General

A conclusion on reliability of an identification result is often of a probabilistic nature. The reasons for this are the following.

B.L. Milman, *Chemical Identification and its Quality Assurance*,
DOI 10.1007/978-3-642-15361-7_3, © Springer-Verlag Berlin Heidelberg 2011

- Analytes may get into samples[1] under analysis in non-deterministic ways; the presence in itself of analytes in samples in amounts sufficient for detection can be considered as a random characteristic.
- Measurands used for identification, related to both (a) a position of an analytical signal in a scale of time, frequency, wavelength, mass, and so on, and (b) an intensity of an analytical signal are random variables, and their values are statistically distributed.
- Chemical compounds are similar in their properties within related compound groups. So there is a chance that experimental data obtained for the purpose of identification will lead to several compounds rather than one of them.

Therefore the use of probability/statistical methods [1, 2] in identification procedures is a natural decision for analysts, irrespective of what analytical techniques and methods are involved. However, statistical methods for a simple yes/no determination (Sect. 3.2), and complicated analytical techniques based on measurements (Sect. 3.3), may be not the same. Many such methods, mainly multivariate ones, have been developed within chemometrics [3–5].

Statistical methods are also used for testing hypotheses, which are the procedures adequately expressing the essence of the identification process (Sect. 3.6). Setting up and testing of identification hypotheses can be also considered as mental operations of the analyst, followed by him/her making a decision with regard to identification results.

## 3.2 Binary Responses of Qualitative Analysis

Answers to the question whether the particular substance presents in the sample are binary responses of the yes/no type. Given that the qualitative method (chemical test) is not absolutely reliable (all methods are such at low amounts of analytes), a replication of an analytical experiment results in a sample of positive (P) and negative (N) responses. The numbers of P or N outcomes of binary chemical test are random variables. The probability of obtaining these results is described by binomial distribution [1]. The example of this distribution is given in Fig. 3.1.

For the numbers of outcomes of a binary chemical test as random quantities, the corresponding statistical error expressed as the confidence intervals $\Delta_c$ for the proportion of the certain outcomes are given by equations as follows [6]:

$$\Delta_c(n_P/n) = z\sqrt{\frac{n_P/n \cdot (1 - n_P/n)}{n}} \tag{3.1}$$

[1]In this book, two similar but not the same meanings of *sample* occur. They are "a part of something to be tested" (this case) and "a subset of random values selected from a population" (statistical issues).

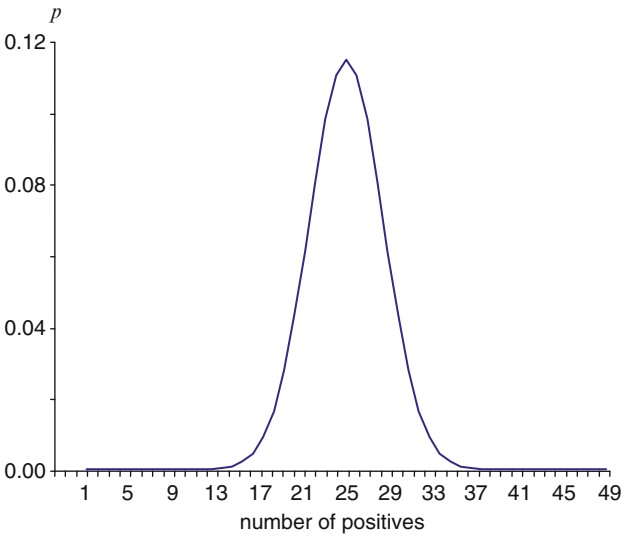



**Fig. 3.1** The probability of the certain number of P responses of qualitative test with 50 replicates, and the same chances of P and N outcomes. The distribution maximum (the mean value) is for half of replicates; this is natural for a method with the same rates for both outcomes

$$\Delta_c(n_N/n) = z\sqrt{\frac{n_N/n \cdot (1 - n_N/n)}{n}}, \tag{3.2}$$

where $n_P$ and $n_N$ are the numbers of $P$ and $N$ test results respectively; $n$ is the overall number of replicates, $n = n_P + n_N$; $z$ is the quantile of the normal distribution for the probability $p$ (1.64, 1.96, 2.58, and 3.29 for $p$ 0.90, 0.95, 0.99, and 0.999 respectively). Equations (3.1) and (3.2) will be used further for estimating trueness rates of qualitative methods (Sect. 4.2.1).

Binomial statistics can also be used in chemical informatics for evaluation of compound proportions which have particular properties significant for qualitative analysis. An abundance/rarity of compounds is the example of the property. For a full set of chemical entities, the mentioned calculation may be laborious. Instead, random sample methods are used. A random sample including $n$ entities is chosen. The number of compounds which have the particular property (also denoted as $n_P$) is counted. The $n_P/n$ ratio is the estimate for the proportion of such compounds in the initial set. The statistical error for the sampling is given by (3.1).

## 3.3 Distribution of Measured Quantities

Unlike qualitative reactions and related tests of target determination with their results as binary responses, analytical instrumental methods lead to values of continuous (or conditionally continuous) quantities/variables. The latter are

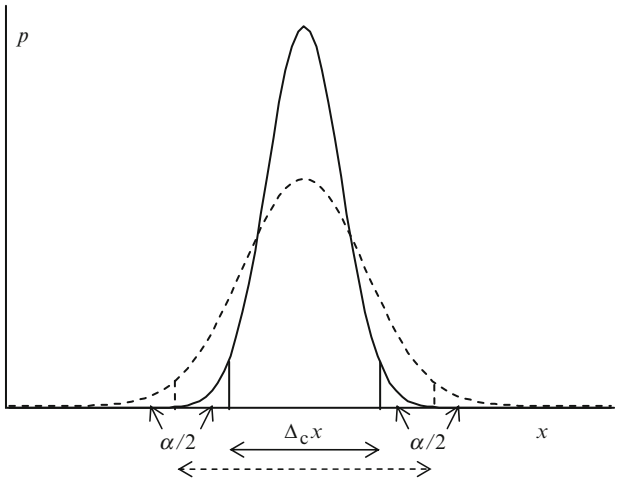



**Fig. 3.2** Normal and *t*-distributions of the quantity $x$ as the property of chemical compounds over probability of its observed values. The second one corresponds to a relatively small sample of values. The central intervals $\Delta_c x$ of both distributions are confidence ones. These can be also treated as distances between the $(\alpha/2)^{\text{th}}$ and $(1-\alpha/2)^{\text{th}}$ quantile points. There are $100(1-\alpha)\%$ values within the confidence intervals, where $\alpha$ is significance level. For $\alpha = 0.05$, there are 95% values of $x$ within the central ranges. The parameter $\alpha$ expresses the probability of very low/high values of the quantity, measured by the area under the distribution curve tails. If the measurand value is in the tail, it may belong to not this compound, with the error (FN) of $\alpha$

retention times in chromatography, wavelengths and frequencies in optical spectroscopy, masses in MS, and intensities (heights, areas) of any analytical signals (Chap. 2). An infinite number of experimental replicates in the same conditions provide a data population. It is commonly supposed that measurand values are normally distributed over their probabilities [1, 2] (Fig. 3.2). The measure of a value spread is here a population standard deviation $\sigma$.

In reality, a sample of values rather than their population is obtained from analytical experiment. Sample distribution is known as Student's *t*-distribution or simply the *t*-distribution [1–5] (Fig. 3.2). This is one from the family of normal distributions for finite samples.

For identification as well as chemical analysis as a whole, the following distribution parameters are calculated from a sample of values:

- Average $\overline{x}$
- Standard deviation[2] $s$
- Dispersion $s^2$
- Confidence interval $\Delta_c x$, and so on

[2]In the case of a large number of replicate measurements, the population standard deviation $\sigma$ can be estimated by the sample parameter $s$.

Properties of chemical compounds required during identification processes are sample distributions of corresponding measurands, or equivalently their mean values combined with deviation parameters ($\delta$ and others). For the purpose of identification, similarity or difference in properties is searched. Either may be significant or insignificant, which is decided according to results of statistical tests. The *t*-tests [1–5] are widespread (Sect. 3.6.5) even though they are criticized by metrological purists; this criticism seems to be disputable [7].

## 3.4 Multivariate Statistics and Chemometrics

Many quantities used in identification, such as spectra, can be considered as multivariate ones. For example, mass spectrum is *n*-variate quantity as a set of mass values $m_i$ and corresponding peak intensities $I_i$ , where $i = 1, 2, \ldots, n$. Thus, mass spectrum can be represented as the vector in multivariate space of mass values, or as a point in this space. The similarity in spectra corresponds to a negligible angle between vectors or a short distance between points (Sect. 4.4.2.1).

Multivariate distributions, their parameters, and corresponding statistical tests [1] are analogous to them for monovariate quantities. For example, Hotelling's $T^2$ statistics can be used instead of simple *t*-tests (Sect. 3.6.5).

Data obtained in the analysis of mixtures can be also represented as multivariate values. Here the univariate quantities are signal intensities $I_i$, e.g., heights or areas of chromatographic peaks of each component of the mixture. Multivariate statistical methods (see below) are used for analysis of such data. This mathematical approach is efficient for qualitative characterization of compound mixtures and samples which are analyzed as a whole under qualitative analysis II (Table 3.1 and Chap. 8).

Table 3.1 lists main multivariate statistical and chemometrical methods [3–5] together with examples of their applications. The use of these methods makes it possible to reduce the observed variates into a smaller number of components/factors (artificial variables). As a result, the statistical picture becomes far more informative and decisive for a data user, and applicable for subsequent mathematical processing.

A similarity of values of new variates between different objects (compounds, mixtures, samples) is used for identification/classification of them. Classes/groups of objects may be predetermined/supervised or unsupervised, i.e., established on performing statistical analysis (see Table 3.1). In supervised classification, training and evaluation data sets are first formed and explored to develop a classification method.

Different methods are illustrated by Fig. 3.3.

## 3.5 Bayesian Statistics

It is emphasized in this section that not only the appearance of an analytical signal and the distribution of its features but also the formation of an analyzed sample itself are related to probability. Analytes originate in a system under analysis from

**Table 3.1** Multivariate statistical and chemometrical methods in classification/authentication of chemical entities and systems [3–5]

| # | Method, procedure | Remarks | Examples of classification/ authentication |
|---|---|---|---|
| 1 | Data reduction/ processing | Conversion of raw data into a more applicable form including fewer dimensions (see 1.1 and 1.2) | |
| 1.1 | Principal component analysis (PCA) | Transformation (by looking for linear combinations) of many possibly correlated variates into a smaller number of uncorrelated variates named principal components | Drinks [8], genes/cells [9], saccharides [10], painting media [11], plants [12], polymeric materials [13], potatoes [14],wastewaters [15], |
| 1.2 | Factor analysis (FA) | Relates to PCA and is distinct from it in some initial assumptions and calculation details | Genes/cells [9], HPLC packings [16], painting media [11], polymeric materials [13], pollution sources [17], |
| 2 | Classification, discrimination | Group of supervised/ unsupervised classification methods/ algorithms (see 2.1 and 2.2) | |
| 2.1 | Supervised learning | Computer learning method for classification of the input objects from training data | |
| 2.1.1 | Discriminant analysis (DA) | Classifying a data set into predefined classes based on a training set; relates to PCA and FA in the use of linear combinations of variates | Bacteria [18], drinks [8], honey [19], potatoes [14], wines [20] |
| 2.1.2 | K-nearest neighbor (k-NN) | Classification of an object by assigning to the class most common amongst its *k* nearest neighbors (points in a multivariate space) | Saccharides [10], proteins [21], potatoes [14], plants [12], wines [20] |
| 2.1.3 | Soft independent modelling by class analogy (SIMCA) | Supervised classification which identifies samples as belonging to multiple/ overlapping classes | Drinks [8], saccharides [10], painting media [11], pharmaceutical solutions [22], wastewaters [15] |
| 2.2 | Unsupervised learning | Classification where initial object classes are unknown and training data are not used | |

(*continued*)

**Table 3.1** (continued)

| # | Method, procedure | Remarks | Examples of classification/ authentication |
|---|---|---|---|
| 2.2.1 | Cluster analysis | Sorting different objects into groups, with the degree of similarity between two objects being maximal if they belong to the same group | Drinks [8], electrophoretic spots [23], genes/cells [9], painting media [11], waters [24] |
| 2.1.4/ 2.2.2 | Artificial neural network (ANN) | Classification model simulating the structural and functional aspects of biological neural networks; may be supervised or unsupervised | Bacteria [25], electrophoretic spots [23], particles [26], plants [12], potatoes [14] |

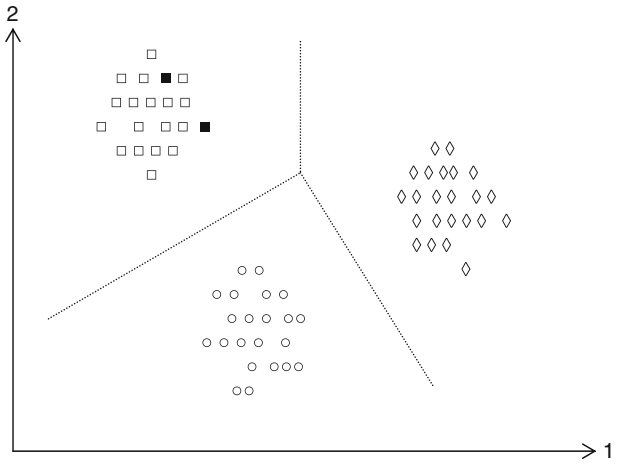



**Fig. 3.3** Classification of chemical entities, e.g., compounds into three groups/classes. Corresponding symbols are *open square*, *open diamond*, and *open circle*. Objects of distinct classes are different in values of two variables. The latter, e.g., two principal components, are obtained on processing initial analytical data. The object groups are predetermined, e.g., by the use of discriminant analysis or formed as the result of statistical analysis (cluster analysis). Dotted lines separate subspaces related to different classes. Unknown objects symbolized *closed square* are very similar in values to ones from upper left subset *open square* and therefore classified as belonging to this group

outside (environmental samples such as air, water, soil, sediment) or within as the result of industrial (materials) or natural (biosamples) processes. In general, it is very difficult to predict their presence in a sample in amounts sufficient for detection and identification. Nevertheless, one could try to express that in terms of random events.

An uncertainty in a priori knowledge about a chance for an analyte to present in a sample is related to the concept of a prior probability. The latter is the probability of an analyte being in a sample estimated before performing chemical analysis. By its nature, a prior probability is a subjective one because it is derived from an analyst's personal judgment about whether a particular compound is likely to be present, and reflects his/her opinion based on skill and past experience. However, this type of probability can be estimated using appropriate quantitative rates (Chap. 6).

Performing an analytical experiment changes the knowledge of the analyst about whether an analyte is present in the sample. The probability of detection and identification under consideration in the book is based on analytical data, and named a posteriori one. The latter also depends on corresponding prior probability.

The statistical method for estimation of probabilities incorporating prior ones is Bayesian statistics (see [1, 3, 27–31]). According to this approach, the posteriori probability $p\,(A_i|\text{result})$ of identification of the compound $A_i$ based on the analysis result, is estimated as follows:

$$p(A_i|\text{result}) = \frac{p(\text{result}|A_i) \cdot p(A_i)}{\sum\limits_i p(\text{result}|A_i) \cdot p(A_i)}, \tag{3.3}$$

where $p(\text{result}|A_i)$ is the conditional probability of obtaining the result given the substance $A_i$ is present in the sample, and $p(A_i)$ is the prior probability that $A_i$ is present there, the index $i$ refers to one from analytes.

This equation can be qualitatively considered. The case is imaginable where the spectral data indicate that several compounds $A_i$ may be components of the sample, i.e., their $p(result \mid A_i) > 0$. Based on the prior data, an analyst can assume that the compound $A_1$ ($i = 1$) is unlikely to be present in the sample. Here, the prior probability, $p(A_1)$, the product $p(result|A_i) \cdot p(A_1)$, and therefore the posterior probability, $p(A_1|result)$, are close to zero. This means that the hypothesis: the analyte is the compound $A_1$, is rejected. The experimental test to confirm this conclusion may be not needed.

Based on the prior data, an analyst also excludes the presence of other compounds $A_i$ but $A_2$, i.e., $p(A_i) = 0$ for $i \neq 2$. This means that only product $p(result|A_2) \cdot p(A_2) \neq 0$. Then $p(A_2|result)$ is evidently close to 1, see eq. (3.3), which implies that the compound $A_2$ is only detected and identified. Depending on applied analytical techniques and methods, this identification result should or should not be further confirmed.

The subjective or numerical estimation of prior probability is connected with consideration of prior data/information, which will be addressed in Chap. 6. For other pertinent applications of Bayesian approach, see for example [32–37] and references in Sect. 8.1.2.

## 3.6 Intellectual Operations, Making Decisions

### *3.6.1 General*

Apart from experiments and statistical calculations, chemical analysis includes also intellectual activity of analysts not reduced to performing experimental and calculational work. In target analysis based on standard methods, rational (mental) activity and creativity are not incorporated to a large degree. However, such a kind of analysis as unknown/non-target determination demands significant intellectual labor, including:

- Selection and adjustment of appropriate techniques, methods, reference data, software, and so on to the analytical problem
- Estimation of trueness of an identification result

In many cases, advanced proposals are first demanded about the nature of unknown analytes, followed by estimating plausibility of these speculations.

In science, such proposals being set up by analysts are named *hypotheses*. So we consider chemical identification to be a process of setting up, testing, and screening of hypotheses. This approach is rather common for procedures of detection (e.g., see [38, 39]), and has only recently been suggested for identification operations [30, 40, 41]. On performing both procedures, an analyst explicitly or implicitly accepts/rejects corresponding hypotheses, i.e., makes one or another decision. Therefore, the theory of hypothesis testing [42], and in part the theory of decision making [4, 43], are of significance to chemists.

Due to the difference in the essence of detection and identification operations, hypotheses connected with these are also not the same. This concerns mainly the statement of a null hypothesis.

### *3.6.2 Hypotheses Connected with Detection*

In the case of detection, the *null* hypothesis $H_0$ ("no difference") means that any kind of difference between the blank and the sample is due to chance. So the tested null $H_0$ and *alternative* $\overline{H}_0$ hypotheses:

$$H_0 : \text{an analyte is absent in the sample} \tag{3.4}$$

$$\overline{H}_0 : \text{an analyte is present in the sample} \tag{3.5}$$

can be stated and tested.

An analyst accepts or rejects the null hypothesis. The second decision stands for accepting the alternative hypothesis. In both cases, the analyst can make a true or

false decision. Thus, there are four different results of detection (or qualitative analysis as a whole): TP, FP, TN, and FN (Table 3.2).

A statistical reason for false results of detection is shown in Fig. 3.4. The blank signal combining detector noise together with chemical noise is inherently distributed over its intensity. This distribution is obtained from numerous analyses of a blank sample. The right tail of the distribution curve falls in the range of the analyte signal for its concentration $x_\beta$. The range about the intersection point at the concentration $x_\alpha$ is one of the possible false results. In that point, the signal may belong to both the noise and signal with the same probability. The value $\alpha$ is set before testing the hypothesis connected with the analyte detection.

**Table 3.2** Hypotheses and errors associated with analyte detection

| Analyst's decision | Reality | |
|---|---|---|
| | An analyte is absent, $H_0$ is true | An analyte is present, $\overline{H}_0$ is true |
| An analyte is absent, $H_0$ accepted, $\overline{H}_0$ rejected | TN probability $(1-\alpha)$ | FN type II error probability $\beta$ |
| An analyte is present, $\overline{H}_0$ accepted, $H_0$ rejected | FP type I error probability $\alpha$ | TP probability $(1-\beta)$ |

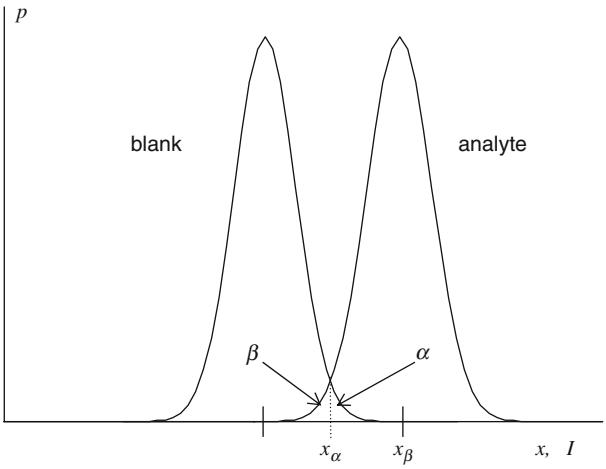



**Fig. 3.4** The probability distribution of the analytical signal intensity for a blank and the analyte concentration $x_\beta$. The horizontal axis is both concentration $x$ and intensity $I$ proportionally related to $x$. For each concentration, e.g., $x_\beta$, there is an inherent distribution with the maximum as the most probable intensity. Signals higher or lower than the maximum one are observed with lower probability, i.e., at $x > x_\beta$ or $x < x_\beta$. The blank distribution is considered in a similar manner. In the case of the analyte concentration $x_\beta$, the other value, $x_\alpha$, is a critical value named a decision limit. At lower concentration, an analyte is assumed to be undetected with the FN probability $\beta$. At a higher amount, it is considered to be detected with the FP probability $\alpha$. Here $\alpha$ and $\beta$ are area fractions under corresponding curves intercepted by the dotted line

The border value $x_\alpha$ is the special point for testing of hypotheses (3.4) and (3.5); see Fig. 3.4. For a lower concentration, i.e., a weaker signal, the hypothesis $H_0$ is accepted and $\overline{H}_0$ is rejected. The possibility that this decision is false (FN result), and that $\overline{H}_0$ is not correctly rejected, is expressed by the type II error with its value $\beta$.

For a concentration higher than $x_\alpha$ and a stronger signal, the null hypothesis $H_0$ is rejected, and the analyte is assumed to be present in the analysed sample. However, this result may be FP, and the corresponding error is $\alpha$ (type I error).

The value $\alpha$, commonly 0.05, can be considered as the chance criterion of accepting the hypothesis. For the single (not distributed) analytical signal which is stronger than one in the point $x_\alpha$, the probability of FP measured by the area under the blank distribution curve intersected by the vertical line right to $x_\alpha$ (not shown in Fig. 3.4) is smaller than $\alpha$. It means that the null hypothesis is rejected, with a lower error for FP than for the signal observed in $x_\alpha$. This fits the statement based on common sense that the probability that a signal is noise is decreased with an increase in signal intensity.

Correspondingly, if a signal is weaker than in the point $x_\alpha$, the intersected area fraction under the blank probability curve is larger than $\alpha$, and the hypothesis $H_0$ is certainly accepted.

For the established $\alpha$, the value $\beta$ depends on the concentration level $x_\beta$ rather than a noise level. In the case of high concentration, its distribution curve is shifted to the right as compared with the curve in Fig. 3.4. The error $\beta$ becomes very low, which naturally means that the probability of falsely rejecting the null hypothesis is insignificant.

### *3.6.3 Identification Hypotheses*

The fact that some compound is detected obviously does not imply its unambiguous identification. Only appearance of an analytical signal can be rigorously stated. However, one can propose that the detected analytical signal belongs to the particular chemical compound/substance or the group of such objects. A proposal of this kind is an *identification hypothesis* [30]. S*tructure hypotheses* are also mentioned [44, 45].

An analyst determining an unknown compound suspects that it may be compound A, and sets up two identification hypotheses:

$$H_0 : \text{the analyte is compound A}, \tag{3.6}$$

$$\overline{H}_0 : \text{the analyte is not compound A} \tag{3.7}$$

The first is the null hypothesis. It states that there is no any difference between the analyte and one of the known compounds, compound A, as the candidate for identification. If the hypothesis (3.6) is rejected, one accepts the alternative one (3.7). The identification results and proper errors are given in Table 3.3.

**Table 3.3** Hypotheses and errors associated with analyte identification

| Analyst's decision | Reality | |
|---|---|---|
| | An analyte is A, $H_0$ is true | An analyte is not A, $\overline{H}_0$ is true |
| An analyte is compound A, $H_0$ accepted, $\overline{H}_0$ rejected | TP probability $(1-\alpha)$ | FP type II error probability $\beta$ |
| An analyte is not compound A, $\overline{H}_0$ accepted, $H_0$ rejected | FN type I error probability $\alpha$ | TN probability $(1-\beta)$ |

Unlike the simple hypothesis (3.6), the alternative proposal (3.7) is a composite one, i.e., a set of simple hypotheses:

$$H_1 : \text{the analyte is compound B} \tag{3.8}$$

or

$$H_2 : \text{the analyte is compound C} \tag{3.9}$$

or some other compound. So, if the null hypothesis is rejected and positive identification is required, the next proposal (3.8) should be tested and so on.

In qualitative analysis II (Chap. 8), identification hypotheses are set up in a similar way. That may be exemplified by the authentication procedure of honeys [33]:

$$H_0 : \text{the honey is Galician,} \tag{3.6a}$$

$$\overline{H}_0 : \text{the honey is non-Galician} \tag{3.7a}$$

In identification, the hypotheses (3.6) and (3.7) can be reversed, and the hypothesis $\overline{H}_0$ can be considered as the null one. Sometimes, it is more important for analyst to find out that the analyte is not the predetermined compound, and more certain identification does not really matter. Thus there is a seeming freedom in choosing the null hypothesis.

However, the choice of the null hypothesis can be substantiated. For example, the fact that type I error is a false rejection of a null hypothesis by definition can be taken into account. Neyman, the author of fundamental works on the theory of hypothesis testing, recommended the term *type I error* for denoting the error which it was more important to avoid [42]. In chemical analysis, this is often a FN result. Indeed, in the screening of samples, positive responses are further confirmed by more reliable methods, whereas negative responses are considered to be final (Sect. 2.9). Therefore, the probability of the FN response should be lower, which means that a type I error is FN. This is the case for identification (Table 3.3), but in contrast to the relationship between errors set up for detection (Table 3.2).

Simplicity can also be a selection criterion for the null hypothesis [46]. For chemical identification, the assumption $H_0$ (3.6) is much simpler than the alternative $\overline{H}_0$ (3.7). Indeed, the latter is a composite one because it consists of many simpler hypotheses – (3.8), (3.9), and so on.

Thus, the concept of simplicity and the statement of "no difference" are in line in defining the same null hypothesis (3.6). Also, the relationship between the two types of errors, FP/FN results and type I/II errors, is reversed for detection (Table 3.2) as compared with identification (Table 3.3). This is due to the different nature of the null hypotheses (3.4) and (3.6). In the literature on identification, both this and the contrary interpretation of the concept of null hypothesis are available, yielding $\alpha \equiv$ FN, $\beta \equiv$ FP [7, 40] (see Table 3.3) and $\alpha \equiv$ FP, $\beta \equiv$ FN [47] (see Table 3.2). The last pair of relationships also occur in books on statistics in chemistry [2, 4]. To avoid confusion in this book, the terms *type I error* and *type II error* will not be used often; instead, the terms FN and FP, which are understood unambiguously, will be preferred.

### 3.6.4 Experimental Hypotheses

In order to find out which of the identification hypotheses (3.6) or (3.7) is true, they should be transformed into hypotheses suited for experimental testing:

$$H_0 : \text{properties of the analyte and compound A are not differentiated,} \tag{3.10}$$

$$\overline{H}_0 : \text{properties of the analyte and compound A are different} \tag{3.11}$$

These can be named *experimental hypotheses*.

Properties of chemical compounds are (a) qualitative features, or (b) values of measured quantities (see Chap. 1). In case (a), the identity of features resulting from the use of validated methods stands for TP of qualitative determination, i.e., the hypothesis (3.10) is accepted. For measurements (b):

- A similarity degree of corresponding numerical values is taken into account or
- A value interval is divided into narrow ranges, each of which is a certain qualitative feature

In any case, the fact that a value for an unknown compound falls in the window range established for compound A means:

- Identity of features or close similarity in values
- Acceptance of the hypothesis (3.10) and
- A chance for the analyte to be identified as compound A

If a value of the measurand of any other compound is outside the range specific for A, and different analytical techniques or their different versions also lead to

values typical for compound A, a possibility of identification is transformed into the analyst's confident decision.

Such *range criteria* (terms of *window criterion* or *tolerance criterion* are synonymous ones) for acceptance of hypotheses applied for retention parameters (time, relative time, index) are widespread in chromatography (Chap. 5, Sect. 7.2); see Fig. 3.5. For multivariate quantities, e.g., spectra, univariate range criteria analogous to those shown in Fig. 3.5 can be applied for each variable (intensity of spectral peak; see Chap. 5). In analytical practice, univariate measures/indices are also calculated which express a similarity of full spectra as multivariate quantities (Fig. 3.6). The window range of high indicator values is one for acceptance of experimental hypotheses. Dissimilarity in spectra is the reason for making the decision to reject an experimental and thus identification hypothesis (Fig. 3.6).

The window range is chosen on the base of available reference data and their measured or typical spreads. Several cases deserve to be noted.

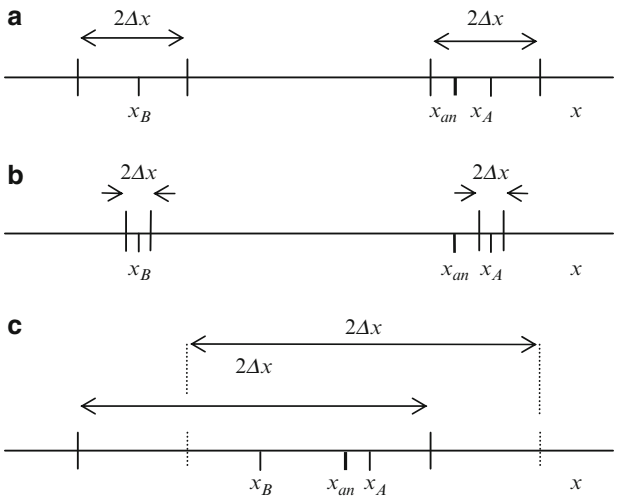



**Fig. 3.5** Range testing of identification hypothesis. There is the case of the univariate measurand $x$, e.g. retention parameter, with $x_{an}$, $x_A$, $x_B$ being the values for the analyte, compounds A and B, respectively. The ranges $2\Delta x_A$ and $2\Delta x_B$ are criterion windows about reference values $x_A$ and $x_B$ and also the measures of identification uncertainty. (**a**) The analyte has the value $x_{an}$ falling within the window range of A, and is therefore identified as that compound. In other words, the hypothesis (3.10) is correctly accepted. (**b**) For the thin range $2\Delta x_A$, the value $x_{an}$ measured with some uncertainty may be fall outside this range. It results in FN, i.e., the false rejection of the null hypothesis (3.10). The case is typical for very accurate reference data and experimental value measured with a bias or vice versa. (**c**) In contrast, widening of the range $2\Delta x_A$ (and also the $2\Delta x_B$ ) may lead to overlapping of criterion tolerances of reference values for compounds A and B, i.e., A and B will be hardly distinguishable by $x$. If unambiguous identification is demanded, it may lead to FP, which is the identification of B instead of A or vice versa. It is the possibility of false acceptance of (3.6) or (3.8). The ambiguous answer that the analyte is compound A or B is true. Due to a spread in real reference data, this is a very practical case

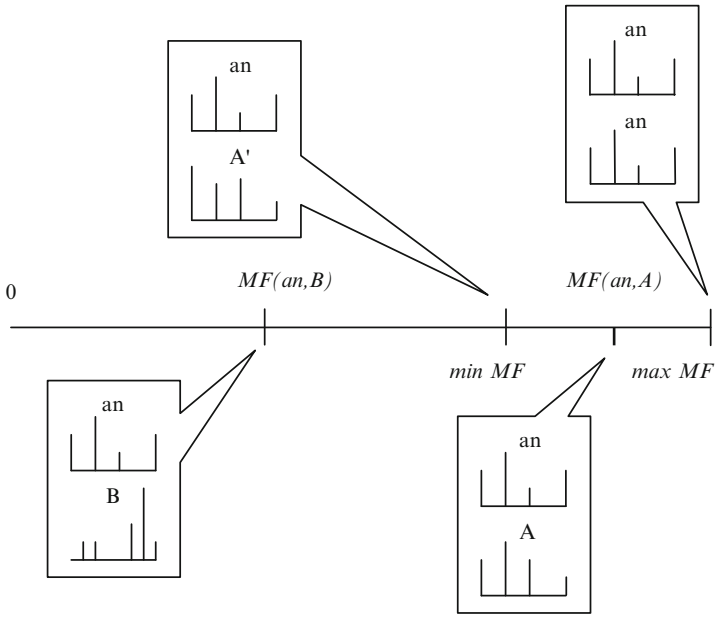



**Fig. 3.6** Range testing of identification hypothesis. There is a case of a multivariate measurand, e.g., mass spectrum, with a match factor *MF* as a value of univariate criterion for accepting/rejecting of experimental hypotheses (identification criterion) and the measure of reliability of identification. The match factors *MF(an,A)* and *MF(an,B)* indicate how close the experimental spectrum of an analyte matches that of A and B respectively. The latter are reference spectra. The range between *max MF and min MF* is the interval of high *MF* values for acceptance of experimental hypotheses. Here, *max MF* is the maximum value (1, 100, or 1,000 depending on the algorithm of MF calculation and the particular software) and *min MF* (e.g., 900 at maximum value 1,000) is the lowest value for acceptance of the null hypothesis, i.e., for positive identification. The comparison of a spectrum to itself gives *max MF*. (1) Further, the point *MF(an,A)* expresses the similarity of the experimental spectrum to a not fully similar spectrum of the compound A. The value *MF(an,A)* falls into that range for identification, so the hypothesis (3.10) is accepted. The factor *MF(an,B )* is much lower and outside the range. The hypothesis that the analyte is compound B is therefore rejected. (2) For relatively high values of *min MF* (not shown), the observed value *MF(an,A)* may be fall outside the range for identification. It results in FN, i.e. the false rejection of the null hypothesis (3.10). (3) In contrast, low values of *min MF* (also not shown) may lead to *MF(an,B )* falling within the range between *max* and *min*. It leads to FP, which is the identification of B instead of A or vice versa, or true ambiguous identification of the pair of compounds. For the last case, the conclusion is that A and B would be hardly distinguishable by this kind of spectrometric technique and these reference data

1. There is a single reference value. The interval $2\Delta x$ (see Fig. 3.5) is a typical range of the measurand observed in reliable experiments. The tolerance also can be calculated from standard deviations of those data. In methods, range/interval criteria for accepting hypotheses are predetermined (Chap. 5).
2. There are two or more reference values. The interval is $x_{max}$–$x_{min}$, which is extracted from those references, with outliers being excluded. This range also can be taken from equivalent methods (Chap. 5).

3. Reference and experimental data are biased to each other due to different experimental conditions when these data sets are obtained. The range for identification should be corrected for the bias value (Sect. 7.2).

The range method for testing of experimental hypotheses is simplest and most efficient if analytical conditions are not very different from those in which reference values were recorded. For the same conditions and same compounds, deviations in results of measurement results are just random. This is the case for statistical control and setting up of statistical hypotheses. These are special types of experimental ones.

### 3.6.5 Statistical Hypotheses

Any property as a measurand can be expressed in the form of a statistical distribution. Accordingly, the hypotheses (3.10) and (3.11) can be formulated in the other way:

$$H_0 : \text{distribution parameters of the variate } x \text{ for the analyte and compound A are not significantly different, i.e., values of } x \text{ belong to the same population,} \tag{3.12}$$

$$\overline{H}_0 : \text{distribution parameters of the variate } x \text{ for the analyte and compound A are significantly different, i.e., values of } x \text{ do not belong to the population for A} \tag{3.13}$$

In a particular case, an analyst finds out whether a single value of an analyte belongs to the distribution for one or another compound (Fig. 3.7).

Methods of testing of statistical hypotheses such as (3.12) and (3.13) depend on the types of distribution and data. The latter can be approximated in the form of normal distribution (see Fig. 3.7), but their population, i.e., the infinite set, is unattainable. All an analyst really has is a sample of values. For this case, the Student's *t*-distribution is usually considered, and the statistical *t*-test is often applied to test hypotheses as follows (see books [1, 2, 4] for this and many other tests with full details).

Given that an analyte is compound A, observed values of a particular variable $x$ for an analyte and A are $x = (x_1, x_2,...,x_m)$ and $x_A = (x_{A1}, x_{A2},...,x_{An})$ respectively; the indices $1, 2, ..., m$, $A_1, A_2, ..., A_n$ refer to individual observations; measurement methods and procedures involved are the same, the condition must be fulfilled:

$$\frac{|\overline{x} - \overline{x}_A|}{s} \cdot \sqrt{\frac{m \cdot n}{m + n}} < t_x(\alpha, \upsilon), \tag{3.14}$$

where $s = \sqrt{\frac{s_{an}^2(m-1)+s_A^2(n-1)}{\upsilon}}$; $s_{an}$ and $s_A$ are standard deviations for an analyte and the compound $A$ respectively; $t_x(\alpha, \upsilon)$ is the critical value for $t$; $\alpha$ is the significance level, which is usually 0.05; $\upsilon = m + n - 2$ is the degrees of freedom.

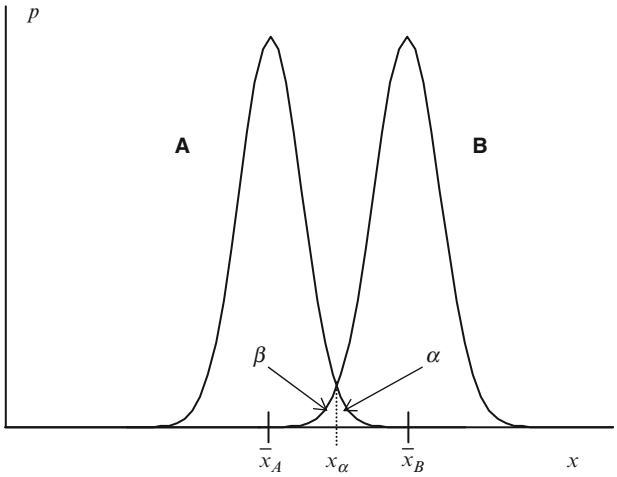



**Fig. 3.7** Testing of hypotheses in which properties of candidate compounds are expressed by the distributions of the variate $x$ (horizontal axis). The mean $x$ values of compounds A and B are $\overline{x}_A$ and $\overline{x}_B$ respectively. The values $x$ of the analyte are individual observations. The various kinds of the null hypothesis are (3.6), (3.10), and (3.12). The left curve is the probability distribution for the compound A. The value $\alpha$, commonly 0.05, is the area fraction under the curve in the range of relatively high values $x$ not specific for A ($x \geq x_\alpha$). If the analyte has the relatively high value $x > x_\alpha$, the decision is made that it is not compound A. This decision may be erroneous with the type I error $\alpha$, which is the probability of FN (see Table 3.3). For $x$ closer to the mean $\overline{x}_A$, i.e., to the left from the border value $x_\alpha$, the hull hypothesis (the analyte is A) is accepted, with the probability of TP being $(1-\alpha)$. The acceptance may be also erroneous with the type II error $\beta$, which is the corresponding area fraction under the curve for the candidate compound B (the probability of FP). If the alternative hypothesis (the analyte is B) is tested, all considerations are analogical. If $x = x_\alpha$, the analyte can be identified as A or B with the same probability of FP. For a significant difference in the mean $x$ between both tested compounds, the B distribution curve shifts to the right (not shown). In this case, if $x = x_\alpha$ ($\alpha = 0.05$), false identification as the compound B is improbable

In contrast, if the condition (3.14) is met, the hypothesis $H_0$ (3.12) is accepted and the alternate hypothesis $\overline{H}_0$ (3.13) is rejected for the significance level $\alpha$. If the relationship is the inverse of (3.14), the alternate hypothesis (3.13) is accepted and the null hypothesis (3.12) is rejected. The lower the $\alpha$ value, the higher the permitted difference in $x$ for accepted hypotheses.

This $t$-criterion for accepting or rejecting the hypothesis is readily transformed to the probability $\alpha$ criterion. If a value $\alpha$ that meets the equation:

$$\frac{|\overline{x} - \overline{x}_A|}{s} \cdot \sqrt{\frac{m \cdot n}{m + n}} = t_x(\alpha, \upsilon) \tag{3.15}$$

is equal to or larger than 0.05, the hypothesis $\overline{H}_0$ is accepted and the alternate hypothesis $\overline{H}_0$ is rejected for this significance level $\alpha$, which is the statistical type I error. Given (3.15) is met for $\alpha < 0.05$, the conclusion on the hypotheses is reversed.

The probability $\alpha$ is not only the statistical parameter but also the type I identification error related to the hypothesis (3.6) and one of the measures of identification reliability. That is the probability of the means $\bar{x}$ and $\bar{x}_A$ being far different given an analyte = A, which results in the erroneous decision that an analyte is not *A*, i.e., FN (Table 3.3). Usually this is not a widespread error, since the large difference between means is due to them belonging to different compounds.

Thus, the value $\alpha \geq 0.05$ is the prerequisite for identification. However, it is a necessary but insufficient criterion for unambiguous identification. The latter also calls for low type II error, i.e., low FP, which means that the probability of similarity in observed sample average of the quantity *x* between an analyte and substances B, C, etc. is low, and therefore falsely accepting hypothesis $H_0$ is unlikely.

The statistical type II error, designated $\beta$ (see Table 3.3), can be evaluated for statistical versions of every individual alternate hypothesis (3.8), (3.9), etc. by using a non-central Student's distribution [1]. An approximate way to do so is to equate $\beta$ to the value $\alpha$ obtained when testing hypotheses $H_1$, $H_2$, etc. This is illustrated by Fig. 3.7, where type I error at testing of the hypothesis: analyte is B, is $\beta$.

Thus, two cases of testing of statistical hypotheses were considered above. First, the theoretical one for normal distribution of reference data and single observations for an analyte is shown in Fig. 3.7. Second, there are *t*-distribution (*t*-test) and two samples of reference and experimental data. The latter is practical testing, where a single value of the measurand without replicates can be used instead of a corresponding sample of values.

For the same reference and experimental data, both tolerance (Sect. 3.6.4) and statistical criteria can be applied for testing of hypotheses. Moreover they are transformed into each other and may lead to a close level of FN and FP [48] (see also Sect. 7.2.5). Range (tolerance) criteria are widespread, because they are simpler to apply. Testing of statistical hypotheses demands a higher degree of analyst skill and experience. However, in this approach errors are numerically evaluated which are of value to estimate an overall reliability of identification (Chap. 4). Therefore, statistical hypotheses should be formulated and tested in particularly important analyses such as

- Interlaboratory comparisons in qualitative analysis
- Chemical analysis of samples originating from various accidents, e.g., disaster, poisoning, and so on

For multivariate data, such as spectra, a suitable hypothesis can be tested by using the Hotelling' $T^2$ statistics and the threshold value $F_x$ [1]:

$$\frac{T^2}{a(\upsilon)} = (\bar{\mathbf{x}} - \bar{\mathbf{x}}_\mathbf{A})' C^{-1} (\bar{\mathbf{x}} - \bar{\mathbf{x}}_\mathbf{A}) < b(\upsilon) \cdot F_x(\alpha, \upsilon), \tag{3.16}$$

$$\frac{T^2}{a(\upsilon)} = (\bar{\mathbf{x}} - \bar{\mathbf{x}}_\mathbf{B})' C^{-1} (\bar{\mathbf{x}} - \bar{\mathbf{x}}_\mathbf{B}) > b(\upsilon) \cdot F_x(\alpha, \upsilon), \tag{3.17}$$

where $\mathbf{x} = (x_{11}, x_{21}, \ldots, x_{n1}, x_{12}, x_{22}, \ldots, x_{n2}, \ldots, x_{1m}, x_{2m}, \ldots, x_{nm})$ is the sample vector of an analyte; index $ij$ refers to univariate $i$ and observation $j$, i.e., $x_{21}$ means the first observation of the variate 2, etc.; the vectors $\mathbf{x_A}$ and $\mathbf{x_B}$, for the compounds identified and unidentified respectively, are defined in a similar manner; $a(\upsilon)$ and $b(\upsilon)$ are the factors depending on degrees of freedom $v$; $C$ is the sample covariance matrix, and $F_x(\alpha,\upsilon)$ is the value of the variable from $F$-distribution for significance level $\alpha$ and degrees of freedom $v$. Threshold values $F_x$ are readily converted to the threshold $\alpha$ as the measures of identification reliability, in line with the univariate formulae (3.15).

Although the Hotelling approach is attractive due to the possibility of estimating identification error, this is not easy to implement. The reasons are that (a) the corresponding software is not widespread, and (b) large data samples (numbers of experimental and/or reference spectra) are required for calculations. So the approach has rarely been applied in routine analysis; for the exception, see for example the work [48].

## References


1. Lloyd E (1984) Handbook of applicable mathematics, vol 6, Statistics. Wiley, Chichester
2. Meier PC, Zund RE (1993) Statistical methods in analytical chemistry. Wiley, New York
3. Sharaf MA, Illman DL, Kowalski BR (1986) Chemometrics. Wiley, New York
4. Massart DL, Vandeginste BGM, Deming SN, Michotte Y, Kaufman L (1988) Chemometrics: a textbook. Elsevier, Amsterdam
5. Varmuza K, Filzmoser P (2009) Introduction to multivariate statistical analysis in chemometrics. CRC Press, Boca Raton, FL
6. Thompson SK (1992) Sampling. Wiley, New York
7. Milman BL, Konopelko LA (2000) Identification of chemical substances by testing and screening of hypotheses. I. General. Fresenius J Anal Chem 367:621–628
8. Jurado JM, Alcázar A, Pablos F, Martín MJ, González AG (2005) Classification of aniseed drinks by means of cluster, linear discriminant analysis and soft independent modelling of class analogy based on their Zn, B, Fe, Mg, Ca, Na and Si content. Talanta 66:1350–1354
9. Pillati M, Viroli C (2010) Gene selection in classification problems using independent factor analysis. http://www2.stat.unibo.it/viroli/publications/articleIFa.pdf. Accessed 1 May 2010
10. Goux WJ (1989) NMR pattern recognition of peracetylated mono- and oligosaccharide structures. Classification of residues using principal-component analysis, *K*-nearest neighbor analysis, and SIMCA class modeling. J Magn Reson 85:457–469
11. Aruga R, Mirti P, Casoli A, Palla G (1999) Classification of ancient proteinaceous painting media by the joint use of pattern recognition and factor analysis on GC/MS data. Fresenius J Anal Chem 365:559–566
12. Hristozov D, Da Costa FB, Gasteiger J (2007) Sesquiterpene lactones-based classification of the family Asteraceae using neural networks and *k*-nearest neighbors. J Chem Inf Model 47:9–19
13. Elomaa M, Lochmüller CH, Kudrjashova M, Kaljurand M (2000) Classification of polymeric materials by evolving factor analysis and principal component analysis of thermochromatographic data. Thermochimica Acta 362:137–144
14. Anderson KA, Magnuson BA, Tschirgi ML, Smith B (1999) Determining the geographic origin of potatoes with trace metal analysis using statistical and neural network classifiers. J Agric Food Chem 47:1568–1575

15. Pell M, Ljunggren H (1996) Composition of the bacterial population in sand-filter columns receiving artificial wastewater, evaluated by soft independent modelling of class analogy (SIMCA). Water Res 30:2479–2487
16. Walczak B, Morin-Allory L, Lafosse M, Dreux M, Chrétien JR (1987) Factor analysis and experiment design in high-performance liquid chromatography. VII. Classification of 23 reversed-phase high-performance liquid chromatographic packings and identification of factors governing selectivity. J Chromatogr A 395:183–202
17. Zeng Y, Hopke PK (1990) Methodological study applying three-mode factor analysis to three-way chemical data sets. Chemometrics Intell Lab Syst 7:237–250
18. Harwood VJ, Whitlock J, Withington V (2000) Classification of antibiotic resistance patterns of indicator bacteria by discriminant analysis: use in predicting the source of fecal contamination in subtropical waters. Appl Environ Microbiol 66:3698–3704
19. Serrano S, Villarejo M, Espejo R, Jodral M (2004) Chemical and physical parameters of Andalusian honey: classification of *Citrus* and *Eucalyptus* honeys by discriminant analysis. Food Chem 87:619–625
20. Moret I, Di Leo F, Giromini V, Scarponi G (1994) Multiple discriminant analysis in the analytical differentiation of Venetian white wines. 4. Application to several vintage years and comparison with the k nearest-neighbor classification. J Agric Food Chem 32:329–333
21. Ankerst M, Kastenmüller G, Kriegel HP, Seidl T (1999) Nearest neighbor classification in 3D protein databases. ISMB-99 Proceedings. http://www.aaai.org/Papers/ISMB/1999/ISMB99-005.pdf. Accessed 2 May 2010
22. Wiberg K, Hagman A, Burén P, Jacobsson SP (2001) Determination of the content and identity of lidocaine solutions with UV-visible spectroscopy and multivariate calibration. Analyst 126:1142–1148
23. Vohradský J (1997) Adaptive classification of two-dimensional gel electrophoretic spot patterns by neural networks and cluster analysis. Electrophoresis 18:2749–2754
24. McNeil VH, Cox ME, Preda M (2005) Assessment of chemical water types and their spatial variation using multi-stage cluster analysis, Queensland, Australia. J Hydrol 310:181–200
25. Chun J, Atalan E, Ward AC, Goodfellow M (1993) Artificial neural network analysis of pyrolysis mass spectrometric data in the identification of *Streptomyces* strains. FEMS Microbiol Lett 107:321–326
26. Song XH, Hopke PK (1999) Classification of single particles analyzed by ATOFMS using an artificial neural network, ART-2A. Anal Chem 71:860–865
27. Sivia DS (2001) Data analysis: a Bayesian tutorial. Oxford University Press, Clarendon
28. Spiehler VR, O'Donnell CM, Gokhale DV (1988) Confirmation and certainty in toxicology screening. Clin Chem 34:1535–1539
29. Ellison SLR, Gregory S, Hardcastle WA (1998) Quantifying uncertainty in qualitative analysis. Analyst 123:1155–1161
30. Milman BL, Konopelko LA (2000) Identification of chemical substances by testing and screening of hypotheses I. General. Fresenius J Anal Chem 367:621–628
31. Milman BL (2005) Identification of chemical compounds. Trends Anal Chem 24:493–508
32. Emerenciano VDP, Ferreira MJP, Branco MD, Dubois JE (1998) The application of Bayes' theorem in natural products as a guide for skeleton identification. Chemometrics Intell Lab Syst 40:83–92
33. Latorre MJ, Peña R, García S, Herrero C (2000) Authentication of Galician (N.W. Spain) honeys by multivariate techniques based on metal content data. Analyst 125:307–312
34. Roussel S, Bellon-Maurel V, Roger JM, Grenier P (2003) Fusion of aroma. FT-IR and UV sensor data based on the Bayesian inference. Application to the discrimination of white grape varieties. Chemometrics Intell Lab Syst 65:209–219
35. Alterovitz G, Liu J, Afkhami E, Ramoni MF (2007) Bayesian methods for proteomics. Proteomics 7:2843–2855

36. Toher D, Downey G, Murphy TB (2007) A comparison of model-based and regression classification techniques applied to near infrared spectroscopic data in food authentication studies. Chemometrics Intell Lab Syst 89:102–115
37. Hibbert DB, Armstrong N (2009) An introduction to Bayesian methods for analyzing chemistry data. II. A review of applications of Bayesian methods in chemistry. Chemometrics Intell Lab Syst 97:211–220
38. Beyermann K (1984) Organic trace analysis. Ellis Horwood, Chicester
39. Currie LA (1995) Nomenclature in evaluation of analytical methods, including detection and quantification capabilities (IUPAC Recommendations 1995). Pure Appl Chem 67:1699–1723
40. Hartstra J, Franke JP, de Zeeuw RA (2000) How to approach substance identification in qualitative bioanalysis. J Chromatogr B 739:125–137
41. Eriksson J, Chait BT, Fenyö D (2000) A statistical basis for testing the significance of mass spectrometric protein identification results. Anal Chem 72:999–1005
42. Neyman J (1968) Introductory course of probability theory and mathematical statistics (In Russian). Nauka, Moscow
43. March JG (1994) Primer on decision making: how decisions happen. Simon and Schuster, New York
44. Vershinin VI, Derendyaev BG, Lebedev KS (2002) Computer-Assisted Identification of Organic Compounds (In Russian). Akademkniga, Moscow
45. Elyashberg M, Blinov K, Williams A (2009) A systematic approach for the generation and verification of structural hypotheses. Magn Reson Chem 47:371–389
46. Easton VJ, McColl JH Statistics Glossary. http://www.stats.gla.ac.uk/steps/glossary/hypothesis_testing.html#h0. Accessed 2 May 2010.
47. Nesvizhskii AI, Vitek O, Aebersold R (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Methods 4:787–797
48. Milman BL, Kovrizhnych MA (2000) Identification of chemical substances by testing and screening of hypotheses. II. Determination of impurities in n-hexane and naphthalene. Fresenius J Anal Chem 367:629–634

# Chapter 4
# Reliability and Errors of Identification

**Abstract** In this chapter, approaches to estimating reliability and errors of detection and identification are considered. Related terminology is presented; reliability of identification is defined as a probability of its true result. False results are demonstrated to be attributes of determination of low analyte amounts by screening methods. Formulas for calculating rates of true and false, positive and negative results are given. The rates are derived both from tests using analytical standards (blank samples) and upon verification of screening results by confirmatory methods/techniques. A replication of analytical determinations is also considered, including Bayesian statistics. Limit characteristics of detection and identification are treated.

It is noted that confirmatory methods based on spectrometry must be free of identification errors. Nevertheless, errors occur if methods are non-targeted, invalidated, or *ad hoc*. True and false results obtained with use of spectral techniques are discussed in terms of matching spectra. A best/good or poor matching resulting in a high or low match factor means a good/fair or poor chance respectively of accepting an identification hypothesis. Different match factors calculated in mass spectrometry and also NMR, IR-, and UV–V is spectroscopy are outlined, with many details with regard to searches in reference spectral libraries. Further, a probability interpretation of match factors is considered, which is essential for identification of peptides and proteins in proteomics. Other approaches to deriving a probability of identification from analytical/spectral data are also noted. This kind of probability, as well as the reported result of identification, can be expressed in words.

## 4.1 General

In previous chapters, identification errors were only briefly considered when testing of hypotheses was discussed. In this chapter, the matter of reliability of detection and identification results and approaches to estimating corresponding error rates will be considered as fully as possible.

B.L. Milman, *Chemical Identification and its Quality Assurance*,
DOI 10.1007/978-3-642-15361-7_4, 

A result of both *detection* and *target identification* is clearly positive or negative and true or false (Table 4.1). Again, Table 4.1 shows the difference between detection and identification:

- Detection is preliminary but decisive identification
- Type I/II errors are contrary to each other (Sect. 3.6.3)

It is easy to conclude that qualitative operations carried out during *screening* procedures (Chap. 1) are far closer in matter to detection than identification (see also Chap. 5). So *detection* is fairly synonymous with *screening*.

An analyst formulates the result of identification (see examples in Table 4.1) based on experimental and reference data, method criteria, theoretical models, reference materials, etc. The same data can be used to estimate chances for trueness or falseness, i.e., error rate of results in qualitative analytical procedures. There are concepts and terminology for reporting those results and errors.

In order to express the extent to which an analyst is confident in chemical identification results, numerous terms occur in the literature on qualitative analysis. There have been *probability*, *reliability*, *uncertainty*, *certainty* (see [1]), *correctness* [2], *confidence* [3] and so on. In science as a whole, there are even more terms with regard to identification (Table 4.2). The author has chosen the frequently used term of *reliability* [4], with the following definition:

*Reliability of identification is a probability of its true result, denoted by p(TP).*

In this book, this term is also used. To some degree, it is the conventional one because other terms also occur often (Table 4.2). To establish the standard terminology in the field, some kind of consensus between analytical chemists is obviously required.

Other terms, FP and FN, are generally accepted. They probably originated in medical diagnostics (see [6]). Then this terminology was transferred to biochemical and chemical analysis, including instrumental techniques and chemo- and bioinformatics (see Chap. 7).

The concepts of reliability, probability, and error are interrelated. As the sum of probabilities $p$ of true and false results is 1 (or 100% if expressed in percent), reliability is $1-p(FP)$ *or* $100\%-p(FP)$ and $1-p(FN)$ *or* $100\%-p(FN)$ for negative and positive results respectively. Thus, an analyst trying to estimate the reliability of a result of qualitative analysis should (a) evaluate the probability of true results in a direct or indirect way, or (b) first take into account the probability of false results. Depending on the particular case, corresponding methods may be or not be confirmatory. In confirmatory methods, $p(FP)$ and $p(FN)$ are taken to be about 0, which has the result that reliability is about 1. Therefore, rates of FP or FN are estimated, and used primarily in screening methods based on various techniques beginning with simple qualitative reactions, chemical tests (Fig. 4.1). These techniques will be treated in Sect. 4.2.

Analytical data obtained by confirmatory methods based on spectrometry reliably determine positive and negative results if the methods are validated. In cases of invalidated (*ad hoc*) methods, spectral data not only lead to statements of results but

**Table 4.1** Terminology for detection (screening) and identification

| Result | Detection (screening) | | Identification | |
|---|---|---|---|---|
| | Definition | Remarks | Definition | Example |
| True positive (TP) | The analyte which is *present* in the sample is *detected* | Positive response for the analyte present in the sample (positive result of a confirmatory method) | The analyte is *truly identified* | The chromatographic peak of compound A *is assigned* to this compound |
| True negative (TN) | The analyte which is *absent* in the sample is *not detected* | Negative response for the analyte absent in the sample (negative result of a confirmatory method) | The analyte which is *absent* in the sample is *not identified* | Any chromatographic peak *is absent* in the RT range typical for compound A which is absent in the sample |
| False positive (FP) | The analyte which is *absent* in the sample is *detected*, type I error | Positive response for the analyte absent in the sample (positive result of a confirmatory method) | The analyte is *falsely identified*, type II error | The chromatographic peak of compound *A* is assigned to compound *B* |
| False negative (FN) | The analyte which is *present* in the sample is *not detected*, type II error | Negative response for the analyte present in the sample (positive result of a confirmatory method) | The analyte which is *present* in the sample is *not* identified, type I error | The chromatographic peak of compound A present in the sample *is not assigned* to this compound |

**Table 4.2** Terminology for a trueness degree of identification and related terms. Corresponding frequency in the scientific literature

| Term | The number of articles | Available word combinations and their occurrences |
|---|---|---|
| Performance[a] | 8,432 | *Identification performance* – 7,560, *performance of* (*the*) *identification* – 872 |
| Error[b] | 5,361 | *Identification error* – 4,870, *error of* (*the*) *identification* – 491 |
| Efficiency | 3,627 | *Identification efficiency* – 2,960, *efficiency of* (*the*) *identification* – 667 |
| Reliability | 2,970 | *Reliability of* (*the*) *identification* – 2,380, *identification reliability* – 590 |
| Probability | 2,360 | *Identification probability* – 1,110, *probability of* (*the*) *identification* – 1,250 |
| Confidence | 1,235 | *Identification confidence* – 813, *confidence of* (*the*) *identification* – 422 |
| Certainty | 903 | *Certainty of* (*the*) *identification* – 752, *identification certainty* – 151 |
| Correctness | 780 | *Certainty of* (*the*) *identification* – 697, *identification correctness* – 83 |
| Uncertainty | 685 | *Identification uncertainty* – 400, *certainty of* (*the*) *identification* – 285 |
| Unreliability | 75 | *Unreliability of* (*the*) *identification* – 68, *identification unreliability* – 7 |

[a]The search in Google Scholar [5]
[b]This term is not fully synonymous with *reliability*

also make it possible to estimate identification errors (see Fig 4.1). That is the subject of the second part of this chapter.

## 4.2 Formal Statistics of False and True Results

### *4.2.1 Statistics of False Results*

Commonly used values characterizing the reliability of a used qualitative method are false rates [6–8] such as false positive rate (*FPR*)

$$FPR = \frac{100 \times n_{FP}}{n_{FP} + n_{TN}}\% \tag{4.1}$$

and false negative rate (*FNR*)

$$FNR = \frac{100 \times n_{FN}}{n_{FN} + n_{TP}}\%, \tag{4.2}$$

where $n_{\mathrm{FP}}$, $n_{TN}$, $n_{FN}$, $n_{TP}$ are the numbers of false positives, true negatives, false negatives, and true positives respectively. These and other terms expressing

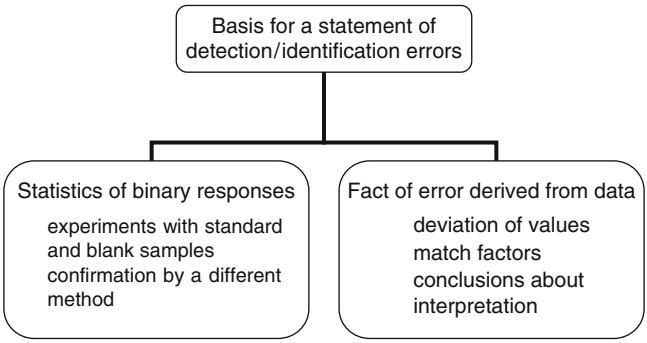



**Fig. 4.1** Ways in which errors of detection and identification are estimated

probabilities of different outcomes of qualitative test are characterized in Table 4.3. Similar terminology is employed in screening procedures performed through the use of information retrieval in data systems, including searches for identification purposes (Chap. 7).

These rates are determined as follows. Identical samples, both containing a certain practically important amount of the analyte, and blank samples containing no analyte (or to be more exact, containing much lower concentration of the analyte), are taken. Both series of samples are tested $n$ times, recording positives and negatives. For the samples containing analyte, true positives and false negatives are recorded (here $n = n_{TP} + n_{FN}$). Conversely, for the blank samples, true negatives and false positives are recorded (in this case, $n = n_{TN} + n_{FP}$). False positive and false negative rates are further calculated by (4.1) and (4.2).

These rates can also be evaluated in two different ways, which are (a) an inspection of screening results with the use of confirmatory methods, and (b) interlaboratory studies.

False rate $FR$ is a random variable following a binomial distribution (see [12, 13] and Sect. 3.2). Depending on the number of replicate experiments (trials), a false rate can be determined with a certain degree of accuracy, expressed by a confidence interval $\Delta_c FR$. This is calculated approximately by the formula (4.3), which is obtained by transforming (3.1) or (3.2).

$$\Delta_c FR = z\sqrt{\frac{FR \times (100 - FR)}{n}}\%, \tag{4.3}$$

where $z$ is the distribution parameter; see its values in Sect 3.2. The confidence intervals, including relative ones, calculated by (4.3) for various $FR$ and $p$ are given in Table 4.4.

Table 4.4 indicates firstly that evaluating the false rate with the same accuracy (same $\Delta_c FR/FR$ values) requires a greater number $n$ of trials for lower error level

**Table 4.3** Basic nomenclature for the metrology of qualitative methods

| Result, rate, value | Notation | Definition, comment |
|---|---|---|
| Positive | P | Presence of the analytical signal |
| Negative | N | Absence of the analytical signal |
| True positive | TP | See Table 4.1 |
| False positive | FP | See Table 4.1 |
| True negative | TN | See Table 4.1 |
| False negative | FN | See Table 4.1 |
| Sensitivity[a] (true positive rate) | *St* (*TPR*) | Percentage of TP of the total of TP and FN; calculated by (4.4) |
| False positive rate | *FPR* | Percentage of FP of the total of FP and TN; calculated by (4.1) |
| Specificity (true negative rate) | *Sp* (*TNR*) | Percentage of TN of the total of TN and FP; calculated by (4.5) |
| False negative rate | *FNR* | Percentage of FN of the total of FN and TP; calculated by (4.2) |
| False rate | *FR* | *FPR* or *FNR* |
| Efficiency[b] | | $100\%\ (TP + TN)/(TP + TN + FP + FN)$ |
| Youden index[b] | | $(St + Sp - 100)\%$ |
| Likelihood ratio[b] | | $(100 - FNR)/FPR$ |
| Positive predictive value[c] | *PPV* | Percentage of TP of the total of TP and FP; calculated by (4.6) or (4.8) |
| Negative predictive value | *NPV* | Percentage of TN of the total of TN and FN; calculated by (4.7) or (4.9) |
| Prevalence | *Pv* | Percentage of samples containing the analyte; corresponds to a prior probability of the presence of this substance in a sample |
| Cumulative positive predictive value | *CPPV* | Percentage of TP of the total of positive responses for duplicate testing; calculated by (4.12) |

[a] The synonym of *recall* is used in estimation of performances for information retrieval (e.g., see [9, 10])

[b] Indices not treated in this book

[c] The synonymic term of *reliability* occurs in reports on information retrieval in data systems, including searches in spectral libraries [9]. This *reliability* is similar in meaning to *reliability of identification* (see above). In the literature on identification of peptides and proteins by mass spectral match, the $(100-PPV)\%$ rate is named *false discovery rate* (*FDR*) (e.g., see [11])

(that is, *FR*). Second, an estimation of *FR* with a reasonable accuracy ($\Delta_c FR/FR$ is at most 50 rel %) requires hundreds and even thousands of replicates. Table 4.5 gives the number of trials necessary for the evaluation of *FR* with the confidence probability of 0.95 and various levels of *FR* and $\Delta_c FR/FR$ calculated by the transformation of (4.3).

This large number of experiments is difficult and often unprofitable to perform. This is especially true for reliable methods having low false rates (1% and 5%, Table 4.5). In this case, many trials should be taken to obtain at least one false response. Table 4.6 gives the number of these replicates calculated using a binomial

**Table 4.4** Confidence intervals for *FR* at various *p*

| FR, % | *n* | $\Delta_c FR$, % | | | | $\Delta_c FR/FR$, rel % | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 0.90 | 0.95 | 0.99 | 0.999 | 0.90 | 0.95 | 0.99 | 0.999 |
| 1 | 10 | 5.2 | 6.2 | 8.1 | 10 | 518 | 617 | 811 | 1035 |
| | 100 | 1.6 | 2.0 | 2.6 | 3.3 | 164 | 195 | 256 | 327 |
| | 1000 | 0.5 | 0.6 | 0.8 | 1.0 | 52 | 62 | 81 | 104 |
| 5 | 10 | 11 | 14 | 18 | 23 | 227 | 270 | 355 | 454 |
| | 100 | 3.6 | 4.3 | 5.6 | 7.2 | 72 | 85 | 112 | 143 |
| | 1000 | **1.1** | **1.4** | **1.8** | **2.3** | **23** | **27** | **36** | **45** |
| 10 | 10 | 16 | 19 | 24 | 31 | 156 | 186 | 244 | 312 |
| | 100 | **4.9** | 5.9 | 7.7 | 9.9 | **49** | 59 | 77 | 99 |
| | 1000 | **1.6** | **1.9** | **2.4** | **3.1** | **16** | **19** | **24** | **31** |

Values set in bold correspond to the condition $\Delta_c FR/FR < 50$ rel %

**Table 4.5** Number of method trials for $p = 0.95$

| *FR*, % | $\Delta_c FR/FR$, rel % | | | |
|---|---|---|---|---|
| | 20 | 30 | 40 | 50 |
| 1 | 9,508 | 4,226 | 2,377 | 1,521 |
| 5 | 1,825 | 811 | 456 | 292 |
| 10 | 864 | 384 | 216 | 138 |

**Table 4.6** The most probable number of trials necessary to obtain at least one false result

| *FR*, % | $P = 0.95$ | $P = 0.99$ |
|---|---|---|
| 1 | 299 | 459 |
| 5 | 59 | 90 |
| 10 | 29 | 44 |

distribution. The best way out of this situation is that one should take a relatively small number of trials of a standard sample containing (or not containing) the analyte in the necessary amount [14]. If no false responses are obtained, one can state that the false rate is below the corresponding value given in Table 4.6. Thus, the absence of false responses in 59 analyses means that the false rate is below 5%, with a confidence probability of 0.95.

The above statistics characterize samples with the same amount of analytes. However, *FR* can be estimated for not only concentration points but also ranges. For example, these rates are calculated for all the samples used for estimating performance functions at different concentrations of analytes. In this case, *FPR* or *FNR* is some central value or a value interval for a concentration/amount range.

The same may refer to a series of samples different in not only levels of analytes but also a composition of matrices, e.g., real biochemical samples. For the latter, *FR* are evaluated by using confirmatory GC–MS methods (Sect. 4.2.7).

For estimating *FR* and related performances in interlaboratory studies, the overall number of laboratories and trials per laboratory and typical numbers of

replicates per material will be outlined in Sect. 9.6. In any case, many tens or even hundreds of replicate experiments are required to provide accurate estimates of rates for qualitative analysis.

### 4.2.2 Statistics of True Results

Equations analogous to (4.1) and (4.2) can be used for calculating true result rates, which in medicine and toxicology [6, 7] are called *sensitivity* and *specificity* (Table 4.3). Sensitivity, *St*, is given by the formula:

$$St = \frac{100 \times n_{TP}}{n_{TP} + n_{FN}}\%. \tag{4.4}$$

The equation for the specificity, Sp, is written as follows:

$$Sp = \frac{100 \times n_{TN}}{n_{TN} + n_{FP}}\%. \tag{4.5}$$

This concept of sensitivity differs from that used in state-of-the-art analytical chemistry and metrology ("the slope of the calibration curve" [15]), and agrees with another interpretation of this term inversely related to a *detection limit* [16]. Indeed, if the qualitative method allows small amounts of a compound to be detected (low detection limit), the sensitivity, i.e., the true positive rate calculated by (4.4), is high.

The above definition of the specificity of a qualitative method coincides with the term *specific* recommended by IUPAC. This term

> . . . expresses qualitatively the extent to which other substances interfere with the determination of a substance according to a given procedure. Specific is considered to be the ultimate of selective,[1] meaning that no interferences are supposed to occur [17].

Indeed, interfering substances cause false positives, which reduces the specificity calculated by (4.5). Thus, treatment of concepts of *sensitivity* and *specificity* originated from medicine, etc, does not conflict with that typical for chemistry. Nevertheless, in order to not mix up the two interpretations, treatments corresponding to (4.4) and (4.5) will be specified below using the term of *statistical* or the notations of *St* (or *TPR*) and *Sp* (*TNR*).

The four rates (false or true, negative or positive) are not independent. They are related in pairs: $FNR + St = 100\%$ and $FPR + Sp = 100\%$. These relationships can be easily derived using (4.1), (4.2), (4.4), and (4.5). Therefore, in characterizing reliability of the qualitative determinations, both false and true rates can be used.

[1] I.e., specificity is 100% selectivity.

### 4.2.3 Replication

In replication of qualitative testing for the same samples, the probabilities of the positive and negative responses are multiplicative (e.g., see [8, 18]). Thus, if the sample contains the analyte, and the probabilities of positive and negative responses are equal to *St* and *FNR* respectively (which is true for reasonably large number of trials), retrying the same test gives the following:

- Probability of the positive response $= St \times St$
- Probability of the negative response $= FNR \times FNR$
- Probability of the uncertain response (positive and negative responses in different order) $= 2 \times St \times FNR$

Setting aside the uncertain responses and evaluating the outcome by the positive-to-negative ratio (i.e., odds), one can conclude that duplicate testing reduces the probability of error.

**Example 4.1** For the analyte present in the sample, qualitative determination with the false negative rate of 5% has the odds in favor of the true (positive) response equal to $St/FNR = (1-0.05)/0.05 = 19/1$. The odds in favor of the true response after two trials is $(0.95/0.05)2 = 361/1$, that is, much higher than in the first case.

Nevertheless, it is unknown a priori whether the analyte is present in the real sample. Therefore, the positive response obtained can be either true or false. The chances are determined by the corresponding probabilities. In this case, this is the sensitivity-to-false positive rate ratio.

**Example 4.2** Screening by the test method with the *FNR* of 5% and *FPR* 3% has the odds in favor of the true positive response equal to $(1-0.95)/0.03 \approx 32/1$. A repeated trial substantially increases the odds (to about 1,000/1).

The latter example implies that prior probabilities of the presence or absence of the analyte in the sample are equal. In general, this condition is not met, and the trueness of chemical testing is to be assessed with the samples with and without the analyte. Therefore, having prior statistical information about the sample composition substantially reduces the unreliability of screening tests (see below). This is also achieved by using confirmatory methods (another simple qualitative method or a more complicated instrumental technique); see Chap. 6 and [6, 18].

Thus, the screening run reveals the positive and negative responses, and the repeated analysis is performed with one of these sample series (as a rule with positives) to evaluate the rate of the true and false responses.

### 4.2.4 Predictive Values

Rates of *St* and *Sp* unambiguously characterize the efficiency of qualititative methods when all the samples contain or do not contain the analyte. For the samples containing the analyte with a certain probability, different numerical characteristics are used.

Chemical testing of an arbitrary sample can give positive or negative responses; in both groups, true and false responses are observed. The proportion of true responses is characterized by the *predictive values* [6, 7, 14]. Two types of predictive values are distinguished, positive predictive value *PPV*

$$PPV = \frac{100 \times n_{TP}}{n_{TP} + n_{FP}}\% \tag{4.6}$$

and negative predictive value *NPV*

$$NPV = \frac{100 \times n_{TN}}{n_{TN} + n_{FN}}\% \tag{4.7}$$

Equations (4.6) and (4.7) pertain to a series of $n$ identical samples or similar ones (the same matrix), where $n = n_{TP} + n_{FP}$ or $n = n_{TN} + n_{FN}$ respectively.

For real samples, (4.6) and (4.7) are hard to use. The reason is that it is impossible to distinguish between true and false responses, i.e., TP from FP and TN from FN. Nevertheless, if the proportion of the samples containing the analyte, i.e., *prevalence Pv* (see Table 4.3) is known, the predictive values can be calculated by the equivalent formulae using sensitivity *St* and specificity *Sp* [7].

$$PPV = \frac{100 \times St \times Pv}{(100 - Sp)(100 - Pv) + St \times Pv}\% \tag{4.8}$$

$$NPV = \frac{100 \times Sp \cdot (100 - Pv)}{(100 - St) \cdot Pv + Sp \cdot (100 - Pv)}\%. \tag{4.9}$$

The pair correspondence of (4.6) and (4.8) with (4.7) and (4.9) can be easily demonstrated using definitions (4.4) and (4.5).

Estimating the *PPV* values being of concern in state-of-art statistical tests of various identification methods (see footnote to Table 4.3) is exemplified further.

The above example requires an additional comment. Positive and negative predictive values for each individual sample can be calculated if the prevalence

**Example 4.3** A qualitative method with performances *St* and *Sp* of 99% is applied to a series of 1,000 samples containing the analyte. These are, e.g., body fluids taken from people with a certain disease, or soil samples taken in a relatively compact area. One would expect 990 (99%) true positives according to (4.4). Since the false positive responses are excluded in this case, the predictive value *PPV* in this case is 100%; see (4.6) or (4.8).

In the other case, the same test is applied to another series of 1,000 samples, of which one tenth ($Pv = 10\%$) contains the detectable amount of the analyte. Analytical procedures followed by calculations by (4.1), (4.2), (4.4), and (4.5) lead to the following results:

$n_{\mathrm{TP}} = (1{,}000 \times 0.1 \times 0.99) = 99$,
$n_{\mathrm{FN}} = (1{,}000 \times 0.1 \times 0.01) = 1$,
$n_{\mathrm{TN}} = (1{,}000 \times 0.9 \times 0.99) = 891$,
$n_{FP} = (1{,}000 \times 0.9 \times 0.01) = 9$.

It is easy to demonstrate that the results of this method correspond to its rates. Indeed, $St = (99 \times 100)/(99 + 1) = 99\%$ and $Sp = (891 \times 100)/(891 + 9) = 99\%$.

However, the positive predictive value derived from analytical experiments, which characterizes not only the method but testing of the individual samples as well, is less than 99% and equals $(99 \times 100)/(99 + 9) = 91.7\%$. Therefore, each positive response can be false with a rather high probability (more than 0.08). This results from a relatively large number of false positives due to a large number of samples containing no analyte (900 samples), although the probability of the false positive response is low (1%). Hence, depending on the previously established criteria of reliability, further confirmation of the determination results may be necessary.

of the analyte is known [(4.8)]. These data are mostly lacking, but they can be obtained for a series of samples of a similar composition from the results of analyzing this series. Indeed, assume that the numbers of positive and negative responses correspond to the values above. In this case $Pv = (n_{TP} + n_{FP}) \times 100/n = (99 + 9) \times 100/1000 = 10.8\%$; this value differs only slightly from the a priori value of 10.0%. Taking into account the prevalence of 10.8%, we obtain a *PPV* of 92.3%, which also differs only slightly from the above value of 91.7%.

Positive predictive values for the high sensitivity *St* of 99% and various specificity *Sp* and prevalence values are given in Table 4.7. It is easy to see that if more than half of the samples contain no analyte ($Pv < 50\%$), the positive response level becomes lower than the nominal specificity. Thus, if the analyte is present in less than 5% of samples and *Sp* is 95%, more than half of the positive responses become false. For the statistical sensitivity and specificity of 90%, this result is obtained for the prevalence below 10% (Table 4.8).

**Table 4.7** Positive predictive values for 99% statistical sensitivity and various specificity and prevalence values

| *Pv*, % | *Sp*, % | | |
|---|---|---|---|
| | 99.9 | 99.0 | 95 |
| 100 | 100 | 100 | 100 |
| 75 | 100 | 99.7 | 98 |
| 50 | 99.9 | 99.0 | 95 |
| 25 | 99.7 | 97 | 87 |
| 10 | 99.1 | 92 | 69 |
| 5 | 98 | 84 | 51 |
| 1 | 91 | 50 | 17 |

**Table 4.8** Positive predictive values for 90% statistical sensitivity and specificity and various prevalence values

| *Pv*, % | *PPV*, % |
|---|---|
| 100 | 100 |
| 75 | 96 |
| 50 | 90 |
| 30 | 79 |
| 25 | 69 |
| 10 | 50 |
| 5 | 32 |
| 1 | 8 |

Hence, if samples are tested in which the analyte is rare in occurrence, its detection frequently can be false. Similar large errors occur in the quantitative determination of the analyte at low concentrations.

### *4.2.5 Bayesian Approach*

Equations (4.8) and (4.9) used for estimating the reliability of qualitative methods can also be derived using the Bayesian approach to probability, taking into account prior data (see Sect. 3.5 and the references therein). Indeed, the prevalence of the samples containing the analyte can be known prior to the experiment as prior information. Analysis of the samples implies its transformation into more reliable a posteriori information as a result of the experiment.

This situation is described by the Bayesian equation (3.3), which can be also represented in the form:

$$p(A|result) = \frac{p(result|A) \cdot p(A)}{p(result|A) \cdot p(A) + p(result|\overline{A}) \cdot p(\overline{A})}, \tag{4.10}$$

where $p(A|result)$ is the conditional probability of the presence of the analyte $A$ in the sample, given the analysis *result* is obtained; $p(result|A)$ is the conditional probability of the positive response in the presence of the analyte; $p(A)$ is the probability of the presence of the analyte; $p(result|\overline{A})$ is the conditional probability

of the positive response in the absence of the analyte; $p(\overline{A})$ is the probability that the analyte is absent.

In this case, the probability $p(result|A)$ equals the proportion of the true positive responses at a reasonably large number of samples tested, each of them containing the analyte, i.e., the sensitivity $St$. In the same conditions, the other probability, $p(result|\overline{A})$, is $FPR$ or $100\%-Sp$. The probability $p(A)$ equals the prevalence, and so on. Thus transforming (4.10) and taking into account (4.8), one obtains the following:

$$p(A|result) = PPV. \tag{4.11}$$

The corresponding Bayesian conditional probability $p(\overline{result}|\overline{A})$, i.e., the probability of the absence of the analyte $A$ given the negative response, is equivalent to the negative predictive value, as can easily be demonstrated by modifying (4.10). This equivalence as well as (4.11) implies that Bayesian a posteriori probabilities are related to the sensitivity, specificity, and prevalence in the same way as predictive values in (4.8) and (4.9).

### 4.2.6 Prior Data and Replication

A prior probability in the Bayesian equation (4.10) can either pertain to the amount of the analyte in the sample known a priori or represent the results of another, initial, screening test. In this case, conditional probabilities, including a posteriori probability $p(A|result)$, pertain to the second analytical experiment performed in parallel with the first one or after it. As shown above, duplicate determination gives more reliable results than each method performed separately. The particular estimation is exemplified below.

**Example 4.4** A qualitative method with a statistical sensitivity and specificity equal to 90% is used for screening of 300 samples, 30% of them containing the analyte. The most probable results are the following:

$n_{TP} = 300 \times 0.3 \times 0.9 = 81$,
$n_{FN} = 300 \times 0.3 \times 0.1 = 9$,
$n_{TN} = 300 \times 0.7 \times 0.9 = 189$,
$n_{FP} = 300 \times 0.7 \times 0.1 = 21$.

The method exhibits the following predictive values:

$PPV = (81 \times 100)/(81 + 21) = 79\%$ (corresponds to the data of Table 4.8),

$NPV = (189 \times 100)/(189 + 9) = 95\%$.

*(continued)*

Therefore, 102 positive responses are obtained, and only 79% of them are true. Repeated testing with the same specificity and sensitivity gives improved results:

$n_{TP} = 102 \times 0.79 \times 0.9 = 73$,
$n_{FN} = 102 \times 0.79 \times 0.1 = 8$,
$n_{TN} = 102 \times 0.21 \times 0.9 = 19$,
$n_{FP} = 102 \times 0.21 \times 0.1 = 2$.

Simultaneous use of the two tests gives the predictive value
$PPV = (73 \times 100)/(73 + 2) = 97\%$.

This implies that each positive response of the chemical test is true with a high probability of 97%. Therefore, further confirmation of the true responses is not necessary.

Here, the prevalence values (30%) were given, whereas they are unknown in general. However, they can be calculated from the results of the first test. Indeed, the number of positive responses is $81 + 21 = 102$, and, therefore, $Pv = 102 \times 100/300 = 34\%$. Using this value instead of 30% for calculating the predictive value gives 82 and 98% positive predictive values for the first and duplicate analysis respectively. These values differ only slightly from those calculated using the prior probability.

In general, the result of the two combined analyses for the case when only the samples with the positive response of the first test are retested can be expressed by the equation obtained by transforming (4.8) or (4.10)

$$CPPV = p(A|result)_{1,2} =. \frac{100 \times St_2 \times PPV_1}{(100 - Sp_2)(100 - PPV_1) + St_2 \times PPV_1}\%, \quad (4.12)$$

where $CPPV$ is the cumulative positive predictive value of these analyses, $p(A|\text{result})_{1,2}$ is the conditional probability of the presence of the analyte in the sample given the certain test results, and subscripts 1 and 2 pertain to the first and second experiment respectively. Cumulative positive predictive values calculated for various combinations of the prior probability of the presence of the analyte in the sample, statistical sensitivity and specificity are given in Tables 4.9 and 4.10.

Table 4.9 shows that the cumulative predictive value of the two independent tests does not depend on the order in which they are performed. Thus, if one method has the sensitivity and specificity of 90%, and the other one has the sensitivity and specificity of 95%, the cumulative value for any order of qualitative analytical experiments is 98.7%. Nevertheless, a different number of samples is tested in this case, and one of the two experiments (the first one is more specific, and the second one is more sensitive) is less laborious [6, 7]; see Example 4.5.

**Example 4.5** Three hundred samples with the analyte prevalence of 30% are taken for analysis. Two qualitative methods are used: the first one having a statistical sensitivity of 90% and specificity of 95%, and the second one having a sensitivity of 95% and specificity of 90%. If the first one is used first, true positive responses are recorded for 81 samples, and false positive responses are recorded for 10 or 11 samples. Hence, only 91 or 92 samples need retesting. If the second method is used first, 85 or 86 true positive and 21 false positive responses are recorded after the first stage. In this case, as many as 106 or 107 samples need retesting, which is less cost-efficient.

**Table 4.9** Predictive values of single and duplicate tests for 30% prevalence and various sensitivity and specificity values

| Method 1 | | Method 2 | | *PPV*, % | | *CPPV*, % |
|---|---|---|---|---|---|---|
| *St* | *Sp* | *St* | *Sp* | method 1 | method 2 | methods 1 + 2 |
| 90 | 90 | 90 | 90 | 79 | 79 | 97.2 |
| 90 | 95 | 90 | 95 | 89 | 89 | 99.3 |
| 95 | 90 | 95 | 90 | 80 | 80 | 97.5 |
| 95 | 95 | 95 | 95 | 89 | 89 | 99.4 |
| 90 | 90 | 95 | 95 | 79 | 89 | 98.7 |
| 95 | 95 | 90 | 90 | 89 | 79 | 98.7 |
| 90 | 95 | 95 | 90 | 89 | 80 | 98.7 |
| 95 | 90 | 90 | 95 | 80 | 89 | 98.7 |

**Table 4.10** Predictive values of single and duplicate tests for sensitivity and specificity of 95% and various prevalence values

| *Pv*, % | *PPV*, % (method 1 or 2) | *CPPV*, % |
|---|---|---|
| 0.1 | 2 | 27 |
| 1 | 16 | 78 |
| 5 | 50 | 95.0 |
| 10 | 68 | 97.6 |
| 20 | 83 | 98.9 |
| 30 | 89 | 99.4 |
| 40 | 93 | 99.6 |
| 50 | 95 | 99.7 |

### *4.2.7 Screening of Real Samples*

Rapid test methods are widely used in biochemical, clinical, and toxicological analysis. Table 4.11 gives the typical false positive and false negative rates and also other performances for the screening detection of abused drugs in urine and other matrixes of humans (average concentrations in the range of 1 ng/ml–1 μg/ml). Different techniques of immunoassay and chromatography were used. Gas chromatography mass spectrometry was used as a confirmatory method unambiguously determining the presence or absence of these drugs down to a low concentration of about 1 ng/ml.

**Table 4.11** Rates for immunoassay screening of sample containing abused drugs. Confirmation by GC–MS

| Sample | Rate, % | | | | | |
|---|---|---|---|---|---|---|
| | *FPR* | *FNR* | *St* | *Sp* | *PPV* | *NPV* |
| Opiates, cocaine, and so on in urine [7][a] | 0.4–16 | 1.5–43 | 44–98.5 | 84–99.7 | 22.5–98.9 | 85–99.6 |
| | 0–3.8 | 18–77 | 23-–82 | 96–100 | 79–100 | 75–96 |
| | 0–2.2 | 21–92 | 4.5–79 | 98–100 | 23–100 | 68–96 |
| | 0–0.4 | 16–29 | 71–84 | 99.6–100 | 94–100 | 82–97.5 |
| Amphetamine, opiates, and so on in urine [19][b] | | | | | 71–100 | |
| Cannabis and codeine in saliva [20][c] | | | 90–100 | 90–100 | | |
| Amphetamine, opiates, and so on in urine [21][d] | | | 32–95 | 41–98 | 58–93 | 62–90 |
| Amphetamine, opiates, and so on in urine [22][e] | 0–3.8 | 0–0.9 | | | | |
| Methamphetamine in hair [23][f] | | | 88–97 | 93–100 | | |
| Cocaine and its metabolites in urine [24][g] | | | 86–98.4 | 88–100 | 7.5–100 | 96.5–100 |

[a]Screening results by TLC, LC, and HPLC in second, third, and fourth lines respectively. Some samples (2–18%) contain analytes under their detection limits (in most cases 25–300 ng/ml)
[b]Detection limits are 200–300 ng/ml. TLC is another confirmatory technique. Ranitidine interferes with determination of amphetamine
[c]The detection limit is 10 ng/ml
[d]The analyte concentration is 15–700 ng/ml
[e]Detection limits are 4–2,000 ng/ml; HPLC–MS ($MS^2$) is another confirmatory technique
[f]Detection limits are 0.75–1 ng/ml
[g]The analyte concentration is 50–300 ng/ml, prevalence is 1–20%

Table 4.11 shows that false rates are inevitable in screening methods. In some cases, false responses occur in more than 10% cases. However, a high false negative rate may result from relatively low concentrations of the analytes at and below detection limits. Hence, these responses are false just relative to gas chromatography mass spectrometry data.

Quantitative estimates of errors of qualitative methods also come into other fields of analysis. Examples of such screening analyses are (see also Chap. 8):

- Environmental health analysis: the determination of lead in air and paint films [13, 25, 26]
- Qualitative analysis II: the determination of hazelnut oil in olive oil based on measuring triglycerides and sterols by GC and HPLC [27]

### *4.2.8 Other Indices*

In toxicology and some other fields, the index of the mean list length (MLL) has been proposed to define the overall efficiency of the identification method/

technique applied to a sample of analytes [28, 29]. This is one of the indices of selectivity. In common screening, a result of identification is unambiguous for some analytes and ambiguous for other detected compounds. In the second case, candidates for identification are partially (within corresponding lists) indistinguishable. By definition, the MLL is an average number of candidate compounds per one analyte from some analyte group: $N/n$, where $N$ is the overall number of candidates and $n$ is the number of compounds in a group. If the MLL is equal to 1, there is the case of unambiguous individual identification of a group of compounds. The value of 2 signifies that an average list of two candidates for identification per each analyte is set up by a method. The index under consideration is estimated for model samples of compounds.

Another index of selectivity, the discriminating power (DP), is also sometimes reported (e.g., see [29]). The DP expresses the probability that two compounds from a sample of analytes can be distinguished by the method/technique. This index is $1-2n_p/n(n-1)$, where $n_p$ is the number of indistinguishable pairs of compounds, n is again their overall number. The index DP clearly varies from 0 (all the compounds from a group are indistinguishable and therefore $n_p = n(n-1)/2$) through 1, i.e., full recognition of all the analytes under analysis and $n_p = 0$.

## 4.3 Concentration Dependence of Detection and Identification Results

In the determination of a low amount/concentration of a substance, an analytical method is near its capability limits. In consequence, detection and especially identification of an analyte may be unreliable. Indeed, a low signal may be related to (a) a noise, or (b) a foreign compound rather than a target one. Their interference, as well as an occurrence of the analytical signal itself, is clearly dependent on analyte concentration.

### *4.3.1 Binary Responses*

General outcomes of screening in terms of true and false rates depend on the concentration (amount) of the analyte. When the concentration decreases, TP rate begins to decrease at a certain level (Fig. 4.2). This results from the following two factors. The first is related to statistics, and consists of the fact that the proportion of weak analytical signals in their intensity distribution increases at this concentration. This can be manifested by the absence of

- The signal because it is masked by the noise background in the instrumental detection and
- Any visible color changes in the sensory detection.

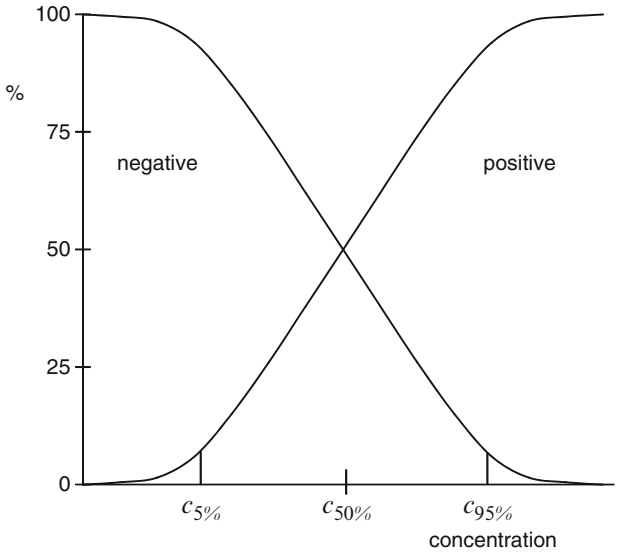



**Fig. 4.2** Percentage of positive and negative responses of a qualitative method as a function of the analyte concentration (amount). The concentrations $c_{5\%}$ and $c_{95\%}$ correspond to 5 and 95% of positive responses. In these two cases, false negative rates are 95 and 5% respectively. For $c_{50\%}$, proportions of positive and negative responses as well as true and false results are the same

The second factor causing false responses at low analyte concentrations is mainly typical of qualitative reactions. These responses result from incompleteness of the reaction and masking effects [14]. Analogous effects can be also observed in instrumental analysis. For example, ESI of many analytes is suppressed by surfactants [30, 31]. Therefore, low analyte concentrations are unavoidably associated with negative responses. The latter can be called conditionally true negatives if the corresponding concentrations are below the established regulatory values or other values of practical importance. Obviously, the false negative level becomes reasonably low for relatively high analyte concentrations.

Conversely, false positives can be detected for blank samples, and they also can be of two sorts. They can arise from the fluctuation of the noise level of the instrument detector, or from the detection of some interfering substances.

Figure 4.2 shows the concentration dependence of results of common chemical tests or other screening techniques. The highest error rate is found at relatively low concentrations where the true rate is nonzero but does not reach 100%. The analyte concentration range where this rate is between 5 and 95%, i.e., from $c_{5\%}$ to $c_{95\%}$ (see Fig. 4.2) is called the *unreliability region* [32, 33]. Similar ranges are also called the region of *unreliable* or *unsteady reaction*, the region of *unreliably detected concentrations* (see [14]), and the *uncertainty interval* [13].

The boundaries of this uncertainty interval can characterize the chemical test (qualitative method). Thus, the procedure for determination of lead in the air (spot

**Table 4.12** Terms for special concentration points in the unreliability region

| Concentration | Terminology |
|---|---|
| $c_{5\%}$ | Cut-off [34], low cut-off [35], low value of unreliability region [36] |
| $c_{50\%}$ | Identification limit [13, 25], mean detectable concentration [37], |
| $c_{95\%}$ | Cut-off [34], high cut-off [35], limit of detection [38], screening limit [34, 35], upper value of unreliability region [36] |
| $c_{99,7\%}$[a] | Reliably detectable concentration, reliably detectable minimum [37] |

[a]Defined in the same way as other points; see Fig. 4.2

test with rhodizonate) has the following performance characteristics for the confidence probability of 0.95 [25]:

- positive > app. 10 μg Pb
- negative < app. 0.6 μg Pb

This implies that for samples with a lead content higher than 10 μg and lower than 0.6 μg, the test will respond positively or negatively respectively, with a probability of 0.95. The reverse is also true: for the positive response, the lead content is higher than 10 μg, and for the negative response, the lead content is lower than 0.6 μg with the same probability.

Various points at the boundaries of unreliability regions and in their centers also have special names, although definitive terminology for binary response determinations (Table 4.12) is not yet established.

The plots in Fig. 4.2 indicating the integral distribution of the positive and negative responses are called performance functions/curves/graphs [4, 14, 39, 40]. These functions are approximated by normal, logarithmically normal, logistic, exponential, the Weibull distribution and other distributions [13, 14, 25, 39, 40]. No significant difference has been found between the approximations; sometimes, preference has been given to the exponential, logarithmically normal, or normal distributions (references, see [14]). In a recent book [40], only two functions, logistic and exponential ones, were proposed to be tested.

## *4.3.2 Measurands*

### 4.3.2.1 Detection

In the previous section, just analytical signals as yes/no responses were treated. Measured signals are also concentration-dependent. The examples of those for low amounts are given in Fig. 4.3. Such limit performances, essential in many analytical problems, e.g., a doping determination, food, and environmental analysis, are further considered.

Signals are observed/recorded only if they exceed a noise level. The low limit concentration (if observed) should correspond to low probability of the noise level coming up to the intensity of such an analytical signal. This probability derived

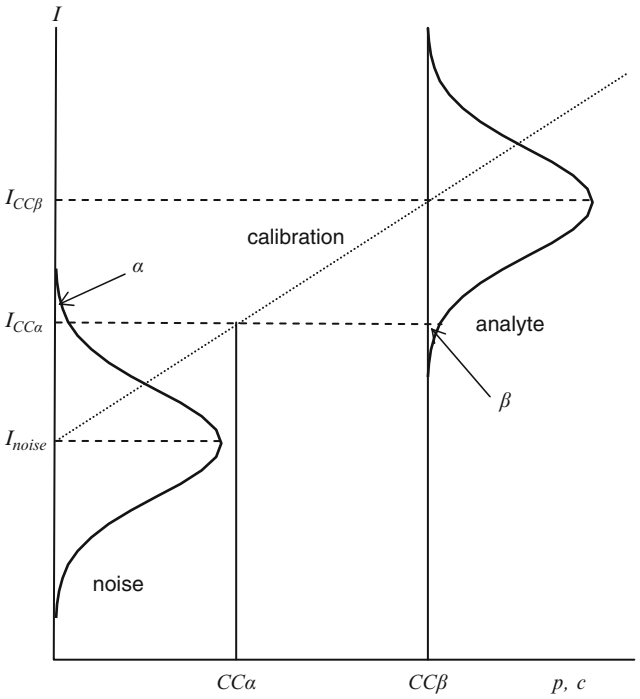



**Fig. 4.3** Probability distributions of an intensity of an analytical signal and a noise. This is Fig. 3.7 adapted to the case (see also [41, 42]). The vertical axis shows the signal intensity. The horizontal axis demonstrates both the probability of the particular signal and the concentration/amount of an analyte. The noise curve is recorded for a blank sample containing a very low amount of an analyte. The most probable level of the noise corresponds to the amplitude $I_{noise}$. Two other signal intensities shown here are related to a low and the most probable signals of an analyte at the limit concentration $CC\beta$. This, and another point $CC\alpha$, are discussed in the text

**Table 4.13** Terms for limit concentration points

| Critical concentration | Terminology |
|---|---|
| $CC\alpha$ | Critical value [43], critical value of the net state variable [41], decision limit [44] |
| $CC\beta$ | Detection capability [41, 44], detection limit [43], minimum detectable value [43] |

from the noise distribution is α, see Fig. 4.3. The α value 0.05 is usually established. The calibration graph shows that corresponding analyte concentration is critical concentration $CC\alpha$. There are several names for this limit concentration (Table 4.13).

Thus, if the analyte concentration is equal to or is higher than $CC\alpha$, there is the probability α that the corresponding signal with the intensity $I_{CC\alpha}$ is the noise.

However, there is a relatively low probability (probability of FP). So, an analyst decides that such a signal refers to the analyte.

With increasing concentration, the signal also rises. For the concentration $CC\beta$, the most probable amplitude of this analytical signal is $I_{CC\beta}$. (Fig. 4.3). Due to the statistical nature of the signal, its amplitude may be as low as $I_{CC\alpha}$, depending on the value $\alpha$. The probability of that outcome (FN) is $\beta$. It should be noted that this is not the only name for the limit quantity $CC\beta$ (see Table 4.13).

Figure 4.3 demonstrates also that the signal in the point $CC\alpha$ refers to both noise and the analyte with the same probability. Here, $I_{CC\alpha}$ is the most probable signal (not shown) for this concentration. There is a relatively low probability (e.g., the same value $\alpha$) that the signal amplitude for the level $CC\alpha$ measures up to $I_{CC\beta}$. For the intensity $I_{CC\beta}$ and higher, there is almost complete certainty that one observes the analytical signal.

The concentrations $CC\alpha$ and $CC\beta$ are calculated based on established values $\alpha$ and $\beta$, standard deviations $\sigma$ (or $s$) of intensities of both noise and analytical signal, and calibration dependence for the involved analytical experiment. In general form, those quantities are (see [43]):

$$CC\alpha = z(1-\alpha)\cdot\sigma_{bl}, \tag{4.13}$$

$$CC\beta = CC\alpha + z(1-\beta)\cdot\sigma_{an}, \tag{4.14}$$

where $z(1-\alpha)$ and $z(1-\beta)$ are the $(1-\alpha)^{\text{th}}$ and $(1-\beta)^{\text{th}}$ quantile respectively of the normal distribution; $\sigma_{bl}$ and $\sigma_{an}$ are the standard deviation of the signal intensity for the noise (blank) and analyte respectively.

Equations (4.13) and (4.14) are transformed into the explicit forms [42–46] to use in practical analysis. Different standard documents, the ISO 11843 standard [45], the German standard DIN 32645 [47]), and the Commission Decision 2002/657/EC [44], lead to the same or similar estimates for both values $CC$ [48].

For the calibration line $I = a\cdot c + I_{\text{noise}}$ (see Fig. 4.3), $\alpha = 0.01$ or 0.05, and $\beta = 0.05$; simple forms of those equations [42, 44] are

$$CC\alpha = \frac{1.645\cdot s_{bl}}{a}(\alpha = 0.05), \tag{4.15}$$

$$CC\alpha = \frac{2.33\cdot s_{bl}}{a}(\alpha = 0.01), \tag{4.16}$$

$$CC\beta = CC\alpha + \frac{1,645\cdot s_{an}}{a}, \tag{4.17}$$

where $s_{bl}$ and $s_{an}$ are the sample estimates for the standard deviation of the signal intensity for the noise (blank) and analyte respectively. The more replications of the measurement are carried out, the more accurate this approximation becomes. At least five replications have been reported [42, 45, 46]. As the level $CC\beta$ is not

known a priori, the analyte concentration and corresponding sample for estimating $s_{an}$ should be specially chosen. This may be [42, 44–46]

- The lowest concentration from samples taken for the calibration
- The level $CC\alpha$ at which the matrix is fortified with the analyte or
- The amount providing several times as much signal-to-noise ratio

For dangerous/banned compounds, permitted/legal/regulated limits are commonly established. In terms of an analytical purpose, the most important thing is to find out whether an analyte content exceeds or is not higher than some established value. In other words, it is necessary to determine a compliance with legal limits. In these cases, a blank level of a concentration (Fig. 4.3) is substituted by a permitted level to be treated [44].

In a more classical approach, a limit performance without calibration data is estimated.

- $CC\alpha$ is the concentration corresponding to the signal amplitude, which is three times the mean signal-to-noise ratio calculated for at least 20 repetitions of the analysis of blank sample or matrix fortified with the analyte compound at the regulated level [44].
- $CC\beta$ is estimated analyzing the samples containing the analyte at and above the level $CC\alpha$. The standard deviation $s_{CC\alpha}$ of the value $I_{CC\alpha}$ for the first from concentrations is further calculated. $CC\beta$ is estimated which, by definition, is the concentration at which the analytical signal is so large that only 5% of the lower signal values for $CC\alpha$ measures up that level. The equation for the calculation is: $I_{CCb} = I_{CC\alpha} + 1.645 \cdot s_{CC\alpha}$. Subsequently, the concentration $CC\beta$ inducing the response $I_{CCb}$ is chosen or calculated after the analysis of other samples and the recording of their signals. Again, at least 20 repetitions of the analysis of corresponding samples are recommended to be carried out [44].

In any case, if calibration data are not specially used, analysis of at least two matrices fortified with an analyte at different concentrations is required to estimate both limit concentrations.

For estimation of limit concentrations $CC$ or analogous quantities, the EU guidance [44] and ISO standard [45] are most commonly used. The document [44] provides rules for arrangement and estimation of characteristics of analytical methods for residues of substances having a pharmacological action and their metabolites transmitted to products of animal origin. Some estimates are shown in Table 4.14. This guidance is also of value for analysts in containing many useful approaches, rules, and criteria for routine chemical identification in the general case, although some items from the document have been criticized (see Sect. 5.6).

There is an analogy between a qualitative determination with the use of yes/no (Fig. 4.2) and measured (Fig. 4.3) responses with regard to critical concentrations. Indeed, there are the same levels of false results in points of $c_{50\%}$ and $CC\alpha$, $c_{95\%}$ and $CC\beta$. Furthermore, measurement results can be expressed not only as

**Table 4.14** Examples of limit concentrations of veterinary drugs in methods. Calculations according to [44]

| Residue, matrix, reference | Technique | Concentration level | $\alpha$ | $\beta$ | $CC\alpha$ | $CC\beta$ |
|---|---|---|---|---|---|---|
| Chloramphenicol in muscle [49] | HPLC–$MS^2$ | Blank | 0.01 | 0.05 | 0.15 µg/kg | 0.22 µg/kg |
| 3-Amino-2-oxazolidinone [a] in food [50] | HPLC–$MS^2$ | Zero | 0.01 | 0.05 | 0.14 µg/kg | 0.18 µg/kg |
| Tetracycline in methanol–water [51] | Fluorimetry | Blank<br>Permitted limit 100 µg/l | 0.05<br>0.05 | 0.05<br>0.05 | 13.1; 20.1[b] µg/l<br>123.6 µg/l | 25.3; 38.5 [b] µg/l<br>136.0 µg/l |
| Five penicillin antibiotics in muscle [52] | HPLC | Spiking from 25 to 300 µg/kg | 0.05 | 0.05 | From 26.6 to 341 µg/kg | From 29.1 to 380 µg/kg |
| Seven tetracycline antibiotics in milk [53] | HPLC | Spiking 100 µg/kg | 0.05 | 0.05 | 101–106 µg/kg | 104–109 µg/kg |
| Ten quinolone antibiotics in milk [54] | HPLC | Spiking from 6–8 to 100–103 µg/kg | 0.05 | 0.05 | From 7.8 to 103 µg/kg | From 9.2 to 105 µg/kg |

[a] Nitrofuran metabolite, calculations according to the ISO standard 11843-2 (2000) [45]
[b]Different methods

values of quantities but also converted into yes/no responses. Three examples of combinations of measurements and responses are noted.

1. For Hg monitoring in soils by atomic absorption spectrometry, the screening method was developed, and a narrow region of unreliability was determined (1.15–1.95 µg/g). The samples falling into this interval were destined to be further analyzed by use of a more accurate technique [55].
2. In detection of pesticide residues in vegetables by GC–ECD, lower and upper screening limits (probably corresponding to values $c_{5\%}$ and $c_{95\%}$; see Table 4.12) were estimated. The samples with responses which were equal to or higher than signals in points of lower screening limits were considered as non-negative. The GC–MS technique was further used for confirmation of an analyte presence in such samples [56].
3. Even such a performance analytical technique as GC–$MS^2$ may be just a screening one if peaks of the only product ion are recorded. Using such screening method and counting positive and negative results, unreliability intervals, from 3.7 (minimum low cut-off) to 22.2 µg/kg (maximum upper cut-off), were estimated for determination of about 130 pesticide residues in vegetables. Two or three other product ions ($MS^2$ transitions) were intended for confirmation of the screening result [36].

#### 4.3.2.2 Identification

Above the concentration $CC\beta$, analytes are reliably detected. However, this does not imply that those are reliably identified as well. For this purpose, another definition is introduced:

> . . . *limit of identification* (LOI). . . is defined as the lowest concentration for which the identification criteria are met [57].

This limit value is analogously treated in [4].

There is another similar definition for the critical value named *lowest concentration for identification* proposed by ISO standard guidelines for the identification of analytes by GC–MS in soils:

> . . . lowest concentration of the target compound, which, if present in the sample, can be identified using the identification criteria. . . [58]

This definition is obviously generalizable to various techniques and matrices. Thus, at the point $CC\beta$ (Fig. 4.3), the criteria established independently from analyte amount may be met or not met. There is also an intermediate case of so-called *indication* when some criteria are met and other ones are not met for the same analyte [58]. This is typical for identification of low amounts of organic compounds by MS based on several identification points. For GC–MS, an indication is treated in Example 4.6. The possibility of such incomplete identification of the pesticide analyte is illustrated in Fig. 4.4. The concentration dependence of how identification criteria are met is given in Example 4.7.

**Example 4.6** is based on the data from [58], and related to identification of the particular compound by GC–MS. The chromatographic peak and three mass peaks of characteristic/diagnostic ions are recorded for both standard and analyzed samples. Identification criteria are:

(a) the relative retention time *RRT* of the analyte differs from that of the standard by less than $\pm$ 0.2%

(b) the relative intensities *I* (relative to the highest characteristic peak in the mass spectrum of the standard solution) of all the characteristic ions in the mass spectrum of the sample do not differ by more than $\pm$ $(0.1 \times I_{std} + 10)$% from those determined in the spectrum of the standard, where $I_{std}$ is *I* of the characteristic ion in the mass spectrum of the standard.

The analytical experiment showed that analyte times *RRT* had deviated by less than $\pm$ 0.2% from the reference value determined with the standard calibration solution. So the *chromatography criterion* is met.

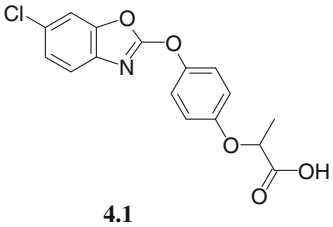
In mass spectra, three characteristic ions are selected. Corresponding ion currents when recording the mass spectrum of the most diluted calibration solution are 13.3, 9.75, and 12.8 (in $10^4$ a.u.). Calculation of relative values

*(continued)*

$I_{std}$ leads to 100, 73, and 96%, respectively. Than tolerance ranges for these values are evaluated. They are $\pm$ (0.1 $\times$ 73 + 10)% = $\pm$ 17.3% and $\pm$ (0.1 $\times$ 96 + 10)% = $\pm$ 19.6% respectively.

In mass spectra of the target analyte, peaks of the same three ions are present. Their currents and relative values $I$ are 1.42, 1.01, 1.02 (in $10^4$ a.u.) and 100, 71, 72% respectively. Subsequently, deviation of $I$ in the sample from the calibration standard are calculated. They are (71–73)/73% = −2.7% and (72–96)/96% = −25%, respectively. One can see that the deviation for the second ion (−2.7%) is within the permitted range ($\pm$ 17.3%, see above). In contrast, the deviation for the third ion (−25%) is outside the maximum range ($\pm$ 19.6%).

Thus, one from the MS identification criteria is not met. According to the rule set up in the standard [58], this is the case of indication rather than full identification, and at least one additional piece of evidence should be required for the reliable recognition of this target analyte. It is hardly surprising because those are low (i.e., not very reproducible) signals as compared to spectral peaks recorded for the calibration solution.




**4.1**

**Example 4.7** [59]. The EC criteria [44] (see Sect. 5.2.2.2) for the confirmation of the presence of anabolic steroids as illegal compounds present in biological matrices at concentrations from 0.5 through 5.0 μg/kg were tested. The analysis was carried out by GC–MS. Four characteristic ions were monitored, and three ion abundance ratios were calculated and compared to the criteria values. The latter were established using either standards or fortified samples as references. The proportion of ion ratios falling within the tolerance ranges as criteria was correlated to the S/N ratio of the least abundant from recorded characteristic ions. In general, it was concluded that at S/N=3 the percentage of ratios within the tolerances was $\leq$ 50%. With more intense signals, S/N$\geq$ 10, the proportion increased to more than 90%. In other words, the identification criteria were met for (a) $\leq$ 50% and (b) $>$ 90% of pairs of (a) very low peaks and (b) not very low peaks respectively.

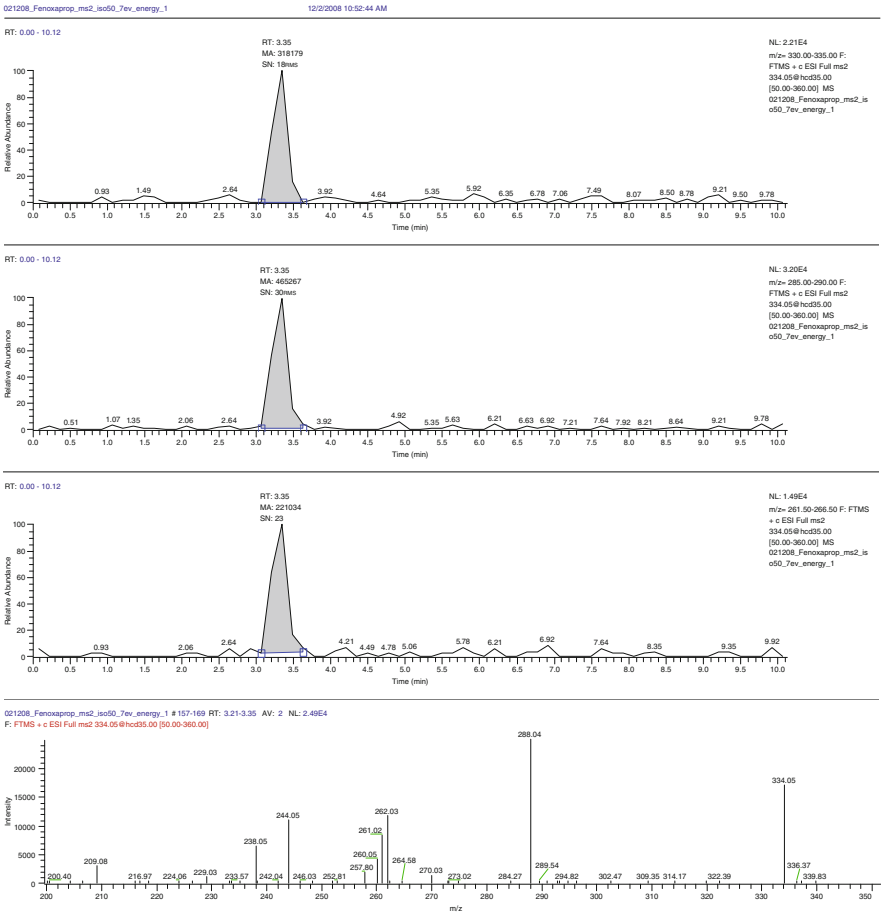



**Fig. 4.4** Mass chromatograms of three of the most intense peaks (*m*/*z* 334, 288, and 262, from *top* to *bottom*) and corresponding tandem mass spectrum of the solution of pesticide fenoxaprop **4.1** The 1.3 ng amount of the compound was injected in HPLC–$MS^2$ instrument (LIT-Orbitrap of Thermo, USA). Relative intensities of mass peaks, app. 70:100:50, will change with a decrease in the analyte amount due to non-zero interferences and the difference in S/N ratios (here 18, 30, and 23 respectively). This is the reason why one or two less intense peaks may be fallen outside of permitted ranges of their relative intensities

For combined qualitative and quantitative analysis, it is essential that analytes would be reliably identified at the least concentration measured (i.e., at the *limit of quantitation,*[2] for example see [57]). In MS and other spectral techniques, a

[2]This is

> the minimum concentration or mass of the analyte that can be quantified with acceptable accuracy and precision [60]

**Table 4.15** Identification thresholds for drug impurities [62]

| Maximum daily dose | Threshold [a] |
|---|---|
| <1 mg | 1.0% or 5 μg TDI, whichever is lower |
| 1–10 mg | 0.5% or 20 μg TDI, whichever is lower |
| >10 mg–2 g | 0.2% or 2 mg TDI, whichever is lower |
| >2 g | 0.10% |

[a]Thresholds given for impurities as degradation products are expressed either as a percentage of the drug substance amount or as total daily intake (TDI) of the degradation product. Lower thresholds can be established if the degradation products are unusually toxic

substance amount may be measured with different spectral peaks. The most intense peak is commonly selected for measurement (so-called *quantifier*). So, at the limit of quantification established for a quantifier signal, the relationship between the intensity of characteristic peaks used for identification must be within tolerance ranges, i.e., limit of identification $\leq$ limit of quantitation. From the above, it is also obvious that detection limit $<$ limit of identification (see also [57]).

The synonymic terms of *limit of identification* and *lowest concentration for identification* should be not confused with another one, *identification threshold*, as if it were analogous to them. The latter is

> a limit above... which a degradation product should be identified [62].

This concept expresses the need for identification rather than identification capability as compared with the above term. Identification threshold is related to the determination of impurities in pharmaceutical products [62] which is essential for registration application of new drugs. Now, the concept is also extended to the manufacture of chemical products (for example, see [63]). Thus if an impurity is contained in a pharmaceutical or chemical at or above an established percentage, that must be identified. Table 4.15 demonstrates the threshold values depending on the amount of drug substance administered per day (maximum daily dose).

In turn, the concept of identification threshold can be ambiguously interpreted. Indeed, this term occurs in publications on MS [64] and NIR [65] to specify a minimum spectral matching providing true result of identification (see Fig. 3.6). However, the term of *identification threshold* in its first understanding (limit

---

The limit of quantification in such definition is close to the concept of *minimum required performance limit* (MRPL), i.e.,

> ... minimum content of an analyte in a sample, which at least has to be detected and confirmed [44].

Modern MRPL values, e.g., setting for pharmaceutical residues in animal products are very low at 0.3 – 1.0 μg/kg [61].

There is another synonymic term for low limit amount which can be measured. It is *lowest calibrated level:*

> the lowest concentration (or mass) of analyte with which the determination system is successfully calibrated... [60]

amount above which analyte should be identified) is "more standard" than as introduced into the guidance [62].

## 4.4 Similarity of Spectra: Match Factors

### *4.4.1 General*

Spectral properties are among the most important for identification because of their uniqueness for many individual compounds and compound mixtures. Similarity of spectra of an unknown analyte and one from reference compounds means, or may mean, that this target is identified as just the reference. Here the key point is a measure (indicator, index, and so on) of similarity called *match factor*, MF (see Fig. 3.6), which is a function (or its value) of a similarity/difference of spectral measurands for two objects (chemical compounds and also their mixtures). An index MF expresses

- A similarity degree between spectra of the same compound (object)
- A difference degree between spectra of different compounds (object), and thus
- A criterion for accepting/rejecting identification hypotheses based on spectral search (Sect. 3.6.3); an identification error (Sect. 3.6.4).

It is a feature of analytical practice to treat MF by type of spectrometry. This is also due to the fact that different measures of similarity better fit to different spectral type. For example, dot products and correlation coefficients efficiently used to estimate a similarity in mass, IR, and UV-Vis spectrometry (see below) have been noted not to be suitable for NMR [66].

### *4.4.2 Mass Spectrometry*

#### 4.4.2.1 Classical Algorithms

Two basic algorithms have been used to calculate a similarity of classical mass spectra (EI–$MS^1$) of low molecules for spectral searches in large MS libraries (Sect. 7.4.1). In both, the match in peak masses and ion abundances is considered.

McLafferty proposed the search algorithm named '*probability-based matching*' (PBM) [67, 68]. The general essence of this approach is that

- The probability of a chance match of mass spectra of different compounds is not very high in most cases, thus
- Similarity of the unknown spectrum with that of a reference compound implies that they are most likely one and the same compound.

In the PBM algorithm, MF is the index (*reliability* or *match quality*; maximum value is 100) which is estimated on the base of the uniqueness of mass values of spectral peaks, the probability of different peak abundances, and other contributions [67, 68].

The second type of MF in mass spectrometry is *cosine function* or *dot product* [69–71]. These measures of mass spectral similarity are calculated by the formulas:

$$FOR = 1000 \cdot \left( \sum\nolimits_i^R \sqrt{I_i^R I_i^U} \right)^2 \Big/ \sum\nolimits_i^U I_i^U \cdot \sum\nolimits_i^R I_i^R \tag{4.18}$$

$$REV = 1000 \cdot \left( \sum\nolimits_i^R \sqrt{I_i^R I_i^U} \right)^2 \Big/ \sum\nolimits_i^R I_i^U \cdot \sum\nolimits_i^R I_i^R, \tag{4.19}$$

where $I_i$ is the peak intensity of the $i^{th}$ mass in the spectrum, and indices $R$ and $U$ refer to the reference and unknown spectrum respectively. Some peaks are present in only a reference or unknown spectrum that is accounted in sums with corresponding upper indices.

The direct factor *FOR* reflects a general similarity between compared spectra. That can be simply rationalized as being directly proportional to the cosine of the angle between two spectra as multidimensional vectors in the space of mass values (Fig. 4.5). $I_i$ is the component of such a vector in the direction of $i^{th}$ mass; see

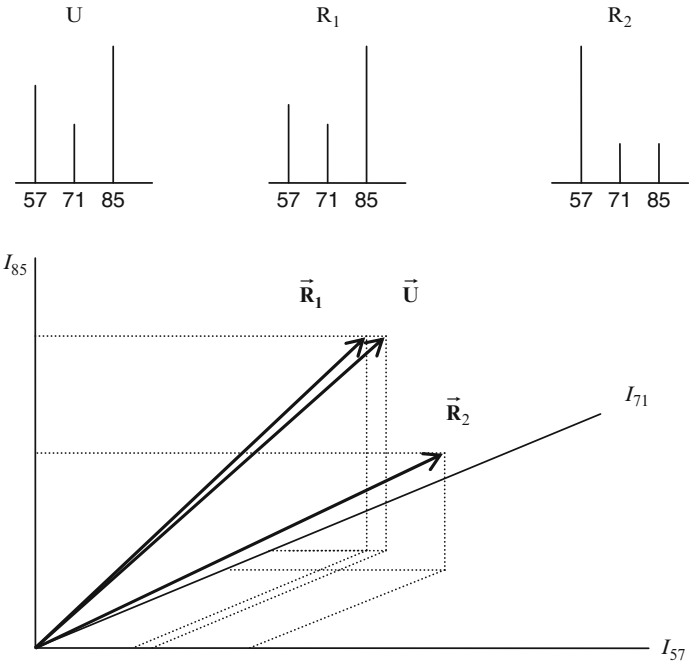



**Fig. 4.5** Three-peak mass spectrum of an unknown compound U and two reference mass spectra $R_1$ and $R_2$ and geometrical rationalization of these spectra as vectors $\vec{\mathbf{U}}$, $\vec{\mathbf{R}}_1$ and $\vec{\mathbf{R}}_2$ respectively. The directions refer to the mass 57, 71 and 85 of recorded ions. The intensities of peaks represent the length of the components of the vectors in these directions. The similarity of spectra U and $R_1$ and the distinction of these from the spectrum $R_2$ are evident. The angles between vectors $\vec{\mathbf{U}}$ and $\vec{\mathbf{R}}_1$,$\vec{\mathbf{U}}$ and $\vec{\mathbf{R}}_2$display the similarity and the distinction under consideration. The indices *FOR* are directly proportional to the cosine of the angles

Fig. 4.5. The reverse index *REV* resembles the *FOR* one as a measure of the degree to which the reference mass spectrum matches the mass spectrum of an unknown substance. *REV* is calculated for mass peaks which are contained only in the reference spectrum.

Another MF is spectral contrast angle [72]. This measure of spectral similarity is analogous to cosine/dot function (Fig. 4.5), but the angle itself rather than its cosine is calculated. Some other measures of mass spectral similarity have been also proposed, including modified cosine function [70].

#### 4.4.2.2 Modifications of MF

Matching factors were first used in computer algorithms to search spectra in spectral databases (libraries). With this purpose, MF formulas have been improved to represent in the best way

- Similarity of spectra of the same compounds recorded in somewhat different conditions.
- Spectral mismatching between different compounds.
- The same characteristics for tandem and/or high resolution mass spectra.
- The same characteristics for high-molecular compounds (proteomics).

One of the examples of such improvements was peak intensity scaling and ion mass weighting, to increase the significance of lower intensity and higher mass peaks respectively [70, 73]. This improvement (*composite algorithm*) was further included in NIST MS Search program for library search. Modern commercial software for searches in mass spectral libraries use this or similar MF and also PBM algorithm (Sect. 4.4.2.1). It should be noted that corresponding formulas are commonly hidden from users of commercial mass spectrometers; main users are analysts carrying out routine identification of volatile compounds using EI–$MS^1$.

Further tests and improvement of formulas for calculating MF were in line with a widespread application of HPLC–$MS^n$ and/or HRMS (Sects. 2.8.4 and 7.4) for both low-molecular and high-molecular compounds. Typical MF used in spectrometry such as the correlation coefficient [74] and the Euclidean distance [75] and also another index ("*R* score" [76]) were introduced and used in $MS^2$ libraries (Sect. 7.4.1.2). In searches with the use of the new $HRMS^n$ library, (a) the match in fragment $m/z$ within the narrow range ($\leq$ 0.01−0.1 Da) between an unknown and reference spectrum, and (b) the corresponding match degree in fragment peak intensities, and the number of (c) fragment ions and (d) matching fragment ions were taken into consideration [77].

In HRMS, the difference between experimental mass measured with high accuracy ($\leq$ a few ppm) and one of the theoretical/formula masses is taken into account. The formulas are generated by the special instrument software. This difference can be combined with that of isotopic ratios between compared data to predict candidate formula(s) required in identification procedures (Fig. 4.6).

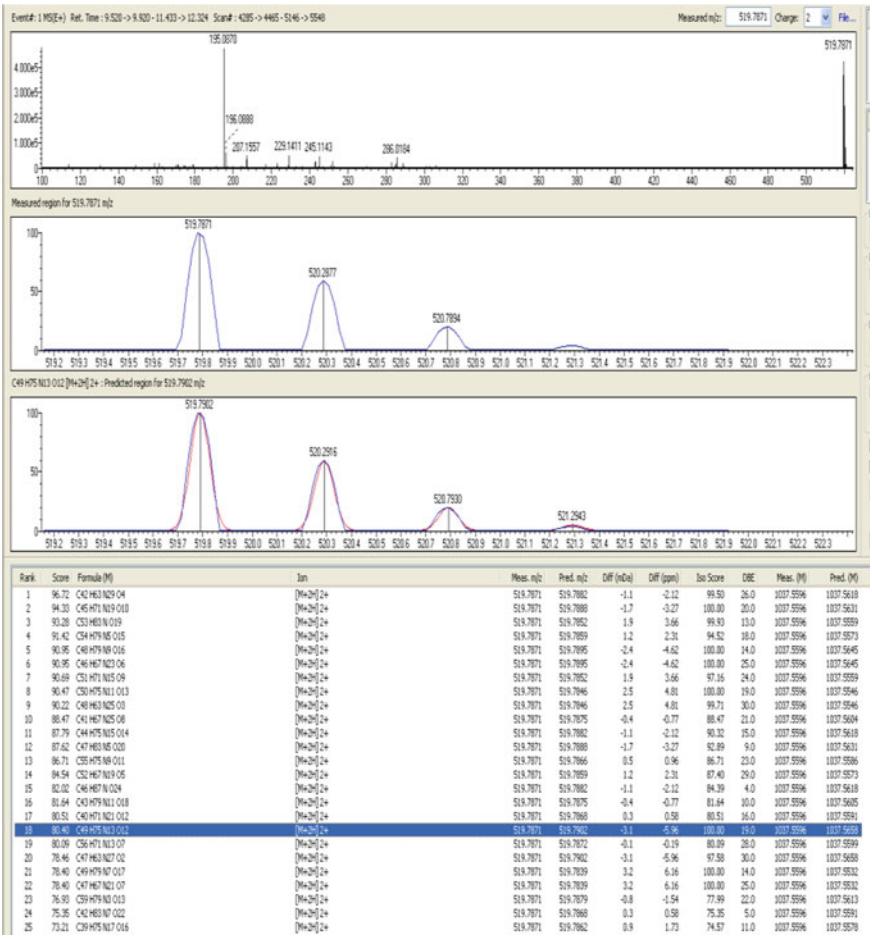
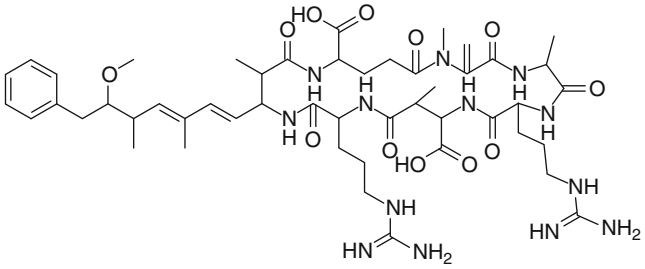


**Fig. 4.6** High-resolution mass spectrum of the blue-green algal hepatotoxin, microcystin RR $C_{49}H_{75}N_{13}O_{12}$ **4.2,** processed by the Formula Predictor software (Shimadzu, Japan). From top to bottom: the experimental spectrum in the wide *m*/*z* range and the range of $[M+2H]^{2+}$ ion, the overlap of the experimental and predicted peaks of that ion, the table with predicted molecular formulas and their characteristics. The formulas are ranked according to their Score index which is the hidden function of the difference in (**a**) accurate mass, Diff(mDa) or Diff(ppm), between experimental and predicted data, and (**b**) the corresponding isotope ratio, Iso Score. There are a number of candidate formulas matching this accurate molecular mass; the target molecule is of only the 18th rank. The selection from a profusion of such candidate formulas is discussed in Sect. 7.4.2.



#### 4.4.2.3 Peptides and Proteins

Protein and peptide molecules are built from residues of 20 common amino acids. Due to very high molecular mass, it is very difficult to record the mass spectra of most proteins. The standard bioanalytical practice is that proteins are cleaved by enzyme treatment into peptides according to the rules depending on the nature of the enzyme; the trypsin enzyme is widely applicable for this purpose [78–80]. Then the mass spectrum of a set of peptides is obtained, which is composed of peaks of (1) protonated (technique of MALDI), including multiply protonated (ESI–$MS^1$) peptide molecules, or (2) those and peptide fragments (ESI–$MS^n$). There are two further approaches to identification. Historically, the comparison to theoretical spectra was first used for identification, with two possible sub-approaches related to spectra of the 1st and 2nd kinds.

In the first case, *peptide mass fingerprinting*, the similarity in mass between peaks in the experimental spectrum of a peptide set and theoretically possible peptides which are formed from all the proteins contained in databases of amino acid sequences, is searched [79]. Similarity in mass is matching within some tolerance range set up before searches in databases. So, the MF named a *score* is the number of matching masses (see Fig. 4.7). That is the simplest kind of similarity measures. In advanced scoring, the probability of a chance match is estimated. For example, if 500 identical peptides from 1,000,000 database entries fall within the mass tolerance range about the experimental peptide mass, the probability of the chance match is 500/1,000,000 = 0.0005. The smaller this probability is, the more significant the match becomes. As usual, the general probability is the product of individual contributions (see Sect. 4.5.4.1) from individual peptides. Furthermore, a statistical weight for each individual peptide match is taken into account [81, 82]. Different peptides (with different size/mass) as counterparts occur in protein molecules with differing probabilities depending on the size of the peptide and the size of the protein. A small peptide originating from a large protein leads to a low score, a large peptide from a small protein results in a high score.

In the second case, *fragment mass fingerprinting,* the similarity in mass between ions in the experimental $MS^2$ spectrum containing peak set of peptide fragments

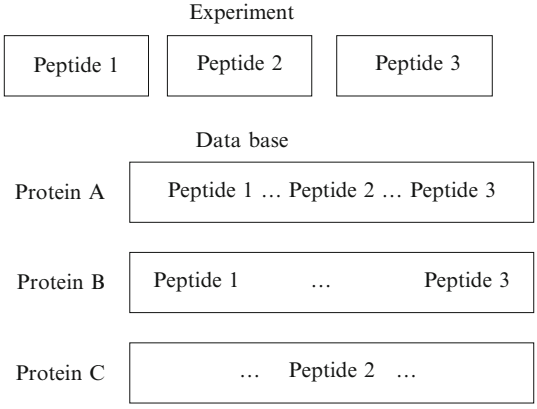



**Fig. 4.7** The example of peptide mass fingerprinting (adopted from [82]). There are three peptide mass matches, which can be assigned to three proteins. The most probable answer is protein A, composed of all the peptide residues and thus providing three matches. However, it is possible that the sample actually contains a mixture of proteins B and C. This is the case where the proteins are digested to peptides without prior separation

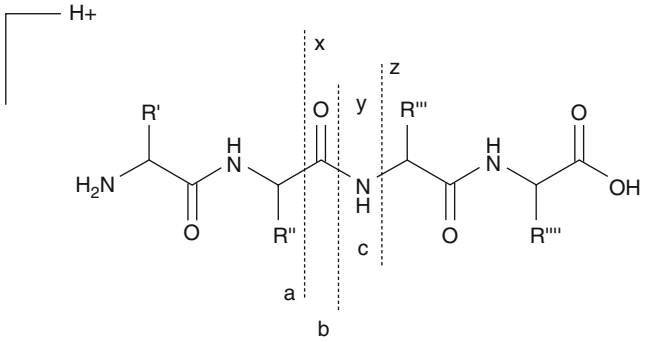



**Fig. 4.8** The bond cleavage leading to basic peptide fragments in tandem mass spectrometer subsequent to protonation of a peptide molecule and collisional activation of this precursor. If the positive charge is retained on the N terminal fragment, the ions are classified as *a*, *b* or *c*. If the charge is retained on the C terminal, the fragment ion class is *x*, *y* or *z*. There are series of such regular ions in longer peptides (three and more amino acid residues). Those may also be ionized by adding two or more protons. Formation of all the ions or a part of them and also some other fragments is incorporated in algorithms of computer MS identification of proteins. For example, the Mascot program (Matrix Science, UK) takes into account by default: (1) *a*, *b*, and *y* ion series, (2) their [*a*-$NH_3$], [*b*-$NH_3$], and [*y*-$NH_3$] fragments, and (3) doubly-charged fragments if the precursor is doubly or multiply charged [83]

and theoretically possible fragment ions that are formed from all the peptides of all the proteins from corresponding database, is searched [78–80]. Theoretical spectra are generated according to the rules of fragmentation (Fig. 4.8). Again, the number

of peak matches is the simplest score which can be further improved with evaluation of match probabilities [11, 82].

Recently, tandem spectral peptides libraries began to be developed. These databases make it possible to use the standard MS approach, which is comparison to reference spectra. As a rule, various forms of cosine function (dot product) have been included in algorithms for unknown peptides searches (e.g., [84–87]. Correlation-type MF [87, 88] has been pointed out as the most robust one [87], and some other similarity measures [85, 87] have also been used.

### *4.4.3 NMR Spectroscopy*

#### 4.4.3.1 Comparison with Predicted Spectra

Expert systems are used in NMR spectroscopy (Sect. 2.8.3) to predict spectra of compounds for their structure elucidation. Trueness of a prediction is determined by a similarity between experimental and predicted spectra. To measure this similarity, MF as the objective function

$$F = \Sigma(W_{Shift} \cdot F_{Shift} + W_{Quant} \cdot F_{Quant} + W_{Mult} \cdot F_{Mult}) \tag{4.20}$$

was invented for $^{1}$H NMR spectra [89]. Here, the terms of $F_{Shift}$, $F_{Quant}$, and $F_{Mult}$ are the degree of similarity between the one and the other spectrum in chemical shift, signal intensity and its multiplicity respectively. $W_{Shift}$, $W_{Quant}$, and $W_{Mult}$ are corresponding weighting factors. The summation is carried out over all of the predicted signals.

When $F$ terms are estimated, a significant dissimilarity of compared quantities, if observed, is expressed in a penalty included in a total score. If the latter is 0 (or a little higher) and 1 (or a little lower), there is *incorrect match* and *correct match* respectively [89]. This predictional approach has been validated for:

- $^{1}$H NMR spectra [89–91],
- Two-dimensional $^{1}$H–$^{13}$C correlation spectra [90, 91].

For related topics, see also Sect. 7.6.

#### 4.4.3.2 Comparison with Reference Spectra

The simple MF for searches in NMR spectral libraries (Sect. 7.6) consists of the average difference in chemical shifts between the query spectrum and the database one, corresponding squared difference, and Euclidean distance; such an MF was named the *hit quality index* (see [92]). Also, complicated functions of differences in chemical shifts between compared spectra have been reported [93–95].

In the general case, the similarity between experimental and reference $^{1}$H NMR spectra was noted to be heavily estimated, because signals are very narrow as

compared to their bias in different experimental conditions [66]. As one of the possible ways of solving the problem, a compression of NMR data by a binning technique has been proposed. The similarity index is further calculated, which depends on the number of H atoms in each bin [66]. The use of another MF, a weighted cross-correlation function, has also been discussed [66].

### 4.4.4 IR Spectroscopy

A number of similarity measures (in some cases, named *hit quality index* like NMR) and related algorithms of librarian searches (IR spectral libraries, see Sect. 7.5) have been tested and used in IR and also Raman spectrometry:

- Euclidean distance [96–98],
- First derivative absolute value (absorbance difference) [96, 98],
- Correlation coefficient [96, 99, 100],
- Absolute value (absorbance difference) [96, 97, 99],
- Squared absolute value (absorbance difference) [96, 97, 99],
- Dot product [99],
- First derivative correlation [96, 100],
- First derivative least squares [96]; see also [101].

In analytical practice, one or another MF is considered as a good or even the best one.

- For a flat baseline at very low IR absorbance, the Euclidean library search algorithm is recommended. If a baseline is bad, (a) it can be corrected before searches, or (b) the first derivative algorithm should be used [96, 98].
- In the case of low signal-to-noise ratios and negative bands/spikes (the case of GC–IR), the correlation MF are recommended [96].
- Correlation coefficients work better than some other MF for the purpose of structure elucidation achieved with the use of the reference IR spectral library [99].

### 4.4.5 UV-V is Spectroscopy

The correlation coefficient is a common form of MF for diode-array detection in HPLC [102]. The corresponding maximum value 1,000 is established. Different similarity indices, such as those mentioned above for IR spectroscopy, can also be used [103]. Another MF is spectral contrast angle [104].

### 4.4.6 Meaning of MF

MF is a measure of how well an unknown spectrum matches a spectrum from the database/library. The maximum value (established as 1,000, 100 or 1 depending on the

technique, software, library, and so on) indicates a perfect match. The minimum possible value is zero, which indicates that a spectrum of the analyte did not match a library spectrum at all. Commonly, a high MF, not lower than 80–90% of maximum values, will be sufficient to take into account the identification hypothesis. However, this value alone does not make it possible to definitely accept the corresponding hypothesis. An analyst should compare MF for the first, second, and possibly some subsequent spectra ranked top by match values with the unknown spectrum. If there is not a large difference between corresponding MF, e.g., it is smaller than the difference between the maximum value and MF of the first rank spectrum (see Fig. 3.6), search results cannot be considered as fully definitive. Correspondingly, the unknown compound could be any one of the compounds related to hit spectra. This is the case of several candidates for identification.

On the whole, a quantity of MF can be rather considered as a measure or one of the measures of identification reliability rather than some (intermediary) variable for ultimate estimating rates of true and false results. However, in some models, spectral similarity factors can be directly related to the probability of true/false identification results, i.e., the identification reliability (see below).

## 4.5 Probabilistic Interpretation of Analytical Data

There is no single approach to estimating an identification reliability, i.e., the probability of true results for any target in any matrix. Furthermore, there should be one or another

- Development of a probability model of phenomena underlying an analytical experiment,
- Probabilistic interpretation of raw or partly processed analytical data,
- Approximation in evaluation of the probability of true and false results.

In most cases, these are hard to implement. Nevertheless, some probability interpretations and models have been developed and are considered below.

### *4.5.1 True and False Rates*

Screening methods are characterized by their true/false result rates, which refer to both detection and identification of analytes (Sect. 4.2). Therefore, a probability of positive or negative identification results can be expressed by corresponding rates (Table 4.3) obtained in numerous analytical trials. Table 4.16 lists the rates and their interpretation just for identification.

**Table 4.16** Probabilistic interpretation of result rates

| Rate | Interpretation |
|---|---|
| Sensitivity (true positive rate), *St* | Probability of true identification of analyte |
| False positive rate, *FPR* | Probability of false identification of analyte |
| Specificity (true negative rate), *Sp* | Probability of true non-identification of analyte[a] |
| False negative rate, *FNR* | Probability of false non-identification of analyte |
| Positive predictive value, *PPV*[b] | Probability of true identification among all positive results |
| Negative predictive value, *NPV* | Probability of true non-identification among all negative results |
| Prevalence, *Pv* | Prior probability of the presence of analyte in the sample [c] |
| Cumulative positive predictive value, *CPPV* | Probability of true identification among all positive results of two combined identification procedures |

[a]The analyte is not present in the sample or not related to the analytical signal under consideration
[b]In identification of peptides and proteins by mass spectral match, the (100−PPV)% rate named false discovery rate (*FDR*), is calculated [11]
[c]For probabilities derived from previous information, see Chap. 6

Now, such characteristics as *St*, *FPR*, and *PPV* are not only estimated in chemical and biochemical tests and chromatography-based methods, but also calculated as performances of searches in spectral databases (Chap. 7). This emphasizes that ultimately all those are techniques/methods of holistic qualitative analysis as the part of analytical science and practice.

### *4.5.2 Type I and II Error*

These originate in false acceptance or rejection of identification hypotheses. These errors are also interpreted as probabilities. The type I error $\alpha$ is the probability of FN. The type II error $\beta$ is that of FP (Sect. 3.6).

### *4.5.3 Confidence Probability*

If an analyte value of some measurand falls in a confidence interval for the known compound A, an analyte may be identified as A (Sect. 3.3). Here the confidence probability, usually 0.95, can be equated to the identification probability. For confirmation of an identification result, an analyst should check whether

- Conditions of measurement in compared cases are the same or very similar,
- Other compounds with a similar property (a similar value of the proper quantity), are not present in the sample, i.e., the chance of FP is insignificant.

## 4.5.4 Spectral Matching and Probability of Identification

### 4.5.4.1 General

Several probability models are based on estimating a probability of spectral matching and a probability that a match is a random one (see below). In general, such probability can be expressed by the formula:

$$P_{match} = \prod_{i=1}^{n} p_{match,i} \tag{4.21}$$

where $P_{match}$ is the full probability of matching the unknown and reference spectrum, $n$ is the number of spectral peaks, and $p_{match,i}$ is the probability of matching the i$^{th}$ peaks. The latter means a match of numerical values within a tolerance range or an equality of rounded values. These are (a) values of spectral variables, i.e., frequency, wavelength, wave number, mass, and so on, and (b) peak intensities, i.e., their heights or areas.

For most types of spectra and most pairs of compounds, $P_{match} << 1$. Thus the match of many peaks, if observed, is due to the fact that compared spectra belong to the same compound. This is the reason to accept the identification hypothesis that the analyte is the compound with the matching spectrum. In this case, $P_{match}$ is the probability of FP.

Most research devoted to this probabilistic approach to identification is related to mass spectrometry (Sects. 4.5.4.2 and 4.5.4.3); for IR spectroscopy, see [105].

### 4.5.4.2 Mass Spectrometry of Low Molecules

*Probability-based matching* (Sect. 4.4.2.1), one of the main algorithms for mass spectral retrieval, results rather in a probability of matching than that in ultimate identification. The same applies to any other pertinent algorithm if the only MF value obtained in library searches is used for identification.

Probability interpretation of spectral match indices becomes more straightforward if all MF of spectra of different compounds presenting in a hit list are taken into account [106]. This is expressed by the general formula:

$$Prob_i = f(\Delta MF_{i,i+1}), \tag{4.22}$$

where $Prob_i$ is the relative probability of unknown identification as the compound belonging to the $i^{th}$ hit in the list, and $\Delta MF_{i,i+1}$ is the MF difference between $i^{\text{th}}$ and $(i+1)^{\text{th}}$ hits; only the hit with the highest MF is taken into account if two or more reference spectra for a compound are included in a library. To transfer to absolute probabilities, values $Prob_i$ should be normalized with regard to all the differences $\Delta MF$ from the hit list [106]. Estimating conventional identification probability of

this sort was entered into the NIST MS Search program for mass spectra librarian searches [107], and is exemplified below in Example 4.8 and Sect. 6.6.

**Example 4.8** Results of the library search, the MF and corresponding *Prob* values, for one (Fig. 4.9a) of the mass spectra of benzene **4.3** as an unknown one, are given in Table 4.17. The most similar are other benzene spectra (Fig. 4.9b and top lines in Table 4.17). Hits of non-benzene compounds are of lower MF and correspondingly *Prob* than benzene spectra. The conventional probability of identification *Prob* of two or three other compounds rather than benzene itself, starting from diacetylene hydrocarbon **4.4** (about 18%, the spectrum in Fig. 4.9c), is not very low. In the case of analysis of a real sample, corresponding identification hypotheses should not be rejected. However, an experienced mass spectrometrist will see the close resemblance of two benzene spectra (Fig. 4.9a, b) and the spectral contrast between that pair and the spectrum of **4.4** (Fig. 4.9c). This is a strong reason to reject the identification hypothesis for **4.4** as well as any hypothesis not connected with benzene **4.3.**

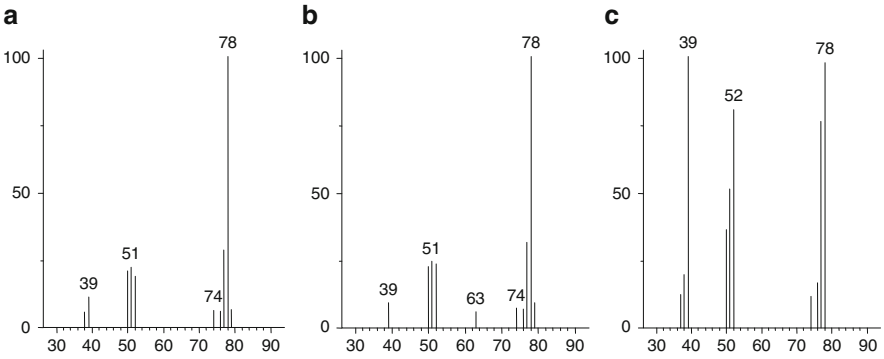
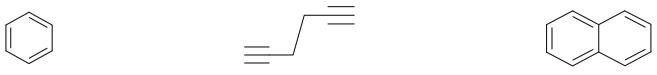



**Fig. 4.9** Mass spectra of benzene **4.3** as an unknown (**a**) and reference (**b**) compound and also 1,5-hexadiyne **4.4 (c)**. The latter spectrum is the most similar to those for benzene. Mass spectra are extracted from the NIST'05 library [107] (reproduced with permission) and reduced to ten main peaks



**4.3** **4.4** **4.5**

**Table 4.17** Spectral retrieval[a] for the mass spectrum of benzene **4.3**

| # | MF (Match) | *Prob* | Name |
|---|---|---|---|
| 1 | 999 (Fig. 4.9a)[b] | 77.2 | Benzene **4.3** |
| | 973 (Fig. 4.9b) | 77.2 | Benzene **4.3** |
| | 965 | 77.2 | Benzene **4.3** |
| | 950 | 77.2 | Benzene **4.3** |
| | 938 | 77.2 | Benzene **4.3** |
| 2 | 932 (Fig. 4.9c) | 17.6 | 1,5-Hexadiyne **4.4** |
| 3 | 897 | 4.42 | 2,4-Hexadiyne |
| | 896 | 4.42 | 2,4-Hexadiyne |
| | 853 | 17.6 | 1,5-Hexadiyne **4.4** |
| 4 | 826 | 0.62 | 2-Butenedinitrile, (*E*)- |
| | 822 | 0.62 | 2-Butenedinitrile, (*E*)- |

[a]The NIST software and MS library [107]
[b]This is self-matching

#### 4.5.4.3 Mass Spectrometry of High Molecules

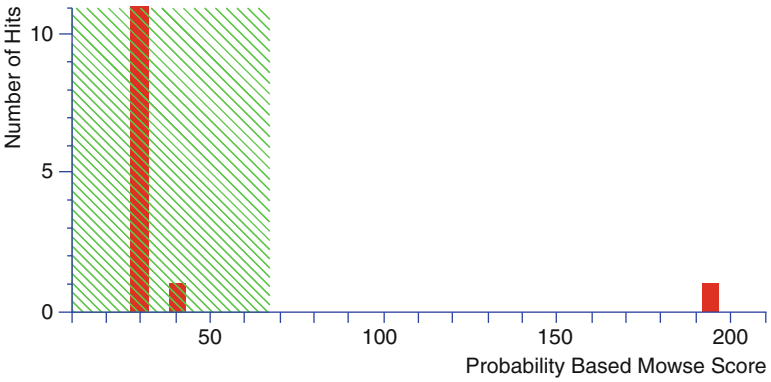
In the case of peptides and proteins, the MF is the number of matching ion masses of peptides or their fragments. The more significant factor is the probability-based score ($-10 \cdot \log P$), where $P$ is the probability of the chance match; see Sect. 4.4.2.3. This score is considered as the measure of the statistical significance of matches calculated in the following way [81].

Possible matches between the experimental mass set and theoretical ones lead to the set of score values. The distribution of these values can be obtained, where the highest score belongs to the most probable protein candidate for identification. It is also important that the best value would fall in the range of low probability of a chance match, e.g., $\leq 0.05$ (see Fig. 4.10). The fact that only one protein is within this range and all other matches are outside the range means that the match for the protein is significant, and strong evidence for its identification is obtained.

It should be also noted that probability-based scores of this sort are evaluated in identification procedures using both peptide mass fingerprinting and peptide fragment mass fingerprinting (Sect. 4.4.2.3). For the latter, peptide matches are grouped into protein ones for their scoring [81].

This or similar probability approaches to estimate the trueness of peptide identification, are widespread. They could be supplemented by other ones, for example the statistical model using discriminant and Bayesian analysis [109].

In any case, the identification method discussed in this subsection, as well as a spectral library search of spectra of low-molecular compounds (Sect. 4.5.4.2 and Chap. 7), is essentially a screening method. Therefore, identification results judged correct according to a non-random match may be partly false. The corresponding percentage, i.e., the *FDR* rate (see Table 4.3), can be statistically estimated [11].




**1** MEPAPAR**SPR PQQDPARPQE PTMPPPETPS EGRQPSPSPS PTERAPASEE**
**51** **EFQFLR**CQQC QAEAKCPKLL PCLHTLCSGC LEASGMQCPI CQAPWPLGAD
**101** TPALDNVFFE SLQRRLSVYR QIVDAQAVCT RCKESADFWC FECEQLLCAK
**151** CFEAHQWFLK **HEARPLAELR** NQSVREFLDG TRKTNNIFCS NPNHRTPTLT
**201** SIYCRGCSKP LCCSCALLDS SHSELKCDIS AEIQQRQEEL DAMTQALQEQ
**251** DSAFGAVHAQ MHAAVGQLGR ARAETEELIR ERVRQVVAHV RAQERELLEA
**301** VDARYQR**DYE EMASR**LGR**LD AVLQR**IRTGS ALVQRMKCYA SDQEVLDMHG
**351** FLRQALCR**LR QEEPQSLQAA VRTDGFDEFK** VRLQDLSS0043I TQGKDAAVSK
**401** KASPEAASTP RDPIDVDLPE EAERVKAQVQ ALGLAEAQPM AVVQSVPGAH
**451** PVPVYAFSIK GPSYGEDVSN TTTAQKRKCS QTQCPRKVIK **MESEEGKEAR**
**501** LAR**SSPEQPR PSTSKAVSPP HLDGPPSPRS PVIGSEVFLP NSNHVASGAG**
**551** **EAEER**VVVIS SSEDSDAENS SSR**ELDDSSS ESSDLQLEGP STLRVLDENL**
**601** **ADPQAEDRPL VFFDLK**IDNE TQKISQLAAV NRESKFRVVI QPEAFFSIYS
**651** KAVSLEVGLQ HFLSFLSSMR RPILACYKLW GPGLPNFFRA LEDINRLWEF
**701** QEAISGFLAA LPLIRERVPG ASSFKLKNLA QTYLARNMSE RSAMAAVLAM
**751** RDLCRLLEVS PGPQLAQHVY PFSSLQCFAS LQPLVQAAVL PRAEARLLAL
**801** HNVSFMELLS AHRRDRQGGL KKYSRYLSLQ TTTLPPAQPA FNLQALGTYF
**851** EGLLEGPALA RAEGVSTPLA GRGLAERASQ QS

**Fig. 4.10** Histogram of the score distribution in the procedure of identification of the PML_HUMAN protein, molecular mass 97489**,** by peptide mass fingerprinting (the Mascot software, Matrix Science, UK [108], reproduced with permission) (*top*) and the amino acid sequence of this protein (*bottom*). There are 15 peptide mass values matched (**bold** in the sequence). They provide the high probability-based score 194, which falls within the range of statistically significant data, i.e., that of the 0.05 probability (*unshaded*). The latter is analogous to the $\alpha$ range for the cases of testing detection (Fig. 3.4) and identification (Fig 3.7) hypotheses. Here it is the range for acceptance of the identification hypothesis: the analyte is the protein with the score which is not random, i.e., within the 0.05 range. This probability is that of FP. Other candidates for identification have scores of 56, 51, 50, 49, 42, 41 and smaller. These rates fall in the *shaded area*, i.e., that of random match. In general, varying of FP probability affects rates of both TP and FP. Dependencies between them are used to estimate a performance of peptide identification algorithms, see Sect 7.4.1.4.

The Mowse (MOlecular Weight SEarch) is the name of the similar algorithm and scoring system which preceded Mascot. There are a number of other programs under consideration intended for the purpose (see [11]).

One-letter codes of amino acids: A is alanine, R is arginine, N is asparagine, D is aspartic acid, C is cysteine, E is glutamic acid, Q is glutamine, G is glycine, H is histidine, I is isoleucine, L is leucine, K is lysine, M is methionine, F is phenylalanine, P is proline, S is serine, T is threonine, W is tryptophan, Y is tyrosine, V is valine.

### 4.5.5 *Spectral Interpretation*

In past decades, computer-assisted systems for structure elucidation were developed [110–112]. Based on spectrum–structure correlations, corresponding rules, logic, and knowledge, and chemometrical methods, a computer expert generates/predicts plausible structures and substructures for experimental spectra recorded by an analyst. Also, a conventional probability is assigned to candidate structures and substructures; see Example 4.9.

**Example 4.9** One of the options of the NIST MS Search program is the generation of candidate substructures from unknown EI mass spectra [113]. Table 4.18 lists substructures or structural features of benzene and its isomer, 1,5-hexadiyne, deduced by the automatic interpretation of library mass spectra of the compounds by the program. There are also corresponding probabilities for the appearance of substructures/features and the same outcomes for another testing compound, naphthalene, with a somewhat different structure. All three compounds **4.3**, **4.4**, and **4.5** are considered as unknown ones.

Table 4.18 classifies the two first compounds as unsaturated hydrocarbons, with the probability of their features being over about 80% (*top lines* in Table 4.18). The compound **4.3** is also concluded to be without alkyl groups (*two bottom lines*). Therefore it is the ring compound, "probably" 6-membered (47%) aromatic (57%) ring, i.e., benzene characteristics are adequately predicted. For the compound **4.4**, the last two features are not so evident (22 and 34%). In contrast, the conclusion that the compound **4.4** is acyclic (*no rings* 61%, *no aromatic rings* 54%) acetylene (*non-ring CC triple bond* 55% and so on, Table 4.18) hydrocarbon can be made. Corresponding features of **4.3** are less apparent or absent.

Further, it is also logical that the compound **4.5** is unsaturated (81%) aromatic (83%) hydrocarbon, non-benzene (83%), and without many features of an acyclic structure (*bottom probabilities*, Table 4.18).

The above example of the correlation between structures of relatively simple compounds and mass spectra disclosed by the computer expert shows the good possibility for discrimination between different classes of compounds, if not individual identification. In any case, expert systems work in such a way that several plausible structures as identification hypotheses are generated, which can be further tested using different techniques/methods (Chap. 7).

**Table 4.18** Probability $p$ of presence (+) and absence (−) of substructures or structural features

| Substructure, structural feature | $p \cdot 100\%$ | | | | | |
|---|---|---|---|---|---|---|
| | Benzene **4.3** | | 1.5-Hexadiyne **4.4** | | Naphthalene **4.5** | |
| | + | − | + | − | + | − |
| Common features of benzene and its linear isomers | | | | | | |
| rings + double bond counts = 4 | 92 | | 94 | | | 97 |
| no branches | 92 | | 92 | | | 96 |
| unsaturated hydrocarbon | 86 | | 93 | | 81 | |
| hydrocarbon (C and H atoms only) | 79 | | 90 | | 72 | |
| Benzene features | | | | | | |
| aromatic ring | 57 | | 34 | | 83 | |
| isolated benzenoid (6-membered) ring | 47 | | 22 | | | 83 |
| Features of acyclic hydrocarbon | | | | | | |
| no aromatic rings | 34 | | 54 | | | 88 |
| no rings | 26 | | 61 | | | 98 |
| non-ring CC triple bond | 29 | | 55 | | ? | |
| 2 CC double/triple bonds | | 29 | 52 | | | 74 |
| ethynyl group | 15 | | 43 | | ? | |
| methylene or methyl group (chain) | | 98 | 27 | | | 99 |
| exactly one ethyl or dimethylene group (chain) | | 84 | 26 | | | 95 |

The software and MS library [107]

**Table 4.19** Scales of word expressions proposed for identification reliability[a]

| *Identification confidence* [3] | Reliability | *Strength of evidence* [115] |
|---|---|---|
| Identified with utmost certainty[b] | ↑ | Very strong evidence |
| Confirmed | | Strong evidence |
| Identified with confidence | | Good evidence |
| Identified | | |
| Indicated | | Fair evidence |
| Tentative identification | | |
| Suspected | | |
| Presumptive | | |
| Non-negative | | Non-match |

[a]The levels are compared by the author
[b]Corresponds to the popular term of *unambiguous identification*

## 4.6 Non-numerical Estimates of Reliability

It would be decisive to express a numerical probability of any identification result. However, this is very hard to perform because of a plethora of factors affecting identification results. In fact, even a perfect spectral match may lead to an erroneous conclusion (FP) if, for example, (a) an "identified" compound is the product of the analyte transformation formed during an analysis rather than the analyte itself [57], or (b) a spectrum of an "identified" compound is similar to the analyte spectrum which is actually absent from the spectral library. In

**Table 4.20** Word expressions for reporting the result of identification or detection

| Terms | Target | Explanation |
|---|---|---|
| Analyte is (a) identified, (b) tentatively confirmed, (c) not identified, (d) not detected, | Pesticides | (a) Selected ion ratios are within established tolerance ranges at given RT, 2–3 expected ion peaks (3) are present in mass spectra at given RT or (c) are not present, (d) RT are outside the established tolerance range, respectively [116] |
| (e) Identification, (f) indication, (g) absence | Target compounds in soil | At least (e) three, (f) one or two, and (g) no identification points are obtained with the use of mass spectrometry, respectively [58] |
| Positive, negative, not detected, none detected | Toxicants | The first result is that the substance is identified according to the laboratory protocols. The other results mean that the analyte or analytes are absent; *none detected* is preferred [117] |
| Unidentified A, unidentified with relative retention of [value] | Pharmaceutical impurities | Examples of descriptive labels for unidentified impurities included in the specifications of new drug substances [62] |
| (h) Positive match, (i) probable match, (j) non-match, and (k) inconclusive | Oils | Standardized degrees of quantitative similarity/difference in ratios of biomarker amounts between samples of two oils. The match is in the (h) narrow, (i) wide, and (j) very wide ratio range respectively; (k) samples available are difficult for definite conclusions [118] (see Sect. 8.2.2) |
| Match, probable match, non-match, and indeterminate | Oils | Standardized degrees of qualitative similarities/differences in composition between samples of two oils. Conclusions are made "in the light of experience and the existing body of knowledge about oil analysis" [119] |
| (l) Yes, (m) no, (n) inconclusive | Hg | Content is respectively (l) above or equal to the high cut-off point of screening method, (m) equal to or below the corresponding low cut-off point, and (n) in the unreliability range [55] (see Sect 4.3.2.1) |
| (o) Presence, (p) negative, (q) non-negative | Pesticides | Content is respectively (o) above the high cut-off point of the screening method (here, the upper screening limit), (p) below the low cut-off point (the lower screening limit), and (q) in the unreliability range (between the screening limits or equal to one of them) [56] (see Sect 4.3.2.1) |

contrast, a spectral match for one and the same substance may be rather poor and lead to FN due to:

- Insufficient selectivity of analysis and interference with matrix compounds.
- Low reproducibility/quality of experimental and reference spectral data and many other reasons (see [4, 57, 114]).

These and other factors are very hard, if not impossible, to uniformly quantify and further transform into identification probabilities.

Nevertheless, an experienced analyst takes into account these factors and expresses, directly or latently, their role as general/fuzzy estimates, using certain word combinations (*prose terms* [3], *verbal conventions* [115]). A series of such word expressions have been proposed (Table 4.19). They can be considered as points on another nominal scale (see Sect. 1.7) specially generated to express the identification reliability.

One of two scales given in Table 4.19 and proposed for human identification [115] can be quantitatively "calibrated" by means of the probability of matching profiles derived from probes. In other cases, professional judgments of trained and experienced analysts based on the knowledge are essential for expressing results and errors of identification in those or other standardized terms/concepts (see also [3, 57].

Word terms are or may be used for reporting results of qualitative analytical operations in standard (standardized, validated) methods (Table 4.20). Here, words expressing the result and reliability of identification are connected to method criteria which should be met. The abundant criteria are related with, for example, the number of identification points and corresponding tolerances for spectral quantities (see Chap. 5) or unreliability concentration range (detection, screening; see Sect. 4.3.1).

## References

1. Milman BL, Konopelko LA (2000) Identification of chemical substances by testing and screening of hypotheses. I. General. Fresenius J Anal Chem 367:621–628
2. Ellison SLR, Fearn T (2005) Characterising the performance of qualitative analytical methods: Statistics and terminology. Trends Anal Chem 24:468–476
3. Bethem R, Boison J, Gale J, Heller D, Lehotay S, Loo J, Musser S, Price P, Stein S (2003) Establishing the fitness for purpose of mass spectrometric methods. J Am Soc Mass Spectrom 14:528–541
4. Milman BL (2008) Introduction to chemical identification (In Russian). VVM, Saint Petersburg
5. Google Scholar. http://scholar.google.com. Accessed 14 Oct 2009.
6. Spiehler VR, O'Donnell CM, Gokhale DV (1988) Confirmation and certainty in toxicology screening. Clin Chem 34:1535–1539
7. Ferrara SD, Tedeschi L, Frison G, Brusini G, Castagna F, Bernardelli B, Soregaroli D (1994) Drugs-of-abuse testing in urine: statistical approach and experimental comparison of immunochemical and chromatographic techniques. J Anal Toxicol 18:278–291

8. Ellison SLR, Gregory S, Hardcastle WA (1998) Quantifying uncertainty in qualitative analysis. Analyst 123:1155–1161
9. McLafferty FW, Stauffer DA, Loh SY, Wesdemiotis C (1999) Unknown identification using reference mass spectra. Quality evaluation of databases. J Am Soc Mass Spectrom 10:1229–1240
10. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27:861–874
11. Nesvizhskii AI, Vitek O, Aebersold R (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Methods 4:787–797
12. Lloyd E (1984) Handbook of applicable mathematics. Wiley, Chichester
13. ASTM E 1828 (1996) Standard guide for evaluating the performance characteristics of qualitative chemical spot test kits for lead in paint
14. Mil'man BL, Konopel'ko LA (2004) Uncertainty of qualitative chemical analysis: general methodology and binary test methods. J Anal Chem 59:1128–1141
15. Sensitivity in metrology and analytical chemistry. In: IUPAC Gold Book. http://goldbook.iupac.org/S05606.html. Accessed 11 May 2010
16. Detection limit in analysis. In: IUPAC Gold Book. http://goldbook.iupac.org/D01629.html. Accessed 11 May 2010
17. Specific in analysis. In: IUPAC Gold Book. http://goldbook.iupac.org/S05788.html. Accessed 11 May 2010
18. Song R, Schlecht PC, Ashley K (2001) Field screening test methods: performance criteria and performance characteristics. J Hazard Mater 83:29–39
19. Dietzen DJ, Ecos K, Friedman D, Beason S (2001) Positive predictive values of abused drug immunoassays on the Beckman Synchron in a veteran population. J Anal Toxicol 25:174–178
20. Jehanli A, Brannan S, Moore L, Spiehler VR (2001) Blind trials of an onsite saliva drug test for marijuana and opiates. J Forensic Sci 46:1214–1220
21. Kadehjian LJ (2001) Performance of five non-instrumented urine drug-testing devices with challenging near-cutoff specimens. J Anal Toxicol 25:670–679
22. Crouch DJ, Hersch RK, Cook RF, Frank JF, Walsh JM (2002) A field evaluation of five on-site drug-testing devices. J Anal Toxicol 26:493–499
23. Miki A, Katagi M, Tsuchihashi H (2002) Application of EMIT(R) d.a.u. (TM) for the semiquantitative screening of methamphetamine incorporated in hair. J Anal Toxicol 26:274–279
24. Cone EJ, Sampson-Cone AH, Darwin WD, Huestis MA, Oyler JM (2003) Urine testing for cocaine abuse: metabolic and excretion patterns following different routes of administration and methods for detection of false-negative results. J Anal Toxicol 27:386–401
25. Ashley K, Fischbach TJ, Song R (1996) Evaluation of a chemical spot-test kit for the detection of airborne particulate lead in the workplace. Am Ind Hyg Assoc J 57:161–165
26. Ashley K, Hunter M, Tait LH, Dozier J, Seaman JL, Berry PF (1998) Field investigation of on-site techniques for the measurement of lead in paint films. Field Anal Chem Technol 2:39–50
27. García-González DL, Viera M, Tena N, Aparicio R (2007) Evaluation of the methods based on triglycerides and sterols for the detection of hazelnut oil in olive oil. Grasas Aceites 58:344–350
28. Schepers PGAM, Franke JP, De Zeeuw RA (1983) System evaluation and substance identification in systematic toxicological analysis by the mean list length approach. J Anal Toxicol 7:272–278
29. Porter SEG, Stoll DR, Paek C, Rutan SC, Carr PW (2006) Fast gradient elution reversed-phase liquid chromatography with diode-array detection as a high-throughput screening method for drugs of abuse. II. Data analysis. J Chromatogr A 1137:163–172
30. Cech NB, Enke CG (2001) Practical implications of some recent studies in electrospray ionization fundamentals. Mass Spectrom Rev 20:362–387

31. Milman BL, Alfassi ZB (2005) Detection and identification of cations and anions of ionic liquids by means of electrospray ionization mass spectrometry and tandem mass spectrometry. Eur J Mass Spectrom 11:35–42
32. Valcárcel M, Cárdenas S, Barceló D, Buydens L, Heydorn K, Karlberg B, Klemm K, Lendl B, Milman B, Neidhart B, Ríos A, Stephany R, Townshend A, Zschunke A (2002) Metrology of qualitative chemical analysis. Report EUR 20605. EC, Luxembourg
33. Ríos A, Barceló D, Buydens L, Cárdenas S, Heydorn K, Karlberg B, Klemm K, Lendl B, Milman B, Neidhart B, Stephany R, Townshend A, Valcárcel M, Zschunke A (2003) Quality assurance of qualitative analysis in the framework of 'MEQUALAN' European project. Accred Qual Assur 8:68–77
34. Pulido A, Ruisánchez I, Boqué R, Rius FX (2003) Uncertainty of results in routine qualitative analysis. Trends Anal Chem 22:647–654
35. Plata MR, Pérez-Cejuela N, Rodríguez J, Ríos Á (2005) Development and validation strategies for qualitative spot tests: application to nitrite control in waters. Anal Chim Acta 537:223–230
36. Garrido Frenich A, González-Rodríguez MJ, Arrebola FJ, Martínez Vidal JL (2005) Potentiality of gas chromatography–triple quadrupole mass spectrometry in vanguard and rearguard methods of pesticide residues in vegetables. Anal Chem 77:4640–4648
37. Komar' NP (1955) Basics of qualitative chemical analysis. Book 1: Ionic equilibria (In Russian). Kharkov University Publisher, Kharkov
38. Trullols E, Ruisánchez I, Ruis FX (2004) Validation of qualitative analytical methods. Trends Anal Chem 23:137–145
39. Panteleimonov AV, Nikitina NA, Reshetnyak EA, Loginova LP, Bugaevskii AA, Kholin YV (2008) Binary response procedures of qualitative analysis: metrological characteristics and calculation aspects (In Russian). Methods Objects Chem Anal 3:128–146, http://www.nbuv.gov.ua/portal/Chem_Biol/moca/2006_2008/pdf/03022008-128.pdf. Accessed 27 Oct 2010
40. Kholin YV, Nikitina NA, Panteleimonov AV, Reshetnyak EA, Bugaevskii AA, Loginova LP (2008) Metrological characteristics of binary response qualitative methods (In Russian). Timchenko, Kharkov
41. ISO Standard 11843-1 (1997) Capability of detection. Part 1: Terms and definitions.
42. Antignac JP, Le Bizec B, Monteau F, Andre F (2003) Validation of analytical methods based on mass spectrometric detection according to the "2002/657/EC" European decision: guideline and application. Anal Chim Acta 483:325–334
43. Currie LA (1995) Nomenclature in evaluation of analytical methods, including detection and quantification capabilities (IUPAC Recommendations 1995). Pure Appl Chem 67:1699–1723
44. Commission Decision 2002/657/EC, August 12, 2002, implementing Council Directive 96/23/EC concerning the performance of analytical methods and interpretation of results (2002) Off J Eur Commun L 221:8-36. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2002:221:0008:0036:EN:PDF. Accessed 14 May 2010.
45. ISO Standard 11843-2 (2000) Capability of detection. Part 2: Methodology in the linear calibration case
46. ISO Standard 11843-4 (2003) Capability of detection. Part 4: Methodology for comparing the minimum detectable value with a given value
47. DIN 32645:2008-11 (2008) Chemische Analytik – Nachweis-, Erfassungs- und Bestimmungsgrenze unter Wiederholbedingungen – Begriffe, Verfahren, Auswertung. http://www.beuth.de/langanzeige/DIN+32645/110729574.html. Accessed 14 May 2010
48. Brüggemann L, Morgenstern P, Wennrich R (2010) Comparison of international standards concerning the capability of detection for analytical methods. Accred Qual Assur 15:99–104
49. Vinci F, Guadagnuolo G, Danese V, Salini M, Serpe L, Gallo P (2005) In-house validation of a liquid chromatography/electrospray tandem mass spectrometry method for confirmation of chloramphenicol residues in muscle according to Decision 2002/657/EC. Rapid Commun Mass Spectrom 19:3349–3355

50. Verdon E, Hurtaud-Pessel D, Sanders P (2006) Evaluation of the limit of performance of an analytical method based on a statistical calculation of its critical concentrations according to ISO standard 11843: Application to routine control of banned veterinary drug residues in food according to European Decision 657/2002/EC. Accred Qual Assur 11:58–62
51. Rodríguez N, Cruz Ortiz M, Herrero A, Sarabia LA (2007) Performance characteristics according to Commission Decision 2002/657/EC in the fluorimetric determination of tetracycline in the absence and in the presence of magnesium. Luminescence 22:518–526
52. Samanidou VF, Nisyriou SA, Papadoyannis IN (2007) Development and validation of an HPLC method for the determination of penicillin antibiotics residues in bovine muscle according to the European Union Decision 2002/657/EC. J Sep Sci 30:3193–3201
53. Samanidou VF, Nikolaidou KI, Papadoyannis IN (2007) Development and validation of an HPLC confirmatory method for the determination of seven tetracycline antibiotics residues in milk according to the European Union Decision 2002/657/E. J Sep Sci 30:2430–2439
54. Christodoulou EA, Samanidou VF (2007) Multiresidue HPLC analysis of ten quinolones in milk after solid phase extraction: Validation according to the European Union Decision 2002/657/EC. J Sep Sci 30:2421–2429
55. Resano M, Garcia-Ruiz E, Aramendia M, Belarra MA (2005) Solid sampling-graphite furnace atomic absorption spectrometry for Hg monitoring in soils. Performance as a quantitative and as a screening method. J Anal At Spectrom 20:1374–1380
56. Aybar-Muñoz J, Fernández-González E, García-Ayuso LE, González-Casado A, Cuadros-Rodríguez L (2005) Semiqualitative method for detection of pesticide residues over established limits in vegetables by use of GC-μECD and GC-(EI)MS. Chromatography 61:505–513
57. Lehotay SJ, Mastovska K, Amirav A, Fialkov AB, Martos PA, de Kok A, Fernández-Alba AR (2008) Identification and confirmation of chemical residues in food by chromatography-mass spectrometry and other techniques. Trends Anal Chem 27:1070–1090
58. ISO Standard 22892 (2006) Soil quality – Guidelines for the identification of target compounds by gas chromatography and mass spectrometry.
59. Stolker AAM, Linders SHMA, Van Ginkel LA, Brinkman UAT (2004) Application of the revised EU criteria for the confirmation of anabolic steroids in meat using GC–MS. Anal Bioanal Chem 378:1313–1321
60. Method validation and quality control procedures for pesticide residues analysis in food and feed (2009) Document No. SANCO/10684/2009. http://ec.europa.eu/food/plant/protection/resources/qualcontrol_en.pdf. Accessed 14 May 2010
61. Commission Decision 2003/181/EC, March 13, 2003, amending Decision 2002/657/EC as regards the setting of minimum required performance limits (MRPLs) for certain residues in food of animal origin Off J Eur Commun L 71:17–18. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:071:0017:0018:EN:PDF. Accessed 14 May 2010
62. Impurities Testing Guideline: Impurities in New Drug Substances (2006) ICH Topic Q 3 B (R2). European Medicines Agency. http://www.ema.europa.eu/pdfs/human/ich/273899en.pdf. Accessed 14 May 2010
63. Grace WR Material Safety Data Sheet. Product Name: ADVA 190. http://www.na.grace-construction.com/concrete/download/ADVA%20190%20D-06621%20_Q_.pdf. Accessed 14 May 2010
64. Savitski MM, Nielsen ML, Zubarev RA (2007) Side-chain losses in electron capture dissociation to improve peptide identification. Anal Chem 79:2296–2302
65. Olsen BA, Borer MW, Perry FM, Forbes RA (2002) Screening for counterfeit drugs using near-infrared spectroscopy. Pharm Technol June: 62, 64, 66, 68, 70–71, 95. http://pharmtech.findpharma.com/pharmtech/data/articlestandard/pharmtech/222002/20241/article.pdf. Accessed 14 May 2010
66. Bodis R (2007) Quantification of spectral similarity: towards automatic spectra verification. Dissertation ETH 17361, Zürich. http://e-collection.ethbib.ethz.ch/eserv/eth:29907/eth-29907-02.pdf. Accessed 15 May 2010

67. McLafferty FW, Stauffer DB (1985) Retrieval and interpretative computer programs for mass spectrometry. J Chem Inf Comput Sci 25:245–252
68. McLafferty FW, Tureĉek F (1993) Interpretation of mass spectra. University Science Book, Sausalito, CA
69. MassLab Version 1.2 User Guide (1993). VG Organic SD 001126
70. Stein SE, Scott DR (1994) Optimization and testing of mass spectral library search algorithms for compound identification. J Am Soc Mass Spectrom 5:859–866
71. Milman BL, Kovrizhnych MA, Konopelko LA (1999) Identification of chemical substances in analytical measurements. Accred Qual Assur 4:185–190
72. Wan KX, Vidavsky I, Gross ML (2002) Comparing similar spectra: from similarity index to spectral contrast angle. J Am Soc Mass Spectrom 13:85–88
73. Stein SE (1999) An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. J Am Soc Mass Spectrom 10:770–781
74. Takegawa Y, Deguchi K, Ito S, Yoshioka S, Sano A, Yoshinari K, Kobayashi K, Nakagawa H, Monde K, Nishimura S (2004) Assignment and quantification of 2-aminopyridine derivatized oligosaccharide isomers coeluted on reversed-phase HPLC/MS by MSn spectral library. Anal Chem 76:7294–7303
75. Kameyama A, Kikuchi N, Nakaya S, Ito H, Sato T, Shikanai T, Takahashi Y, Takahashi K, Narimatsu H (2005) A strategy for identification of oligosaccharide structures using observational multistage mass spectral library. Anal Chem 77:4719–4725
76. Zhang H, Singh S, Reinhold VN (2005) Congruent strategies for carbohydrate sequencing. 2. FragLib: an MSn spectral library. Anal Chem 77:6263–6270
77. Oberacher H, Pavlic M, Libiseller K, Schubert B, Sulyok M, Schuhmacher R, Csaszar E, Köfeler HC (2009) On the inter-instrument and the inter-laboratory transferability of a tandem mass spectral reference library: 2 Optimization and characterization of the search algorithm. J Mass Spectrom 44:494–502
78. Kinter M, Sherman NE (2000) Protein sequencing and identification using tandem mass spectrometry. Wiley, New York
79. Aebersold R, Goodlett DR (2001) Mass spectrometry in proteomics. Chem Rev 101:269–295
80. Sechi S (2007) Quantitative proteomics by mass spectrometry. Humana Press, Totowa, NJ
81. Mascot Scoring Schemes. http://www.matrixscience.com/help/scoring_help.html. Accessed 15 May 2010
82. Cottrell J. Database searching for protein identification and characterization. http://www.matrixscience.com/pdf/asms_tutorial_2005.pdf. Accessed 15 May 2010
83. Mascot Search Fields. Instrument. http://www.matrixscience.com/help/search_field_help.html#INSTRUMENT. Accessed 15 May 2010
84. Tabb DL, MacCoss MJ, Wu CC, Anderson SD, Yates JR (2003) Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. Anal Chem 75:2470–2477
85. Craig R, Cortens JC, Fenyo D, Beavis RC (2006) Using annotated peptide mass spectrum libraries for protein identification. J Proteome Res 5:1843–1849
86. Lam H, Deutsch EW, Eddes JS, Eng JK, King N, Stein SE, Aebersold R (2007) Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics 7:655–667
87. Liu J, Bell AW, Bergeron JJM, Yanofsky CM, Carrillo B, Beaudrie CEH, Kearney RE (2007) Methods for peptide identification by spectral comparison. Proteome Sci 5:3. doi:10.1186/1477-5956-5-3
88. Yates JR, Morgan SF, Gatlin CL, Griffin PR, Eng JK (1998) Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis. Anal Chem 70:3557–3565

89. Golotvin SS, Vodopianov E, Lefebvre BA, Williams AJ, Spitzer TD (2006) Automated structure verification based on 1H NMR Prediction. Magn Reson Chem 44:524–538
90. Golotvin SS, Vodopianov E, Pol R, Lefebvre BA, Williams AJ, Rutkowske RD, Spitzer TD (2007) Automated structure verification based on a combination of 1D 1H NMR and 2D 1H–13C HSQC spectra. Magn Reson Chem 45:803–813
91. Keyes P, Hernandez G, Cianchetta G, Robinson J, Lefebvre B (2009) Automated compound verification using 2D-NMR HSQC data in an open-access environment. Magn Reson Chem 47:38–52
92. ACD/1D NMR Manager. http://www.acdlabs.com/products/adh/nmr/1d_man/databasing.php. Accessed 15 May 2010
93. Bally RW, Van Krimpen D, Cleij P, Van'T Klooster HA (1984) An automated library search system for $^{13}$C-n.m.r. spectra based on the reproducibility of chemical shifts. Anal Chim Acta 157:227–243
94. Farkas M, Bendl J, Welti DH, Pretsch E, Dütsch S, Portmann P, Zürcher M, Clerc JT (1988) Similarity search for a $^{1}$H-NMR spectroscopic data base. Anal Chim Acta 206:173–187
95. Smith SK, Cobleigh J, Svetnik V (2001) Evaluation of a 1H-13C NMR spectral library. J Chem Inf Comput Sci 41:1463–1469
96. Algorithms. https://ftirsearch.com/help/algo.htm. Accessed 15 May 2010
97. ACD/UV-IR Manager. http://www.acdlabs.com/products/adh/uvir/uvir_man. Accessed 27 Oct 2010
98. Search Strategies for IR Spectra – Normalization and Euclidean Distance vs. First Derivative Algorithm (2008) Bio-Rad application note 94034-REV200801. http://www.knowitall.com/literature/application_notes/an-search-strat-1.pdf . Accessed 15 May 2010
99. Varmuza K, Karlovits M, Demuth W (2003) Spectral similarity versus structural similarity: infrared spectroscopy. Anal Chim Acta 490:313–324
100. McCreery RL, Horn AJ, Spencer J, Jefferson E (1998) Noninvasive identification of materials inside USP vials with Raman spectroscopy and a Raman spectral library. J Pharm Sci 87:1–8
101. Oberreuter H, Seiler H, Scherer S (2002) Identification of coryneform bacteria and related taxa by Fourier-transform infrared (FT-IR) spectroscopy. Int J Syst Evol Microbiol 52:91–100
102. Huber L (1989) Application of diode-array detection in high performance liquid chromatography. Hewlett-Packard Co. Publication number 12-5953-2330.
103. Hristozov D, Penchev P, Andreev G. Searching in UV/VIS spectral library. http://web.uni-plovdiv.bg/plamenpenchev/art14.pdf. Accessed 15 May 2010
104. Waters ACQUITY UPLC Photodiode Array Detector 71500108703 / Revision A. http://www.waters.com/webassets/cms/support/docs/71500108703ra.pdf. Accessed 15 May 2010
105. Ellison SLR, Gregory SL (1998) Predicting chance infrared spectroscopic matching frequencies. Anal Chim Acta 370:181–190
106. Stein SE (1994) Estimating probabilities of correct identification from results of mass spectral library searches. J Am Soc Mass Spectrom 5:316–323
107. NIST Mass Spectral Search Program, version 2.0d, and NIST/EPA/NIH Mass Spectral Library (2005)
108. Mascot Search Results. http://www.matrixscience.com/cgi/master_results.pl?file=../data/F981122.dat. Accessed 19 Nov 2009
109. Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 74:5383–5392
110. Vershinin VI, Derendyaev BG, Lebedev KS (2002) Computer-assisted identification of organic compounds (In Russian). Akademkniga, Moscow
111. Steinbeck C (2004) Recent developments in automated structure elucidation of natural products. Nat Prod Rep 21:512–518

112. Elyashberg M, Blinov K, Molodtsov S, Smurnyy Y, Williams AJ, Churanova T (2009) Computer-assisted methods for molecular structure elucidation: realizing a spectroscopist's dream. J Cheminformatics 1:3. doi:10.1186/1758-2946-1-3
113. Stein SE (1995) Chemical substructure identification by mass spectral library searching. J Am Soc Mass Spectrom 6:644–655
114. Milman BL (2005) Identification of chemical compounds. Trends Anal Chem 24:493–508
115. Evett IW, Gill P (1991) A discussion of the robustness of methods for assessing the evidential value of DNA single locus profiles in crime investigations. Electrophoresis 12:226–230
116. FAO/WHO Codex Alimentarius. Guidelines on the use of mass spectrometry (MS) for identification, confirmation and quantative determination of residues (2005) CAC/GL 56-2005. http://www.codexalimentarius.net/web/standard_list.jsp. Accessed 16 May 2010
117. SOFT/AAFS Forensic Laboratory Guidelines (2006). http://www.soft-tox.org/docs/Guidelines%202006%20Final.pdf. Accessed 16 May 2010
118. Faksness LG, Weiss HM, Daling PS (2002) Revision of the Nordtest methodology for oil spill identification. SINTEF Report STF66 A02028. http://www.nordicinnovation.net/nordtestfiler/tec498.pdf. Accessed 16 May 2010
119. ASTM D 3415 (1998) Standard practice for identification of waterborne oils

# Chapter 5
# Target Identification in Methods

**Abstract** Target identification is considered in detail. A qualitative analysis of this type is mostly performed according to validated methods which are screening and confirmatory. An identification result is the conclusion based on criteria. Those for screening identification are not very rigorous and not numerous. An example is the presence of a particular mass chromatographic peak in a rather wide range of the retention parameter. Most chromatographic techniques are suitable for screening. For confirmation of identity, more analytical data are required, e.g., three or four mass peaks and matching tolerance/range criteria for peak intensities. Any such value is named an identification point. An analyst should gather the required number of points. Chromatography and mass spectrometry and their combinations are the most appropriate techniques for the purpose. Different versions of the techniques, as well as other types of spectroscopy, are considered. The requirements and guidelines for setting up identification criteria presented in a number of laboratory guidances which have been issued by various organizations and agencies are outlined in detail. These are not the same in different documents; that is the reason for criticizing them. The system of identification points itself and the evident or suspected invalidity of tolerance criteria has also been criticized. The criticism is partly accepted, and some objections are also presented here. In general, the guidelines are regularly tested through a global analytical practice, and new improvements of identification criteria are reported.

## 5.1 General

Target identification (see Sect. 1.5.1) is mostly performed according to validated (standard, standardized) methods. The fact that methods are already developed and established provides

- Analytical selectivity, i.e., separation of interfering compounds leading to FP
- Identification criteria set up during a method development and confirmed by its validations

B.L. Milman, *Chemical Identification and its Quality Assurance*,
DOI 10.1007/978-3-642-15361-7_5, 

- The potential for confirmation of identification results by co-analysis (see Table 1.4) with analytical standards (reference materials, solutions) prescribed in methods.

Therefore, it is not very hard to avoid FP and FN when determining target analytes by methods. This statement refers to screening methods to a lesser degree, because they are not free from errors. This is the reason why screening and confirmatory methods will be separately considered.

## 5.2 Screening

***General***. Two main attributes of screening methods are (1) a high throughput and (2) rather low *FNR* (Sect. 2.9). Many analytical techniques, first of all simple analytical tests (Chap. 2), make it possible to obtain fast results. Modern chromatography techniques also provide high throughput performance, i.e., a separation of tens and hundreds of components in several minutes. The second characteristic, a low level of FN (e.g., $\leq$5% for pharmaceutical and pesticide residues in food [1, 2]), is established during the development and validation of a method. Further, negative results cannot be confirmed but should be controlled by accompanying measurements (e.g., by determining recovery; see [2]).

***Criteria for identification*** in screening methods are related to a presence of an analytical signal

- Within the value range of a measurand, e.g., RT in chromatography, $m/z$ in mass spectrometry, and so on (Sect. 3.6.3)
- Above the level corresponding to low limit concentration/amount (decision/detection limit, see Sect. 4.3), e.g., the level of S/N = 3:1

The second of the two criteria is typical for a detection procedure, and connected with one or another estimate for detection limit (Sect. 4.3.2.1). For example, according to the EC guidance [1], the corresponding limit value $CC\beta$ is determined by the mentioned 5% $\beta$ error (*FNR*, see above).

The first criterion resembles that for confirmation procedures. The differences are that here

- Only one or two quantities (e.g., the retention parameter plus $m/z$ of abundant characteristic ion in chromatography mass spectrometry) are required, whereas more individual criteria (three and more points, see below) are commonly intended for confirmation
- Criteria may be less rigorous, e.g., the tolerance for RT in GC is 1.5–3% vs about 0.2% in the case of confirmation [3]

***Screening and confirmation***. When proceeding to confirmation in chromatography mass spectrometry, a measurement for one mass number is supplemented by (e.g., see [4–6])

- Increasing number of recorded characteristic ions
- Recording full $MS^1$ scans
- Recording $MS^n$ spectra
- Measuring accurate ion masses

In some cases, the availability of one to two identification points may be sufficient for unambiguous recognition of a substance. However, this is not the case for fast *multi-target* determinations, where the same experimental conditions and similar criteria cannot be chosen to be equally effective for all the analytes (see [7]).

In any case, positive screening results, particularly "a suspected non-compliant result" [1], are or may be validated using confirmatory methods (Sect. 5.3). General requirements for analytical methods to be met for the use in confirmation of initial results are that they should be independent (orthogonally selective [8]). If confirmatory methods are used without screening ones from the very beginning, the former should be more specific then the latter. Individual requirements depend on the sort of analytes and the type of analytical technique. This can be demonstrated by two examples. The first is related to methods of residue/trace analysis of products of animal origin (Table 5.1). If requirements for confirmatory methods, or rather techniques underlying methods, are not met (see the 3rd column of

**Table 5.1** Requirements for confirmatory methods for residues [1]

| Technique | Group of substances[a] | Requirement to be met |
|---|---|---|
| GC–MS, LC–MS | A, B | Chromatography provides separation<br>Full mass spectra or at least three (group B) or four (group A) identification points |
| GC–IR, LC–IR[b] | A, B | Specific requirements for IR spectrometry (see below) |
| GC–electron capture detection | B | Two columns of different polarity[c] |
| LC–UV–Vis (full scan, DAD) | B | Specific requirements for UV spectrometry (see below) |
| LC–UV–Vis (single wavelength) | B | At least two different chromatographic systems or a second, independent detection technique[c] |
| LC–fluorescence | B | For compounds that exhibit native fluorescence or become fluorescent after either transformation or derivatisation |
| LC–immunogram | B | At least two different chromatographic systems or second, independent detection technique[c] |
| TLC–UV–Vis (full scan) | B | Two-dimensional HPTLC and co-chromatography |

[a]Group A contains the banned substances ("which are prohibited from use in food-producing animals in the EU"), e.g., substances having an anabolic effect. Group B contains "many pharmacologically active substances which may be authorised for use in food-producing animals in the EU", e.g., certain pesticides [9]

[b]Techniques rarely used

[c]Analytical signal of different type

Table 5.1), the method/technique becomes less specific and classified as a screening one.

The second example, taken from pesticide analysis (Table 5.2), demonstrates a variety of choice of confirmatory techniques for a particular method. However, the demand remains that the confirmatory method should provide complementary information, i.e., should not be the same and inconclusive.

For relatively high analyte amounts and simple mixtures, spectrometry techniques without combination with chromatography can also underlie confirmatory methods suitable for some analytical problems. In general,

> . . . the rigorousness required of a confirmation depends to some extent on the importance of the analytical finding and circumstances of the case [10].

This rule is of more value in unknown analysis (Chap. 7) but it should be taken into account in developing methods.

***Co-analysis***. In some cases, a confirmatory method does not work well for the purpose. To obtain the ultimate result, one can check the identification criteria using matrix-matched standards or other techniques [3]. It should always be taken into account that the most conclusive/confirmatory identification procedure is co-analysis, first of all co-chromatography (see Sect. 1.6). This is the procedure where the sample to be analysed or material extracted from it is fortified with an analytical standard of a suspected/candidate compound contained in a sample. Then prescribed analytical procedures are fulfilled. A candidate for identification can be considered as identified, if [1]

**Table 5.2** Recommended screening and confirmatory technique for pesticides [3]

| Confirmatory technique | Screening | | | | | |
|---|---|---|---|---|---|---|
| | GC, specific detectors[a] | GC–MS | LC–MS | LC–UV–Vis (full scan) | LC–UV–Vis (single wavelength) | LC–fluorescence |
| GC, specific detectors | +[b] | +[c] | + | + | + | + |
| GC–MS | + | +[d] | + | + | + | + |
| LC–MS | + | + | | + | + | |
| $MS^n$, HRMS, non-EI | + | + | + | + | + | + |
| LC–UV–Vis (full scan) | + | + | + | | + | + |
| LC–UV–Vis (single wavelength) | + | + | | | | + |
| LC–fluorescence | + | + | | + | + | |
| Derivatization | + | + | + | + | + | + |

TLC-enzyme assay also may serve as both the screening technique and the confirmatory one for all other techniques. If TLC is used in both steps, mobile or stationary phases should be of different polarity

[a]Detectors belong to the group consisting of ECD, NPD, FPD, and PFPD

[b]Column of different polarity or another specific detector should be used

[c]Column of different polarity should be used

[d]Another GC-MS method

- Only one peak is observed, the peak height (or area) being enhanced
- For GC or LC methods, the new peak width at half-height falls within the 90–110% range of the original width, and the retention times are similar within the 5% deviation margin
- For TLC technique, only the spot assigned to the analyte should be intensified, without any other changes of the visual appearance

In less simple cases (no validated methods, unknown analysis), this kind of tests should be repeated using a different chromatography system, e.g., a column of different polarity.

## 5.3 Confirmation

For this purpose, several range criteria are usually required (Sect. 3.6). It should be repeated here that a value of a measurand $x_i$, e.g., a retention parameter, a relative intensity of spectral peak, and so on should fall within a relatively narrow range from $x_{ri} - \Delta x_{ri}$ to $x_{ri} + \Delta x_{ri}$, where $x_{ri}$ is the reference value of the $i^{th}$ measurand, and $\Delta x_{ri}$ is its acceptable deviation. The set of criteria depends on both the particular technique and the chemical nature of the analyte. Chromatography mass spectrometry is the principal confirmatory technique.

The *confirmation* term is here used in the sense of *confirmation of identity* (e.g., see [11]), *confirmation of identification result*. Its second interpretation is "confirmation of an analyte presence in a sample" [8]. For both confirmatory purposes, not only different techniques but also different extracts of the same or a second sample should be used [8, 10].

The essential requirements for development, validation, and status of confirmatory methods are given in many laboratory guidance documents issued by national or international organizations/agencies (Sects. 5.4 and 5.5).

## 5.4 EPA Confirmatory Methods

The US Environmental Protection Agency have developed pertinent analytical methods for many years. Methods based on GC–MS (Table 5.3) have become classical and widespread in the analysis of volatile and semivolatile compounds. Identification criteria included are related to the window ranges about (GC) the reference RT or RRT and (EI–MS[1]) the relative peak intensities (ion abundances) of at least three characteristic ions. If the values of those quantities fall outside the ranges, a negative result is concluded.

A few EPA methods will be also cited below.

**Table 5.3** The identification criteria in classical EPA methods for GC–MS and volatile/semi-volatile compounds

| GC | MS | General |
|---|---|---|
| Window for the analyte peak about RT or RRT of the standard at calibration: ± 3 *s* of mean RT [12]; ± 5 s [13], 10 s [14, 15], 30 s [16]; ± 0.06 RRT [17–19] | The presence of:<br>- all ions having relative abundance >10% in the standard spectrum and some minor ions of special importance (molecular ion) [12–15],<br>- three ions with preset *m*/*z* [16],<br>- three characteristic/diagnostic ions with highest abundances or any ions with abundance >30% [17–19],<br>The agreement in peak relative abundances within absolute 20% [12–16] or 30% [17–19].<br>Abundances of the characteristic ions of an analyte maximize in the same scan or within one scan of each other [16, 19] | For the co-eluting analytes, identification may be based on the reference spectrum containing extra ions contributed by the co-eluting compound [12–15, 17–19]<br>For chromatographic resolution of the structural isomer peaks less than 25% of the sum of the two peak heights, the compounds are identified and reported as isomeric pairs [12–14, 16–19] |

There also are (a) the particular RRT window and (b) 30% tolerance range for the ratio of two characteristic ion abundances for every analyte in the GC–HRMS method [20]

## 5.5 Confirmation: Guidances and Methods of Various Organizations and Agencies

### *5.5.1 General*

The popular international and US laboratory guides considering identification procedures in a standard way and describing them in many detail are given in Table 5.4. It is easy to see that they refer to life-critical and socially significant fields of chemical analysis such as food, environment, toxicology, and sport. It should again be noted that chromatography mass spectrometry is the principal confirmatory technique, with individual independent (or conditionally independent) criteria set up for each of two parts of this combination. Many documents recommend that laboratories should establish their own criteria for identification based on the corresponding guidelines.

Some other guides have been issued which are useful for an implementation qualitative analysis, including identification, and its quality assurance. First of all, there was the report on the MEQUALAN (formed from MEtrology of QUALitative chemical ANalysis) European project [27]. In this report, many basic and applied

**Table 5.4** Documents of international and national organizations containing of rules or guidelines for identification

| Organization, agency | Year and reference | Analytes, matrices | Techniques | Details regarding identification |
|---|---|---|---|---|
| EU | 2002 [1] | Residues in products of animal origin | Numerous, including chromatography mass spectrometry | Requirements for screening methods, including estimating α and β errors, and confirmation methods, first of all identification criteria ("performance criteria") |
| FDA Center for Veterinary Medicine | 2003 [21] | Residues in products of animal origin | Chromatography mass spectrometry | Guidance for identification, including identification/confirmation criteria |
| AORC | 2003 [22] | Doping samples | Chromatography mass spectrometry | Guidelines for identification, including identification criteria |
| WADA | 2003[23] | Doping samples | Chromatography mass spectrometry | Guidelines for identification, including identification criteria |
| FAO/WHO Codex Alimentarius Commission | 2005 [3] | Pesticide residues in food | Mainly chromatography mass spectrometry | Guidelines for identification, including identification/confirmation criteria |
| ASTM | 2006 [24] | Organic compounds in water | Gas chromatography mass spectrometry | Standard guide for identification (without criteria) |
| SOFT/AAFS | 2006 [10, 25] | Drugs and other related to forensic toxicology | Numerous, including chromatography mass spectrometry | Guidelines for screening and confirmation methods (tests), with some identification criteria |
| ISO | 2006 [26] | Organic compounds in soils and other environmental samples | Gas chromatography mass spectrometry | Guidelines for identification, including identification criteria |
| EU | 2009 [2] | Pesticide residues in products of plant and animal origin | Chromatography mass spectrometry | Requirements for confirmation methods, with identification criteria |

approaches to qualitative analytical procedures, including chemical identification, and estimation of trueness of corresponding results, are treated.

A concise document treating uncertainties in qualitative analysis, without formulas and quantitative criteria, was developed by the EURACHEM Measurement Uncertainty Working Group [28].

From national documents, the guide on the best practice in qualitative analysis prepared by LGC (UK) [29] is worth mentioning. The document mainly describes chemical tests, but can also be extended to techniques of chromatography and spectrometry. In particular, the following requirements to analytical procedures were included.

- The classification criteria, i.e., identification ones, should be defined.
- Analytical methods should be documented, validated, and fit to the purposes. Rates of specificity, sensitivity, and misclassification (i.e., false results) should be known and controlled.
- Results of computational procedures should be carefully taken, reviewed, and checked for true conclusions. This requirement refers to computer libraries and also expert systems.

In general, these items agree with guides specified in Table 5.4 and issued later. The LGC document will be cited in great detail in Chap. 9, devoted to quality control and assurance. Requirements for identification procedures or corresponding recommendations presented in the documents listed in Table 5.4 will be considered below by technique.

### *5.5.2 Chromatography*

Common chromatography criteria are related to tolerance ranges about reference values of RT or RRT obtained at calibration (Table 5.5). These values are less reproducible in HPLC than with capillary GC. The acceptable RT of the analyte under identification is not less than twice RT, corresponding to the void volume of the column. For measuring RRT, some kind of an internal standard is used which is a substance related to RT close to this value of the analyte. Further, RRT values are reproduced better than corresponding absolute values of RT.

These regularities were taken into account when criteria were established, though the most rigorous criteria were set just for RT ( $\pm$ 1 s, see Table 5.5). The last criterion may be unrealistically rigid, in spite of the existence of the suitable technique named "retention time lacking" [30]. It should be noted here that the issues of RT reproducibility and also the related challenge of aligning chromatograms (see [31]) are very important for quality assurance of qualitative analysis.

In usual chromatography mass spectrometry analysis, several chromatograms for different ions (mass chromatograms) are recorded or extracted from the full recorded data. Therefore, maximums of chromatographic peaks of the same analyte

**Table 5.5** Permitted ranges about RT or RRT of reference compounds

| Guide | GC | HPLC[a] |
|---|---|---|
| EU[b] [1, 2] | RRT, ± 0.5% | RRT, ± 2.5% |
| FDA [21] | RRT, ± 2% | RRT, ± 5% |
| AORC[c] [22] | RRT, ± 1% | RRT, ± 2% |
| | RT, ± 1% or 6 s (whichever is the greater) | RT, ± 2% or 12 s (whichever is the greater) |
| WADA [23] | RT, ± 1% or ± 0.2 min (whichever is smaller) [c] | RT, ± 2% or ± 0.4 min (whichever is smaller) |
| FAO/WHO[d] [3], ISO[e] [26] | RT, ± 1 s (RT<500 s), RRT, ± 0.2% (RT 500–5,000 s), RT, ± 6 s (RT>5,000 s) | |

[a]The example can be added where the range of ± 15 s is suggested; see the EPA method [32]
[b]In chromatography mass spectrometry, peaks in mass chromatograms should be of S/N > 3:1 and of similar retention time and shape to those obtained from a calibration standard at comparable concentration. Chromatographic peaks of different characteristic ions for the same analyte must overlap with each other. In the case of significant chromatographic interference, residues must be not identified. Subtraction of background spectra may be required to remove chemical noise [2]
[c]If the chromatography system is overloaded by the sample, these criteria may be relaxed [23]. This factor also affects ratios of peak intensities of characteristic ions [2, 26]
[d]In subsequent injections of solutions of the analyte and the standard of the same compound, both are matrix extracts, the difference between the analyte and the standard in RRT typically is < 0.1%
[e]Difference in RT of peaks of all characteristic ions of the analyte ≤ ± 20% of the peak half-widths or ± 1 s. See also Example 4.6

should fall within a narrow RT range. This is another criterion for identification (see bottom lines, Tables 5.3, and the footnotes[b,e], Table 5.5).

Here and below (mass spectrometry), all considered criteria are supplemented by the requirement to S/N ratios which should not be very low, mostly not lower than 3:1.

For the concept of identification points with regard to chromatography, see Sect. 5.5.3.2.

## *5.5.3 Mass Spectrometry*

### 5.5.3.1 Full Scans and Selective Monitoring

General and many particular requirements for mass spectrometry identification are given in Table 5.6. The maximum permitted tolerances for relative peak intensities of selective ions are presented. They are expressed as percents of the base peak intensity (relative abundances, *I*). The percentage tolerances may depend on intensities. Such a version of the criteria is separately given in Table 5.7. Furthermore, the number of mass peaks taken into account during identification depends on what MS technique is used. The relationship between the effective number of peaks/ions, i.e., identification points and the technique type, is given in Table 5.8.

**Table 5.6** MS characteristics for identification criteria[a]

| Guide | MS, full scans | $MS^1$, SIM | $MS^n$, SRM |
|---|---|---|---|
| EU [1] | The presence of all ions (≥ four ions in $MS^1$) having $I > 10\%$ in the standard spectrum<br>Tolerances for $I$ are given in Table 5.7 | The selected characteristic ions preferably contain molecular ion, and originate from different parts of the molecule<br>The number of ions corresponds to a minimum four or three (compounds of group A or B, respectively; see Table 5.1) identification points; see text. Tolerances for $I$ are given in Table 5.7. | |
| EU [2] | The presence of three and more characteristic ions[b]. Tolerances for $I$ are given in Table 5.7. In $MS^n$, at least two product ions. Also, matrix-matched standards may need to be used for measuring ratios of ion abundances. | | |
| FDA [21] | The presence of three and more characteristic (structurally-specific) ions [c]<br>In $MS^n$, if the precursor completely fragments to product ions, at least two products should appear | $I$ for three or at least four characteristic ions match standard within ± 10 or ± 15 absolute % respectively | $I$ for two (precursor is fragmented) or at least three characteristic product ions match standard within ± 10 and ± 20 absolute %, respectively |
| AORC [22] | The presence of all ions (≥ three ions in $MS^1$) having $I > 10\%$ in the standard spectrum[d] | ≥ four ions with stricter tolerances than required for full scan | |
| WADA [23] | The presence of all ions having $I > 10\%$ in the standard $MS^1$ spectrum or lower peaks of characteristic ions (<5% in $MS^n$)[e]<br>Tolerances for $I$ of three characteristic ions in $MS^1$ (≥2 ions in $MS^n$) are given in Table 5.7 | The presence of three or more characteristic ions<br>Tolerances for $I$ of three characteristic ions are given in Table 5.7 | Tolerances for $I$ of more than one product ion are given in Table 5.7 |
| Other | | [f, g, h] | |

[a]Unless otherwise stated, standard conditions of gas chromatography mass spectrometry ($MS^1$) analysis are implied, which are EI, electron energy of 70 eV, several scans per chromatographic peak, and so on (e.g., see [24, 26]). The combination of HPLC and ESI is very typical for $MS^n$. As a rule, only signals with S/N >3:1 are considered

[b]Molecular or related ion should be included in identification procedure whenever possible. In general, high $m/z$ ions are more characteristic than low $m/z$ ions. Ions arising from loss of water or that sort of low molecules may be useless. Selected characteristic ions are recommended to be selected from different parts of the analyte molecule

[c] Ions related to water loss and isotopic peaks are discouraged

[d]Molecular or analogous ion is taken into account if >5% in $MS^1$. In $MS^n$, the precursor ion is selected if insufficient ions are present; $I$ should be in the range of 10–80%. If <3 suitable ions, different techniques or derivatizations may be used. Tolerances in $MS^1$ ($MS^n$): match standard within ± 10 (20) absolute % or ± 30 (40) relative %, whichever is the greater

[e]If three characteristic ions with ($I > 5\%$ in $MS^1$, full scans) or any characteristic ions ($MS^n$) are not available, a second ionization/fragmentation technique or a second derivative yielding different ions should be used, with ($MS^1$) two or more characteristic ions in each spectrum. In $MS^2$, the precursor ion should be present. In some cases a single precursor–product pair may be characteristic

[f]The tolerance ranges for the two or three ion abundance ratios match standard "within the limits of ± 30% of absolute ion abundances ratios" [3]
[g]Characteristic ions match standard within an average ± 20 and ± 25–30 relative % in GC–MS and LC–MS respectively [10]
[h]Characteristic ions with high $m/z$ values, especially the molecular ion, and high abundances (>15%) are preferred due to their higher significance. Even mass fragment ions are preferred over odd ones. For characteristic isotope clusters, e.g., chlorine, two characteristic ions are selected from the same cluster, and so on. The presence of three characteristic ions, with $I$ matching standard within $\pm (0.1 \cdot I_{std} + 10)\%$, where $I_{std}$ is the relative intensity (in absolute %) of the peak of the corresponding characteristic ion in the standard spectrum [26]. See also Example 4.6

**Table 5.7** Maximum permitted tolerances for selective ions

| $I$, % | GC–EI–MS | | GC–CI–MS, GC–MS$^n$, LC–MS, LC–MS$^n$ | |
|---|---|---|---|---|
| | EU [1, 2] | WADA [23] | EU [1, 2] | WADA [23] |
| >50% | ± 10% (relative) | ± 10% (absolute) | ± 20% (relative) | ± 15% (absolute) |
| ≥25 to 50% | | ± 20% (relative) | | ± 25% (relative) |
| >20 to 50% | ± 15% (relative) | | ± 25% (relative) | |
| >10 to 20% | ± 20% (relative) | | ± 30% (relative) | |
| <25% | | ± 5% (absolute) | | ± 10% (absolute) |
| ≤10% | ± 50% (relative) | | ± 50% (relative) | |

**Table 5.8** The number of identification points for different MS techniques [1, 33]

| Technique[a] | IP per ion |
|---|---|
| MS[1] | 1.0 |
| MS$^n$, precursor ion | 1.0 |
| MS$^n$, product ion | 1.5 |
| HRMS[1] | 2.0 |
| HRMS$^n$, precursor ion | 2.0 |
| HRMS$^n$, product ion | 2.5 |

[a]Unless otherwise stated, low-resolution mass spectrometry

Thus, full spectra or selected ions are used or recommended for identification in methods (Tables 5.6–5.8). In the first case, visual inspection of an analyte spectrum and its comparison to a reference one is performed (Table 5.9). Searches in computer spectral libraries, followed by automatic calculation of MF, are also permitted. However, library matches must be reviewed by a qualified experienced analyst (Table 5.9). In general, the guidance demonstrate a cautious attitude to the use of computer spectral libraries in target analysis, and do not disregard them, because these documents cover to some degree challenges of non-target analysis (see Chap. 7).

In selective ion detection (SIM, SRM), different (a) maximum tolerance ranges (Tables 5.6 and 5.7) and (b) numbers of ions/IP (Tables 5.6 and 5.8) have been proposed. Tolerances in different documents are not fully agreed (Fig. 5.1).

**Table 5.9** Full mass spectra in identification[a]

| Guide | Requirements and guidelines to procedures and operations |
|---|---|
| EU[1] | Computer spectral *libraries* searching *may be used*. MF for mass spectrum of the analyte and that of the standard has to exceed a "critical" level. This factor is determined during the validation process for every analyte. |
| | Spectral variations "caused by the sample matrix and the detector performance" are checked. |
| EU [2] | *Reference spectra* should be generated using the same instruments as employed for analysis of the samples. These spectra should be validated if they significantly differ from published spectra. The reference spectrum can be recorded during a previous injection without matrix components, preferably from the same experimental batch (e.g., with calibration solutions). Signal of an analyte must not overload the detector. |
| FDA[b] [21] | There should be a general correspondence between analyte and standard in relative/ranked abundances. The mass spectrum of a suspect compound should visually match that of a contemporaneous standard. |
| | All characteristic ions are present above a minimum level. The last is established by the method developer based on either *I* or S/N. |
| | "*Library* search algorithms should *not be used* to confirm identity". "Strict numerical criteria [i.e., MF – author] need not be applied". |
| | Appearance of abundant ions other than from targets can be explained, for example, by the presence in blanks |
| WADA [23] | The use of the computer *libraries is permitted*. Criteria based on MF are established in the laboratory, but there is no guarantee of identification based on MF. Thus, all library matches must be reviewed by a qualified scientist |
| ASTM [24] | Computerized spectral matching and manual searching of mass/intensity matches are included. |
| | The *library* of reference spectra should contain spectra of all organic compounds that may be present in the samples. The spectra should be recorded on the same instrument and under the same conditions as the unknown. A mass spectrometrist should be capable of evaluating the information provided with the spectral matches. |
| | The peak-by-peak comparison of the full mass spectrum of the analyte with that of an authentic sample of the suspected compound should be also used. |
| SOFT/ AAFS [10] | Searches in commercial/user-compiled *libraries* are performed. MF "must be used as guides only and are not sufficiently reliable to be used as the final determinant of identification." Finally, an experienced toxicologist must critically review spectral matches. |
| | For a positive match, all the abundant characteristic ions present in the reference spectrum must be present in the spectrum of "unknown". Ions that are non-abundant in the reference spectra may be missing in the latter. Additional major ions from a co-eluting substance or "background" may be present. |
| ISO [26] | Agreement with the mass spectrum of the pure compound and absence of other ions. |

[a]Unless otherwise stated, EI–$MS^1$
[b]$MS^1$ and $MS^n$ full scan.

An analyst can choose the criteria from the suitable guidance or any criteria from Tables 5.6–5.8, given that they are checked for ultimate performance (e.g., the absence of FP and FN) on the method validation. Examples of estimating tolerances and counting IP are as follows (Examples 5.1 and 5.2, see also Sect. 4.3.2.2).

**Example 5.1**

The reference spectrum of a compound contains three peaks of characteristic ions, intensities *I* of which are 100, 50, and 20%. An analyst should estimate tolerance ranges to use this spectrum as the reference one for identification. The dependence of ranges on *I* is taken into account there. The percentage ranges necessary for calculations are given in Table 5.7. Arithmetic calculations are placed in Table 5.10.

Tolerances obtained from calculations somewhat differ between guides for the same technique(s). An analyst may choose the range set (a) according to the specialization of the laboratory (e.g., determination of residues or doping), or (b) in an arbitrary way. In the second case, the range criteria are checked for acceptance for identification.

**Table 5.10** Examples of permitted tolerances for three selective ions

| *I*, % | GC–EI–MS | | GC–CI–MS, GC–MS$^n$, LC–MS, LC–MS$^n$ | |
|---|---|---|---|---|
| | EU [1, 2] | WADA [23] | EU [1, 2] | WADA [23] |
| 100% | 100% | 100% | 100% | 100% |
| 50% | (50 ± 0.15·50) % = (50 ± 7.5) % | (50 ± 0.2·50) % = (50 ± 10) % | (50 ± 0.25·50) % = (50 ± 12,5) % | (50 ± 0.25·50) % = (50 ± 12.5) % |
| 20% | (20 ± 0.2·20) % = (20 ± 4) % | (20 ± 5) % | (20 ± 0.3·20) % = (20 ± 6) % | (20 ± 10) % |

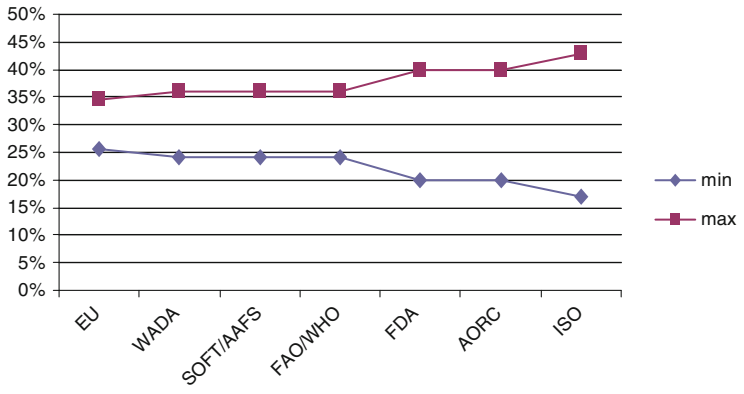



**Fig. 5.1** The tolerance ranges about $I = 30\%$ of an arbitrary mass peak, from minimum to maximum limits, according to different guidances (Tables 5.6 and 5.7)

### 5.5.3.2 Identification Points

The concept of IP came into being not long ago [1, 33]. Originally they were mainly related to confirmation by MS (Table 5.8). The example of counting IP for various MS techniques is as follows.

**Example 5.2**

An analytical mass spectrometrist would like to know whether the certain number of selective ions detected using various MS techniques provide sufficient IP and therefore true identification of analytes in the samples. One can easily do it calculating the number of IP with the use of initial relationships; see Table 5.8. The typical outcomes for required three/four points or their larger numbers are given in Table 5.11.

**Table 5.11** The number of identification points for various techniques (adapted from [1])

| Technique | The number of ions | The number of IP |
|---|---|---|
| GC–EI–MS, GC–CI–MS, LC–ESI–MS[a], LC–APCI–MS[a] | 3 | 3 |
| | 4 | 4 |
| | n | n |
| GC–EI–MS and GC–CI–MS | 2 (EI) + 2 (CI) | 4 |
| GC–EI–MS or GC–CI–MS, analyte and derivative or two derivatives | 2 (analyte) + 2 (derivative) | 4 |
| | 2 (derivative A) + 2 (derivative B) | |
| GC–MS and LC–MS[a] | 2 + 2 | 4 |
| GC–$MS^2$, LC–$MS^2$ | 1 precursor and 2 product ions | 4 |
| GC–$MS^2$, LC–$MS^2$ | 2 precursors, each with 1 product ion | 5 |
| LC–$MS^3$ | 1 precursor, 1 product, and 2 second-generation product ions[b] | 5,5 |
| HRMS | 2 | 4 |
| | *n* | 2*n* |
| GC–MS and LC–HRMS | 2 + 1 | 4 |

[a]There may be the only characteristic ion unless in-source fragmentation is used
[b]Ions are only counted once

The concept of IP is also applicable to chromatography. It has been noted [1] that "a maximum of one identification point' may be obtained with the following techniques":

- Combinations of HPLC with DAD or fluorescence detection
- HPLC coupled to an immunogram
- Two-dimensional TLC coupled to spectrometric detection

Recently, the concept of IP in the broad sense was introduced into the ISO standard for GC–MS identification of target compounds in soil samples and also other environmental samples [26]. According to this standard, additional IP are assigned also to a chromatographic signal recorded with another column or signal of specific detector, a specific chromatographic pattern, even prior data, and so on (Table 5.12). Additional IP are required in a multistep identification procedure, if an

**Table 5.12** The number of identification points in GC–MS (adapted from [26])

| Technique, device, procedure, or information[a] | Remark, example | IP |
|---|---|---|
| MS, characteristic ion | every ion | 1 |
| GC–EI–MS and GC–CI–MS CI, positive/negative ion | 1 (EI) + 2 (CI) | 3 |
| GC–$MS^2$ | 1 precursor and 2 product ions | 4 |
| HRMS | every ion | 2 |
| Column with other polarity[b] | extra RT as GC criterion. | 1 |
| Spike/standard addition[c] | | 1 |
| Isotope dilution | | 1 |
| Chromatographic pattern[d] | i.e., PCB, PAH, dioxins | 1 |
| Other techniques[e] | | 1 |
| Expectation, plausibility, previous investigations[f] | | 1 |

[a] As above, the condition of S/N>3 must be met for analytical signals. One IP is also provided with "absence of any other ions in full scan" [26]; there is no explicitation for this source of IP
[b]Not valid for non-separated isomers (e.g. chrysene/triphenylene, *m*/*p*-xylene)
[c]This is the initial part of co-analysis. In fact, a standard addition, followed by chromatography mass spectrometry analysis, automatically results in several IP and provides the strongest ultimate evidence for identification
[d]I.e. a peak belonging to the compound group easily identified by its fingerprint
[e]Every other selective detector (e.g., ECD for organochlorine compounds) or technique (e.g., LC–fluorescence for PAH)
[f]In this book, these are named prior data; see Chap. 6

attempt to gather the necessary three IP from the initial mass spectrum has not been successful (Fig. 5.2).

It should be noted that in this standard [26] a sound IP is assigned to prior information ("expectation, plausibility, earlier investigations"). This is not in line with the author's view that prior data are sources for setting up hypotheses rather than confirmatory evidences (see Chap. 6).

#### 5.5.3.3 High Resolution

The use of HRMS is recommended in some documents (Table 5.13), first of all the guidance on MS for confirmation of the identity of animal drug residues [21]. That is applicable to not only this group of analytes but also to a general unknown analysis by HRMS. Now, this MS technique is fast progressing (see below). In the future, new guidelines will be devoted to HRMS to a greater extent.

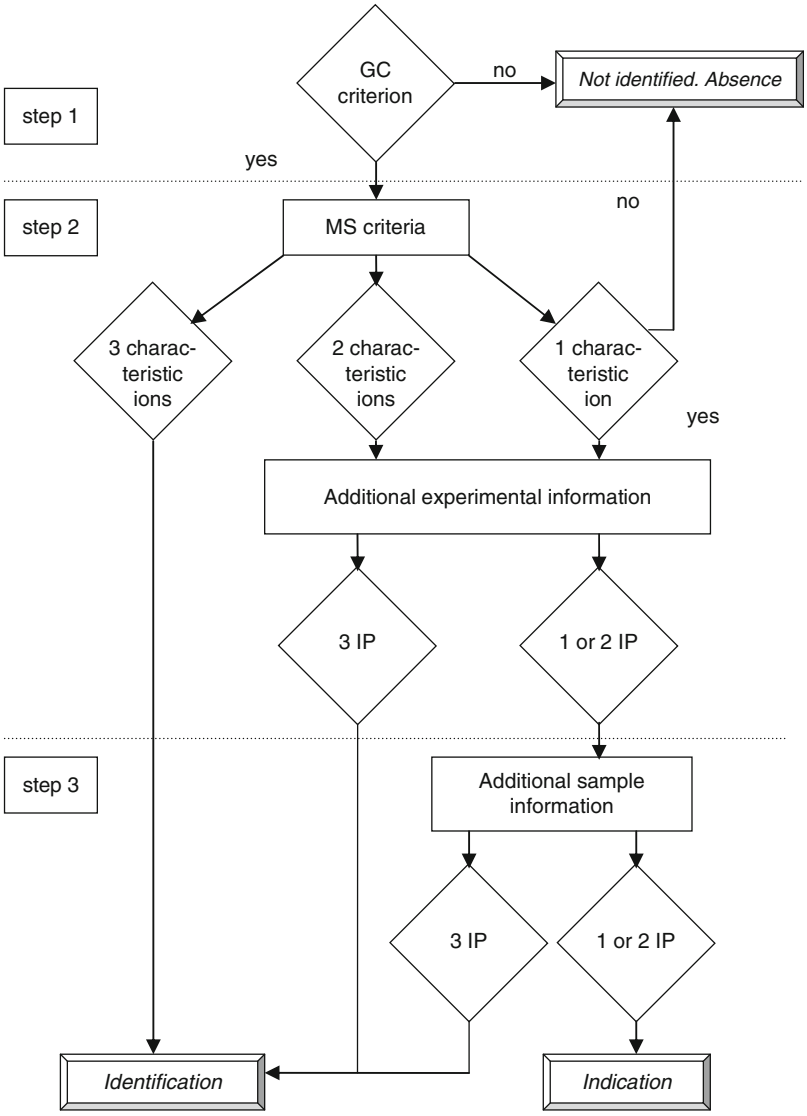



**Fig. 5.2** The flow scheme for identification procedure of environmental analytes (adapted from [26]). The first step is chromatographic screening. If the range criteria related to RT (see Table 5.5) are not met, the peak is considered *not identified*. If met, identification of corresponding compounds is confirmed by MS in the second step. Three tolerance criteria for *I* of characteristic ions (Table 5.6, bottom footnote) should be met. For three ions with the abundances within tolerances, identification hypothesis is accepted. Full non-match means no identification. Only one or two IP may be also observed. Corresponding ions are outside tolerances due to some factors, e.g., very low absolute intensity of peaks. The second possible reason is that spectra of such compounds as some PAH do not contain peaks of abundant ions but the only one. In any case, the gathering IP should be continued. They may come from additional experimental evidences or, if unsuccessful, a prior data (see Table 5.12). Finally, the set of IP becomes full or is kept incomplete. The result is classified as *identification* or *indication*, respectively

**Table 5.13** HRMS in guidance documents

| Guide | Requirements and guidelines to procedures and operations |
|---|---|
| EU [1] | Resolution >10,000 at 10% valley |
| FDA [21] | Mass resolution and peak purity should be sufficient to provide only one predominant ion formula per mass peak.<br>Standard of known formula/composition is used to demonstrate acceptable mass accuracy in method.<br>Accuracy is reported in the unit of ppm.<br>At low mass, $< m/z$ 500, the criterion of < 5 ppm difference is acceptable to confirm a unique molecular formula. At higher mass, this criterion does not provide unambiguous confirmation.<br>The possibility of other candidate compositions within the mass accuracy of the instrument is evaluated for C, H, N, O and some other elements. The range of candidate compositions should be shown. In addition, multiple candidates are evaluated for their logicality. The numbers of heteroatoms, the isotopic patterns, or other characteristics are taken into account. The possibility of alternative compositions diminishes the importance of exact mass measurements. |
| WADA [23] | Resolution >3,000 at 10% valley |
| FAO/WHO [3], ISO [26], EU [2] | The fact of the use of HRMS for confirmation is noted.<br>Requirements for identification: ≥2 characteristic ions (preferably including the molecular or related ion), at least one fragment ion. Mass accuracy <5 ppm. |

The EPA method [20] can be added where resolution >5,000 at 10% valley and two characteristic ions, are recommended

## *5.5.4 Other Techniques*

### 5.5.4.1 GC with Specific Detectors

The status of these techniques in relation to methods is specified in Sect. 5.2. Some requirements for the determination of an analyte by GC with ECD are noted in the document [1]. General requirements for GC, including the 0.5% tolerance for RRT and a possibility for co-chromatography, remain valid (see above).

For certain substances, selective detectors may be not sufficiently selective, resulting in FP. An example is that positive responses of ECD are observed not only to halogen compounds but also phthalate esters [3].

### 5.5.4.2 HPLC–UV–Vis

According to the EU guidance [1], there are the following requirements and criteria for the technique.

- A 2.5% tolerance interval for RRT and other general requirements to chromatography methods are proposed, see Sect. 5.5.2.

- In conditions of diode array detection, the absorption maxima in the spectrum of the analyte typically match those of the corresponding analytical standard within ± 2 nm wavelength intervals.
- For above 220 nm and the spectral regions with a relative absorbance ≥10%, the spectrum of the analyte and that of the standard are not visibly different, i.e., the same maxima are present and the difference between absorbances at any point is not larger than 10% of the absorbance in the standard spectrum.
- In the case of searching in computer spectral libraries, the analyte spectrum matches that of the standard solution above a critical MF. The latter is determined during the validation procedure for every analyte. Spectral variability caused by various factors is checked.

Here, the tolerance of ± 2 nm is somewhat wider than corresponding values of resolution, accuracy, and precision of modern diode array detectors. For example, optical resolution, wavelength accuracy, and wavelength repeatability are 1.2, ± 1.0, and ± 0.1 nm respectively [34]. So this tolerance range is suitable, if only instrumental factors affect a spread in wavelength. However, this is not the case when conditions for recording spectra of analyte and standard are not the same, e.g., there is some difference in pH. Therefore, the range of that tolerance should be checked during validation of the method.

A detector of this type is also used in TLC (next Section)

#### 5.5.4.3 Thin Layer Chromatography

Requirement established for this techniques are the following ([1]; see also Sect. 5.2).

- Such technique options as two-dimensional HPTLC and co-chromatography are considered mandatory.
- The tolerances for the RF values of analyte are ± 5% with reference to those for the analytical standards.
- The spot of the analyte is visually indistinguishable from that of the standard. The same is true for corresponding spectra recorded at full-scan UV–Vis detection.
- The separation of spots of the same color should be so effective that centers of the spot of the analyte and the nearest one are separated by a distance of not less than half the sum of the spot diameters.

For detection by UV–Vis, see also Sect. 5.5.4.2.

#### 5.5.4.4 IR Spectroscopy

The EU guide [1] (see also Sect. 5.2) exploits the concept of adequate peaks, which are absorption maxima in the IR reference spectrum of an analytical (calibration) standard fulfilling a number of the requirements.

- Absorption maximum is in the wavenumber range of 500–4,000 $cm^{-1}$.
- Relative intensity of absorption, with respect to (a) zero absorbance or (b) peak base line, is not less than (a) 12.5% or (b) 5% respectively of the absorbance of the most intense peak in the region noted above.
- There are a minimum of six adequate peaks in the reference spectrum of the standard. If there are less than six peaks, the spectrum of the standard is not qualified as the reference one.

Criteria for identification are the following [1] (see also [35]).

- Absorption is present in all regions of the analyte spectrum corresponding with adequate peaks of the standard.
- At least 50% of the adequate peaks are found in the IR spectrum of the analyte.
- Correspondence of peaks in the IR spectrum of the analyte with adequate peaks in the spectrum of the standard is determined within a tolerance range [1] of $\pm$ 1 $cm^{-1}$.
- "Where there is no exact match for an adequate peak, the relevant region of the analyte spectrum shall be consistent with the presence of a matching peak."
- The requirement of S/N $\geq$ 3:1 is applicable to absorption peaks of the analyte.

The remark should be made that the use of IR spectroscopy for a determination of residues concerned in the guide [1] is not fully appropriate due to a relatively low sensitivity of the technique. The IR spectral technique is far more effective for identification of materials which are accessible in large amounts. Therefore procedures of IR spectral analysis have been widespread in qualitative analysis II (Chap. 8).

#### 5.5.4.5 NMR Spectroscopy

There are a few standard methods using NMR, as compared with the abundance of those based on chromatography and mass spectrometry. However, a number of methods have been developed in ASTM and some other organizations (see [37]). Usually, the chemical shifts of particular groups in analyte molecules are main quantities used for identification as a qualitative part of methods. Three points should be noted which may give an advantage for NMR over other techniques for developing standard methods.

- NMR is very sensitive in relation to changes in molecular stereochemistry where different spectral techniques, first MS, meet with failure. One of the new examples is the use of $^{13}$C-NMR for regiospecific analyses of triacylglycerols to differentiate between fish oils for their authentication [38].
- A specificity of analytical results is further provided by the resonance techniques for the nuclei of $^{15}$N, $^{17}$O, $^{19}$F, $^{23}$Na, $^{29}$Si, $^{31}$P and some other elements, if these are contained in the respective molecules.

[1] This is probably a too rigorous criterion, see [36].

- This technique is required in metabolomics and also proteomics, which are among the principal challenges for modern analytical science and practice (Chap. 7).

## 5.6 Testing and Criticism of Guidances

There have been many scientific reports criticizing the guides considered in this chapter, firstly the requirements of the EU document [1]. An example of relevant sentences is as follows:

> What could be the scientific basis for the above criteria? As regards the tolerance windows, the general opinion is that the repeatability of ion ratio measurements decreases with lower RIs [relative intensities, i.e. *I* – author], but the guidelines do not provide substantiating references nor other indications how they arrived at the tolerance windows [in Tables 5.6 and 5.7 – author]. It seems that the latter are just based on arbitrary decisions. The reasons for switching between absolute and relative differences [Table 5.7 – author], as advocated by WADA, also remains unexplained.. . . obviously the widely divergent criteria between the Guidelines for number of ions to be monitored and tolerance windows in SIM are scientifically unsound and legally untenable. It cannot be that one and the same test result may lead to a 'positive' identification when using Guideline A and a 'negative' identification when using Guideline B. However, even if the identification result remains the same under all guidelines, we cannot guarantee that the result is correct [39].

Another judgment is also worth mentioning:

> . . . the identification-points system is not scientific. . . . as in the case of essentially all identification guidelines to date, a critical drawback is that a rigorous assessment has not been conducted to determine the uncertainty of the approach(es). For example, what are the differences in the rates of false positives and false negatives by requiring four IP for banned substances [group A, see Table 5.1 – author] over three IP for registered compounds [group B, Table 5.1 – author]? Why should a high-resolution ion always be worth two points in the IP system, and $MS^2$ ions always be worth 1.5, whereas the (pseudo)-molecular ion is only worth 1? [8]

The quotations show that the choice of (1) a length of tolerance ranges and (2) the number of ions/IP is the attackable target. Also, the use of tolerances themselves as the type of criteria has been put in doubt [40].

In general, one would find difficulty not to accept many things from the criticism. It is an indisputable fact that every analyte (or more exactly, the analyte in a particular matrix at a certain concentration) is the individual analytical target. Uniform requirements and guidelines do not provide unambiguous true identification in all such cases, e.g.,

> . . . many exceptions can be found that indicate a three-ion requirement is either too strict or not strict enough [8].

Further, it should be accepted that there is an uncertainty as to what particular level of FP and FN is related to tolerance criteria (see [8, 40]).

Nevertheless, the EU guide [1] and other documents (see Table 5.4) are based on the results of research by numerous reputed laboratories (e.g., see [33]). There

appears to be little doubt about the reliable empirical basis for many of the above requirements and many (but not all) proper analytes and matrices. There is something else that supports the discussed system of IP and the accompanying tolerances. This is the simplicity of criteria. According to our observations, many chemists engaged in routine analysis prefer simple rules and simple decisions. There are simple and uniform criteria that are consistent with expectations of analysts.

However, testing the recommendation and rules for identification in confirmatory methods has generated exemptions. Careful registration of the latter would lead to "the list of rules" supplemented with "the list of exemptions" as another basis for true identification or no identification. In turn, it would be some kind of a step to a new system of criteria. "The registration of exemptions" could be carried out in a natural way if it accompanies validation of qualitative methods in the spirit of proposals of the critical review [8].

Also, different documents should be mutually adjusted.

One of the aspects of the criticism is rather inconsistent. It was noted [8, 41] that uniform rigid tolerances were narrow as compared to experimental spreads of a ion abundance ratios, e.g., standard deviations about corresponding mean values. That may lead to a significant rate of FN (Fig. 5.3a). In contrast, other workers are concerned about excessively wide tolerance ranges established by guidelines and a subsequent chance for FP [42]; see Fig. 5.3b. It has been concluded that statistical criteria based on real data spreads for analytes in one or another sample would be narrower and therefore more correct [42].

However, it has been demonstrated that statistical processing of data does not improve results of qualitative determination of pesticide residues [43]. In our research, window/tolerance (RI in GC, MF in mass spectrometry) and statistical (*t*-test for the same data) criteria led to similar results of impurity identification [44]. So, a substitution of rigid windows by statistical criteria may be efficient or inefficient depending on the case and is not a panacea from all troubles, especially taking into account possible deviations from normal distribution [43]. Further research of this topic is certainly essential.

Another aspect of requirements and guidelines for identification and confirmation is due to the progress in MS techniques and the wide use of $MS^n$ and HRMS in numerous analytical laboratories. Corresponding range criteria (Tables 5.6–5.8 and 5.11–5.13) for those have been further tested for trueness of identification. Two findings will be outlined.

In pesticide residue analysis using HPLC–$MS^2$, there was the false determination of a coeluting interfering compound instead of the pesticide sebuthylazine [45]. The identification criteria matching the EU guide [1] included two $MS^2$ transitions (one precursor and two product ions; see Table 5.11 for the criterion) and just one abundance ratio of product ions (Table 5.7). It was found that those criteria were insufficient to obtain the true result. The FP error was detected by anything from (a) the third fragmentation of the precursor, (b) GC–MS, or (c) UHLC–HRMS [45].

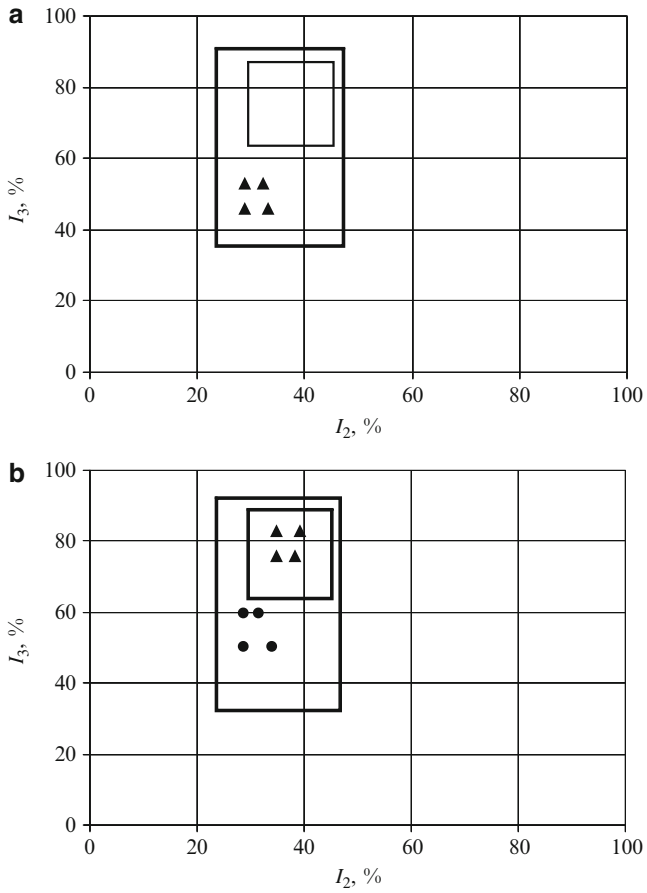



**Fig. 5.3** Adapted from [42]. Relative intensities of two mass peaks for two analytes (*triangles* for compound A and *circles* for analyte B, four experimental values per every analyte) and two criteria for *identification of A* as *rectangle* tolerances of different sizes. Any values of intensities are arbitrary. Criteria are "soft" (*large rectangle*) and "rigorous" (*small rectangle*). (**a**) The case where compound B is absent and the measurement result for A is somewhat biased. If the rigorous criterion is accepted, the identification result is FN (*triangles* are outside the *small rectangle*). In other words, one can falsely reject the hypothesis that analyte is compound A. For the soft criterion (*large rectangle*), *triangles* are inside and TP is declared. (**b**) Both A and B are present in the sample. Applying the rigorous criterion, one obtains the identification result which is the combination of TP (*triangles* are within the *small rectangle*) and TN (*circles* are outside). For the soft criterion, the result is FP in relation to compound B due to both groups of *points* being within the *large rectangle*

The second case is related to HRMS, particularly $HRMS^n$. Most relevant criteria for the technique, Table 5.13, look outdated and need to be improved. One of the proposals includes [46]:

- Minimum mass resolution measured for width at half peak maximum [2] (rather than at 10% valley, see Table 5.13) ≥10,000
- Different resolution values for screening ≥10,000, the reliable confirmation (up to ≥20,000), and unknown analysis (by HRMS[2], up to ≥70,000)
- Dependence of the number of IP (1.5 or 2 IP per ion vs 2–2.5 IP/ion; Table 5.8) earned by different ions on mass resolution
- Not only mass measurement but also calculation of at least one ion abundance ratio, likewise low-resolution mass spectrometry
- Different mass tolerances for screening and confirmation (± 50 and 5 mDa respectively were proposed)

In our opinion, the central ideas of the proposals [46] are healthy, and the particular values of tolerances and different criteria may depend on many factors, such as the molecular mass of analyte, its concentration, the nature of the matrix, and so on (see above).

## References

1. Commission Decision 2002/657/EC, August 12, 2002, implementing Council Directive 96/23/EC concerning the performance of analytical methods and interpretation of results (2002) Off J Eur Commun L 221:8-36. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2002:221:0008:0036:EN:PDF. Accessed 14 May 2010
2. Method validation and quality control procedures for pesticide residues analysis in food and feed (2009) Document No. SANCO/10684/2009. http://ec.europa.eu/food/plant/protection/resources/qualcontrol_en.pdf. Accessed 14 May 2010
3. FAO/WHO Codex Alimentarius. Guidelines on the use of mass spectrometry (MS) for identification, confirmation and quantitative determination of residues (2005) CAC/GL 56-2005. http://www.codexalimentarius.net/web/standard_list.jsp. Accessed 16 May 2010
4. Nielsen KF, Smedsgaard J (2003) Fungal metabolite screening: database of 474 mycotoxins and fungal metabolites for dereplication by standardised liquid chromatography–UV–mass spectrometry methodology. J Chromatogr A 1002:111–136
5. Gentili A, Perret D, Marchese S (2005) Liquid chromatography-tandem mass spectrometry for performing confirmatory analysis of veterinary drugs in animal-food products. Trends Anal Chem 24:704–733
6. Mueller CA, Weinmann W, Dresen S, Schreiber A, Gergov M (2005) Development of a multi-target screening analysis for 301 drugs using a QTrap liquid chromatography/tandem mass spectrometry system and automated library searching. Rapid Commun Mass Spectrom 19:1332–1338
7. Jiménez C, Ventura R, Segura J (2002) Validation of qualitative chromatographic methods: strategy in antidoping control laboratories. J Chromatogr B 767:341–351
8. Lehotay SJ, Mastovska K, Amirav A, Fialkov AB, Martos PA, de Kok A, Fernández-Alba AR (2008) Identification and confirmation of chemical residues in food by chromatography-mass spectrometry and other techniques. Trends Anal Chem 27:1070–1090
9. EC Food Safety. http://ec.europa.eu/food/food/chemicalsafety/residues/third_countries_en.htm#2. Accessed 16 May 2010

[2]More exactly, this quantity is named *mass-resolving power* [47].

10. SOFT/AAFS Forensic Laboratory Guidelines (2006). http://www.soft-tox.org/docs/Guidelines%202006%20Final.pdf. Accessed 28 Oct 2010
11. EURACHEM Guide (1998) The fitness for purpose of analytical methods. http://www.eurachem.org/guides/valid.pdf. Accessed 28 Oct 2010
12. US EPA Method 524.2 (1995) Measurement of purgeable organic compounds in water by capillary column gas chromatography/mass spectrometry. Revision 4.1
13. US EPA Method 525.2 (1995) Determination of organic compounds in drinking water by liquid–solid extraction and capillary column gas chromatography/mass spectrometry. Revision 2.0
14. US EPA Method 525.1 (1991) Determination of organic compounds in drinking water by liquid-solid extraction and capillary column gas chromatography/mass spectrometry. Revision 2.2
15. US EPA Method 548.1 (1992) Determination of endothall in drinking water by ion-exchange extraction, acidic methanol methylation and gas chromatography/mass spectrometry. Revision 1.0
16. US EPA Method 625. Base/neutrals and acids. http://www.epa.gov/waterscience/methods/method/organics/625.pdf. Accessed 16 May 2010
17. US EPA Method 8270C (1996) Semivolatile organic compounds by gas chromatography/mass spectrometry(GC/MS). Revision 3
18. US EPA Method 8275A (1996) Semivolatile organic compounds (PAHs and PCBs) in soils/sludges and solid wastes using thermal extraction/gas chromatography/mass spectrometry (TE/GC/MS). Revision 1.
19. US EPA Method 8260B (1996) Volatile organic compounds by gas chromatography/mass spectrometry (GC/MS). Revision 2
20. US EPA Method 1698 (2007) Steroids and hormones in water, soil, sediment, and biosolids by HRGC/HRMS. http://www.epa.gov/waterscience/methods/method/files/1698.pdf. Accessed 18 May 2010
21. FDA Center for Veterinary Medicine Guidance for Industry (2003) Mass spectrometry for confirmation of the identity of animal drug residues. http://www.fda.gov/downloads/AnimalVeterinary/GuidanceComplianceEnforcement/GuidanceforIndustry/UCM052658.pdf. Accessed 17 May 2010
22. AORC Guidelines for the minimum criteria for identification by chromatography and mass spectrometry (2003). http://cobra.vdl.iastate.edu/aorc-2/AORC%20MS%20Criteria.pdf. Accessed 17 May 2010
23. WADA Technical Document TD2003IDCR (2003) Identification criteria for qualitative assays incorporating chromatography and mass spectrometry. http://www.wada-ama.org/rtecontent/document/criteria_1_2.pdf. Accessed 17 May 2010
24. ASTM D 4128 (2006) Standard guide for identification and quantitation of organic compounds in water by combined gas chromatography and electron impact mass spectrometry
25. Penders J, Verstraete A (2006) Laboratory guidelines and standards in clinical and forensic toxicology. Accred Qual Assur 11: 284–290
26. ISO Standard 22892 (2006) Soil quality – Guidelines for the identification of target compounds by gas chromatography and mass spectrometry
27. Valcárcel M, Cárdenas S, Barceló D, Buydens L, Heydorn K, Karlberg B, Klemm K, Lendl B, Milman B, Neidhart B, Ríos A, Stephany R, Townshend A, Zschunke A (2002) Metrology of qualitative chemical analysis. Report EUR 20605. EC, Luxembourg
28. Ellison SLR (2000) Uncertainties in qualitative testing and analysis. Accred Qual Assur 5:346–348
29. Hardcastle WA (1998) Qualitative analysis: a guide to the best practice. LGC. http://www.rsc.org/ebooks/archive/free/BK9780854044627/BK9780854044627-00001.pdf. Accessed 17 May 2010
30. Giarrocco V, Quimby B, Klee M (2000) Retention time locking: concepts and applications. Agilent Application 5966-2469E. http://www.chem.agilent.com/Library/applications/59662469.pdf. Accessed 18 May 2010

31. Rohrback BG, Ramos LS (2003) Aligning Chromatograms. Gulf Coast Conference, 2003. http://www.infometrix.com/apps/GCC2003_Align_L.pdf. Accessed 18 May 2010
32. US EPA Method 1694 (2007) Pharmaceuticals and personal care products in water, soil, sediment, and biosolids by HPLC/MS/MS. http://www.epa.gov/waterscience/methods/method/files/1694.pdf. Accessed 18 May 2010
33. Andre F, De Wasch KKG, De Brabander HF, Impens SR, Stolker LAM, Van Ginkel L, Stephany RW, Schilt R, Courtheyn D, Bonnaire Y, Furst P, Gowik P, Kennedy G, Kuhn T, Moretain JP, Sauer M (2001) Trends in the identification of organic residues and contaminants: EC regulations under revision. Trends Anal Chem 20:435–445
34. Waters ACQUITY UPLC Photodiode Array Detector 71500108703 / Revision A. http://www.waters.com/webassets/cms/support/docs/71500108703ra.pdf. Accessed 15 May 2010
35. De Ruig WG, Weseman JM (1990) A new approach to confirmation by infrared spectrometry. J Chemom 4:61–77
36. Ellison SLR, Gregory SL (1998) Predicting chance infrared spectroscopic matching frequencies. Anal Chim Acta 370:181–190
37. PNA Test methods performed. http://www.process-nmr.com/test_methods_performed.htm. Accessed 18 May 2010
38. Standal IB, Axelson DE, Aursand M (2009) Differentiation of fish oils according to species by 13C-NMR regiospecific analyses of triacyglycerols. J Am Oil Chem Soc 86:401–407
39. De Zeeuw RA (2004) Substance identification: the weak link in analytical toxicology. J Chromatogr B 811:3–12
40. Bethem R, Boison J, Gale J, Heller D, Lehotay S, Loo J, Musser S, Price P, Stein S (2003) Establishing the fitness for purpose of mass spectrometric methods. J Am Soc Mass Spectrom 14:528–541
41. De Boer WJ, Van der Voet H, De Ruig WG, Van Rhijn JA, Cooper KM, Kennedy DG, Patel RKP, Porter S, Reuvers T, Marcos V, Munoz P, Bosch J, Rodriguez P, Grases JM (1999) Optimizing the balance between false positive and false negative error probabilities of confirmatory methods for the detection of veterinary drug residues. Analyst 124:109–114
42. Faber NM (2009) Regulations in the field of residue and doping analysis should ensure the risk of false positive declaration is well-defined. Accred Qual Assur 14:111–115
43. Soboleva E, Ahad K, Ambrus Á (2004) Applicability of some mass spectrometric criteria for the confirmation of pesticide residues. Analyst 129:1123–1129
44. Milman BL, Kovrizhnych MA (2000) Identification of chemical substances by testing and screening of hypotheses. II. Determination of impurities in n-hexane and naphthalene. Fresenius J Anal Chem 367:629–634
45. Schürmann A, Dvorak V, Crüzer C, Butcher P, Kaufmann A (2009) False-positive liquid chromatography/tandem mass spectrometric confirmation of sebuthylazine residues using the identification points system according to EU directive 2002/657/EC due to a biogenic insecticide in tarragon. Rapid Commun Mass Spectrom 23:1196–1200
46. Nielen MWF, Van Engelen MC, Zuiderent R, Ramaker R (2007) Screening and confirmation criteria for hormone residue analysis using liquid chromatography accurate mass time-of-flight, Fourier transform ion cyclotron resonance and orbitrap mass spectrometry techniques. Anal Chim Acta 586:122–129
47. Marshall AG, Hendrickson CL (2008) High-resolution mass spectrometers. Annu Rev Anal Chem 1:579–599

# Chapter 6
# Prior Data for Non-target Identification

**Abstract** This chapter is devoted to prior information required to set up and test identification hypotheses. According to its type, the relevant information is divided into meaning and statistical data. Knowledge with regard to the origin, properties, and use of chemical compounds is very essential in order to be able to propose and reject candidate compounds for identification. Prior information about samples analyzed is important in order to gather full evidence of the trueness of an identification result. Plausibility of qualitative analytical results is also taken into account to confirm conclusions made by analysts. Much of such data are extracted from chemical databases outlined in this chapter. These data sources are also used to calculate statistical rates of occurrence and co-occurrence of chemical compounds in the literature. The occurrence rate is the direct measure of the abundance of chemical compounds, and the related possibility of presenting in samples to be analyzed. Rare compounds are filtered out by means of this rate, and further excluded from consideration for identification purposes. Most known compounds are rare ones, as proved by respective statistical data. Facts and rates of the co-occurrence of chemical compounds in the literature provide the possibility of a priori prediction of a group of compounds available in the same samples analyzed. Different methods of estimating these rates are described; examples of their use for identification are given.

## 6.1 General

This type of analysis will be considered in Chap. 7 with all possible details. In this chapter (and also the succeeding one), it will be demonstrated that data/information play the main role in obtaining true results of non-target/unknown identification. In this chapter, information about (a) the origin, properties, and use of chemical compounds, (b) their popularity (abundance, occurrence) and quantitative measures for them will be considered.

These data are/may be required for both setting up and screening candidates for identification (identification hypotheses).

B.L. Milman, *Chemical Identification and its Quality Assurance*,
DOI 10.1007/978-3-642-15361-7_6, 

## 6.2 A Variety of Prior Data

Candidates for identification can be predicted before or during chemical analysis by means of various prior data; see Table 6.1. Such information may be conditionally classified as *statistical data* (Table 6.1, the bottom line) and *meaning data* (the bulk of Table 6.1). Information analysis makes it possible to set up the particular candidate(s) for identification based on properties and characteristics of (a) the sample under analysis, or (b) already identified analyte(s) [1–3]; see Fig. 6.1.

**Table 6.1** Prior data as sources for identification hypotheses[a]

| Data | Remarks, examples |
|---|---|
| Method targets | The particular sample/matrix may contain compounds commonly determined in this matrix. Such compounds listed in corresponding methods can be considered as candidates for identification |
| Previous and current analyzes, information about samples | Analytical reports from the home or different laboratory may contain valuable information on possible analytes. Further, the presence of some compounds in the sample means that other compounds of the same groups can also be detected (e.g., PAH, PCB, PCDD/F) |
| Thermodynamic and kinetic data | Unstable compounds can be often excluded from consideration |
| Compounds synthesized/transformed in nature by regular rules. Compounds rare for natural matrices | (1) DNA determines amino acid sequence of proteins synthesized in the living organisms. (2) Metabolites are formed according to the regularities [6–9]. (3) Alkenes are not appreciable components of oil. (4) Cyclopropane derivatives are very rare substances in nature |
| Databases/reference books on known compounds | See Sect. 6.3. The example is that all compounds available in the database and having the certain molecular masses are considered as candidates for identification |
| Compositions and formulations. Origin and use of particular compounds | A composition of commercial mixtures of substances (e.g., glues or drug dosages) is relatively easy to search in corresponding reference books, databases, patent and other literature. Such formulations can therefore be considered as "mixtures" of candidates for identification to be tested during analysis of real samples. Related information on origin and use of the particular compounds can be found in some databases (see Sect. 6.3) |
| Lists of regulatory chemicals | As a rule, they are widely occurring compounds/substances. Collected in the CHEMLIST data base (see Table 6.2) |
| Hit lists | Search of matching spectra in corresponding spectral libraries results in the ranked list of the spectra of compounds which are advanced candidates for identification |
| *Occurrence* and *co-occurrence*[b] of chemical compounds in databases and the literature | Popular compounds, i.e., ones with high rates of occurrence in the chemical literature and databases, are present in a sample with a greater probability than rare compounds. Constituents of the same sample are or may co-occur with each other and with the name of the matrix in the chemical literature. So data on occurrence and co-occurrence are useful in setting up identification hypotheses or, equally, predicting composition of samples [1–3]. See Sects. 6.4 and 6.5 |

[a]For references, see [4, 5] unless otherwise noted

[b]The synonymic terms of *citation* and *co-citation* (*co-reference*), respectively are also used by the author [1, 2]

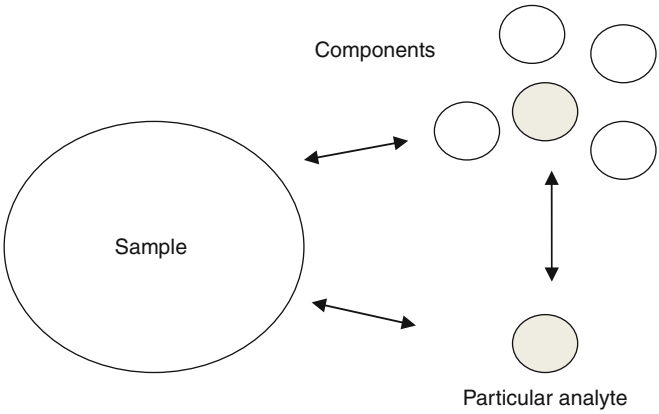



**Fig. 6.1** Information relations between a sample and its components. Analysis of a sample leads to identification of an individual analyte or all (many) sample components. The same analysis may determine the identity/authenticity of a sample itself (qualitative analysis II, Chap. 8). Results of previous analyzes have been directly or indirectly introduced into documents (protocols, reports, articles, database entries, and so on) which can be used to predict a sample composition from such retro-information on a sample or its individual components. This kind of prediction is based on not only analytical data but also similarity in properties between the compounds as sample components. This is expressed in co-occurrence of related compounds both in the same samples and in the same documents; see Sects. 6.4 and 6.5. Thus, if an analyst knows about the presence of one or several analytes, he/she may predict the existence of some other sample components and determine the nature of the sample itself. The kinds of reverse conclusions are also true.

In the discussion of various kinds of chemical information, it is important to find out what is the set of popular compound/substances which should be taken into account for the purpose of identification of components in real samples.

## 6.3 Set of Abundant Compounds

In spite of huge sets of known and especially possible compounds (Sect. 1.5.4), an analytical chemist cannot consider most of them which are rare ones. The latter are compounds synthesized or isolated from natural sources (detected in them), as a rule in subgram amounts, in a few laboratories. Such substances are not present in common samples chemically analyzed, such as environmental, food, or biochemical ones. Those differ from industrial chemicals, solvents and other abundant compounds (see Sect. 1.5.4) which can be detected in many matrices.

Abundance of chemical compounds can be estimated by their occurrence in the literature and databases [1–3]. Principal relevant databases are given in Table 6.2; there are also many other e-sources of chemical information, see [26, 27]. For the purpose of determining popularity/abundance, various observations or indicators can be used. From our point of view, the compound is considered abundant if it occurs

**Table 6.2** Chemical and biochemical databases and on-line data sources

| Name | Content | Intended use for identification |
|---|---|---|
| CAS[a] [10] | Abstracts of all articles from chemical and related scientific literature and patents in the CA. Numerous indices. The registry of all known compounds/substances (>110 million). Now available as CA on CD-ROM, or accessible on-line | Finding whether the compound is known and where it originated and is used. Different searches of compounds and their properties. Estimation of occurrence and co-occurrence rates |
| The Beilstein Database[a] [11] | Numerous reference data in organic chemistry. Covers 10 million organic compounds, 10.5 million reactions, and 320 million items of experimental property data | Finding whether the compound is known and where it originated and is used. Different searches of compounds |
| The Combined Chemical Dictionary[a] [12, 13] | Over 225,000 entries containing chemical, physical, and structural data, uses, including biological use and sources on more than 570,000 compounds. Combines Dictionaries of Analytical Reagents (16,000 compounds), Carbohydrates (29,000 compounds), Drugs (48,000 compounds), Inorganic and Organometallic Compounds (106,000 compounds), Natural Products (207,600 compounds) and Organic Compounds (286,500 compounds) | Finding whether the compound is known and where it originated and is used. Different searches of compounds and their properties, including the search for accurate monoisotopic mass |
| The Merck Index[a] [14] | About 4,000 of the entries covering drugs and pharmaceuticals, 2,000 describing organic chemicals and reagents, 2,000 covering natural substances, 1,000 focusing on the elements and inorganic chemicals, and approximately 1,000 covering agricultural substances. Numerous features and properties | Finding whether the (bio) compound is known and abundant, and where it originated and is used. Different searches of compounds |
| KEGG [15] | 16,220 Metabolites and other small molecules (KEGG COMPOUND), 9,454 drugs (KEGG DRUG), 8,169 biochemical reactions (KEGG REACTION), and so on | Finding whether the analyte is biocompound /metabolite/drug |
| NIST Chemistry WebBook [16] | Chemical and physical property data on over 40,000 compounds | Spectra and retention indices |
| PubChem [17] | Data on the biological activities of small molecule compounds. More than 26 million unique structures[b] | Searches for formula, molecular mass and so on. Searches of related structures |

| | | |
|---|---|---|
| ChemSpider [18] | Chemical structures[b] with many properties | Different searches of compounds, including search for accurate monoisotopic mass. Searches of isomers and similar structures |
| CHEMCATS[a] [19] | Combined catalogs. More than 40 million purchasable compounds[c]. More than 1,000 suppliers of chemicals | Purchase of pure compounds for recording reference data and co-analysis |
| ZINC [20] | Combined catalogs. More than 13 million purchasable compounds[c] "for virtual screening." Numerous suppliers of chemicals | Purchase of pure compounds for recording reference data and co-analysis |
| ChemIDplus [21] | Properties of over 370,000 chemicals | Search of features and properties of compounds |
| Internet [22, 23] | Various items of information on many chemical compounds. Often scattered, deficient, disordered, and hidden data | Search of features and properties of compounds. Estimation of their occurrence rates. |
| GMELIN97[a] [24] | Numerous (and not updated) reference data in inorganic and organometallic chemistry | Search of features and properties of inorganic and organometallic compounds |
| SureChem [25] | Gateway for chemical patent search on full text databases of USPTO, EPO, and WIPO. More than 9 million unique compounds | Finding where the compound is used. |

[a]Commercial information source
[b]Some of them seem to be virtual
[c]The number of unique compounds is substantially smaller

1. In the database of the most known compounds such as the Dictionary of Organic Compounds and related Dictionaries (Table 6.2).
2. Very frequently in the large chemical/general data system containing multiple records/entries for any abundant compound, such as CAS or even the Internet as a whole (see Table 6.2 and Sects. 6.4 and 6.5).
3. In several chemical/spectral databases (see Sect. 6.6).

Abundant and rare compounds are differed by their occurrence rates. There are common distributions of chemical entities over that rate. Two such distributions are shown in Fig. 6.2. In the data sources, the great majority of compounds occur only once! They are just rare or relatively rare substances.

The set of abundant compounds is not so large, but they occur in the literature far more frequently than rare ones. Indeed, the example of the sample from CA

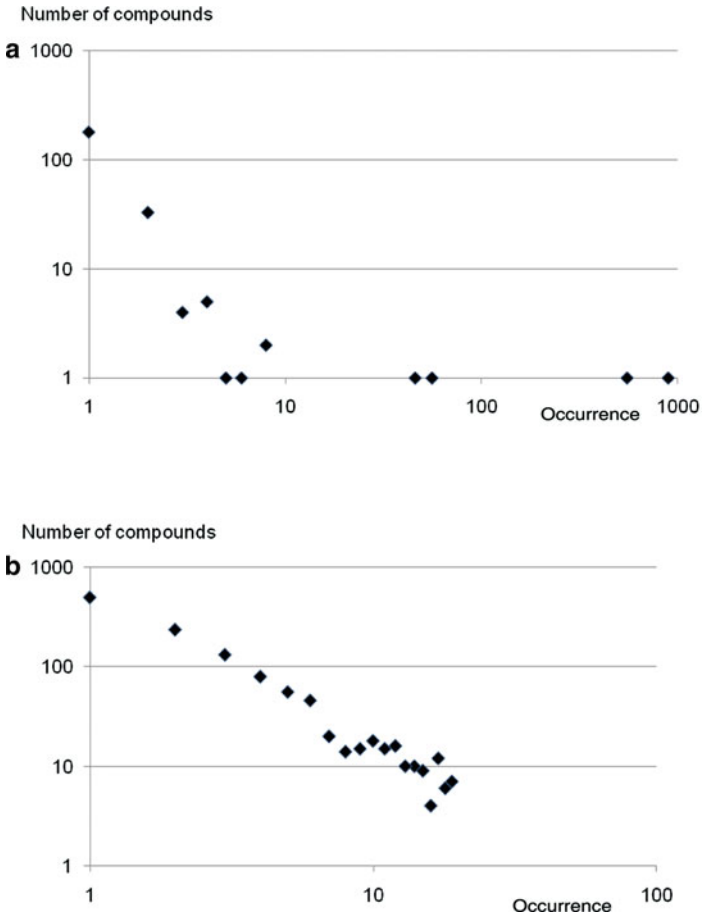



**Fig. 6.2** Distribution of unique chemical compounds over their occurrences in the database or reports/lists. (**a**) The random sample of 300 compounds from CA up to 2003 (for sampling, see [3]). (**b**) Reports on organic analysis of water for 1970–1976 containing 1,258 compounds in 175 lists [28]. The most frequently occurring compounds are not shown

(Fig. 6.2a) demonstrates that only 1.3% (more exactly, not larger than 2.7% with the probability 0.95) of compounds occur at least ten times, but these occur cumulatively in 83% of cases [29]! Taking ten citations as the threshold for the abundance, the overall number of such compounds can be estimated as $50 \times 10^6 \times 0.013 = 650{,}000$, where $50 \times 10^6$ is the number of known low molecules (Sect. 1.5.4). It is appropriate to compare this value to sizes of other sets of popular compounds.

- More than 248,000 inventoried/regulated substances covered by the CHEMLIST data base [30].
- More than 570,000 compounds are included in the last version on Dictionaries; see Table 6.2.
- The exact number of commercially available unique chemicals is unknown, but certainly measured in millions (see Table 6.2).

So the upper limit for the number of widely occurring low-molecule substances as potential analytes and candidates for identification in common analytical problems seems to be within the order-of-magnitude range of $10^5$–$10^6$.

It has been proved that there is positive correlation between the occurrence rate of chemical compounds and their presence in the sample to be analyzed [1–3], see below. So the chance of detecting an abundant compound in a corresponding matrix is relatively high, and rare compounds can be often disregarded in setting up identification hypotheses.

In relation to high molecules, related statistics seem to be less clear and more misleading. In May 2010, CAS registered 61,885,559 sequences [31]. There are nucleic acids, proteins, possibly polysaccharides, and also partial nucleotide (genes) and amino acid sequences [32, 33], i.e., fragments of molecules. Furthermore, this data system contains not only real compounds, i.e., ones having experimentally proved formula and structure, but also virtual/predicted molecules, e.g., "sequences deduced from gene translations" [32]. Certainly, a chemist should not refer possible rather than real-world structures and molecular pieces to molecules of known individual chemical compounds!

Some other databases on sequences make it possible to estimate the number of known high-molecule compounds more precisely. For example, in June 2010 one of the main protein data banks, UniProtKB/Swiss-Prot, contained 517,100 sequence entries [34]. Only for 69,384 sequences (13.4%) was "evidence at protein level" obtained which

> . . . indicates that there is clear experimental evidence for the existence of the protein. The criteria include partial or complete Edman sequencing, clear identification by mass spectrometry, X-ray or NMR structure, good quality protein–protein interaction or detection of the protein by antibodies [35].

The largest fraction of proteins, 363,688 or 70.8%, was "inferred from homology," i.e.,

> . . . the existence of a protein is probable because clear orthologs [some kind of analogy – author] exist in closely related species [35].

Therefore, it is reasonable to suppose that the existence of only a minority of the predicted high molecules have been proved. The set of $10 \times 10^6$ compounds (16% of registered sequences, see above) would be the reasonable conventional estimate for the overall number of actually known high molecules (any compounds, rather than only abundant ones).

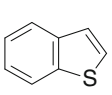
## 6.4 Occurrence and Co-Occurrence Rates

### *6.4.1 Kinds of rates*

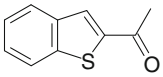
The abundance of analytes is estimated by occurrence and co-occurrence rates which are the absolute or relative number of corresponding records (entries) in databases (Table 6.3). The first of these rates is a direct measure of abundance of chemical compounds, which is evaluated by simple counting. In evaluating co-occurrences, the information is retrieved for a pair of words which are the names of (1) two compounds (one of them may be the known component of the sample), or (2) a compound and a matrix; the two words must present in the same record (Table 6.3).

### *6.4.2 Databases*

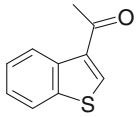
The rates were mainly evaluated for the Chemical Abstract database, including its printed version and the recent version on CD [36]. The abstract extracted from CA is shown in Fig. 6.3. For a simple occurrence rate, documents are retrieved for names of compounds in queries. The number of documents returned is just the value of the rate. For complicated or multiple names, unambiguous RN should be requested. Recently, searches in both CA and PubChem (see Table 6.1) for molecular formulas were also carried out; corresponding occurrence rates belong to groups of isomers combined by the same formulas [37]. Different databases consisting of multiple records per an abundant compound or compound group can be also used in statistical calculations of occurrence and co-occurrence rates. In general, a search of the Internet seems to be the simple way (e.g., Table 6.3), but it should be further tested. The preliminary research shows that only semiquantitative correlation of occurrences is observed between rates estimated by CA and the Internet (Fig. 6.4). In any case, corresponding values differ between the two data systems (see also Table 6.3). In information analysis, one should compare rates of different chemical entities which are obtained using the same data source for the same time periods.




**6.1**




**6.2**




**6.3**

**Table 6.3** Occurrence and Co-Occurrence Rates (adapted from [2])

| Rate | Definition, explanation | Example |
|---|---|---|
| Occurrence, individual compound | The number of different documents (papers, patents, notes, etc.) recording the particular compound. Evaluated as the number of (1) abstracts in abstract database/journal, (2) entries in index to this database, or (3) records in other databases counted for the compound | Information retrieval in the Chemical Substance Index (CSI) to CA for January to June, 1987 resulted in 403 entries for *cyclohexane*. The analogous search in 2009 using the Google Scholar engine [23] returned 8,510 articles[a] with *cyclohexane*. So, the occurrence rates of the compound estimated for different data sets in different ways are 403 and 8,510, respectively |
| Occurrence, compound group | The number of different documents (papers, patents, notes, etc.) recording the compound group. Evaluated as the number of (1) abstracts in abstract database/journal, (2) entries in index to this database, or (3) records in other databases counted for the compound group | Searches in CA on CD for 2008 resulted in 271 documents for $C_{22}H_{17}N_3O_5$, i.e., the occurrence rate of the group of isomers with that formula for the period is 271. The analogous search in Pubchem [17] returned 764 records for that molecular formula |
| Co-occurrence, compound-to-compound | The number of different documents mutually recording the two compounds. Evaluated as the number of (1) abstracts in abstract database/journal, (2) entries in index to this database, or (3) records in other databases counted for the pair of compounds | In that issue of CSI (see above), there were 113 different abstracts belonging to both *hexane* and *cyclohexane*, as indicated by the same abstract numbers. Hence, the *cyclohexane-to-hexane co-occurrence* rate was 113 in January–June, 1987. This rate evaluated using Google Scholar (see above) was 3,750 |
| Relative co-occurrence, compound-to-compound | Compound-to-compound co-occurrence in reference to the occurrence rate of one of the compound pair | The relative cyclohexane-to-hexane co-occurrences for cyclohexane are 113/403 and 3,750/8,510 (403 and 8,510 are occurrence rates, see above), i.e., 0.28 and 0.44, respectively |
| Co-occurrence, compound-to-compound group | The sum of individual co-occurrence of the compound with every member of the group of compounds | Literature co-occurrence of *9-methylanthracene* and several *PAH* detected in waste gas is calculated. The count of entries in CSI to CA for January to June, 1995 led to the co-occurrence of this anthracene derivative as follows: nine, with anthracene; five, with pyrene, and so on. The sum of individual contributions (27) is the *9-methylanthracene-to-PAH* rate |

(*continued*)

**Table 6.3** (continued)

| Rate | Definition, explanation | Example |
|---|---|---|
| Relative co-occurrence, compound-to-compound group | Compound-to-compound group co-occurrence in reference to occurrence of an individual compound | The relative *9-methylanthracene-to-PAH* co-occurrence is 27 divided by 15 (occurrence rate), i.e., 1.8 |
| Co-occurrence, compound-to-matrix group | The number of different documents mutually recording names of the compound and the matrix. Evaluated as the number of abstracts in abstract database/journal or records in other databases counted for the pair of the compound and the matrix. This rate can be separately calculated for only articles in analytical chemistry, namely for determination of the particular compound in the matrix. | For *cyclohexane*, CSI to CA for the first half of 1987 reveals three different abstracts associated with determination of this hydrocarbon in *air* (one of matrices of environmental type). Hence, the rate of *cyclohexane-to-air* co-occurrence is three for the time range. In another search in 2009 (Google Scholar, see above), there were 3,810 articles with the words *cyclohexane* and *environmental*; that value is the *cyclohexane-to-environmental* co-occurrence |
| Relative co-occurrence, compound-to-matrix group | Compound-to-matrix co-occurrence in reference to the occurrence of a compound. | This indicator estimated for *cyclohexane* and *air* matrix for the first half of 1987 is 3/403 = 0.0074, where 403 is the occurrence. The *cyclohexane-to-environmental* co-occurrence rate in 2009 is 3,810/8,500 = 0.45 |

[a]Including cyclohexane derivatives

**Title**

Acyl derivatives of thianaphthene

**Author**

Royer, Rene; Demerseman, Pierre; Cheutin, Andree

**Organization**

Inst. Radium, Paris

**Publication Source**

Bulletin de la Societe Chimique de France (**1961**) 1534-42

…

**Abstract**

Acetylation of 0.25 mole **thianaphthene** (I) with 0.25 mole AcCl in 300 cc. C6H6 and 0.25 mole SnCl4 and distn. of the product, b14 170.5°, gave a mixt. of **2-acetyl deriv.** (II) of I, m. 45°, and largely **3-acetyl deriv.** (III) of I, m.64°,sepd. by repeated fractional crystn. from EtOH...

…

**Accession Number**

1962:45923 CAPLUS

…

**Fig. 6.3** The fragment of typical CA abstract recording names of thianaphthene **6.1** and its derivatives, first 2-acetyl- **6.2** and 3-acetyl thianaphthene **6.3** (names in **bold**). At least, the three compounds are present in the same reaction mixture, i.e., they are components of the same system. The abstract expresses the unit occurrence of compounds **6.1**–**6.3** and co-occurrence of any pair from them. Counting abstracts (entries) results in overall rates

### *6.4.3 The Co-Occurrence Rate*

The co-occurrence rate is more complicated in interpretation than the simple occurrence. There are two general factors: (a) a similarity in different features/properties, and (b) a presence in the same sample/system, causing the existence of names of chemical compounds within the same database record (Table 6.4). For example, the reaction mixture described in the older article (Fig. 6.3) had contained thianaphthene **6.1** and two of its derivatives **6.2** and **6.3**. The reason of their presence in the article and the corresponding abstract was *the same reaction system* (Table 6.4).

It has been reasonably proposed that analytes which occurred in the same samples under analysis also have high co-occurrence rates in literature/databases [1–3]. Indeed, any factor noted in Table 6.4 can lead to facts of real analyte co-occurrence in samples. It can be exemplified by the following cases.

- Similarity in such properties as solubility in water may result in mutual presence of many industrial chemicals in significant amounts in the same samples of waste water.
- Closeness in boiling points causes corresponding impurities in chemicals purified by distillation/rectification.

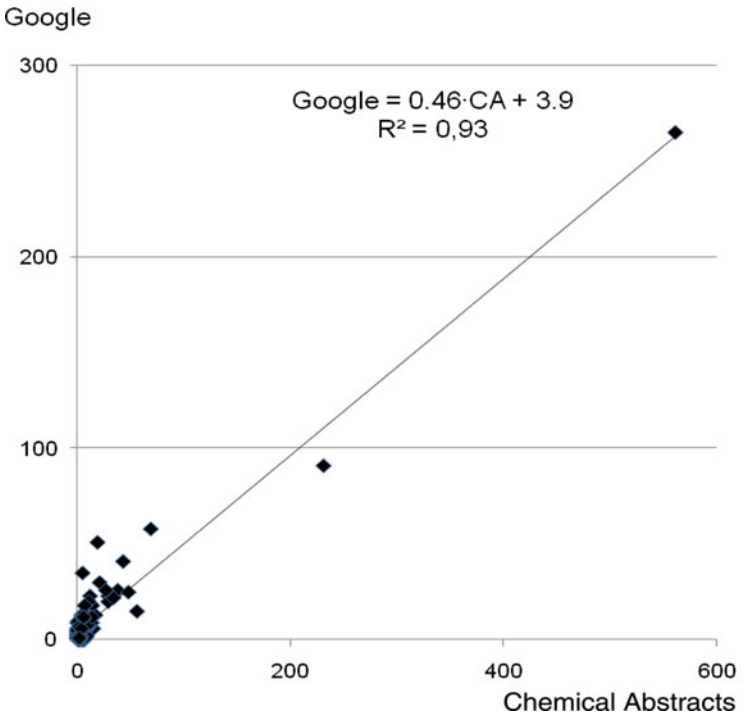



**Fig. 6.4** The correlation between Chemical Abstracts and the Internet in the number of occurrence of different molecular formulas. The data for the artificial sample consisting of 114 formulas (25 pesticides and compounds most similar in molecular mass to the former) are shown. The pesticides were randomly sampled from their list [38]. The formulas of other compounds were generated using the NIST Formula Generator of the NIST MS Search 2.0 software [39]. The occurrence rates for all the molecular formulas were evaluated by a search (1) in the CA edition on CD for 2007 (see [36)], and (2) using the Google engine [22]. Formulas without citation in both information sources are excluded. In the case where widely occurring formulas (>100 times) are also excluded, correlation between the two rates becomes worse ($R^2 = 0.60$)

- PAH formation in the same combustion reactions leads to mixtures of these hydrocarbons being detected in environmental samples (e.g., see [40].
- Synthetic precursors and decomposition products of chemicals are their native impurities.

A co-occurrence rate can also be calculated for a word pair denoting a chemical compound and matrix in which that compound may be present. Different key words for various matrixes suitable for information searches are listed in Table 6.5. The particular analyte often presenting in the particular kind of samples also seems to have a high co-occurrence rate of compound-to-matrix type [1–3] (Table 6.3). The reverse statement implies that the rate, if high enough, can be taken into account to set up the hypothesis on the presence of the compound in a sample of the corresponding sort (see below).

**Table 6.4** Reasons of co-occurrence [3]

| Factor | % | Examples |
|---|---|---|
| Similarity in properties, activity, structure, use | 58 | PAH, PCB, PCDD/F; drugs/pharmaceuticals of the same class |
| The same reaction system | 13 | Reactants and products belonging to the same reaction |
| The same mixture/solution/matrix/sample | 17 | The same formulation; impurities of the same product; pollutions in the same water sample |
| Other/hidden | 12 | |

**Table 6.5** Different matrixes and related key words

| Matrix type | Word |
|---|---|
| Biomedical | Bacteria, bile, blood, breath, saliva, tissue, urine |
| Coal | Coal, coal gases, coal tar |
| Environmental | Aerosols, air, environmental, dust, gases, sediments, soil, water |
| Petroleum | Gasoline, naphtha, oil, petroleum, petroleum products |
| Pharmaceutical | Capsules, dosage forms, drugs, pharmaceuticals, suppositories, suspensions, tablets, transdermal systems |
| Waste | Combustion gas, exhaust, refuse, smoke, waste |

If the co-occurrence rate is estimated only for cases where chemical systems are analyzed ("analytical entries" in databases), words such as *analysis*, *determination*, *assay*, and so on are further introduced in search queries

### *6.4.4 Methodological Aspect*

Exploration of occurrences in chemical databases [1–3] is methodologically related to citation (occurrence) and co-citation (co-occurrence) analysis of text constituents, i.e., bibliographic references, words/terms, and author names, which is widely used in the science of science, information science, sociology, etc. to explore intellectual and social structure of science [41–48]. Co-word networks created by text mining in databases have also been used for the generation of advanced hypotheses and the discovery of new relationships between phenomena in biomedicine [49–52]. In those researches, some words belonged to names of chemical compounds depicting research specialties [45–47] or having biological activity [49–51]. The MEDLINE database is another information source [49–52] which can be also used for a generation of identification hypotheses discussed here.

## 6.5 Identification Hypotheses and Occurrence/Co-Occurrence Rates

These rates are required for the generation and deletion of identification hypotheses for unknown analytes which are shown in Fig. 6.5. The essence of the approach is rather simple.

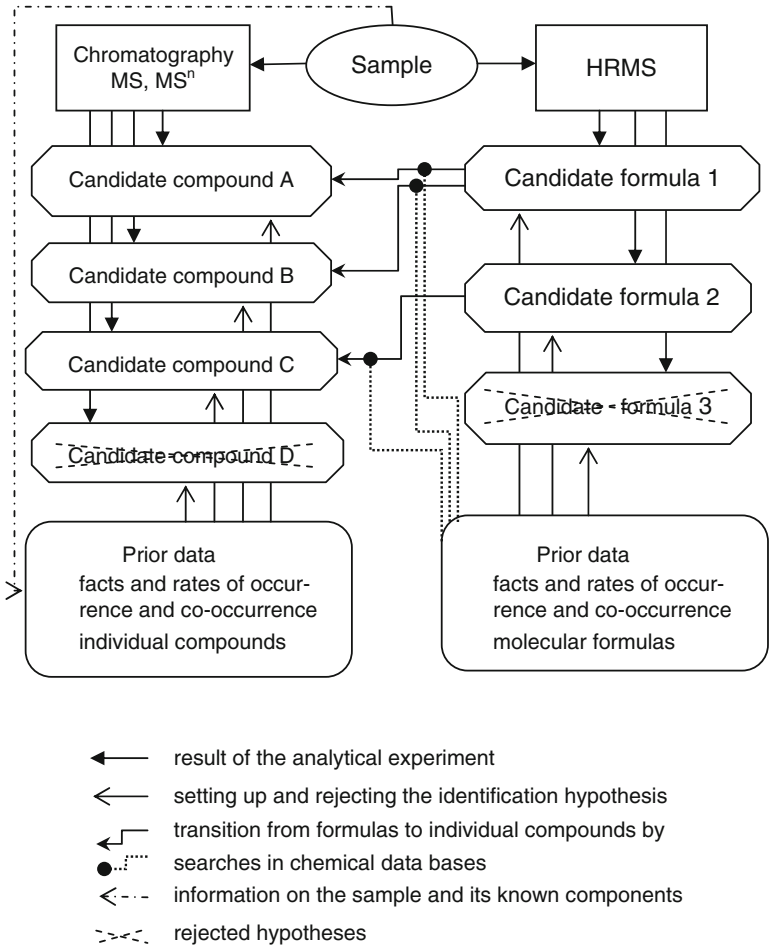



**Fig. 6.5** Schematic of the use of prior data in unknown analysis. Candidate compounds and formulas are generated in analytical experiments by various chromatographic and mass spectral techniques. Prior data, including occurrence and co-occurrence are used (**a**) to set up candidates for identification independently from experimental data, and (**b**) to remove redundant hypotheses related to rare formulas and compounds, i.e., in the cases of no occurrence or a low rate of occurrence. If a sample composition is partly known, co-occurrence with the known sample components may be efficient in generating new hypotheses. Some candidate compounds can be retried in searches in chemical databases for candidate formulas. Definitive identification of candidates; see text in Sect 6.5 and Chap. 7

### *6.5.1 Redundant Hypotheses*

Chromatographic and MS analysis using reference retention data and mass spectra very often does not provide unambiguous identification, i.e., the identification results (candidate compounds) are multiple computer answers per each analyte.

Rare (rarely occurring) compounds, e.g., ones with occurrences which are smaller than 5% of the sum of entries for all the candidates [37], cannot be considered. In the same manner, rare formulas are filtered out. The 95% and 5% occurrences can be considered as rates of TP and FN respectively, derived only from prior data.

### *6.5.2 Deficient Hypotheses*

There are no advanced hypotheses for some analytes. However, some components of the sample are a priori known or established in the initial analytical stages. As a rule, the matrix kind is also clear. Further, the candidate compounds and/or their formulas which co-occurred with known analytes or the matrix itself can be searched. Candidates are tested vs retention values and spectra recorded before or after such operations of setting up hypotheses.

Definitive conclusions about identification are stated when the number of analytes and candidate compounds is consistent, all possible hypotheses are tested, and conditions/criteria of identification are met; see Sect. 7.1.

In the general case, the list of advanced hypotheses to be further tested is formed according to high values of corresponding occurrence and co-occurrence rates. For a long list of hypotheses, they are tested in the order of decreasing rates, starting from the highest ones. In the cases of unit rates, all selected hypotheses are tested.

The methods of setting up candidates for identification under consideration were validated by the comparison of the rates evaluated for (1) groups of analytes, including ones specially identified for the researches, and (2) reference groups of compounds similar to the analytes in RI, mass spectra, or accurate molecular masses (Table 6.6). In almost all tests, mean rates of the first group are higher than in the case of reference compounds. However, not all the differences were statistically significant [$\alpha < 0.05$, (3.15)].

The overall conclusion derived from the researches cited in Table 6.6 is that significance was observed for 2/3 and 3/4 tests of occurrence and co-occurrence rates, respectively. Therefore, in order not to miss advanced hypotheses, there should be some redundancy of candidate compounds. The use of the combination of rates of different type (Table 6.3) may also be efficient in setting up identification hypotheses. The efficiency and productivity of the occurrence and co-occurrence approach in qualitative chemical analysis also depend on the availability of special software for on-line processing records in databases.

Below, two examples of evaluating these rates are given. The first of them (Example 6.1) refers to occurrence rates which are estimated for molecular formulas.

The use of co-occurrence rates is further illustrated for reverse prediction of the impurity composition [3] (Example 6.2). Here advanced identification hypotheses are not efficiently selected by high rates, but consideration of all facts of co-occurrences ensures that no analyte is missed.

The procedures of evaluation and use of statistical rates based on occurrences in chemical databases are somewhat analogous to corresponding intellectual

**Table 6.6** Different groups of analytes and compared compounds. Searches of occurrences and co-occurrences in CA

| Group of analytes | Reference group | References |
|---|---|---|
| Impurities in *n*-hexane | (1) Impurities in naphthalene. (2) Candidates for impurities[a] in *n*-hexane. (3) Candidates for impurities[a] in naphthalene. (4) Pharmaceuticals | [1, 2] |
| Impurities in naphthalene | (1) Impurities in *n*-hexane. (2) Candidates for impurities[a] in *n*-hexane. (3) Candidates for impurities[a] in naphthalene. (4) Pharmaceuticals | [1, 2] |
| PAH, unambiguously identified | (1) PAH, ambiguously identified. (2) PAH and their isomers, candidates for identification[a]. PAH and their isomers, without reference GC and MS data | [2] |
| Impurities in 17 chemical/ pharmaceutical products | (1) Random sample from known compounds. (2) Candidates for impurities[b]. (3) The same group, without rare compounds | [3] |
| Impurities in three chemical/ pharmaceutical products | (1, 2) Candidates for impurities[b], two groups. (3,4) The same groups, without rare compounds | [3] |
| 25 pesticides | Candidates for identification[c] | [53] |
| 25 pesticides and their known isomers[d] | Candidates for identification[c] | [53] |
| 18 pesticides | Candidates for identification[c] | [37] |
| 18 pesticides and their known isomers[d,e] | Candidates for identification[c] | [37] |

[a]Compounds with similar RI and mass spectra
[b]Compound co-occurred with products but co-occurring impurities
[c]Compounds with similar molecular masses
[d]Compounds with the same molecular masses
[e]Search in both CA and PubChem (see Table 6.2)

operations, reasoning, and discourse, and so on, specific for experienced researchers. Indeed, when solving identification problems, chemists traditionally tend to consider (a) widely occurring compounds and (b) compounds similar to already detected ones in some or other features. This approach, proposed in reports [1–3] and described in this section, and partly simulating intellectual activity of a scientist, gains advantage because it

- Is free from subjectivism and specialization inherent to any chemist as well as any expert
- Uses huge data sets
- Requires little personal skill to initially process database information

**Example 6.1**
The new $MS^2$ library (see Chap. 7), TaMaSA, supplemented with HRMS and searches in chemical databases, was examined for screening/identification of organic compounds, as exemplified by pesticides [37]. Occurrence rates obtained by the search in Pubchem [17] were used for filtering out rare formulas/compounds. The model data set consisted of formulas of 18 pesticides (model analytes from real-world samples) and 153 candidate formulas

*(continued)*

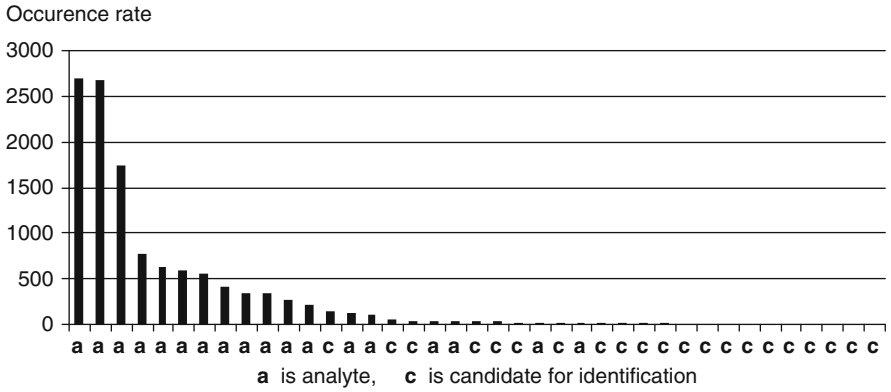



**Fig. 6.6** Occurrence rates for formulas of 18 pesticides (analytes, *a*) and compounds (candidates, *c*) similar to analytes in molecular mass (differences <5 ppm). The pesticides are azoxystrobin, carbendazim, carbofuran, carboxin, chlormequat, chlorsulfuron, cloquintocet-mexyl, ethaboxam, fenhexamid, fenoxaprop, flumioxazin, glyphosate, imidacloprid, ipconazole, iprodione, iprovalicarb, metsulfuron-methyl, and pyriproxyfen. Candidate formulas were generated using the software of the LIT-Orbitrap instrument (LTQ Orbitrap XL, Thermo, USA). Occurrences were searched in the Pubchem database [17]. Only rates $\geq 2$ are shown here

(see Table 6.6). Figure 6.6 shows that the former occurred far more frequently than candidates for analytes similar in molecular mass. Mean occurrence rates were approximately 643 and three in the two groups respectively, which is a very significant difference ($t$-test, $\alpha = 0.0000$). The proportions of overall occurrences were 94 and 6%, respectively. The threshold of 5% of the overall rate (the conditional limit for rare chemical entities) was not exceeded by the vast majority of candidate formulas (148 from 153), i.e., about 97% from them were filtered out. On the other hand, 12 from 18 pesticide formulas exceed the threshold of 95%, i.e., here other candidate formulas cannot be taken into account. Statistics for occurrences of pesticides themselves and their isomers covered by pesticide formulas will be reported below (Sect. 7.4.2). To search individual compounds corresponding to known formulas, CAS databases are suitable.

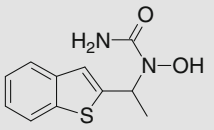
**Example 6.2**
2-Acetylbenzothiophene (2-acetyl thianaphthene) **6.2**, the material for the synthesis of drug zileuton **6.4**, contains nine impurities **6.5**–**6.13** [54] which resemble **6.2** in structure: the common substructure is the benzene ring bonded to the sulfur atom. One of the reasons leading to relatively frequent co-occurrence of compounds is their structural similarity observed here for the main component **6.2** and those impurities. Therefore it is no wonder that

*(continued)*

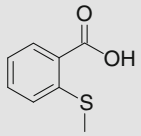
**Table 6.7** Rejected identification hypotheses

| Feature | Technique for testing hypotheses | Rejections |
|---|---|---|
| Molecular mass | MS, common resolution | 29 |
| | MS, high resolution | 3 |
| Molecular mass, isotope pattern | MS, common resolution | 4 |

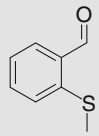
compounds **6.5**–**6.13** occurred at least one time together with **6.2** in the chemical literature, as found out by searches in CA for 1964–2003 [3]. However, there are 36 other compounds which co-occurred with **6.2** in corresponding abstracts. From them, the compounds **6.14**–**6.19** more frequently co-occurred with the compound **6.2**. Here, both compound groups, nine impurities which were detected and identified, and 36 candidate compounds are hardly differentiated by occurrence or (relative) co-occurrence rates, because mean values in the two groups are comparable. Let us suppose the case that the impurities are unknown. It is evident that all the hypotheses which related to 45 (9 + 36) candidate compounds should be tested. The MS method is very efficient for testing the hypotheses. In most cases, integer molecular mass measured by low-resolution mass spectrometry makes it possible to accept/reject identification hypotheses. To differentiate impurity **6.7** from candidates for identification of **6.20**–**6.22** having the same integer molecular mass, high-resolution mass spectrometry for measuring different accurate mases is necessary; monoisotopic mass: 232.06 (**6.7**), 232.01 (**6.20**), 232.09 (**6.21** and **6.22**). Four candidate chlorine/bromine compounds can be also removed because of isotope patterns different from those of impurity molecules. Statistics of rejected identification hypotheses are given in Table 6.7.



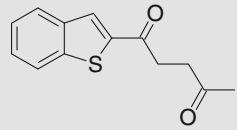

**6.4**



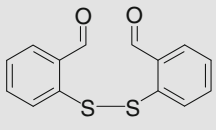

**6.5**



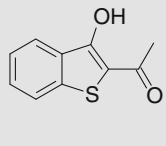

**6.6**



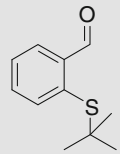

**6.7**



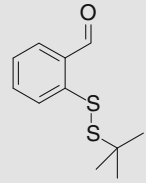

**6.8**




**6.9**

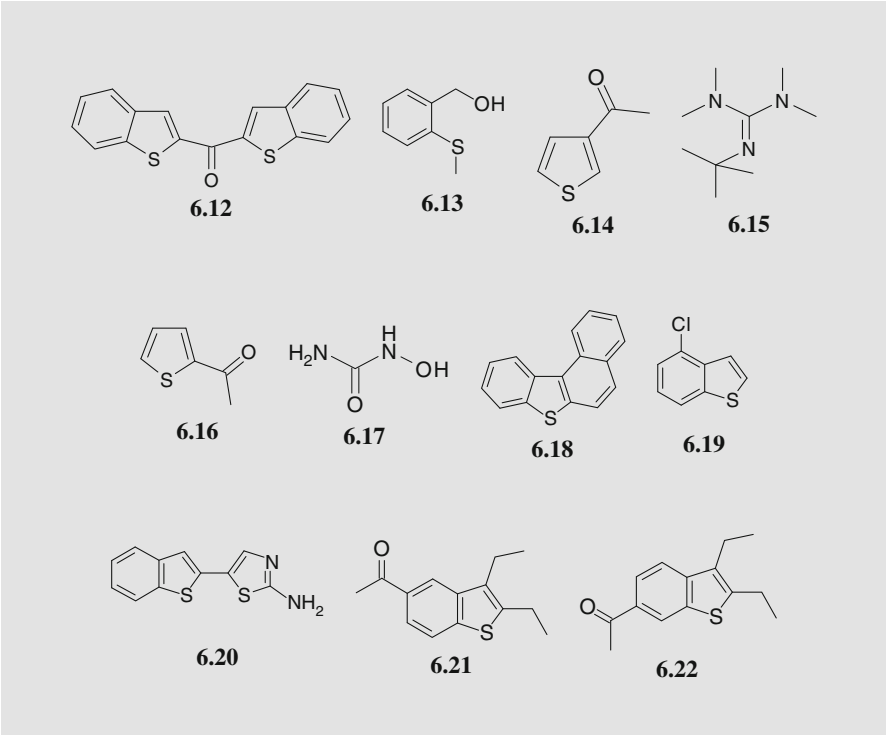



**6.10**




**6.11**

*(continued)*




6.12 6.13 6.14 6.15

6.16 6.17 6.18 6.19

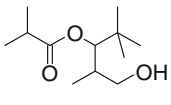
6.20 6.21 6.22

## 6.6 Prior Data Involved in Analytical Procedures

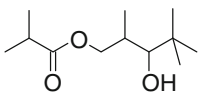
Prior data are or may be included in identification procedures in different ways. Using statistics of occurrence rates (see above) is one of the approaches. Related or different ones are considered in this section.

### *6.6.1 Searches in Databases*

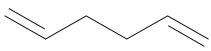
Common information retrieval in chemical databases can be used for identification purposes. An example is the analysis of the indoor air by GC–MS carried out by the author; some components proved to be hard to recognize. For certain identification, different volatile compounds detected in air before that analysis were retrieved in CA using a query containing *analysis* and *air*. Two esters of isobutyric acid **6.23** and **6.24**, possibly metabolites from microorganisms, were suitable findings [55, 56] fitting to mass spectra of two unknowns.




6.23




6.24



6.25

### 6.6.2 Penalty for Rare Compounds

Using NIST MS library searching program [39], an analyst can penalize rare compounds. It means that MF for spectra of these compounds will be reduced up to 50 out of 1,000 units. A compound is classified as rare if contained in only a few chemical databases. The penalty value depends on the number of such databases. This approach to the consideration of a prior probability is exemplified for mass spectrum of 1,5-hexadiene **6.25** as the unknown analyte (Table **6.8**). The software "concluded" that all candidate compounds, but 1,5-hexadiene itself and cyclohexane are rare ones (see Table **6.8**). Corresponding MF were properly decreased, leading to a somewhat higher probability of identification (see Sect. 4.5.4.2) of three candidate compounds (see Table **6.8**). Compounds determined as rare ones could be excluded from further testing of identification hypotheses, i.e., from confirmation of an identity screened by library searches.

### 6.6.3 Information About the Sample

In complicated identification problems, the ISO standard [57] (see Sect. 5.5.3.2) calls for

> gathering additional identification points using knowledge and interpretation of this knowledge about the sample or sampling site.

**Table 6.8** Comparison in library searches[a]

| Name | Regular search | | | Penalizing rare compounds[b] | | |
|---|---|---|---|---|---|---|
| | No. | MF (Match) | *P* (Prob.) | No. | MF (Match) | *P* (Prob.) |
| 1,5-Hexadiene | 1 | 944 | 79.4 | 1 | 944 | ↑ 80.9 |
| 1,5-Hexadiene | 2 | 917 | 79.4 | 2 | 917 | ↑ 80.9 |
| Bicyclo[3.1.0]hexane | 3 | 888 | 4.95 | 4 | ↓ 838 | ↓ 1.80 |
| Cyclopropane, 1-propenyl- | 4 | 887 | 4.76 | 5 | ↓ 837 | ↓1.73 |
| 1,1'-Bicyclopropyl | 5 | 865 | 1.88 | 6 | ↓ 835 | ↓ 1.59 |
| Cyclopropane, 1,2-dimethyl-3-methylene- | 6 | 864 | 1.80 | 18 | ↓ 814 | ↓ 0.70 |
| 1,1'-Bicyclopropyl | 7 | 846 | 1.88 | 16 | ↓ 816 | ↓ 1.59 |
| 1,4-Pentadiene, 2-methyl- | 8 | 841 | 0.66 | 9 | ↓ 831 | ↑ 1.35 |
| Cyclohexene | 9 | 841 | 0.66 | 3 | 841 | ↑ 2.03 |
| Cyclopropene-1,3,3-trimethyl | 10 | 834 | 0.50 | 33 | ↓784 | ↓ 0.23 |

[a]The NIST MS Search 2.0 program and NIST'05 MS library [39]

[b]The symbols ↑ and ↓ depict the increase and decrease respectively of the MF and probability

The kinds of such information are as follows.

- The component is identified in earlier samples from the same site...
- From historical investigation, it was shown that presence of the component was expected.
- Other samples from the same site give positive identification.

In [57], this information is considered as significant as direct experimental evidence (obtained by GC–MS). Our point of view is that any prior data is rather a prompt for an analyst, a source for hypotheses. The standard [57] also takes into account this judgment.

> Strictly taken, an identification point obtained in Step 3 [gathering additional information, see Fig. 5.2 – author] is of another order than the identification points obtained in Steps 1 and 2 [gathering identification points using analytical procedures – author].

The value of the sample information is noted for pesticide determination.

> The need for confirmatory tests may depend upon the type of sample or its known history. In some crops or commodities, certain residues are frequently found. For a series of samples of similar origin, which contain residues of the same pesticide, it may be sufficient to confirm the identity of residues in a small proportion of the samples selected randomly. Similarly, when it is known that a particular pesticide has been applied to the sample material, there may be little need for confirmation of identity, although a number of randomly selected results should be confirmed [58].

Circumstances accompanying sampling which are properly documented should be taken into account in the arrangement of analytical operations. Toxicologists provided the instance related to the medication digoxin.

> In practice, the extent and nature of methods used to "confirm" the presence of a particular analyte will depend in part on the type of case and nature of the analyte. A "holistic" approach is required. For example, in a well-documented suicide where a note is found with an empty container of digoxin that was prescribed to that person, an appropriately validated RIA [radioimmunoassay – author] for digoxin may be all that is required. However, a digoxin-related death where there was no suspicion of suicide and where the medication was not prescribed to that individual may require much more extensive testing, including LC/MS [59].

### *6.6.4 Plausibility of Analytical Results*

Critical considerations of results of chemical analysis (may) require comprehensive data about the practical use of chemical products.

> The use of certain compounds for a particular crop grown in a given environment requires some plausibility. Accordingly, it is difficult to explain e.g., high concentrations of a long-forbidden herbicide in plants where an application makes no sense or large amounts of old outdated insecticides such as dicrotophos in produce from countries with high agricultural standards, e.g., peppers from Holland [60].

## References


1. Milman BL, Kovrizhnych MA (2000) Identification of chemical substances by testing and screening of hypotheses II. Determination of impurities in n-hexane and naphthalene Fresenius. J Anal Chem 367:629–634
2. Milman BL (2002) A Procedure for decreasing uncertainty in the identification of chemical compounds based on their literature citation and cocitation. Two case studies. Anal Chem 74:1484–1492
3. Milman BL (2005) Literature-based generation of hypotheses on chemical composition using database co-occurrence of chemical compounds. J Chem Inf Model 45:1153–1158
4. Milman BL (2005) Identification of chemical compounds. Trends Anal Chem 24:493–508
5. Milman BL, Konopelko LA (2000) Identification of chemical substances by testing and screening of hypotheses I. General. Fresenius J Anal Chem 367:621–628
6. Anari MR, Baillie TA (2005) Bridging cheminformatic metabolite prediction and tandem mass spectrometry. Drug Discov Today 10:711–717
7. Baranczewski P, Stańczak A, Kautiainen A, Sandin P, Edlund PO (2006) Introduction to early in vitro identification of metabolites of new chemical entities in drug discovery and development. Pharmacol Rep 58:341–352
8. Staack RF, Hopfgartner G (2007) New analytical strategies in studying drug metabolism. Anal Bioanal Chem 388:1365–1380
9. Roger S, Scheltema RA, Girolami M, Breitling R (2009) Probabilistic assignment of formulas to mass peaks in metabolomics experiments. Bioinformatics 25:512–518
10. Chemical Abstracts Service. http://www.cas.org/expertise/cascontent/index.html. Accessed 23 May 2010
11. CrossFire Beilstein. http://www.info.crossfirebeilstein.com. Accessed 30 Oct 2010
12. The Combined Chemical Dictionary on DVD. http://www.crcpress.com/product/isbn/9780412820205. Accessed 29 Oct 2010
13. CHEMnetBASE. http://www.chemnetbase.com. Accessed 23 May 2010
14. The Merck Index. http://www.merckbooks.com/mindex. Accessed 23 May 2010
15. KEGG: Kyoto Encyclopedia of Genes and Genomes. http://www.genome.jp/kegg. Accessed 23 May 2010
16. NIST Chemistry WebBook. http://webbook.nist.gov/chemistry. Accessed 23 May 2010
17. PubChem. http://pubchem.ncbi.nlm.nih.gov. Accessed 6 July 2009
18. ChemSpider. http://www.chemspider.com. Accessed 23 May 2010
19. CHEMCATS. http://www.cas.org/expertise/cascontent/chemcats.html. Accessed 23 May 2010
20. ZINC http://zinc.docking.org. Accessed 23 May 2010
21. ChemIDplus. http://chem.sis.nlm.nih.gov/chemidplus. Accessed 23 May 2010
22. Google http://www.google.com. Accessed 22 March through 03 April 2008
23. Google Scholar. http://scholar.google.com. Accessed 1 Jan 2010
24. GMELIN97 http://www.cas.org/ASSETS/DB25829EA4F94816AB0A152D24863B92/gmelin97.pdf. Accessed 23 May 2010
25. SureChem http://www.surechem.org. Accessed 23 May 2010
26. Chemical databases http://www.google.com/Top/Science/Chemistry/Chemical_Databases. Accessed 23 May 2010
27. Drug databases http://www.drugbank.ca. Accessed 30 Oct 2010
28. Schaeffer DJ, Janardan KG (1980) Abundance of organic compounds in water. Bull Environ Contam Toxicol 24:211–216
29. Milman BL (2008) Introduction to chemical identification (In Russian). VVM, Saint Petersburg
30. CHEMLIST http://www.cas.org/expertise/cascontent/regulated/index.html. Accessed 23 May 2010

31. CAS Registry http://www.cas.org/expertise/cascontent/registry/regsys.html. Accessed 23 May 2010
32. Protein sequences in the CAS Registry file on STN – exact and pattern searching (2004) CAS2052-1104 http://www.stn-international.com/uploads/tx_ptgsarelatedfiles/protseq.pdf. Accessed 23 May 2010
33. CAS Registry: Exact and pattern searching of nucleic acid sequences (2008) CAS2536-1108. http://www.cas.org/ASSETS/4CE1649F453A44E78DC4763702375D92/nucleic.pdf. Accessed 23 May 2010
34. UniProtKB/Swiss-Prot protein knowledgebase release 2010_06 statistics. http://expasy.org/sprot/relnotes/relstat.html. Accessed 24 May 2010
35. Protein existence (2008) http://www.uniprot.org/manual/protein_existence. Accessed 24 May 2010
36. CA Abstracts. http://www.cas.org/products/print/ca/abstracts.html. Accessed 24 May 2010
37. Milman BL, Zhurkovich IK (2009) Tandem mass spectral library of pesticides and its use in identification. Proceedings of the 18th International Mass Spectrometry Conference, Bremen
38. Compendium of Pesticide Common Names. http://www.alanwood.net/pesticides/index.html. Accessed 24 May 2010
39. NIST Mass Spectral Search Program, version 2.0d, and NIST/EPA/NIH Mass Spectral Library (2005)
40. Mastral AM, Callén MS (2000) A review on polycyclic aromatic hydrocarbon (PAH) emissions from energy generation. Environ Sci Technol 34:3051–3057
41. Small H (1973) Co-citation in the scientific literature: a new measure of the relationship between two documents. J Am Soc Inf Sci 24:265–269
42. Small H, Sweeney E (1985) Clustering the Science Citation Index using co-citation I. A comparison of methods. Scientometrics 7:391–409
43. Small H, Sweeney E, Greenlee E (1985) Clustering the Science Citation Index using co-citation II. Mapping science. Scientometrics 8:311–340
44. Milman BL, Gavrilova YA (1993) Analysis of citation and co-citation in chemical engineering. Scientometrics 27:53–74
45. Law J, Bauin S, Courtial JP, Whittaker J (1988) Policy and the mapping of scientific change: a co-word analysis of research into environmental acidification. Scientometrics 14:251–264
46. Peters HPF, Hartmann D, Van Raan AFJ (1988) Monitoring advances in chemical engineering. Informetrics 87(88):175–195
47. Milman BL, Gavrilova YA (1994) Science news in business journals as the source of information on applied and strategic research and science policy (In Russian). Sci Technol Inf 1(7):17–26
48. Wolfram D (2003) Applied informetrics for information retrieval research. Library Unlimited, Westport
49. Smalheiser NR, Swanson DR (1998) Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. Comput Methods Programs Biomed 57:149–153
50. Weeber M, Klein H, Jong-van D, den Berg LTW, Vos R (2001) Using concepts in the literature-based discovery: simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. J Am Soc Inform Sci Technol 52:548–557
51. Wren JD, Bekeredjian R, Stewart JA, Shohet RV, Garner HR (2004) Knowledge discovery by automated identification and ranking of implicit relationships. Bioinformatics 20:389–398
52. Jenssen TK, Öberg LMJ, Andersson ML, Komorowski J. Methods for large-scale mining of networks of human genes. http://www.egeeninc.com/u/vilo/edu/2004-05/Bioinformaatika/Bioinf/Komorowski_siam2001.pdf. Accessed 30 Oct 2010
53. Milman BL (2008) Unpublished data
54. Shada DM, Wong CF, Elrod L, Morley JA, Gay CM (1996) Determination of 1-benzo[*b*] thien-2-ylethanone and related impurities by high performance liquid chromatography. J Pharm Biomed Anal 14:501–510

55. Sunesson AL, Nilsson CA, Andersson B, Blomquist G (1996) Volatile metabolites produced by two fungal species cultivated on building materials. Ann Occup Hyg 40:397–410
56. Zhou S, Ma J, Wang S, Chen Z (1991) Qualitative analysis of organic compounds in enclosed air by gas chromatography/mass spectrometry (In Chinese). Fenxi Huaxue 19:1115–1121. CA (1992) 116:135267
57. ISO Standard 22892 (2006) Soil quality - Guidelines for the identification of target compounds by gas chromatography and mass spectrometry
58. FAO/WHO Codex Alimentarius. Guidelines on the use of mass spectrometry (MS) for identification, confirmation and quantative determination of residues (2005) CAC/GL 56-2005. http://www.codexalimentarius.net/web/standard_list.jsp. Accessed 16 May 2010
59. SOFT/AAFS Forensic Laboratory Guidelines (2006). http://www.soft-tox.org/docs/Guidelines%202006%20Final.pdf. Accessed 17 May 2010
60. Schürmann A, Dvorak V, Crüzer C, Butcher P, Kaufmann A (2009) False-positive liquid chromatography/tandem mass spectrometric confirmation of sebuthylazine residues using the identification points system according to EU directive 2002/657/EC due to a biogenic insecticide in tarragon. Rapid Commun Mass Spectrom 23:1196–1200

# Chapter 7
# Non-target Identification. Chromatography and Spectrometry

**Abstract** The content of this chapter are focused on unknown analysis when a chemist answers the question of what compounds are present in the sample. The true result of identification is provided by at least two independent (orthogonal) methods. The most general approach to the identification of non-targets is based on chromatography mass spectrometry. Gas chromatographic parameters, widely used for identification, are retention indices. To a lesser degree, retention indices are applicable in liquid chromatography. Now, retention parameters are required in proteomics. In mass spectrometry, volatile analytes are preferably identified by means of reference libraries of electron ionization mass spectra. For identification of nonvolatile compounds, libraries of tandem/product mass spectra have been built. Their use is especially effective when combined with high-resolution mass spectrometry which provides candidate molecular formulas. Interpretation of mass spectra is also possible but not widely applied. NMR and IR spectroscopy are comparable to MS in identification potential if there are a relatively large amount of analytes and a simple composition of a sample under analysis. In NMR, algorithms of spectral prediction as well as respective spectral databases have been rapidly developed. Analytical metabolomics and proteomics are individually discussed, with the focus on approaches to identification, identification criteria, the problems arising due to a great complexity of analytes and unavailability of analytical standards, and interlaboratory comparisons. For all the techniques, information about reference spectral libraries/databases is tabled. Quality assurance of identification is widely covered in the chapter.

## 7.1 General

***Definitions.*** *Non-target identification* is a qualitative determination of analytes unknown to a chemist before performing analytical procedures (Sect. 1.5.1). When determining unknown, an analyst answers the questions of what compounds are present in the sample, or what the nature of the sample is.

B.L. Milman, *Chemical Identification and its Quality Assurance*,
DOI 10.1007/978-3-642-15361-7_7, 

There is not just one approach to this sort of chemical analysis, and just *ad hoc* methods are used by necessity. A large number of candidate analytes have some chance of presenting in a sample. Tens of millions of compounds/substances are known to chemists; hundreds of thousands of them are widespread and practically useful (see Sects. 1.5.4 and 6.3). Many compounds resemble each other closely in structure and properties. This circumstance leads to numerous FP results of identification. However, most popular/abundant/common compounds differ in properties, which demands a variety of analytical methods, procedures, and operations. A use of the particular method for determination of a "foreign" analyte, i.e., in the case not fitted for the purpose leads to FN. However, some guidances (see Sect. 5.5) describing common experimental conditions for identification of targets and corresponding identification criteria can be also applied to unknown analysis.

In many various fields of chemical analysis, identification of unknown is very essential. The first example is toxicology, where this type of analysis is called *general unknown analysis* [1]. The number of unknown compounds in toxicological analysis is measured in thousands [2]. This is a considerable but not very large number if compared to the overall number of known substances. So the type of the analysis for toxicology could be named "semi-unknown analysis". Environmental research and technology [3] and processing of wastes [4], as well as food/residue analysis, are other examples of fields which demands identification of unknowns.

In the recent literature, the terms of *non-target analysis* or *non-target screening* have often occurred instead of *unknown analysis*. So those can be treated as synonyms. However, the terms are sometimes differentiated. One can easily find some semantic difference between the terms in the quotation:

> In pesticide residue analysis (PRA), as well as traditional quantitative analysis of target compounds – mainly pesticides in their parent form – there is now remarkable interest in screening pesticides in a comprehensive way, including not only common pesticides but also less common or relatively new pesticides (non-target) or unknown transformation products (unknowns) [5].

So, "more unknown" and "less unknown" compounds can be seen. In this book, these groups of compounds, as well as *unknown analysis* and *non-target analysis*, are not specially differentiated.

***General approaches.*** The most general approach to identify unknowns is based on using chromatography mass spectrometry. The simple chart for individual identification of organic compounds is given in Fig. 7.1 (see also Fig. 2.3). Methods of isolation of analytes from a sample and techniques of subsequent analysis depend on the type of sample (solid, liquid, gas) and analyte (volatile, non-volatile, polar, unpolar, acidic, basic, and so on) and on whether all or only some sample components are of interest for an analyst. If all components are to be determined, they should be transferred into solution(s) without affecting the analytes. For solids, sample disintegration and digestion, followed by inorganic/element analysis, may be required. If high-molecular compounds are determined, size-exclusion

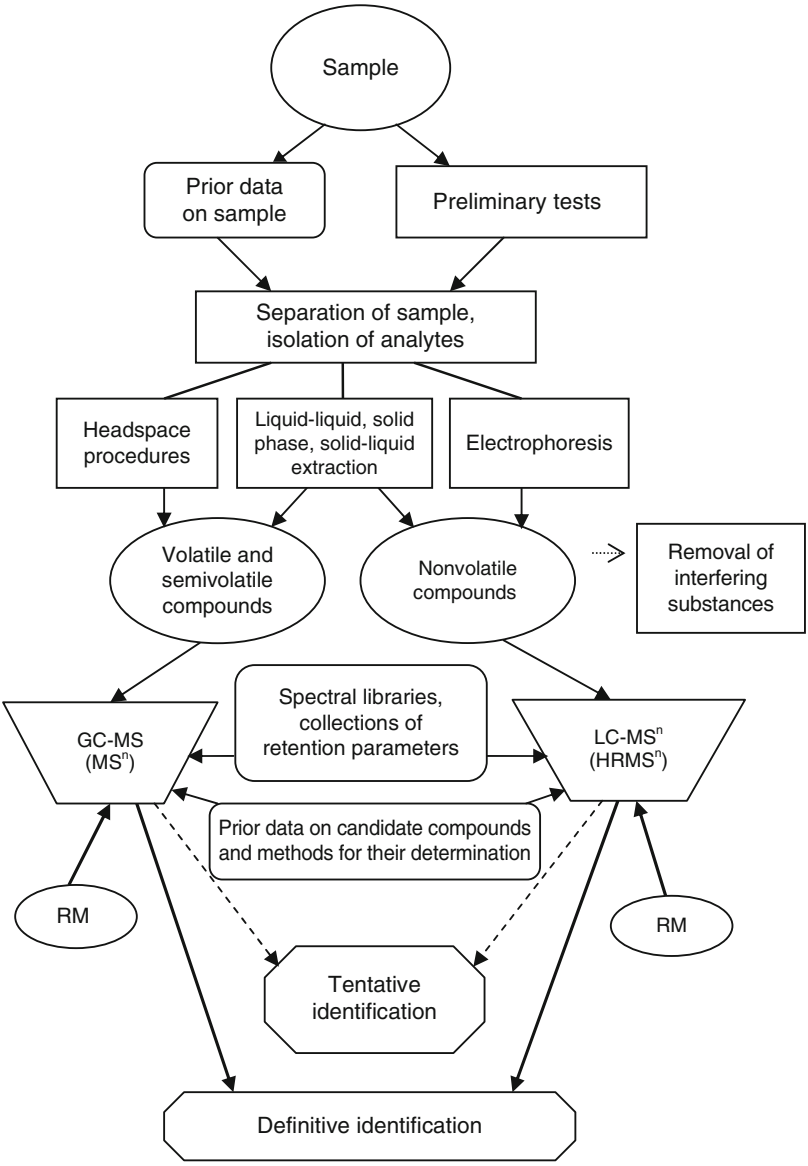



**Fig. 7.1** The flow chart of identification of unknown organic/bioorganic compounds using techniques of chromatography mass spectrometry. To choose the particular separation/isolation methods and analytical techniques, one should search prior information on the sample, visually inspect it, and test it for solubility, burning, and so on. If a sample is an organic substance, it is analyzed further as a vapor or solution in organic solvents. In the case of water samples, non-organic solids/quasi-solids (such as soils and sediments) and biosamples (urine, serum, tissues and so on), unknown organic compounds are extracted from the sample. GC–MS is intended for volatile or relatively volatile compounds; non-volatile analytes are determined using LC–MS techniques. Chromatograms and spectra are recorded in the widest range of conditions: the

chromatography, dialysis, ultra filtration, and so on are used for the isolation of the sample component. All possible methods are also used for separation from those compounds which are unsuitable for the analysis by the particular analytical technique. In a determination of individual compounds having certain properties (partial analysis of a sample), selective extraction of components of interest is applicable.

Qualitative determination of individual[1] volatile compounds is the simplest exercise in identification (see Sect. 2.8.4). Identification of non-volatile compounds is a far more complicated problem (see also Sect. 2.8.4). Here, availability of RM and the use of $HRMS^n$, various databases, and expert systems are of the most value. Identification becomes a yet harder task for high molecules, first proteins, due to an almost total absence of RM (pure compounds). Incidentally, there may be no standards available also for some low-molecular substances, e.g., designer drugs, new prescription drugs, and drug metabolites [6], and emerging pollutants.

The criterion for true identification carried out without analytical standards is that

- The true result is provided by at least two independent (orthogonal) methods fitting the valid determination of the analyte with the given properties.

Properly, at least two techniques are/may be also required when an identity is confirmed with the use of analyte standards for candidate compounds which appeared after tentative identification.

Independent methods are mainly based on techniques where observed analytical signals are due to different physical and chemical processes. GC, LC, and MS can be considered as independent or almost independent techniques. Different methods connected to the same technique may be to some degree independent. In MS, such "partly orthogonal" methods are related to

---

**Fig. 7.1** (continued) range of programming column temperature in GC, the percentage of organic mobile phase for gradient elution in LC, the range of mass numbers in MS. Comparison of mass spectra and retention parameters obtained in analytical experiments with those from spectral libraries and databases of parameters/indices results in identification hypothesis(es) or tentative identification(s) at once. In some cases (e.g., an analyte has unique properties), the latter may be even a definitive answer to the qualitative problem. Nevertheless, additional/prior information is often required to accept/reject identification hypotheses and remove redundant candidate analytes. Furthermore, having a few candidate compounds, one can find validated methods for their true identification, and therefore verify conditions of the analytical experiment and change it if necessary. A clarified list of candidates is further reduced to the final version by incorporating RM in co-analytical experiments, i.e., top quality confirmatory ones

[1]Multi-analyte determination, e.g., in metabolomics (Sects. 7.4.1.3 and 7.7.1), is far more challenging.

- Different ionization (EI vs ESI)
- Different chemical forms of compounds (an analyte and its derivative)
- Algorithms of comparison of experimental and theoretical spectra (proteomics, see Sects. 4.4.2.3, 7.4.1.4, and 7.7.2) and so on

For some analytes, an availability of even partly independent methods may provide reliable identification.

In general, agreement in identification results achieved by several methods implies a definitive identification, and does not demand confirmation with RM if evidence gathered for every method excludes FP. This is the case when, for example, there are no other compounds other than the basic candidate for identification with (a) spectra similar to the reference one and (b) RI within the target range. Correspondingly, other candidates for identification have (a) spectra little resembling the reference one and (b) RI which are far outside the reference range. However, the confirmation of identification results in co-analysis procedures is essential for cases involving legal responsibility, e.g., in analysis of samples originating from such accidents as disaster, poisoning, and so on.

***Co-analysis.*** It was noted above (Sects. 1.6 and 5.2) that co-analysis using RM (analytical standards) provides the strongest evidence for true identification. This is valid not only for target determination but also for unknown analysis. The difference between the two cases is that this analytical approach cannot be initially applied for the non-target determinations, because an analyte is unknown. So tentative identification should be achieved and some identification hypotheses (candidate compounds) should be set up and tested. The use of two and more orthogonal methods/techniques in co-analytical procedures is also appropriate.

***Information.*** The discussion of the problem of unknown/non-target identification shows that data/information play the main role for achieving the true results. There are several groups of data:

- Information about properties, origin, and use of chemical compounds
- Data on their abundance/popularity/occurrence
- Spectral libraries, first of all mass spectral ones
- Collections of chromatographic retention parameters, first of all RI in GC

The first two information groups required for setting up identification hypotheses (prior data on a sample 1, Fig. 7.1) and screening and accepting/rejecting candidates for identification (prior data on candidate compounds) were considered in Chap. 6. The same information may be used in different stages of the identification process.

Other data, noted in the centre of Fig. 7.1, are used for comparison with experimental data. Reference chromatographic and spectral information, with programs required for data use, corresponding data bases/systems, and related non-information topics, will be considered in this chapter in terms of different approaches to identification of unknowns.

## 7.2 Gas Chromatography Retention Indices

### 7.2.1 Index Types

Retention times as measured quantities in chromatography depend on too many factors to be specific enough for identification of analytes.[2] Influence of many factors is removed if retention parameters are calculated using a relative scale. These chromatographic parameters named *indices* (see reviews [8, 9] and the website [10]) are the essential supplements to mass spectra in non-target analysis. To obtain true results of GC identification, different types of RI, reference data for them, and correct choice of reference values and identification criteria, should be taken into consideration.

Kovats indices (KI) were the first to be introduced in the practice of qualitative chromatography analysis (see [8]). KI denoted by $I_x$ are measured under isothermal column conditions and calculated by the formula:

$$\frac{I_x}{100} = n + \frac{\lg t'_{R_x} - \lg t'_{R_n}}{\lg t'_{R_{n+1}} - \lg t'_{R_n}}. \tag{7.1}$$

where $x$ is the analyte, $n$ and $n+1$ are the number of carbon atoms of the reference $n$-alkanes which bracket the retention time of the analyte, $t'_{R_x}$, $t'_{R_n}$, and $t'_{R_{n+1}}$ are the adjusted retention times of the analyte and reference $n$-alkanes ($t'_{R_n} < t'_{R_x} < t'_{R_{n+1}}$); the adjusted retention time $t'_R = t_R - t_M$, $t_R$ is the retention time, and $t_M$ is the gas hold-up time.

For temperature-programmed GC, linear retention indices (LRI) $I_x^T$ have been proposed (see [9]). For a linear ramp temperature program, the indices are calculated by the following formula:

$$\frac{I_x^T}{100} = n + \frac{t^T_{R_x} - t^T_{R_n}}{t^T_{R_{n+1}} - t^T_{R_n}} \tag{7.2}$$

where different $t_R^T$ are the corresponding retention times; $x$, $n$, and $n+1$ are specified above.

The version of LRI (Lee indices) with aromatic hydrocarbons as reference compounds was proposed for identification of PAH (see [11]). The references are benzene, $n = 1$, $t_R^T = 100$; naphthalene, $n = 2$, $t_R^T = 200$; phenanthrene, $n = 3$, $t_R^T = 300$; chrysene, $n = 4$, $t_R^T = 400$; picene, $n = 5$, $t_R^T = 500$.

[2] A collection of RT can be also used for identification if they are perfectly reproduced [7].

### 7.2.2 Reference Data

Different collections of RI have been issued as reference books and databases on CD, or can be accessed through respective on-line versions (Table 7.1); also see [12, 35, 36] for these and some other databases. Collections are divided into general and special (field-oriented) ones. The NIST collection is the largest from general collections; it encloses the most known RI values (Table 7.1). The example of the record from NIST database on CD [37] is given in Fig. 7.2. In a similar way to the citation of compounds in CAS and other chemical databases (Sect. 6.3), index values are not uniformly distributed over different compounds [12]. More than one half of collected compounds have only been measured once. On the other hand, the minority of 2.4% abundant compounds with no less than 46 replicate indices for each compound provide more than 50% of the total of RI values [12].

Field-oriented reference data (see Table 7.1) have been mainly developed in:

- Toxicology
- Flavor and fragrance compounds, odorants, related substances
- Metabolomics

In new databases, RI are included together with mass spectra (see Table 7.1). In metabolomics (Sects. 7.4.1.3 and 7.7.1), reference data essentially are chromatographic profiles with RI values and MS tags. In some cases, metabolite compounds are not identified [26–28].

### 7.2.3 Choice of Reference Values

The choice first depends on chromatographic conditions used in the identification experiment. In isothermal conditions, Kovats indices are determined, which mainly depend on the composition of the stationary phase and the column temperature [8, 9, 12]. For the same nominal phase and temperature, KI are reproduced relatively well between columns, instruments, and laboratories. So reference data used in qualitative chromatographic analysis must refer to the same temperature and column (phase). On the other hand, experimental conditions can be chosen according to reliable reference data.

For several reference values available in databases for the same compound, phase, and temperature, a mean value may be used for comparison with an experimental RI. Means and corresponding standard deviations are suitable for statistical estimation (Sect. 3.6.5). With this, outliers should be discarded using statistical tests or analyzing distributions of reference values [38].

If reliable reference data for some columns and temperatures are absent or not representative enough for a statistical estimation, different values may be used, corrected for

**Table 7.1** Large collections of RI

| Name, reference | Collection | Index type | Columns | Mass spectra |
|---|---|---|---|---|
| NIST database [12] (2008 on CD[a] [13], Chemistry WebBook [10]) | 237,206 index values for 21,940 compounds[b] | Various | Various | + |
| Sadtler library[a] [14] | Over 2,000 compounds | KI, LRI | OV-1, SE-54, CW-20M | – |
| Pseudo-Sadtler [15] | 3,195 entries | Recalculated from Sadtler etc. | | – |
| Russian reference book [16] | Hydrocarbons and oxygen-containing compounds | KI | Various | – |
| LRI & Odour Database [17] | Over 9,000 records for over 5,000 volatile compounds of foods. | LRI | Various | – |
| ESO 2000 (update 2006)[a] [18] | About 2,500 essential oil components | | Various | – |
| Essential oils [19] | About 2,000 essential oil components | KI | SE-30, CW-20M | – |
| Chromaleont / FFNSC Wiley[a] [20] | 1,831 flavor and fragrance compounds | LRI | SLB-5MS | + |
| Flavornet [21] | 738 odorants | KI | OV-101, DB-5, OV-17, CW-20M | – |
| Terpenoids Library [22] | About 2,000 terpenoids and related components of essential oils | LRI | DB-1 | + |
| Book on identification of essential oil components [23] | Over 1,200 essential oil components | RT | DB-5 | + |
| Book on analysis of flavour and fragrance volatiles [24] | Over 1,150 flavor and fragrance compounds | KI | Methyl silicone[c], polyethylene glycol | + |
| Pherobase [25] | 17,000 records on over 7,000 pheromone compounds and semiochemicals | KI | Various | + |
| GMD [26–28] | 6,205 non-unique metabolites, not fully identified | | Rtx-5Sil MS | + |
| FiehnLib[a] [29, 30] | 1,200 values for over 1,000 identified metabolites | | DB5-MS, Rtx-5Sil MS | + |
| Wiley data on toxicological compounds[a] [31] | 7,840 datasets on drugs, poisons, pesticides, pollutants, and their metabolites | Temperature program | Methyl silicone[c] | + |
| Wiley data on designer drugs[a] [32] | 5,866 values for designer drugs | KI | | + |
| Toxicology [33] | Over 6,000 toxicological compounds, | KI, LRI | CW, SE-30, OV-1, dimethyl silicone [c] | – |
| Toxicology [34] (see [12]) | 4,500 toxicological compounds | | Dimethyl silicone[c] | – |

[a]Commercial data base

[b]Literature data, including recalculated ones. They are supplemented by theoretical values estimated with relatively large uncertainties. In total, there are 293,247 RI values for 44,008 compounds [13]

[c]The group of poly(dimethylsiloxane) phases

128. Value: 653 iu
Column Type: Capillary
Column Class: Standard non-polar
Active Phase: OV–1
Column Length: 17.5 m
Column Diameter: 0.2 mm
Phase Thickness: 0.15 um
Data Type: Kovats RI
Program Type: Isothermal
Start T: 50 C
Source: Johansen, N.G.; Ettre, L.S. **Retention index values of hydrocarbons on Open-tubular columns coated with methylsilicone liquid phases** *Chromatographia, 15*(10), **1982**, 625-630.

**Fig. 7.2** One of the records for benzene from NIST 05 database [37] (reproduced with permission)

- Column phase, with the use of typical or specially calculated differences in RI of specified compound classes between different phases and
- Temperature differences, using the temperature increment found in the literature or calculated for a given or similar compound; see Example 7.1

**Example 7.1.** In order to confirm the presence of methylcyclopentane as the impurity in n-hexane, KI of the compound was determined by GC–MS with the use of the poly(5% diphenyl-95% dimethylsiloxane) column at 30°C [39]. There were no literature reference values for just this phase and temperature. So data determined for (a) this column at different temperatures, and (b) different phases at various temperatures were used and corrected for both characteristics (Table 7.2). Reference indices were recalculated to 30°C with the use of the temperature increments, ΔRI/10°C, extracted from the respective reports (see [39]). Further, the phase correction of 4.0 i.u was added to experimental values according to the difference in KI of cycloalkanes between the two phases. Corrected values were finally averaged (Table 7.2).

**Table 7.2** Corrected KI of methylcyclopentane [39, 40]

| | Literature data | | RI correction, i.u. | | RI corrected, i.u. |
|---|---|---|---|---|---|
| RI, i.u.[a] | RI, i.u.[a] | ΔRI, i.u./10°C | 30°C | phase | |
| 624.4[a] | 30 | | 0 | + 4.0 | 628.4 |
| 626.0[a] | 30 | | 0 | + 4.0 | 630.0 |
| 628.2[a] | 45 | 1.45 | −2.2 | + 4.0 | 630.0 |
| 635.0[b] | 60 | 1.66 | −5.0 | 0 | 630.0 |
| 626.6[a] | 40 | 1.64 | −1.6 | + 4.0 | 629.0 |
| | | | | mean | 629.5 |

[a]Poly(dimethylsiloxane)
[b] Poly(5% diphenyl-95% dimethylsiloxane)

The LRI values depend on the larger number of factors than just KI [9]. Therefore, it is very important to choose (and standardize) conditions for recording chromatograms in temperature-programmed GC as well as isothermal conditions. The parameter $S$, which should be constant, was proposed for such standardization: $S = r_T \cdot t_M/\beta_I$, where $r_T$ is the heating rate; $t_M$, see (7.1); $\beta_c$ is the column phase ratio; $\beta_c \sim d_c/4d_f$, $d_c$ is the column inner diameter, $d_f$ is the thickness of stationary phase film (see [9]). However, that formula is not applicable to all possible cases. In particular, it discards a start temperature of a program. It is simpler to implement other approaches to compare indices in a correct way.

- Experimental conditions which are the same or similar to those used for recording reference data, can be chosen.
- Another approach is that the temperature program would be set in such a way to match reference indices of known components of a mixture under analysis [41].
- The known compounds from a mixture can act as the special markers to calculate the LRI difference between the experimental and reference data. This difference can subsequently be used as the correction to experimental indices.
- Those components can be recognized during the analysis itself, e.g., by using MS (in GC–MS analysis).
- Corrections can be required to remove the effects of irreproducibility of temperature programs [33]. The compounds used to recalculate RI by the formula (7.3) can initially be introduced into analyzed mixtures as secondary references/standards.

$$I_x^T = I_{R_1}^T + \frac{t_{R_x}^T - t_{R_1}^T}{t_{R_2}^T - t_{R_1}^T}\left(I_{R_2}^T - I_{R_1}^T\right) \tag{7.3}$$

  where $t_{R_x}^T$, $t_{R_1}^T$, and $t_{R_2}^T$ are RT of the unknown analyte and secondary standards, respectively; $t_{R_1}^T < t_{R_x}^T < t_{R_2}^T$ [33].

- (a) Indices measured in conditions of not the same temperature programs and even (b) KI and LRI may be not very different from one another. Thus, they can eventually be combined for statistical (or rather, quasi-statistical) estimates [38]. Corresponding mean index values, together with the standard deviations, are suitable for both tolerance and statistical tests of identification hypotheses (Sect. 3.6).

Indices estimated from different physical quantities and molecular features can be also used as reference RI. Current methods of estimation are based on:

- Correlations of RI with physical chemical properties, for example [42]
- Quantitative structure–property (retention) relationships (QSRRs) [43]
- The use of molecular statistical methods [44], and some others (see [43, 45])

Some quantitative predictions have resulted in a good accuracy of several i.u. [43], which is comparable with an uncertainty in many experimental reference RI. Thus, upon a careful examination of the origin, calculated indices can be used in identification procedures.

### 7.2.4 *Identification Criteria*

Criteria of identification based on RI are related to their reference values and corresponding tolerance/range windows, $\pm$ $\Delta$RI, consistent with a typical or specially estimated spread of values for particular experimental conditions. There are several ways of setting reasonable criteria.

**General tolerance criteria.** If experimental and reference data refer to (a) the same stationary phase and index type (isothermal or programming temperature) and (b) not very different other conditions, the general identification ranges of $\pm$ $\Delta$RI may be set up. They are (e.g., see [45]):

- 5–10 i.u. for standard non-polar, poly(dimethylsiloxane), and slightly polar, poly (5% diphenyl-95% dimethylsiloxane), phases
- 15–25 i.u. for standard polar, polyethylene glycol, columns

These ranges resemble values of interlaboratory reproducibility (see [43]). Further, a conventional view is that methylsiloxane columns provide better reproducibility of GC data. Therefore non-polar and slightly polar columns should be preferred to solve identification problems if the columns provide sufficient separation of mixtures under analysis. That may be not the case, e.g., see [46].

**Special criteria.** These are used for identification of particular groups of substances in similar conditions of chromatographic analysis. Some examples are presented below.

- Narrow tolerances can be established for identification of, for example, alkanes, using non-polar phases where the RI reproducibility is better (even far better) than $\pm$ 5 i.u., e.g., see [47].
- Metabolites [27, 28] and components of essential oils [48] are identified in the LRI window of $\pm$ 3 i.u.
- The range of $\pm$ 5 i.u is suitable for plant volatile compound KI calculated for temperature programming [49].
- LRI of food aroma volatile compounds are reproduced in the range $\leq$10 i.u. at different temperature programs [50].

**Criteria changed on correcting for RI.** Corrected indices are more accurate, and therefore fall into narrower ranges of values. This is the case for recalculating indices with the use of secondary indices [33]; see above. The LRI collection of compounds of toxicological interest measured for packed columns was used for identification using capillary columns and the criterion of $\pm$ 50 i.u. The wide range was due to interlaboratory irreproducibility of temperature programming. The recalculations of indices made it possible to narrow down the range to $\pm$ 25 i.u. [33].

**Statistical criteria.** These save analysts the trouble of selecting suitable tolerance ranges. The drawback of such criteria is that (a) conditions of recording current and reference data must be the same, and (b) there must be a sample of experimental and/or reference RI rather than individual indices. An example of statistical estimation is given below (Example 7.2).

**Example 7.2.** For the impurity in *n*-hexane [39], the replicate KI of 628.5 and 629.7 were obtained. The analytical purpose was to identify the impurity. The identification hypothesis was that the impurity is methylcyclopentane. This hypothesis was accepted if experimental and reference indices were insignificantly different, i.e., at $\alpha > 0.05$ (see Sect. 3.6.5). The reference indices for the compounds are represented in Table 7.2. The $\alpha$ values can be returned using numerous computer programs, starting with Excel. For those RI, in the cases of two-side distribution and the same dispersions, $\alpha$ was 0.58. So the hypothesis is accepted. The identification hypotheses for other candidate compounds with the same molecular formula of $C_6H_{12}$ were rejected by this *t*-test. The identification result was fully consistent with MS data [39].

The reference indices which are in the above example were extracted from the literature. Now there is the large NIST database containing RI for numerous compounds from different sources (Table 7.1). This dataset makes it possible to generate RI samples required for statistical tests. For example, the mean values of multi-literature RI, corresponding standard deviations, and confidence intervals have been estimated for components of essential oils [38]. Any interested analyst may make such estimations for his/her own researches based on the free version of the RI collection [10]. Depending on whether sample values (a) refer to the same RI type, columns, and temperatures/temperature programs, or (b) represent different index types and a wide range of experimental conditions, tests using these data are considered as (a) statistical or (b) quasi-statistical ones respectively (see [51]).

Samples from reference RI data are also useful for establishing range identification criteria. Indeed, it is appropriate to consider 95% confidence intervals as reasonable criteria for identification or no identification of unknown compound with RI which fall in or outside the ranges.

### *7.2.5 GC–MS*

Mass spectrometry is the most reliable technique for identification (see Sect. 7.4). However, at least two different techniques are required for unknown identification (Sect. 7.1). Thus, the combination of GC (reference RI) and MS (reference full EI mass spectra or selected ions) provide unambiguous identification of many volatile combinations, e.g., see [27, 28, 39, 41, 45–48, 50, 52]. The role of RI is that a group of compounds with similar mass spectra may be differentiated by these chromatographic parameters. Thus, the windows of $\pm$ $\Delta$RI can be considered as filters for MS data [48, 53].

**Table 7.3** Tolerances for identification of alkanes, cycloalkanes, and alkenes

| Technique | Parameter, quantity | Reliability level | | |
|---|---|---|---|---|
| | | High | Medium | Low |
| GC | $\alpha$, *t*-test | $\geq 0.05$ | 0.01–0.05 | $< 0.01$ |
| GC | $\pm$ ΔKI | $< 5$ | 5–10 | $> 10$ |
| MS | $\alpha$, $T^2$ statistics | $\geq 0.05$ | 0.01–0.05 | $< 0.01$ |
| MS | MF | $> 950$–1,000 | 925–950 | $< 925$ |

Extracted from [39] with a minor changes. The terms of high, medium, and low identification reliability accepted in the book correspond to low, medium, and high identification uncertainty [39] respectively

For all that GC–MS is very popular in identification operations, there are a few reports where rigorous criteria for identification by each from the two techniques are set. The three examples below refer to such researches.

- Metabolites were identified with the MS match factor (see Sect. 4.4.2) $> 650$ and the LRI window $\pm$ 3 [28].
- The pair of the criteria for unambiguous identification of PAH was: MF $\geq$ 800 and Δ(Lee RI) $\pm$ 2 [52].
- Different numerical criteria were established for different reliability of identification, see Table 7.3.

## 7.3 HPLC and Related Techniques

### *7.3.1 Introductory Note*

In the popular techniques of HPLC–UV–Vis, unknowns can be identified by their UV–Vis spectra and chromatography retention parameters. Also, retention parameters are now engaged in qualitative analytical procedures in proteomics. Parameters analogous to retention times are of value for identification in capillary electrophoresis and related techniques. In combination with MS, retention/migration parameters act as filters for redundant candidates for identification derived from mass spectrometry.

### *7.3.2 Libraries of UV–Vis Spectra*

Use of libraries of UV–Vis spectra is not very widespread as compared with MS or IR spectral libraries (see below). However, several data collections are available for comparison with experimental spectra acquired together with chromatograms (Table 7.4); see also [10, 57, 58].

**Table 7.4** Libraries of UV–Vis spectra

| Database[a] | Remarks |
|---|---|
| HaveItAll UV–Vis [54] | 20,928 spectra of pure organic compounds. Contains three databases: 14,464 spectra for the range of 200–350 nm, 5,559 spectra for 200–500 nm, and 905 spectra for 200–800 nm. |
| UV/Vis$^+$ [55] | About 5,600 spectra of about 900 substances |
| HPLC-DAD Data Base [56] | 3,270 toxicologically relevant substances, HPLC-DAD detector, RRT |
| NIST Chemistry WebBook [10] | 1,600+ spectra |

[a]Commercial databases excluding NIST collection

It would be perfect for an analyst if the HPLC-DAD instrument could be directly provided with the spectral library for on-line fast identification. That is available in toxicology, where reference spectra were recorded under the same experimental conditions [56] (see Table 7.4).

### *7.3.3 Retention Parameters and Their Reproducibility*

It is very hard if not impossible to identify unknowns using only RT, due to their general irreproducibility caused by variations in column properties, composition of mobile phase, temperature, and so on. In a similar way to GC, influence of many factors is removed if retention parameters are calculated in relative terms. In reversed-phase liquid chromatography, RI calculated by formulas (7.1) and (7.2) for isocratic and gradient conditions respectively were proposed for identification purposes; see [59–61] and references therein. Alkan-2-ones, alkyl aryl ketones, and 1-nitroalkanes were mainly used for scaling RI (see [59–61]).

However, some effects of irreproducibility, e.g., caused by different brands and batches of column [60], also remain with index parameters. As in GC (see Sect. 7.2), these effects can be partly removed using secondary standards when indices are recalculated by (7.3). In toxicology, different corrections were made for (a) acidic and neutral drugs and (b) basic drugs. This increased the interlaboratory reproducibility of RI estimated for 62 acidic and basic compounds. Without the corrections, the reproducibility was as large as $\pm$ 25 i.u. The spread of corrected values was diminished to $\pm$ 10 i.u. Nevertheless, the general tolerance of $\pm$ 30 i.u. was recommended for identification [60] where RI were combined with UV spectra or wavelength maxima [59–61]. The standardization of LC conditions, with the special focus on pH, is indispensable for interlaboratory reproducibility of RI values [62].

Recently, the approach related to secondary standards was combined with the chemometrical method of the resolution of overlapped chromatographic peaks. This was done for the purpose of the high-throughput screening of drugs by

HPLC-DAD, using the library of RI and UV spectra [63, 64]. A very high RT repeatability of better than 0.002 min in consecutive runs was achieved. Identification criteria for the method included (a) the spectral correlation coefficient $\geq 0.98$ and (b) the RI range $\pm$ 12 i.u. Rather low FP and FN rates were recorded that resulted in *St* 92% and *Sp* 94% (see Sect. 4.2 for the terminology and notions). The MLL rate (Sect. 4.2.8) with regard to compounds contained in the library was 1.255, i.e., relatively close to the limit value of 1 and far better than such rates for earlier techniques of toxicological analysis [65]. In contrast to this rate, two single criteria based on either RI or UV spectra led to ambiguous identification, as reflected by the MLL values of 3.26 and 3.72 respectively [63, 64].

A relatively good (home) RI reproducibility was observed in screening of 474 mycotoxins and fungal metabolites [66]. That was $\pm$ 1–2 i.u. for most compounds. However, some alkaloid- and amino acid-derived substances showed a variation in RI up to 10–20 i.u. and even the bias to 30 i.u. due to interactions with silanol groups of columns. The maximum bias in RI (up to 50 i.u.) was caused by ion-pairing with the component of the mobile phase [66]. In that report, RI are used for identification together with maxima of UV absorption and accurate ion masses.

Another screening method of toxicological analysis [67] has the similar selectivity as expressed by MLL of 1,253. The approach to identification was based on combination of the RRT collection and the library of UV spectra for 2,682 toxicologically relevant substances. Testing of the library resulted in 60 and 84% of compounds being unambiguously identified by UV spectra alone and by the pair data of the spectrum and RRT respectively [67]. Spectra of 3,270 compounds are contained in the newer version of the library (see Table 7.4).

Just as other types of chromatography RI, the HPLC indices are predictable, e.g., with ANN models. However, the prediction accuracy of $\geq$ 30 i.u. [68] seems to exceed typical values of experimental irreproducibility (see above).

The appearance of HPLC–MS ($MS^n$, HRMS) has shifted focus of identification procedures away from the use of the combination UV with RI to different types of mass spectra (see below). Nevertheless, mass spectrometric identification is not always unambiguous, and therefore retention parameters included in combined analytical procedures may provide more certain results. It is also clear that the use of RI will progress further with better reproducibility of RT and RI in liquid chromatography.

### *7.3.4 Retention Parameters of Peptides and Proteins*

Recent advances in proteomics related to sequence analysis of many proteins and peptides have been mainly due to peptide mass fingerprinting by MS (HPLC–MS). Also, the possibility that the chromatographic behavior of peptides can assist protein identification has been successfully explored; see reviews [69, 70] and references therein. Reference retention data used in proteomics are mainly predicted ones.

The peptide retention times, combined with $MS^2$ data analysis, enlarge the reliability of peptide identifications. Various approaches for peptide RT prediction in HPLC have been proposed [70]. Just as in the case of RI for volatile compounds (Sect. 7.2.3), many models for the prediction of peptide retention times are based on QSRRs, e.g., regression methods and artificial neural networks [69, 70]. Both peptide sequences and physicochemical features such as peptide hydrophobicity and molecule length have had an influence on peptide RT [70].

Predictive models, particularly those based on ANN, are relatively efficient. The RRT values named there *normalized elution time* were predicted with the error < 3–4% [69]. An even lower average error of 1.5% has been observed [71]. For an RT of 10 min, the error of 3% corresponds to the accuracy of about 20 s. This is quite similar to the experimental repeatability (reproducibility) of peptide RT expressed by the confidence interval ($\pm$ 6 ÷ 22 s) [72]. Special programs are developed to calculate sequence RT [73, 74].

Predicted values are used in a combination with MS data as a filter for candidate peptide mass matches (See Sects. 4.4.2.3, 7.4.1.4, and 7.7.2), which leads to a significant increase in the proportion of identified (correctly identified) peptides [75]. Preliminary filtering of candidate peptides by predicted RT increased the number of positive peptide identifications by as much as 50% at the false discovery rate (percentage of FP, see Table 4.3) of 3% [76]. In other research, concomitant filtering out FP with the use of predicted RT led to a ~19% increase in the number of positive peptide identifications achieved by MS [77]. The improvement of identification results may also be related to a large reduction of FP without a significant effect on the number of TP [78].

The use of HPLC, including corresponding retention parameters, is not necessarily related to proteomics. It was proved that RT of hemoglobins are diagnostic and reproducible, which makes it possible to reliably detect and identify more than 30 different hemoglobin variants; in many cases, it is superior to the efficiency of conventional electrophoresis procedures [79].

### *7.3.5 Migration Parameters in Electromigration Techniques*

Capillary electrophoresis (CE) and also related techniques are widespread in genomics and proteomics. To a somewhat lesser degree, they are applied for chemical determinations of low-molecule compounds. In toxicology, CE and micellar electrokinetic chromatography (MEKC) have been considered as methods of choice [80–84]. Migration parameters, such as migration time, relative migration, and mobility values (see [82]) are analogous to retention parameters in chromatography.

As reference data for CE, values of relative migration and electrophoretic mobility for more than 650 drug compounds have been reported [82]. Reproducibility of such parameters was improved when they were corrected (in a similar way

to HPLC) [80, 83–85]. However, neither CE nor MEKC (as with GC and HPLC alone) provided unambiguous identification of drugs, even with corrected parameters. This is expressed in relatively high MLL rates, 4.32 and 1.44 respectively, when these electromigration methods have been used (see Sect. 4.2.8 for the definition of MLL rates). Nevertheless, pair combinations of CE and MEKC with each other and with GC or HPLC led to more certain identification, with MLL of 1.0–1.2 [80].

Migration times in CE as reference data for identification are reasonably well predictable by the ANN technique. In identification of metabolites based on predicted values, true results appear among the best three candidate compounds in 78% of cases [86].

As is the case for HPLC–MS, the combination of CE with MS is the preferred technique for identification [85, 87]. Here, migration parameters are also considered. Examples can be given from metabolomics and proteomics. In the research [88], metabolites were tentatively identified by HRMS, taking into account accurate masses, isotopic patterns (see below), and electrophoretic mobilities of analytes. The relative migration time predicted based on absolute mobility and dissociation constant was in less than 2.0% agreement with corresponding experimental data for cationic metabolites [89]. That inaccuracy is consistent with experimental repeatabilities of such parameters: less than 1.5–4.5 % for migration times [85].

The contribution of CE in combined identification by CE–MS may be not only the relatively accurate estimation of migration values but also the prediction of migration order. It has been observed that the migration order of the phosphopeptides was correlated closely with their isoelectric points [87]. In the general case, such kinds of regularities are usable for differentiation between a limited number of candidates for identification.

## 7.4 Mass Spectrometry

### *7.4.1 Libraries*

For identification of unknowns, analysts use spectral libraries and also go to spectral interpretation. The versions of the MS technique used in these approaches to identification depend on the volatility of compounds and the complexity of systems/samples under chemical analysis.

Volatile compounds are mostly identified by GC–EI–$MS^1$. So mass spectral databases for analyses of these analytes contain electron ionization spectra. For determination of non-volatiles, the technique of HPLC-ESI–$MS^n$ is widespread. Here, $MS^2$ and other $MS^n$ spectra are the most useful for identification. Such spectra have been included in reference libraries, which are of the special type in metabolomics and proteomics.

#### 7.4.1.1 EI–MS[1]

***Reference libraries.*** The main libraries are listed in Table 7.5; see also [36, 96]. Standard conditions for recording mass spectra are quadrupole instruments and the 70 eV energy of ionizing electrons. EI mass spectra obtained in such conditions are reproduced well between time intervals, instruments, and laboratories, presenting the "gold standard" of mass spectral reproducibility. However, large spectral collections (NIST, Wiley; see Table 7.5) contain also replicate spectra with some variations in ion abundances. Replicates improve library searches [97]. Further, users of MS databases should take into account that there are some exemptions from standard conditions for mass spectrometric experiments; see Table 7.5.

Two large libraries, Wiley and NIST, consist of hundreds of thousands of mass spectra (Table 7.5). Furthermore, the publisher Wiley distributes a combined MS

**Table 7.5** EI mass spectral libraries

| Name | Compounds | Spectra | Remarks |
|---|---|---|---|
| Wiley[a,b] [90] | 667,000 | 796,000 | General dataset |
| NIST 08[a] [13 ] | 192,108 | 220,460 | General dataset |
| NIST Chemistry WebBook [10] | | Over 15,000 | General dataset |
| HaveItAll MS[a] [91] | | 199,000 | NIST 02 and several other collections |
| SDBS [92] | | Appr. 24,500 | General dataset. Magnetic and double-focusing sector mass spectrometers, electron energy 75 eV |
| AAFS Drug Library [93] | Hundreds | Over 2,700 | Drugs and metabolites |
| Pherobase [25] | $\leq$ About 2,500 | | Pheromones, semiochemicals |
| Terpenoids Library[a] [22] | Appr. 2,000 | | Constituents of essential oils. Double-focusing sector mass spectrometer |
| MassBank [94] | | About 13,000 | Metabolites and related compounds. Partly (a) CI spectra and (b) ToF mass spectrometer |
| GMD [26–28] | | Over 2,000 | Metabolites spectra, including 1,206 unique spectra and 535 identified unique spectra. Quadrupole and ToF mass spectrometers |
| FiehnLib[a] [29, 30] | Over 1,000 | 1,200 | Identified metabolites |

[a]Commercial library

[b]Registry of Mass Spectral Data, 9th edn with NIST 08. Smaller collections are also issued which relate to organic compounds from combinatorial synthesis, designer drugs, pharmaceuticals and agrochemicals, steroids, flavors and fragrances, pesticides, volatiles in food, geochemicals and petrochemicals, biomarkers [90, 95]

collection [90]. In general, libraries of such sizes are sufficient for solving most identification problems found by organic analysts when determining volatile/semi-volatile compounds. However, mass spectra of some targets may be absent or be erroneous in large collections. Thus, smaller databases should not be disregarded, especially in such areas as toxicology, pharmaceuticals, metabolomics, and so on (Table 7.5).

To search reference spectra in MS libraries and compare the analyte spectra with reference ones, there are special computer programs. The NIST MS Search (current version 2.0f [13]) is specifically designated to search in NIST libraries (see Table 7.5). The Wiley and NIST libraries have been also issued in software formats of the main manufacturers of mass spectrometers and other companies [13, 90].

Given that very complex mixtures are commonly analyzed by chromatography mass spectrometry, there is the need in programs for automated spectral deconvolution, i.e., detection and extraction of the spectrum of each component in a mixture for its further identification. The NIST AMDIS [13] is just such a program for GC–MS, and not the only one. For example, one of the modules of the Mass Frontier program (HighChem) enables component detection and spectra deconvolution from GC–MS and LC–MS ($MS^n$) data [98]. Such software has been developed specially for metabolomics; see [99] and other references in Sect. 7.4.1.3.

***Evaluation of Databases.*** The success of the identification procedure is critically dependent on a library quality provided by evaluation and testing of a library. The methods for critical evaluation of a large mass spectral library at its building from many individual collections have been described [100]. The main points of evaluation methods used in the NIST practice are the following.

- The quality is controlled by experts in the course of a spectrum-by-spectrum review of the library. A gain in quality depends critically on the expertise of the reviewers/evaluators.
- Chemical structures which are required for the evaluation of their mass spectra must be digitally represented for all compounds in the library. The correctness of compound names is also important.
- When reviewing spectra, the evaluator identifies peaks in terms of fragmentation reactions, and decides whether to accept, edit, flag (for example, as low quality), or delete the spectrum. If there is a problem how to act best, the spectrum is considered by a second evaluator. Evaluators should come to an agreement upon an action.
- A spectrum is included if it is consistent with the structure of the molecule and contains the majority of the characteristic peaks. An assigned name, structure, and the spectrum itself should be consistent. Incomplete spectra are included only if the compounds are of special interest for builders of the library. Further, the isotope ratios should be correct for both the molecular ion and its major fragments. The major peaks are examined to be reasonable for the particular structure of the compound under consideration. The expertise of the evaluator is based on an intimate knowledge of rules of fragmentation in EI ion source (e.g., see [101]).

- Spectra are edited if there are correctable errors such as (a) peaks due to impurities, (b) effects of ion–molecule collisions in an ion source, (c) displacements of peaks in the mass scale by one unit, (d) spurious peaks, and some others. Problems arising from low volatility and/or reactivity of compounds or impurities in the sample are specially treated. For example, the region of the molecular ion peak is examined to confirm that the compound does not decompose on vaporization.

The procedures [100] described above are not only human/manual but also computer-assisted, e.g., an inspection of isotope ratios. Somewhat similar approaches are used in the evaluation of Wiley libraries [97].

The so-called *quality index* was introduced, which takes into account higher molecular mass impurities, illogical neutral losses, isotopic abundance inaccuracy, number of spectral peaks, and some other factors leading to false results [102]. Later, the index was modified to express an overall data base quality [97]. The indices in both versions are estimated by computer. Nevertheless, the author agrees with the point of view that a full quality of spectral library can not be provided without expert inspection of spectra [100].

The practice of the use of spectral libraries is that the experimental spectrum is compared vs. library ones and then a ranked list of matching reference/library spectra is generated (e.g., Tables 4.17 and 6.8). Modern large libraries and searching algorithms and programs provide correct answers placed in the top lines of search result lists, i.e., answers of high ranks. According to testing results of libraries and searching algorithms summarised in Table 7.6, appr. three quarters of searches led to correct identification at the 1st rank[3] (the highest MF and upper line in hit lists).

**Table 7.6** Performance of searches in EI mass spectral libraries

| Algorithm[a] | Reference library | Test spectral set | % Correct answers (TP) at rank | |
|---|---|---|---|---|
| | | | 1 | 1–3 |
| Dot product [103] | 62,235 | 12,593 | 72.9 | 90.8 |
| NIST dot product[b] [103] | 62,235 | 12,593 | 75.7 | 92.5 |
| PBM [103] | 62,235 | 12,593 | 64.7 | 84.8 |
| NIST dot product[b] [104] | 62,235 | 12,593 | 77.0 | 91.6 |
| PBM[c] [104] | 62,235 | 12,593 | 74.9 | 86.4 |
| NIST dot product[b] [104] | 228,998 | 370 | 75.4 | 87.8 |
| PBM[c] [104] | 228,998 | 370 | 77.0 | 88.9 |
| PBM [97] | 229,000 | 310[d] | 79 | |

[a]See Sects. 4.4.2.1 and 4.4.2.2
[b]So-called composite algorithm in the article [103]
[c]Commercial PBM according to [104]
[d]There are 1,421 spectra of these compounds included in the reference library

[3]Interlaboratory comparison resulted in the lower rate, but the test spectra sample was very small [105].

The performance of spectral searches affects the choice of identification criteria. Beginning analytical mass spectrometrists often consider the upper lines in the search lists as definitive identification results. Table 7.6 shows that this is false for about 20–25% of cases. The low MF threshold (*min MF* in Fig. 3.6) seems to be a more reliable criterion. For not very high threshold values, the MF of most analytes falls in the tolerance range for identification which is between *min MF* and *max MF* (Fig. 3.6).

On the other hand, this range collects not only TP but also FP, which are mainly compounds similar to analytes in structure/spectrum. The relationship between *TPR* and *FPR* is another performance of spectral database combined with searching algorithms. Corresponding two-dimensional plots are named *receiver operating characteristics* (ROC) [106]. There are plots of dependencies of *TPR* from *FPR* (or related quantities). Values of both variables are determined by changing the MF threshold, including changes due to modifications in searching algorithms and formulas for the calculation of MF. In previous research on retrieval of spectra (e.g., [97, 103]), dependencies of this kind were named recall–reliability plots; see Table 4.3 for terminology, and Example 7.3.

**Example 7.3.** The library consisted of 200 mass spectra of 200 different compounds to be tested. Experimental/test spectra were specially acquired for the test sample of 100 compounds, reference spectra of which were contained in the library. Library searches (Table 7.7), i.e., comparisons of each query spectrum (from the test set) vs. all the reference spectra, were performed at different MF thresholds starting with the almost maximum value of 990 (the maximum is 1,000). A computer answer of the 1st rank (the top line) was a positive result. If matching spectra belonged to the same compound, it was TP. At the highest threshold, only 30 top reference spectra were retrieved (Table 7.7). Most of them (29 items) were spectra of test compounds. In this series of searches (see Table 4.3 for notations), $TPR = St = recall$ is $29/100 = 0.29$ or 29 %, *FPR* is $1/100 = 1\%$, $PPV = reliability$ is $29/30 = 97\%$. By diminishing the MF threshold, the number of retrieved reference spectra was increased as well as result rates, with the exception of the proportion of true positives (*PPV, reliability*); see Fig. 7.3. In unknown analyses, this means that most computer answers will need a confirmation.

Figure 7.3a shows that *TPR* values were far from 100% at low FP (high MF; see Table 7.7). A perfect librarian search would be achieved in the case of 100% *TPR* at any level of *FPR*. The perfect ROC would coincide with the vertical axis and the line parallel to horizontal axis at the 100% level (upper line, Fig. 7.3a). The area under the ROC curve is the special measure of the integral performance of the search algorithm/software (and the library quality), independent of the choice of the particular threshold MF or corresponding *FPR*, e.g., see [107, 108]. For the perfect searching method, the area is $100\% \times 100\% = 100$ area%. In the case under consideration (curve in Fig. 7.3a), the area under ROC was about 85%.

**Table 7.7** True and false results of MS librarian search depending on MF

| MF | TP | FP | TP+FP |
|---|---|---|---|
| 990 | 29 | 1 | 30 |
| 975 | 47 | 3 | 50 |
| 950 | 77 | 25 | 102 |
| 900 | 90 | 60 | 150 |
| 850 | 96 | 85 | 181 |
| 800 | 99 | 99 | 198 |

The above in this section shows that the 1st rank of matching spectrum is not the perfect criterion for identification. Setting a MF threshold is a better decision. However, it is followed by (a) appearance of a group of candidates for identification and (b) a need to test related identification hypotheses by other methods. Some of the redundant hypotheses can be rejected using prior information (Chap. 6).

#### 7.4.1.2 $MS^n$

Electrospray ionization (ESI), atmospheric pressure chemical ionization (APCI), laser desorption (MALDI), and other "soft" ionization techniques, have increased the potential of MS and its combination with liquid chromatography in analysis of non-volatile compounds (see Sect. 2.8.4 and reference below). Over time, several ESI/APCI–$MS^1$ (conventional single-stage mass spectrometry) libraries containing up to approximately 1,600 spectra of hundreds of compounds have been generated [109] (see also [110] and Table 7.8). In comparison with $MS^1$, the combination of new ionization techniques with tandem mass spectrometry ($MS^2$), with the interface of collision-induced dissociation (CID), led to more reliable identification because of (a) better control of fragmentation of ions containing intact molecules of analytes by varying the collision energy, and (b) removal of background ions arising from solvents, etc.

***Reference libraries.*** For purposes of identification, ESI–$MS^2$ libraries (product-ion libraries) of different sets of compounds ranging up to several thousands of substances have been generated (Table 7.8). Libraries may be classified according to such characteristics as large/small, universal/field-specific, transferable/home, and so on.

The NIST 08 is the largest[4] of the universal libraries (and possibly the only current universal library). It seems to be transferable to a greater or lesser degree between instruments and laboratories, because it consists of tandem mass spectra from various collections, acquired on tandem instruments of different types. However, the library size is yet inadequate for the purpose of identification of many significant substances. Indeed, if one assumes that at least tens of thousands of

[4]The largest collection [120] (see Table 7.8) was reported without many details about this home library.

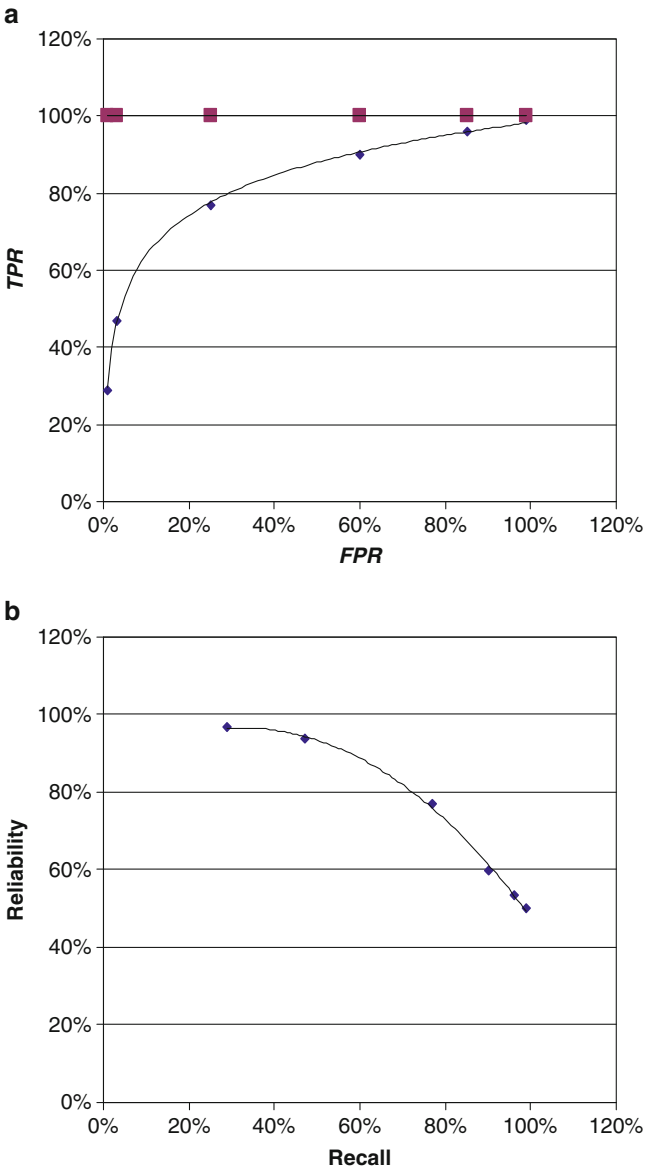



**Fig. 7.3** (**a**) Receiver operating characteristic and (**b**) recall–reliability plot for the same library searches, Table 7.7. For conceptions and terms, see Table 4.3. The two lines in Fig. 7.3a are: real librarian searches (the *bottom curve*, see data in Table 7.7) and perfect results (the *upper line*)

**Table 7.8** $MS^2$ mass spectral libraries

| Title, institute/company | Compounds | Spectra | Remarks |
|---|---|---|---|
| NIST 08[a] [13] | Over 5,000 | 14,802 | Organic compounds of various classes. Spectra from NIST and some other laboratories and literature. Mainly ESI–TQ–$MS^2$ and also $MS^1$, $MS^{>2}$, different ionization/collision energies/instruments. 3,898 positive and 1,410 negative precursor ions |
| Centre for Chemical Sciences, Royal Holloway University of London[b] [111] | Over 600 | Over 1,000 | Drugs, drug metabolites, pesticides, natural products and dyes. ESI or APCI, positive mode. One collision energy |
| Institute of Legal Medicine, University of Freiburg [a,c] [112–114] | | | |
| –New collection | 1,253 | Over 5,600 | Drugs and pharmaceuticals. ESI–TQ–LIT, three collision energies and energy spread. Positive, negative modes. |
| –Previous collection | 800 | Thousands | Drugs and pharmaceuticals. ESI–TQ, three collision energies. Positive, negative modes. |
| Universities/Institutes in Taipei [115] | over 800 | Thousands | Toxicological compounds. ESI–IT, four collision energies and energy spread. RT |
| Institute of Legal Medicine, Innsbruck Medical University [107, 116, 117] | 402 [107], Appr. 900 [117] | 3,759 [107], 9,275 [117] | Drugs. ESI–Q–ToF, ten collision energies. Positive, negative modes. |
| Department of Forensic Medicine, University of Helsinki[d] [118] | About 400 | | b-blocking drugs. ESI–TQ, from one to three collision energies |
| Bruker Daltonics [119] | 127 | 311 | Toxicological compounds. IT, 127 $MS^2$ and 184 $MS^3$ spectra. RT |
| Bristol–Myers Squibb [120] | | > 68,000 | Drugs, metabolites. IT |
| MassBank [94] | | Over 8,000 ($MS^2$) and 3,000 ($MS^1$) | Metabolites and related compounds. ESI–TQ–$MS^2$ > ESI–Q–$MS^1$ > ESI–Q–ToF–$MS^2$, etc |
| HMDB [121, 122] | 2654 | Thousands | Metabolites. ESI–TQ, different collision energies. Positive, negative modes. |
| METLIN [123] | 2057 | 11,100 | Metabolites. Positive, negative modes. Different collision energies |
| HighChem[a] [124] | Over 1,000 | Thousands | Pharmaceuticals, natural compounds, metabolites. IT–$MS^n$, n = 1 ÷ 3. Positive, negative modes |
| RIKEN Plant Science Center ESI–$MS^2$ library [125, 126] | 860 | 61,920 ($MS^2$, $MS^1$) | Plant metabolites. ESI–TQ–$MS^n$, $n$ = 1,2. Positive, negative modes |

| | | | |
|---|---|---|---|
| MMD [36, 127] | 788 | 1,500 | Metabolite standards. ESI–HRMS$^2$, two collision energies. Positive, negative modes |
| Korea Research Institute of Bioscience and Biotechnology [128] | Over 930 | Over 1,000 | Over 500 flavonoids and 430 microbial metabolites. ESI–IT, one collision energy. Positive, negative modes. |
| Syngenta Crop Protection [129] | 1,020 | 1,020 | Natural products. ESI–IT. One collision energy. Positive, negative modes. |
| Graduate School of Science, Hokkaido University [130] | | | Oligosaccharides. IT–MS$^n$, $n = 1 \div 3$. Retention parameters |
| National Institute of Advanced Industrial Science and Technology [131] | | | Oligosaccharides. MALDI–Q–IT– MS$^n$, $n = 1 \div 3$. |
| FragLib [132] | | | Oligosaccharides. ESI–IT, MS$^n$, $n=1 \div 5$. Positive mode |
| TaMaSA [133] | 482 | 1,381 | Pesticides and related compounds. Adopted spectra from Internet and literature. Mainly ESI–TQ–MS$^2$ and also EI, IT, MS$^{3-4}$, other ionization/ instruments. Positive, negative modes |
| US EPA test library [110] | 129 | | Organic compounds of environmental concern. ESI–Q–MS$^1$, ESI–TQ–MS$^2$ |

Full spectra libraries. See also [109, 134–141]

[a]Commercial library

[b]Probably contained in the NIST library

[c]There are also small MS$^2$ library for designer drugs and ESI/APCI–MS$^1$ libraries for drugs, pesticides and explosives

[d]There also were MS$^1$ libraries

compounds are abundant (this is an underestimation, see Sect. 6.3) and takes into account that up to 10 different spectra are required for reliable recognition of an analyte by $MS^2$ [107, 142], the conclusion can be made that hundreds of thousands of spectra is a minimum requirement for the comprehensive universal $MS^2$ library of low molecules.

Field-specific databases are also relatively large, ranging to thousands or even tens of thousands of spectra (Table 7.8). They have been developed within metabolomics, which is one of the moving forces of the progress in modern analytical science/practice, chemo- and bioinformatics. Other prominent fields for the use of $MS^n$ libraries are toxicology, food, and environment (Table 7.8). Most libraries are home-made, with MassBank [94] and TaMaSA [133] as exceptions. Possibilities of the use of home-made libraries in other laboratories should be explored.

MS libraries under consideration have been generated using various instrumental platforms. ESI is the most popular ionization method, although APCI is also used. Reference EI/CI spectra for GC–$MS^n$ are rather rare items [13, 133]). As for mass analyzers, most library spectra were acquired on TQ (tandem in space, special cell of collision activation for fragmentation) and classical IT (tandem in time). In recent years, the contribution of Q–ToF and combinations including LIT, ICR and Orbitrap (see Sect. 2.8.4) has been growing.

As a rule, the number of reference library spectra is larger than the number of corresponding compounds, i.e., there are several replicate spectra for each unique compound. Replicates may be "true" ones, i.e., generated in the same or similar experimental conditions, or "quasi-replicates", acquired at different collision energies on the same/different tandem instrument. Replicates of the second kind are both very dissimilar to each other, and very essential to provide the best matching experimental spectra to reference ones (see below). It should be noted that a spectral appearance is critically dependent on the conditions of CID, first collision energy [143].

Algorithms and programs of the library search are or may be the same as in the case of EI–$MS^1$. It would be appropriate if the modified versions, e.g., providing the option of fractional *m*/*z* tolerances of precursor and/or product ions (see [13, 116]) were available for advanced retrieval of reference spectra.

***General evaluation of libraries.*** Quality assurance of tandem MS libraries has not yet been concerned with all the details which are considered in building EI–$MS^1$ libraries [100]. However, some aspects of the quality of $MS^2$ reference data have been controlled.

When generating the research library consisting of 3,766 $MS^n$ spectra of 1,743 compounds [142] and included later in NIST MS library [13] and the TaMaSA $MS^n$ library of pesticides [133], the author deleted spectra of poor quality and edited/processed many intended spectra. A poor quality of a spectrum[5] means that

- Intense peaks are annotated with incorrect *m*/*z* values or are not annotated at all

[5]The $MS^2$ spectrum with a few peaks is sometimes considered as of low quality [129]. However, this may be just a case of fragmentation where one or a few reactions prevail.

- There are intense peaks with *m*/*z* exceeding the precursor ion mass
- There are intense fragment peaks related to "illogical" neutral losses
- In a series of spectra obtained at a gradual collision energy ramp, the peak intensity ratio is very different from that for "neighboring" spectra recorded at close energies
- The noise is intense (at least several percent relative to a basic peak) and distributed over the full spectral range, and so on

The edition of spectra is commonly accompanied by the removal of individual noise, spurious, or artifact peaks. The latter may appear in experimental spectra when a computer program processes anomalously wide and/or tailed peaks of precursor ions. Spectral editing similar to that has been reported [107].

***Spectral reproducibility.*** In general, the use of product-ion mass spectra for identification of unknowns may be limited by insufficient spectral reproducibility in different tandem mass spectrometers. Specific inter-laboratory studies have been conducted to compare tandem spectra and to evaluate the uncertainty in compound identification based on spectral matching (Table 7.9). As a result of standardization of experimental CID conditions with a selected tuning compound, $MS^2$ spectra acquired on three [144] or four [145] TQ mass spectrometers from two manufacturers were rather similar. Recently, a good match of $MS^2$ spectra was observed between the TQ and Q–LIT instruments where standardized collision energies and probably one-type collision cells were used [114] (Table 7.9).

Earlier, good reproducibility for those types of tandem instruments was not observed [146]. In general, a spectral similarity degree between the tandem in space (TQ) and time (IT and other instruments) used without their combinations seems to be poor [134, 135, 138, 139] (see Table 7.9). Correspondingly, that affects a potential of identification. On the whole, $MS^2$ spectra are reproduced much worse than EI–$MS^1$ ones and the match degree, e.g., demonstrated by Fig. 7.4, should be declared as good or even very good.

***Testing and validation of libraries.*** Spectral reproducibility is indirectly revealed by testing libraries and estimating rates of true or false computer answers. Two transferable libraries [133, 142] (see Table 7.9) were evaluated and validated in such a way. The method for such evaluation of libraries is that all the test $MS^2$ spectra were in turn incorporated into the test sample of "unknowns" spectra and into the reference library. Spectra of any given compound were considered as test ones if originated from at least two different sources, i.e., laboratories or publications. At the beginning of the particular search for any test compound, the subset of its spectra originated from the same source was extracted from the library, followed by searching in the $MS^n$ library with every spectrum from this subset. The search results were combined and the best match over the subset was selected. Then this subset was returned to the library and the next subset for the particular compound was extracted (Fig. 7.5). In statistics/chemometrics, a similar procedure is named *cross-validation* (see Chap. 8).

The rank of the correct computer answer depends or may be dependant on the number of replicate ("quasi-replicate") spectra involved in librarian searches, a

**Table 7.9** Reproducibility of product-ion spectra for different instruments/laboratories

| Year, reference | Comparison | Conclusion |
|---|---|---|
| 2000 [144] | ESI–$MS^2$ spectra of drugs. Three (two different) TQ from two manufacturers, three laboratories. Tuning collision energies | Good reproducibility of mass spectra in the similar CID conditions |
| 2004 [145] | ESI–$MS^2$ spectra of 30 drug compounds. (a) The TQ instrument over 4-year period. (b) Four different TQ of two manufacturers, four laboratories. Tuning energies | Fair/good reproducibility in all cases, with the average MF 78–95% from 100%. The lower reproducibility for different instrument brands |
| 2004 [135] | ESI–$MS^2$ spectra of 19 steroids. Q vs IT, one manufacturer, one same laboratory | Interinstrument reproducibility: elucidation of "gross structure" and diminished possibility to differentiate stereo- and regioisomers |
| 2004 [134] | ESI/APCI–$MS^2$ spectra of 20 test compounds. Five different instruments (three IT, one TQ, one FT ICR) from three manufacturers. Different laboratories. Tuning to abundance 10–50% of the $[M + H]^+$ ion | Match of *m/z* of at least three from five most abundant ions in pair interinstrument comparisons in appr. 65–90% of cases |
| 2005 [142] | Test spectra selected from the set of 1,018 $MS^2$ spectra of 193 compounds vs interlaboratory reference library of appr. 3,766 $MS^n$ spectra of 1,743 compounds, including 3,126 $MS^2$ spectra. Different ionization (ESI, APCI, CI, etc.) and instruments (IT, TQ, Q–ToF, etc), large variations in experimental conditions and the number of replicates | Correct answers as the 1st rank in up to 60% of the searches (all cases, on average 2.2 "unknown" vs 6.2 reference replicates per compound). With two or more replicates of both "unknown" and reference spectra (the average numbers of replicates 4.0 and 7.8 respectively), 77% of correct answers of the 1st rank. |
| 2005 [146] | ESI–$MS^2$ spectra of eight drugs. Q–LIT with two different scan modes and two TQ from different manufacturers, one laboratory. Standardization of CID conditions. | Significant differences in relative ion intensities between scan modes and instruments. The effects of an analyte concentration and contamination in the ionization source on the relative intensity of ion peaks |
| 2008 [139] | Negative ESI–$MS^2$ spectra of four dyes. Seven different instruments (three IT, two TQ, two Q–ToF), four manufacturers, different laboratories. Tuning energies and other experimental conditions | Differences in the relative abundances of product ions between instruments, mainly between IT and Q-based platforms |
| 2008 [138] | ESI–$MS^2$ spectra of 48 compounds. Eleven instruments: six IT, two TQ, one Q–LIT, and two Q–ToF, four manufacturers, different laboratories. Tuning CID conditions. | Interinstrument reproducibility: (a) poor in general, (b) fair between IT and between Q-based analyzers , (c) better for IT of the same brand. Correspondingly (a) 27%, (b) 57-58%, and (c) 80% of the spectra matched to MF >70% |
| 2009 [116] | 418 test $MS^2$ spectra of 22 compounds. Different instruments (TQ, Q–LIT, Q–ToF, LIT–FT ICR) from three laboratories. Interinstrument comparison of test spectra vs reference library of 3,759 $MS^2$ spectra of 402 compounds (see Table 7.8) acquired at up to | Relative peak intensities somewhat varied. Correct answers as the 1st rank in 98%. More different spectra obtained from LIT–FT ICR, with 95% of correct answers. |

| | | |
|---|---|---|
| | ten energies. No common standard CID conditions. Special matching algorithm and spectral processing | |
| 2009 [107] | Two test samples of 710 and 98 $MS^2$ spectra of control compounds vs the above library of 3,759 $MS^2$ spectra. Spectra of controls partly not available in the library | Most control compounds truly identified (sensitivity *St* 95–96%) or not identified (specificity *Sp* 94–96%). 98% specificity at the match of the precursor *m*/*z* between compared spectra |
| 2009 [133] | Test spectra selected from the subset of 562 $MS^2$ spectra of 151 compounds belonging to the TaMaSA library (Table 7.8). Comparison of test spectra vs combined TaMaSA (excluding test spectra) and NIST 05 (about 5,000 $MS^2$ spectra of about 1,950 compounds) libraries | In general, correct answers as the 1st rank in (a) 90% and (b) 72% of the searches, (a) with and (b) without matching precursor *m*/*z*, respectively. For the subset of at least two replicates of both "unknown" and reference spectra (the average number 3.0 and 3.6 respectively), (a) 95%[a] and (b) 90% of correct answers. |
| 2009 [114] | ESI–$MS^2$ spectra of 15–25 compounds. Interinstrument comparison of Q–LIT with TQ and two Q–LIT from two other laboratories, one manufacturer. Three standard particular energies and also energy spread | Good reproducibility between Q–LIT and TQ (MF 87-93%) and high similarity within two pairs of Q-LIT (MF 90-98 %) |

See also [110, 128, 147]

[a]Corrected value instead of 97% from [133]

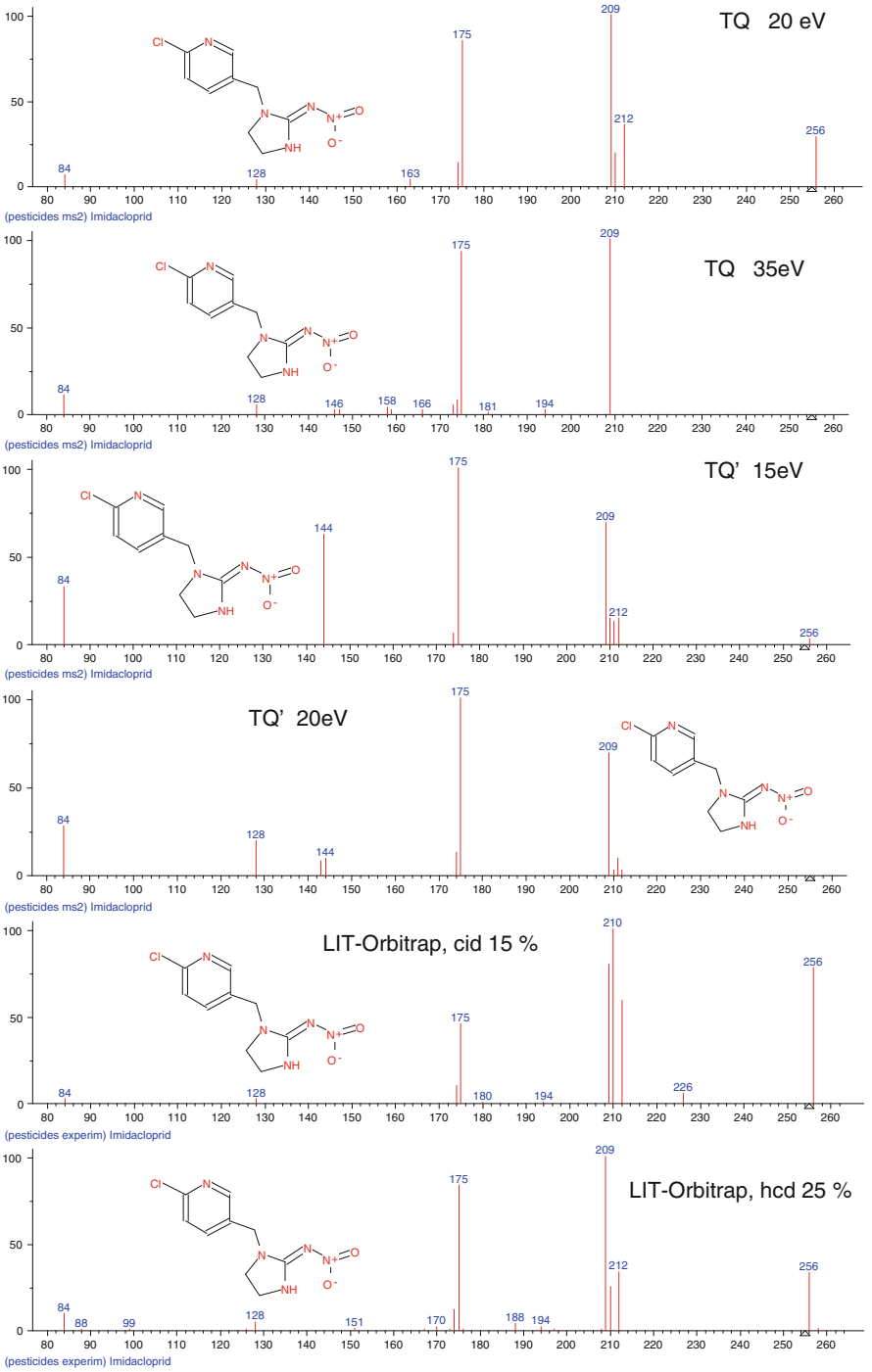



**Fig. 7.4** Tandem mass spectra of the pesticide imidacloprid extracted from TaMaSA [133] (*four top spectra*) and acquired in the author's laboratory (*two bottom*). Mass analyzers and collision energies are specified. TQ and TQ' are different triple quadrupoles. The acronyms of *cid* and *hcd* denote different modes of $MS^2$ scans of the LIT–Orbitrap instrument [148]

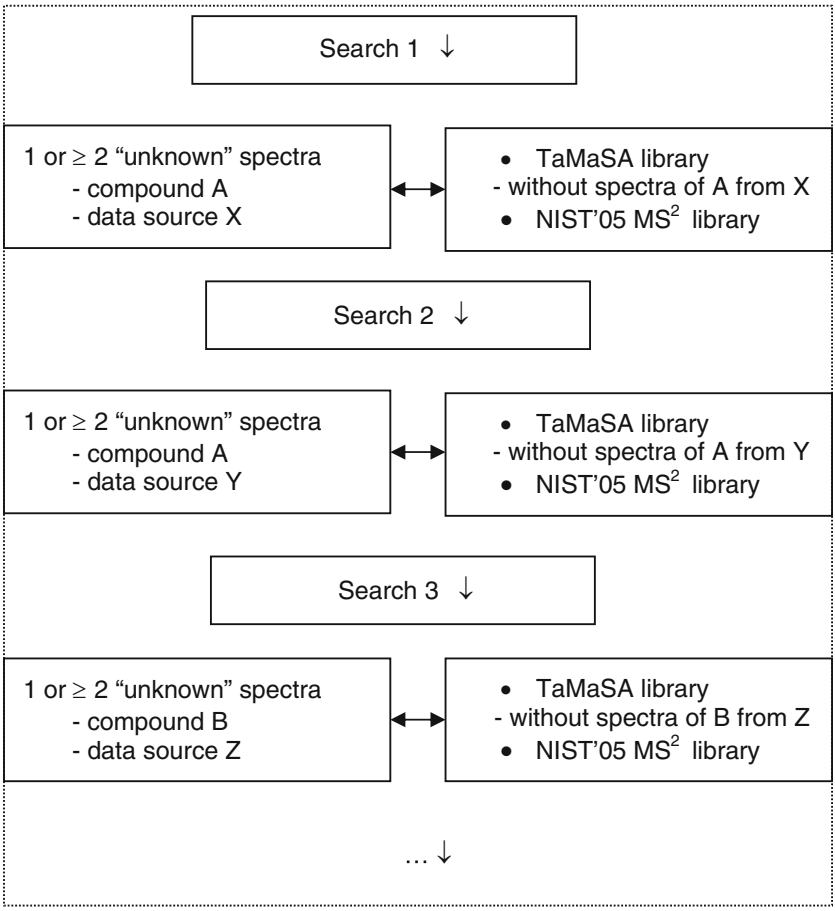



**Fig. 7.5** Chart for library searches in the TaMaSA library of pesticide spectra [133]. These originate from the literature sources and Internet. The search result is the best matching, its MF value and the rank of MF in the hit list. A similar method of testing the library was used in the research [142] where spectral subcollections from not only articles but different laboratories were entered into the database

difference in the type of tandem instrument, and so on. For the library of 3,766 $MS^n$ spectra, the first rank was nothing more than fair ($\leq$60%) in general, and growing to the level of 77% in searches with two and more replicates of both "unknown" spectra and two references [142] (see Table 7.9). Therefore, the efficiency of identification by means of $MS^2$ reference libraries may be close to that using the standard EI–$MS^1$ library, where the percentage of 1st rank correct answers was 79% (Table 7.6). The latter result was provided with an average of 4.6 reference entries vs. the only "unknown" spectrum [97] For the tandem library under consideration, the result of 77% for searches was based on 4.0 spectra of "unknown" and 7.8 reference spectra on average (Table 7.9).

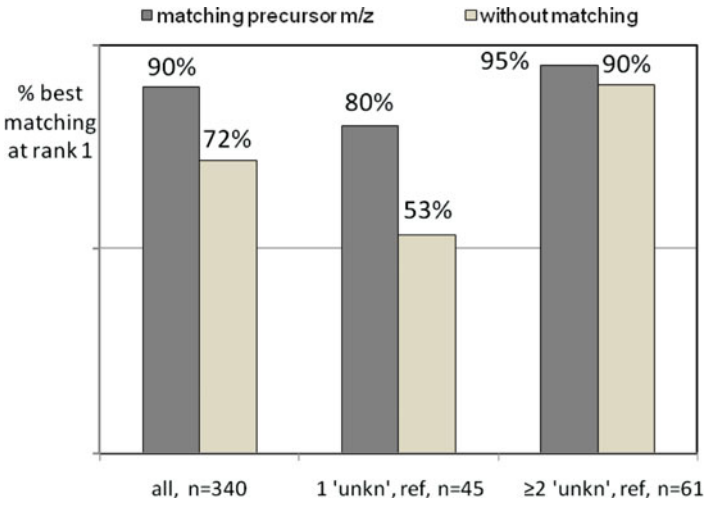



**Fig. 7.6** The results of library searches; *n* is the number of searches, *'unkn'* is "unknown" spectra, *ref* is reference ones. Replicate "unknown" spectra were mainly acquired at different collision energies

Testing of the library of the lesser size, TaMaSA, resulted in even better performance [133] (Table 7.9). In librarian searches performed with two and more replicates of both sorts (an average of 3.0 "unknown" and 3.6 reference spectra), 90% of correct answers of the 1st rank were observed. The addition of the criterion of matching precursor *m*/*z* increased the latter percentage up to a "suspicious" 95%. The dependence of the percentage on the number of both replicates is shown in Fig. 7.6.

Interlaboratory comparisons of TQ vs. TQ and IT vs. IT in the researches [133, 142] were not fully conclusive, because some spectral samples were not sufficiently large. However, the tendency can be seen that TQ spectra are more similar to each other than those recorded on IT. Cross-comparisons, i.e., TQ vs. IT and v.v. provided search results closer to those obtained for samples of TQ [133] or IT [142] spectra.

Recent studies [107, 116] have confirmed that (a) a large number of replicate spectra (up to 10 "quasi-replicates" recorded at different energies) and (b) the requirement of matching precursor *m*/*z* within the tolerance established before MS analysis led to high rates of true positive and negative results of librarian searches, close to 100%. That library was one of the first ones for $HRMS^n$, which combines the identification powers of HRMS and $MS^n$. It should be added that in the article [107] as well as in other reports, e.g., [131, 133], the identification performance of $MS^2$ libraries is discussed in terms of true/false result rates and correspondingly selectivity/specificity (for the terminology, see Chap. 4).

Thus, some recent tests [107, 133] demonstrated a very low rate of FP if only computer answers of the 1st rank, with matching *m*/*z* of precursor ions, are taken into account as the identification criteria. However, it would be correct to check the robustness of this conclusion over a wide range of libraries, including transferable

ones, analytes, analyte concentrations, and matrices. Without such confirmations, the "good old" criterion attached to the MF threshold as in EI–MS (see above) seems to be a good solution when analytes are identified by the use of $MS^n$ libraries.

This is exemplified by unknown screening in toxicology, where the $MS^1$ library built for the technique of CID-in-source (fragmentation in the ion source, the special collision cell is absent) was used. There,

- MS criterion of the reverse MF $>$ 60% (100% is the maximum value) and
- LC criterion of RRT being within $\pm$ 20% of the reference value

were applied [149]. These MS criteria can be established as corresponding to one or another level of TP and FP based on a dependence type shown in Fig. 7.3. Example 7.4 demonstrates how to make such estimations.

**Example 7.4.** In testing the TaMaSA library [133] (see Tables 7.8 and 7.9), TP and FP were recorded. The first of them are related to the best match spectra of target compounds (*pesticides*) retrieved at the 1st rank. The second are *non-pesticide compounds*, with spectra matched to those of test compounds also at the 1st rank. For example, pair comparisons between three "unknown" and four reference $MS^2$ spectra of diuron resulted in the best match of MF 592. Meanwhile, the spectrum of another compound was similar to a greater degree to each of three test diuron spectra, providing FP with MF 605 which is the best matching.

Two MF sets for TP and FP, i.e., for two sets of identification hypotheses, are distributed (Fig. 7.7). The distribution of FP is not surprisingly shifted to a lower MF. The threshold value of 275 can be established for the identification criteria. That provides *TPR* 95%. FP data may also be observed above the threshold, which leads to only 87 % proportion of TP in reference to the overall number of TP and FP (the *PPV* value, see Table 4.3).

The lower and higher thresholds can be set up. In the case of the former established at MF = 200 (not shown in Fig. 7.7), *TPR* of 100% with *PPV* of 86% is produced. At the higher threshold of 620, the *FPR* index decreased to 5%, but some TP can be missed in the general case (Fig. 7.7). The presence of FP at each MF border value means that identification results obtained with the use of such an $MS^2$ library must be confirmed by at least one other method/technique.

#### 7.4.1.3 Metabolomics

Collections of mass spectra of metabolites are described in Tables 7.5 and 7.8. They are useful for both target analysis and metabolite profiling (see Sect. 7.7.1). For the intended purpose of metabolomics, reference metabolite profiles, e.g., pair sets of

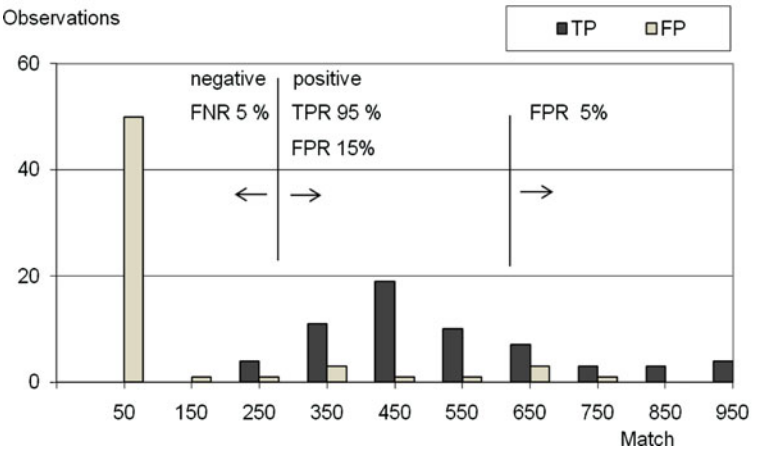



**Fig. 7.7** The histogram of MF for TP and FP results of searches in the TaMaSA library [133]. The NIST MS Search program [13], the search options of MS/MS identity and matching precursor $m/z$, the case of two and more both test and reference spectra available per one test compound. Two lines are at threshold values of MF (275 and 620). Above the first and second borders, 95% of TP and 5% of FP matches were observed, respectively

retention parameters and mass spectra, are of more value than single chromatographic or MS data. The profiles have been also included in such databases as GMD [26–28], FiehnLib [29, 30], RIKEN Plant Science Center library [125, 126], and MMD [36, 127]; see Tables 7.5 and 7.8.

Not all metabolite components of such data samples, both references and data obtained from experimental workflows, are definitely identified. Non-identified components are or may be annotated by retention parameters and mass spectra, including accurate mass values, e.g., see [36, 99, 150]. Special software has been developed for rather similar purposes of (a) the management and editing of metabolite mass spectral libraries [151], (b) systematically cataloguing metabolite peaks and their further identifying [152], and (c) alignment of large metabolite profile sets into data useful for identification [99].

#### 7.4.1.4 Proteomics

In the standard way, proteins are identified by peptide ion mass fingerprinting (Sects. 4.4.2.3, 4.5.4.3, and 7.7.2). Non-spectral databases consisting of amino acid sequences of proteins have been required for respective identification of these high-molecules (e.g., see [153–156]). In proteomics, libraries of $MS^2$ peptide spectra and corresponding identification procedures have been also developed.

The libraries are built from spectra of peptides which are (a) produced by protein digestion, and (b) identified by peptide fragment fingerprinting of theoretical $MS^2$ spectra generated from reference sequences on the basis of rigorous criteria [108];

see also [157] and [158]. This approach to building peptide $MS^2$ libraries can be considered as the typical one.

If purified protein standard samples are selected for digestion [159], peptides produced can be considered as transferable to identities of initial proteins. From the perspective of metrology, both libraries eventually generated just from protein/peptide standards and the identification approach itself based on the use of such libraries are of the upper reliability for proteomics. For protein reference materials, see also [160].

In most experimental spectral collections constituting raw data, there are many spectral replicates per peptide. The "best" spectrum can be selected, and so-called best-replicate libraries have been proposed (see [161]). In another way, the peptide MS library was created by averaging different experimental spectra contained in the large proteome database (the Global Proteome Machine Database, see below) [162]. Averaging was demonstrated to provide a better combination of performances for library searches than individual replicates [161, 162]. Such combined/averaged spectra, called "consensus spectra," have been included in the SpectraST library and NIST Peptide Mass Spectral Libraries [161, 163, 164].

Among many other subapproaches to protein identification, the use of the $MS^2$ peptide library consisting of spectra simulated by means of the kinetic model [165], should be also noted.

Building of peptide tandem MS libraries and their use has called for new special software (see [166, 167] and references in this Subsect.).

There is a global repository, the Global Proteome Machine Database [168], containing MS data on 888,874 of unique ("distinct") proteins. The data sets have been contributed by researchers from many countries and selected by database developers for data quality, biological interest, and so on. The Peptidome is another public resource that collects, archives, and freely distributes tandem mass spectra for peptide/protein identification [169]. For references to other public data repositories, see also [163].

Just as with other MS libraries, peptide ones are evaluated in terms of quality and performance. In quality control, problems of noisy, contaminated, and singly observed spectra and also contradictory identifications have been addressed [161, 163]. There are other characteristics of poor quality spectra such as low ion abundances, a few peak spectra, and ones with very short peptides; see [170] and also Sect. 7.7.2. Spectra of insufficient quality have been filtered out before being entered in libraries [157, 161, 163, 170].

The library quality affects the performance of spectral searching. The performances themselves have been expressed as rates of TP, FP, and so on, and also ROC curves (see Sect. 7.4.1.1 and Table 7.10). All of those variables were recorded with varying MF (score) thresholds established for true peptide identification (Table 7.10).

The spectrum library-based approach for peptide identification was compared with the conventional one of matching theoretical spectra. It was concluded that the former is more rapid in implementation, and results in the larger number of truly identified peptides [158, 162, 163]. Therefore, the conclusion can be made that the identification procedure based on libraries is preferable for target

**Table 7.10** ROC curves, $y = f(x)$, as performances of searches in peptide $MS^2$ libraries

| Reference | *y* | *x* |
|---|---|---|
| [108] | *TP* | *FP* |
| [157] | *St* (*recall*) | *1-Sp (FPR)* |
| [158] | *St* | *1-Sp (FPR)* |
| [161] | *TP* | *FDR* |

Notations, see Table 4.3. Different MF (spectral similarity scores) are varying threshold parameters

proteins, and the approach with searches in amino acid sequence databases is applicable for both proteins already studied (with a somewhat lower performance) and new sequences.

### *7.4.2 HRMS*

Various topics related to identification with the use of HRMS were considered above:

- Instruments, Sect. 2.8.4
- The difference between experimental and theoretical/formula masses ($\leq$ a few ppm) as the spectral MF and the criterion for identification, Sect. 4.4.2.2
- Requirements for HRMS in methods, Sect. 5.5.3.3
- The schematic role of this technique in unknown identification (Figs 6.5 and 7.1)
- Appearance of HRMS libraries (spectra mostly acquired by ToF instruments, see Tables 7.5 and 7.8)

As the value of the technique has been highly increased, HRMS-based identification procedures should be addressed in more detail. This is especially significant for the approach where identification begins with the use of HRMS (the right part, Fig. 6.5). This mass spectrometric technique is commonly combined with ESI to ionize and analyze non-volatile compounds.

There is a serious handicap to identification using only this technique. As a rule, HRMS overgenerates molecular formulas of analytes corresponding to the measured ion mass and the accuracy of its measurement. For a mass within the *m*/*z* range of 300 ÷ 1,000 and a reasonable mass accuracy of 5 ppm, the number of elemental compositions of ions containing C, H, N, O and other common elements is measured in hundreds and thousands (Table 7.11). Even very high mass accuracy commonly achieved by FT ICR ($\leq$ 1 ppm, e.g., see [171]) cannot provide an unambiguous determination of molecular formulas. Thus, incorrect formulas must be removed in one or another way. Several operations have been introduced to filter these formulas and reduce their number, to achieve certainty in solving identification problems (Fig. 7.8).

**Table 7.11** The approximate number of possible molecular formulas for elements: C, H, N, O, S, P, F, Br, Cl, and Na [136]

| Mass accuracy, ppm | Number of formulas for the ion mass, Th | | | | |
|---|---|---|---|---|---|
| | *m/z* 100 | *m/z* 200 | *m/z* 300 | *m/z* 400 | *m/z* 1000 |
| 10 | 5 | 100 | 300 | 1,000 | 10,000 |
| 5 | 3 | 50 | 150 | 500 | 5,000 |
| 2 | 2 | 20 | 60 | 200 | 2,000 |
| 1 | 1 | 10 | 30 | 100 | 1,000 |

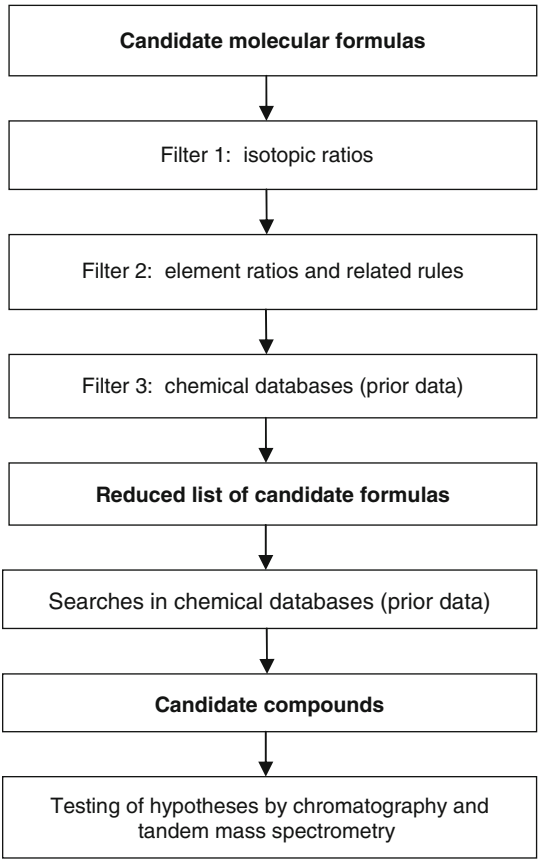



**Fig. 7.8** Flow diagram of filtering out of redundant candidate formulas and further pathway to candidate compounds. Filter 1 may not be very efficient, see text. The diagram can be incorporated into the general schematic for unknown identification, see Figs. 6.5 and 7.1. In the case of a plethora of candidate compounds, the technique of $MS^n$ can be also used first before applying HRMS, see Example 7.5 below

***Filter 1: abundance of isotope peaks.*** The first filtering out of improper formulas is based on criteria of isotope ratios. There is the (different) natural abundance of isotopes of D, $^{13}C$, $^{15}N$, $^{18}O$ and others which form peaks of the $[M+1]^{+\cdot}$ and $[M+2]^{+\cdot}$ ions in EI mass spectra and $[M+2]^+$ and $[M+3]^+$ species in ESI/MALDI spectra. So isotope ratios/patterns, first of all $[M+1]^+/M^{+\cdot}$ or $[M+2]^+/[M+1]^+$, are characteristic of elemental compositions of molecules. In turn, tolerances of these ratios are

filters for removing molecular formulas, with compositions significantly differing from the correct analyte formulas (e.g., see [172, 173] and references therein).

From early years of the technique, especially with the appearance of double-focusing sector mass spectrometers and later ToF, FT ICR, and Orbitrap, analytical and organic mass spectrometrists have taken into account isotopic ratios to perform structure elucidations. In recent years, the approach to identification incorporating measuring isotopic abundances has been developed by metabolomists [172, 174, 175] (for metabolomics, see Sects. 7.4.1.3 and 7.7.1). It was demonstrated that a low percentage tolerance for isotopic ratios might theoretically be a powerful filter for enriching a set of candidate formulas with a few advanced ones. For example, a substance with molecular mass of 500 Da may have 266 and 64 elemental compositions at a mass accuracy of 10 and 3 ppm respectively. The 2% isotopic abundance accuracy added to the last mass window reduces the set of 64 formulas to only three [172].

In another report, 5% tolerance as some kind of standard window for isotopic ratio deviation was considered [175]. Unfortunately, this accuracy level of the relative intensity of isotopic peaks is hardly approachable for most common high-resolution mass spectrometers in routine analytical practice (complicated mixtures, high-throughput determinations, a few or no replicates, low signals of analytes). This follows both from the author's experience of analyses using such instruments as IT–ToF and LIT–Orbitrap and from the literature [174]. For ToF, deviations from correct isotope ion ratios were up to 50, 20, and 10% in the case of peak intensities of $< 60$, $60 \div 200$, and $> 200$ counts respectively [176]. Therefore, special efforts should be made to efficiently use the filter under consideration; special software has been developed for the purpose [177].

Commonly, there are no problems with isotopes of chlorine and bromine (and also silicon and sulfur). The presence of atoms of these elements in many molecules can be easily determined without accurate measuring ion masses and abundance ratios [101].

In any case, commercial software processing data obtained by HRMS may provide scoring differences between experimental and formula isotopic ratios (e.g., in instruments manufactured by Shimadzu; see Fig. 4.6, and Waters, see [178]).

***Filter 2: element ratios.*** An experienced organic chemist knows that, as a rule, in most molecules, the number of hydrogen atoms exceeds that of carbon atoms, and the number of nitrogen and oxygen heteroatoms is smaller than two first numbers. In general, the ranges for possible element ratios are relatively wide, but probably limited by the estimates presented in Table 7.12. According to our observations, a common software which generates candidate formulas from experimental accurate ion masses and established mass tolerances does, as a rule, mostly provide plausible elemental compositions within limits of ranges indicated in Table 7.12. The reason seems to be that unusual combinations of elements lead to rare molecular masses deviating from those of analytes. The fact of "correct' elemental ratios is demonstrated in Table 7.13, where formulas generated by the Thermo Xcalibur software for one from the unknown components of the sample

**Table 7.12** Element ratios in molecules of abundant compound [175]

| Element ratio | Value range |
|---|---|
| H/C | 0.2–3.1 |
| F/C | 0–1.5 |
| Cl/C | 0–0.8 |
| Br/C | 0–0.8 |
| N/C | 0–1.3 |
| O/C | 0–1.2 |
| P/C | 0–0.3 |
| S/C | 0–0.8 |
| Si/C | 0–0.5 |

Correspond to 99.7% of 45,000 formulas from the Wiley MS library

**Table 7.13** Candidate formulas for ion with *m/z* 426.2425

| Theo. Mass | Delta (ppm) | Ion composition for $[M+2H]^{2+}$ | Molecular formula | Occurrence[a] | |
|---|---|---|---|---|---|
| | | | | PubChem | Google |
| 426.2426 | −0.23 | [12]C26 H54 O5 N29 | $C_{26}H_{52}N_{29}O_5$ | no | no |
| 426.2426 | −0.23 | [12]C27 H60 O10 N22 | $C_{27}H_{58}N_{22}O_{10}$ | no | no |
| 426.2426 | −0.24 | [12]C28 H66 O15 N15 | $C_{28}H_{64}N_{15}O_{15}$ | no | no |
| 426.2423 | 0.35 | [12]C43 H70 O14 N3 | $C_{43}H_{68}N_3O_{14}$ | no | no |
| **426.2423** | **0.35** | [12]**C42 H64 O9 N10** | $\mathbf{C_{42}H_{62}N_{10}O_9}$[b] | **9(5)** | **3** |
| 426.2423 | 0.36 | [12]C41 H58 O4 N17 | $C_{41}H_{56}N_{17}O_4$ | no | no |
| 426.2430 | −1.21 | [12]C42 H54 N21 | $C_{42}H_{52}N_{21}$ | no | no |
| 426.2430 | −1.21 | [12]C43 H60 O5 N14 | $C_{43}H_{58}N_{14}O_5$ | 1(1) | no |
| 426.2430 | −1.22 | [12]C44 H66 O10 N7 | $C_{44}H_{64}N_7O_{10}$ | no | no |
| 426.2430 | −1.23 | [12]C45 H72 O15 | $C_{45}H_{70}O_{15}$ | 7(6) | 6 |
| 426.2419 | 1.33 | [12]C26 H64 O14 N18 | $C_{26}H_{62}N_{18}O_{14}$ | no | no |
| 426.2419 | 1.34 | [12]C25 H58 O9 N25 | $C_{25}H_{56}N_{25}O_9$ | no | no |
| 426.2433 | −1.80 | [12]C28 H56 O6 N26 | $C_{28}H_{54}N_{26}O_6$ | no | no |
| 426.2433 | −1.81 | [12]C29 H62 O11 N19 | $C_{29}H_{60}N_{19}O_{11}$ | no | no |
| 426.2433 | −1.82 | [12]C30 H68 O16 N12 | $C_{30}H_{66}N_{12}O_{16}$ | no | no |
| 426.2417 | 1.92 | [12]C41 H68 O13 N6 | $C_{41}H_{66}N_6O_{13}$ | no | no |
| 426.2417 | 1.93 | [12]C40 H62 O8 N13 | $C_{40}H_{60}N_{13}O_8$ | no | no |
| 426.2417 | 1.94 | [12]C39 H56 O3 N20 | $C_{39}H_{54}N_{20}O_3$ | no | no |
| 426.2437 | −2.78 | [12]C44 H56 O1 N18 | $C_{44}H_{54}N_{18}O_1$ | no | no |
| 426.2437 | −2.79 | [12]C45 H62 O6 N11 | $C_{45}H_{60}N_{11}O_6$ | no | no |
| 426.2437 | −2.80 | [12]C46 H68 O11 N4 | $C_{46}H_{66}N_4O_{11}$ | 2(2) | no |
| 426.2413 | 2.90 | [12]C25 H68 O18 N14 | $C_{25}H_{66}N_{14}O_{18}$ | no | no |
| 426.2412 | 2.91 | [12]C24 H62 O13 N21 | $C_{24}H_{60}N_{21}O_{13}$ | no | no |
| 426.2412 | 2.92 | [12]C23 H56 O8 N28 | $C_{23}H_{54}N_{28}O_8$ | no | no |

Data for the extract of the algal biomass sampled in the Sestroretskij Razliv lake, Saint-Petersburg region, Russia, in summer of 2008. ESI mass spectrum was acquired on the Orbitrap HPLC–HRMS. Theo. Mass is the formula mass, Delta (ppm) is the difference between experimental and predicted/formula mass

[a]Searches on 12 March, 2010. Data sources, see Table 6.2. The number of formulas is noted, with the number of unique compounds/structures in parentheses

[b]The tentative identification as anabaenopeptin F, monoisotopic mass 850.4701, based on several criteria (see below)

analyzed are verified against the rules under consideration. One can see that there is no elemental composition with ratios falling outside the ranges given in Table 7.12.

***Filter 3: Chemical databases.*** It was proved in Sect. 6.2 that a presence of an analyte in sample can be predicted if the corresponding compound widely occurs both in real-world samples and in chemical databases, with many entries in the case of the most abundant compounds. A relevant search can be carried out by their names or other identifiers, e.g., formulas. Ultimately, filtering of molecular formulas generated from high-resolution mass spectrometry data according to the presence (and the number of occurrences) in databases shortens the list of candidate formulas for their testing in subsequent identification procedures. In recent years, the search in chemical databases has become a common practice in many laboratories using HRMS, e.g., see [133, 175, 176, 179].

Table 7.13 shows that most formulas generated by software for high-resolution mass spectrometers are those of nonexistent or very rare chemical compounds. Only four from 25 formulas generated for the ion with $m/z$ 426.2425 which is the $[M+2H]^{2+}$ doubly charged ion of the biomass component occur in the particular chemical database and Internet.

***Candidate compounds.*** Molecular formulas are ambiguous identifiers (Sect. 1.5.3). Thus, each candidate formula may cover more than one individual compound. Further searches in chemical databases are required to find out candidate compounds instead of candidate formulas for definite identification. The number of compounds exceeds the number of formulas, and the difference between them determining a set of hypotheses to be subsequently tested varies widely.

Two cases considered, see Table 7.13 and Example 6.1 with regard to pesticides, demonstrate statistics of formula/compound occurrences in two databases.

- For relatively heavy organic compounds, there may not very many known candidate compounds for each formula (more than two compounds in only two cases, Table 7.13). Ultimately, 14 compounds should be further tested, most of which are easily rejected without experimental tests (see below).
- Pesticide formulas are far more ambiguous. On average, the formula covers 71 isomer compounds (searches in CA for 2 years, see Table 7.14). Nevertheless, this does not mean that the same number of hypotheses should be tested. The most cited candidates for identification, i.e., test pesticides themselves, occur about 100 times more frequently than other candidate/isomer compounds. In most cases, the latter are so rare that each from them occurs in no more than 5% of the overall citation count for the particular formula. So this test proves the

**Table 7.14** Statistics of database occurrences of 18 pesticides and their isomers and formulas

| Quantity | Value |
|---|---|
| Number of formulas | 18 |
| Average number of candidate compounds per one formula | 71 |
| Average number (%) of database occurrences | 163 (62%) |
| pesticides | 1.7 (1.4%) |
| other candidate compounds | |

Searches in CA (CD editions) for 2007–2008. See Example 6.1 for other details

statement that the compound which occurs most in the database may turn out to be analyte presenting in the sample.

***HRMS$^n$.*** The database occurrence filter removes many redundant candidates for identification and not all of them. Combination of HRMS with another technique, first of all $MS^2$, will lead to more definite results. In the general case, a combined procedure can be implemented starting with each one, e.g., see [180]. For a particular analytical problem, the order of application of the techniques (or the order of data analysis if data are simultaneously acquired) depends on whether spectra for the analytes are present in $MS^2$ libraries. If this is the case, it would be more productive to start with librarian searches; see Example 7.5 below. Libraries consisting of spectra with accurate mass value would be the best solution, but they are only beginning to emerge. Nevertheless, common libraries for low-resolution mass spectrometry with integer mass values (most in Table 7.8) are also fit to the purpose.

The searches may result in very low MF for hit records. It means that corresponding reference spectra probably are not available in a library. In this case, candidate formulas and corresponding compounds should be found in chemical libraries. Advanced/filtered candidate compounds are further tested as identification hypotheses by means of (a) interpretation of tandem mass spectra, (b) prediction of retention parameters, and/or (c) co-analysis with the proper reference materials; see Example 7.6 in Sect. 7.4.3.

**Example 7.5.** Here, the identification potential and productivity of techniques of HRMS and $MS^2$ are estimated for the test set of 18 pesticides (see Example 6.1, Fig. 6.6, and Table 7.14). ESI–$HRMS^{1,2}$ spectra were simultaneously acquired for corresponding reference materials (the Orbitrap instrument). So (a) accurate $m/z$ of $[M+H]^+$ ions and (b) $MS^2$ spectra, six per compound, were obtained.

If identification starts with HRMS data (see Fig. 7.8), it would result in 171 candidate formulas without their filtering. Among them, 48 formulas occurred in PubChem at least one time, and 23 passed the filter of 5% of the overall occurrence rate; see Example 6.1. In any case, 23 or 48 elemental compositions correspond to hundreds of individual compounds (Table 7.14). To continue identification, these compounds should be found in chemical databases and reference $MS^2$ spectra should be also found, recorded, or predicted for widely occurring compounds from those hundreds of structures. Next, reference or predicted spectra should be compared to experimental ones. However, many spectra are unlikely to be found or predicted with the necessary accuracy. Even if one supposes that the abundant candidate compounds are not very numerous and the required spectral data are available, it would be a lengthy procedure to test this volume of hypotheses from the beginning.

*(continued)*

In general, the conclusions about the use of only HRMS are that (a) individual identification is impossible, (b) all pesticide formulas meet the criterion of low difference between experimental and formula mass (measured difference $< 4$ ppm), and (c) 2/3 formulas are identified as true ones with low *FPR*, i.e., other candidate formulas have less than 5% of the overall occurrence rate (see Example 6.1).

The start with experimental $MS^2$ spectra is more productive and advanced. The special TaMaSA library was developed for pesticides (Table 7.8), evaluated for true and false result rate (Fig. 7.7), and used for this case. The results of library searches are given in Table 7.15. Fifteen from 18 results are reliable true positive or negative ones which are based on the low probability of a false result. The remainder are false or less certain results. Combining these results with those obtained with the use of HRMS changes that relationship.

The identity of eight pesticides reliably recognized by $MS^2$ is confirmed by the use of HRMS, with low probability of FP estimated by means of occurrence searches in the database (Table 7.15). The other two from ten reliable TP go into the group of less certain results ($< 95\%$ "true" occurrences). The overall conclusion about the use of both techniques is that the 13 from 18 results are true, with a low false result rate. It is just the level of screening methods to which identification based on database searches refers in most cases. It should be noted that $MS^2$ spectra of other candidate compounds, if available, could affect the above estimations.

**Table 7.15** Test for identification of 18 pesticides

| Number of compounds | $MS^2$ Result[a] | Reason | Number of compounds | $MS^2$ and HRMS Combined result | Reason[b] |
|---|---|---|---|---|---|
| **10** | **TP** | 1st rank of the spectrum in hit list, $FPR < 5\%$ | **8** | **TP** and **TP** | $> 95\%$ occurrences |
| | | | 2 | **TP** and ~TP | $< 95\%$ occurrences |
| 2 | ~TP | 1st rank of the spectrum in hit list, $FPR > 5\%$ | 1 | ~TP and **TP** | $> 95\%$ occurrences |
| | | | 1 | ~TP and ~TP | $< 95\%$ occurrences |
| 1 | FP | 3rd rank of the spectrum in hit list | 1 | FP and ~TP | $< 95\%$ occurrences |
| **5** | **TN** | No spectra in the library, TNR $< 5$ %. | | | |

[a]The ranges for MF corresponding to different positive and negative results are shown in Fig. 7.7. ~TP specifies the particular TP result which is relatively uncertain, i.e. *FPR* is larger than 5%
[b]Searches in PubChem; see Example 6.1. Here ~TP specifies the TP result with <95% of overall occurrences in that database

Some other approaches to identification based on a very challenging combination of HRMS and $MS^2$ have been reported, e.g., see [133, 140, 176, 178, 181–183].

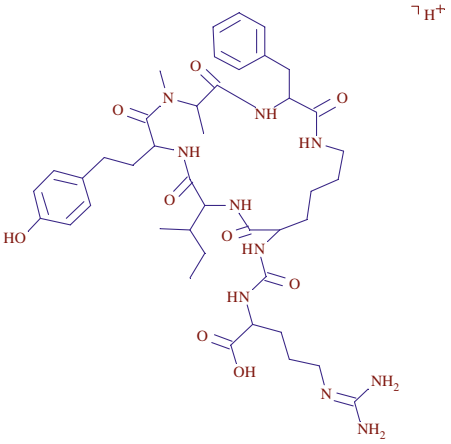
***High molecules.*** This section is mostly devoted to low molecules. Applications of HRMS for identification of high-molecular compounds, e.g., in proteomics is also very essential. Reduction in the number of candidate formulas of peptides is not the only gain obtained from the use of this technique. A new analytical

methodology has been developed which includes an initial characterization of intact protein ions by accurate masses ('top-down" approach, see [184, 185]). High resolution is required for resolving overlapping isotope peaks of multi-charged protein ions ($[M+nH]^{n+}$) typical for ionization of these high-molecular compounds in conditions of ESI [154, 155]. The distance between peaks in such spectra expressed in the mass unit is very short (1/n Da); see Fig. 7.9. Thus, high-resolution measurements take it possible to resolve a peak and to determine a charge of an ion and further generate possible molecular formulas.

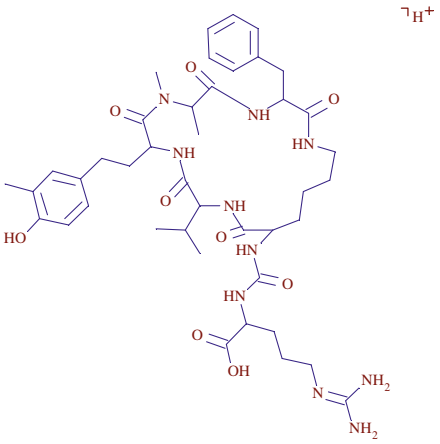
### 7.4.3 *Spectral Interpretation*

In addition to searches on spectral databases, non-target identification can be also provided by means of computer-assisted methods commonly used for the structure elucidation of new organic compounds. This approach is of more value for identification of compounds whose spectra are unavailable in libraries. The methods of *computer-aided structure elucidation* based on the use of expert systems are briefly considered in Sect. 4.5.5. Modern systems of this kind are mainly intended for structure interpretation of NMR spectra, and approaches related to other types of spectra have been also developed [187–189].

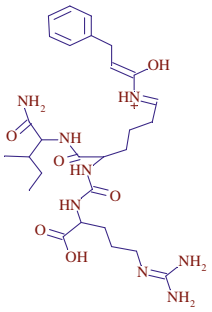
In mass spectrometry practice, especially in that of $MS^2$, computer systems which are directed to the prediction of mass spectra from structures of organic compounds, based on known fragmentation rules, have been increasingly used. The most popular programs probably are Thermo/HichChem Mass Frontier [98] (applications, e.g., see [128, 129]) and ACD MS Fragmenter [190]; some other programs have also been created [191]. The use of the MS Frontier software is demonstrated in Example 7.6. In this and many other cases of complex structures of analytes, such a program of the predictor type has a heuristic value, rather than providing strong evidence for absolutely reliable identification.



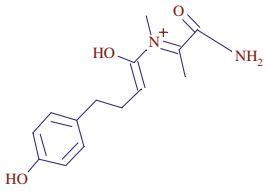

**7.1**




**7.2**




7.3




7.4

In determinations where an analyte amount is sufficiently large to record IR and/or NMR spectra and an analyte is sufficiently pure, MS could be used in combination with those spectral techniques [187, 189, 193].

## 7.5 IR Spectroscopy

IR, and to a lesser degree Raman spectroscopy, have traditionally been required for structure elucidation of pure new organic compounds using reference data on full spectra or individual specific bands; see references in Sect. 2.8.2 and also [194, 195]. This is the kind of general approach 4 to identification; see Table 1.4. Traditional methods of such spectral interpretation rely on the analyst's experience. Software tools have appeared to assist in verification and interpretation of IR spectra, which may provide more rapid and reliable identification. Software of this sort uses spectral interpretation rules based on spectra–structure correlations, e.g., see [196–198]. It is also possible to obtain information about substructures of unknown analytes from similar spectra of other compounds retrieved by searches in IR spectral libraries [199, 200].

Reference IR spectral libraries (including NIR and Raman, Table 7.16) are mostly used for direct identification of compounds by matching their spectra with reference ones (general approach 2, Table 1.4) as well as in mass spectrometry. The use of IR spectroscopy is especially appropriate when there is a rather large amount of a substance and no need to determine minor components of complex mixtures. It is the case of product quality control and authentication of various samples, identification of microorganisms, and so on (Chap. 8).

Each of the large commercial databases (Table 7.16) can be supplied as an integrated package, or one of the special libraries devoted to, for example, polymers and related compounds, organic, inorganic and organometallic compounds, industrial chemicals, compounds of forensic and environmental interest, pesticides, dyes, pigments, coatings, vapors/gases, and so on. Typical spectra are ones recorded by FTIR. Special libraries are or may be originated from different producers/vendors;

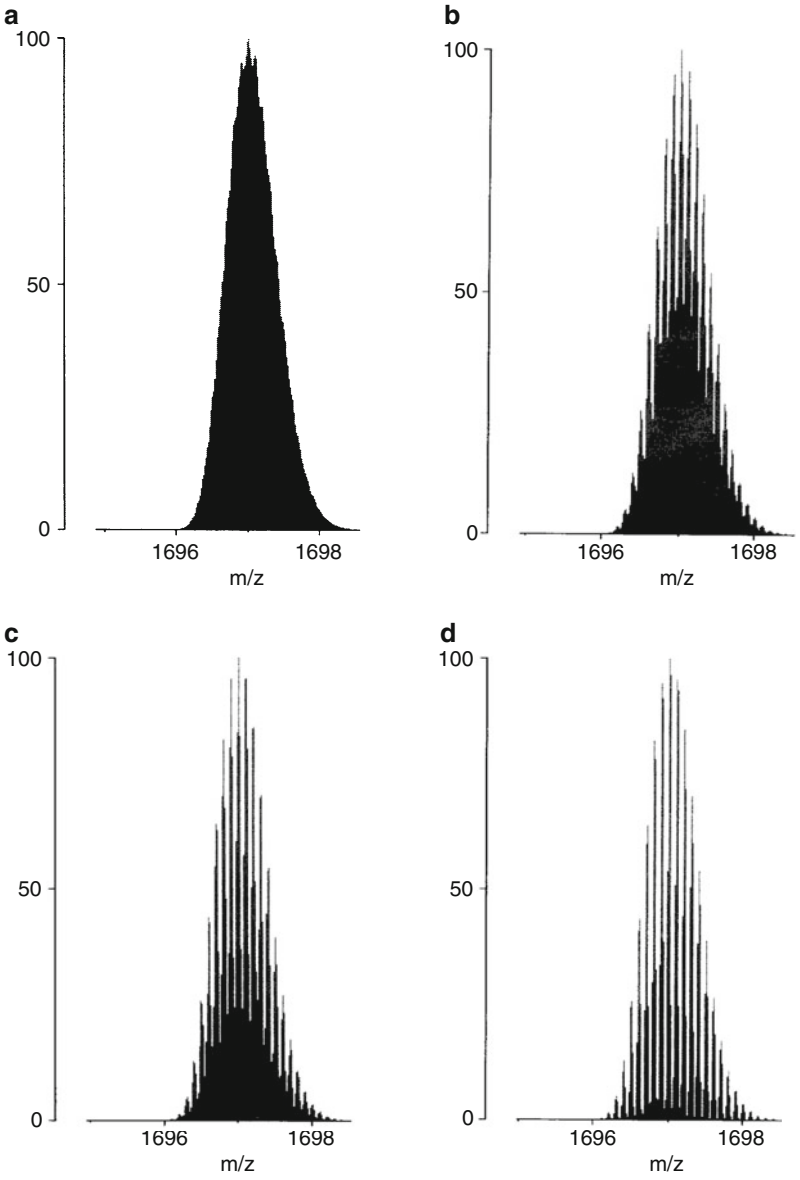



**Fig. 7.9** The peaks view of the $[M+10H]^{10+}$ ion of horse myoglobin ($C_{769}H_{1212}N_{210}O_{218}S_2$, monoisotopic mass 16,940.96) depending on the resolution: (a) 15,000, (b) 25,000, (c) 30,000, and (d) 50,000. The spectrum was generated by the use of the MS-Isotope program [186]. A resolution not less than 50,000 is evidently required for more accurate measurements

**Example 7.6.** This is identification of the component of algal biomass by means of the high-resolution tandem mass spectra of the $[M+2H]^{2+}$ precursor ion (Fig 7.10). With the accurate mass of this ion and the criterion of Delta (ppm) $\leq 3$, all possible formulas were generated (see Table 7.13). Only four formulas and 14 corresponding compounds were found in PubChem (this database, see Table. 6.2). Searches in Internet did not enlarge this number of formulas.

Three-step fragmentations of the $[M+H]^+$ ions of eight compounds were further generated using the Mass Frontier software, version 5.1.0.1. Six isomers of the $C_{45}H_{70}O_{15}$ formula were not taken into account because they did not contain nitrogen atoms, and only molecules of basic (mostly N-containing) compounds form doubly and multicharged ions (multiply protonated molecules). Ion masses of fragments predicted by the program were compared with experimental mass values (see Fig. 7.10). Here, it should be noted that Mass Frontier does not fragment doubly protonated molecules and only handles single-charged ions. Therefore, the comparison of the experimental cleavages in the $[M+2H]^{2+}$ ion with predicted fragmentations of $[M+H]^+$ ions was some kind of speculation.

The highest number of matches, precisely nine from the 15 most abundant single-charged ions in the spectrum (Fig. 7.10), was observed for anabaenopeptin F **7.1**, a compound with the formula $C_{42}H_{62}N_{10}O_9$ which was isolated from cyanobacteria contained in the sample of just the same type [192]. In the case of the second hit, isomeric anabaenopeptin E **7.2**, there were two fewer matches. These refer to the fragments **7.3** and **7.4** predicted only for the precursor **7.1**. It looks very plausible because one can easily see that these ions, **7.3** with sec-butyl group and **7.4** with unsubstituted *p*-hydroxyphenyl, are fragments of **7.1** rather than **7.2**. The same matches as for **7.2** were observed in the case of its isomer, where a methyl substitute of p-hydroxyphenyl group is transferred to oxygen atom. There are far fewer matches for five other compounds considered.

Thus, anabaenopeptin F **7.1** is the most probable result of identification. However, (1) the speculation was done (see above) and (2) it is theoretically possible for a large number of isomers of **7.1** to exist. So a rigorous analyst will consider this identification as a tentative one needing to be confirmed, for example by co-analysis with the reference material.

replicates of the same collections are or may be presented in different integrated libraries.

There are also other large non-commercial spectral collections.

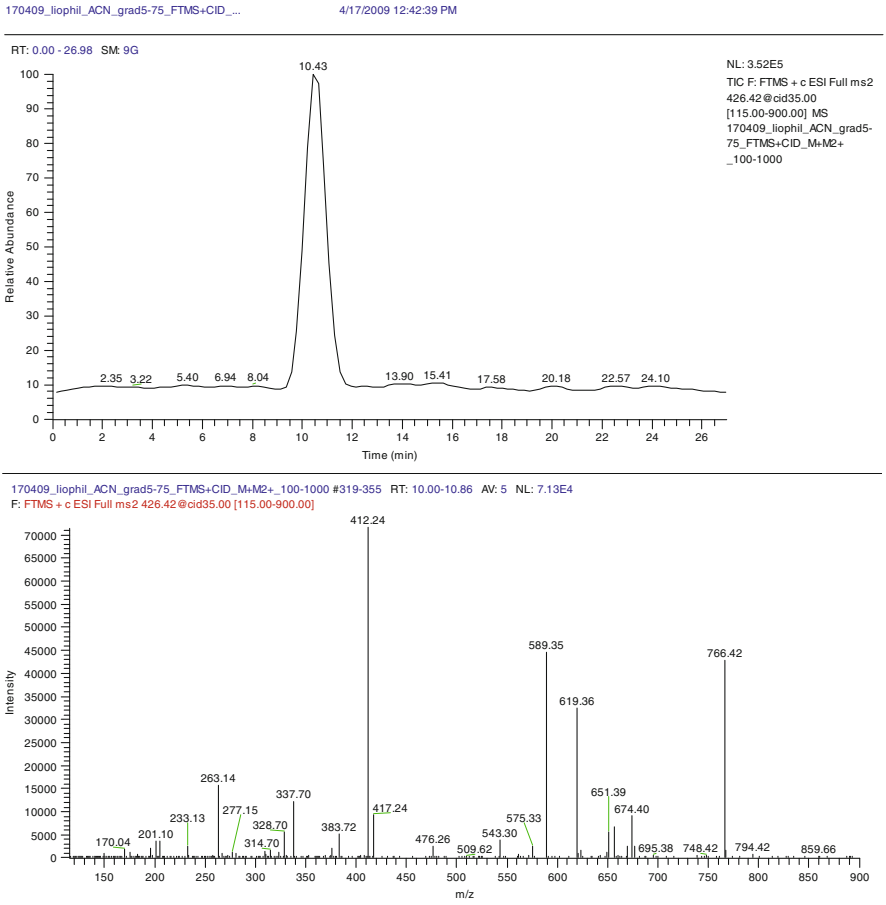



**Fig. 7.10** Mass chromatogram and $MS^2$ spectrum of the component of algal biomass sampled in the Sestroretskij Razliv lake near Saint-Petersburg, Russia (summer, 2008). Mass spectra were acquired on the LIT–Orbitrap high-resolution tandem mass spectrometer. The chromatogram was recorded for the precursor $[M+2H]^{2+}$ doubly charged ion with $m/z$ 426.2425. In this mode of $MS^2$, fragment ions were formed in IT and further separated by accurate $m/z$ and detected in Orbitrap. In the initial ESI–$MS^1$ spectrum, the peak of common $[M+H]^+$ ions was also observed. Corresponding fragments were less characteristic

- Database of National Institute of Advanced Industrial Science and Technology (AIST, Japan) [92], free on-line searches. The database contains about 52,100 FT–IR spectra and 3,500 Raman spectra.
- SpecInfo [96], with access for eligible users. There are 21,000 IR spectra.

Special databases should be also selected:

- Gas/vapor IR spectral collections [207, 208] (vapor spectra of organic compounds are also included in some databases, noted in Table 7.16) which are

**Table 7.16** Large commercial IR spectral libraries [a]

| Name | Spectra | Remarks |
|---|---|---|
| HaveItAll IR [201] | over 233,000 | 88 databases, including NIR and Raman spectra |
| NICODOM IR [202] | 150,758 | Smaller packages/libraries of various vendors, including Raman and NIR collections |
| ACD/IR and Raman [203] | Over 100,000 | 13 libraries, including Raman spectra |
| Nicolet, Aldrich, and related IR libraries [204] | Over 88,000 | Ten libraries, including Raman and high-resolution spectra (4 $cm^{-1}$ resolution) |
| Thermo IR [205] | Over 41,000 | 15 libraries of various vendors. |
| FDM FTIR/Raman [206] | Over 33,000 | 20 databases, including Raman spectra |

[a]May be available also in other collections

characterized by the highest resolution achieved, as compared to spectra for other phase states and intended for atmospheric environmental monitoring and

- The library of spectra of coryneform bacteria: 730 reference strains, covering 220 different species from 46 genera [209]

For the given library, a search result may depend on the search algorithm connected with the particular MF. Various commercial algorithms make it possible to find compounds of similar structure if spectra of unknowns are not available in the library, or reduce the effects of offset or slope in the baseline which increases differences between spectra compared; see Sect. 4.4.4. A high-resolution spectral library improves the spectral match between the unknown sample and library references as compared with common resolution. For example, characteristic shoulders of a stronger absorbing peak may be seen only in a high-resolution library spectrum [210].

In contrast to the above MS libraries, there have not been many reports on performances of librarian searches in IR spectral libraries. The exceptions can be exemplified by (a) the evaluation of an IR spectral library searching for the purpose of identification of automotive paints [211], and (b) performances of the Raman spectral library built for 309 pharmaceutical reference materials [212]. In the second case, a significant fluorescent signal observed for some test samples prevented identification (*TPR* 88–96% depending on the search algorithm). Without those cases, *TPR* was up to 100% [212].

## 7.6 NMR Spectroscopy

This type of spectroscopy, mainly proton and $^{13}C$ NMR, has been traditionally used for structure elucidation of new organic compounds by organic chemists, based on reference tables of chemical shifts and spin–spin coupling constants; see references in Sect. 2.8.3 and [195]. Now, special computer expert systems containing structure generators, spectra predictors (Table 7.17), and other software modules assist

**Table 7.17** Prediction of NMR chemical shifts

| Predictor | Algorithm[a] | Nuclei | Accuracy, ppm[b] |
|---|---|---|---|
| NMRPredict [217, 218] | HOSE code, ANN, functional groups based | $^{1}H$, $^{13}C$, $^{15}N$, $^{19}F$, $^{31}P$, $^{17}O$, $^{29}Si$, 2D | $^{1}H$: down to 0.14 [219] $^{13}C$: 1.4[c] [220] |
| ACD/NMR predictors [216, 221] | additivity rules, ANN, HOSE code | $^{1}H$, $^{13}C$, $^{15}N$, $^{19}F$, $^{31}P$, 2D ($^{1}H$-$^{13}C$) | $^{13}C$: 1.59[c] (ANN) [222], 1.8 (additivity rules, ANN) [216] |
| CSEARCH [218] | HOSE code, ANN | $^{13}C$ | 2.19–2.22[c] |
| Upstream Solutions NMR prediction [223] | Additivity rules | $^{1}H$, $^{13}C$ | $^{1}H$: 0.2–0.3 (90% of $CH_x$ groups), $^{13}C$: 3.8 (> 95% shifts) |
| SpecInfo [96, 224] | Rule-based, database-based | $^{1}H$, $^{13}C$, $^{19}F$, $^{31}P$, $^{17}O$ | |

[a]The authors' names of algorithms are given
[b]Average deviation between calculated and experimental chemical shifts
[c]The NMRShiftDB data set, see Table 7.19, was used to evaluate the prediction accuracy

chemists in setting up, evaluating, and accepting/rejecting identification/structure hypotheses [213].

An initial hypothesis is a known molecular formula or one of the compounds/structures corresponding to this formula. Similar structures, as other candidates for elucidation, can be further generated. For each structure, spectra are predicted and compared with experimental ones. Hypotheses are ranked by the MF (Sect. 4.4.3) for these structures. Compounds having the best MF are the most probable candidates for identification [188, 189, 197, 213–216].

There are several methods for predicting NMR spectra which are based on additivity rules (incremental approach), correlations between structures and spectra, ANN, so called hierarchical organization of spherical environments (HOSE) code, and so on; see references in Table 7.17. The HOSE code algorithm uses data from databases where chemical structures and their NMR spectra are present. Also, quantum mechanical calculations of chemical shifts can be found in the literature [216].

In the right column of Table 7.17, evaluations of the shift prediction accuracy are given; the accuracy for $^{1}H$ chemical shifts of 0.17–0.18 ppm, has also been reported [225]. These are indicators of the predictability efficiency of different algorithms, but not the direct measure of reliability of identification/structure elucidation.

Estimations of identification trueness have been made by matching experimental $^{1}H$ and $^{13}C$ spectra of test structures with theoretical spectra predicted for a set of known structures; structures corresponding to the best matching were considered as true results [226]. Another evaluation method was based on statistics of MF (see Sect. 4.4.3) for experimental and predicted spectra of test structure sets. The low, medium, and high MF were initially referred to as negative, ambitious, and positive results respectively; positives and negatives definitely appeared to be true or false [227–229]. All these rates of prediction are collected in Table 7.18. The rates depended on what techniques, test sets, and prediction algorithms had been used. In most cases, more than 50% of predictions were correct.

**Table 7.18** Evaluations of prediction efficiency

| Reference | Technique | Result |
|---|---|---|
| [227] | $^{1}$H | *St* 97–100%, *Sp* 60–100%, not including 2-50% ambiguous results |
| [226] | $^{1}$H, $^{13}$C | *St* 49–63% and 65–92% (known molecular formula) |
| [228] | $^{1}$H | *St* 91–100%, *Sp* 67–100%, not including 7–17% ambiguous results[a] |
| | $^{1}$H and 2D $^{1}$H-$^{13}$C | *St* 92–100%, *Sp* 94–100%, not including $\leq$ 7% ambiguous results |
| [229] | $^{1}$H | *Sp* 28%, at 10% ambiguous results |
| | $^{1}$H and 2D $^{1}$H-$^{13}$C | *St* 56%, *Sp* 21%, at 23% ambiguous results |

See also [230, 231]
[a]Without results for very small samples

Another approach to identification, direct matching of reference NMR spectra as well as in MS and IR spectroscopy, was made possible with an appearance of the respective databases/libraries (Table 7.19; see also [125, 242]). Special libraries have been built in metabolomics and proteomics (Table 7.19). Together with MS, NMR spectroscopy belongs to the main instrumental techniques of metabolomics (e.g., see [243]). Both platforms can lead to complementary results of metabolite analysis [244] or correlate between each other [245]. The term of *metabonomics* is sometimes used instead of *metabolomics* if just NMR is the analytical technique (e.g., see [244]), although there is some semantic difference between the two concepts [246].

In early evaluation of $^{13}$C NMR libraries, *TPR* was up to 94% [247]. Recently, evaluations were made for identification of metabolites contained in biofluids of complex composition by means of 2D NMR spectrometry, with the use of a specially built library consisting of reference spectra of about 500 compounds and special software [248]. The rate of TP was about and over 80%, which can be evaluated as a high value, given that the analytes were not isolated [248]. A bit earlier, even higher estimates of *TPR* and *TNR* were reported for the same instrumental technique and similar samples [249].

## 7.7 "Omics"

Advances in genetics and gen*omics* have originated a plenty of research fields combined with the "*omics*" suffix in their names and the "*ome*" suffix in the names of corresponding subjects of scientific research.

> The goal of 'omic' approaches is to acquire comprehensive, integrated understanding of biology by studying all biological processes to identify the different players (e.g., genes, RNA, proteins and metabolites) rather than each of those individually [250].

Among different "omics", metabol*omics* (metabol*ome*, the particular study of metabolites; see Sects. 7.7.1 and also 7.4.1.3) and prote*omics* (proteome, the

**Table 7.19** NMR spectral databases

| Name | Spectra/records | Remarks |
|---|---|---|
| SpecInfo [96] | 359,000 $^{13}$C, 130,000 $^{1}$H, 90,000 heteroatoms | Heteroatoms: $^{15}$N, $^{17}$O, $^{19}$F, $^{11}$B, $^{31}$P. Access for eligible users. Prediction of spectra |
| HaveItAll NMR, HaveItAll XNMR [232] | Over 438,000 $^{13}$C, over 51,000 $^{1}$H, 71,000 heteroatoms | Heteroatoms: $^{19}$F, $^{31}$P, $^{15}$N, $^{17}$O, $^{11}$B, $^{29}$Si. Prediction of solvent-specific chemical shifts. Also 1,060 spectra of metabolites and 740 those of monomers and polymers. |
| ACD/NMR [233] | Over 200,000 $^{13}$C, 210,000 $^{1}$H, over 53,000 heteroatoms, and others | Heteroatoms: $^{15}$N, $^{19}$F, and $^{31}$P. Eight databases, including Aldrich NMR library. Prediction of spectra |
| NMRPredict [234] | 465,349 | $^{13}$C, $^{1}$H, $^{19}$F, $^{31}$P, $^{15}$N, $^{17}$O, $^{11}$B, $^{29}$Si NMR. Three subcollections. Prediction of spectra. |
| CSEARCH [235, 236] | 75,000 | $^{13}$C NMR spectra. Access is to be allowed, prediction of spectra. |
| NMRShiftDB [237] | Over 47,000 | 38,802 compounds. $^{13}$C, $^{1}$H, $^{15}$N, $^{31}$P. Free on-line access, prediction of spectra. |
| SDBS [92] | 13,500 $^{13}$C, 15,200 $^{1}$H | Free on-line access |
| MMCD [141] | 20,306 compounds | Metabolites, $^{1}$H and $^{13}$C, including 2D. Free on-line access |
| HMDB [121] | 1,800 compounds[a] | Metabolites, $^{1}$H and $^{13}$C. Free on-line access |
| MDL [238] | | Access for privileged users |
| BMRB[b] [239] | 3,964,515 $^{1}$H, $^{13}$C and $^{15}$N chemical shifts of proteins and peptides and over 54,000 $^{13}$C, $^{1}$H, $^{15}$N, $^{31}$P shifts of nucleic acids | Repository for NMR spectra of biomolecules. Free on-line access |
| RefDB [240, 241] | 2,162 files | Evaluated $^{1}$H, $^{13}$C and $^{15}$N chemical shifts of proteins from BioMagResBank. Free on-line access |

[a]Probably, not all compounds are unique
[b]It was concluded that nearly 40% of protein entries deposited in this data bank had at least one erroneous assignment of the chemical shift [241]

particular study of proteins, Sects. 7.7.2 and also 4.4.2.3, 4.5.4.3, and 7.4.1.4) are from widespread "omics" fields, and directly correlate to issues of chemical identification, even if not fully relevant to the goals of the book. In these and other "omics", there are great analytical challenges relating to:

- The general complexity of the sample
- A vast number of target analytes (hundreds, thousands, tens of thousands in the same samples)
- Large diversity of their structures
- Large dynamic ranges of abundance of different analytes, up to $10^{8}$–$10^{12}$
- Very low concentrations of many biocompounds ($\leq$ 1 pg/l $\div$ 1 ng/l)
- Their rapid transformations in living systems, and so on [250].

### 7.7.1 Metabolomics

This is the new field of bioanalytical and biological research [180, 243, 250–252]. Corresponding definitions, including these of analytical procedures, are given in Table 7.20. Those can be accomplished with the quotation:

> Metabolomics can be used for two major and very different purposes: the screening for differences between global metabolic fingerprints of cohorts of populations, which is often referred to as metabonomics, or efforts to understand the regulatory structure of metabolic pathways, its connectivity, control of cellular concentrations and fluxes of metabolites, and partitioning of metabolic products between cellular compartments and excretion [180].

Table 7.20 shows that there are three principal groups of analytical procedures in metabolomics: targeted analysis, metabolic profiling, and metabolic fingerprinting. The first two are related to identification of individual compounds. Most analyses are performed with GC–MS, LC–$MS^n$ ($HRMS^n$), and NMR techniques [180, 243, 250–252]. GC–MS is mainly used for plant metabolite analysis, e.g., see [125, 126]. In this field of metabolomics, derivatization of metabolites is often needed to obtain volatile and thermostable compounds. NMR spectroscopy (Sect. 7.6) has been used for metabolic profiling, e.g., that of physiological fluids (serum, urine) though the sensitivity of this technique is not the highest. This limitation also applies to IR spectroscopy. The technique of CE–MS can be used instead of LC–MS for determination of polar (ionizable) metabolites [85].

In general, identification of components of very complicated biochemical samples is a prerequisite for solving the inherent problems of metabolomics. Different types and subtypes of qualitative analysis (see Chap.1) as well as outcomes of identification operations may be:

**Table 7.20** Definitions related to metabolomics (adapted from [243, 251, 252])

| Concept | Definition |
|---|---|
| Metabolome | The complete collection of metabolites produced by cells[a] present in an organism |
| Metabolomics | The analysis and the study of metabolome |
| Metabolite target analysis | Determination of one or a few selected metabolites |
| Metabolite profiling | Determination of a relatively large number of metabolites resulting in a (graphic) form of biochemical profiles commonly recorded by combined separation and detection techniques (mainly, chromatography mass spectrometry) |
| Metabolic fingerprinting | Rapid, global analysis of biological samples based on their patterns/ "fingerprints", commonly without separation into fractions, to classify samples according to their origin, state (physiological and disease states), and so on |
| Metabolic footprinting | Fingerprinting referring to extracellular metabolites |

[a]There are also metabolites of foreign compounds (xenobiotics)

- Identification of known compounds
- Structure elucidation of new ones (named *de novo* identification in the literature on “omics”)
- Unambiguous, tentative, and group/class identification
- Ultimate identification failures

All the items are very common for chemical analysis of low-molecular compounds, though considered [180] as something attributive to this field.

Two strategies of qualitative analysis have been proposed [180] (see also Example 7.5 above).

- The start from HRMS determining accurate ion masses and isotope ratios, then generation of candidate molecular formulas and search for candidate compounds in chemical databases.
- The start from searches in MS libraries.

Ultimately, both lead to profiles of metabolite samples annotated with identifiers of chemical compounds: unambitious identifiers (see Chap. 1) or even retention parameters and mass spectra. In this context, metabolomists pull together concepts of *identification* and *annotation*.

Groups of metabolite analytical signals such as chromatographic profiles or integral MS/NMR/IR spectra of biosamples act as fingerprints in metabolic fingerprinting [137, 253–255], which is one of the numerous methodologies of quantitative analysis II (Chap. 8).

Structures of metabolites produced in living organisms can be predicted based on metabolic reaction rules (Table 6.1). This can facilitate identification of compounds under consideration.

An ambitious goal of metabolomists aimed at global determination of a metabolome seems not to be very attainable.

> Metabolic profiling (sometimes referred to as untargeted analysis or metabolite profiling) provides a more or less holistic study of a metabolome with detection of hundreds or thousands of metabolites... Although metabolic profiling has been described as unbiased and global, in reality all methods of sample preparation and all analytical platforms introduce a level of chemical bias [36].

A complexity of a metabolome, poor reproducibility of analytical signals of metabolite traces, and an absence of analytical standards for many metabolome components are important factors precluding true metabolic results. However, many new information technologies and analytical methodologies have been developed within metabolomics (as well as proteomics, see below) which are very useful in respect of general progress of analytical chemistry.

### *7.7.2 Proteomics*

By analogy with metabolomics (Table 7.20), proteomics is defined as the analysis and the study of proteomes. In turn, a proteome is a set of proteins produced by

a genome present in cell, tissue, or organism. Researchers begun to study proteomes earlier than metabolomes, being guided by considerations such as the following ones.

> Proteomics can be viewed as an experimental approach to explain the information contained in genomic sequences in terms of the structure, function, and control of biological processes and pathways. Proteomics attempts to study biological processes comprehensively by the systematic analysis of the proteins expressed in a cell or tissue. Mass spectrometry (MS) is currently proteomic's most important tool [155].

Identification of proteins holds a central position in proteomics [155, 156]. In Sects. 4.4.2.3 and 4.5.4.3, two main versions of the MS approach (the so-called *bottom-up* one) for protein identification are treated:

- Peptide mass fingerprinting
- Peptide fragment mass fingerprinting

Both subapproaches are based on matching an experimental mass spectrum with theoretical ones which are generated from all amino acid sequences contained in special data banks. Criteria for identification are related to high score values for the sequence from peptides/proteins, low scores for all other candidate sequences, and low probabilities of a random match for an identified peptide/protein (see Sects. 4.4.2.3 and 4.5.4.3).

There are also other general approaches to MS identification of amino acid sequences.

- *De novo* sequencing [155, 256] where fragment ion spectra are interpreted according to the rules/regularities derived from studies of fragmentation of protonated peptide molecules; see Fig. 4.8 for types of fragments.
- A version of the method of peptide fragment mass fingerprinting, with reference tandem spectra retrieved from corresponding libraries (Sect. 7.4.1.4) rather than generated from amino acid sequences.
- "Top-down" identification based initially on the mass analysis of intact protein ions (see Sect. 7.4.2).
- Hybrid approaches; see [166].

It should be noted once again that decisions on trueness of protein identification by existing techniques have been made basing on statistical estimations of significance of spectral matching (Sect. 4.5.4.3). The criterion of at least two peptide matches per protein should also be applied (e.g., see [108, 257, 258]). In establishing the strongest rules for protein identification, some other criteria expressing the spectral quality have been added [108]:

- Identified peptides should be sufficiently long (length at least seven amino acid residues)
- Peptide fragments are predictable (at least 30% of *b*/*y* ions, see Fig. 4.8)
- Peptides are fully tryptic, i.e., produced by the particular rules of protein cleavage when trypsin is used; see Sect. 4.4.2.3

These and other similar rules [163] have been set for identification of peptides preceding the inclusion of their spectra in MS libraries.

Top quality identification with the use of reference materials (Table 1.4) widespread in qualitative analysis of small molecules is practically unattainable in the case of most high molecules, due to a vast number of theoretically possible amino acid sequences and therefore unavailability of corresponding references. Without such co-analytical data, a decrease of FP rates can be achieved by a combination of

- Different identification approaches
- Not the same algorithms of searches in databases and subsequent match scoring
- Data from different ionization techniques, precursor ions, or $MS^n$ scans (different $n$)
- Analytical results obtained for different pieces of the same protein
- MS data and auxiliary information,

and so on [166, 259–261]. In this context, retention parameters (Sect. 7.3.4) are considered as an important new type of auxiliary data.

The potential of every method/approach/algorithm is evaluated by interlaboratory comparisons in a conclusive way. Some comparisons are shown next.

- Two MS methods based on platforms of MALDI–IT–$MS^2$ and ESI–ToF–$HRMS^2$ were compared in respect of the analysis of protein samples isolated by the common technique of 2D gel electrophoresis. A reliable identification, i.e., one performed by both methods, was observed for 85 from the total of 128 identified proteins (66%) [257].
- Comparisons were carried between several search algorithms for matching spectra, using the same test subset of $MS^2$ spectra and the estimation of true and false result rates. Peptides were conventionally considered correctly identified if the scores were at the 1st rank hit and fell in the range of low probability for random match; see Fig. 4.10. Out of 608 peptides correctly identified by at least one algorithm, only 335 peptides (55%) were recognized by all four algorithms. Mascot (Fig. 4.10) was the most efficient algorithm/software [262].
- Establishing the most rigorous identification criterion for proteins, with two or more matching distinct peptides instead of one or more peptides, led to 37% decrease in the number of discovered proteins. Again, only 55–57% of proteins were identified by all algorithms. Also, there may be little difference in the output of the popular search algorithms if improved and consistent scoring methodology is used [258].
- In an interlaboratory study of reliability of protein identification, 120 laboratories requested the standard mixture of 49 human proteins. Seventy four laboratories (62%) reported identification results. On the average, 60% of results (estimated by the author) from those laboratories were TP. The conclusion was reached that “success was possibly experience- or technical ability-dependent” [263].
- The study was repeated for another mixture containing phosphorylated and non-phosphorylated proteins. Problems in relation to phosphorylation site

identification, laboratory evaluation, and creation of standard materials were highlighted [264].

- One more interlaboratory study of that sort was carried out by another group. The test sample consisting of 20 human proteins was distributed to 27 laboratories for identification. Initially, only seven participants reported true identification of all 20 proteins. Nevertheless, all the proteins had in fact been detected in all the participating laboratories, but were not reported. The sources of such FN were noted to be problems that laboratories had in regard to database searches and spectral data matching [265].

The three last studies are of special value because their results were evaluated in conditions of traceability to the known identity of original proteins.

On the whole, the comparative experiments showed that there are many laboratories with the potential of true identification of proteins. Nevertheless, algorithms of MS identification by ion mass fingerprinting are still imperfect, and may be hard to adopt in laboratories. The following conclusion belongs to proteomists themselves.

> Despite the high-mass accuracy of modern mass spectrometers, the general perception of the reliability of MS-based proteomics is that it is low [265].

Nevertheless, the progress in proteomics directly or indirectly affects developments in a general methodology of analytical chemistry.

## 7.8 Comparison of Spectral Techniques

In this section, brief conclusions will be reached regarding the efficiency of different spectral techniques in unknown/non-target analysis. General evaluations of different constituents and features of analysis are given in Table 7.21; particular comparisons of potentialities of various techniques can be found in the literature, e.g., see [253].

General conclusions can be divided into two parts, differing according to the kind of sample.

- A sample contains a complex mixture of analytes being present in low amounts. Combinations of chromatography with mass spectrometry are indispensable. Either (a) GC/LC and at least one MS technique or (b) several different MS techniques should be applied to achieve reliable results for identification.
- A sample of relatively simple composition: the only or a few analytes being present in rather large amounts. Another case is characterization of just a sample rather than individual analytes. All three techniques under discussion are comparable in overall performance and should be used in a combination with each other or a different technique to obtain reliable identification results.

**Table 7.21** Relative potentialities of identification with different techniques[a]

| Item | MS | NMR | IR |
|---|---|---|---|
| Sample preparation[b] | From + to +++ | +++ | +++ |
| Sensitivity[c] | +++ | From + to ++ | From + to ++ |
| Compatibility with chromatography | +++ | + | + |
| Cost[d] | | | |
| Instrument | $$$ | $$$ | $ |
| Sample analysis | From $ to $$$ | From $ to $$ | $ |
| Software, databases | Comparable | Comparable | Comparable |
| Throughput of samples | From + to ++ | From + to ++ | +++ |
| Approaches to identification[e] | | | |
| Co-analysis | No or +++ | No or +++ | No or +++ |
| Comparison to library | From ++ to +++ | ++ | ++ |
| Comparison to calculation/predicted data | From + to ++ | From ++ to +++ | + |
| Spectral interpretation | ++ | +++ | ++ |

[a]Interpretation of symbols unless otherwise stated: +++ high, ++ medium, + low
[b]Symbols of +++ and + specify that common analytical practice require no or minimum sample treatment (e.g., dissolution) and numerous preparative operations respectively
[c]A high sensitivity corresponds to a low detection/identification limit and v.v
[d]Interpretation of symbols: $$$ high, $$ medium, $ low
[e]Symbols specify a high +++, medium ++, and low + usability/efficiency of the approach. The absence of corresponding analytical standards limits the use of co-analysis

Independently of the particular technique(s), a confirmation of identity by subsequent co-analysis (Table 1.4, Sect. 5.2) is required in crucial analyses, e.g., following mass poisoning or disaster.

## References


1. De Zeeuw RA, Franke JP (2000) 'General unknown' analysis. In: Smith RM (ed) Handbook of analytical separations, vol 2. Elsevier, Amsterdam, pp 567–599
2. Rivier L (2003) Criteria for the identification of compounds by liquid chromatography–mass spectrometry and liquid chromatography–multiple mass spectrometry in forensic toxicology and doping analysis. Anal Chim Acta 492:69–82
3. Richardson SD (2001) Mass spectrometry in environmental sciences. Chem Rev 101:211–254
4. Waste requiring special processing. http://www.dehs.umn.edu/hazwaste_chemwaste_umn_cwmgbk_sec5.htm#asoebocceb. Accessed 27 May 2010
5. García-Reyes JF, Hernando MD, Molina-Díaz A, Fernández-Alba AR (2007) Comprehensive screening of target, non-target and unknown pesticides in food by LC-TOF-MS. Trends Anal Chem 26:828–841
6. Ojanperä S (2008) Drug analysis without primary reference standards. Application of LC-TOFMS and LC-CLND to biofluids and seized material**.** Dissertation, University of Helsinki. https://oa.doria.fi/bitstream/handle/10024/42995/danalysi.pdf?sequence=2. Accessed 27 May 2010
7. Kinton VR, Pfannkoch EA, Whitecavage JA, Thorp J (2003) Coupling retention time locked methods and libraries to automated SPME or SBSE for analysis of flavors and fragrances.

Gerstel Application Note 7. http://www.gerstel.de/pdf/p-gc-an-2003-07.pdf. Accessed 27 May 2010
8. Tarján G, Nyiredy S, Györ M, Lombosi ER, Lombosi TS, Budahegyi MV, Mészáros SY, Takács JM (1989) Thirtieth anniversary of the retention index according to Kováts in gas–liquid chromatography. J Chromatogr A 472:1–92
9. Gonzales FR, Nardillo AM (1999) Retention index in temperature-programmed gas chromatography. J Chromatogr A 842:29–49
10. NIST Chemistry WebBook. http://webbook.nist.gov/chemistry. Accessed 23 May 2010
11. Castello G (1999) Retention index systems: alternatives to the *n*-alkanes as calibration standards. J Chromatogr A 842:51–64
12. Babushok VI, Linstrom PJ, Reed JJ, Zenkevich IG, Brown RL, Mallard WG, Stein SE (2007) Development of a database of gas chromatographic retention properties of organic compounds. J Chromatogr A 1157:414–421
13. NIST/EPA/NIH Mass Spectral Library with Search Program: (Data Version: NIST 08, Software Version 2.0f). http://www.nist.gov/data/nist1a.htm. Accessed 3 Nov 2010
14. The Sadtler standard gas chromatography retention index library (1985) Sadtler Research Laboratories, Philadelphia
15. Richmond R (1997) Database of structures and their gas chromatography retention indices, tagged with individual search windows. J Chromatogr A 758:319–323
16. Bogoslovsky YN, Anvaer BN, Vigdergaus MS (1978) Chromatographic constants in gas chromatography–hydrocarbons and O-containing compounds (in Russian). Standards Publisher, Moscow
17. LRI and Odour database. http://www.odour.org.uk. Accessed 28 May 2010
18. ESO 2000 (update 2006). http://www.leffingwell.com/baciseso.htm. Accessed 28 May 2010
19. RI essential oil components (in Russian). http://viness.narod.ru/ret_ind.htm. Accessed 28 May 2010
20. Mondello L (2008) FFNSC 1.3 – Flavors and fragrances of natural and synthetic compounds – Mass spectral database. http://www.chromaleont.it/site/index.php?option=com_content&view=article&id=3&lang=en. Accessed 3 Nov 2010
21. Flavornet. http://www.flavornet.org/flavornet.html. Accessed 28 May 2010
22. König WA, Joulain D, Hochmuth DH. Terpenoids and related constituents of essential oils. http://massfinder.com/wiki/Terpenoids_Library. Accessed 28 May 2010
23. Adams RP (2007) Identification of essential oil components by gas chromatography/mass spectrometry, 4th edn. Allured Publishing Corporation, Carol Stream
24. Jennings W, Shibamoto T (1980) Qualitative analysis of flavour and fragrance volatiles by glass capillary gas chromatography. Academic, London
25. Pherobase Kovats retention index of organic compounds. http://www.pherobase.com/database/kovats/kovats-index.php. Accessed 28 May 2010
26. GMD. The Mass spectral (MS) and retention time index (RI) libraries. http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/msri/gmd_msri.html#mtop. Accessed 28 May 2010
27. Wagner C, Sefkow M, Kopka J (2003) Construction and application of a mass spectral and retention time index database generated from plant GC/EI–TOF–MS metabolite profiles. Phytochemistry 62:887–900
28. Schauer N, Steinhauser D, Strelkov S, Schomburg D, Allison G, Moritz T, Lundgren K, Roessner-Tunali U, Forbes MG, Willmitzer L, Fernie AR, Kopka J (2005) GC–MS libraries for the rapid identification of metabolites in complex biological samples. FEBS Lett 579:1332–1337
29. FiehnLib. http://fiehnlab.ucdavis.edu/projects/FiehnLib/index_html. Accessed 28 May 2010
30. Kind T, Wohlgemuth G, Lee DY, Lu Y, Palazoglu M, Shahbaz S, Fiehn O (2009) FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. Anal Chem 81:10038–10048

31. Maurer HH, Pfleger K, Weber AA (2007) Mass spectral and GC data of drugs, poisons, pesticides, pollutants and their metabolites. http://www.wiley-vch.de/publish/en/books/ISBN978-3-527-31538-3. Accessed 3 Nov 2010
32. Rösner P (2010) Mass spectra of designer drugs. http://www.sisweb.com/software/ms/wiley.htm#designerdrugs. Accessed 28 May 2010
33. Franke JP, Bogusz M, De Zeeuw RA (1993) An overview on the standardization of chromatographic methods for screening analysis in toxicology by means of retention indices and secondary standards. Fresenius J Anal Chem 347:67–72
34. Gas chromatographic retention indices of toxicologically relevant substances on packed or capillary columns with dimethylsilicone stationary phases (1992) Report XVIII of the DFG Commission for Clinical-Toxicological Analysis, 3rd edn. VCH, Weinheim
35. Zellner BA, Bicchi C, Dugo P, Rubiolo P, Dugo G, Mondello L (2008) Linear retention indices in gas chromatographic analysis: a review. Flavour Fragr J 23:297–314
36. Brown M, Dunn WB, Dobson P, Patel Y, Winder CL, Francis-McIntyre S, Begley P, Carroll K, Broadhurst D, Tseng A, Swainston N, Spasic I, Goodacre R, Kell DB (2009) Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. Analyst 134:1322–1332
37. NIST Mass Spectral Search Program, version 2.0d, and NIST/EPA/NIH Mass Spectral Library (2005)
38. Babushok VI, Zenkevich IG (2009) Retention indices for most frequently reported essential oil compounds in GC. Chromatography 69:257–269
39. Milman BL, Kovrizhnych MA (2000) Identification of chemical substances by testing and screening of hypotheses. II. Determination of impurities in n-hexane and naphthalene. Fresenius J Anal Chem 367:629–634
40. Valcárcel M, Cárdenas S, Barceló D, Buydens L, Heydorn K, Karlberg B, Klemm K, Lendl B, Milman B, Neidhart B, Ríos A, Stephany R, Townshend A, Zschunke A (2002) Metrology of qualitative chemical analysis. Report EUR 20605. EC, Luxembourg
41. Richmond R, Pombo-Villar E (1997) Gas chromatography–mass spectrometry coupled with pseudo-Sadtler retention indices, for the identification of components in the essential oil of *Curcuma longa* L. J Chromatogr A 760:303–308
42. Wang YH, Wong PK (2003) Correlation relationships between physico-chemical properties and gas chromatographic retention index of polychlorinated-dibenzofurans. Chemosphere 50:499–505
43. Héberger K (2007) Quantitative structure–(chromatographic) retention relationships. J Chromatogr A 1158:273–305
44. Buryak AK (2002) The use of molecular–statistical methods for the calculation of thermodynamic characteristics of adsorption for identification of organic compounds by gas chromatography–mass spectrometry. Russ Chem Rev 71:695–706
45. Milman BL (2008) Introduction to chemical identification (In Russian). VVM, Saint Petersburg
46. Ruther J (2000) Retention index database for identification of general green leaf volatiles in plants by coupled capillary gas chromatography–mass spectrometry. J Chromatogr A 890:313–319
47. Steward EM, Pitzer EW (1988) Gas chromatographic analyses of complex hydrocarbon mixtures void of n-paraffin retention index markers using joint mass spectral and retention index libraries. J Chromatogr Sci 26:218–222
48. Mondello L, Dugo P, Basile A, Dugo G, Bartle KD (1995) Interactive use of linear retention indices, on polar and apolar columns, with a ms-library for reliable identification of complex mixtures. J Microcol Sep 7:581–591
49. Lucero M, Estell R, Tellez M, Fredrickson E (2009) A retention index calculator simplifies identification of plant volatile organic compounds. Phytochem Anal 20:378–384

50. Bianchi F, Careri M, Mangia A, Musci M (2007) Retention indices in the analysis of food aroma volatile compounds in temperature-programmed gas chromatography: database creation and evaluation of precision and robustness. J Sep Sci 30:563–572
51. Milman BL, Konopelko LA (2000) Identification of chemical substances by testing and screening of hypotheses. I. General. Fresenius J Anal Chem 367:621–628
52. Milman BL (2002) A procedure for decreasing uncertainty in the identification of chemical compounds based on their literature citation and cocitation. Two case studies. Anal Chem 74:1484–1492
53. Shellie R, Marriott P, Zappia G, Mondello L, Dugo G (2003) Interactive use of linear retention indices on polar and apolar columns with an MS-library for reliable characterization of Australian tea tree and other *Melaleuca* sp oils. J Essent Oil Res 15:305–312
54. Bio-Rad/KnowItAll HaveItAll UV-Vis. http://www.knowitall.com/literature/docs/96331-Bio-Rad_HaveItAll_UV-Vis_Spectral_Database.pdf#zoom=90%. Accessed 4 June 2010
55. Science-softCon UV/Vis$^{+}$ Spectra Data Base (2010). http://www.science-softcon.de/software-e.htm#2010. Accessed 28 May 2010
56. Bakdash A, Herzler M, Herre S, Erxleben BT, Rothe M, Pragst F. The HPLC–DAD Data Base. UV spectra of pharmaceuticals and toxic compounds. http://pharmascops-sy.org/PDF%20Files/UV%20Library.pdf. Accessed 28 May 2010
57. The Combined Chemical Dictionary on DVD. http://www.crcpress.com/product/isbn/9780412820205. Accessed 23 May 2010
58. UV/Vis Spectral Data. http://chemistry.library.wisc.edu/subject-guides/spectroscopy.html. Accessed 28 May 2010
59. Bogusz M, Wu M (1991) Standardized HPLC/DAD system, based on retention indices and spectral library, applicable for systematic toxicological screening. J Anal Toxicol 15:188–197
60. Bogusz M, Franke JP, De Zeeuw RA, Erkens M (1993) An overview on the standardization of chromatographic methods for screening analysis in toxicology by means of retention indices and secondary standards. Fresenius J Anal Chem 347:73–81
61. Bogusz M, Erkens M (1994) Reversed-phase high-performance liquid chromatographic database of retention indices and UV spectra of toxicologically relevant substances and its interlaboratory use. J Chromatogr A 674:97–126
62. Bogusz M, Hill DW, Rehorek A (1996) Comparability of RP–HPLC retention indices of drugs in three databases. J Liq Chromatogr Relat Technol 19:1291–1316
63. Stoll DR, Paek C, Carr PW (2006) Fast gradient elution reversed-phase high-performance liquid chromatography with diode-array detection as a high-throughput screening method for drugs of abuse. I. Chromatographic conditions. J Chromatogr A 1137:153–162
64. Porter SEG, Stoll DR, Paek C, Rutan SC, Carr PW (2006) Fast gradient elution reversed-phase high-performance liquid chromatography with diode-array detection as a high-throughput screening method for drugs of abuse. II. Data Analysis. J Chromatogr A 1137:163–172
65. Maier RD, Bogusz M (1995) Identification power of a standardized HPLC-DAD system for systematic toxicological analysis. J Anal Toxicol 19:79–83
66. Nielsen KF, Smedsgaard J (2003) Fungal metabolite screening: database of 474 mycotoxins and fungal metabolites for dereplication by standardised liquid chromatography–UV–mass spectrometry methodology. J Chromatogr A 1002:111–136
67. Herzler M, Herre S, Pragst F (2003) Selectivity of substance identification by HPLC–DAD in toxicological analysis using a UV spectra library of 2682 compounds. J Anal Toxicol 27:233–242
68. Albaugh DR, Hall LM, Hill DW, Kertesz TM, Parham M, Hall LH, Grant DF (2009) Prediction of HPLC retention index using artificial neural networks and I-Group E-state indices. J Chem Inf Model 49:788–799
69. Shinoda K, Sugimoto M, Tomita M, Ishihama Y (2008) Informatics for peptide retention properties in proteomic LC–MS. Proteomics 8:787–798

70. Bączek T, Kaliszan R (2009) Predictions of peptides' retention times in reversed-phase liquid chromatography as a new supportive tool to improve protein identification in proteomics. Proteomics 9:835–847
71. Petritis K, Kangas LJ, Yan B, Monroe ME, Strittmatter EF, Qian WJ, Adkins JN, Moore RJ, Xu Y, Lipton MS, Camp DG 2, Smith RD (2006) Improved peptide elution time prediction for reversed-phase liquid chromatography–MS by incorporating peptide sequence information. Anal Chem 78:5026–39
72. Mason CJ, Johnson KL, Muddiman DC (2005) Reproducibility of retention time using a splitless nanoLC coupled to an ESI–FTICR mass spectrometer. J Biomol Tech 16:412–420
73. rt. http://www.ms-utils.org/rt.html. Accessed 29 May 2010
74. Sequence Specific Retention Calculator. http://hs2.proteome.ca/SSRCalc/SSRCalc.html. Accessed 29 May 2010
75. Pfeifer N, Leinenbach A, Huber CG, Kohlbacher O (2007) Statistical learning of peptide retention behavior in chromatographic separations: a new kernel-based approach for computational proteomics. BMC Bioinformatics 8:468. doi:10.1186/1471-2105-8-468
76. Klammer AA, Yi X, MacCoss MJ, Noble WS (2007) Improving tandem mass spectrum identification using peptide retention time prediction across diverse chromatography conditions. Anal Chem 79:6111–6118
77. Pfeifer N, Leinenbach A, Huber CG, Kohlbacher O (2009) Improving peptide identification in proteome analysis by a two-dimensional retention time filtering approach. J Proteome Res 8:4109–4115
78. Xu H, Yang L, Freitas MA (2008) A robust linear regression based algorithm for automated evaluation of peptide identifications from shotgun proteomics by use of reversed-phase liquid chromatography retention time. BMC Bioinformatics 9:347. doi:10.1186/1471-2105-9-347
79. Joutovsky A, Hadzi-Nesic J, Nardi MA (2004) HPLC retention time as a diagnostic tool for hemoglobin variants and hemoglobinopathies: a study of 60 000 samples in a clinical diagnostic laboratory. Clin Chem 50:1736–1747
80. Boone CM, Ensing K (2003) Is capillary electrophoresis a method of choice for systematic toxicological analysis? Clin Chem Lab Med 41:773–781
81. Muijselaar PG (1997) Retention indices in micellar electrokinetic chromatography. Chromatogr A 780:117–127
82. Hudson JC, Golin M, Malcolm M, Whiting CF (1998) Capillary zone electrophoresis in a comprehensive screen for drugs of forensic interest in whole blood: an update. Can Soc Forensic Sci J 31:1–29
83. Boone CM, Franke JP, De Zeeuw RA, Ensing K (2000) Intra- and interinstrument reproducibility of migration parameters in capillary electrophoresis for substance identification in systematic toxicological analysis. Electrophoresis 21:1545–1551
84. Boone CM, Manetto G, Tagliaro F, Waterval JCM, Underberg WJM, Franke JP, De Zeeuw RA, Ensing K (2002) Interlaboratory reproducibility of mobility parameters in capillary electrophoresis for substance identification in systematic toxicological analysis. Electrophoresis 23:67–73
85. Ramautar R, Somsen GW, De Jong GJ (2009) CE–MS in metabolomics. Electrophoresis 30:276–291
86. Sugimoto M, Kikuchi S, Arita M, Soga T, Nishioka T, Tomita M (2005) Large-scale prediction of cationic metabolite identity and migration time in capillary electrophoresis mass spectrometry using artificial neural networks. Anal Chem 77:78–84
87. Nesbitt CA, Zhang H, Yeung KKC (2008) Recent applications of capillary electrophoresis–mass spectrometry (CE–MS): CE performing functions beyond separation. Anal Chim Acta 627:3–24
88. García-Villalba R, León C, Dinelli G, Segura-Carretero A, Fernández-Gutiérrez A, Garcia-Cañas V, Cifuentes A (2008) Comparative metabolomic study of transgenic

versus conventional soybean using capillary electrophoresis-time-of-flight mass spectrometry. J Chromatogr A 1195:164–173
89. Lee R, Ptolemy AS, Niewczas L, Britz-McKibbin P (2007) Integrative metabolomics for characterizing unknown low-abundance metabolites by capillary electrophoresis–mass spectrometry with computer simulations. Anal Chem 79:403–415
90. Wiley: All Titles in Mass Spectrometry. http://eu.wiley.com/WileyCDA/Section/id-350204.html. Accessed 29 May 2010
91. Bio-Rad/KnowItAll HaveItAll MS. http://www.knowitall.com/literature/docs/95381-HIA_MS_DS.pdf#zoom=90%. Accessed 4 June 2010
92. AIST Spectral Database for Organic Compounds (SDBS). http://riodb01.ibase.aist.go.jp/sdbs/cgi-bin/cre_index.cgi. Accessed 5 June 2010
93. AAFS Mass Spectrometry Database Committee. http://www.ualberta.ca/~gjones/mslib.htm. Accessed 29 May 2010
94. MSSJ MassBank. http://www.mssj.jp. Accessed 29 May 2010
95. Sparkman OD (2009) A review of electronic mass spectral databases from John Wiley and Sons. J Am Soc Mass Spectrom 20:R22–R27
96. SpecInfo. http://cds.dl.ac.uk/cds/datasets/spec/specinfo/specinfo.html. Accessed 29 May 2010
97. McLafferty FW, Stauffer DA, Loh SY, Wesdemiotis C (1999) Unknown identification using reference mass spectra. Quality evaluation of databases. J Am Soc Mass Spectrom 10:1229–1240
98. Mass Frontier. http://www.highchem.com/massfrontier/mass-frontier.html. Accessed 29 May 2010
99. Luedemann A, Strassburg K, Erban A, Kopka J (2008) TagFinder for the quantitative analysis of gas chromatography–mass spectrometry (GC–MS)-based metabolite profiling experiments. Bioinformatics 24:732–737
100. Ausloos P, Clifton CL, Lias SG, Mikaya AI, Stein SE, Tchekhovskoi DV, Sparkman OD, Zaikin V, Zhu D (1999) The critical evaluation of a comprehensive mass spectral library. J Am Soc Mass Spectrom 10:287–299
101. McLafferty FW, Tureĉek F (1993) Interpretation of mass spectra. University Science Book, Sausalito, CA
102. Speck DD, Venkataraghavan R, McLafferty FW (1978) A quality index for reference mass spectra. Org Mass Spectrom 13:209–213
103. Stein SE, Scott DR (1994) Optimization and testing of mass spectral library search algorithms for compound identification. J Am Soc Mass Spectrom 5:859–866
104. McLafferty FW, Zhang MY, Stauffer DB, Loh SY (1998) Comparison of algorithms and databases for matching unknown mass spectra. J Am Soc Mass Spectrom 9:92–95
105. Silva-Wilkinson RA, Burkhard LP, Sheedy BR, DeGraeve GM, Lordo RA (1999) A simple comparison of mass spectral search results and implications for environmental screening analyses. Arch Environ Contam Toxicol 36:109–114
106. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27:861–874
107. Oberacher H, Pavlic M, Libiseller K, Schubert B, Sulyok M, Schuhmacher R, Csaszar E, Köfeler HC (2009) On the inter-instrument and the inter-laboratory transferability of a tandem mass spectral reference library: 2. Optimization and characterization of the search algorithm. J Mass Spectrom 44:494–502
108. Frewen BE, Merrihew GE, Wu CC, Noble WS, MacCoss MJ (2006) Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. Anal Chem 78:5678–5684
109. Milman BL (2005) Identification of chemical compounds. Trends Anal Chem 24:493–508
110. Rosal C, Betowski D, Romano J, Neukom J, Wesolowski D, Zintek L (2009) The development and inter-laboratory verification of LC–MS libraries for organic chemicals of environmental concern. Talanta 79:810–817
111. Baumann C, Cintora MA, Eichler M, Lifante E, Cooke M, Przyborowska A, Halket JM (2000) A library of atmospheric pressure ionization daughter ion mass spectra based on

wideband excitation in an ion trap mass spectrometer. Rapid Commun Mass Spectrom 14:349–356
112. Institute of Legal Medicine, University of Freiburg. http://www.chemicalsoft.de. Accessed 31 May 2010
113. Dresen S, Kempf J, Weinmann W (2006) Electrospray-ionization MS/MS library of drugs as database for method development and drug identification. Forensic Sci Int 161:86–91
114. Dresen S, Gergov M, Politi L, Halter C, Weinmann W (2009) ESI-MS/MS library of 1, 253 compounds for application in forensic and clinical toxicology. Anal Bioanal Chem 395:2521–2526
115. Liu HC, Liu RH, Lin DL, Ho HO (2010) Rapid screening and confirmation of drugs and toxic compounds in biological specimens using liquid chromatography/ion trap tandem mass spectrometry and automated library search. Rapid Commun Mass Spectrom 24:75–84
116. Oberacher H, Pavlic M, Libiseller K, Schubert B, Sulyok M, Schuhmacher R, Csaszar E, Köfeler HC (2009) On the inter-instrument and inter-laboratory transferability of a tandem mass spectral reference library: 1. Results of an Austrian multicenter study. J Mass Spectrom 44:485–493
117. Pavlic M, Schubert B, Libiseller K, Oberacher H (2010) Comprehensive identification of active compounds in tablets by flow-injection data-dependent tandem mass spectrometry combined with library search. Forensic Sci Int 197:40–47
118. Gergov M, Robson JN, Duchoslav E, Ojanperä I (2000) Automated liquid chromatographic/tandem mass spectrometric method for screening beta-blocking drugs in urine. Mass Spectrom 35:912–918
119. Mylonas R, Mauron Y, Masselot A, Philippe O, Binz PA, Budin N, Fathi M, Viette V, Hochstrasser DF, Lisacek F, Goetz S, Vagts J, Baessmann C (2009) A new approach for acute clinical toxicology based on ion trap LC/MSMS library search. Proceedings of the 18th International Mass Spectrometry Conference, Bremen
120. Josephs JL, Grubb MF, Shipkova P, Langish RA. (2005) A comprehensive strategy for the characterization and optimization of metabolic profiles of compounds using a hybrid linear ion trap/FTMS. Proceedings of the 53rd ASMS Conference on Mass Spectrometry and Allied Topics, San Antonio
121. HMDB. http://www.hmdb.ca. Accessed 31 May 2010
122. Wishart DS, Knox C, Guo AC et al (2009) HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res 37:D603–D610
123. Scripps Center for Mass Spectrometry METLIN. http://metlin.scripps.edu. Accessed 31 May 2010
124. HighChem MS/MS Spectral Libraries. http://www.highchem.com/leading-edge-technologies/ms/ms-spectral-libraries.html. Accessed 31 May 2010
125. Platform for RIKEN Metabolomics. http://prime.psc.riken.jp. Accessed 31 May 2010
126. Sawada Y, Akiyama K, Sakata A, Kuwahara A, Otsuki H, Sakurai T, Saito K, Hirai MY (2009) Widely targeted metabolomics based on large-scale MS/MS data for elucidating metabolite accumulation patterns in plants. Plant Cell Physiol 50:37–47
127. Manchester Metabolomics Database (MMD). http://dbkgroup.org/MMD. Accessed 31 May 2010
128. Lee JS, Kim DH, Liu KH, Oh TK, Lee CH (2005) Identification of flavonoids using liquid chromatography with electrospray ionization and ion trap tandem mass spectrometry with an MS/MS library. Rapid Commun Mass Spectrom 19:3539–3548
129. Fredenhagen A, Derrien C, Gassmann E (2005) An MS/MS library on an ion-trap instrument for efficient dereplication of natural products. Different fragmentation patterns for $[M + H]^+$ and $[M + Na]^+$ ions. J Nat Prod 68:385–391
130. Takegawa Y, Deguchi K, Ito S, Yoshioka S, Sano A, Yoshinari K, Kobayashi K, Nakagawa H, Monde K, Nishimura S (2004) Assignment and quantification of 2-aminopyridine derivatized oligosaccharide isomers coeluted on reversed-phase HPLC/MS by $MS^n$ spectral library. Anal Chem 76:7294–7303

131. Kameyama A, Kikuchi N, Nakaya S, Ito H, Sato T, Shikanai T, Takahashi Y, Takahashi K, Narimatsu H (2005) A strategy for identification of oligosaccharide structures using observational multistage mass spectral library. Anal Chem 77:4719–4725
132. Zhang H, Singh S, Reinhold VN (2005) Congruent strategies for carbohydrate sequencing. 2. FragLib: an MSn spectral library. Anal Chem 77:6263–6270
133. Milman BL, Zhurkovich IK (2009) Tandem mass spectral library of pesticides and its use in identification. Proceedings of the 18th International Mass Spectrometry Conference, Bremen
134. Bristow AW, Webb KS, Lubben AT, Halket J (2004) Reproducible product-ion tandem mass spectra on various liquid chromatography/mass spectrometry instruments for the development of spectral libraries. Rapid Commun Mass Spectrom 18:1447–1454
135. Josephs JL, Sanders M (2004) Creation and comparison of MS/MS spectral libraries using quadrupole ion trap and triple-quadrupole mass spectrometers. Rapid Commun Mass Spectrom 18:743–759
136. Ferrer I, Fernandez-Alba A, Zweigenbaum JA, Thurman EM (2006) Exact-mass library for pesticides using a molecular-feature database. Rapid Commun Mass Spectrom 20:3659–3668
137. Højer-Pedersen J, Smedsgaard J, Nielsen J (2008) The yeast metabolome addressed by electrospray ionization mass spectrometry: Initiation of a mass spectral library and its applications for metabolic footprinting by direct infusion mass spectrometry. Metabolomics 4:393–405
138. Hopley C, Bristow T, Lubben A, Simpson A, Bull E, Klagkou K, Herniman J, Langley J (2008) Towards a universal product ion mass spectral library – reproducibility of product ion spectra across eleven different mass spectrometers. Rapid Commun Mass Spectrom 22:1779–1786
139. Volná K, Holcapek M, Kolárová L, Lemr K, Cáslavský J, Kacer P, Poustka J, Hubálek M (2008) Comparison of negative ion electrospray mass spectra measured by seven tandem mass analyzers towards library formation. Rapid Commun Mass Spectrom 22:101–108
140. Hogenboom AC, Van Leerdam JA, De Voogt P (2009) Accurate mass screening and identification of emerging contaminants in environmental samples by liquid chromatography-hybrid linear ion trap Orbitrap mass spectrometry. J Chromatogr A 1216:510–519
141. Madison–Qingdao Metabolomics Consortium Database (MMCD). http://mmcd.nmrfam.wisc.edu. Accessed 1 June 2010
142. Milman BL (2005) Towards a full reference library of $MS^n$ spectra. Testing of a library containing 3126 MS2 spectra of 1743 compounds. Rapid Commun Mass Spectrom 19:2833–2839
143. McLuckey SA (1992) Principles of collisional activation in analytical mass spectrometry. J Am Soc Mass Spectrom 3:599–614
144. Weinmann W, Gergov M, Goerner M (2000) MS/MS-libraries with triple quadrupole-tandem mass spectrometers for drug identification and drug screening. Analysis 28:934–941
145. Gergov M, Weinmann W, Meriluoto J, Uusitalo J, Ojanpera I (2004) Comparison of product ion spectra obtained by liquid chromatography/triple-quadrupole mass spectrometry for library search. Rapid Commun Mass Spectrom 18:1039–1046
146. Jansen R, Lachatre G, Marquet P (2005) LC–MS/MS systematic toxicological analysis: comparison of MS/MS spectra obtained with different instruments and settings. Clin Biochem 38:362–372
147. Kienhuis PG, Geerdink RB (2002) A mass spectral library based on chemical ionization and collision-induced dissociation. J Chromatogr A 974:161–168
148. Thermo Scientific LTQ Orbitrap XL. http://www.analiticaweb.com.br/thermo/AdMS/Orbitrap/LTQOrbitrapXL_PS.pdf. Accessed 3 June 2010
149. Venisse N, Marquet P, Duchoslav E, Dupuy JL, Lachâtre G (2003) A general unknown screening procedure for drugs and toxic compounds in serum using liquid chromatography-electrospray-single quadrupole mass spectrometry. J Anal Toxicol 27:7–14

150. Matsuda F, Yonekura-Sakakibara K, Niida R, Kuromori T, Shinozaki K, Saito K (2009) MS/MS spectral tag-based annotation of non-targeted profile of plant secondary metabolites. Plant J 57:555–577
151. Thielen B, Heinen S, Schomburg D (2009) mSpecs: a software tool for the administration and editing of mass spectral libraries in the field of metabolomics. BMC Bioinformatics 10:229. doi:10.1186/1471-2105-10-229
152. Styczynski MP, Moxley JF, Tong LV, Walther JL, Jensen KL, Stephanopoulos GN (2007) Systematic identification of conserved metabolites in GC/MS data for metabolomics and biomarker discovery. Anal Chem 79:966–973
153. UniProtKB/Swiss-Prot protein knowledgebase release 2010_06 statistics. http://expasy.org/sprot/relnotes/relstat.html. Accessed 24 May 2010
154. Kinter M, Sherman NE (2000) Protein sequencing and identification using tandem mass spectrometry. Wiley, New York
155. Aebersold R, Goodlett DR (2001) Mass spectrometry in proteomics. Chem Rev 101:269–295
156. Sechi S (2007) Quantitative proteomics by mass spectrometry. Humana Press, Totowa, NJ
157. Hummel J, Niemann M, Wienkoop S, Schulze W, Steinhauser D, Selbig J, Walther D, Weckwerth W (2007) ProMEX: a mass spectral reference database for proteins and protein phosphorylation sites. BMC Bioinformatics 8:216. doi:10.1186/1471-2105-8-216
158. Liu J, Bell AW, Bergeron JJ, Yanofsky CM, Carrillo B, Beaudrie CE, Kearney RE (2007) Methods for peptide identification by spectral comparison. Proteome Sci 5:3. doi:10.1186/1477-5956-5-3
159. Falkner JA, Kachman M, Veine DM, Walker A, Strahler JR, Andrews PC (2007) Validated MALDI-TOF/TOF mass spectra for protein standards. J Am Soc Mass Spectrom 18:850–855
160. sPRG. http://www.abrf.org/index.cfm/group.show/ProteomicsStandardsResearchGroup.47.htm. Accessed 4 June 2010
161. Lam H, Deutsch EW, Eddes JS, Eng JK, Stein SE, Aebersold R (2008) Building consensus spectral libraries for peptide identification in proteomics. Nat Methods 5:873–875. doi:10.1038/nmeth.1254
162. Craig R, Cortens JC, Fenyo D, Beavis RC (2006) Using annotated peptide mass spectrum libraries for protein identification. J Proteome Res 5:1843–1849
163. Lam H, Deutsch EW, Eddes JS, Eng JK, King N, Stein SE, Aebersold R (2007) Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics 7:655–667
164. Tasman N (2009) SpectraST: a spectral library building and searching tool for proteomics. http://www.proteomecenter.org/april.09.weblectures/3.tasman.SpectraST.4.09.pdf. Accessed 4 June 2010
165. Yen CY, Meyer-Arendt K, Eichelberger B, Sun S, Houel S, Old WM, Knight R, Ahn NG, Hunter LE, Resing KA (2009) A simulated MS/MS library for spectrum-to-spectrum searching in large scale identification of proteins. Mol Cell Proteomics 8:857–869
166. Nesvizhskii AI, Vitek O, Aebersold R (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Methods 4:787–797
167. Frewen B, MacCoss MJ (2007) Using BiblioSpec for creating and searching tandem MS peptide libraries. Curr Protoc Bioinf: Chapter 13, Unit 13.7. doi: 10.1002/0471250953.bi1307s20
168. The global proteome machine organization proteomics database and open source software. http://www.thegpm.org. Accessed 4 June 2010
169. Slotta DJ, Barrett T, Edgar R (2009) NCBI Peptidome: a new public repository for mass spectrometry peptide identifications. Nat Biotechnol 27:600–601. doi:10.1038/nbt0709-600
170. Morey J, Rogers I, Chen C (2006) Filtering out MS/MS spectra of insufficient quality before database searching. Proceedings of the 54st ASMS Conference on Mass Spectrometry and Allied Topics, Seattle. http://www.bioinformaticssolutions.com/products/peaks/db_bsipaper.php. Accessed 4 June 2010

171. Han J, Danell RM, Patel JR, Gumerov DR, Scarlett CO, Speir JP, Parker CE, Rusyn I, Zeisel S, Borchers CH (2008) Towards high-throughput metabolomics using ultrahigh-field Fourier transform ion cyclotron resonance mass spectrometry. Metabolomics 4:128–140
172. Kind T, Fiehn O (2006) Metabolomic database annotations *via* query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. BMC Bioinformatics 7:234. doi:10.1186/1471-2105-7-234
173. Stoll N, Schmidt E, Thurow K (2006) Isotope pattern evaluation for the reduction of elemental compositions assigned to high-resolution mass spectral data from electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. J Am Soc Mass Spectrom 17:1692–1699
174. Breitling R, Pitt AR, Barrett MP (2006) Precision mapping of the metabolome. Trends Biotechnol 24:543–548
175. Kind T, Fiehn O (2007) Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Bioinformatics 8:105. doi:10.1186/1471-2105-8-105
176. Ibáñez M, Sancho JV, Pozo OJ, Niessen W, Hernández F (2005) Use of quadrupole time-of-flight mass spectrometry in the elucidation of unknown compounds present in environmental water. Rapid Commun Mass Spectrom 19:169–178
177. MassWorks sCLIPS. http://www.cernobioscience.com/products/sClips.pdf. Accessed 4 June 2010
178. Farré M, Gros M, Hernández B, Petrovic M, Hancock P, Barceló D (2008) Analysis of biologically active compounds in water by ultra-performance liquid chromatography quadrupole time-of-flight mass spectrometry. Rapid Commun Mass Spectrom 22:41–51
179. Gilbert JR, Lewer P, Duebelbeis DO, Carr AW, Snipes CE, Williamson RT (2003) Identification of biologically active compounds from nature using liquid chromatography/mass spectrometry. ACS Symp Ser 850:52–65
180. Fiehn O (2007) Cellular metabolomics: the quest for pathway structure. In: Lindon JC, Nicholson JK, Holmes E (eds) The handbook of metabonomics and metabolomics. Elsevier, Amsterdam
181. Ferrer I, Thurman EM (Eds) (2003) Liquid chromatography/mass spectrometry, MS/MS and time of flight MS: Analysis of emerging contaminants. ACS, Washington DC, ACS Symp Ser V. 850
182. Grimalt S, Pozo OJ, Sancho JV, Hernández F (2007) Use of liquid chromatography coupled to quadrupole time-of-flight mass spectrometry to investigate pesticide residues in fruits. Anal Chem 79:2833–2843
183. García-Reyes JF, Hernando MD, Ferrer C, Molina-Díaz A, Fernández-Alba AR (2007) Large scale pesticide multiresidue methods in food combining liquid chromatography-time-of-flight mass spectrometry and tandem mass spectrometry. Anal Chem 79:7308–7323
184. Bogdanov B, Smith RD (2005) Proteomics by FTICR mass spectrometry: top down and bottom up. Mass Spectrom Rev 24:168–200
185. Marshall AG, Hendrickson CL (2008) High-resolution mass spectrometers. Annu Rev Anal Chem 1:579–599
186. The Regents of the University of California ProteinProspector. MS-Isotope. http://prospector.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msisotope. Accessed 4 June 2010
187. Vershinin VI, Derendyaev BG, Lebedev KS (2002) Computer-assisted identification of organic compounds (In Russian). Akademkniga, Moscow
188. Steinbeck C (2004) Recent developments in automated structure elucidation of natural products. Nat Prod Rep 21:512–518
189. Elyashberg M, Blinov K, Molodtsov S, Smurnyy Y, Williams AJ, Churanova T (2009) Computer-assisted methods for molecular structure elucidation: realizing a spectroscopist's dream. J Cheminformatics 1:3. doi:10.1186/1758-2946-1-3
190. ACD/MS Fragmenter. http://www.acdlabs.com/products/adh/ms/ms_frag. Accessed 4 June 2010

191. Heinonen M, Rantanen A, Mielikäinen T, Kokkonen J, Kiuru J, Ketola RA, Rousu J (2008) FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data. Rapid Commun Mass Spectrom 22:3043–3052
192. Shin HJ, Matsuda H, Murakami M, Yamaguchi K (1997) Anabaenopeptins E and F, two new cyclic peptides from the cyanobacterium *Oscillatoria agardhii* (NIES-204). J Nat Prod 60:139–141
193. Kornakova TA, Bogdanova TF, Piottukh-Peletskii VN (2008) Evaluation of the efficiency of the concurrent use of IR and mass spectrometry databases for structure elucidation. J Struct Chem 49:224–234
194. Coates J (2000) Interpretation of infrared spectra, a practical approach in encyclopedia of analytical chemistry. In: Meyers RA (ed) Encyclopedia of analytical chemistry, pp. 10815-10837. Wiley, Chichester. http://infrared.als.lbl.gov/BLManual/IR_Interpretation.pdf. Accessed 4 June 2010
195. Pretsch E, Bühlmann P, Badertscher M (2009) Structure determination of organic compounds, 4th edn. Springer, Berlin
196. Debska B, Guzowska-Swider B, Cabrol-Bass D (2000) Automatic generation of knowledge base from infrared spectral database for substructure recognition. J Chem Inf Comput Sci 40:330–338
197. Hachey MRJ (2004) Tautomerism and expert systems in spectroscopy. Spectroscopy 19:44. http://spectroscopyonline.findanalytichem.com/spectroscopy/data/articlestandard//spectroscopy/192004/94284/article.pdf. Accessed 4 June 2010
198. Boruta M, Hachey M, Bogomolov A, Karpushkin E, Williams T**.** Computer assisted structure verification and interpretation of Infrared and Raman Spectra. http://www.acdlabs.com/download/publ/2004/facss_verif_interp_raman.pdf. Accessed 4 June 2010
199. Varmuza K, Karlovits M, Demuth W (2003) Spectral similarity versus structural similarity: infrared spectroscopy. Anal Chim Acta 490:313–324
200. Derendyaev BG, Bogdanova TF, Piottukh-Peletsky VN, Makarov LI (2004) Fast taxonomy of chemical structures selected from IR spectral database. Anal Chim Acta 509:209–216
201. Bio-Rad/KnowItAll HaveItAll IR. http://www.knowitall.com/literature/docs/95379-Bio-Rad_HaveItAll_IR_Datasheet.pdf#zoom=90%. Accessed 4 June 2010
202. NICODOM IR Professional Package. http://www.ir-spectra.com/download/NICODOM_IR_prof_pac1.htm. Accessed 4 June 2010
203. ACD/IR and Raman Databases. http://www.acdlabs.com/products/adh/uvir/ir_raman_db. Accessed 5 June 2010
204. Sigma-Aldrich spectral libraries. http://www.sigmaaldrich.com/catalog/Lookup.do?N5=All&N3=mode+matchpartialmax&N4=spectral+libraries&D7=0&D10=spectral+libraries&N1=S_ID&ST=RS&N25=0&F=PR. Accessed 5 June 2010
205. Thermo IR spectral libraries. http://www.thermoscientific.com/wps/portal/ts/products/catalog?categoryId=81851. Accessed 4 June 2010
206. FDM Reference Spectra Databases. http://www.fdmspectra.com/index.html. Accessed 5 June 2010
207. NIST Standard Reference Database 79. http://www.nist.gov/srd/nist79.cfm. Accessed 3 Nov 2010
208. Pacific Northwest National Laboratory Northwest-Infrared. https://secure2.pnl.gov/nsd/nsd.nsf/Welcome. Accessed 5 June 2010
209. Oberreuter H, Seiler H, Scherer S (2002) Identification of coryneform bacteria and related taxa by Fourier-transform infrared (FT-IR) spectroscopy. Int J Syst Evol Microbiol 52:91–100
210. Improving search results using high resolution libraries (2007) Thermo Application note AN50745_E 11/07M. http://www.thermo.com/eThermo/CMA/PDFs/Articles/articlesFile_7205.pdf. Accessed 5 June 2010
211. Chang WT, Yu CC, Wang CT, Tsai YY (2003) A critical evaluation of spectral library searching for the application of automotive paint database. Forensic Sci J 2:47–58

212. McCreery RL, Horn AJ, Spencer J, Jefferson E (1998) Noninvasive identification of materials inside USP vials with Raman spectroscopy and a Raman spectral library. J Pharm Sci 87:1–8
213. Meiler J, Will M (2002) Genius: a genetic algorithm for automated structure elucidation from $^{13}$C NMR spectra. J Am Chem Soc 124:1868–1870
214. Meiler J, Köck M (2004) Novel methods of automated structure elucidation based on $^{13}$C NMR spectroscopy. Magn Reson Chem 42:1042–1045
215. Bodis R (2007) Quantification of spectral similarity: towards automatic spectra verification. Dissertation ETH 17361, Zürich. http://e-collection.ethbib.ethz.ch/eserv/eth:29907/eth-29907-02.pdf. Accessed 15 May 2010
216. Elyashberg M, Blinov K, Williams A (2009) A systematic approach for the generation and verification of structural hypotheses. Magn Reson Chem 47:371–389
217. Modgraph NMRPredict overview. http://www.modgraph.co.uk/product_nmr.htm. Accessed 5 June 2010
218. NMRPredict server. http://nmrpredict.orc.univie.ac.at. Accessed 5 June 2010
219. Modgraph Press Release. http://www.modgraph.co.uk/best_proton_press_release.htm. Accessed 5 June 2010
220. Modgraph NMRPredict versus ACD CNMR/Predictor. http://www.modgraph.co.uk/product_nmr_shiftdb.htm. Accessed 5 June 2010
221. ACD/NMR Predictors. http://www.acdlabs.com/products/adh/nmr/nmr_pred. Accessed 5 June 2010
222. Blinov KA, Smurnyy YD, Elyashberg ME, Churanova TS, Kvasha M, Steinbeck C, Lefebvre BA, Williams AJ (2008) Performance validation of neural network based (13)c NMR prediction using a publicly available data source. J Chem Inf Model 48:550–555
223. Upstream Solutions NMR prediction. http://www.upstream.ch/products/nmr.html#PredictionQuality. Accessed 5 June 2010
224. SpecSurf XS. http://cds.dl.ac.uk/cds/manuals/specsurf/i-guide.html. Accessed 5 June 2010
225. Kuhn S, Egert B, Neumann S, Steinbeck C (2008) Building blocks for automated elucidation of metabolites: Machine learning methods for NMR prediction. BMC Bioinformatics 9:400. doi:10.1186/1471-2105-9-400
226. Dunkel R, Wu X (2007) Identification of organic molecules from a structure database using proton and carbon NMR analysis results. J Magn Reson 188:97–110
227. Golotvin S, Vodopianov E, Lefebvre B, Williams AJ, Spitzer TD (2006) Automated structure verification based on $^{1}$H NMR prediction. Magn Reson Chem 44:524–538
228. Golotvin SS, Vodopianov E, Pol R, Lefebvre BA, Williams AJ, Rutkowske RD, Spitzer TD (2007) Automated structure verification based on a combination of 1D $^{1}$H NMR and 2D $^{1}$H–$^{13}$C HSQC spectra. Magn Reson Chem 45:803–813
229. Keyes P, Hernandez G, Cianchetta G, Robinson J, Lefebvre B (2009) Automated compound verification using 2D-NMR HSQC data in an open-access environment. Magn Reson Chem 47:38–52
230. Smith SK, Cobleigh J, Svetnik V (2001) Evaluation of a $^{1}$H-$^{13}$C NMR spectral library. J Chem Inf Comput Sci 41:1463–1469
231. Meiler J, Sanli E, Junker J, Meusinger R, Lindel T, Will M, Maier W, Köck M (2002) Validation of structural proposals by substructure analysis and $^{13}$C NMR chemical shift prediction. J Chem Inf Comput Sci 42:241–248
232. Bio-Rad/KnowItAll NMR Databases. http://www.knowitall.com/literature. Accessed 6 June 2010
233. ACD/NMR Databases. http://www.acdlabs.com/products/adh/nmr/nmr_db. Accessed 6 June 2010
234. Modgraph C13 NMR and X-Nuclei Reference Database. http://www.modgraph.co.uk/product_nmr_database.htm. Accessed 6 June 2010
235. CSEARCH-NMR Database Description. http://homepage.univie.ac.at/wolfgang.robien/csearch_main.html. Accessed 6 June 2010

236. Schütz V, Purtuc V, Felsinger S, Robien W (1997) CSEARCH-STEREO: A new generation of NMR database systems allowing three-dimensional spectrum prediction. Fresenius J Anal Chem 359:33–41
237. NMRShiftDB. http://www.ebi.ac.uk/nmrshiftdb/portal/js_pane/P-Help. Accessed 6 June 2010
238. NMR metabolomics database of Linkoping (MDL). http://www.liu.se/hu/mdl/main. Accessed 6 June 2010
239. Biological Magnetic Resonance Data Bank (BMRB). http://www.bmrb.wisc.edu. Accessed 6 June 2010
240. Re-referenced Protein Chemical Shift Database (RefDB). http://redpoll.pharmacy.ualberta.ca/RefDB. Accessed 6 June 2010
241. Zhang H, Neal S, Wishart DS (2003) RefDB: a database of uniformly referenced protein chemical shifts. J Biomol NMR 25:173–195
242. ChemSpider. http://www.chemspider.com. Accessed 6 June 2010
243. Dunn WB, Ellis DI (2005) Metabolomics: current analytical platforms and methodologies. Trends Anal Chem 24:285–294
244. Wilson ID, Plumb R, Granger J, Major H, Williams R, Lenz EM (2005) HPLC–MS-based methods for the study of metabonomics. J Chromatogr B 817:67–76
245. Biais B, Allwood JW, Deborde C, Xu Y, Maucourt M, Beauvoit B, Dunn WB, Jacob D, Goodacre R, Rolin D, Moing A (2009) $^{1}$H NMR, GC-EI-TOFMS, and data set correlation for fruit metabolomics: application to spatial metabolite analysis in melon. Anal Chem 81:2884–2894
246. Dunn WB, Bailey NJC, Johnson HE (2005) Measuring the metabolome: current analytical technologies. Analyst 130:606–625
247. Bally RW, Van Krimpen D, Cleij P, Van 'T Klooster HA (1984) An automated library search system for $^{13}$C-n.m.r. spectra based on the reproducibility of chemical shifts. Anal Chim Acta 157:227–243
248. Xia J, Bjorndahl TC, Tang P, Wishart DS (2008) MetaboMiner – semi-automated identification of metabolites from 2D NMR spectra of complex biofluids. BMC Bioinformatics 9:507. doi:10.1186/1471-2105-9-507
249. Xi Y, De Ropp JS, Viant MR, Woodruff DL, Yu P (2008) Improved identification of metabolites in complex mixtures using HSQC NMR spectroscopy. Anal Chim Acta 614:127–133
250. Lay JO, Borgmann S, Liyanage R, Wilkins CL (2006) Problems with the "omics". Trends Anal Chem 25:1046–1056
251. Villas-Bôas SG, Mas S, Akesson M, Smedsgaard J, Nielsen J (2005) Mass spectrometry in metabolome analysis. Mass Spectrom Rev 24:613–646
252. Dettmer K, Aronov PA, Hammock BD (2007) Mass spectrometry-based metabolomics. Mass Spectrom Rev 26:51–78
253. Ellis DI, Dunn WB, Griffin JL, Allwood JW, Goodacre R (2007) Metabolic fingerprinting as a diagnostic tool. Pharmacogenomics 8:1243–1266
254. Idborg H, Zamani L, Edlund PO, Schuppe-Koistinen I, Jacobsson SP (2005) Metabolic fingerprinting of rat urine by LC/MS Part 1. Analysis by hydrophilic interaction liquid chromatography–electrospray ionization mass spectrometry. J Chromatogr B 828:9–13
255. Idborg H, Zamani L, Edlund PO, Schuppe-Koistinen I, Jacobsson SP (2005) Metabolic fingerprinting of rat urine by LC/MS Part 2. Data pretreatment methods for handling of complex data. J Chromatogr B 828:14–20
256. Bafna V, Edwards N (2003) On *de novo* interpretation of tandem mass spectra for peptide identification. Proceedings of the 7th annual international conference on Research in computational molecular biology, Berlin. http://proteomics.ucsd.edu/papers/on_de_novo.pdf. Accessed 6 June 2010
257. Arrigoni G, Fernandez C, Holm C, Scigelova M, James P (2006) Comparison of MS/MS methods for protein identification from 2D-PAGE. J Proteome Res 5:2294–2300

258. Balgley BM, Laudeman T, Yang L, Song T, Lee CS (2007) Comparative evaluation of tandem MS search algorithms using a target–decoy search strategy. Mol Cell Proteomics 6:1599–1608
259. Price TS, Lucitt MB, Wu W, Austin DJ, Pizarro A, Yocum AK, Blair IA, FitzGerald GA, Grosser T (2007) EBP, a program for protein identification using multiple tandem mass spectrometry datasets. Mol Cell Proteomics 6:527–536
260. Alves G, Wu WW, Wang G, Shen RF, Yu YK (2008) Enhancing peptide identification confidence by combining search methods. J Proteome Res 7:3102–3113
261. Zubarev RA, Zubarev AR, Savitski MM (2008) Electron capture/transfer versus collisionally activated/induced dissociations: solo or duet? J Am Soc Mass Spectrom 19:753–761
262. Kapp EA, Schütz F, Connolly LM, Chakel JA, Meza JE, Miller CA, Fenyo D, Eng JK, Adkins JN, Omenn GS, Simpson RJ (2005) An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. Proteomics 5:3475–3490
263. Andrews PC, Arnott DP, Gawinowicz MA, Kowalak JA, Lane WS, Lilley KS, Martin LT, Stein SE. ABRF-sPRG2006 Study: A proteomics standard. http://www.abrf.org/ResearchGroups/ProteomicsStandardsResearchGroup/EPosters/ABRFsPRGStudy2006poster.pdf. Accessed 7 June 2010
264. Andrews PC, Arnott DP, Gawinowicz MA, Kowalak JA, Lane WS, Lilley KS, Loo RRO, Martin LT, Stein SE. sPRG2007: Development and evaluation of a phosphoprotein standard. http://www.abrf.org/ResearchGroups/ProteomicsStandardsResearchGroup/EPosters/Gawinowicz_sPRG07_032707.pdf. Accessed 7 June 2010
265. Bell AW, Deutsch EW, Au CE, Kearney RE, Beavis R, Sechi S, Nilsson T, Bergeron JJ, HUPO Test Sample Working Group (2009) A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. Nat Methods 6:423–430

# Chapter 8
# Chemical Qualitative Analysis II

**Abstract** Qualitative analysis II is identification/classification/authentication of such objects as foodstuffs, products, specimens, materials, pollutions, living organisms, and others. Typical procedures of this sort are authentication of food, determination of its adulteration, oil spill identification, and that of microorganisms. Identification of an object is based on recognition of its indicative component(s), measuring ratios between several components of a sample, or fingerprinting overall sample signals. Almost all analytical techniques are applicable for the purpose, with an indispensable role for chemometrics/multivariate statistics in processing of analytical data. In the same way as in identification of individual chemical compounds, quality of identification II is assured by validation of methods, the use of reference materials, and availability of standard/valid reference data. Examples of qualitative analysis of vegetable oils, honey, wine, and some non-food samples are given.

## 8.1 General

### *8.1.1 Concepts and Definitions*

In its second implementation, chemical identification is recognition of such objects as foodstuffs, products, specimens, materials, pollutions, living organisms, and so, on rather than individual chemical compounds and their simple mixtures (see Chap. 1). Following [1–3], the term of qualitative analysis II is used here. Up to about ten not fully identical versions of such qualitative procedures occur (Fig. 8.1). Some typical procedures can be used as examples:

- Authentication of food [6–18]
- Oil spill identification [19–22]
- Identification of microorganisms [23–32]

The type of identification under consideration can be also named *qualitative analysis of complex chemical mixtures* [33]. This is not a very exact definition in all

B.L. Milman, *Chemical Identification and its Quality Assurance*,
DOI 10.1007/978-3-642-15361-7_8, 

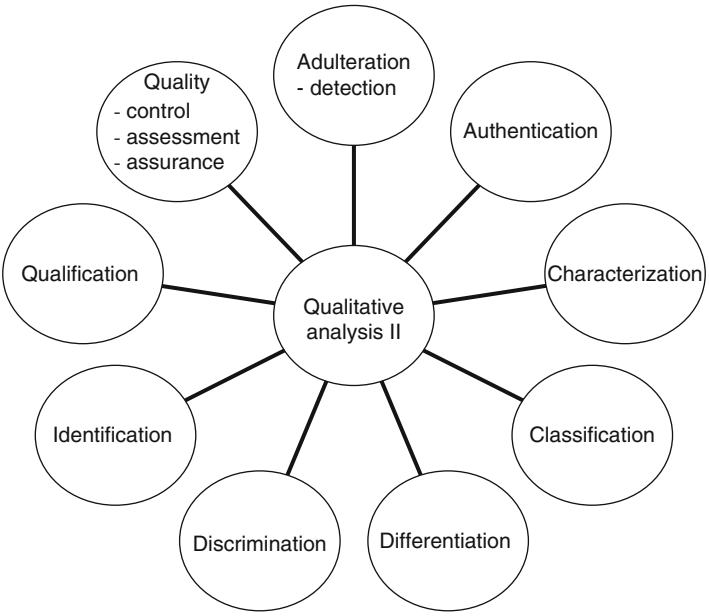



**Fig. 8.1** Terminology of qualitative analysis II: names of procedures including analytical operations. Names are partly synonymic and carry different shades of meaning. *Identification* (Sect. 1.2), *classification*, and *qualification* are rather similar in meaning. *Authentication* is a confirmation of identity, "verification of the claimed identity. . ." [4]. *Characterization* is rather a determination or description of characteristics/properties/features of an object for its further identification. The terms of *differentiation* and *discrimination* have been used to emphasize a possibility of distinction between objects or their states under identification. *Adulteration* is understood to be ". . .replacing valuable ingredients with inferior ones" [5]; the respective analytical procedure for determination of those components (*adulterants*) is named *detection* or *determination of adulteration*. Lastly, these procedures are required when *quality* of food, materials, raw products, and so on is controlled/assessed

cases, because an object such as a bacterium or fruit is something far more complex than complex mixtures of chemical compounds. However, if one focuses on samples extracted from objects identified instead of objects themselves, the definition related to complex chemical mixtures seems to fit the case.

What is determined is the object type or the object having the particular feature(s). Detection of features related to an origin and state of an object is predominant in qualitative analysis II (Fig. 8.2). There is another feature/characteristic closely related to *authenticity* and *origin*. It is *traceability* (a "non-metrological" procedure; see Sect. 1.7) which is defined by ISO as "ability to trace the history, application, or location of that which is under consideration" (see [34]). A special attention has been focused on *food traceability* [34].

In a very broad sense, any chemical analysis related to detection of controlled chemicals in food and so on, and targeted to subsequent characterization of a sample as hazardous/dangerous or safe according to compliance with some acts,

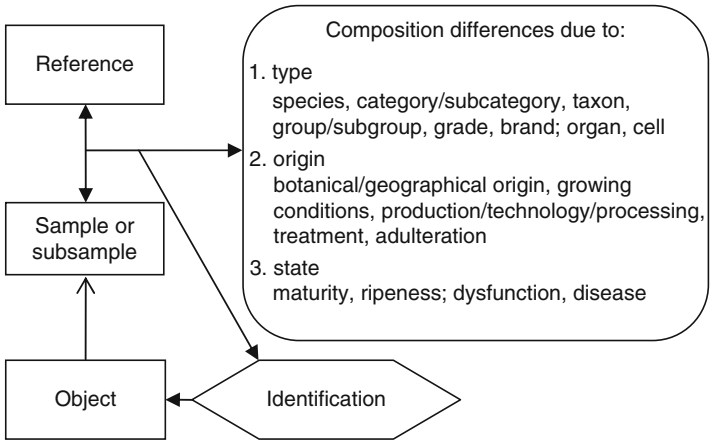



**Fig. 8.2** Schematic of qualitative analysis II. One identifies a subject of a certain type, origin, and state, analyzing its sample using techniques and methods of chemical identification and reference data/materials

standards, rules, regulations, and so on, can be considered as a version of qualitative analysis II.

### *8.1.2 Analytical Approaches, Techniques, and Methods*

**Approaches.** Identification of an object is based on identification of

- A single sample component (indicative component, marker)
- Group of components or
- Fingerprinting its composition without certain determination of its constituents

Examples are given in Table 8.1.

**Techniques.** In contrast to identification I, where individual compounds in complicated mixtures are predominantly determined by chromatography mass spectrometry, there is a wide variety of analytical techniques for chemical qualitative analysis II; see books [7, 18, 38]. The number of proper applications of GC, GC–MS, HPLC, HPLC–MS, and $MS^n$ is certainly very large. The same holds true for NIR and other versions of IR spectroscopy [9, 12, 13, 17]. The following techniques are also widely or rather widely used.

- NMR [39–41].
- Isotopic ratios mass spectrometry [36, 40], pyrolysis mass spectrometry [42] and direct injection/infusion ESI mass spectrometry (see [43]) which are a very special branch of MS.

**Table 8.1** Different approaches in qualitative analysis II

| Analyte, measurand | Remark | Examples |
|---|---|---|
| Individual component, its amount | Specific compound in the sample which characterizes the object | 1. Typical phenolic compounds in food authentication, e.g., phloretin 2′-xylosylglucoside and phloretin 2′-xylosylglucoside for apples [10]<br>2. Anthocyanins for detecting the adulteration of expensive fruit purees (see [35]). |
| Diagnostic ratio of components | Specific ratio between amounts of the particular components characterizes the object | 1. Diagnostic ratios of biomarkers for spill identification of oil and its fractions [20, 22]<br>2. Ratios of the heavy to light isotopes of the same elements ($^{13}C/^{12}C$, $^{18}O/^{16}O$, $^{87}Sr/^{86}Sr$ and other) for origin assignment and adulteration detection of wines [36]. |
| Fingerprinting | Identification of objects by matching their complex analytical signals with the use of multivariate statistics | 1. Mass chromatograms of oil hydrocarbons processed by PCA for characterization of oil spill samples [33].<br>2. MALDI mass spectra of milk proteins for differentiation between industrial processes or milk samples from different mammal species, observation of milk adulteration, and so on [37] |

- Fluorescence spectroscopy [14].
- Electrochemical sensors [44], including electronic nose [45] and electronic tongue [46].
- DNA analysis-based methods [15].

What particular technique should be used for qualitative analysis II? This depends on the kind of object and the version of analysis. In the general case, the use of several techniques is appropriate, because their advantages and disadvantages are commonly mismatched (Table 8.2) and different indicative components of the sample are not determined by the same technique/method (e.g., see [47]). Techniques of chemical analysis have been also applied, together with physical characterization [48] and sensory analysis [47] of samples.

**Standard Methods** play the same important role as in analysis I**.** Some important examples are as follows (see also Chap. 5).

- The European Commission regulates methods of analysis of food, e.g., olive oil and olive-residue oil [49, 50].
- Methods of testing foodstuffs are described or noted in the food standards, guidelines, and related texts of the Codex Alimentarius Commission (operating under the aegis of FAO/WHO) [51].

**Table 8.2** Comparative advantages and disadvantages of different spectroscopic techniques applied for the determination of the quality/identity of undiluted milk (adapted from [47])

| Technique | Sensitivity | Information content | Interferences | Repeatability | Light scatter |
|---|---|---|---|---|---|
| NIR | Intermediate | Intermediate | Many | Intermediate | Intermediate |
| FT-IR | High | High | Many | Good | Intermediate |
| Fluorescence | High | Low | Few | Intermediate | Severe |
| NMR | Intermediate | Intermediate | Many | Intermediate | No |

Techniques are compared with each other. Terminology is not adapted. Information content can be defined as the amount of information useful for the purpose of qualitative analysis (see Sect. 2.1). Availability of interferences characterizes the degree of selectivity of techniques. Commonly, light scattering limits performances of spectroscopic analysis

- Among standards for identification waterborne oil samples are ASTM documents, e.g., [21] (see also [20]).
- ASTM standard general methods exist which are developed for identification of different materials; see [52].
- There are standard methods of NIR analysis of foods, e.g., determination of cereal proteins [9].
- In the collection of official methods of AOAC International, methods for the authenticity testing of fruit juices, syrups, and honey and the revelation of the origin of ethyl alcohol are presented [53]. These methods are based on measurements of the isotope ratio of carbon $^{13}C/^{12}C$ (the value $\delta^{13}C$, ‰) in carbon dioxide formed upon combustion of organic samples.

**Multivariate Analysis and Chemometrics.** Another difference between identifications I and II is that in the second case multivariate/chemometrical methods (see Sect. 3.4 and [14, 33, 47, 48]) have been far more often applied. There are all possible methods listed in Table 3.1, and also Bayesian approach [54–57], with the widely occurring combination of PCA (see Fig. 8.3) and various variants of DA.

There is a standard for implementation of NIR qualitative analysis by means of multivariate statistical methods [59].

As in the use of other methods/techniques, the potentialities of different statistical ones are best found out from their methodical comparisons.

- Chemometric analysis of the Vis and NIR spectra by (a) DA (as discriminant partial least squares regression, PLS DA), (b) *k*-NN, and (c) SIMCA, (see Table 3.1) has been carried out to discriminate between unadulterated honey samples and those adulterated with fructose and glucose. This is a widespread case of adulteration. The DA version led to the most accurate results as compared to the other two methods [60]
- In solving the same analytical problem, different versions of DA such as PLS DA, linear (LDA) and quadratic discriminant analysis were tested and compared. Good classification results were obtained with all DA variants tested [57].
- Honey samples have been also differentiated by their floral origin. Techniques of Vis and NIR spectroscopy with both PLS DA and LDA made it possible to discriminate between two groups of honey samples with up to 100% trueness [61].

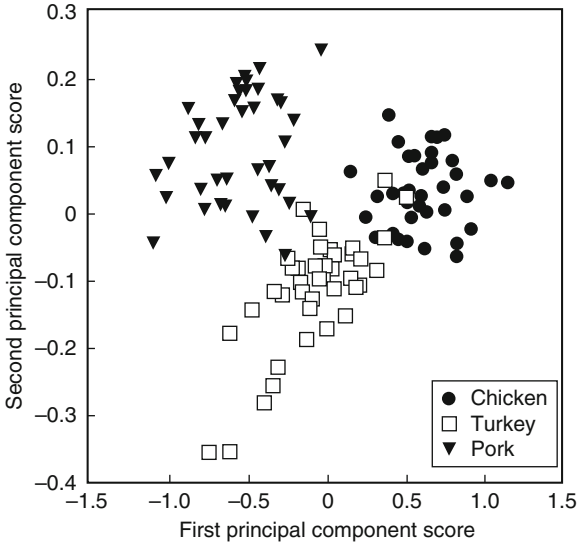



**Fig. 8.3** Mid-infrared spectroscopy and PC analysis discriminating between meats of fresh chicken, turkey, and pork [58]. These data can be further processed by technique of discriminant analysis for accurate classification and estimating corresponding errors (reproduced by permission of Elsevier)

### *8.1.3 Reliability of Results*

**Method validation.** Chemometric methods proving classification of unknown samples are developed and tested using training (calibration) and test (validation) sets of reference samples [14, 41, 54, 56, 58]. The first set is commonly employed to optimize mathematical models and derive classification rules for the attribution of unknown objects, whereas the second one is used to validate the classification/identification reliability of the ultimate optimized statistical model.

The approach to verification of classification models named cross-validation[1] is widespread [12, 14, 47, 54, 57, 58]. Here, the original sample set is selected, which is further partitioned into subsets. The subsets, with one exception, are used for training, and the single/excluded subset is intended for testing/validating the model. The validation process is then repeated with each of the subsets treated exactly once as the test data. The results can be further averaged/combined to estimate classification power.

There may also be a double validation: (a) cross-validation (an internal one) and (b) subsequent verification procedure with another sample set (named *external*

[1]This procedure is also used in validation of bioanalytical methods [62] and evaluation of MS libraries (without using this term [63, 64]; see Chap. 7).

**Table 8.3** Classification of wastes

| Sample | Classified as (%) | | | Correct classification (*St*) (%) | *FPR* (%) |
|---|---|---|---|---|---|
| | Wood | Plastic | Stone | | |
| Wood | 98.6 (TP) | 0.7 (FP) | 0.7 (FP) | 98.6 | 1.4 |
| Plastic | 0.0 (FP) | 96.9 (TP) | 3.1 (FP) | 96.9 | 3.1 |
| Stone | 2.9 (FP) | 2.5 (FP) | 94.6 (TP) | 94.6 | 5.4 |

Adapted from [2])

*validation*) [65, 66]. Correspondingly, two estimates for reliability will be produced (see below).

For the number of tests needed to reliably estimate an error level, see Sect. 4.2.1.

**Errors of identification/classification.** As the result of any type of qualitative analysis may be positive/negative and true/false, the identification terminology related to its reliability and errors (Chap. 4) holds valid for the classification of objects under discussion. The common concept is *correct classification* expressed in percents, i.e., percent of sample truly classified [12, 13, 56, 61, 66]. It is obviously the same as *TPR* (or statistical *sensitivity*; see Sect. 4.2.2 and Table 4.3). Estimation of these indices incorporated into a validation of chemometric procedures, is shown below. Example 8.1 is that of a triple classification.

Another example refers to the binary classification, with some cases unclassified.

Among other terms, there are for example *recognition ability* (%) and *prediction ability* (%) [54]. These are the percentages of correct classifications of the members of the training and evaluation sets respectively, with the classification rules deriving from training procedure. Such rates (80–100%) were evaluated for authentication of Galician (Spain) honeys by several multivariate methods based on processing of results of metal determination in the samples; the evaluation sample set was formed according to the cross-validation rule [54].

**Example 8.1**
In the report [2], NIR analysis of wastes carried out with the purpose of their classification into three classes (wood, plastic, and stone) was considered. Variables were abundances at six wavelengths selected from the range of 1,154–1,700 nm. LDA was used for classification of data obtained. Classification rates were evaluated with the test samples of all three groups. The identification results as normalized ones are given in Table 8.3. The ultimate error level is not high.

**Example 8.2**
Wine vinegar has been adulterated by adding alcohol vinegar. To discriminate between wine vinegar and the product containing adulterant, NIR spectroscopy and chemometric techniques, PCA, and the potential functions method were used [65]. Eventually, both types of individual samples were

*(continued)*

correctly classified in 80–100% of cases, depending on the mathematical model selected for classification. This is just the index of sensitivity or *TPR*, see Table 4.3. In these cases, the statistical specificity or *TNR* was estimated for correct unclassification of samples. The index of *TNR* was in the wide range of 0–100%. The best statistical models demonstrated *St* and *Sp* of 96–100% and 100%, respectively. The practically essential rate of *TPR* for vinegar blends was from about 94 to 100% [65].

These estimates were obtained for the particular validation set; see above. The cross-validation (internal one) led to not very different values.

### *8.1.4 Reference Materials*

The most reliable approach to classification of objects is when an analyst compares analysis results of unknown samples and reference materials. The latter are required to train/validate chemometric and other methods. Other purposes are also declared.

> Such CRMs also provide measurement traceability for food exports to facilitate acceptance in foreign markets, assess compliance with legal limits, and improve the accuracy of label information that is provided to assist consumers in making sound choices [67].

Main types of proper standards are characterized in Table 8.4. There have been no available standards for many plant/animal materials because their compositions depend on many factors, e.g., plant growing conditions and growing site. So those materials are just substances of variable composition, launching a challenge against the analyst's skill and experience. The challenge is that co-analysis of unknown and reference samples (see Table 1.4) may be impossible due to an absence of a reference of exactly the same composition.

Sometimes, analysts make reference materials themselves from proper raw materials in their own laboratories (Table 8.4). Some methods of food authentication, e.g., based on polymerase chain reaction, may not need corresponding reference materials [15].

### *8.1.5 Reference Data on Sample Composition*

Among information of this sort, reference data related to food compositions quantitatively prevail over those for other objects. Examples are the following.

- Numerous standards, tables, and other documents of the FAO/WHO Codex Alimentarius Commission [51].
- The FAO International Network of Food Data Systems (INFOODS) [75, 76].

**Table 8.4** Types of reference material for qualitative analysis II

| Type | Example |
|---|---|
| Individual component (markers) of samples | Oil hydrocarbons [68] , triglycerides [69] |
| Reference substance of nominal isotope ratio | Standard mean ocean water (isotopes of H and O), atmospheric air (N), Pee Dee belemnite (limestone, C), Canyon Diablo triolite (iron meteorite, S) [36] |
| Authentic food sample from outside | Established reference oil (petroleum) [70], plant oils [69], poultry samples [71] |
| Home-made standard of food, drinks, raw | Dried samples of medicinal herbs [72], "model wine" [73], plant oil from seeds [69] |
| Food-matrix standard reference material characterizing nutrient, element, and contaminant concentrations | Fatty acids and cholesterol in a frozen diet composite, trace elements in spinach leaves, whole milk powder and other NIST and non-NIST standards/certified reference materials [67, 74] |

- The USDA Nutrient Database for Standard Reference of so named *key foods* [76, 77].
- The particular data, e.g., the database of authentic poultry samples originated from the proper research reports [71].

Composition of other objects for qualitative analysis II, e.g., petroleum is searchable; see [20, 22, 78].

## 8.2 Objects

### *8.2.1 Food*

A profusion of reports on authentication/classification/characterization/ of food samples have been published, e.g., see [6–18]. The emphasis was put on analysis of plant oils (first, olive oil), honey, wines and some other foodstuffs and beverages because of their sales volume and the practice of their adulteration.

In the context of the authenticity of edible oils and fats, three main areas have to be differentiated:

- Economic adulteration, i.e., blending of cheaper oils with commodities of higher economic value.
- Minimally processed (non-refined, cold-pressed) oils.
- Characterisation and denomination of geographical origin [79].

**Oils and fats** [6–8, 12, 15, 39–41, 45, 66]. Here, detection of adulteration of olive oil is the challenge for analysts. Adulterants are cheaper oils such as soybean oil, sunflower oil, walnut oil, hazelnut oil. The latter is not easily detectable due to its similarity to olive oils in composition [80].

In general, a modern analytical method is based on separate detection of triglycerides, main components of plant oils. Triglycerides of different fatty acids and their different ratios characterize various oils. There are different types of LC which are very efficient for analysis of corresponding mixtures (e.g., see [69, 81]), although GC may also be usable to some degree [81]. Using the technique of UPLC with different detectors, one can detect and identify different oil impurities but not more than a few percents of other seed oils in olive oil samples [82]. Most other techniques and methods, including chemometric ones (Sect. 8.1.2), have also been required to analyze olive and other vegetable oils.

**Honey** [12, 13, 48, 53, 54, 60, 61]. Several examples of honey authentication were cited above.

The adulteration of honey has been commonly performed by its extension with sugar solutions, syrups, and also simple mixtures of fructose and glucose. Techniques of near- and mid-IR spectroscopy followed by chemometric processing of data, correctly classify 88–100% of honey authentic samples and samples mixed with those adulterants [12, 13]. The use of isotope analysis as the standard method [53] should be noted once again.

**Wine** [65, 73, 83, 84]. Again, isotopic ratio mass spectrometry is a valuable technique for authentication of wines, e.g., by measuring $^{18}O/^{16}O$ and $^{13}C/^{12}C$ isotope ratios in glycerol, the significant by-product of wine fermentation [73]. There are also other applications of this technique to detect sugar additions and wine watering, and to determine product origin/traceability [83, 85]. It has been proposed that grapevine/wine identification be performed by DNA analysis, e.g., see [86]. Different techniques combined with chemometrics are also outlined in Table 8.5.

**Other foodstuffs and food raw materials** tested for purposes of qualitative analysis II are: coffee [87], essential oils [88], fruit juices [12, 13, 40], meat/fish [71], milk and dairy products [11, 18, 37, 47], organic food [89], tea [90], and transgenic foods [17].

### *8.2.2 Oil Spills*

Oil spills became a global problem many years ago. The characterization/identification of spilled oils is a conclusive part of assessments related to the liability

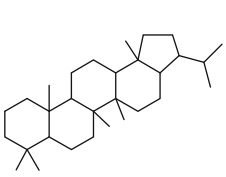
**Table 8.5** Some techniques combined with PCA and DA for qualitative analysis of grape and wine

| Techniques | Purpose |
|---|---|
| Electronic nose | Wine classification, wine spoilage, classification of aroma compounds |
| Electronic nose and tongue | Wine classification |
| NIR | Discrimination of Riesling and Chardonnay wines |
| NIR and MIR | Wine grading |

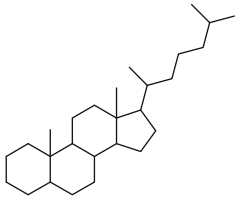
Adapted from [84])

for oil released into the environment. There are a number of methods for such identification. The most reliable methods are based on the combination of chromatography and mass spectrometry [19–21].

Figure 8.4 shows the chart for the accepted protocol of oil spill identification using GC–FID and GC–MS [22]; see similar procedures in [19–21]. According to this methodology, (a) chromatography profiles or abundance ratios of *n*-alkanes, PAH, (b) hydrocarbon biomarkers of terpane series, e.g., hopane **8.1**, $C_{30}H_{52}$, and the sterane series, e.g., cholestane **8.2**, $C_{27}H_{48}$ (different enantiomers of both hydrocarbons), and (c) other characteristic compounds are compared between oil extracts of the spilled sample and suspected sources. Matching corresponding analytical signals implies TP in regard to the suspicion. No match recorded means that the result is negative, but this may be FN if composition differences are caused by weathering and/or degradation of oil in water and soil. Thus, the origin of differences must be made clear, e.g., special supplemental experiments should be carried out with "intact" oil samples undergoing artificial weathering.



**8.1**



**8.2**

There are different types of matching. A semi-quantitative match is sufficient for alkanes and PAH. A certain match is required for diagnostic ratios of biomarker hydrocarbons recorded by GC–MS. The match degree between the spill and an oil source can be estimated by the use of *t*-criterion (3.14) and (3.15); analogous estimations are made with the use of confidence intervals [22]. A significant/*positive match* means that the criterion (3.14) is fulfilled for difference between ratios at the level of $\alpha \geq 0.05$. *Possible match* means that a larger difference in ratio values, at $\alpha \geq 0.02$, is obtained. A significant difference ($\alpha < 0.02$) leads to the conclusion of *no match*. The fourth kind of conclusions an *inconclusive* result, is made when, for example, heterogeneity of oil samples is observed. Many details of this methodology are placed in its updated version [22]; the terms in italics are from this report.

The cited [22] and other recent methodologies have been largely focused on just biomarkers [91] as oil components most resistant to weathering/biodegradation. For example, it was proved in the case of heavily biodegraded oils that diagnostic ratios of some triaromatic steranes and high molecular-mass terpane and sterane did not reveal significant changes during biodegradation [92].

The above methods certainly fit the purpose of identification of not only crude oil but also its fractions (oil fuels), e.g., spilled diesel [68, 70]. For this fuel, it has been demonstrated that the number of biomarker diagnostic ratios can be reduced using PCA [93]. For chemometrics applied to the problem of spilled oils, see also [33].

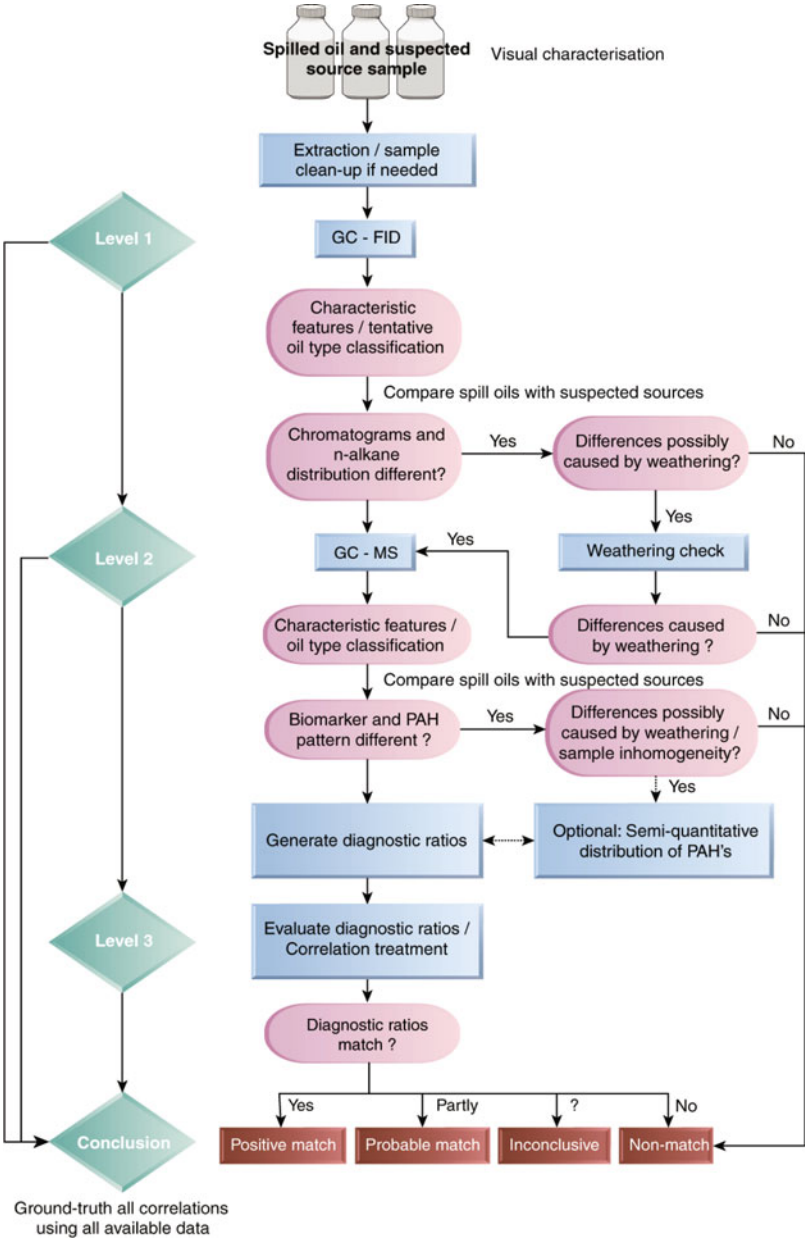



**Fig. 8.4** Oil spill identification flow chart [22] (reproduced by permission)

### 8.2.3 Microorganisms

The possibility of the deployment of biological weapons, including microorganisms (viruses, bacteria, ricketsia, fungi), asks for rapid and efficient methods for characterization of these objects at the level of the species, subspecies, strain, and others. Together with traditional tests and methods of DNA analysis (e.g., see [23]), IR spectroscopy [24–26] and mass spectrometry [27–32] have provided advanced methods for this field of qualitative analysis II.

The identification approach by IR is based on librarian searches [24, 26] or fingerprinting using chemometrical methods of data processing and classification [25]. Spectral database of coryneform bacteria (730 reference strains, covering 220 different species from 46 genera) makes it possible to correctly identify 98% of the validation set of 208 strains at the species level. This was the result of internal validation, where single replicates of strain spectra available in the spectral library were tested vs. the database. The second validation was an internal one, with strains absent in the spectral library. The result was the correct identification of 87 and 95% at the species and genus level respectively [24]. In both cases, the identification criteria were the best matching and high MF values (see Chap. 4). Equally good identification results were achieved in some other researches using this method (see references in [24]).

Recently, approaches based on MS techniques and methods of proteomics (Chap. 7) have become popular.

> Confident identification of an organism can be achieved by top-down proteomics following identification of individual protein biomarkers from their tandem mass spectra. In bottom-up proteomics, rapid digestion of intact protein biomarkers is again followed by MS/MS to provide unambiguous bioagent identification and characterization [32].

The identification is related to organism-specific protein biomarker molecules (the first approach in Table 8.1) and/or specific distribution of a protein group (the third kind of identification, Table 8.1). Unique proteins are recognized by a peptide mass or fragment mass fingerprinting (Sect. 4.4.2.3) or even by direct mass measuring if corresponding sequences are short and have no isomers. Techniques of MALDI [27, 29, 31, 32] and ESI–$MS^n$ [28, 30] have been required for recording mass spectra of intact protein ions and their fragments, respectively.

Multi-peak protein MALDI spectra are processed by means of chemometrics, as well as any other multidimensional data (see above), which results in up to 100% *TPR* and *PPV* [31]. Such spectra can be also used as reference ones in library searches; a library of about 3,500 spectra, with replicate strain ones for most species, was built [29]. The rate of correct identification (*TPR*) with the use of this database ranged between 33 and 100%. The lower percentage was explained by "poor representation of some species within the database" [29]. For bacterial identification based on MALDI and database of protein masses, see also [27].

Identification of bacteria using $MS^2$ was carried out by comparing experimental peptide fragment spectra with those predicted based on the special proteome

database constructed from 87 [28] and 170 [30] sequenced bacterial genomes; true positive results of classification were reported for most test cases.

### 8.2.4 *"Omics"*

Metabolic Sect. 7.7.1 and peptide/protein (proteomics, Sect. 7.7.2) profiles are characteristics of different metabolomes and proteomes, originated from different species, organisms, organs, tissues, populations of cells, individual cells and their different states ("normal" function, dysfunction, good health, disease, status of metabolism, and so on), see [94]. Numerous researches have been carried out to develop and validate methods of profiling. For example, the capability of differential profiling up to 30,000 different ions of both metabolite and peptide molecules by LC–MS has been demonstrated [95].

Multisignal metabolic profiles have been processed by techniques of multivariate statistics to classify/differentiate biosamples, e.g., according to their classes/types, responses of organism to toxicity or disease, and so on [43, 96–101]. Appropriate software has also been engaged to detect the contribution of individual compounds to multidimensional data, e.g., to discover biomarkers [100]. Just as in metabolomic data, MS– and LC–MS-based proteomic profiles/patterns have been used for clinical diagnosis and biomarker discovery, e.g., see [95, 98].

### 8.2.5 *Other Objects*

Methodologies of qualitative analysis II similar to those outlined above have also been applied to paintings [102, 103], counterfeit drugs [104], powder residues [105], and so on.

## References

1. Milman BL (2008) Introduction to chemical identification (In Russian). VVM, Saint Petersburg
2. Valcárcel M, Cárdenas S, Barceló D, Buydens L, Heydorn K, Karlberg B, Klemm K, Lendl B, Milman B, Neidhart B, Ríos A, Stephany R, Townshend A, Zschunke A (2002) Metrology of qualitative chemical analysis. Report EUR 20605. EC, Luxembourg.
3. Valcárcel M, Cárdenas S, Simonet BM, Carrillo-Carrión C (2007) Principles of qualitative analysis in the chromatographic context. J Chromatogr A 1158:234–240
4. Erudium Glossary. http://www.erudium.polymtl.ca/html-eng/glossaire.php#A. Accessed 5 April 2010
5. Princeton University WordNet. http://wordnetweb.princeton.edu/perl/webwn?s=adulterate&sub=Search+WordNet&o2=&o0=1&o7=&o5=&o1=1&o6=&o4=&o3=&h=00. Accessed 4 Nov 2009

6. Aparicio R, Aparicio-Ruíz R (2000) Authentication of vegetable oils by chromatographic techniques. J Chromatogr A 881:93–104
7. Aparicio R, Harwood J (2000) Handbook of olive oil: analysis and properties. Aspen Publishers, Gaithersburg
8. Baeten V, Aparicio R (2000) Edible oils and fats authentication by Fourier transform Raman spectrometry. Biotechnol Agron Soc Environ 4:196–203
9. Osborne BG (2000) Near-infrared spectroscopy in food analysis. In: Meyers RA (ed) Encyclopedia of analytical chemistry. Wiley, Chichester, http://www2.hcmuaf.edu.vn/data/phyenphuong/Near%20Infrared%20Spectroscopy%20in%20Food%20Analysis.pdf. Accessed 4 Nov 2010
10. Nollet L (2004) Food authentication by HPLC. http://www.labint-online.com/fileadmin/pdf/pdf_general/food-authentication-by-hplc.pdf. Accessed 12 June 2010
11. Cserhati T, Forgacs E, Deyl Z, Miksik I (2005) Chromatography in authenticity and traceability tests of vegetable oils and dairy products: a review. Biomed Chromatogr 19:183–190
12. Downey G, Kelly JD (2006) Food authentication using infrared spectroscopic methods. Project RMIS No. 4907 Final report. Ashtown Food Research Centre, Dublin. http://www.teagasc.ie/research/reports/foodprocessing/4907/eopr-4907.pdf. Accessed 5 April 2010
13. Downey G, Kelly JD, Rodriguez CP (2006) Food authentication – has near infrared spectroscopy a role? Spectrosc Eur 18:10–14, http://www.spectroscopyeurope.com/images/stories/ArticlePDfs/NIR_18_3.pdf. Accessed 12 June 2010
14. Sadecka J, Tothova J (2007) Fluorescence spectroscopy and chemometrics in the food classification – a review. Czech J Food Sci 25:159–173
15. Mafra I, Ferreira IMPLVO, Oliveira MBPP (2008) Food authentication by PCR-based methods. Eur Food Res Technol 227:649–665
16. Sun DW (2009) Infrared spectroscopy for food quality analysis and control. Academic, Burlington
17. Alishahi A, Farahmand H, Prieto N, Cozzolino D (2010) Identification of transgenic foods using NIR spectroscopy: a review. Spectrochim Acta A 75:1–7
18. Nollet LML, Toldra F (2010) Handbook of dairy foods analysis. CRC Press, Boca Raton, FL
19. NORDTEST method NT CHEM 001, 2nd edn. Oil spill identification. http://www.nordicinnovation.net/nordtestfiler/chem001.pdf. Accessed 12 June 2010
20. Wang Z, Fingas M, Page DS (1999) Oil spill identification. J Chromatogr A 843:369–411
21. ASTM D 5739 (2006) Standard practice for oil spill source identification by gas chromatography and positive ion electron impact low resolution mass spectrometry
22. Faksness LG, Weiss HM, Daling PS (2008) Revision of the Nordtest methodology for oil spill identification. SINTEF Report STF66 A02028. http://www.nordicinnovation.net/nordtestfiler/tec498.pdf. Accessed 12 June 2010
23. Bohaychuk VM, Gensler GE, King RK, Wu JT, McMullen LM (2005) Evaluation of detection methods for screening meat and poultry products for the presence of foodborne pathogens. J Food Prot 68:2637–2647
24. Oberreuter H, Seiler H, Scherer S (2002) Identification of coryneform bacteria and related taxa by Fourier-transform infrared (FT–IR) spectroscopy. Int J Syst Evol Microbiol 52:91–100
25. Preisner O, Lopes JA, Guiomar R, Machado J, Menezes JC (2007) Fourier transform infrared (FT–IR) spectroscopy in bacteriology: towards a reference method for bacteria discrimination. Anal Bioanal Chem 387:1739–1748
26. Bruker Bacteria Identification. http://www.brukeroptics.com/bac.html. Accessed 12 June 2010
27. Wang Z, Dunlop K, Long SR, Li L (2002) Mass spectrometric methods for generation of protein mass database used for bacterial identification. Anal Chem 74:3174–3182
28. Dworzanski JP, Snyder AP, Chen R, Zhang H, Wishart D, Li L (2004) Identification of bacteria using tandem mass spectrometry combined with a proteome database and statistical scoring. Anal Chem 76:2355–2366

29. Keys CJ, Dare DJ, Sutton H, Wells G, Lunt M, McKenna T, McDowall M, Shah HN (2004) Compilation of a MALDI–TOF mass spectral database for the rapid screening and characterisation of bacteria implicated in human infectious diseases. Infect Genet Evol 4:221–242
30. Dworzanski JP, Deshpande SV, Chen R, Jabbour RE, Snyder AP, Wick CH, Li L (2006) Mass spectrometry-based proteomics combined with bioinformatic tools for bacterial classification. J Proteome Res 5:76–87
31. Hsieh SY, Tseng CL, Lee YS, Kuo AJ, Sun CF, Lin YH, Chen JK (2007) Highly efficient classification and identification of human pathogenic bacteria by MALDI–TOF MS. Mol Cell Proteomics 7:448–456
32. Demirev PA, Fenselau C (2008) Mass spectrometry in biodefense. J Mass Spectrom 43:1441–1457
33. Christensen JH (2005) Chemometrics as a tool to analyse complex chemical mixtures. Environmental forensics and fate of oil spills. PhD Thesis, Roskilde University. http://www2.dmu.dk/1_viden/2_Publikationer/3_Ovrige/rapporter/phd_jch.pdf. Accessed 12 June 2010
34. Golan E, Krissoff B, Kuchler F (2004) Food traceability: One ingredient in a safe and efficient food supply. http://www.ers.usda.gov/amberwaves/april04/pdf/featureFoodTraceability.pdf. Accessed 12 June 2010
35. Dennis MJ (1998) Recent developments in food authentication. Analyst 123:151R–156R
36. Ghidini S, Ianieri A, Zanardi E, Conter M, Boschetti T, Iacumin P, Bracchi PG (2006) Stable isotopes determination in food authentication: a review. Ann Fac Medic Vet di Parma 26:193–204, http://www.unipr.it/arpa/facvet/annali/2006/193_204.pdf. Accessed 12 June 2010
37. Guy PA, Fenaille F (2006) Contribution of mass spectrometry to assess quality of milk-based products. Mass Spectrom Rev 25:290–326
38. Sun DW (2008) Modern techniques for food authentication. Academic, Burlington
39. Fauhl C, Reniero F, Guillou C (2000) $^{1}$H NMR as a tool for the analysis of mixtures of virgin olive oil with oils of different botanical origin. Magn Reson Chem 38:436–443
40. Ogrinc N, Košir IJ, Spangenberg JE, Kidrič J (2003) The application of NMR and MS methods for detection of adulteration of wine, fruit juices, and olive oil. A review. Anal Bioanal Chem 376:424–430
41. Vigli G, Philippidis A, Spyros A, Dais P (2003) Classification of edible oils by employing 31P and 1H NMR spectroscopy in combination with multivariate statistical analysis. A proposal for the detection of seed oil adulteration in virgin olive oils. J Agric Food Chem 51:5715–5722
42. Goodacre R, Kell DB (1996) Pyrolysis mass spectrometry and its applications in biotechnology. Curr Opin Biotechnol 7:20–28
43. Ellis DI, Dunn WB, Griffin JL, Allwood JW, Goodacre R (2007) Metabolic fingerprinting as a diagnostic tool. Pharmacogenomics 8:1243–1266
44. Eggins BR (2002) Chemical sensors and biosensors. Wiley, Chichester
45. Cosio MS, Ballabio D, Benedetti S, Gigliotti C (2006) Geographical origin and authentication of extra virgin olive oils by an electronic nose in combination with artificial neural networks. Anal Chim Acta 567:202–210
46. Vlasov Y, Legin A, Rudnitskaya A, Di Natale C, D'Amico A (2005) Nonspecific sensor arrays ("electronic tongue") for chemical analysis of liquids. Pure Appl Chem 77:1965–1983
47. Karoui R, De Baerdemaeker J (2007) A review of the analytical methods coupled with chemometric tools for the determination of the quality and identity of dairy products. Food Chem 102:621–640
48. Arvanitoyannis IS, Chalhoub C, Gotsiou P, Lydakis-Simantiris N, Kefalas P (2005) Novel quality control methods in conjunction with chemometrics (multivariate analysis) for detecting honey authenticity. Crit Rev Food Sci Nutr 45:193–203

49. Comission regulation (EEC) No 2568/91 of 11 July 1991 on the characteristics of olive oil and olive-residue oil and on the relevant methods of analysis (1991). Off J L 248:1–112. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:1991R2568:20080101:EN:PDF. Accessed 4 Nov 2010
50. Commission Regulation (EC) No 656/95 of 28 March 1995 amending Regulation (EEC) No 2568/91 on the characteristics of olive oil and olive-residue oil and on the relevant methods of analysis and Council Regulation (EEC) No 2658/87 on the tariff and statistical nomenclature and on the Common Customs Tariff (1995) Off J L 069:1–12. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995R0656:EN:HTML. Accessed 4 Nov 2010
51. FAO/WHO Food Standards. Codex Alimentarius. http://www.codexalimentarius.net/web/index_en.jsp#. Accessed 13 June 2010
52. ASTM Standards and Engineering Digital Library. http://www.astm.org/DIGITAL_LIBRARY/index.shtml. Accessed 13 June 2010
53. Official Methods of Analysis of AOAC International, 18th edn (2005, revis 2006) AOAC, Gaithersburg. Ch. 37:21–22, 33–37, Ch. 44:33–36, 38–39
54. Latorre MJ, Peña R, García S, Herrero C (2000) Authentication of Galician (N.W. Spain) honeys by multivariate techniques based on metal content data. Analyst 125:307–312
55. Roussel S, Bellon-Maurel V, Roger JM, Grenier P (2003) Fusion of aroma, FT–IR and UV sensor data based on the Bayesian inference. Application to the discrimination of white grape varieties. Chemom Intell Lab Syst 65:209–219
56. Ruoff K, Luginbühl W, Künzli R, Bogdanov S, Bosset JO, Von der Ohe K, Von der Ohe W, Amado R (2006) Authentication of the botanical and geographical origin of honey by front-face fluorescence spectroscopy. J Agric Food Chem 54:6858–6866
57. Toher D, Downey G, Murphy TB (2007) A comparison of model-based and regression classification techniques applied to near infrared spectroscopic data in food authentication studies. Chemom Intell Lab Syst 89:102–115
58. Downey G (1998) Food and food ingredient authentication by mid-infrared spectroscopy and chemometrics. Trends Anal Chem 17:418–424
59. ASTM E 1790 (2000) Standard practice for near infrared qualitative analysis
60. Downey G, Fouratier V, Kelly JD (2003) Detection of honey adulteration by addition of fructose and glucose using near infrared transflectance spectroscopy. J Near Infrared Spectrosc 11:447–456
61. Corbella E, Cozzolino D (2005) The use of visible and near infrared spectroscopy to classify the floral origin of honey samples produced in Uruguay. J Near Infrared Spectrosc 13:63–68
62. FDA Center for Veterinary Medicine Guidance for Industry Guidance for Industry. Bioanalytical Method Validation (2001). http://www.docstoc.com/docs/24786912/Bioanalytical-method-validation. Accessed 13 June 2010
63. Milman BL (2005) Towards a full reference library of $MS^n$ spectra. Testing of a library containing 3126 MS2 spectra of 1743 compounds. Rapid Commun Mass Spectrom 19:2833–2839
64. Milman BL, Zhurkovich IK (2009) Tandem mass spectral library of pesticides and its use in identification. Proceedings of the 18th International Mass Spectrometry Conference, Bremen
65. Saiz-Abajo MJ, Gonzales-Saiz JM, Pizarro C (2004) Classification of wine and alcohol vinegar samples based on near-infrared spectroscopy. Feasibility study on the detection of adulterated vinegar samples. J Agric Food Chem 52:7711–7719
66. Aparicio R Chemometrics as an aid in olive oil authentication. http://www.eurofedlipid.org/divisions/oliveoil/chemometrics.pdf. Accessed 13 June 2010
67. Sharpless KE, Thomas JB, Christopher SJ, Greenberg RR, Sander LC, Schantz MM, Welch MJ, Wise SA (2007) Standard reference materials for foods and dietary supplements. Anal Bioanal Chem 389:171–178

68. Wang Z, Hollebone B, Yang C, Fingas M, Landriault M. Source identification of an unknown spill (2002) from Quebec by the multi-criterion analytical approach and lab simulation of the spill sample. http://www.iosc.org/papers/IOSC%202005%20a157.pdf. Accessed 13 June 2010
69. Lísa M, Holčapek M (2008) Triacylglycerols profiling in plant oils important in food industry, dietetics and cosmetics using high-performance liquid chromatography-atmospheric pressure chemical ionization mass spectrometry. J Chromatogr A 1198–1199:115–130
70. Wang Z, Yang C, Hollebone B, Brown C, Landriault M. Source identification of spilled diesel using diagnostic sesquiterpanes and diamondoids. http://www.iosc.org/papers/2008%20051.pdf. Accessed 13 June 2010
71. Kelly S (2007) The development of methods to determine the geographical origin of poultry. UK Food Standards Agency Report Q01086. http://www.foodbase.org.uk/admintools/reportdocuments/268-1-489_Q01086%5BIFR%5D_GEOPOULTRY_Final_Report.pdf. Accessed 13 June 2010
72. Hu P, Liang QL, Luo GA, Zhao ZZ, Jiang ZH (2005) Multi-component HPLC fingerprinting of Radix Salviae Miltiorrhizae and its LC–MS–MS identification. Chem Pharm Bull 53:677–683
73. Jung J, Jaufmann T, Hener U, Münch A, Kreck M, Dietrich H, Mosandl A (2006) Progress in wine authentication: GC–C/P–IRMS measurements of glycerol and GC analysis of 2, 3-butanediol stereoisomers. Eur Food Res Technol 223:811–820
74. Phillips KM, Wolf WR, Patterson KY, Sharpless KE, Amanna KR, Holden JM (2007) Summary of reference materials for the determination of the nutrient composition of foods. Accred Qual Assur 12:126–133
75. INFOODS. http://www.fao.org/infoods/index_en.stm. Accessed 13 June 2010
76. Merchant AT, Dehghan M (2006) Food composition database development for between country comparisons. Nutr J 5:2. doi:10.1186/1475-2891-5-2
77. USDA Agricultural Research Service. Nutrient Data Laboratory. http://www.ars.usda.gov/main/site_main.htm?modecode=12-35-45-00. Accessed 13 June 2010
78. Speight JC (1999) The chemistry and technology of petroleum, 3rd edn. Marcel Dekker, New York
79. Ulberth F, Buchgraber M (2000) Authenticity of fats and oils. Eur J Lipid Sci Technol 102:687–694
80. Peña F, Cárdenas S, Gallego M, Valcárcel M (2005) Direct olive oil authentication: detection of adulteration of olive oil with hazelnut oil by direct coupling of headspace and mass spectrometry, and multivariate regression techniques. J Chromatogr A 1074: 215–221
81. Buchgraber M, Ulberth F, Emons H, Anklam E (2004) Triacylglycerol profiling by using chromatographic techniques. Eur J Lipid Sci Technol 106:621–648
82. Lee PJ, Di Gioia AJ (2009) Rapid seed oil analysis using UPLC for quality control and authentication. Lipid Technol 21:112–115
83. Flamini R, Panighel A (2006) Mass spectrometry in grape and wine chemistry II: The consumer protection. Mass Spectrom Rev 25:741–774
84. Cozzolino D, Cynkar WU, Shah N, Dambergs RG, Smith PA (2009) A brief introduction to multivariate methods in grape and wine analysis. Int J Wine Res 2009:123–130
85. Fauhl-Hassek C (2009) Trends in wine authentication. Bull de l'OIV 82:93–100
86. Siret R, Boursiquot JM, Merle MH, Cabanis JC, This P (2000) Toward the authentication of varietal wines by the analysis of grape (*Vitis vinifera* L.) residual DNA in must and wine using microsatellite markers. J Agric Food Chem 48:5035–5040
87. Downey G, Briandet R, Wilson RH, Kemsley EK (1997) Near- and mid-infrared spectroscopies in food authentication: coffee varietal identification. J Agric Food Chem 45:4357–4361
88. Hanneguella S, Thibault JN, Naulet N, Martin GJ (1992) Authentication of essential oils containing linalool and linalyl acetate by isotopic methods. J Agric Food Chem 40:81–87

89. Molkentin J (2009) Authentication of organic milk using $\delta^{13}C$ and the α-linolenic acid content of milk fat. J Agric Food Chem 57:785–790
90. Engelhardt UH (2007) Authenticity of tea (*C. sinensis*) and tea products. ACS Symp Ser 952:138–146
91. Wang Z, Stout SA, Fingas M (2006) Forensic fingerprinting of biomarkers for oil spill characterization and source identification. Environ Forensics 7:105–146
92. Yang C, Wang Z, Hollebone B, Brown CE, Landriault M. Application of statistical analysis in the selection of diagnostic ratios for forensic identification of an oil spill source. http://www.iosc.org/papers/2008%20052.pdf. Accessed 13 June 2010
93. Gaines R, Hall G, Frysinger G, Gronlund W, Juaire K (2006) Chemometric determination of target compounds used to fingerprint unweathered diesel fuels. Environ Forensics 7:77–87
94. Lay JO, Borgmann S, Liyanage R, Wilkins CL (2006) Problems with the "omics". Trends Anal Chem 25:1046–1056
95. Roy SM, Becker CH (2007) Quantification of proteins and metabolites by mass spectrometry without isotopic labeling. Methods Mol Biol 359:87–105
96. Dunn WB, Bailey NJC, Johnson HE (2005) Measuring the metabolome: current analytical technologies. Analyst 130:606–625
97. Idborg H, Zamani L, Edlund PO, Schuppe-Koistinen I, Jacobsson SP (2005) Metabolic fingerprinting of rat urine by LC/MS Part 2. Data pretreatment methods for handling of complex data. J Chromatogr B 828:14–20
98. Hilario M, Kalousis A, Pellegrini C, Müller M (2006) Processing and classification of protein mass spectra. Mass Spectrom Rev 25:409–449
99. Kind T, Tolstikov V, Fiehn O, Weiss RH (2007) A comprehensive urinary metabolomic approach for identifying kidney cancer. Anal Biochem 363:185–195
100. Styczynski MP, Moxley JF, Tong LV, Walther JL, Jensen KL, Stephanopoulos GN (2007) Systematic identification of conserved metabolites in GC/MS data for metabolomics and biomarker discovery. Anal Chem 79:966–973
101. Ramautar R, Somsen GW, De Jong GJ (2009) CE–MS in metabolomics. Electrophoresis 30:276–291
102. Clark RJH (2002) Pigment identification by spectroscopic means: an arts/science interface. C R Chimie 5:7–20
103. Zadrożna I, Połeć-Pawlak K, Głuch I, Ackacha MA, Mojski M, Witowska-Jarosz J, Jarosz M (2003) Old master paintings – A fruitful field of activity for analysts: Targets, methods, outlook. J Sep Sci 26:996–1004
104. Olsen BA, Borer MW, Perry FM, Forbes RA (2002) Screening for counterfeit drugs using near-infrared spectroscopy. Pharm Technol 26:62–71
105. MacCrehan WA, Smith KD, Rowe WF (1998) Sampling protocols for the detection of smokeless powder residues using capillary electrophoresis. J Forensic Sci 43:119–124

# Chapter 9
# Good Identification Practice

**Abstract** Good identification practice is considered as an underlying system of particular requirements and guidelines with regard to laboratories, personnel, instruments, and methods directed to quality assurance and quality control of chemical identification (qualitative analysis). Terminological standardization and "metrologization" of qualitative analysis are stated to be general prerequisites for consistency and comparability of identification results between analytical/bioanalytical chemists and laboratories. Requirements and guidelines concerning quality assurance and control of identification procedures which are contained in official laboratory guidances are considered. According to principles of good identification practice, criteria for detection and identification in target methods, screening and confirmatory ones, should be formulated and validated. Accepted levels of false result rates are established. In non-target/unknown analysis, approaches to identification should be validated, which include evaluation of pertinent databases, spectral libraries, predictor programs, identification/classification algorithms, and so on. Interlaboratory studies provide assessment of laboratory performances and evaluation (validation) of identification methods/approaches.

## 9.1 General

Processes for obtaining true, unambiguous, accurate, precise, and reproducible results of chemical analysis, i.e., high-quality analytical data have been supported by both requirements to performance and quality of analytical procedures incorporated in analytical methodology itself, and supplemental quality assurance and quality control (QA/QC) programs [1–3] based on international [4] and national standard requirements and related to the GLP principles [5]. Different aspects of QA/QC as applied to chemical identification and other procedures of qualitative analysis are shown in Fig. 9.1 and will be considered in this chapter, starting with issues of terminology and metrology.

B.L. Milman, *Chemical Identification and its Quality Assurance*,
DOI 10.1007/978-3-642-15361-7_9, 

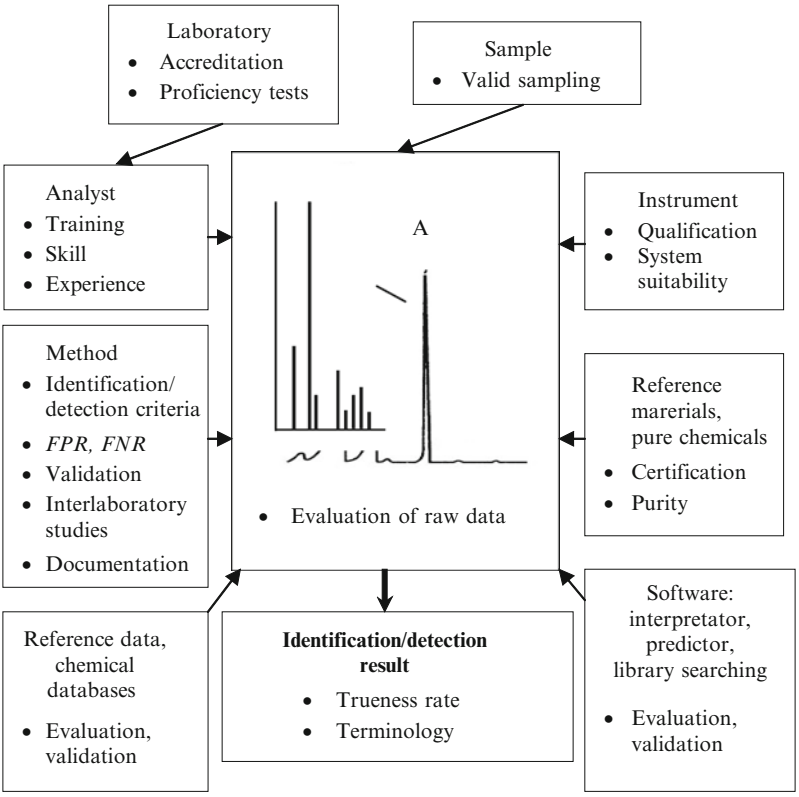



**Fig. 9.1** An underlying system of good identification practice in implementation of special procedures of qualitative analysis in combination with different aspects of QA/QC. Methods are intended for target analysis (Chap. 5). Searches in chemical databases, matching reference chromatographic and spectral data, spectral interpretation, prediction of spectra, and ultimately co-analysis with reference materials, are used in non-target/unknown analysis (Chap. 7)

## 9.2 Standardization of Terminology

As in any other field of science, technology, and practice, qualitative chemical analysis needs clear, definite, and unambiguous terminology. Current problems with regard to the terminology can be divided into three groups.

1. Concepts and terms are absent in basic glossaries such as IUPAC Compendium on Analytical Nomenclature [6] and Compendium of Chemical Terminology [7]. This is particularly true for *identification* itself, although there is a good definition for qualitative analysis [8]:

   Analysis in which substances are identified or classified on the basis of their chemical or physical properties, such as chemical reactivity, solubility, molecular weight, melting point, radiative properties (emission, absorption), mass spectra, nuclear half-life, etc.

2. Concepts and terms are ambiguous alone or in relation to other ones. This can be exemplified by *sensitivity* and *specificity*, which are equivocal (Sect. 4.2.2). Another case is type I (type II) error which may be FP or FN (FN or FP) depending on what qualitative procedure, detection or identification, is considered (Sect. 3.6).
3. There are multiple terms for the same concepts (synonymy). The main example is *reliability of identification* (Sect. 4.1). This term was selected from numerous synonyms and eventually preferred by the author. Other kinds of such disorder occur among statistical terms, e.g., both *sensitivity* and *recall* signifies *TPR* (Table 4.3).

In this book, the author selected one or another synonym on the traditional or logic base, or used several terms if the choice was not easy. In any case, an individual worker may prefer and suggest various concepts, terms, and notations, but they are discussed, modified, and eventually accepted or rejected by the chemical and metrological community. In a good way, a terminology is established in the spirit of its harmonization in accordance with international standards (e.g., see [9]).

## 9.3 Metrology for Chemical Identification

Issues of theoretical metrology, traceability and nominal scales, were treated in Chap. 1. Here, other (more practical) metrological aspects of chemical analysis will be considered.

***Reference materials (analytical standards).*** These are used to obtain reference data (Chap. 7) and for the purpose of co-analysis (Sects. 1.6, 5.2, and 7.1). Various RM [10, 11] are of different metrological quality/status. The top place of the metrological hierarchy is occupied by RM produced and/or certified by National or International metrological institutions (CRM, standard RM):

- The Institute for Reference Materials and Measurements (IRMM, Europe) [12]
- National Institute of Standards and Technology (NIST, USA) [13]
- LGC (UK) [14]
- Federal Institute for Materials Research and Testing (BAM, Germany) [15]
- The Ural Research Institute for Metrology (UNIIM) [16] and D.I. Mendeleyev Institute for Metrology (VNIIM, both Russia) [17]
- National Metrology Institute of Japan (NMIJ, Japan) [18], and so on [11]

The fact that these standards are of the highest quality first of all means the high accuracy of certified values of the target compound amount. There are no doubts about the identity of the targets in CRM/standard RM, and the same also applies to most pure chemicals produced by numerous chemical companies. However, the use of CRM in qualitative analysis is limited because they are produced only for a minority of known chemicals. Commonly, pure chemicals are analytical standards

(references) in identification operations. There should be periodical monitoring of their purity according to the special standard operating procedure developed in the laboratory [19].

There are not less than about two thousand main suppliers of commercial chemicals in the world [20]. There were more than 42 million commercially available chemicals in June 2010 [21]. The accurate number of chemicals with unique identifiers is unknown, because the above number is the sum of records in catalogs. However, the number of unique commercial substances is obviously measured in millions (5.7 million "different products" in June 2010 [20]). So the vast majority of abundant analytes can be identified with the use of chemicals available in the market.

In order to find out whether chemicals as analytical standards are accessible to analysts, it is appropriate to search for information on them and their vendors in various chemical databases (Fig. 9.2). In such searches, databases and sites compiling e-catalogs of many companies, e.g., Chemcats (ACS, USA) [21] and ChemExper (Belgium) [20] (see also Fig. 9.2) are especially useful.

Shortcomings of many analytical standards are that concomitant impurities are not specified. Sensitivity of analytical instruments in reference to some impurity compounds may be higher than for base components. Thus, the signal of an impurity may be the most intensive, e.g., in the cases of specific detectors in GC or highly ionizable compounds in ESI MS. This results in a false identification.

In chemical analysis, matrix effects, i.e., an influence of matrix interfering compounds on intensity and the shape of an analytical signal, are widespread. In turn, this leads to both inaccurate results of quantitative analysis and false identification. To avoid matrix effects, matrix-matched standards of the target and suspected analytes should be used in identification procedures, e.g., for confirmation [24].

For progress in proteomics, it is important to develop protein standard samples [25] (see Sect. 7.7.2).

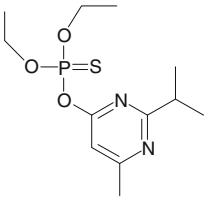
Chemicals used for calibration of analytical instruments are other kinds of references essential for qualitative analysis (see below).

**Accuracy in measurand values.** Accuracy (trueness and precision) of measurement results directly determine the quality of experimental and reference data and therefore reliability of identification. The effect of inaccurate data is briefly as follows.

- Low accuracy, i.e., a large bias of experimental or reference values to each other means a real possibility of a significant divergence between them for the same compound, followed by FN.
- Low precision, i.e., a large uncertainty of experimental or reference values means a possibility of apparent similarity between values of different compounds, which results in FP or ambiguous identification.

The accuracy of measurements for qualitative analysis is provided by the timely calibration of analytical tools (e.g., see [26, 27]). Corresponding technical aids and chemicals are listed in Table 9.1. In spectrometers, scales of wavelengths (UV-Vis),

**a**



**b**

APAC Sourcing Solutions: 827704
Aurora Feinchemie: kasf-167599, kasf-126617
Bosche Scientific: D5663
MicroSource: 00330017
National Cancer Institute: 8938
NCI Plated 2007: 8938
PubChem: 3017, 180695
spectrum: 00330017
Toronto Research Chemicals: D416882
ref: mol2, SDF, SMILES, flexibase
3.58,8.62,-17.2,0,5,0, 304.352, 7

Similar to: 5575979, 5649524, 5862152, 5862888, 6070194

**c**

Substance Vendors: 9 Links
ChemSpider ( 2 )
SID 36528219 - External ID: 13861378
SID 29222164 - External ID: 2909
Sigma-Aldrich ( 6 )
ZINC ( 1 )
SID 58106937 - External ID: ZINC00001309

**d**

Click on a product name to get more information on that compound, on a supplier name to get more information on that supplier.

| Supplier | Description | Reference | | |
|---|---|---|---|---|
| orgchem | Diazinon | 333-41-5 | on request | **Get offer** |
| orgchem | Diazinon | on request | | **Get offer** |
| sinoqf | Diazinon, 98% | on request | | **Get offer** |

**Fig. 9.2** (**a**) Formula of the pesticide diazinon and examples of information about its vendors in chemical databases, (**b**) ZINC [22], (**c**) Pubchem [23], and (**d**) ChemExper [20]. There are names of different databases and sites of chemical companies and hyperlinks to them

wave numbers (IR), chemical shifts (NMR), and $m/z$ values (MS) are calibrated. In high-resolution spectrometers, the foremost NMR and MS instruments (Chap. 2), high-precision scales are established.

High-performance chromatography, including GC with capillary columns and HPLC/UPLC, is somewhat analogous to high-resolution spectrometry. In both cases, a better separation of peaks is obtained and thinner peaks are recorded than

**Table 9.1** Standards for calibration and testing of analytical instruments

| Technique | Standard |
|---|---|
| UV-Vis, NIR | Light-emitting sources, e.g., mercury and deuterium lamps, filters, solutions of potassium dichromate, rare earth compounds, and others [28–31] |
| IR | IR-emitting sources, filters, polystyrene films [32, 33] |
| NMR | Tetramethylsilane ($^{1}$H, $^{13}$C, $^{29}$Si), trichlorofluoromethane ($^{19}$F), $H_3PO_4$ ($^{31}$P), and so on [34, 35] |
| MS | Perfluorotributylamine, perfluorokerosene and other fluorine organic substances (EI); CsI, polyethylene glycol and its ethers, other chemicals (ESI); peptides, proteins, other compounds, and special mixtures (MALDI) [34, 36–40] |
| GC[a] | Grob mixture: fatty acid methyl esters $C_{10}$–$C_{12}$, 2,3-butanediol, dicyclohexylamine, 2,6-dimethylaniline, 2,6-dimethylphenol, 2-ethylhexanoic acid, nonanal, 1-octanol, undecane, decane; other mixtures [41] |
| HPLC, UPLC | NIST standard RM: amitriptyline hydrochloride, ethylbenzene, quinizarin, toluene, uracil, and other mixtures[a] [42, 43]; mixture of five peptides[b] [40] |

[a]For controlling the performance of chromatography columns and chromatography systems at a whole
[b]For standardization of RT in LC–MS

in conditions of low resolution when older/simpler instruments are used. It is correspondingly important that accurate retention times can be measured. Chemical mixtures for controlling the performance and standardization of chromatography system are also stated in Table 9.1. For standard chemicals used in calculations of RI, see Sects. 7.2 and 7.3.

## 9.4 Instrumental Parameters

In target identification by methods, instrumental parameters should be optimized, standardized, and documented. In pharmaceutical/biochemical analysis and some other analytical fields, the concept of system suitability has been introduced in laboratory practice and QA programs [24, 44–47].

> *System Suitability* - The fitness of analytical instruments for the purpose at hand, based on manufacturer specifications, instrumental Standard Operating Procedure, or specific requirements of the analytical method. Suitability may be established through verification of relevant instrumental parameters such as calibration, pressure, flows, temperature, multiplier gain, etc., or through verification of method-specific parameters such as signal-to-noise level for a known amount injected, peak shape, test spectra, etc. [45].

There are standard tests for system suitability, e.g., monitoring of gas chromatographs with special chemical mixtures for pesticide analysis [46], which are or may be a part of analytical protocols, as well as any routine conditions of maintenance of analytical equipment. One or another version of system suitability testing, e.g.,

verification of resolution in chromatography or sensitivity in spectrometric methods, are required also in non-target analysis. Insufficiently good instrumental parameters may lead to non-detection of unknown analytes or their false identification, i.e., one or other FN/FP.

System suitability tests are related to some degree with an overall process of instrument qualification [47].

As modern analytical instruments are under computer control, computer/computer system validation is also essential (see [48]).

## 9.5 Laboratory Practice and Quality Assurance

Commonly, chemical identification (qualitative analysis) has been poorly considered in general treatises on QA/QC. Many requirements for QA of qualitative analytical procedures can be taken from some special documents.

**Special documents related to qualitative analysis.** One of the first guides issued for the purpose under consideration was the LGC document [49], which placed procedures of qualitative analysis in the context of common laboratory practice. Below are some extracts from this guidance, with more details referring to identification operations, there named *classification*.

> **Definition.** *Qualitative Analysis*: The classification of objects against specified criteria to meet an agreed requirement.
>
> **Health and Safety.** Laboratory staff should be made aware of any potential dangers associated with the collection, analysis or storage of sample materials.
>
> **Establishing the Requirements. An Agreed Requirement.** The analyst and customer should agree on the business requirement and the technical solution.
>
> **The Customer Requirement.** The analyst and customer need a clear mutual understanding of the requirement to ensure that the work done meets the customer's needs.
>
> **The Technical Requirement.** The methodology selected should be technically capable of satisfying the business needs… Criteria should be selected to give demonstrably acceptable performance against the business needs. The performance of the methodology on the subject materials must therefore be taken into account when selecting criteria and setting decision criteria. The test method must itself satisfy performance criteria which will ensure that it is capable of establishing whether the technical criteria have been met.
>
> **The sample. Collection.** Samples should be collected in such a way as to provide a representative portion of the source material free from contamination by the sampling process. In addition, the sampling process should not contaminate the source material.
>
> **Containment.** Sample containers should provide a safe and secure environment for the storage and transport of their contents. A container should not in any way alter the composition of its contents.
>
> **Preservation.** The sample should remain unaltered with respect to the unknown(s) of interest between the time of collection and the time of analysis.
>
> **Identification.** All samples should be uniquely identified [i.e., labeled – author] in some way.
>
> **Documentation.** All materials intended for analysis should be accompanied by sufficient information so as to enable their correct storage, appropriate analysis and safe handling.

**Packaging.** Where sample containers are required to be packaged for transportation, the method of packaging and the packaging materials themselves should be chosen so as to minimize the likelihood of damage to the containers or contamination of their external surfaces.

**Transportation.** The method of transportation should be chosen with regard to the stability of the sample or components of interest and any known hazards.

**Receipt.** Laboratories should have a clear and workable procedure for the reception of samples.

**Opening.** Sample containers should be opened in a manner which maintains the integrity of their contents and ensures the safety of staff.

**Sub-sampling.** Where it is necessary to take a sub-sample from an existing sample, such sub-samples should be uniquely identified and should be traceable to the parent sample.

**Storage.** Material received for analysis should be stored in a safe and secure manner and under conditions which preserve its integrity.

**Disposal.** Laboratories should have a policy covering the retention and disposal of analyzed samples.

**The Analytical Process. Classification Criteria**. The classification criteria should be sufficiently well defined to enable an unambiguous result to be obtained by appropriate methods. Qualitative analysis is based on *criteria*. Criteria may be, for example, the presence of a particular analyte above a specified level, physical properties within particular limits, a match between two spectra, or combinations of features. Whatever the nature of the criteria, however, it is important that the criteria be unambiguous, clearly stated and, as far as is possible, objective. For example, "the melting point should match the reference value" is insufficient because it does not specify the degree of match acceptable, "The melting point should match the reference value *within 1°C and the material should melt entirely within a range less than 0.5°C*" is a clear, unambiguous and objective statement. . .The criteria should demonstrably distinguish adequately between different classes and permit unambiguous assignment of an object or material. The criteria used to interpret each category of observational or experimental data should be recorded and available for reference by relevant staff as required.

**Method Selection.** All methods used for qualitative analysis (and also for quantitative analysis) should be documented and fit for their intended purpose.

**Method Validation. Specificity.** The specificity of a test for identity should always be known.

**Detection Limit.** Any qualitative test employed must be sensitive enough to detect the species of interest at the concentration levels of interest.

**Misclassification Rate.** Rates of misclassification must be known and controlled.

**Computations.** There should be a policy for reviewing computational procedures and checking the correctness of results obtained.

**Confirmation.** If appropriate, a confirmatory test should be employed to substantiate the conclusions from the primary test or tests. It is important that any confirmatory test used is completely independent of the primary test. It should also be noted that, just as it is possible to obtain a false positive result, it is also possible to obtain a false negative result. The use of confirmatory tests is therefore not restricted to confirming only positive results.

**Quality Assurance. Environment.** The working environment should be actively controlled to ensure that all relevant parameters are within appropriate limits.

**Equipment.** Equipment used in the course of an analysis should be well maintained and operated within its design parameters.

**Reagents.** The purity of all reagents employed in a method should be known, at least approximately, and all materials used should be tested to ensure they do not contain substances which would interfere in the analytical method.

**Operators.** All analyzes should be conducted by personnel who have demonstrated the relevant competencies.

**Documentation of Methods.** Analytical methods should be documented sufficiently well as to enable their successful use by competent analysts unfamiliar with them.

**Laboratory Records.** A record should be kept of all observations made and of all data generated during an analysis. All information necessary to perform the analysis should also be recorded.

**Quality Control.** Quality control measures should be in place which cover all activities likely to affect a result.

**Independent Assessment.** There should be independent assessment of performance.

**Interpretation of Results.** The interpretation of observations or experimental data should provide a classification consistent with the available information... Interpretation criteria will typically specify the basis for prediction and the quality of match between prediction and observation... Interpretation of observations or experimental data also requires some skill and in some instances, considerable skill... Interpretation should be consistent with the observations, the established criteria and relevant quality control and assurance data.

**Analysis Report.** The analysis report should clearly indicate to the customer how the analysis has met his/her requirements.

**Field Testing.** Field testing is subject to additional problems to those found in the laboratory and staff working in the field should be aware of these before they set out [49].

Another important document treating qualitative analysis in relation to the QA/metrology principles was issued as the report on the MEQUALAN project [26, 27]. Some principal conclusions are as follows.

- The traceability concept is applicable to qualitative analysis, which requires an availability of analytical standards, e.g., pure compounds and matrix standards.
- Calibration of analytical instruments as the metrological procedure is necessary not only in quantitative analysis but also for identification.
- *Unreliability* as the analog of the term of *uncertainty*, especially for qualitative analysis, has been proposed. The unreliability region is the low concentration interval where the percentage of false results is high (Sect. 4.3).
- To validate qualitative/identification methods, selectivity/specificity, the limit of detection, *FPR*, and *FNR* and other characteristics should be evaluated.
- Control cards are proposed for internal quality control of screening yes–no procedures.
- In external quality control, special proficiency tests are performed where false results are counted and further summed to obtain the value of *z*-score [26, 27].

The aspects of QA outlined in MEQUALAN were further developed [50–52].

**Laboratory Guides.** Modern analytical guidances (see Sect. 5.5) contain a lot of details with regard to QA/QC requirements. The FDA guidance on MS applied for confirmatory of identity of animal drug residues [45] specified the following requirements for QC:

- Preliminary establishment of system suitability (see above), and reanalysis of samples if system suitability was not adequate
- At least one negative control and one fortified control sample injected per day, meeting fail and pass identification criteria respectively

- Provision of the control that analysis of standards/fortified control samples is not followed by carryover causing false positives
- The specification and use of confirmation criteria without any substitution of the criteria after analytical experiments have been performed

In *ad hoc* cases where unanticipated situations arise and fully validated procedures are unavailable, additional procedures for control samples are prescribed, and good training and high expertise in the laboratory are of great concern [45]. It should be noted that this is the same case as for unknown/non-targeted analysis (Chap. 7).

Forensic toxicology guidelines [19] and other documents direct attention to proficiency tests (see below) which highlight false results. False positive errors treated as more serious, and also false negative results, should be investigated. As for control samples commonly used in QA/QC, it is proposed that each batch of specimens under analysis should include at least 10% positive and negative controls. There may be open (identity is known to the analyst) and blind (identity is unknown to the analyst) controls. The latter is more suitable for maintaining quality control [19].

**Data quality.** The modern principles of GLP have been also extended to acquisition and processing of electronic raw data [53].

There are also such aspects of QC/QA as the particular quality of experimental raw data obtained in chemical analysis (chromatograms, spectra). The guide [19] states that the validity of analytical data, e.g., shape and signal-to-noise ratio of chromatographic peak, should be reviewed by scientific personnel. Verification of chromatographic peak purity may show how unambiguous are identification results. Low-quality mass spectra having high noise, low S/N ratios, and anomalous/irreproducible relationships between peak intensities should be paid special attention. Such spectra cannot provide reliable identification, and therefore must be excluded from data processing and analysis. This is of special value for determination of low amounts of complex molecules (proteomics, see Sect. 7.7.2).

## 9.6 Validation of Methods and Approaches

According to the ISO/IEC 17025 standard [4],

> Validation is the confirmation by examination and the provision of objective evidence that the particular requirements for a specific intended use are fulfilled.

There may be (a) validation of methodology,[1] data, and samples, and (b) in-house validation and interlaboratory studies [1–3, 55–59]. For the particular aspects of method validation in qualitative analysis, see [47, 50, 52, 60, 61]. For

[1]There is also the *method verification* term, see [54]. A verification is a series of tests demonstrating that a standard method has specified performances when a laboratory starts to use it.

unknown/non-target analysis, validated methods are absent by definition. Here an analyst faces the challenge of validation of some kind of general methodology, i.e., various rather general approaches to identification.

### *9.6.1 Methods*

For validation of screening and confirmatory qualitative methods, the European Commission Decision [62] (see Chap. 5) established that their following performance characteristics have to be determined.

- *Detection limit CCβ* and *decision limit CCα* (only for confirmatory methods); see Sect. 4.3.2.1.
- *Selectivity/specificity*. A validation of these characteristics results in the power of differentiation between the analytes and closely related substances such as isomers, transformation products, and matrix components, being tested. Effects of the interferences and chances of false identifications are examined for blank samples and those fortified with the proper amounts of compounds which may interfere with the identification (and quantitative determination) of the analytes.
- *Ruggedness (robustness* is used in other documents). Poor ruggedness means that an analytical method is susceptible to changes in experimental conditions which may include the composition of the sample, sample preparation conditions, the reagent stability, pH, temperature, etc. Effects of these changes on parameters of methods should be studied and indicated.

According to [62], it is mandatory to estimate *FPR* at the *CCβ* level in *screening* procedures; the rate is limited (<5%). Evaluation of other true/false result rates was not explicitly specified.

In the guidance for method validation and QC of pesticide determinations [63], a *screening FNR* threshold (not specified in [62]) of 5% at the concentration of interest is recommended. There should be a duplicate analysis of ten different samples referring to each group of commodities; samples are spiked with analytes at the lowest concentration level. FP should be excluded when analyzing unspiked/blank samples. However, if confirmation is further applied, the limitation for the number of FP is not strict. Usually, screening positives require confirmation, and negatives need no confirmation [63].

*Confirmation* of identity is best performed using MS (Chaps. 5 and 7). The corresponding guidance, intended for determination of animal drug residues, contains the following recommendations [45].

- Five control samples are taken for in-house validation. These are duplicated for an interlaboratory study, with two concentration levels and five samples at each level.
- What should be demonstrated are (a) zero *FPR* for negative control and (b) $FNR \leq 10\%$ at or above the tolerance/safe concentration level for fortified and incurred samples.

- Validation of ruggedness/robustness is provided with the demonstration that the suitable rates come on 2 or more days. This helps to ensure that the method is rugged and under control.
- Validation for specificity is that the analyte should be demonstrated not to be contaminated by other animal drugs and matrix components of control samples from two and more individual animals.

The last guideline hold true for other types of analytes and matrices, though the number of samples taken for validation of selectivity/specificity may be different. In bioanalytical methods, there is the necessity for analyzes of blank samples of the appropriate biological matrix (plasma, urine, and so on) obtained from six or more sources [64]. In the case of LC–ESI–$MS^n$, matrix effects affect ion currents and relative intensities of different component ions of samples (e.g., see [65]), and should therefore be investigated throughout method validation [47, 64].

It is specially noted that validation of MS methods should be carried out for all analytes, with necessary estimation of false result rates for blind fortified and blank samples [66].

There are both quantitative and qualitative aspects of selectivity/specificity. Interference of foreign matrix components with an analyte changes not only the amplitude of the analytical signal, e.g., chromatography peak height/area, but also the corresponding spectrum if a combination of chromatography with mass spectrometry is used. So it is important to confirm not only that this effect is insignificant in terms of quantitative determination (response in blank samples should be $<30\%$ of limit of quantitation [63]) but also that spectral distortion will be sufficiently small not to significantly change abundances of characteristic ions (see identification criteria, Sect. 5.5.3).

### *9.6.2 Approaches*

In non-target analysis (Chap. 7), at least at the start of identification procedures, valid methods may be unavailable "by definition." In such analyzes, different versions of four general approaches (Sect. 1.6) are applied, and one or another *ad hoc* method is adjusted. To validate approaches to identification, corresponding constituent parts such as individual operations, databases, algorithms, and so on, should be evaluated. These are:

- Setting up identification hypotheses by using prior data from chemical databases (Chap. 6 and Sect. 7.4.2)
- Reference GC RI (see Sect. 7.2)
- Mass spectrometry libraries (Sect. 7.4.1)
- Generation of candidate formulas by HRMS (Sect. 7.4.2)
- Prediction of NMR spectra (Sect. 7.6)
- Algorithms of protein identification (Sects. 4.4.2.3, 4.5.4.3, 7.4.1.4, and 7.7.2)
- Chemometrical methods for classification in qualitative analysis II (Sect. 8.1.2), and others

The respective approaches are mainly based on computer databases and software. Therefore, unknown identification approaches are tending to be more valid with the progress in chemoinformatics, e.g., as chemical and related databases become more complete, correct, and updated. This is particularly significant for biochemical analysis (e.g., see [67, 68] and references in Sect. 7.7).

> As with any "omics" discipline, metabolomics is highly dependent on the availability and quality of electronic databases. Furthermore, because metabolomics combines molecular biology with chemistry and physiology, there is a need for not just one type of database, but a wide variety of electronic resources [68].

For chemical databases, it is important that compound identifiers are unambiguous, and that origins of chemical compounds are specified to differ between real-world and virtual molecules.

As methods of non-target analysis are difficult to standardize and proper identification results certainly need some kind of verification and confirmation, minimum reporting standards (sample preparation, experimental analysis, quality control, and so on) proposed for metabolomics [69] seems to be a good idea which is applicable in other cases of qualitative chemical analysis. Another general approach to unknown identification is that initial *ad hoc* identification procedures performed by multiple analytical techniques and statistical calculations are followed by confirmatory co-analysis with reference materials (Sects. 1.6, 5.2, and 7.1). The last step is usually inapproachable in proteomics, where standards of most proteins are not available.

Unknown analysis is difficult to describe in terms of standard operation and formal requirement for method validation and laboratory accreditation. However, this kind of qualitative analysis is indirectly taken into account in the guidance for accreditation.

> It is accepted that sometimes it is not practicable for laboratories to use a fully documented method in the conventional sense, which specifies each sample type and determinand. However the laboratory must have a generic method or procedure for the use of the instrument in question, which includes a protocol defining the approach to be adopted when different sample types are analyzed. Full details of the procedures, including instrumental parameters and ad hoc validation, must be recorded at the time of each analysis such as to enable the procedure to be repeated in precisely the same manner at a later date [70].

## 9.7 Proficiency Tests, Interlaboratory Comparisons

These procedures are intended for (e.g., see [3, 71]):

- QA/QC of participating laboratories, assessment of laboratory performances
- Development/evaluation/validation/ of analytical methods and individual constituents of methodologies
- Certification of reference materials
- Gaining experience, and so on

Reference/control/test and similar laboratories must participate regularly in proper proficiency tests [19, 44, 63]. As a rule, this refers to quantitative determination. Nevertheless, requirements for qualitative operations and procedures to be studied in interlaboratory experiments also appear. For example, levels of FP and FN should be found out in interlaboratory studies related to pesticide determination [63]. Earlier, the count of false results was recommended for $z$-scoring in such tests [26, 27, 72]. More complicated scoring systems have also been proposed [73, 74].

In the case of qualitative analysis performed by the microbiological method, requirements for interlaboratory studies intended for evaluation of laboratory and methodical performances are as follows [44].

- At least ten valid laboratories reporting data for each food type are needed.
- Six test subsamples per analyte concentration level for each food type and six negative (uninoculated) control subsamples for each type, all blind coded, are required.
- By counting positive and negative results and using a statistical test (a chi square test), it is estimated whether (a) any laboratory shows results which significantly differ from the determinations in the other laboratories, and/or (b) the test method is statistically different from the established reference method.

The common statistical indicators, sensitivity, specificity, false negative rate and false positive rate (for rates, see Table 4.3), together with the test for significance of differences, provide a basis for the assessment of test methods and laboratories.

This system of interlaboratory studies seems also to be suitable for chemical identification performed by chromatography and spectrometry.

At present, the best known proficiency tests for qualitative analysis (identification) are those organized by the Organisation for the Prohibition of Chemical Weapons (OPCW) and provided on a regular basis since 1996 [75]. Participating laboratories must detect and identify chemicals relevant to the Chemical Weapons Convention [76] present in the samples. Laboratories successfully completing the tests prove their competence in the analysis of chemicals related to the Convention. Principles of the tests are briefly as following.

- ***Samples.*** The participants receive two subsamples of test, control, and blank samples of unknown composition. Participants analyze all the samples for presence of possible chemicals as spikes, volatile or non-volatile. Various matrices such as soil, water, waste, etc., often with a high background, are tested.
- ***Methods.*** Neither methods of sample preparation nor analytical methods/techniques are prescribed to participating laboratories. Identification results must be provided by two or more different analytical techniques consistent with each other; a spectrometric technique must be used.
- ***Report.*** Results are thoroughly reported, including "unbroken chain of evidence linking each test sample to each reported chemical in the entire report". There are only 15 calendar days allowed for performing tests and sending the report.
- ***True positives.*** Among identified compounds, "only the chemicals relevant to the aim of the test should be reported."

**Table 9.2** Interlaboratory studies: qualitative determination, identification

| Year, reference | Study | Technique |
|---|---|---|
| 1984 [77] | Standard operating conditions for the acquisition of $MS^2$ spectra | $TQ–MS^2$ |
| 1987 [78] | Reproducibility of chemical shifts in NMR spectra | $^1H$ and $^{13}C$ NMR |
| 1990 [79] | Limits of detection | NCI–MS |
| 1993 [80] | Reproducibility of RI | HPLC |
| 1993- [81–83] | Determination of drug of abuse in hair | Various, including GC–MS |
| 1996- [75] | Identification of chemicals related to the chemical weapons convention | Spectrometry and others |
| 1996- [84] | Determination of pesticide residues in fruits and vegetables | GC, HPLC, GC–MS ($MS^2$), HPLC–MS ($MS^2$) |
| 1997 [85] | Determination of $^{13}C$ in sugars and fruit juice pulp | MS |
| 1997 [86] | Determination of environmental contaminants, different matrices | GC–MS libraries |
| 1998 [87] | Determination of abused drugs in urine | Immunoassay, GC–MS |
| 1999 [88] | Determination of veterinary drug residues in bovine urine | Immunoassay, GC–MS |
| 1999 [89] | Identification of organic compounds in solution | GC–MS libraries |
| 2000 – see Table 7.9 | Reproducibility of tandem mass spectra, efficiency of $MS^n$ libraries | Mainly ESI–$TQ–MS^2$ or ESI–$IT–MS^2$ |
| 2001- [90, 91] | Reproducibility of ESI mass spectra, efficiency of MS libraries | $ESI–MS^1$ |
| 2001- [92–94] | Polystyrene molecular mass distribution and other characteristics | MALDI |
| 2002 [95] | Reproducibility of migration parameters | CE |
| 2002- [96, 97] | Oil spill identification | GC, GC–MS |
| 2003 [98] | Identification of gunshot residues | Scanning electron microanalysis |
| 2004 [99] | Determination of cholesterol oxidation products (COP) in food | GC, HPLC, GC–MS, MS |
| 2005 [100] | Identification of bacterium in clinical samples | PCR |
| 2005 [101] | Identification of yeasts | PCR |
| 2005 [102] | Determination of antibiotics in milk | Immunoassays, biosensors |
| 2005 [103] | Search algorithms for matching mass spectra in protein identification | $MS^2$ |
| 2006- [104–106] | Identification of test protein mixture | $HPLC–MS^2$ |
| 2006 [107] | Detection of peanut proteins in cookies | Dipstick tests |
| 2007 [108] | Detection of hazelnut oil in olive oil | Column and gas chromatography |
| 2008 [109] | Detection of animal proteins in feed | Microscopy, PCR, immunoassay |
| 2009 [110] | Study of artworks | Various, including pyrolysis–GC–MS |
| 2009 [111] | Optimal performances of $HPLC–MS^2$ instruments in proteomics | $HPLC–MS^2$ |

Some studies are on both determinations, qualitative and quantitative

- ***False positives***. Any chemical compound which (a) is not contained in the sample, or (b) could not be formed in the corresponding matrix from analytes, or is identified on the base of "erroneous or misinterpreted analytical data," is considered as FP. Reporting any FP means failure of the test for the laboratory.
- ***Individual rates.*** Each of (a) TP, (b) the product of full degradation of TP (original spiking special chemical), and (c) group identification of special nerve agents, even without full structure elucidation, is scored as +1 point. Non-detection of one spiking compound or corresponding degradation product is FN, scored as −1 point.
- ***Laboratory ratings.*** Result scores for all the samples are combined. The maximum laboratory rating specified as A is that all chemicals were identified. If the laboratory identified (1) all analytes but one, (2) more, or (3) less than 50% of the chemicals, the rating codes are (1) B, (2) C, or (3) D respectively. A failure is coded as F.
- ***Conclusions.*** A laboratory can be designated for analysis of authentic samples if it participated in one or more proficiency tests per year and scored $A_3$ or $A_2B$, i.e., unidentified no or only one analyte, in the last three successive tests [75].

OPCW proficiency tests are good examples of how interlaboratory studies in chemical identification can be performed. These and other interlaboratory tests (mainly round robins) in qualitative analysis/identification known to the author are listed in Table 9.2.

A series of these studies have been carried out by means of routine analytical methods; corresponding areas of analysis are of great social value (drug of abuse determinations). These were parts of external QA/QC programs. In many other cases, collaborative researches have been related to development, evaluation, and expansion of new analytical methodologies such as new MS techniques and new applications of MS. Evaluation of laboratories participating in such comparisons is that of their receptivity to new techniques/methods/approaches. It should be noted that the converse type of comparison, namely comparing different analytical methods/techniques in the same laboratory, also provides improvements in methodology of chemical/biochemical analysis and its quality; see Chap. 7.

## References

1. Taylor JK (1987) Quality assurance of chemical measurements. CRC Press, Boca Raton, FL
2. Hibbert DB (2007) Quality assurance for the analytical chemistry laboratory. Oxford University Press, New York
3. Konieczka P, Namieśnik J (2009) Quality assurance and quality control in the analytical chemical laboratory. CRC Press, Boca Raton, FL
4. ISO/IEC Standard 17025 (1999) General requirements for the competence of testing and calibration laboratories
5. OECD Principles of good laboratory practice (1998) OECD, Paris. http://indiaglp.gov.in/docs/No1.pdf. Accessed 17 June 2010

6. IUPAC Compendium on analytical nomenclature (the Orange Book, 1997). http://old.iupac.org/publications/analytical_compendium. Accessed 17 June 2010
7. IUPAC Compendium of chemical terminology (the Gold Book). http://goldbook.iupac.org. Accessed 17 June 2010
8. IUPAC Compendium of chemical terminology (the Gold Book). http://goldbook.iupac.org/Q04973.html. Accessed 17 June 2010
9. Wright SE, Strehlow RA (1995) Standardizing and harmonizing terminology: theory and practice. ASTM, Philadelphia, PA
10. Zschunke A (2000) Reference materials in analytical chemistry: a guide for selection and use. Springer, Berlin
11. Ulberth F (2006) Certified reference materials for inorganic and organic contaminants in environmental matrices. Anal Bioanal Chem 386:1121–1136
12. IRMM http://irmm.jrc.ec.europa.eu/html/homepage.htm. Accessed 12 Oct 2009
13. NIST http://www.nist.gov/index.html. Accessed 12 Oct 2009
14. LGC http://www.lgc.co.uk. Accessed 12 Oct 2009
15. BAM http://www.bam.de/index_en.htm. Accessed 12 Oct 2009
16. UNIIM http://www.uniim.ru/content/view/81/167. Accessed 12 Oct 2009
17. VNIIM http://www.vniim.ru/index.en.html. Accessed 12 Oct 2009
18. NMIJ http://www.nmij.jp/english. Accessed 12 Oct 2009
19. SOFT/AAFS Forensic Laboratory Guidelines (2006) http://www.soft-tox.org/docs/Guidelines%202006%20Final.pdf. Accessed 17 May 2010
20. ChemExper http://www.chemexper.com. Accessed 15 June 2010
21. CAS Chemcats. http://www.cas.org/expertise/cascontent/chemcats.html. Accessed 15 June 2010
22. ZINC http://zinc.docking.org. Accessed 23 May 2010
23. PubChem http://pubchem.ncbi.nlm.nih.gov. Accessed 6 July 2009
24. FAO/WHO Codex Alimentarius. Guidelines on the use of mass spectrometry (MS) for identification, confirmation and quantitative determination of residues (2005) CAC/GL 56-2005. http://www.codexalimentarius.net/web/standard_list.jsp. Accessed 16 May 2010
25. Proteomics standards research group (sPRG). http://www.abrf.org/index.cfm/group.show/ProteomicsStandardsResearchGroup.47.htm. Accessed 17 June 2010
26. Valcárcel M, Cárdenas S, Barceló D, Buydens L, Heydorn K, Karlberg B, Klemm K, Lendl B, Milman B, Neidhart B, Ríos A, Stephany R, Townshend A, Zschunke A (2002) Metrology of qualitative chemical analysis. Report EUR 20605. EC, Luxembourg
27. Ríos A, Barceló D, Buydens L, Cárdenas S, Heydorn K, Karlberg B, Klemm K, Lendl B, Milman B, Neidhart B, Stephany R, Townshend A, Valcárcel M, Zschunke A (2003) Quality assurance of qualitative analysis in the framework of the European project 'MEQUALAN'. Accred Qual Assur 8:68–77
28. Allen MW (2007) Wavelength accuracy – measurement and effect on performance in UV-Visible spectrophotometry. Thermo Technical Note 51171. http://www.sorvall.com/eThermo/CMA/PDFs/Articles/articlesFile_1181.pdf. Accessed 12 Oct 2009
29. NIST SRM archive https://www-s.nist.gov/srmors/certArchive.cfm. Accessed 12 Oct 2009
30. Starna NIST Traceable UV/Vis/NIR Reference Sets. 2008 Catalog. http://starnacells.com/d_download/RefCat.pdf. Accessed 12 Oct 2009
31. Burgess C, Hammond J (2007) Wavelength standards for the near-infrared spectral region. Spectroscopy. Apr 1. http://spectroscopyonline.findanalytichem.com/spectroscopy/Near-IR + Spectroscopy/Wavelength-Standards-for-the-Near-Infrared-Spectra/ArticleStandard/Article/detail/421824. Accessed 12 Oct 2009
32. Clarke FJJ, Birch JR, Chunnilall CJ, Smart MP (2002) FTIR measurements – standards and accuracy. Vib Spectrosc 30:25–29
33. Gupta D, Wang L, Hanssen LM, Hsia JJ, Datla RU (1995) Polystyrene films for calibrating the wavelength scale of infrared spectrophotometers - SRM 1921. NIST Special Publication 260-122. http://ts.nist.gov/MeasurementServices/ReferenceMaterials/upload/SP260-122.pdf. Accessed 12 Oct 2009

34. Gordon AJ, Ford RA (1972) The chemist's companion: a handbook of practical data, techniques and references. Wiley, New York
35. Sigma-Aldrich NMR Reference Standards. http://www.sigmaaldrich.com/chemistry/stable-isotopes-isotec/stable-isotope-products.html?TablePage = 12460647. Accessed 12 Oct 2009
36. Pramanik BN, Ganguly AK, Gross ML (eds) (2002) Applied electrospray mass spectrometry. Marcel Dekker, New York
37. Bruker MALDI Kits. http://www2.bdal.de/modux3/modux3.php?pid = 007,000,000,01,02,050,004,0&rid = 000,000,000,01,02,001,001,0. Accessed 6 Nov 2009
38. Sigma-Aldrich Mass Spectrometry Standards. http://www.sigmaaldrich.com/analytical-chromatography/spectroscopy/mass-spectroscopy/ms-markers.html. Accessed 13 Oct 2009
39. Garofolo F (2004) LC–MS instrument calibration. In: Chan CC, Lee YC, Lam H (eds) Analytical method validation and instrument performance verification. Wiley, Hoboken, NJ
40. Protea Overview of Peptide Standards for Mass Spectrometry. http://dichrom.com/downloads/Protea/Phosphopeptide%20Standards_web.pdf. Accessed 6 Nov 2009
41. Barry EF, Grob RL (2007) Columns for gas chromatography: performance and selection. Wiley, Hoboken, NJ
42. Sander LC, Wise SA (2003) A new standard reference material for column evaluation in reversed-phase liquid chromatography. J Sep Sci 26:283–294
43. Smith RM, Dube S (2005) A certified reference material for HPLC. Chromatographia 61:325–332
44. Feldsine P, Abeyta C, Andrews WH (2002) AOAC International methods committee guidelines for validation of qualitative and quantitative food microbiological official methods of analysis. J AOAC Int 85:1187–1200
45. FDA Center for Veterinary Medicine Guidance for Industry (2003) Mass spectrometry for confirmation of the identity of animal drug residues. http://www.fda.gov/downloads/AnimalVeterinary/GuidanceComplianceEnforcement/GuidanceforIndustry/UCM052658.pdf. Accessed 18 May 2010
46. Soboleva E, Ambrus Á (2004) Application of a system suitability test for quality assurance and performance optimisation of a gas chromatographic system for pesticide residue analysis. J Chromatogr A 1027:55–65
47. Careri M, Mangia A (2006) Validation and qualification: the fitness for purpose of mass spectrometry-based analytical methods and analytical systems. Anal Bioanal Chem 386:38–45
48. Chan CC, Lee YC, Lam H, Zhang XM (2004) Analytical method validation and instrument performance verification. Wiley, Hoboken, NJ
49. Hardcastle WA (1998) Qualitative analysis: a guide to best practice. LGC. http://www.rsc.org/ebooks/archive/free/BK9780854044627/BK9780854044627-00001.pdf. Accessed 18 May 2010
50. Cárdenas S, Valcárcel M (2005) Analytical features in qualitative analysis. Trends Anal Chem 24:477–487
51. Ríos A, Téllez H (2005) Reliability of binary analytical responses. Trends Anal Chem 24:509–515
52. Plata MR, Pérez-Cejuela N, Rodríguez J, Ríos Á (2005) Development and validation strategies for qualitative spot tests: application to nitrite control in waters. Anal Chim Acta 537:223–230
53. Hassler S, Donze G, Esch PM, Eschbach B, Hartmann H, Hutter L, Timm U, Saxer HP (2006) Good laboratory practice (GLP) – guidelines for the acquisition and processing of electronic raw data in a GLP environment. Qual Assur J 10:3–14
54. ALACC Guide (2007) How to meet ISO 17025 requirements for method verification. http://www.aoac.org/alacc_guide_2008.pdf. Accessed 18 May 2010
55. EURACHEM Guide (1998) The fitness for purpose of analytical methods: a laboratory guide to method validation and related topics. http://www.eurachem.org/guides/pdf/valid.pdf. Accessed 18 May 2010

56. Christopher B (2000) Valid analytical methods and procedures. The Royal Society of Chemistry, Cambridge
57. Thompson M, Ellison SLR, Wood R (2002) Harmonized guidelines for single-laboratory validation of methods of analysis (IUPAC Technical Report). Pure Appl Chem 74:835–855
58. De Bièvre P, Günzler H (eds) (2005) Validation in chemical measurement. Springer, Berlin
59. González AG, Herrador MÁ (2007) A practical guide to analytical method validation, including measurement uncertainty and accuracy profiles. Trends Anal Chem 26:227–238
60. Gonzalez C, Prichard E, Spinelli S, Gille J, Touraud E (2007) Validation procedure for existing and emerging screening methods. Trends Anal Chem 26:315–322
61. Trullols Soler E (2006) Validation of qualitative analytical methods. Thesis Universitat Rovira i Virgili, Tarragona. http://www.tdr.cesca.es/TESIS_URV/AVAILABLE/TDX-0525106-095917//EstherTrullols.pdf. Accessed 18 June 2010
62. Commission Decision 2002/657/EC, August 12, 2002, implementing Council Directive 96/23/EC concerning the performance of analytical methods and interpretation of results (2002) Off J Eur Commun L 221:8–36. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2002:221:0008:0036:EN:PDF. Accessed 14 May 2010
63. Method validation and quality control procedures for pesticide residues analysis in food and feed (2009) Document No. SANCO/10684/2009. http://ec.europa.eu/food/plant/protection/resources/qualcontrol_en.pdf. Accessed 14 May 2010
64. FDA Guidance for industry (2001) Bioanalytical method validation. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm070107.pdf. Accessed 18 June 2010
65. Stüber M, Reemtsma T (2004) Evaluation of three calibration methods to compensate matrix effects in environmental analysis with LC–ESI–MS. Anal Bioanal Chem 378:910–916
66. Lehotay SJ, Gates RA (2009) Blind analysis of fortified pesticide residues in carrot extracts using GC–MS to evaluate qualitative and quantitative performance. J Sep Sci 32:3706–3719
67. Wishart DS, Tzur D, Knox C et al (2007) HMDB: the human metabolome database. Nucleic Acids Res 35:D521–D526
68. Wishart DS, Knox C, Guo AC et al (2009) HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res 37:D603–D610
69. Sumner LW, Amberg A, Barrett D et al (2007) Proposed minimum reporting standards for chemical analysis. Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics 3:211–221
70. UKAS Accreditation for Chemical Laboratories. http://www.bvsde.paho.org/bvsacd/cd52/accredi.pdf. Accessed 18 June 2010
71. Thompson M, Ellison SLR, Wood R (2006) The International Harmonised Protocol for the proficiency testing of analytical chemistry laboratories. Pure Appl Chem 78:145–196
72. Analytical Methods Committee (2007) Handling false negatives, false positives and reporting limits in analytical proficiency tests. Analyst 122:495–497
73. Ellison SLR, Fearn T (2005) Characterising the performance of qualitative analytical methods: Statistics and terminology. Trends Anal Chem 24:468–476
74. Schilling P, Powilleit M, Uhlig S (2006) Macrozoobenthos interlaboratory comparison on taxonomical identification and counting of marine invertebrates in artificial sediment samples including testing various statistical methods of data evaluation. Accred Qual Assur 11:422–429
75. Dubey V, Velikeloth S, Sliwakowski M, Mallard G (2009) Official proficiency tests of the organisation for the prohibition of chemical weapons: current status and future directions. Accred Qual Assur 14:431–437
76. OPCW Schedules of Chemicals. http://www.opcw.org/chemical-weapons-convention/annex-on-chemicals/b-schedules-of-chemicals. Accessed 19 June 2010
77. Dawson PH, Sun WF (1984) A round robin on the reproducibility of standard operating conditions for the acquisition of library MS/MS spectra using triple quadrupoles. Int J Mass Spectrom Ion Process 55:155–170

78. Chujo R, Hatada K, Kitamaru R, Kitayama T, Sato H, Tanaka Y (1987) NMR measurement of identical polymer samples by round robin method. I. Reliability of chemical shift and signal intensity measurements. Polym J 19:413–424
79. Arbogast B, Budde WL, Deinzer M, Dougherty RC, Eichelberger J, Foltz RD, Grimm CC, Hites RA, Sakashita C, Stemmler E (1990) Interlaboratory comparison of limits of detection in negative chemical ionization mass spectrometry. Org Mass Spectrom 25:191–196
80. Bogusz M, Franke JP, De Zeeuw RA, Erkens M (1993) An overview on the standardization of chromatographic methods for screening analysis in toxicology by means of retention indices and secondary standards. Fresenius J Anal Chem 347:73–81
81. Welch MJ, Sniegoski LT, Allgood CC (1993) Interlaboratory comparison studies on the analysis of hair for drugs of abuse. Forensic Sci Int 63:295–303
82. Montagna M, Polettini A, Stramesi C, Groppi A, Vignali C (2002) Hair analysis for opiates, cocaine and metabolites. Evaluation of a method by interlaboratory comparison. Forensic Sci Int 128:79–83
83. Jurado C, Sachs H (2003) Proficiency test for the analysis of hair for drugs of abuse, organized by the Society of Hair Testing. Forensic Sci Int 133:175–178
84. Medina-Pastor P, Rodríguez-Torreblanca C, Andersson A, Fernández-Alba AR (2010) European Commission proficiency tests for pesticide residues in fruits and vegetables. Trends Anal Chem 29:70–83
85. Rossmann A, Koziet J, Martin GJ, Dennis MJ (1997) Determination of the carbon-13 content of sugars and pulp from fruit juices by isotope-ratio mass spectrometry (internal reference method). A European interlaboratory comparison. Anal Chim Acta 340:21–29
86. Wong DCL, Van Compernolle R, Chai EY, Fitzpatrick RD, Bover WJ (1997) A multi-laboratory evaluation of analytical methods for estimating bioconcentratable contaminants in effluents, tissues and sediments. Environ Toxicol Chem 16:617–624
87. Badia R, De la Torre R, Corcione S, Segura J (1998) Analytical approaches of European Union laboratories to drugs of abuse analysis. Clin Chem 44:790–799
88. De Boer WJ, Van der Voet H, De Ruig WG, Van Rhijn JA, Cooper KM, Kennedy DG, Patel RKP, Porter S, Reuvers T, Marcos V, Munoz P, Bosch J, Rodriguez P, Grases JM (1999) Optimizing the balance between false positive and false negative error probabilities of confirmatory methods for the detection of veterinary drug residues. Analyst 124:109–114
89. Silva-Wilkinson RA, Burkhard LP, Sheedy BR, DeGraeve GM, Lordo RA (1999) A simple comparison of mass spectral search results and implications for environmental screening analyses. Arch Environ Contam Toxicol 36:109–114
90. Milman BL (2005) Identification of chemical compounds. Trends Anal Chem 24:493–508
91. Rosal C, Betowski D, Romano J, Neukom J, Wesolowski D, Zintek L (2009) The development and inter-laboratory verification of LC–MS libraries for organic chemicals of environmental concern. Talanta 79:810–817
92. Guttman CM, Wetzel SJ, Blair WR, Fanconi BM, Girard JE, Goldschmidt RJ, Wallace WE, Vanderhart DL (2001) NIST-sponsored interlaboratory comparison of polystyrene molecular mass distribution obtained by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry: statistical analysis. Anal Chem 73:1252–1262
93. Guttman CM, Wetzel SJ, Flynn KM, Fanconi BM, VanderHart DL, Wallace WE (2005) Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry interlaboratory comparison of mixtures of polystyrene with different end groups: statistical analysis of mass fractions and mass moments. Anal Chem 77:4539–4548
94. Nagahata R, Shimada K, Kishine K, Sato H, Matsuyama S, Togashi H, Kinugasa S (2007) Interlaboratory comparison of average molecular mass and molecular mass distribution of a polystyrene reference material determined by MALDI–TOF mass spectrometry. Int J Mass Spectrom 263:213–221
95. Boone CM, Manetto G, Tagliaro F, Waterval JCM, Underberg WJM, Franke JP, De Zeeuw RA, Ensing K (2002) Interlaboratory reproducibility of mobility parameters in capillary

electrophoresis for substance identification in systematic toxicological analysis. Electrophoresis 23:67–73
96. Faksness LG, Da1ing PS, Hansen AB (2002) Round robin study – oil spill identification. Environ Forensics 3:279–291
97. Sørheim KR, Faksness LG, Almås IK (2008) Round robin oil comparison study – 2008. SINTEF Report SINTEF A8539. http://www.sintef.no/Home/Publications/Publication?page=28511. Accessed 19 June 2010
98. Niewoehner L, Andrasko J, Biegstraaten J, Gunaratnam L, Steffen S, Uhlig S, Antoni S (2008) GSR2005 – Continuity of the ENFSI proficiency test on identification of GSR by SEM/EDX. J Forensic Sci 53:162–167
99. Appelqvist LÅ (2004) Harmonization of methods for analysis of cholesterol oxides in foods – the first portion of a long road toward standardization: interlaboratory study. J AOAC Int 87:511–519
100. Taha MK, Alonso JM, Cafferkey M et al (2005) Interlaboratory comparison of PCR-based identification and genogrouping of *Neisseria meningitidis*. J Clin Microbiol 43:144–149
101. De Baere T, Van Keerberghen A, Van Hauwe P, De Beenhouwer H, Boel A, Verschraegen G, Claeys G, Vaneechoutte M (2005) An interlaboratory comparison of ITS2–PCR for the identification of yeasts, using the ABI Prism 310 and CEQ8000 capillary electrophoresis systems. BMC Microbiol 5:14. doi:10.1186/1471-2180-5-14
102. Gaudin V, Cadieu N, Sanders P (2005) Results of a European proficiency test for the detection of streptomycin/dihydrostreptomycin, gentamicin and neomycin in milk by ELISA and biosensor methods. Anal Chim Acta 529:273–283
103. Kapp EA, Schütz F, Connolly LM, Chakel JA, Meza JE, Miller CA, Fenyo D, Eng JK, Adkins JN, Omenn GS, Simpson RJ (2005) An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. Proteomics 5:3475–3490
104. Andrews PC, Arnott DP, Gawinowicz MA, Kowalak JA, Lane WS, Lilley KS, Martin LT, Stein SE. ABRF-sPRG2006 Study: A proteomics standard. http://www.abrf.org/ResearchGroups/ProteomicsStandardsResearchGroup/EPosters/ABRFsPRGStudy2006poster.pdf. Accessed 7 June 2010
105. Andrews PC, Arnott DP, Gawinowicz MA, Kowalak JA, Lane WS, Lilley KS, Loo RRO, Martin LT, Stein SE. sPRG2007: Development and evaluation of a phosphoprotein standard. http://www.abrf.org/ResearchGroups/ProteomicsStandardsResearchGroup/EPosters/Gawinowicz_sPRG07_032707.pdf. Accessed 7 June 2010
106. Bell AW, Deutsch EW, Au CE, Kearney RE, Beavis R, Sechi S, Nilsson T, Bergeron JJ, HUPO Test Sample Working Group (2009) A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. Nat Methods 6:423–430
107. Van Hengel AJ, Capelletti C, Brohee M, Anklam E (2006) Validation of two commercial lateral flow devices for the detection of peanut proteins in cookies: interlaboratory study. J AOAC Int 89:462–468
108. García-González DL, Viera M, Tena N, Aparicio R (2007) Evaluation of the methods based on triglycerides and sterols for the detection of hazelnut oil in olive oil. Grasas y Aceites 58:344–350
109. Van Raamsdonk LWD, Hekman W, Vliege JM, Pinckaers V, Van der Voet H, Van Ruth SM (2008) The 2008 Dutch NRL / IAG proficiency test for detection of animal proteins in feed. RIKILT – Institute of Food Safety Report 2008.007. RIKILT, Wageningen. http://library.wur.nl/way/bestanden/clc/1876397.pdf. Accessed 6 Nov 2010
110. Van Keulen H (2009) Gas chromatography/mass spectrometry methods applied for the analysis of a Round Robin sample containing materials present in samples of works of art. Int J Mass Spectrom 284:162–169
111. Paulovich AG, Billheimer D, Ham AJ et al (2010) Interlaboratory study characterizing a yeast performance standard for benchmarking LC–MS platform performance. Mol Cell Proteomics 9:242–254

# Index




B.L. Milman, *Chemical Identification and its Quality Assurance*,
DOI 10.1007/978-3-642-15361-7, © Springer-Verlag Berlin Heidelberg 2011