ELSEVIER

Journal of Chromatography B, 739 (2000) 125–137

JOURNAL OF CHROMATOGRAPHY B

www.elsevier.com/locate/chromb

How to approach substance identification in qualitative bioanalysis

J. Hartstra, J.P. Franke, R.A. de Zeeuw*

Department of Toxicology and Bioanalysis, University Centre for Pharmacy, A. Deusinglaan 1, 9713 AV Groningen, The Netherlands

Abstract

The ultimate goal in qualitative analysis in the biosciences is to demonstrate with acceptable probability that for an unknown constituent in a sample only one substance comes into consideration and that all other substances can be rejected. In the biosciences, identification of relevant substances in complex matrices through database retrieval is frequently required. Yet, despite its importance, the subject has not received much attention, so that progress has been limited and relevant literature is scarce. As a result, one can conclude from many publications and reports that qualitative analysis in practice is often not being addressed properly. In this paper, some fundamental aspects of qualitative analysis will be discussed and a general approach is provided for the correct identification of organic substances in complex matrices through database retrieval. Special attention is given to the choice of proper analytical techniques and their inter-laboratory standard deviations, as well as to match factors and decision criteria based on applying multiple analytical techniques, also if the latter have different dimensions (e.g. retention data and spectral data). In addition, the requirements for suitable databases are outlined and the need for inter-laboratory cooperation is emphasized.

Keywords: Qualitative analysis; Similarity index; Match factor; Database retrieval; Identification

1. Introduction

Qualitative analysis deals with the detection and identification of one or more compounds in a sample. The compounds of interest are called analytes and the remainder of the sample is called matrix. This type of work is of vital importance in the analysis of all kinds of products (e.g. pharmaceuticals and foodstuffs, including impurity profiling) but also in areas such as intoxications, drug-abuse testing, doping, environmental pollution, occupational health, metabolic profiling, etc. Often, the concentrations of the analytes are very low (ppm–ppb range) and the matrices may be rather complex (biological fluid, soil, waste water).

De Zeeuw [1] suggested that qualitative analysis can be subdivided into *directed* searches and *undirected* searches. The former is aimed at one or more specific compounds (e.g. does the sample contain amphetamines). The undirected search comprises the screening for a large variety of substances of interest, for example in systematic toxicological analysis (STA), which is aimed at any substance that may be harmful. Furthermore, a distinction can be made between different types of identification:

- Identification of a (relatively) pure substance, usually by powerful spectrometric techniques like NMR, IR and MS, is called *structure elucidation*. This type of identification cannot be applied when

*Corresponding author. Tel.: +31-50-3633-336; fax: +31-50-3637-582.
E-mail address: r.a.de.zeeuw@farm.rug.nl (R.A. de Zeeuw)

0378-4347/00/$ – see front matter
PII: S0378-4347(99)00427-2

only small amounts of sample are available, or when the sample is a complex matrix.

- If there is a priori information about the specific identity of the substance(s) in the sample, establishing the presence of one or more substances may be done through *confirmation*, i.e. by comparing the properties of the expected substance(s) with those of an appropriate reference substance. For instance, when the police confiscates a streetdrug sample from a cocaine dealer, a directed search for cocaine can be carried out and the outcomes of the tests can be compared with those given by the reference substance cocaine[1].
- If there is no a priori information about the presence and the identity of the analyte(s) in the sample an undirected search must be carried out. Identification may then be achieved through database retrieval, i.e. the properties of the unknown(s) are being compared with those of a large number of reference substances in a database and a match is being sought. This process is called *recognition*.

It should be stressed that many cases initially requiring a directed approach often need an undirected follow up. For example, after having confirmed cocaine in the above streetdrug sample, an undirected search will indicate whether other relevant analytes may be present, e.g. XTC or caffeine.

Despite the importance of qualitative analysis in various bioscientific areas, the subject has not received much attention in recent years, so that progress has been limited and relevant literature is scarce. As a result, one must conclude from many publications and reports that qualitative analysis in practice is often being addressed inadequately. In this paper, some fundamental aspects of qualitative analysis are being discussed and an approach is provided for the correct identification of organic substances in complex matrices by means of database retrieval. For simplicity reasons, we will focus on STA, but the principles are equally applicable to other areas.

[1]Note that this process is based on finding a satisfactory match between the identification parameter of the unknown with that of the reference substance presumed to be present. Usually, confirmation procedures do not (adequately) address the question whether other substance(s) may also give satisfactory matches.

2. Identification

In database retrieval the properties of an unknown substance are compared with the properties of reference substances present in a database. If the properties of an unknown substance adequately match the properties of only one single reference substance the unknown substance has been identified. The identification process will, however, usually result in a list of substances, the properties of which are more or less matching those of the unknown. Then the identification process has to continue until only one substance remains in the list. In other words, to reject all substances in the database, except one, is the ultimate goal in the identification process. Important factors involved in this identification process are the choice of the analytical methods, the identification parameters and the size and the scope of the available database.

2.1. Detection and identification methods

Qualitative analytical methods can be subdivided into two categories: (i) classification methods; and (ii) identification methods.

Classification methods yield a selective signal for a class or group of substances, whereas identification methods yield a selective signal for a single substance. Although classification methods do not furnish the identity of a specific substance, they can be used for class detection or to narrow the number of possible candidates, and thus can play an important role in the identification process. Examples of classification methods are color reactions, such as the Marquis spot test [2], immunoassays [2] and receptor assays [3].

Identification methods yield a signal or signals for a substance which reflect a particular property of the substance. Analytical signals are for instance, retention behaviour in a chromatographic technique or a spectrum in a spectroscopic technique. There exists a large difference in identification power (IP) be-

tween various methods [4]. Obviously, the retention in TLC will give much less information about the identity of a substance than retention in GC combined with a mass spectrum. However, also within a technique, there may be considerable differences in IP: A given TLC system A can be much better suitable for identification than a TLC system B. Several mathematical methods have been described to determine the IP of a single analytical method, or the IP of combinations of analytical methods [5–9]. On the basis of IP the best methods for identification can be selected [10,11].

2.2. Identification parameters

For the identification of analytes in a complex matrix, the more suitable methods are separation techniques with a non-selective detector, e.g. GC–FID, or hyphenated techniques where a separation method is combined with a selective detector, e.g. GC–MS and LC–DAD. TLC combined with color reactions can also be considered [12].

When the retention behaviour in a chromatographic system, i.e. the mobility corresponding with the center point (apex) of a chromatographic peak, is used for identification purposes, this retention should be standardized in such a way that within the laboratory (intralaboratory) and/or from laboratory-to-laboratory (interlaboratory) parameters are obtained that are reliable and reproducible. Standardization of retention behaviour is called calibration. For instance in GC, retention times are usually converted to the Kovats retention index (RI) by calibration using *n*-alkanes in combination with relevant drugs [13]. The resulting RI values are used in the identification process and are called identification parameters.

A single identification method can yield a one-dimensional identification parameter (i.e. a number such as the RI) or a multidimensional identification parameter, i.e. an array or vector, such as a spectrum.

Multidimensional identification parameters may also be obtained from two or more identification methods, each yielding a one-dimensional identification parameter, e.g. two or more TLC systems run in parallel, each producing a standardized R_f value.

2.3. Databases

Once the most suitable identification method(s) has (have) been selected and after it has been established which way the identification parameters have to be determined and standardized, a database can be formed by collecting data of as many as possible *relevant* substances. It must be realized that a substance that is not present in the database cannot be found by database retrieval (false negative). Even worse, when the unknown substance is not present in the database a reference substance in the database may give such a match that a false identification is made (false positive).

Depending on the area of interest and the methods preferentially applied in that area, the size and scope of a suitable database may be different: In doping analysis, the relevant substances will be those that are banned in sports and the methods applied will focus particularly on substances such as anabolic steroids or stimulants. Yet, when trying to assess environmental pollution, the focus will be much more on halogenated hydrocarbons and pesticide residues.

It should be stressed, however, that any database should be as large as possible. Not only should it include the parent compounds, but also decomposition products and metabolites. Furthermore, substances that may interfere in the method should be present, such as matrix components, plasticizers, antioxidants. Finally, the selection of relevant substances should be made as broad as possible, be universal and the database should be kept up to date.

2.4. Mathematical description of the identification process

In the mathematical identification process the set of all substances in the database is R. $R = \{r_1, r_2, \ldots, r_i, \ldots, r_N\}$, where N is the total number of substances in the database. Thus, R can also be denoted as the a priori set of candidates.

Let $U = \{u_1, u_2, \ldots, u_k, \ldots, u_L\}$ be the set of L relevant analytes in the sample. Since each unknown substance must be present in the database, U is a subset of R. Identification can now be described by the process leading to the decision that the unknown

u_k is substance r_i (i.e. $u_k = r_i$). The aim of the identification process is to reduce the a priori set of candidates (set R) to a set comprising only a single substance. In practice, however, usually a set comprising a small number of candidates remains for each of the L unknowns.

In the ideal situation, an identification method yields different signals for all substances in the database. This gives a set $Y = \{y_1, y_2, \ldots, y_i, \ldots, y_N\}$, comprising N possible signals. For an imaginary database of five substances, Fig. 1a shows such an ideal situation where unambiguous identification of the substances occurs. For instance, when the signal is the molecular weight as determined by CI–MS and each substance in the database has a different molecular weight, identification will be straightforward.

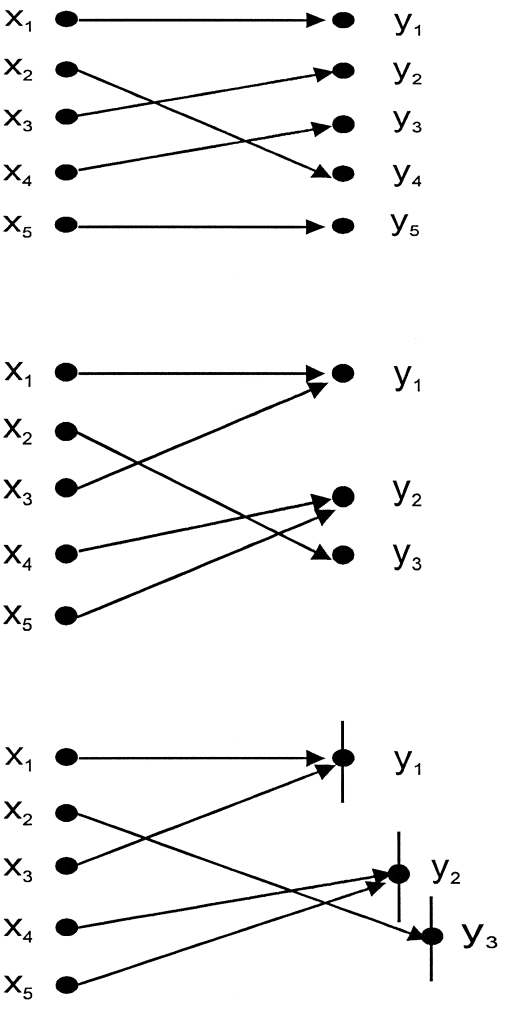

Fig. 1. Input/output graphs. (a, top) Unambiguous identification; (b, middle) substances may give the same signal; (c, bottom) each substance may give a range of signals, ranges of substances may overlap.

In practice, however, especially with larger databases, some substances may yield an identical signal, or signals which cannot be distinguished. Fig. 1b represents such a situation. In a database comprising several hundreds of compounds, a certain R_f value (e.g. 0.56) may belong to a number of substances. Unambiguous identification is no longer possible.

It must be realized that the *measurement error* of the identification parameter is an additional factor of utmost importance. If $\mu(r_i)$ represents the 'true' value (the expectation value) for the identification parameter of substance r_i, the parameter measured for the substance, $y(r_i)$, can be represented by:

$$y(r_i) = \mu(r_i) \pm \epsilon_i, \quad (1)$$

where ϵ_i is the measurement error of the signal of substance r_i. Fig. 1c represents this situation: Each substance can give a range of signals. The size of the range is determined by the measurement error (i.e. the reproducibility of the method).

Usually, it is assumed that the measurement of the signal (identification parameter) of a substance r_i is Gaussian distributed with $\mu(r_i)$ as the center and σ_i as the standard deviation (the size of the error).

2.5. Similarity and dissimilarity

The similarity between the identification parameters measured for the unknown substance and the candidate in the database is usually in the range of 0–1 [9]. When comparing spectra, the correlation coefficient is a useful measure for similarity [13].

To determine the degree of dissimilarity, which is the opposite of similarity [14,15], a so-called distance function is employed. For scalar, zero-order, identification parameters, such as the chromatographic retention, the absolute difference is a suitable parameter:

$$d_i = |y(u_k) - \mu(r_i)| \quad (2)$$

Another useful distance function is the absolute

eccentricity, z, since it takes into account the reproducibility of the measurement of the identification parameter involved:

$$z_i = \left| \frac{y(u_k) - \mu(r_i)}{\sigma_i} \right| \quad (3)$$

Similarities and dissimilarities can be used to develop meaningful decision criteria, as shown in Sections 3 and 4.

3. Database retrieval using a univariate identification parameter

3.1. Window retrieval

Up till now, many database retrieval procedures in use are based on so-called window retrieval. In window retrieval the decision whether the reference substance r_i is a candidate for identification is determined by the following decision function:

$$\text{retain } r_i \text{ if } d_i \leq C \text{ discard } r_i \text{ if } d_i > C \quad (4)$$

When $C = 2\sigma_i$, assuming a normal distribution, and if $u_k = r_i$, there is a 95% probability of measuring an identification parameter for u_k that falls within the window. In other words, the true candidate will be discarded in 1 out of 20 cases (false negative). When the window is enlarged to $3\sigma_i$, the probability increases to 99.6% and hence the true candidate will only be discarded in about 1 out of the 400 cases. However, enlarging the window will result in an increase of the number of possible candidates, thus enhancing the chances for a false positive identification!

Another (major) disadvantage of the window retrieval approach is that no differentiation is made between the candidates. It is obvious that a candidate whose identification parameter equals the signal obtained from the unknown is a more likely candidate than one whose identification parameter is found at the border of the window. Furthermore, a substance with, in one method an outcome just outside the window, and in other methods with outcomes right on the dot, will be lost in a window retrieval approach. In order to overcome these disadvantages, a probabilistic approach is recommended.

3.2. A probabilistic approach towards identification

The identification problem can also be described by the null hypothesis (H_0), or by the alternative hypothesis (H_A):

$$H_0 : \mu_k = r_i \quad (5A)$$

$$H_A : \mu_k \neq r_i \quad (5B)$$

Based on the evidence provided by the analyses we decide on either accepting or rejecting H_0. This may yield four possible results (e.g. [16,17]) as depicted in Table 1.

In statistical hypothesis testing, the tests are designed so that the probability of rejecting H_0, when in fact it is true, is equal to the so-called significance level (α) of the test. The probability of accepting the null hypothesis when in fact it is false is called the power, β, of the test.

There are basically two approaches towards hypothesis testing:

Table 1
Possible outcomes of a hypothesis test

		Unknown truth	
Decision	$H_0 : \mu_k = r_i$	$H_0 : \mu_k = r_i$ Correct decision True positive $p = 1 - \alpha$	$H_0 : \mu_k \neq r_i$ Erroneous decision False positive $p = \beta$
	$H_0 : \mu_k \neq r_i$	Erroneous decision False negative $p = \alpha$	Correct decision True negative $p = 1 - \beta$

1. By defining acceptance and rejection regions under the assumption of H_0 being true and by setting a certain value of α.
2. By calculating the credibility that H_0 is true (p-value), and by subsequently rejecting H_0 if the p-value is smaller than a predetermined value α.

In database retrieval, each of the L unknowns has to be compared with all N substances in the database. Thus, the identification process involves a total of $L \times N$ hypothesis tests. Actually, the tests are based on the question whether the parameter measured for the unknowns can be related to the given candidate, i.e. how similar is the parameter of the unknown compared with the true value of the parameter of the candidate. Thus, the hypotheses from Eq. (5) can be restated as:

$$H_0 : \mu(u_k) - \mu(r_i) = 0 \quad (6A)$$

$$H_A : \mu(u_k) - \mu(r_i) \neq 0 \quad (6B)$$

This indicates the need for a measure of the dissimilarity between the identification parameters measured for the unknown, $y(u_k)$, and the true value of the candidate, $\mu(r_i)$. Furthermore, there is a need for a limit: If the parameters are sufficiently similar, H_0 cannot be rejected and substance r_i remains a candidate for unknown substance u_k.

Under the assumption that the identification parameter of a substance r_i, due to errors in the measurement, is normally distributed with mean $\mu(r_i)$ and standard deviation σ_i, the probability can be calculated that H_0, according to Eq. (6A) is true. This probability is the shaded area under the curve in Fig. 2 and can be expressed as [17]:

$$p(y(u_k)|H_0) = 2 \int_{z_i}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-1/2(z_i)^2} dz \quad (7)$$

where z_i can be calculated according to Eq. (3). This probability can also be obtained from statistical tables of the standard normal distribution. For instance, in the example of Fig. 2, the value found for the identification parameter of the unknown, $y(u_k)$, is 950. The value of the identification parameter of the reference substance in the database, $\mu(r_i)$, is 1000 and the standard deviation, representing the measure-

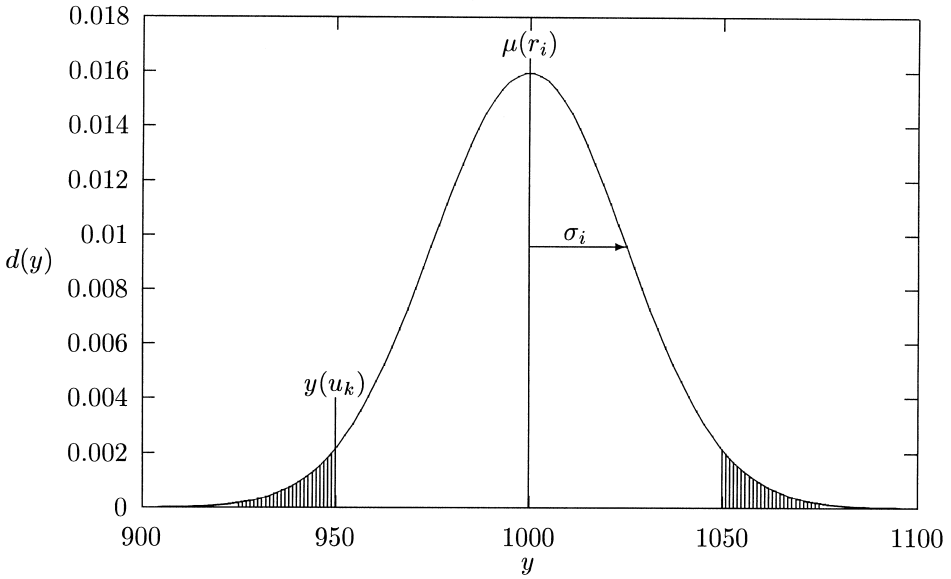

Fig. 2. Probability density function, $d(y)$, of the signal of reference substance r_i with mean, $\mu(r_i)$, 1000 and standard deviation (σ_i), 25. The shaded area represents the credibility (p-value) that the null hypothesis (the mean signal of the unknown substance, u_k equals the mean signal of the reference substance) is true when 950 is found for the unknown.

ment error of the identification parameter is assumed to be 25. According to Eq. (3) $z_i = |950-1000|/25 = 2.00$ and from statistical tables it can be found that the probability under the null hypothesis is $2\times0.023=0.046$. When α has been set at 0.050, the reference substance with value 1000 will *not* be a candidate. For the unknown substance with a value of 950 the probability found is smaller than the critical value of 0.050. The measurement error of the identification parameter as mentioned above relates to the deviations that may occur between the measured data for the unknown and those stored in the database. In essence, it is the interlaboratory standard deviation for a particular analytical technique after assessing the behavior of a suitable number of analytes (mean and SD) in the technique by a selection of laboratories all over the world under their instrumentation and circumstances. It should be noted that an error will also occur when the unknown sample is being measured but the latter is negligible in comparison to the above interlaboratory error.

The probability according to Eq. (7), is also called the similarity index (SI). This SI represents a match, a similarity, between the value found for the unknown substance compared to the mean value of a reference substance in the database [9,18,19]. It can easily be seen from Fig. 2 that the SI (the shaded area) approaches 1 when $y(u_k)$ approaches $\mu(r_i)$. When $y(u_k)$ deviates increasingly from $\mu(r_i)$, the shaded area becomes smaller and SI approaches zero.

4. Database retrieval using UV spectra

For more complex signals, such as spectra, the absolute difference d_i as given in Eq. (2) is unsuitable. For these first-order signals alternative distance functions are available, such as the Euclidean distance:

$$d_i = \sqrt{\sum_{v=1}^{w}(y_v(u_k) - \mu_v(r_i))^2} \quad (8)$$

where the signals are vectors consisting of w discrete measurements [14].

As a similarity function for UV spectra, correlation coefficients (R), such as Pearson's product moment correlation coefficient, are frequently used [13]. Other useful distance functions are the Minkowski metric, the Canberra metric, and the Czekanowski coefficient [15].

The correlation coefficient R (or other suitable function), obtained from comparing a spectrum of substance u_k and a spectrum of reference substance r_i, is of another magnitude than the SI defined in Section 3.2. For instance, a value of 0.80 for a correlation coefficient is a bad match, whereas a value of 0.80 for the SI represents a good match.

In order to decide whether $u_k = r_i$ on the basis of spectra matching, the following hypotheses can be considered:

$$H_0 : \rho(r_i) = 1 \quad (9A)$$

$$H_A : \rho(r_i) \neq 1 \quad (9B)$$

The probability distribution of R, with mean $\rho(r_i)$, is not known. However, R can be transformed (Fischer Z-transformation) and the resulting z is standard normal distributed [20]:

$$T = 2\ln\{(1+R)/(1-R)\};$$

$$s_T = 1/\sqrt{(w-3)}; \quad z = T/s_T \quad (10)$$

where w is the number of data pairs with which R is calculated.

The Fischer Z-transformation assumes that $\rho=0$ and this is not in agreement with the hypotheses given in Eq. (9). Moreover, in the vicinity of $R=1$, z approaches infinity and is therefore not useful for our purpose. For hypothesis testing, a dissimilarity such as z_i from Eq. (3) is needed and therefore the hypotheses, compare Eq. (6), can be transformed into:

$$H_0 : 1 - \rho(r_i) = 0 \quad (11A)$$

$$H_A : 1 - \rho(r_i) \neq 0 \quad (11B)$$

Under the assumption that $d = 1 - R$ and $0<R<1$, the same Fischer Z-transformation can be performed and when d is not too far from 0, the following approximation can be made:

$$s_d = 1/\sqrt{(w-3)}; z_i = d/s_d \quad (12)$$

Using Eq. (12) decisions can now be made on the hypotheses in Eq. (11). As an example, UV spectra

from a diode array detector with 100 diodes ($w=100$, $s_d=0.1$) from an unknown and a reference substance were compared resulting in a correlation coefficient of 0.95 ($d=0.05$). According to Eq. (12), for z_i a value of 0.50 is calculated. The tables of the standard normal distribution give for $z_i=0.50$ a one-sided probability of 0.309 and therefore, a SI of 0.618 is obtained (see also Section 3.2). This SI value represents the credibility (p-value) of the null hypothesis stated in Eq. (11) and, since this value is larger than the critical value (α) of 0.050, the reference substance is a possible candidate for identification. In other words, the hypothesis $u_k = r_i$ cannot be rejected.

5. Multivariate database retrieval

In the previous paragraphs the identification parameter was from a single identification method yielding a scalar, unidimensional, identification parameter or yielding an array of information (spectrum). In practice, however, unequivocal identification, resulting in one single candidate, almost always requires the use of more than one (or a multitude) of identification methods, each yielding a scalar identification parameter or an array of information.

The identification parameters obtained for a single substance from K univariate methods can be regarded as a K-dimensional random variable. This means that for reference substance r_i, for each of the K methods, identification parameters $\mu_j(r_i)$ ($i=1,\ldots,N$, $j=1,\ldots,K$) are present in the database. Subsequently, for the unknown substance in each of the K methods, data are collected: $y_j(u_k)$.

Multivariate mathematical techniques like Principal Components Analysis (PCA) and pattern recognition have been used for such identification purposes. Meglen discussed the use of PCA for the examination of large databases [21]. Musumarra et al. used PCA for identification using TLC data [22,23]. Pattern recognition is actually a collection of multivariate techniques with two main types: Unsupervised pattern recognition (unsupervised learning) and supervised pattern recognition [24–26]. In unsupervised pattern recognition clustering is the major technique. Clustering or numerical taxonomy was used by Massart and De Clerq for the selection of identification methods [27].

For the supervised pattern recognition, the class membership for a set of objects is known. This is the so-called test-set or learning-set. Based on properties measured for these objects, a membership function is deduced. This membership function can be used to assign unknown objects to a specific class. Methods frequently used are the K-nearest neighbor method (KNN), the linear learning machine (LMM), statistical linear discriminant analysis (SLDA), ALLOC, SIMCA, etc. [26]. In both forms of pattern recognition classification plays an important role. Identification by database retrieval may be seen as a complex form of supervised pattern recognition, where each reference substance represents a particular class of its own. The unknown substance should be classified using the classes established by the reference substances.

A major disadvantage of the above multivariate methods is that usually all identification parameters for the unknowns have to be available: For the unknown all methods in the database have to be carried out. Thus, a sequential identification process, where the results of the first identification method determines the choice of the second identification method, etc. becomes impossible. Yet, such a sequential process is the approach of choice in STA and many related areas in the biosciences. In other words, the above multivariate methods are less suitable when only a selection of the available identification methods is employed. Alternative approaches are being discussed in the following sections.

5.1. Discrepancy index

If the analytical methods are totally independent of each other, the sum of the squares of the eccentricities (z_j) is χ^2 distributed with K degrees of freedom [16].

The eccentricities are calculated according to Eq. (8), one eccentricity for each of the K methods. The test statistic G^2 can then be calculated:

$$G^2 = z_1^2 + z_2^2 + \cdots + z_j^2 + \cdots + z_K^2 \qquad (13)$$

Since G^2 is χ^2 distributed, the credibility (p-

value) of the joint null hypothesis that, in the K methods the values found for the identification parameters of the unknown are equal to those of the reference substance, can be calculated.

The parameter G^2 is also called the discrepancy index (DI) [28]. Using the critical value obtained from statistical χ^2 tables and on the basis of the DI, it can be decided whether a substance is a candidate for identification ($DI \leq \chi^2_{K;\alpha}$) or not ($DI > \chi^2_{K;\alpha}$). The DI can be a useful parameter. However, the critical value on which decisions have to be made is dependent on the degrees of freedom, i.e. on the number of methods used. Moreover, handling of missing values can only be performed manually.

5.2. Multivariate similarity index

Under the assumption that the K methods are independent of each other, a joint probability can be calculated by multiplication of the probabilities (SI_{ij}; $i = 1, \ldots, N$; $j = 1, \ldots, K$) found in each of the K methods. This is comparable with throwing a dice: First throw a five (probability = 1/6), second throw a two (probability = 1/6); The joint probability of throwing first a five and then a two is equal to $1/6 \times 1/6 = 1/36$.

Each of the null hypotheses

$$\mu_j(u_k) - \mu_j(r_i) = 0 \tag{14}$$

can be tested separately, resulting in K p-values for each of the reference substances $r_{i.}$. The joint probability for the K null hypotheses is then the product of the K p-values.

However, the critical value of accepting or rejecting the joint null hypothesis that in the K methods the values found for the identification parameters of the unknown are equal to those of the reference substance, is dependent on the number of methods used. Moreover, the more methods used, the smaller the probabilities become. To overcome this problem two approaches were developed:

- Summing the joint probabilities over all substances in the database and dividing the individual probabilities by this sum. In this way, relative probabilities (F_i) are obtained [29].
- Instead of multiplication, the geometric means of the probabilities is determined [19]:

$$SI_i = \sqrt[K]{\prod_{j=1}^{K} SI_{ij}} \tag{15}$$

Relative probabilities have the disadvantage that if only one or two substances are found with low similarity, these substances will have a high relative probability. It gives a false feeling that a good match is obtained. The opposite is also true: A list of ten substances with very high p-values results in very low (<0.1) relative probabilities for each of them. For these reasons, we recommend the use of the joint similarity index based on the geometric means of the SI's of the individual methods used.

On the other hand, it should be noted that if for a substance the joint multivariate SI_i is approaching 1, this does not mean that the latter is the only candidate for identification. There may be more substances with similar SI_i. In the latter case in database retrieval, a list of substances will be obtained and additional methods have to be utilized to obtain a list with a single candidate.

According to Eq. (14), each of the methods used is treated likewise. In practice, however, the identification power of, for instance, a retention index (RI) in GC is not comparable with a UV spectrum. Obviously, a RI, as a unidimensional parameter, has not the same weight as a multidimensional parameter such as a spectrum. Therefore, it is reasonable to give a spectrum more weight in the joint SI than a single retention parameter. This can be achieved by raising each of the single SI's to the power w_j, where for the retention parameter $w_j = 1$ and for a spectrum, a w_j of 2–3 can be applied. In this way, a spectrum weighs 2–3 times more than a single retention parameter. The joint similarity index will then become:

$$SI_i = \left(\prod_{j=1}^{K} SI_{ij}^{w_j} \right)^{1/\sum_{j=1}^{K} w_j} \tag{16}$$

Weight factors for the individual analytical techniques can be determined on the basis of their identification power (IP), as assessed by the mean list length (MLL).

The joint similarity index has the advantage of being easily interpretable as a match factor. How-

ever, the critical value, on the basis of which a substance has to be accepted or rejected as a candidate for identification, has to be developed in practice. At this point, after various simulations, using the geometric mean, a value of 0.05 seems appropriate. The latter is illustrated in the example given in Section 6. It should be noted that the analyst may choose his own critical value as well as the weight factors he wants to use, based on his own expertise or judgement. In doing so, he can either narrow or broaden his search.

6. Substance identification in practice

6.1. A simplified example of substance identification in STA

A plasma sample of an intoxicated patient is extracted and the extract, after evaporation and redissolution in a suitable solvent, is injected in a GC–FID and in a HPLC–DAD under standardized conditions [10,30]. For the GC analysis one peak is obtained with a RI of 2075 after calibration with *n*-alkanes, and with the HPLC system one peak is obtained with a RI of 510 after calibration with 1-nitroalkanes. Moreover, a diode array spectrum is generated from 200–360 nm, resulting in 160 data pairs. With these results, database retrieval is performed with the algorithms developed by Hartstra and utilizing a database of more than 1300 toxicologically relevant substances [19].

On the basis of the RIs present in the GC-database, the computer selects ten substances as possible candidates. These are presented in the first column of Table 2. Note that the 'window' used for retrieval is larger than 3σ, which prevents outliers to be rejected. For each of the ten candidates, the corresponding RIs in HPLC are retrieved and the UV spectra are compared with that of the unknown, resulting in correlation coefficients *R*. The latter two parameters are given in the second and third column of Table 2.

Table 3 shows the results of the calculations according to the procedures given in Sections 3.2, 4 and 5.2. By comparing the values found with those listed, the eccentricities, z_i, and the probabilities, *p*, are then obtained for the individual methods. Finally, the joined similarity indices (SI_i) are generated for the combined methods. In the latter, the spectra have been given double weight as compared to that of the RIs:

$$SI_i = \{p_{i1} * p_{i2} * (p_{i3})^2\}^{1/4} \quad (17)$$

It can be seen from Table 3 that after performing GC and HPLC three substances are still candidates for identification ($SI_i > 0.05$). After introducing the spectra, only one candidate remains, i.e. substance 5. Substance 7 falls below the limit of 0.050.

Table 2
Substances selected from the database as candidates for identification, based on the RIs in GC, the corresponding RIs in HPLC and the correlation coefficients of their UV spectra with that of the unknown (see Section 6.1 for details)

Substance number	RI–GC $\sigma_i = 20$	RI–HPLC $\sigma_i = 10$	*R* $w = 160$
1	1980	800	0.312
2	2005	720	−0.005
3	2010	600	0.560
4	2040	520	0.650
5	2065	505	0.996
6	2070	605	0.654
7	2080	490	0.805
8	2100	670	0.734
9	2150	400	0.245
10	2170	425	0.368

6.2. Multiple unknowns

In the daily practice of substance identification, database retrieval is more difficult than in the above example because the samples are usually more complex: When, for instance, GC and HPLC are carried out on extracts of the same sample, a number of peaks may be seen in each chromatogram. Yet, the number of peaks may not be the same, nor will this be the case for the elution order. Hence, it is hard to establish which GC peak will correspond to which HPLC peak. Therefore, all combinations of GC peaks with HPLC peaks have to be tested separately in the database retrieval process. Other complications may be that peaks may coincide in one run but not in the other, that a substance will not show up with one technique because it is below the detection limit, that matrix components will give peaks or interfere with the spectral data, etc. Furthermore, when a third

Table 3
Results of the calculations using the data in Table 2 in the process of identifying the unknown

Subst. r_i	GC		HPLC		SI_i GC+HPLC	Spectrum		SI_i GC+HPLC+UV
	z_{i1}	p_{i1}	z_{i2}	p_{i2}		z_{i3}	p_{i3}	
1	4.75	0.000	29.0	0.000	0.000	10.6	0.000	0.000
2	3.50	0.000	21.0	0.000	0.000	37.5	0.000	0.000
3	3.25	0.002	9.00	0.000	0.000	5.92	0.000	0.000
4	1.75	0.080	1.00	0.318	0.159	4.72	0.000	0.000
5	0.50	0.618	0.50	0.618	0.618	0.05	0.960	0.770
6	0.25	0.802	9.50	0.000	0.000	4.52	0.000	0.000
7	0.25	0.802	2.00	0.046	0.192	2.61	0.010	0.044
8	1.25	0.212	16.0	0.000	0.000	3.42	0.000	0.000
9	3.75	0.000	11.0	0.000	0.000	12.34	0.000	0.000
10	4.75	0.000	8.50	0.000	0.000	9.33	0.000	0.000

chromatographic technique is being used (e.g. a TLC system, or a second HPLC system) the situation becomes even more complex. Fortunately, information from detectors, such as UV and MS data, are easier to combine with chromatographic data, since it is known to which peak this spectral information belongs.

For these complex situations suitable computer programming must be available. For toxicological analysis we have developed programs that are capable of dealing with the above issues and that can handle TLC data combined with color reactions on the plate, GC data combined with molecular weights from CI–MS and HPLC data combined with diode array spectra [19]. The software is also commercially available [32]. The program has been set up in such a way that it can also accommodate the results of immunoassays and receptor assays, as well as more comprehensive mass spectra.

Some practical examples in STA have been given by Hartstra et al. [31].

7. Discussion and conclusions

The ultimate goal in the qualitative analysis in complex matrices is to demonstrate with acceptable probability that for an unknown constituent of that matrix only one substance comes into consideration and that all other substances can be rejected.

In the above paragraphs we have outlined how substance identification in complex matrices in the biosciences can best be approached. Database retrieval using a probabilistic approach is to be preferred. For all identification parameters generated, dissimilarities between the unknown(s) and the reference substances in the database need to be established. These dissimilarities can then be used in hypothesis testing. The probability (reliability, credibility) that the null hypothesis is true, the so-called *p*-value or similarity index (SI), can be used as the decision criterion whether a reference substance is a candidate for identification or not.

When more than one analytical method or detection system is used, multivariate database retrieval is required. This can be done by using the discrepancy index (DI) or the joint similarity index (Si_j). The latter is the method of choice because it is more straightforward in deducing whether a substance is a candidate for identification or not.

It will be obvious that for meaningful substance identification in the biosciences one must follow a systematic, concise and well planned approach in all three steps of the qualitative analysis, i.e. in:

- Sample work up and concentration (usually by techniques such as liquid–liquid extraction or solid-phase extraction).
- Differentiation and detection (usually by competitive binding assays and by chromatographic techniques and their related detection systems).
- Identification by multivariate database retrieval [33].

When applying these steps, the following prerequisites are of vital importance:

1. Retain all relevant substances, yet remove as much of the non-relevant substances and interferences (matrix) in the sample work up.
2. Obtain maximum differentiation in a minimum amount of time and effort. Detect with optimum universality and sensitivity, yet also try to differentiate in the detection phase.
3. Maintain comprehensive databases for all relevant analytical techniques and all relevant substances. The latter should also include metabolites, matrix interferences, omnipresent contaminants, etc.

With regard to the differentiation and detection step, the following criteria are crucial in the selection of analytical methods most suitable for substance identification [9]: (i). The substances measured in the analytical system should cover the total range and be evenly distributed over that range (e.g. for TLC the R_f-range 0–100). (ii). The measurement errors (standard deviations) should be as small as possible and the identification parameter must be calibrated in such a way that good reproducibility is obtained on an interlaboratory level. (iii). When more than one method is used the correlation between these methods should be low.

Once the best methods have been established for a given application area in the biosciences, the latter should be designated as recommended methods. Not until then will it be realistic to start building up comprehensive databases for these methods. Data should be entered in a standardized way and the interlaboratory measurement error must be included for each method.

When considering the above prerequisites, one must conclude that qualitative analysis in the biosciences is a very complex issue that has been receiving far too little attention. Consequently, the tools available for meaningful and reliable substance identification in biological matrices leave much to be desired. Advances can only be made by extensive, concerted interlaboratory efforts under international guidance and consensus.

References

[1] R.A. de Zeeuw, Toxicol. Lett. 102 (1998) 103–108.
[2] A.C. Moffat, Clark's Isolation and Identification of Drugs, Pharmaceutical Press, London, 1986.
[3] K. Ensing, I.J. Bosman, A.C.G. Egberts, J.P. Franke, R.A. de Zeeuw, J. Pharm. Biomed. Anal. 12 (1994) 53–58.
[4] J.P. Franke, R.A. de Zeeuw, in: T.A. Gough (Ed.), The Analysis of Drugs of Abuse, J. Wiley and Sons, Chichester, 1991, pp. 93–120.
[5] D.L. Massart, J. Chromatogr. 46 (1970) 274–279.
[6] K.W. Smaldon, A.C. Moffat, J. Forens, Sci. Soc. 13 (1973) 291–295.
[7] F. Dupuis, A. Dilkstra, Anal. Chem. 47 (1975) 379–383.
[8] R.K. Müller, W. Möckel, H. Wallenborn, A. Weihermüller, C. Weihermüller, I. Lauerman, Beitr. Gerichtl. Med. 34 (1976) 265–269.
[9] P.G.A.M. Schepers, J.P. Franke, R.A. de Zeeuw, J. Anal. Toxicol. 7 (1983) 272–278.
[10] R.A. de Zeeuw, J.P. Franke, H.H. Maurer, K. Pfleger, Gas Chromatographic Retention Indices of Toxicologically Relevant Substances on Packed or Capillary Columns with Dimethylsilicone Stationary Phases, 2nd ed., VCH Verlag, Weinheim, 1992.
[11] R.A. de Zeeuw, J.P. Franke, F. Degel, G. Machbert, J.H. Schütz, Thin-layer Chromatographic R_f Values of Toxicologically Relevant Substances on Standardized Systems, 2nd ed., VCH Verlag, Weinheim, 1992.
[12] J.P. Franke, M. Bogusz, R.A. de Zeeuw, Fresenius J. Anal. Chem. 347 (1993) 67–72.
[13] L. Huber, Application of Diode Array Detectors in High-Performance Liquid Chromatography, Hewlett–Packard, Waldbronn, Germany, 1989.
[14] M. Zuercher, J.T. Clerc, M. Farkas, E. Pretsch, Anal. Chim. Acta 206 (1988) 161–172.
[15] W.J. Krzanowski, Principles of Multivariate Analysis, Oxford Statistical Science Series, Vol. 3, Clarendon Press, Oxford, 1988.
[16] E. Kreyszig, Introductory Mathematical Statistics, J. Wiley and Sons, New York, 1970.
[17] T. Wonnacott, R. Wonnacott, Introductory Statistics, 5th ed., J Wiley and Sons, New York, 1990.
[18] J. Parker, J. For. Sci. Soc. 6 (1966) 33–39.
[19] J. Hartstra, Computer-aided identification of toxicologically relevant substances by means of multiple analytical methods, Ph.D. Thesis, Groningen, 1997.
[20] O.L. Davies, P.L. Goldsmith, Statistical Methods in Research and Production, 4th ed., Longman Group, London, 1977.
[21] R.R. Meglen, J. Chemom. 5 (1991) 163–179.
[22] G. Musumarra, G. Scarlata, G. Romano, S. Clementi, J. Anal. Toxicol. 7 (1983) 286–292.
[23] G. Musumarra, G. Scarlata, G. Romano, S. Clementi, S. Wold, J. Chromatogr. Sci. 22 (1984) 538–547.
[24] T. Blaffert, Anal. Chim. Acta 161 (1984) 135–148.
[25] M.A. Sharaf, D.L. Illman, B.R. Kowalski, Chemometrics, J Wiley and Sons, New York, 1986.

[26] D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte, L. Kaufman, Chemometrics, A Textbook, Elsevier, Amsterdam, 1988.
[27] D.L. Massart, H. de Clerq, Anal. Chem. 46 (1974) 1988–1992.
[28] R. Gill, B. Law, C. Brown, A.C. Moffat, Analyst 110 (1985) 1059–1065.
[29] J. Akkerboom, P. Schepers, J. van der Werf, Stat. Neerlandica 34 (1980) 173–187.
[30] R.A. de Zeeuw, J. Hartstra, J.P. Franke, J. Chromatogr. A 674 (1994) 3–13.
[31] J. Hartstra, J.P. Franke, R.A. de Zeeuw, GIT Labor-Med. 18 (1995) 272–279.
[32] Merck Tox Screening System (MTSS), version 3.10. E. Merck, Darmstadt, 1995.
[33] R.A. de Zeeuw, J. Chromatogr. B 689 (1997) 71–79.