Vol. 112 No. 2 June 15, 1979 ISSN 0378-4304

Computer Techniques and Optimization, Vol. 3 No. 2)

ANALYTICA CHIMICA ACTA

International journal devoted to all branches of analytical chemistry

COMPUTER TECHNIQUES AND OPTIMIZATION

ELSEVIER SCIENTIFIC PUBLISHING COMPANY

ANALYTICA CHIMICA ACTA

International journal devoted to all branches of analytical chemistry
Revue internationale consacrée à tous les domaines de la chimie analytique
Internationale Zeitschrift für alle Gebiete der analytischen Chemie

PUBLICATION SCHEDULE FOR 1979 (incorporating the section on Computer Techniques and Optimization).

	J	F	M	A	M	J	J	A	S	O	N	D
Analytica Chimica Acta	104/1	104/2	105	106/1	106/2	107	108	109/1	109/2	110/1	110/2	111
Section on Computer Techniques and Optimization			112/1			112/2			112/3			112/4

Scope. *Analytica Chimica Acta* publishes original papers, short communications, and reviews dealing with every aspect of modern chemical analysis, both fundamental and applied. The section on *Computer Techniques and Optimization* is devoted to new developments in chemical analysis by the application of computer techniques and by interdisciplinary approaches, including statistics, systems theory and operation research. The section deals with the following topics: Computerized acquisition, processing and evaluation of data. Computerized methods for the interpretation of analytical data including chemometrics, cluster analysis, and pattern recognition. Storage and retrieval systems. Optimization procedures and their application. Automated analysis for industrial processes and quality control. Organizational problems.

Submission of Papers. Manuscripts (three copies) should be submitted to:
for *Analytica Chimica Acta:* Dr. A. M. G. Macdonald, Department of Chemistry, The University, P.O. Box 363, Birmingham B15 2TT, England;
for the section on *Computer Techniques and Optimization:* Dr. J. T. Clerc, Universität Bern, Pharmazeutisches Institut, Sahlistrasse 10, CH-3012 Bern, Switzerland.

Information for Authors. Papers in English, French and German are published. There are no page charges. Manuscripts should conform in layout and style to the papers published in this Volume. Authors should consult Vol. 102, p. 253 for detailed information. Reprints of this information are available from the Editors or from: Elsevier Editorial Services Ltd., Mayfield House, 256 Banbury Road, Oxford OX2 7DE (Great Britain).

Reprints. Fifty reprints will be supplied free of charge. Additional reprints (minimum 100) can be ordered. An order form containing price quotations will be sent to the authors together with the proofs of their article.

Advertisements. Advertisement rates are available from the publisher.

Subscriptions. Subscriptions should be sent to: Elsevier Scientific Publishing Company, P.O. Box 211, 1000 AE Amsterdam, The Netherlands. The section on *Computer Techniques and Optimization* can be subscribed to separately.

Publication. *Analytica Chimica Acta* (including the section on *Computer Techniques and Optimization*) appears in 9 volumes in 1979. The subscription for 1979 (Vols. 104–112) is Dfl. 1179.00 plus Dfl. 135.00 (postage) (Total approx. U.S. $641.00). The subscription for the *Computer Techniques and Optimization* section only (Vol. 112) is Dfl. 131.00 plus Dfl. 15.00 (postage) (Total approx. U.S. $71.20). Journals are sent automatically by air mail to the U.S.A. and Canada at no extra cost and to Japan, Australia and New Zealand for a small additional postal charge. All earlier volumes (Vols. 1–95) except Vols. 23 and 28 are available at Dfl. 144.00 (U.S. $70.20), plus Dfl. 10.00 (U.S. $4.90) postage and handling, per volume.

Claims for issues not received should be made within three months of publication of the issue, otherwise they cannot be honoured free of charge.

Customers in the U.S.A. and Canada who wish to obtain additional bibliographic information on this and other Elsevier journals should contact Elsevier/North Holland Inc., Journal Information Center, 52, Vanderbilt Avenue, New York, NY 10017. Tel: (212) 867-9040.

Analytica Chimica Acta, 112 (1979) 97—122
Computer Techniques and Optimization

REVIEW

OPTIMIZATION BY STATISTICAL LINEAR DISCRIMINANT ANALYSIS IN ANALYTICAL CHEMISTRY

D. COOMANS and D. L. MASSART*
Farmaceutisch Instituut, Vrije Universiteit Brussel, Bosstraat, B-1090 Jette (Belgium)
L. KAUFMAN
Centrum Statistiek en Operationeel Onderzoek, Vrije Universiteit Brussel (Belgium)
(Received 20th October 1978)

SUMMARY

The application of statistical linear discriminant analysis in analytical chemistry is discussed. In addition to a general discussion of the theory of the method, which is illustrated by some examples, its suitability for problem solving in analytical chemistry is demonstrated by a review of published applications. A more mathematical point of view is added as an appendix.

In recent years optimization and classification techniques have been applied to a growing number of problems in analytical chemistry. An important cause of the attention given to these techniques may be found in the increasing difficulties encountered in interpreting the results of analytical techniques which produce simultaneous information concerning a large number of parameters. Such multicomponent methods include, for instance, gas chromatography, i.r. spectrometry, activation analysis, etc. Automatic apparatus in clinical analysis often allow the determination of 12 parameters at a time. The ultimate object of such a method is very often a classification, for example, classification of patient samples into categories of illness, of food-stuffs according to origin or quality, of oil spills according to origin, etc.

Multicomponent methods generate large quantities of data. Two major problems must be considered.

(a) The classification problem. In many cases, the potential of the method is not fully exploited because only one or a few parameters that seem to be the best at first sight are used; a certain amount of information is therefore lost. Optimal combination of all parameters leads to better identification because more information is used. In pattern recognition, two different situations are considered according to whether the classes into which individual samples must be classified are known or not. In the first instance, one speaks of supervised learning and in the second of unsupervised learning. Some authors call the former pattern recognition and the latter pattern cogni-

tion. Only supervised learning is of interest here. Supervised learning means that a learning or training set, i.e., a number of classified individuals or samples, is developed, and this is used to define a classification rule which is subsequently applied to the classification of unknown samples.

(b) The feature selection or optimization problem. The response from multicomponent methods contains a lot of noise, because of the parameters that permit little discrimination or contain little information. It is desirable to remove the redundant parameters to simplify the ultimate interpretation. This problem is especially important in clinical chemistry. The cost of health care has increased so much that there is a growing interest in reducing the number of parameters or tests. The selection of a restricted set of measurements would also lead to a more rational and economically sounder use of chemical analysis. The optimization problem here is therefore the selection of a restricted number of parameters, i.e., feature selection.

Both problems are interrelated and can be solved by pattern recognition techniques. Two examples from this laboratory will be used for illustration, and will be referred to as the milk problem and the thyroid problem. The milk problem [1] in its simplest form consists of classifying milk samples as goat's milk or sheep's milk on the basis of the gas–liquid chromatographic (g.c.) spectrum of the fatty acids. The thyroid problem [2] consists (also in its simplest form) of distinguishing hypothyroid and euthyroid persons from the results of five chemical tests. In the milk problem, it is desirable to utilize the information present in the g.c. spectrum to the fullest possible extent, whereas in the thyroid example, the real question is to establish whether or not diagnostic results of the same quality can be obtained at a lower price, i.e., with only two or three tests instead of five.

PATTERN RECOGNITION METHODS

The major task of pattern recognition is to define criteria in order to classify individuals into known groups (such as goat's milk or hypothyroid persons). The individuals are characterized by a pattern of parameters or variables and the differences in these patterns are used to discriminate between the groups. For instance, when clear discrimination has been obtained between some thyroid diseases on the basis of 5 laboratory tests, each patient can be classified on the basis of a pattern consisting of the values obtained for these 5 tests. Many techniques have been developed to recognize such differences in patterns. These techniques have the property of defining the boundaries of the different groups. Such boundaries are often called decision boundaries. This problem can be displayed in a pattern space in which each parameter represents a coordinate and each individual a point in this space. The position of each individual can be represented by a pattern vector which for R parameters can be presented as follows

$$\mathbf{x} = x_1, x_2, \ldots x_i, \ldots x_R \qquad (1)$$

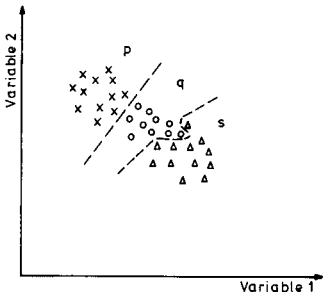


Fig. 1. Linear and non-linear separability between groups p (×), q (○) and s (△).

where each x_i is a scalar representation of the particular value associated with the ith dimension of the pattern space. Each group defines a cluster of related individuals, and the task of pattern recognition is to define the best boundaries between the different groups so that as many individuals as possible for which the groups are known, fall within the bounds of the right group. Pattern recognition techniques differ according to whether they lead to linear (or hyperplane in multidimensional pattern spaces) or non-linear boundaries. If the boundaries are linear, they are obtained by linear combination of the predicting variables. These linear combinations are called linear discriminant functions. Figure 1 shows a spatial representation for a simple two-parameter case with three groups. An optimal boundary has been drawn between the groups. Groups p and q can be completely separated by a linear boundary, but groups q and s cannot.

Pattern recognition methods can also be divided into parametric and non-parametric methods. Parametric classification techniques assume that the forms of the distribution of individuals in each group are known. Non-parametric techniques start from the point that the distributions are unknown. Parametric classification methods are often based on normal distributions and utilize mainly statistical parameters; in contrast, non-parametric methods are chiefly based on purely mathematical and geometrical derivations.

A pattern recognition method generally consists of several steps.

(1) Construction of the pattern space, i.e., selection of the individuals constituting the learning groups and the initial R parameters.

(2) Feature selection: a pattern consisting of many parameters often contains a lot of noise, i.e., useless information. These useless parameters tend to obscure the difference between classes and therefore render the separation more difficult. The simplification of the pattern space to a reduced so-called feature space is therefore necessary. The unnecessary parameters are then eliminated. The feature vector can be presented analogously to the pattern as follows

$$\mathbf{x} = x_1, x_2, \ldots, x_i, \ldots x_M \qquad (2)$$

where each x_i represents the particular value associated with the ith dimension

of the feature space and M indicates the total number of dimensions of this space so that $M \leqslant R$.

(3) Feature display (extraction or transformation): this consists of a visual representation of the pattern or feature space, usually by transformation of the variables in order to reduce the space to two or three dimensions.

(4) Classification: a classification or decision rule is developed with the aid of the data from the training set; this rule is then used to classify each unknown sample into one of the classes.

The place of statistical linear discriminant analysis in pattern recognition

Statistical linear discriminant analysis (s.l.d.a.) is one of the parametric classification methods of pattern recognition. It uses linear boundaries between the groups and is based on normal distributions of the parameters. The model used in s.l.d.a. contains some other assumptions (see below) from statistical multivariate analysis.

There are differences in terminology with pattern recognition. Shoenfeld and De Voe [3], in a classification of applications of statistical methods to analytical chemistry, noted that much frustration is due to non-uniformity of nomenclature. It is therefore necessary to consider these differences. The linear discriminant functions which are considered in non-parametric pattern recognition have properties of separating or defining classes distinct from those of s.l.d.a.; the object of the latter is to reduce the pattern space with minimal loss of differentiation between the classes. The functions which in non-parametric pattern recognition are called linear discriminant functions are called classification functions in s.l.d.a. The sequences of steps that may be observed in s.l.d.a. can be summarized as follows:

(1) As in other pattern recognition methods, the first step is to construct the pattern space (see above).

(2) Feature transformation: this establishment of the discriminant space consists of transforming the pattern space by linear combination of the different parameters of the pattern. The linear discriminant functions obtained in this way are used to define the axes of the discriminant space.

(3) Classification step: formulation of a classification rule in the discriminant space to classify unknown individuals; this is done by using classification functions.

(4) Feature selection: this can be done before or after step (2). In the former case so-called "stepwise" methods are used, and in the latter "direct" procedures (see *Selection of variables*).

STATISTICAL LINEAR DISCRIMINANT ANALYSIS

Construction of the pattern space

Consider a problem in which two groups p and q are investigated; each group consists of individuals or samples with a known but different pathology. Suppose that the results of two laboratory tests are available for each

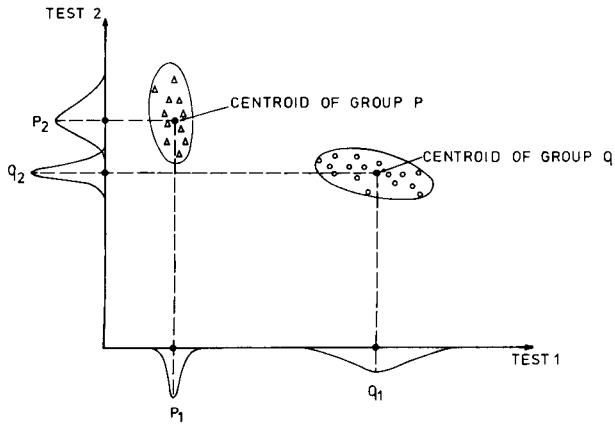


Fig. 2. Spatial representation of two pathological groups p and q on the basis of two laboratory measurements and their distribution characteristics on each axis.

individual. In a spatial representation, two groups are obtained, more or less distinct and characterized by the position of their centroid and by their internal dispersion (within-class variance) (Fig. 2). This can be generalized to R dimensions and leads to a representation of each individual in the R-dimensional pattern space. In the thyroid example, a 5-dimensional pattern space is obtained where each dimension is a laboratory test. These are: RT3U (T3 resin uptake), T4(D) (total serum thyroxine), T3RIA (total serum triiodothyronine), TSH (total serum thyroid stimulating hormone) and ΔTSH (increase of TSH after injection of thyrotropin-releasing hormone). Two learning groups (hypothyroidy and euthyroidy) are considered. These groups contain 30 and 150 individuals, respectively, thus 900 data are available. The milk example deals with a 15-dimensional pattern space, each dimension representing the percentage content of a fatty acid in the milk fat, as determined by g.c. The learning set here consists of 20 samples from sheep and 20 from goats.

Feature transformation

The objective in this step of s.l.d.a. is to weight and combine linearly the discriminating parameters so that the groups are forced to be as distinct as possible and the number of dimensions are reduced with minimum loss of differentiation between the groups. The new dimensions of the reduced discriminant space are obtained by linear combinations of the original dimensions.

In the special case of two groups, one linear combination or one discriminant axis is obtained, and when two variables are used the discriminant function can be represented by

$$f(x_1, x_2) = v_1 x_1 + v_2 x_2 \qquad (3)$$

The discriminant score of individual i is then represented by $DS_i = v_1 x_{1,i} + v_2 x_{2,i}$, where $x_{1,i}$ and $x_{2,i}$ are the values for patient i of variables, x_1 and x_2, respectively, and where v_1 and v_2 are the weights attached to these variables. For R variables the discriminant scores can be represented by $DS_i = \Sigma_{j=1}^{R} v_j x_{j,i}$. When more groups are considered, the maximum number of functions that can be derived is either one less than the number of groups or equal to the number of discriminating variables, if there are more groups than variables. A few tests for determining the number and relative importance of discriminant functions are available and will be considered in more detail below.

The effect of the weight coefficient can be illustrated by Fig. 3, which shows two discriminant functions f_1 and f_2 (i.e., two functions of the form of eqn. (3) with different weights) and the discriminant scores of the individuals of both groups p and q. The discriminant scores represent the orthogonal projection of each sample on the discriminant axis. It can be easily seen that f_2 provides a better separation between the groups than f_1. Clearly, some linear combinations provide better separations than others and therefore the weight attached to each variable must be optimized, so that orientation will be used which will provide maximum differentiation between the groups. Since separation between the groups is improved when the distance between the groups is larger (i.e., the between-groups variance is larger) and the clusters are tighter (i.e., the within-group variances are smaller), the optimal weight coefficients are obtained by maximization of the ratio of between-groups variance of the individuals to their within-group variances. Furthermore, the correlation between the parameters must be as small as possible. The information that can be obtained from two strongly correlated parameters is not much larger than that obtainable from one of these parameters. When two parameters are highly correlated, one of them can be considered redundant.

Thus, the information required to derive the weights consists of the population means and the population variances and covariances. However, "popu-

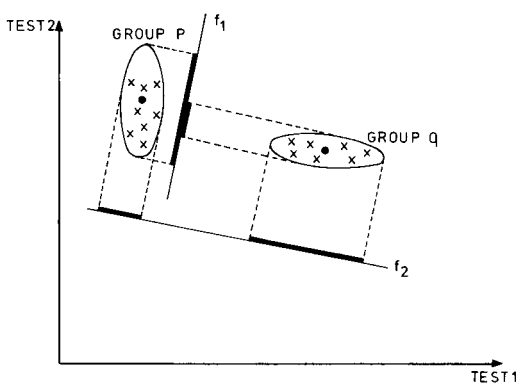


Fig. 3. The influence of the direction (i.e. the choice of the weight coefficients) of a discriminant axis on the discrimination between two groups p and q.

lation" refers to an infinite number of individuals and these parameters must be used to predict the population parameters, i.e., the means of the learning groups and variance—covariance matrix or a sum of squares matrix derived from it (see below).

The variables must be normally distributed in each class and the classes must have identical variance—covariance matrices. These requirements are necessary to determine an optimal set of weight coefficients or to obtain a discriminant space where the groups are optimally separated. Moreover, these requirements are not absolute but should be considered rather as conditions for achieving ideal circumstances. The choice of the classification procedure depends on the agreement with those restrictions. The more ideal the circumstances are, the better the classification procedure becomes. Normal distributions are also required for different statistical tests used during the procedure of s.l.d.a. (see Appendix). However, moderate deviations from the conditions given above can be allowed.

Classification

From the learning groups, a set of classification functions can be derived to allow the classification of new cases with unknown membership. The classification functions have the property that they divide the reduced discriminant space into regions, each region corresponding to a group. The classification functions can be represented in a scalar form as

$$g_p\,(x_1,\ldots,x_R) = \sum_{n=1}^{R} c_{n,p}x_n + c_{R+1,p} \tag{4}$$

and the classification score of individual i for group p is then represented by

$$CS_{p,i} = \sum_{n=1}^{R} c_{n,p}x_{n,i} + c_{R+1,p} \tag{5}$$

where the $c_{n,p}$ values are the classification weight coefficients, $x_{n,i}$ represents the values of the discriminating variables, and $c_{R+1,p}$ is a constant factor. Function g can be derived from the linear discriminant functions. The classification functions are defined in such a way that for all individuals or samples i within the region describing a group p $CS_{p,i} \geqslant CS_{q,i}$ for all $q \neq p$. This expression is called the classification rule.

Practically, a classification rule may be used as follows: assign an individual i to the group whose centroid is nearest to that individual. In the special case of two groups p and q, the discriminant space is given by a single discriminant function and is therefore one-dimensional (Fig. 3). An individual or sample i can be classified according to its position DS_i on the discriminant axis as compared to the positions of the centroids $\overline{DS}_p$ and $\overline{DS}_q$ of groups p and q, respectively. In Fig. 3 the individual i is assigned to group p when the distance between DS_i and $\overline{DS}_P$ is smaller than the distance between DS_i and $\overline{DS}_q$. The classification rule can be formulated as follows: assign individual i to group p if $|DS_i - \overline{DS}_p| < |DS_i - \overline{DS}_q|$; otherwise assign it to group q.

When the classification concerns more than two groups, a discriminant space with more dimensions is needed. The classification functions and classification rule can be formulated in the same way but the quadratic Euclidean distance between individual and group centroids is now used. A detailed discussion has been given by Andrews [4].

On the assumption of a multivariate normal distribution, the classification score can also be converted into probabilities of group membership. The rule of assigning a case to the group with the highest score is then equivalent to assigning the case to the group for which it has the greatest probability of membership (see Appendix). A Bayesian adjustment of this probability is often desirable when the costs of misclassification into certain groups are very high, when the groups are of grossly different sizes, or when it is desirable to take advantage of a priori knowledge of group membership probabilities. For example, in the thyroid problem, it is possible to take into account the prevalence of euthyroidy in the population.

As a check on the adequacy of the classification functions, the original learning set can be classified to see how many are correctly classified by the variables used. To obtain a more correct estimate [5, 6], the data set can be divided into two subsets, one for the development of the discriminant and classification functions (learning set) and one to check the correctness of the classification (test set). Since the individuals of the test set are not used in the development of the classification rules, it can be expected that the result of this check will be less optimistic than it is for the procedure where the test and learning sets are the same. The so-called leave-one-out procedure [7] can also be used, but it is very time-consuming.

Selection of variables (feature selection)

To trace redundant variables, several criteria are available. These criteria attach numerical values to the discriminating power of each individual variable or in relation to others. Some of the criteria are based on statistical distributions and tests and are called "parametric". Coomans et al. [2] discussed some parametric selection criteria in an application concerning the diagnosis of thyroid diseases. There are two kinds of criteria: criteria that do not take into account correlated information and those that do. Indeed there are two main causes for the appearance of redundant variables: (1) the variables contain little individual information (i.e., by themselves they permit little discrimination between the groups studied), or (2) they contain correlated information (the variable in question gives little additional information compared to another variable). Another difference between parametric selection procedures is their position in the s.l.d.a. procedure (see *Place of s.l.d.a. in pattern recognition*, above). One can consider criteria based on the discriminant functions obtained; selection procedures controlled by these criteria are applied after the feature transformation step in s.l.d.a., and only "direct" selection methods are used here.

One can also consider criteria that are independent of the discriminant

functions; these are used principally in "stepwise" selection procedures. Since no knowledge about the discriminant functions is necessary, such methods precede the feature transformation step.

Criteria based on discriminant functions give a larger importance to a variable when the (absolute) value of the corresponding weight coefficient is higher, provided that the variables have been standardized. The criteria trace both kinds of redundancy, i.e., little individual information and correlated information. A direct selection method involves ranking the variables in order of decreasing importance according to the given criterion. Another direct method is to determine the contribution percentage of each variable to the total distance D^2 in the discriminant space, which is the distance between the centroids of the groups which are considered. The contribution percentage of variable n is given by $100 \times |v_n \delta_n| / D^2$, where v_n is the weight coefficient of the discriminant function for variable n and $\delta_n = (\bar{x}_{n,p} - \bar{x}_{n,q})$. $\bar{x}_{n,p}$, $\bar{x}_{n,q}$ are the mean values of x_n for groups p and q, respectively. Thus $D^2 = \sum_{n=1}^{R} |v_n \delta_n|$.

Criteria which are independent of the discriminant functions differ according to whether they take into consideration both individual and correlated information or only the former [2]. Although these criteria also permit direct selection methods, stepwise selection methods are mainly used. A variety of criteria can be found in the SPSS [8] computer program. The following criteria are available for controlling the stepwise selection of the best set of discriminating variables: the smallest Wilks' lambda, the largest Mahalonobis distance between closest groups, the largest F value between closest groups, the residual variation and the largest Rao's V value. Each of these emphasises a different aspect of the separation.

Stepwise selection begins by choosing the single variable which has the best value (i.e., the highest or the lowest depending on the criterion chosen) for the selection criterion. This initial variable is then paired with each of the other available variables, one at a time, and the selection criterion is computed again. The new variable which produces the best new criterion value in conjunction with the initial variable is selected as the second variable. In this way, the process is continued until a certain number of variables is reached or a certain discrimination is obtained.

The use of a stepwise procedure results in a near-optimal or optimal set of variables being selected. The completely optimal result will not always be obtained with this procedure; it will be obtained with certainty only when all the combinations of the variables are considered. If, for example, variable 3 is the first to be selected, only combinations with variable 3 are considered from then on. However, it may be that the best combination of two variables does not comprise variable 3. The same problem occurs when information theory is applied to the selection of sets of g.c. phases or t.l.c. systems. It was shown by Eskes et al. [9] that the best set of two phases or systems does not necessarily include the individually best one.

Reduction of the number of variables can lead to better classification be-

TABLE 1

Effectiveness of the HYPO/EU discrimination after stepwise selection by Rao's V criterion

Laboratory tests used for classification	% correctly classified
5 VAR : T4 + ΔTSH + RT3U + TSH + T3RIA	96.6
4 VAR : T4 + ΔTSH + RT3U + TSH	97.2
3 VAR : T4 + ΔTSH + RT3U	97.8
2 VAR : T4 + ΔTSH	99.4

cause noise from redundant variables may obscure the classification in such a way that useful information is lost among the noise. As an example, Table 1 gives the results of the classification after feature selection for the HYPO/EU discrimination of the thyroid example [2]. S.l.d.a. was applied after each deletion of the less important laboratory test. It can be seen that the number of individuals correctly classified increases with decreasing number of laboratory tests. Clearly RT3U, TSH and T3RIA do not contribute to the discrimination, but only add noise.

LITERATURE

A few books are listed to provide an introduction to the techniques discussed. General treatments of multivariate statistical analysis are found in the books by Cooley and Lohnes [10], Tatsuoka [11], and Morrison [12]. Each of these contains a chapter on s.l.d.a. Cooley and Lohnes describe computer programs in FORTRAN computer language. These authors and Tatsuoka put the emphasis on the derivation of discriminant functions and statistical tests, and Morrison on s.l.d.a. as a classification method. Other books of value in this context have been written by Van Ryzin [13], Harris [14], Seal [15], Kendall and Stuart [16] and Dagnelie [17].

A good elementary text on linear algebra and matrix theory, providing the background necessary to read the statistical literature, has been written by Noble [18]. Summaries of the more important properties of matrices are also given in the books on multivariate statistics [10—12].

A useful book confining itself to the domain of s.l.d.a. and containing the computer programs for computation in Fortran language was written by Romeder [19]. Lachenbuch [20] described similar material. S.l.d.a. in the context of pattern recognition has been discussed by Andrews [4] and by Duda and Hart [21], Patrick [22], Fukunaga [23], Meisel [24] and Young and Calvert [25]. A book on optimization techniques in analytical chemistry, written by Massart et al. [26], contains chapters on pattern recognition and s.l.d.a. The review articles of Kowalski and Bender [27—29], Pratt et al. [30], MacDonald and Levine [31] and Albano et al. [32] concerning the application of pattern recognition to chemistry are also of great interest.

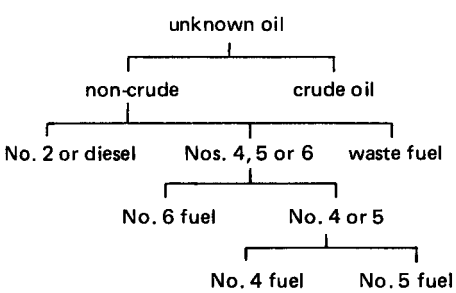
The computations with regard to s.l.d.a. (analysis, classifications and

feature reduction) can be carried out by using commercially available programs among which the best known are: SPSS (Statistical Package for the Social Sciences) [8] and BMDO7M (Biomedical computer programs) [33]. ARTHUR [34], a collection of pattern recognition and general data analysis FORTRAN programs also allows computations concerning s.l.d.a. General information and comments are available in the book by Kowalski [35].

APPLICATIONS

S.l.d.a. has been used very often in the social and economical sciences. Although there are some applications in analytical chemistry and food analysis, the emphasis in this domain of sciences is on distribution-free non-parametric methods. One of the domains in which pattern recognition has been applied in analytical chemistry, is in the analysis of petroleum. Kawahara and Yang [36] used s.l.d.a. for the characterization and identification of petroleum pollutants on the basis of infrared spectrophotometric data with a 99% correct classification of numerous known and unknown oil samples. The known samples constitute three learning groups, namely asphalt, crude oil residue and No. 6 fuel oil samples. "Unknown" samples include weathered oils, or asphalts, the identity of which was determined by other instrumental analysis methods. The 41 discriminating parameters (absorption at particular wavelength positions) caused only two misclassifications for 178 known samples and none for the 16 unknowns. Earlier and related investigations of Kawahara et al. [37], Julian [38] and Santner [39] supported the fact that s.l.d.a. allows more of the information in the laboratory to be utilized and results in a better method of distinguishing between petroleum materials.

A more recent application described by Mattson et al. [40] also deals with the classification of petroleum pollutants based on infrared spectral patterns. Digitized transmission i.r. spectra of 194 oils (62 crude oils, 60 No. 2 and diesel fuels, 28 No. 6 fuels, 22 waste crankcase lubes, 12 No. 4 fuels and 10 No. 5 fuels) were used and several "decision-tree" schemes were tested to develop a high recognition power (% samples correctly classified). The following "decision-tree" scheme resulted in 97.5% correct recognition.



Other applications concern food analysis. Powers and Keith [41] used s.l.d.a. as an aid in the classification of the flavour quality of coffee and

potato chips based on gas chromatographic data. Four lots of roasted coffee with different organoleptical qualities, were examined by g.l.c. By calculating all possible ratios among peak heights and subjecting these ratios to s.l.d.a., the coffee could be classified into the four flavour categories. The stepwise procedure was used to select those ratios that contributed most to the differentiation. An investigation concerning the determination of the origin of peppermint oils by Hartmann and Hawkes [42] affirms the applicability of linear discriminant analysis in this domain. Hartmann and Hawkes were able to classify all 45 pure oils from ten origins correctly on the basis of 12 discriminating parameters obtained by g.l.c. An extension to binary mixtures of oils was obtained by considering all possible combinations of the mean compositions of two geographical origins in proportions from 5% to 95% in steps of 5%. The 450 binary mixtures obtained in this way were all correctly assigned to the correct pair of origins in the correct proportions. Smeyers-Verbeke et al. [1] used s.l.d.a. for the identification of milk samples. The set of predicting variables comprised gas chromatographic data concerning the distribution of fatty acids in milk fat; samples from cows, sheep and goats provided the three learning groups. Between the pure samples, s.l.d.a. allowed complete discrimination. Extension of the investigation to mixtures of these samples showed a high degree of correct classifications between the mixtures and the pure milks. Some degree of reduction of the g.l.c. data pattern was also possible without considerable loss of discrimination efficiency between the groups.

Vandeginste and Van Iersel [43] applied s.l.d.a. to the classification of water quality. The learning groups consisted of 5 groups of water quality from "badly polluted" to "not polluted" as determined by biological quality. As parameters they used chemical and physical quantities such as BOD, COD, ammonium content, pH, temperature, etc. They showed that s.l.d.a. makes it possible to classify surface waters by using these chemical parameters into the right or nearest category of biological quality with 85% success.

Another important domain where the suitability of s.l.d.a. has been shown is clinical chemistry. Applications in this field of analytical chemistry may be found in the review articles by Solberg [44] and De Waard [45]. It is not possible to describe the complete literature in this field here. A few illustrative examples are given below.

Amenta and Harkins [46] used discriminant functions to evaluate the contributions of various phosphate clearance tests in distinguishing between patients with hyperparathyroidism and control patients. They demonstrated that discriminant functions can optimize diagnostic classification and can be important tools in evaluating the effectiveness of laboratory procedures. In a related investigation, Fraser et al. [47] utilized s.l.d.a. to differentiate between some (chiefly neoplasmatic) diseases causing hypercalcemia; biochemical parameters such as inorganic phosphate, alkaline phosphatase, chloride, hydrogen carbonate and urea were used as the input. In 90.4% of

cases (for a total of 218) the s.l.d.a. classification coincided with the detailed retrospective diagnosis.

The diagnostic effectiveness of electrophoresis and specific protein assays for different diseases was evaluated in the same way by Werner et al. [48]. Seven diagnostic categories were considered (normal, cirrhosis, hepatitis, A-plasmocytoma, G-plasmacytoma, systematic lupus erythematosus and nephritis). The discriminating variables were 12 specific proteins and 5 electrophoresis fractions.

Biochemical liver function tests were tested with regard to their functional specificity by Baron [49]. Linear discriminant functions of 12 tests discriminated between 6 learning sets for different diseases: acute infective hepatitis, alcoholic cirrhosis, primary biliary cirrhosis, active chronic hepatitis, inactive cryptogenic cirrhosis, and active cryptogenic cirrhosis. A suitable separation between the groups could not be obtained.

Ramsoë et al. [50] applied s.l.d.a. to clinical laboratory tests to distinguish between cirrhotic and healthy individuals and demonstrated that measurements of bromsulphalein excretion and γ-globulin allow discrimination between the two groups with the same effectiveness as all the 9 laboratory tests considered. A related but more extended investigation was done by Winkel et al. [51]. Two classification problems were considered: a 3-fold discrimination between hepatitis, fatty liver and chronic liver disease and a 5-fold discrimination which was an extension of the preceding 3-fold one. The percentage of patients correctly classified was generally rather poor in the 3-group problem with 6 routine liver tests (80%) and in the 5-group problem with the same routine liver tests (63%). This analysis demonstrated the need for more discriminating liver function tests in the diagnosis of chronic liver diseases and liver tumours. A supplementary study was carried out by Solberg et al. [52] for 8 liver diseases and 28 laboratory tests. Validation of the results obtained by s.l.d.a. was done by means of cluster analysis [53]. An alternative approach to liver diseases is the determination of the origin of jaundice. Carlström et al. [54] used s.l.d.a. to do this.

Different thyroid function tests were investigated by Barnett et al. [55]. A classification of 204 patients with suspected thyroid disorders (thyrotoxic, euthyroid and hypothyroid), was carried out on the basis of all test results (10 variates). The problem was subdivided into two binary cases (i.e., thyrotoxic vs. euthyroid and euthyroid vs. hypothyroid). Two in vitro tests (PBI and T-3 uptake) sufficed to give a good separation of thyrotoxic from euthyroid patients and permitted fairly good distinction of hypothyroid patients. The addition of a third test (^{131}I uptake) made it possible to classify most patients correctly (95.5%). Other tests, including clinical questionnaires, were discarded as poor discriminants.

Mayron et al. [56] applied the technique to separate drug abusers and normal controls on the basis of 9 laboratory variables. They made 4% errors in the control groups and 14% in the abusers group. Very illustrative too is the paper of Wilding et al. [57] where the influence of drugs for treatment

of rheumatoid arthritis on biochemical and haematological data was demonstrated by s.l.d.a.

S.l.d.a. has also been used in the diagnosis of cancer [58], shock [59] and myocardial infarction [60].

CONCLUSION

S.l.d.a. should find widespread applications in analytical chemistry since it provides answers to several multivariate problems. It permits removal of redundant variables, optimal linear combination of meaningful variables, and easy formulation of classification rules. However, there are restrictions to this technique.

Two criticisms can be formulated concerning the classification rules used in s.l.d.a. First, when deviations from normality of the distributions and non-equality of the variance–covariance matrices are large, the linear decision boundary situated halfway between the group centroids is not the optimal boundary. Secondly, since a linear decision boundary does not take into account the presence of outliers in the data set, the classification rule may be regarded as somewhat rudimentary in contrast to some more recent methods such as SIMCA [32, 61, 62].

The authors thank I. Broeckaert, M. Jonckheer, P. Blockx, R. Bosman and J. Smeyers-Verbeke for making their data available.

MATHEMATICAL APPENDIX

Definitions

Pattern vector, data matrices and parameter values. To agree with general statistical and pattern recognition literature, each vector will be defined as a column vector. A pattern vector consisting of R parameters for an individual i belonging to group p can be represented as

$$\mathbf{x}_{pi} = \begin{pmatrix} x_{pi1} \\ \vdots \\ x_{pij} \\ \vdots \\ x_{piR} \end{pmatrix} = (x_{pij})_{R \times 1}$$

A row vector, the transpose of a column vector, is indicated by t and written as follows: $\mathbf{x}_{pi}^t = (x_{pi1}, \ldots, x_{pij}, \ldots, x_{piR})$. Suppose that there are K groups (with indices $p = 1, 2, \ldots, K$) with a total number of N individuals, so that $N = \sum_{p=1}^{K} N_p$, where N_p is the number of individuals belonging to group p. The individuals of group p are indicated by $i = 1, 2, \ldots, N_p$. Suppose also there are R parameter values for each pattern with indices $j = 1, 2, \ldots, R$. The $(N_p \times R)$ data matrix for group p is then represented by:

$$X_p = \begin{pmatrix} x_{p11} & \cdots & x_{p1R} \\ \vdots & & \vdots \\ x_{pN_p1} & \cdots & x_{pN_pR} \end{pmatrix} = \begin{pmatrix} x_{p1}^t \\ \vdots \\ x_{pN_p}^t \end{pmatrix} = (x_{pij})_{N_p \times R} \tag{A.1}$$

The mean values of the jth parameter in group p and for all individuals are given, respectively, by

$$x_{p.j} = \sum_{i=1}^{N_p} x_{pij}/N_p \text{ and } x_{..j} = \sum_{p=1}^{k} N_p x_{p.j}/N$$

Within-group SSCP matrix and between-groups SSCP matrix. The $(R \times R)$ within-group sum of squares and cross products (SSCP) matrix of each group p is of the form

$$W_p = \begin{pmatrix} w_{p11} & \cdots & w_{p1R} \\ \vdots & & \vdots \\ w_{pR1} & \cdots & w_{pRR} \end{pmatrix}$$

where an element of the matrix is

$$W_{pjk} = \sum_{i=1}^{N_p} (x_{pij} - x_{p.j})(x_{pik} - x_{p.k}) \tag{A.2}$$

where j and k stand for the jth and kth parameters. The total within-group SSCP matrix is then

$$W = \sum_{p=1}^{K} W_p \tag{A.3}$$

Analogously, the $(R \times R)$ between-groups SSCP matrix of each group p is of the form

$$B_p = \begin{pmatrix} b_{p11} & \cdots & b_{p1R} \\ \vdots & & \vdots \\ b_{pR1} & \cdots & b_{pRR} \end{pmatrix}$$

where each element of the matrix is $b_{pjk} = N_p (x_{p.j} - x_{..j})(x_{p.k} - x_{..k})$. The pooled between-groups SSCP matrix is then given by

$$B = \sum_{p=1}^{K} B_p \tag{A.4}$$

If b_{jk} is an element of B defined by $b_{jk} = \Sigma_{p=1}^{K} b_{pjk}$ and $h_{pj} = x_{p.j} - x_{..j}$, so that $\Sigma_{p=1}^{K} N_p h_{pj} = 0$, then b_{jk} will be given by

$$b_{jk} = \sum_{p=1}^{K} (N_p^{\frac{1}{2}} h_{pj})(N_p^{\frac{1}{2}} h_{pk}) \tag{A.5}$$

If H now represents the $(K \times R)$ matrix

$$H = (N_p^{\frac{1}{2}} h_{pj})_{K \times R} \tag{A.6}$$

it can be seen that B is given by $B = H^t H$.

Within-group sum of squares and between-groups sum of squares. The within-group sum of squares for variable j (SS_{jw}) is defined here by

$$SS_{jw}(x) = \sum_{p=1}^{K} \sum_{i=1}^{N_p} (x_{pij} - x_{p.j})^2 \tag{A.7}$$

This represents a diagonal element of W (eqn. A.3). The between-groups sum of squares for variable j (SS_{jb}) is

$$SS_{jb}(x) = \sum_{p=1}^{K} N_p (x_{p.j} - x_{..j})^2 \tag{A.8}$$

which represents a diagonal element of B (eqn. A.4).

Normal distributions. When an infinite population P is assumed which is normally distributed in R parameters and from which group p is a sample, the R-variate normal distribution function of P is of the form

$$\gamma_p(\mathbf{x}) = \{1/[(2\pi)^{R/2}|\Sigma_p|^{\frac{1}{2}}]\} \exp[-\tfrac{1}{2}(\mathbf{x} - \vec{\mu}_p)^t \Sigma_p^{-1} (\mathbf{x} - \vec{\mu}_p)] \tag{A.9}$$

The distribution is fully determined by the statistical parameter $\vec{\mu}_p$ [$(R \times 1)$ population mean vector] and Σ_p [$(R \times R)$ population variance–covariance matrix]. When the population parameters are usually unknown and when group p is a representative sample of population P, the population parameters are predicted by the following statistical sample parameters: $\hat{\Sigma}_p$, the estimator of Σ, is given by $\hat{\Sigma}_p = W_p(N_p - 1)$; $\hat{\mu}_p$, the estimator of $\vec{\mu}_p$, is given by $\vec{\hat{\mu}}_p = (\hat{\mu}_{p.j})_{R \times 1} = (x_{p.j})_{R \times 1}$ or $\hat{\mu}_{p.j} = x_{p.j}$, and $x_{p.j}$ is defined as above.

Since each population P is normally distributed in its R parameters, this is also true for the total data set. Σ and $\vec{\mu}$ are then estimated by $\vec{\hat{\mu}} = (\hat{\mu}..j)_{R \times 1} = (x..j)_{R \times 1}$ and $\hat{\Sigma} = W/(N - K)$.

Computation of the discriminant functions

One or several linear functions

$$f(x_1, \ldots, x_R) = \mathbf{v}^t\mathbf{x} = v_1x_1 + \ldots + v_Rx_R \tag{A.10}$$

of R variables are sought for which the variance among the groups is as significant as possible, i.e. the first step is to formulate a criterion for measuring differences between the group means. Such a criterion is given by the familiar F-ratio for testing the significance of the overall difference among several group means in a single variable. As in analysis of variance, it is generally assumed that the variables have to be normally distributed in each group (population) with identical variance–covariance matrices or $\Sigma = \Sigma_1 = \ldots = \Sigma_k$. When individual i of group p is characterized by its pattern vector $\mathbf{x}_{pi}$, for each function $f(x_1, \ldots, x_R)$, a discriminant score DS_{pi} is obtained so that $DS_{pi} = \mathbf{v}^t\mathbf{x}_{pi}$.

The coefficient vectors v are chosen for which the F-ratio reaches a maximum value. The F-ratio is defined by $F = SS_b(DS)/SS_w(DS)$; $SS_b(DS)$ and $SS_w(DS)$ are analogous to those of eqns. (A.7) and (A.8) where the single variates (x values) are replaced by the representative linear combined values

which are also single variates (DS values). For instance, $x_{..j}$ will be replaced by $\overline{DS}$ where $\overline{DS} = \Sigma_{p=1}^{K} N_p \overline{DS}_p / N$ and $\overline{DS}_p = \Sigma_{i=1}^{N_p} DS_{pi}/N_p$.
When the conditions of a valid analysis of variance, mentioned before, are not fulfilled, it is still possible to obtain a vector v that maximizes the F-ratio. Furthermore, it can be shown [11] that the $SS_b(DS)/SS_w(DS)$ ratio can be written as a function of the $(R \times 1)$ vector of combining weights $\mathbf{v} = (v_j)_{R \times 1}$:

$$F = SS_b(DS)/SS_w(DS) = \mathbf{v}^t B\mathbf{v}/\mathbf{v}^t W\mathbf{v} \tag{A.11}$$

The F-ratio is a criterion for measuring the differentiation between groups along the dimension specified by the vector v. The F-ratio is maximized by seeking the appropriate vector v. Therefore the partial derivative of the ratio with respect to each component v_j of v is equated to zero. It can be shown that vector v is determined except for a constant factor [10]. To obtain only one v value, the supplementary condition $\mathbf{v}^t W\mathbf{v} = N - K$ is added; Cooley and Lohnes [10] prefer the condition $\mathbf{v}^t\mathbf{v} = 1$.

If a Lagrange multiplier (ρ) and the condition expressed by eqn. (A.14) are used, the maximization is obtained by

$$\frac{\delta}{\delta \mathbf{v}} [(\mathbf{v}^t B\mathbf{v}/(N-K)) - \rho(\mathbf{v}^t W\mathbf{v} - (N-K))] = 0 \tag{A.12}$$

This leads to

$$B\mathbf{v} = \rho(N-K)\, W\mathbf{v} \tag{A.13}$$

It may be assumed that W is non-singular [11] so that an inverse matrix W^{-1} can be computed. Equation (A.13) can then be transformed to $A\mathbf{v} = \rho(N-K)\,\mathbf{v}$ where $A = W^{-1}B$. When the non-singular $(R \times R)$ matrix A is characterized by S non-zero eigenvalues λ_s (for $s = 1, 2, \ldots, S$) so that $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_s$, then the F-ratio will be maximized when $\rho(N-K)$ corresponds to the largest eigenvalue λ_1 of A. Indeed, it follows from eqn. (A.13) that

$$\mathbf{v}^t B\mathbf{v}/\mathbf{v}^t W\mathbf{v} = \rho(N-K) \tag{A.14}$$

and it is clear from the above argument that $\rho(N-K)$ and v indicate, respectively, the eigenvalues and eigenvectors of A. The Lagrange multiplier for the first eigenvalue λ_1 will then be $\rho = \lambda_1/(N-K)$, and v, indicated by $\mathbf{v}_1$, will be the eigenvector of A corresponding to eigenvalue λ_1 which corresponds to the first (and most discriminating) weight vector. The other weight vectors are determined in the same way.

It can be shown that the number of non-zero eigenvalues of a square matrix A is equal to the rank of A. In the present context, with $W^{-1}B$, the number of non-zero eigenvalues depends on the rank of $W^{-1}B$ and therefore of B, since the rank of the product of two matrices cannot exceed the smaller of the two factor matrix ranks, and since W^{-1}, being non-singular, must be of full rank R, while the rank of B is usually smaller than R. The rank of B is at most that of H (eqn. A.6) and this is at most the smaller of either $K-1$ or R. Indeed, if each row of H is multiplied by $N_p^{\frac{1}{2}}$ (for $p = 1, 2, \ldots, K$) and then

the elements of the columns are summed over $j = 1, 2, \ldots, R$, the sums are equal to zero. This can be written as $\Sigma_{p=1}^{K} N_p h_{ipj} = 0$ which implies that the rank of H is at most equal to the smallest of either $K - 1$ or R.

It can be shown, if the variance–covariance matrix is common to the K groups, that the following equation is valid:

$$\sum_{p=1}^{K} \left(\frac{N_p}{N-K} \right) S_p = I \tag{A.15}$$

where $S_p = (S_{pst})_{S \times S}$ and $S_{pst} = 1/N_p \; \Sigma_{i=1}^{N_p} (DS_{pis} - DS_{p.s})(DS_{pit} - DS_{p.t})$. Here DS_{pis} and DS_{pit} indicate the discriminant score of individual i from group p, respectively, on discriminant axes s and t or mathematically, it can be formulated as follows: $DS_{pis} = \mathbf{v}_s^t \mathbf{x}_{pi}$ and $DS_{pit} = \mathbf{v}_t^t \mathbf{x}_{pi}$. $DS_{p.s}$ and $DS_{p.t}$ represent, respectively, the discriminant scores for the centroid of group p. S_p represents the $(S \times S)$ variance–covariance matrix of the discriminant scores calculated for the S non-zero eigenvectors and I an $(S \times S)$ identical matrix. Equation (A.15) shows that the discriminant scores have variance one and covariance zero.

It can also be shown that the between-groups variance–covariance matrix Λ_c of the S discriminant scores is given by:

$$\Lambda_c = \left(\frac{1}{N-K} \sum_{p=1}^{K} N_p (DS_{p.s} - DS_{..s})(DS_{p.t} - DS_{..t}) \right) = \begin{pmatrix} \lambda_1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_s & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & \lambda_S \end{pmatrix} \tag{A.16}$$

$DS_{p.s}$ here indicates a coordinate of the centroid of groups p on discriminant axis s of the discriminant space and $DS_{..s}$ a coordinate of the centroid of a cluster defining the K different group centroids on the same discriminant axis. Equations (A.16) show that the cluster of the K group centroids in the S-dimensional discriminant space has covariances equal to zero and variances equal to λ_s $(s = 1, 2, \ldots, S)$.

When only two groups are considered, the computations for determining the single discriminant function can be simplified. The single discriminant function can be obtained without solving an eigenvalue problem [10]. The weight vector v for the discrimination of groups p and q is given by

$$\mathbf{v} = \hat{\Sigma}^{-1} m (\vec{\mu}_p - \vec{\mu}_q) \tag{A.17}$$

This equation shows that v is determined except for an arbitrarily chosen multiplier m. A computational example of how to do this in practice is given by Kendall [63].

Determination of the number of significant discriminant functions

Up to this point, the dimensionality of the discriminant space has been considered to be equal to the number of non-zero eigenvalues of $W^{-1} B$

which is the smaller of the two numbers, $K-1$ and R. In fact, the number of significant discriminant dimensions is often smaller. Several tests for determining the number of significant discriminant functions are available. One of these is the relative percentage of the eigenvalue associated with the function. An important criterion for eliminating discriminant functions is to test for the statistical significance of discriminating information not already accounted for by the earlier functions.

For this purpose Wilks' Λ criterion is used, because there is an algebraic relation between this criterion and the eigenvalues for the successive discriminant functions. This relation can be expressed by

$$1/\Lambda = (1+\lambda_1)(1+\lambda_2)\ldots(1+\lambda_S) \tag{A.18}$$

where $\lambda_1, \lambda_2, \ldots \lambda_S$ are the non-zero eigenvalues of A. The significance of an observed Λ value can be tested by means of Bartlett's V statistic [11], which can be expressed as $V = [N-1-(R+K)/2] \ln 1/\Lambda$. This statistic is distributed approximately as a χ^2-statistic with $R(K-1)$ degrees of freedom. The sth component now approximately follows a χ^2-distribution with $R+K-2s$ degrees of freedom:

$$V_s = [N-1-(R+K)/2] \ln(1+\lambda_s) \tag{A.19}$$

When V_1, V_2 and so on are sequentially subtracted from V, the remainder is also a χ^2-variable with $R(K-1)-(R+K-2)\ldots-(R+K-2s)$ degrees of freedom. These successive remainders become appropriate statistics for testing whether the residual discrimination is statistically significant. As soon as the residual, after removal of some discriminant functions, becomes smaller than the prescribed significance level α of the appropriate χ^2-distribution, it may be concluded that only these discriminant functions are significant at the α-level.

Classification of unknown cases

Now consider an individual whose group membership is unknown, for which one has measured R variables and wishes to assign it to one of the K groups. Generally, the classification functions are defined in such a way that for all $\mathbf{x}$ within the region described by a cluster or group p, there exists a function $g_p(\mathbf{x})$ so that $g_p(\mathbf{x}) > g_q(\mathbf{x})$ for all $p \neq q$. The surface separating region p from q is given by $g_p(\mathbf{x}) - g_q(\mathbf{x}) = 0$. There are $K(K-1)/2$ such separating surfaces in a K group problem.

Besides the scalar form mentioned under *Classification* a classification function can also be represented in vector form $g_p(\mathbf{x}) = \mathbf{c}_p^t\mathbf{x} + c_{R+1,p}$, where $\mathbf{c}_p$ is a $(R \times 1)$ column vector called the weight vector. Particularly, in multivariate statistical analysis and if S eigenvectors are used, the decision rule may be: assign each individual i to group p if

$$\sum_{s=1}^{S} [\mathbf{v}_s^t(\mathbf{x}_i - \mathbf{x}_p)]^2 = \min_q \sum_{s=1}^{S} [\mathbf{v}_s^t(\mathbf{x}_i - \mathbf{x}_{.q})]^2 \tag{A.20}$$

where $\mathbf{v}_s$ is the eigenvector of A corresponding to eigenvalue λ_s (see above). $\mathbf{x}_{.p}$ and $\mathbf{x}_{.q}$ are the mean vectors of groups p and q, respectively. The probability of misclassification can be estimated if it is assumed that the x values are R-variate normal.

Other decision rules that are also used are the minimum χ^2-rule and the minimum Mahalonobis distance rule. The classification rule expressed in terms of the sample Mahalonobis distance, $D^2_{p.i} = (\mathbf{x}_i - \mathbf{x}_{.p})^t\, \hat{\Sigma}^{-1}\, (\mathbf{x}_i - \mathbf{x}_{.p})$, of the unknown observation from the mean of the pth group is: assign $\mathbf{x}_i$ to group p if $D^2_{pi} = \min D^2_{1i}, \ldots, D^2_{Ki}$, so that the classification rule is satisfied if $D^2_{pi} < D^2_{qi}$ for all $q \neq p$. D^2_{pi} is a classification function which makes it possible to decide if individual i is a member of group p. In the discriminant space, the Mahalonobis distance reduces to a simple Euclidean distance $D^2_{pi} = (\mathbf{x}_i - \mathbf{x}_{.p})^t\, (\mathbf{x}_i - \mathbf{x}_{.p})$. The minimum χ^2-rule represents a different approach but is in principle the same as the minimum Mahalonobis distance classification [11].

Other classification rules are based on an estimation of an a posteriori probability of group membership and use the Bayes estimation [16]. It can be shown that the classification can then be formulated as follows:

$$CS_{pi} = \mathbf{x}_i^t \hat{\Sigma}^{-1} \mathbf{x}_{.p} - \tfrac{1}{2} \mathbf{x}_{.p}^t \hat{\Sigma}^{-1} \mathbf{x}_{.p} \qquad \text{(A.21)}$$

If the variance–covariance matrix $\hat{\Sigma}$ is the same for all groups and the a priori probability is equal for each group, the a posteriori probability of group membership is then

$$P(p/\mathbf{x}_i) = \exp\,(CS_{pi}) / \sum_{n=1}^{K} \exp\,(CS_{ni}) \qquad \text{(A.22)}$$

An example of classification carried out by means of the minimum Mahalonobis distance and the a posteriori probability of group membership is given below.

Discriminant functions containing a constant

Sometimes equations are found containing a constant factor. A discriminant function such as described by eqn. (A.10) is developed on the basis of unstandardized (raw) data. It is also possible to develop discriminant functions after normalizing the variables (standardizing among the groups). The discriminant scores are of the form $DS_i = \Sigma_{n=1}^{R} v'_n z_{ni}$, where the z values are the standardized values of the R variables. Such a standardized discriminant score can be transformed to an unstandardized one, in which a constant factor then occurs and where the raw values of the variables are used to compute the discriminant score for each individual i. The transformed score is of the form $DS_i = \Sigma_{n=1}^{R} v''_n x_{ni} + v_{n+1}$, where $v''_n = v'_n/\hat{\sigma}_n$ and the constant factor $v_{n+1} = -\Sigma_{n=1}^{R} v'_n x_{.n}/\hat{\sigma}_n$. The total standard deviation and the total mean value (i.e., concerning all groups) are given by $\hat{\sigma}_n$ and $x_{.n}$, respectively.

An example of results obtained with the SPSS programs [8]

In this section some practical computational examples are given for the EU/HYPER(HYPO/150/35/30) cases discrimination. It is an extension of the thyroid example [2]. The differentiation between three thyroid functional states is studied here on the basis of five laboratory tests. Basic statistical data are given in Table 2.

Linear discriminant functions. Since a 3-fold classification problem is involved, two discriminant functions can be computed. For this purpose the standardized variables are used. The first discriminant function f_1 corresponds to an eigenvalue λ_1 of 3.85 and the second discriminant function f_2 to an eigenvalue λ_2 of 0.73. The discriminant scores for an individual i along the optimal discriminant function axes f_1 and f_2 are given by $DS_{1,i}$ and $DS_{2,i}$, respectively, where

$$DS_{1,i} = -0.329z_{\mathrm{RT3U},i} + 1.438z_{\mathrm{T4},i} + 0.162z_{\mathrm{T3RIA},i} - 0.233z_{\mathrm{TSH},i} - 0.590z_{\Delta\mathrm{TSH},i}$$

and

$$DS_{2,i} = 0.026z_{\mathrm{RT3U},i} - 0.486z_{\mathrm{T4},i} - 0.617z_{\mathrm{T3RIA},i} - 0.901z_{\mathrm{TSH},i} - 0.598z_{\Delta\mathrm{TSH},i}$$

Instead of the standardized z values, the raw data can be used. For this purpose, the weight coefficients must be converted (see above) in such a way that the final discriminant scores remain unchanged. For $DS_{1,i}$ the converted weight coefficients appear as follows:

$$DS_{1,i} = -\frac{0.329}{13.15}x_{\mathrm{RT3U},i} + \frac{1.438}{46.95}x_{\mathrm{T4},i} + \frac{0.162}{14.19}x_{\mathrm{T3RIA},i} - \frac{0.233}{61.18}x_{\mathrm{TSH},i} - \frac{0.590}{80.69}x_{\Delta\mathrm{TSH},i} + v_6$$

$$v_6 = \frac{0.329 \times 109.60}{13.15} - \frac{1.438 \times 98.09}{46.95} - \frac{0.162 \times 20.50}{14.19} + \frac{0.233 \times 28.80}{61.18} + \frac{0.590 \times 41.94}{80.69}$$

TABLE 2

Mean values and standard deviations

	Mean values				Standard deviations			
	HYPO	EU	HYPER	Total	HYPO	EU	HYPER	Total
RT3U	121.70	110.51	95.29	109.60	11.06	8.10	18.76	13.15
T4	36.00	91.99	177.46	98.09	17.56	20.43	41.61	46.95
T3RIA	10.63	17.31	42.63	20.50	5.56	4.75	22.54	14.19
TSH	129.20	13.17	9.74	28.80	123.87	4.98	4.00	61.18
ΔTSH	175.33	25.09	−0.20	41.94	155.06	19.63	2.70	80.69

$$DS_{1,i} = -0.250 \times 10^{-1} x_{RT3U,i} + 0.306 \times 10^{-1} x_{T4,i} + 0.114 \times 10^{-1} x_{T3RIA,i} - 0.382 \times 10^{-2} x_{TSH,i} - 0.731 \times 10^{-2} x_{\Delta TSH,i} - 0.767 \times 10^{-1}$$

In the same way $DS_{2,i}$ becomes

$$DS_{2,i} = 0.198 \times 10^{-2} x_{RT3U,i} - 0.103 \times 10^{-1} x_{T4,i} - 0.435 \times 10^{-1} x_{T3RIA,i} - 0.147 \times 10^{-1} x_{TSH,i} - 0.741 \times 10^{-2} x_{\Delta TSH,i} + 2.424$$

Furthermore, it is essential to know whether both or only one of these discriminant functions are statistically significant. Therefore, as a first approximation, the χ^2-statistic (Barlett's V) is computed by taking in account both discriminant functions:

$$V = [N - 1 (R + K)/2] \ln(1 + \lambda_1) (1 + \lambda_2)$$
$$= [215 - 1 - (5 + 3)/2] \ln(1 + 3.85) (1 + 0.73)$$
$$= 210 \ln 8.39 = 446.69$$
$$d.f. = R (K - 1) = 5 (3 - 1) = 10$$

For significance level $\alpha = 0.001$, χ^2 equals 29.59. Thus it can be seen that the V value far exceeds the significance level 0.001 of the χ^2-distribution with 10 degrees of freedom which is 29.59. Therefore the hypothesis is accepted that the differentiation among the groups (EU, HYPER, HYPO) on the basis of both discriminant functions is significant and not due to chance or sampling errors.

A subsequent step is to establish whether the differentiation among the groups is due to both or only one discriminant function (representing the highest eigenvalue λ_1). For this Barlett's V is computed and tested after removal of function f_1. The approximate χ^2-statistic which takes into account the second discriminant function f_2 (i.e. V after deletion of the first one) is computed as follows:

$$V - V_1 = 446.69 - [N - 1 - (R + K)/2] \ln(1 + \lambda_1)$$
$$= 446.69 - [215 - 1 - (5 + 3)/2] \ln(1 + 3.85)$$
$$= 446.69 - 210 \ln 4.48 = 115.10$$
$$d.f. = R (K - 1) - (R + K - 2s) = 5 (3 - 1) - (5 + 3 - 2) = 4$$

For significance level $\alpha = 0.001$, χ^2 equals 18.47. Since $V - V_1$ far exceeds the significance level 0.001 of the χ^2-distribution with 4 degrees of freedom which is 18.47, the hypothesis is accepted that both discriminant functions contribute significantly to the differentiation among the 3 thyroid functional states. Consequently both functions are retained for further analysis.

Classification

Figure 4 shows a map of the individuals of the three groups and the corresponding group centroids in the 2-dimensional discriminant space

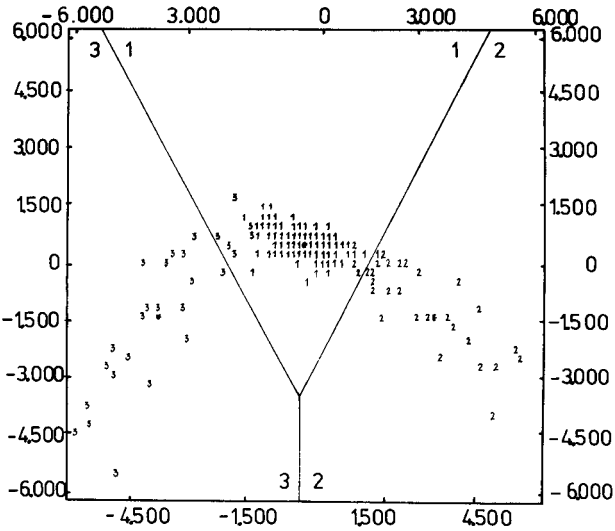


Fig. 4. Plot and territorial map of discriminant score 1 (horizontal) versus discriminant score 2 (vertical) for the HYPO/EU/HYPER thyroidy discrimination. (*) Group centroid; (1) EU individual; (2) HYPER; (3) HYPO.

together with a territorial diagram of each group. It is a constitution of two diagrams obtained by the SPSS programs [8]. The territorial diagram contains linear boundaries drawn orthogonally on half the distance between each pair of group centroids. In this way three regions can be observed, one for each thyroid functional state. Graphically, the classification rule can be formulated as follows: assign patient i to the EU group if he falls into region 1 or on the boundary 1/2; or to the HYPER group if he falls into region 2 or on the boundary 2/3; or to the HYPO group if he falls into region 3 or in the boundary 1/3.

Mathematically, this classification rule can be converted to the following rule: assign patient i to EU if $A_i \leqslant B_i$ and $A_i < C_i$; HYPER if $B_i \leqslant C_i$ and $B_i < A_i$; HYPO if $C_i \leqslant A_i$ and $C_i < B_i$

when

$$A_i = [(DS_{1,i} - \overline{DS}_{1,\mathrm{EU}})^2 + (DS_{2,i} - \overline{DS}_{2,\mathrm{EU}})^2]^{\frac{1}{2}}$$

$$B_i = [(DS_{1,i} - \overline{DS}_{1,\mathrm{HYPER}})^2 + (DS_{2,i} - \overline{DS}_{2,\mathrm{HYPER}})^2]^{\frac{1}{2}}$$

$$C_i = [(DS_{1,i} - \overline{DS}_{1,\mathrm{HYPO}})^2 + (DS_{2,i} - \overline{DS}_{2,\mathrm{EU}})^2]^{\frac{1}{2}}$$

This classification corresponds to the Mahalonobis distance classifier discussed above. Table 3 gives an example for individual 200 of the thyroid data set.

The previous classification procedure was also applied to all 215 patients, and the number of correctly classified individuals was determined. Written as a percentage, this expresses the efficiency (prediction) of the classification

TABLE 3

Allocation of a patient by the Mahalonobis distance classifier

	DS_1	DS_2	
Centroid of			
EU	−0.637	0.559	$A_{200} = [(-3.162 + 0.637)^2 + (0.250 - 0.559)^2]^{\frac{1}{2}} = 2.544$
HYPER	3.424	−1.218	$B_{200} = [(-3.162 - 3.424)^2 + (0.250 + 1.218)^2]^{\frac{1}{2}} = 6.748$
HYPO	−3.676	−1.372	$C_{200} = [(-3.162 + 3.676)^2 + (0.250 + 1.372)^2]^{\frac{1}{2}} = 1.702$
Patient 200	−3.162	0.250	

(a priori known to be HYPO)

Classification: patient 200 is assigned to the region of HYPO thyroidy: $A_{200} < B_{200}$ and $A_{200} < C_{200}$

TABLE 4

Prediction results for the EU—HYPER—HYPO thyroidy classification of 215 patients

A priori group membership	N of cases	A posteriori (predicted) group membership		
		EU	HYPER	HYPO
EU	150	149 (69.3%)	1 (0.5%)	0 (0.0%)
HYPER	35	5 (2.3%)	30 (14.0%)	0 (0.0%)
HYPO	30	6 (2.8%)	0 (0.0%)	24 (11.1%)

TABLE 5

Allocation of a patient on the basis of Bayes estimation (eqns. 5 and A.22)[a]

	$c_{n,\mathrm{EU}}$	$c_{n,\mathrm{HYPER}}$	$c_{n,\mathrm{HYPO}}$	$x_{n,200}$[b]	$c_{n,\mathrm{EU}} x_{n,15}$	$c_{n,\mathrm{HYPER}} x_{n,15}$	$c_{n,\mathrm{HYPO}}$
RT3U	1.027	0.936	1.114	131.0	134.537	122.616	145.934
T4	0.112	0.237	0.215×10^{-1}	27.0	3.024	6.399	0.581
T3RIA	0.373	0.490	0.415	8.0	2.984	3.920	3.320
TSH	0.743×10^{-2}	0.203×10^{-1}	0.497×10^{-1}	99.0	0.736	2.010	4.920
ΔTSH	-0.192×10^{-3}	-1.125×10^{-1}	0.405×10^{-1}	47.0	−0.009	−5.529	1.904
c_0	−65.204	−76.224	−77.147		−65.204	−76.224	−77.147
					76.068	53.192	79.512
					$= g_{\mathrm{EU},200}$	$= g_{\mathrm{HYPER},200}$	$= g_{\mathrm{HYPO}}$

[a]The chance that patient 200 belongs to EU is:
$P(\mathrm{EU}/\mathbf{x}_{200}) = e^{76.068}/(e^{76.068} + e^{53.192} + e^{79.512}) = 0.031$

The chance that patient 200 belongs to HYPER is:
$P(\mathrm{HYPER}/\mathbf{x}_{200}) = e^{53.192}/(e^{76.068} + e^{53.192} + e^{79.512}) \cong 0$

The chance that patient 200 belongs to HYPO is:
$P(\mathrm{HYPO}/\mathbf{x}_{200}) = e^{79.512}/(e^{76.068} + e^{53.192} + e^{79.512}) = 0.969$

On the basis of previous results patient 200 may be assigned to HYPO if a certainty of 97% is acceptable.

[b]$x_{n,200}$ is the measurement of variable n for patient 200.

procedure. The classification results are given in Table 4. It must be noted that when this method is used, the efficiency is somewhat too optimistic (94.4% of the 215 patients correctly classified). More realistic procedures are discussed under *Classification* in the main text. However, the same results were obtained when the classification was based on Bayes estimation (see eqn. 5 and A.22). A computational example for patient 200 is given in Table 5.

REFERENCES

1 J. Smeyers-Verbeke, D. L. Massart and D. Coomans, J. Assoc. Off. Anal. Chem., 60 (1977) 1382.
2 D. Coomans, I. Broeckaert, M. Jonckheer, P. Blockx and D. L. Massart, Anal. Chim. Acta, 103 (1978) 409.
3 P. S. Shoenfeld and J. R. de Voe, Anal. Chem., 48 (1976) 403R.
4 H. C. Andrews, Introduction to Mathematical Techniques in Pattern Recognition, Wiley—Interscience, New York, 1972.
5 R. E. Frank, W. F. Massy and D. G. Morrison, J. Market Res., 2 (1965) 250.
6 D. G. Morrison, J. Market Res., 6 (1969) 156.
7 B. R. Kowalski and C. F. Bender, Anal: Chem., 44 (1972) 1405.
8 N. H. Nie, C. H. Hull, J. G. Jenkins, K. Steinbrenner and D. Bent, Statistical Package for the Social Sciences (SPSS), McGraw-Hill, New York, 1975.
9 A. Eskes, P. Dupuis, A. Dijkstra, H. de Clercq and D. L. Massart, Anal. Chem., 47 (1975) 2168.
10 W. W. Cooley and P. R. Lohnes, Multivariate Data Analysis, Wiley, New York, 1971.
11 M. M. Tatsuoka, Multivariate Analysis: Techniques for Educational and Psychological Research, Wiley, New York, 1971.
12 D. F. Morrison, Multivariate Statistical Methods, McGraw-Hill, New York, 1976.
13 J. van Ryzin, Classification and Clustering, Academic Press, New York, 1977.
14 R. Harris, A Primer of Multivariate Statistics, Academic Press, New York, 1975.
15 H. Seal, Multivariate Statistical Analysis for Biologists, Methuen, London, 1964.
16 M. G. Kendall and A. Stuart, The Advanced Theory of Statistics, Vol. III, Hafner, New York, 1946.
17 P. Dagnelie, Analyse Statistique à Plusieurs Variables, Les presses agronomiques de Gembloux, Brussels, 1975.
18 B. Noble, Applied Linear Algebra, Prentice-Hall, Englewood Cliffs, New Jersey, 1969.
19 J. M. Romeder, Méthodes et programmes d'analyse discriminante, Dunod, Paris, 1973.
20. P. A. Lachenbuch, Discriminant Analysis, Hafner, New York, 1975.
21 R. O. Duda and P. E. Hart, Pattern Classification and Scene Analysis, Wiley—Interscience, New York, 1973.
22 E. A. Patrick, Fundamentals of Pattern Recognition, Prentice-Hall, Englewood Cliffs, New Jersey, 1972.
23 K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, New York, 1972.
24 W. S. Meisel, Computer-oriented Approaches to Pattern Recognition, Academic Press, New York, 1972.
25 T. Y. Young and T. W. Calvert, Classification, Estimation and Pattern Recognition, Elsevier, New York, 1974.
26 D. L. Massart, A. Dijkstra and L. Kaufman, Evaluation and Optimization of Laboratory Methods and Analytical Procedures, Elsevier, Amsterdam, 1978.
27 B. R. Kowalski and C. F. Bender, J. Am. Chem. Soc., 95 (1973) 686.
28 B. R. Kowalski and C. F. Bender, J. Am. Chem. Soc., 94 (1972) 5632.

29 B. R. Kowalski, Anal. Chem., 47 (1975) 1152A.
30 Pratt, Moore, Parsons and J. B. Anderson, Research/Development, (1978) 52.
31 J. C. MacDonald and R. A. Levine, Pattern recognition in laboratory medicine, in S. Subtrata (Ed.) Proceedings of the Fourth New England Bioengineering Conference, Pergamon, New York, 1976.
32 C. Albano, W. Dunn, U. Edlund, E. Johansson, B. Norden, M. Sjöström and S. Wold, Anal. Chim. Acta, 103 (1978) 429.
33 W. J. Dixon, Biomedical computer programs (BMDO7M), University of California Press, Berkeley, 1971.
34 D. L. Duewer, J. R. Koskinen and B. R. Kowalski, ARTHUR, available from B. R. Kowalski, Laboratory for Chemometrics, Department of Chemistry BG-10, University of Washington, Seattle, Washington 98195.
35 B. R. Kowalski, Chemometrics, theory and applications, ACS Symposium Service 52, American Chemical Society, 1977.
36 F. K. Kawahara and Y. Y. Yang, Anal. Chem., 48 (1976) 651.
37 F. K. Kawahara, J. F. Santner and E. C. Julian, Anal. Chem., 46 (1974) 266.
38 E. C. Julian, Application of Discriminant Function Analysis to Infrared Data, Newsletter, No. 14, p. 8, July 1972. (Analytical Quality Control Laboratory, NERC, Environmental Protection Agency, 1014 Broadway, Cincinnatti, Ohio 45202.)
39 J. F. Santner, Status of Research Report; memorandum to W. J. Benoit, EPA, Cincinnatti, 1972.
40 J. S. Mattson, C. S. Mattson, M. J. Spencer and F. W. Spencer, Anal. Chem., 49 (1977) 500.
41 J. J. Powers and E. S. Keith, J. Food Sci., 33 (1968) 207.
42 N. Hartmann and S. J. Hawkes, J. Chromatogr. Sci., 8 (1970) 610.
43 B. G. M. Vandeginste and P. van Iersel, Paper presented at the 4th SAC Conference on Analytical Chemistry, Birmingham, 1977.
44 H. E. Solberg, Scand. J. Clin. Lab. Invest., 35 (1975) 705.
45 F. de Waard, Folia Med. Neerl., 15 (1972) 29.
46 J. S. Amenta and M. L. Harkins, Am. J. Clin. Path., 55 (1971) 330.
47 P. Fraser, M. Healy, N. Rose and L. Watson, Lancet, (1971) 1314.
48 M. Werner, S. M. Brooks and G. Cohnen, Clin. Chem., 18 (1972) 116.
49 D. N. Baron, Ann. Clin. Biochem., 7 (1970) 100.
50 K. Ramsoë, N. Tygstrup and P. Winkel, Scand. J. Clin. Lab. Invest., 26 (1970) 307.
51 P. Winkel, K. Ramsoë, J. Lyngbye and N. Tygstrup, Clin. Chem., 21 (1975) 71.
52 H. E. Solberg, S. Skrede and J. P. Blomhoff, Scand. J. Clin. Lab. Invest., 35 (1975) 713.
53 H. E. Solberg, S. Skrede, K. Elgjo, J. P. Blomhoff and E. Gjone, Scand. J. Clin. Lab. Invest., 36 (1976) 81.
54 E. Carlström, Y. Edlung and H. A. Hansen, Scand. J. Clin. Lab. Invest., Suppl. (1963) 3.
55 D. B. Barnett, A. A. Greenfield, P. J. Howlett, J. C. Hudson and R. M. Smith, Br. Med. J., 2 (1973) 144.
56 L. W. Mayron, E. Kaplon, S. Alling and J. Bectel, Clin. Chem., 20 (1974) 172.
57 P. Wilding, M. J. Kendall, R. Holder, J. A. Grimes and M. Farr, Clin. Chim. Acta, 64 (1975) 185.
58 R. D. Bulbrook, F. C. Greenwood, J. L. Hayward and C. C. Spicer, Lancet, (1960) 1154.
59 P. Winkel, A. A. Afifi, L. D. Cady, M. H. Weil and H. Shubin, J. Chronic. Dis., 24 (1971) 61.
60 J. M. Chapman, A. H. Coulson, A. C. Virginia and E. R. Borun, J. Chronic Dis., 23 (1971) 631.
61 S. Wold, Technical Report no. 357, Department of Statistics, University of Wisconsin, 1974; Pattern Recognition, 8 (1976) 127.
62 S. Wold and M. Sjöström, in B. R. Kowalski (Ed.), Chemometrics: Theory and Application, A.C.S. Symposium series no. 52, 1978, p. 243.
63 M. Kendall, Multivariate Analysis, Charles Griffin and Co., London, 1975.

Analytica Chimica Acta, 112 (1979) 123—132
Computer Techniques and Optimization

A FULLY AUTOMATED MASS SPECTROMETER FOR THE ANALYSIS OF ORGANIC SOLIDS

HEINRICH HILLIG, HENDRIK KÜPER* and WOLFGANG RIEPE

Institut für Spektrochemie und angewandte Spektroskopie (ISAS), Postfach 778, D-4600 Dortmund 1 (Federal Republic of Germany)

HANS PETER RITTER

Bayer AG, D-5090 Leverkusen (Federal Republic of Germany)

(Received 1st December 1978)

SUMMARY

Automation of a mass spectrometer—computer system makes it possible to process up to 30 samples without attention after sample loading. An automatic sample changer introduces the samples successively into the ion source by means of a direct inlet probe. A process control unit determines the operation sequence. Computer programs are available for the hardware support, system supervision and evaluation of the spectrometer signals. The most essential precondition for automation — automatic evaporation of the sample material by electronic control of the total ion current — is confirmed to be satisfactory. The system operates routinely overnight in an industrial laboratory, so that day work can be devoted to difficult analytical problems. The cost of routine analyses is halved.

During the past decade, computer techniques have been introduced into practically all fields of instrumental analytical chemistry. In particular, organic mass spectrometry (m.s.) makes use of computer support extensively and successfully. The advantages are faster handling of spectral data, higher efficiency of the spectrometer, greater reliability of the results and the possibility of extracting additional information from the spectra. Of course, m.s.—computer systems are comparatively expensive, and where economic aspects have to be considered, e.g. in industrial laboratories, it is important to utilize the instrumentation outside normal working hours by automatic operation.

Several attempts have already been made to create systems for processing series of samples without human support. Byrd [1] for example, described a microcomputer-based data acquisition and control system that allows unattended mass spectral analyses of 24 gas samples. Brunnée et al. [2] developed a computerized mass spectrometer for automatic measurements of isotopic abundance ratios in inorganic samples.

The direct insertion technique is widely used to analyze organic solids routinely, and the aim of the work described here was to develop a self-contained system suitable for overnight operation. Thus laboratory personnel

would be relieved from routine work and could give more attention to difficult problems requiring individual treatment. In principle, there were two possibilities: either development of a completely new mass spectrometer–computer system, or extension and modification of existing equipment. Realization of an entirely new system was ruled out not only because of time and cost, but because the system had to allow for occasional manual operation in dealing with unusual problems. Therefore, it was decided to automate an existing commercial system comprising a Varian CH7 single focussing mass spectrometer with a conventional direct insertion unit and a Varian SS100MS data acquisition and processing system which has a disc with a storage capacity of 2.3 million words. The final configuration (Fig. 1) includes a sample changer designed and constructed for this particular purpose and a process controller.

APPARATUS

Sample changer

The most essential condition for unattended processing of solid samples is controlled evaporation of the material. This is difficult to achieve by applying fixed temperature programs to the probe furnace, because such programs usually lead to signals which change significantly with time; it is nearly impossible to preselect a single temperature program suitable for a whole series of different samples. Optimum evaporation, however, can be obtained by a technique developed several years ago [3], in which the total ion current is held at a constant preselected value by an electronic feed-back control of the heating power supply (Fig. 2); no manual adjustment is required. This technique makes high demands on the probe — low heat capacity, effective cooling and a close contact between sample container and furnace — but these requirements are met by the probe in the Varian CH7 spectrometer, which was therefore adopted as an unmodified part of the sample changer.

The widely used gold crucibles are replaced by aluminium crucibles (Fig. 3) which are used only once. After introduction of the sample, each crucible is closed with a cap which has a central bore of 0.3-mm diameter. The crucible thus represents a kind of Knudsen cell which allows fractional evaporation of mixtures [4]. The caps have an annular groove for removal of the crucible from the probe, and the crucibles have tapered bottoms to ensure that they

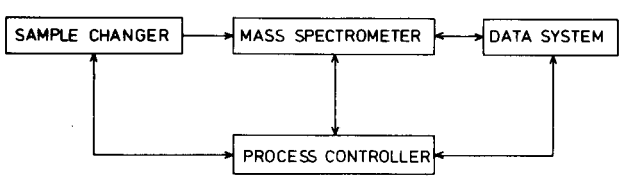


Fig. 1. Configuration of the automatic mass spectrometer. The arrows indicate the direction of control/data paths.

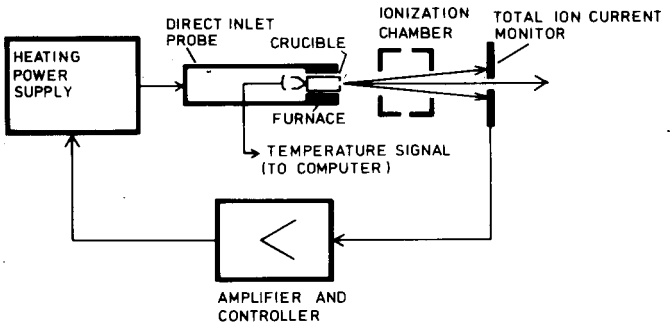


Fig. 2. Automatic total ion current control of the direct inlet system.

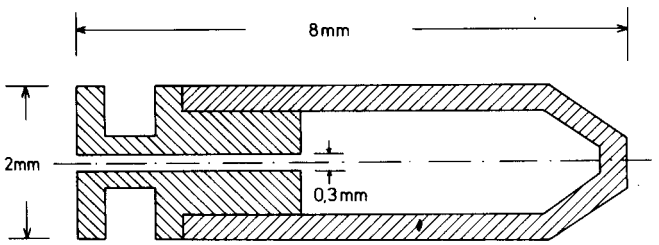


Fig. 3. Aluminium sample crucible.

are reliably picked up by the probe. Up to thirty of these crucibles are stored in a magazine from which they are successively transferred to the ion source. The fully loaded magazine can be dealt within ca. 8 h, provided that each crucible contains a ca. 10^{-6}-g sample that is evaporated completely. The operation cycle repeated for each sample can be roughly subdivided as shown in Fig. 4.

Particular attention was paid to the problem of inserting the crucibles into the probe furnace and removing them after withdrawal of the probe from the source. These steps must be performed with high mechanical precision to prevent any damage to the probe or other parts. All movements of the sample changer are therefore controlled by stepper motors. Figure 5 shows the steps in manipulating the crucible. The magazine is an oblong metal plate with thirty holes in a line. The crucibles are located in these holes as indicated in Fig. 5. After movement of the magazine has aligned a crucible with the probe furnace (1), the probe slips over the crucible (2), which is removed from the magazine because it is held by a spring projecting into the furnace (3). The magazine is then removed from the path of the probe to the source. After withdrawal of the probe (4), the crucible is removed from the furnace (5, 6) by means of a two-pronged fork which fits the groove of the crucible cap. The cap must obviously fit firmly in the crucible.

The functioning of the sample changer is illustrated in Fig. 6 (top views). Each of the moving parts — magazine, lock carriage and probe carriage — has

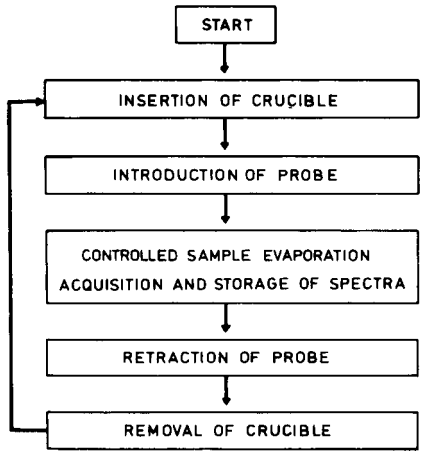


Fig. 4. Operation cycle.

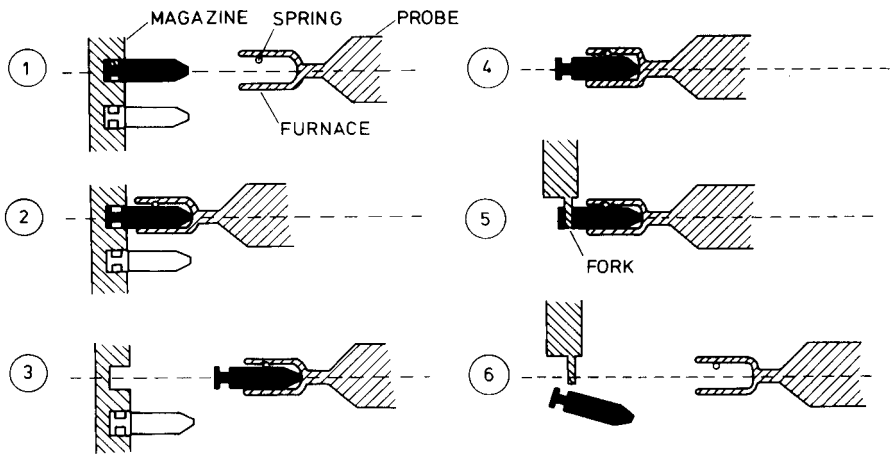


Fig. 5. Insertion (1—3) and removal (4—6) of the sample crucible.

its own stepper motor. The manually operated lock valve is replaced by an electrically activated ball valve. The magazine moves vertically into position on two guiding rods; the lock and probe carriages are guided parallel to the direct inlet path by four rods. Figure 6 (1) shows the changer in its initial position, with the sample magazine raised between the probe and inlet port. To pick up the crucible, the lock carriage moves against the magazine and backward. The probe follows this movement because its stepper motor is not energized. After the magazine has been removed (Fig. 6 (2)), the lock carriage closes the lock volume by moving against an O-ring seal. When the lock has been pumped down, the ball valve is opened and the probe carriage is moved by its stepper motor so that the probe is introduced into the ion source (Fig. 6 (3)).

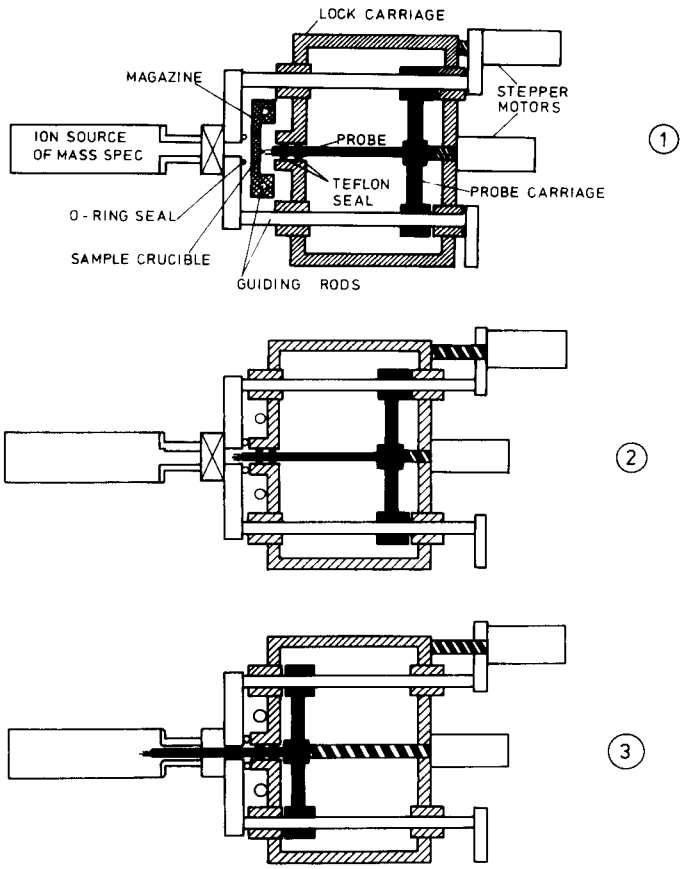


Fig. 6. Sample changer. (1) Starting position before insertion of a crucible; (2) evacuation of the lock volume; (3) probe inserted into the ion source.

After the sample has evaporated and the spectra have been acquired and stored, the probe is withdrawn by a backward movement of its carriage. The ball valve is then closed and the lock volume vented. After that, the lock carriage moves to its original position, thus allowing the magazine to be raised and the attached fork positioned for removal of the crucible.

Modifications of the m.s.—computer system

A mass spectrometer — like any other analytical instrument — will provide reliable results only if the entire procedure is standardized. For automatic operation, all the parameters that influence a spectrum must be controlled. Modern mass spectrometers offering great sensitivity suffer not infrequently from a lack of proportionality with respect to the ion source pressure. Severe variations in the source pressure may even lead to deviations of the mass scale which cannot always be immediately recognized if a data acquisition

system is used. This problem is very likely to arise if samples are evaporated by means of a direct insertion probe because the vapor pressure depends exponentially on the temperature. Achieving a fairly stable ion source pressure can often be arduous.

The signal-to-noise ratio of the spectrometer is another parameter that affects the information content of a spectrum. Its upper limit is set either by the amplitude response of the analog channel or by the above-mentioned lack of proportionality with respect to the source pressure. In order to minimize peak errors caused by noise, an automated system must utilize the admissible amplitude range fully, regardless of the sample.

Both problems are largely overcome by the technique described for controlled sample evaporation. Both the total ion current and the source pressure are kept nearly constant. Moreover, the sample material is evaporated very gently because at all times the temperature can have only the value actually required for an adequate ion current. In general, the technique also provides an optimum signal-to-noise ratio. However, it cannot prevent the analog channel of the spectral signals from becoming overloaded occasionally if the spectra of the samples or of their components contain very different numbers of peaks. Furthermore, the technique, of course, cannot compensate for a decrease in the amplification of the secondary electron multiplier. For that reason, it was desirable to have an expanded dynamic range of the analog input of the computer interface. This was justified because the effective 14 bits of the analogue-to-digital conversion system (10-bit ADC and 4-bit auto-ranging amplifier) do not fully utilize the dynamic range of about 10^5 of the spectrometer. A modification of the auto-ranging electronics would have required a computer word length greater than the available 16 bits and a new peak recognition software. Therefore, a computer-controlled amplifier/ attenuator was inserted in the analog channel. Figure 7 shows the basic circuit diagram. Table 1 gives the different amplifications available depending on the state of the switches $S_1 \cdots S_5$ (actually field effect transistors) which are operated by a digital-output module of the computer. The choice of 10 amplification gains ranging from 0.33 to 3.33 and each differing from the adjacent ones by a constant factor of $10^{1/9} = 1.29$ proved to be quite useful.

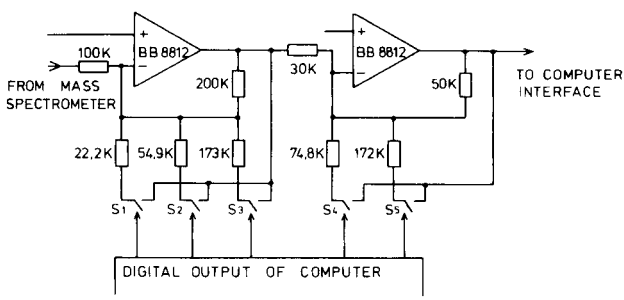


Fig. 7. Basic circuit diagram of the computer-controlled amplifier/ attenuator.

TABLE 1

Amplifications (or attenuations) V according to the positions of switches $S_1 \cdots S_5$ (see Fig. 7; x = switch closed)

S_1	S_2	S_3	S_4	S_5	V
x					0.33
	x		x		0.43
	x			x	0.56
	x				0.72
		x	x		0.93
		x		x	1.20
		x			1.55
					2.00
				x	2.58
					3.33

The proper amplification gain is selected according to the base peak height of the preceding spectrum. It may therefore be necessary to reject a spectrum if the base peak exceeds the maximum height, but the procedure for spectra acquisition ensures that sufficient spectra are measured.

Resolution, mass scale and background of the spectrometer are operating conditions which may severely affect the spectra and therefore have to be checked continuously when the system runs automatically. Otherwise, many spectra would be useless, sample material would be wasted and the system could become futile. A computer program was developed which determines the resolution by inspecting all peaks with heights greater than 10% of that of the base peak. It utilizes the fact that during the acquisition of a spectrum with an exponentially increasing magnetic field, a constant number of signal samples is taken from each peak. The reciprocal of this number measures the resolution directly. Deviations of the mass scale can be caused by changes of the ion source potentials resulting from deposits of sample material. Therefore, the mass scale is tested after each spectrum, and if the given tolerance interval of ±2/16 a.m.u. is exceeded a mass correction is carried out [5].

A background test, carried out just before insertion of each sample, examines all peaks within the preselected mass range to check that their heights remain under 0.6% of the maximum possible value except for those of masses 18(H_2O), 28(N_2) and 32(O_2), for which 3% can be tolerated. The test is repeated as long as the result is negative.

Process controller

To utilize fully the capabilities of the data system, process control is carried out by a separate system which not only guides all activities of the sample changer but also takes care of the proper interaction between sample changer, spectrometer and data system. Moreover, the controller performs several steps not mentioned above, e.g. for safety purposes which may be very im-

TABLE 2

Portion of the process control instruction set

210	WAIT 30 SECONDS	PUMPING TIME
220	START SCAN CYCLE	
230	SWITCH ON PROBE FURNACE POWER SUPPLY	COLLECTION OF SPECTRA
240	PROBE FURNACE AT LIMITING TEMPERATURE NO → 220 YES → 250	COMPLETE SAMPLE EVAPORATION
250	STOP SCAN CYCLE	
260	WAIT 2 MIN	BAKING OF PROBE
270	SWITCH OFF PROBE FURNACE POWER SUPPLY	
280	WAIT 2 MIN	COOLING OF PROBE
290	CHART RECORDER OFF	
300	MOVE PROBE INTO ZERO POSITION	
310	SHUT MAIN VALVE	

portant during initial tests. Although the device used here (ENC Type SLO SYN Numerical Control) is obsolescent, it meets all requirements, especially flexibility, and is cheap. The discrete process steps are effected according to coded instructions stored on paper tape and transferred to the controller. A keyboard enables manual input and control. Table 2 shows a small part of the instruction set executed after the probe has been inserted into the source. The system not only runs automatically but can be operated manually. This allows the spectrometer to be employed without the sample magazine, and in case of a breakdown, a normal operation might be maintained or single process steps could be realized for troubleshooting.

Acquisition and processing of spectra

The specific application of the system is the routine control of synthetic products. This involves recognition of the presence of impurities (e.g., initial compounds, additional reaction products or solvents) and, as far as possible, retrieval of their spectra. Electronically controlled fractional evaporation from the crucibles described above had been applied for nearly three years before the system was automated, and proved entirely suitable for these purposes. A series of 50 ··· 100 spectra is collected consecutively from the start of heating the sample until its exhaustion. During this time, the evaporation temperature is recorded, and the slope of the line reveals the presence and type of a mixture as explained previously [4]. As an example, Fig. 8 shows the slopes of both temperature and total ion current curves for a pure material. After the heating power supply has been switched on, the total ion current moves towards the preselected value and is kept constant as long as sample material is still present in the crucible. The temperature is adjusted

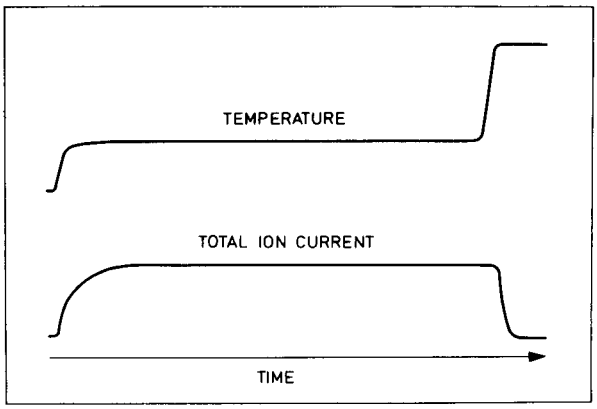


Fig. 8. Slopes of the temperature and total ion current curves.

automatically to the value required to provide a vapor jet stream corresponding to the total ion current. Since a pure sample represents one component in one thermodynamic phase, the temperature remains constant. After exhaustion of the sample, the temperature increases rapidly because of the lack of ion current and then stops at a limiting value.

Inspection of the temperature slopes for a set of automatically processed samples provides a quick survey for characterization of the individual materials, and immediate information about which series of spectra should be examined in detail. In most cases, the fractionating effect of the inlet system suffices to provide component spectra that are already largely free from interference or can be retrieved by simple arithmetic procedures. Where component spectra cannot be isolated completely, there are nearly always groups of characteristic peaks to be found for reliable interpretation.

DISCUSSION

The system has been in routine use for one year. There has been only one break-down caused by a defect in the vacuum system which had no connexion with the spectrometer modifications. The sample throughput has been increased to such an extent that the cost per analysis is half what it was before automation. The flexibility of the system and the possibility of direct access to all functions are regarded as particularly advantageous. There was initially some concern with regard to contamination of the ion source which might become intolerably high with the increased sample throughput; this turned out to be groundless. However, the preselected total ion current should not substantially exceed a value that ensures a sufficient signal-to-noise ratio.

At present, further developments of spectral evaluation are being studied, with the intention of achieving automatic extraction of the component spectra and a library search. A method of disentangling overlapping spectra is also under investigation. The well-known "Mass Max" method [6, 7],

which has been suggested as a solution to this problem, presents certain fundamental difficulties when it is applied to the specific patterns of the mass fragmentograms.

An additional 10-bit ADC already transfers the evaporation temperature signal to the computer for preselection of series of spectra originating from mixtures. The fact that the evaporation temperature is printed out in the interscan report and stored with each spectrum as additional information has also proved useful.

The authors are grateful to the Ministry for Science and Research of Northrhine—Westphalia for the support of this work.

REFERENCES

1 J. S. Byrd, Appl. Spectrosc., 30 (1976) 27.
2 C. Brunnée, G. Kappus, H. Rache, E. U. Seiler and B. Windel, Int. Lab., March/April (1978) 89.
3 J. Franzen, H. Küper, W. Riepe and D. Henneberg, Int. J. Mass Spectrom. Ion Phys., 10 (1972/73) 353.
4 J. Franzen, H. Küper and W. Riepe, Anal. Chem., 46 (1974) 1683.
5 H. Hillig, W. Riepe and J. Kwiatkowski, Org. Mass Spectrum., 9 (1974) 1039.
6 J. E. Biller and K. Biemann, Anal. Lett., 7 (1974) 515.
7 J. E. Biller and K. Biemann, 22nd Annual Conf. on Mass Spectrom. and Allied Topics, San Diego, Calif., (1974) 430.

Analytica Chimica Acta, 112 (1979) 133—141
Computer Techniques and Optimization

OPTIMUM SCALING OF MASS SPECTRA FOR COMPUTER-MATCHING

R. GEOFF DROMEY

Department of Computing Science, The University of Wollongong, P.O. Box 1144, Wollongong, N.S.W. 2500 (Australia)

(Received 22nd January 1979)

SUMMARY

A scaling procedure that minimizes effects caused by mass discrimination and other instrumental distortion in computer-matching of mass spectra is described. It is shown how spectra should be matched only when they have been scaled to be at their minimum "distance" with respect to the similarity index in use for the measurement.

Computer-matching of mass spectra is widely used for identifying unknown compounds [1—6]. Variability in instrumental performance can considerably complicate this task. For example, spectra of the same compound measured on mass spectrometers with different mass discrimination and focussing biases may show appreciable differences in their intensity profiles. This in turn can lead to uncertainties when attempts are made to identify unknown spectra. The problem is particularly relevant when a large library of spectra derived from a number of sources is employed as the standard for computer-matching.

Methods of modifying currently used matching procedures to compensate for instrumental distortion and variability are therefore important. This paper describes a relatively straightforward way of minimizing this problem and so of enhancing confidence in the "degree-of-match" measures obtained. The method recommended relies on an optimum scaling procedure that involves a least-squares model and analysis.

The matching surface and scaling

In preprocessing spectra for comparison one of two standard approaches is generally used. Probably the commonest method is to normalize the spectra being compared so that the most intense ion in each case is set to a constant value (usually 100 or 1000). The other approach is to compute the total ion sum for each spectrum and then adjust the peak intensities so that this sum is always a constant (usually 1 or 100). It might be argued that the second method is preferable to the first because it relies on all peaks for derivation of the normalization factor. At a glance this would seem better than relying on the accuracy of just one peak, the most intense in the spectrum but there is no clear argument to support the idea that total-ion-sum normalization is the better way.

The approach that is proposed here as a viable alternative is based on a somewhat sounder premise. It is suggested that spectra should be compared only when they have been systematically scaled to be at their minimum distance apart as measured by the similarity index employed; as an example the distance-apart or similarity index might be the sum of the squared differences of peak intensities of two spectra.

If the scaling of one spectrum with respect to another is varied over a suitable range, then for some particular value of the scaling parameter(s), the two spectra are at their closest "distance" from each other; i.e., for all smaller and larger scaling parameters, the two spectra are further apart, the matching surface for the two spectra being almost parabolic with a well-defined minimum value. When the distance between two spectra is plotted as a function of the scaling parameter applied to one of the spectra, near parabolas such as those shown in Figs. 1 and 2 are obtained. The matching surface is generally not symmetric about the minimum and so is not strictly parabolic.

The distance indices obtained after largest-peak normalization and total-ion-sum normalization will be arbitrarily located on the matching surface in relation to the minimum distance and optimum scaling. This implies that the amplitude of the distance index taken by either of these methods of scaling will be somewhat arbitrary and inconsistent, because the matching surfaces for each pair of spectra are independent. That is, the "parabola" for the two spectra A and C will be different from the parabola for A and B.

The idea of evaluating the distance index between two spectra only when they have been scaled to be at their minimum distance seems to be a satisfactory way of avoiding inconsistencies in similarity index measurements. The three

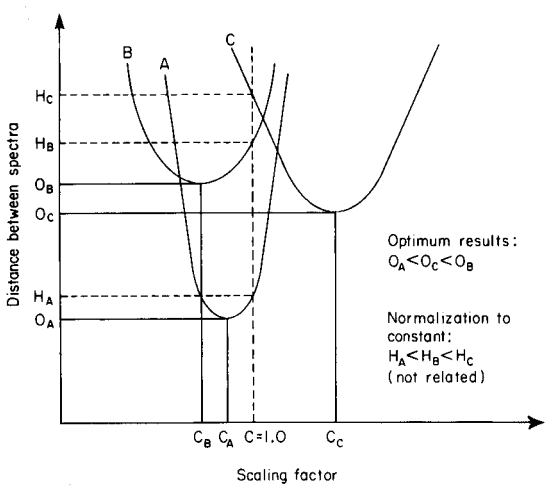


Fig. 1. Three independent "parabolic" matching surfaces indicating the arbitrary relationship among the distance indices H_A, H_B, and H_C evaluated at constant normalization ($C = 1.0$). The similarity indices O_A, O_B and O_C taken at their optimum scaling are directly related.

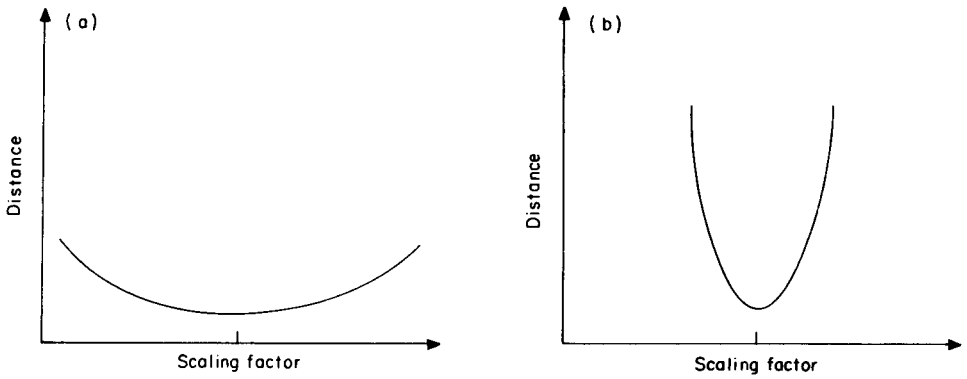


Fig. 2. (a) Matching surface for two disparate spectra. (b) Matching surface for two closely similar spectra.

matching surfaces depicted in Fig. 1 illustrate and emphasize the nature of the scaling problem. Clearly there is an arbitrary relativity for the distance indices H_A, H_B and H_C that have been scaled with respect to the unknown so that their most intense peaks are equal in scale value; i.e., $A_{max} = B_{max} = C_{max}$ = constant (e.g. 100) and the scaling factor $C = 1$. In contrast, the similarity indices O_A, O_B and O_C, taken at their respective optimum scalings, are relative because they are always calculated at their minimum distance.

DERIVATION OF OPTIMUM SCALING CONDITIONS FOR MATCHING SPECTRA

Scaling spectra to be at their minimum distance is usually relatively straightforward. It does however depend on the unit of distance employed. The simplest possible approach to optimum scaling is to find the constant factor by which all peaks in a spectrum must be multiplied to minimize the distance from some other spectrum. In this case, it is easiest to use the mean square distance. In general, the distance between the unknown spectrum U and the reference spectrum R is then given by $D = \Sigma(U_m - R_m)^2$, where the sum is taken over all peaks in both spectra. If c is the scaling factor, then $D = \Sigma(U_m - cR_m)^2$ and c will be optimum in a least-squares sense when $\partial D/\partial c = 0$. Thus

$$\partial D/\partial c = \Sigma\, 2cR_m^2 - \Sigma\, 2U_mR_m = 0 \text{ and } c = \Sigma\, U_mR_m/\Sigma R_m^2$$

Because of the effects of mass discrimination and variations in sample concentration (i.e. in g.c.—m.s.) it is almost always more desirable to work with a mass-dependent scaling factor, so that gradual differences in intensity that are a function of mass can be compensated. The simplest way to introduce this compensation is to use scaling factors that are linearly dependent on mass. A model of this kind should give reasonable compensation against intensity distortions with either increasing or decreasing mass.

Derivation of optimum mass-dependent scaling can be developed in a manner

similar to the method used above. The distance then becomes $D_M = \Sigma[U_m - (c + md)\, R_m]^2$. The distance will be minimal when $\partial D_m/\partial c = 0$ and $\partial D_m/\partial d = 0$. Application of these two conditions gives

$$\sum R_m^2 c + \sum R_m^2 md = \sum U_m R_m \text{ and } \sum R_m^2 mc + \sum R_m^2 m^2 d = \sum U_m R_m m$$

These relationships yield the following optimum values for c and d:

$$c = (\sum U_m R_m - \sum R_m^2 md)/\sum R_m^2$$

$$d = (\sum U_m R_m \sum R_m^2 m - \sum U_m R_m m \sum R_m^2)/(\sum R_m^2 m \sum R_m^2 m - \sum R_m^2 m^2 \sum R_m^2)$$

If the distance between spectra U and R is then calculated by substituting c and d into the relationship given for D_M above, the minimum linear mass-dependent distance between the two spectra can be obtained.

Characteristics of matching profiles

Before the scaling methods derived above are evaluated and compared, it is appropriate to take a closer look at the characteristics of matching surfaces, to gain insight into the relevance of optimum scaling. Two types of matching surface are relevant in this consideration, one where the two spectra being matched are clearly disparate, and the other where the two spectra are quite similar. For clearly disparate spectra, the parabolas are shallow and broad about their minima (Fig. 2a) so that scaling has only a small influence on the magnitude of the distance between spectra. When two spectra are very similar however, the distance index is very sensitive to how they are scaled with respect to one another. This is borne out by the deep and narrow matching surfaces (Fig. 2b) observed for closely similar pairs of spectra. Generally, only the integrity of the distance indices for spectra that are very similar to the unknown is of interest, and so the application of optimum scaling should be considered only in these circumstances, i.e., optimum scaling should only be applied to rank the small subset of spectra that are most similar to the unknown. These considerations are discussed further after the effects of optimum scaling on some examples have been considered.

COMPARISON OF AN UNKNOWN WITH LIBRARY SPECTRA

The two questions "which library spectrum is most like a given unknown?" and "which library spectrum is a given unknown most like?" are quite distinct and separate questions when considered in relation to optimum scaling and optimum matching, though the distinction may seem subtle. When a set of library spectra is compared with an unknown, the reference spectra being scaled in relation to the unknown in this process, then the distance indices are meaningfully related in a relative sense (provided that each comparison has been made at the minimum distance between the spectra). However, in the complementary case when the unknown is scaled to each of the library spectra in turn, the situation is completely different. The distance indices for the

matches of the unknown with the different library spectra are no longer meaningfully related in terms of magnitude, even when the unknown has been optimally scaled in each case. Thus a distance of, say, 100 between spectrum A and spectrum B does not necessarily carry the same weight as a distance 100 between spectrum A and spectrum C. Relativity is lost because the same (unknown) point of reference is not used in each case.

The results in Table 1 emphasize this point concerning the two different approaches to matching spectra and computing distance indices between and among spectra. Table 1 (upper half) shows results for four different methods of scaling the reference spectra with respect to an unknown spectrum. The distance measure is based on the relation $D = \Sigma (U_m - R_m)^2$, as indicated above. The "unknown" spectrum (3-hydroxybenzoic acid methyl ester) and three closely related spectra are shown in Fig. 3. These results show that only when optimum mass-dependent matching is used does reference spectrum 3 come seriously into consideration, i.e. it assumes a distance of 43.27 compared with that of 40.06 for reference spectrum 1. This example proves that the improvements in matching given by optimum mass-dependent scaling can make the difference between a successful match and failure. Also optimum mass-dependent scaling will always tend to cancel out and minimize instrumental intensity distortions. The results for optimum constant scaling will always be better than the two non-optimum methods but in many instances optimum constant scaling is not really as effective as the mass-dependent method.

The results in the bottom half of Table 1 are for the same set of spectra but here the unknown has been scaled to each reference spectrum in turn. There are significant differences in these results compared to the results discussed above. Obviously there can be no change for constant highest intensity normalization. With these results, unlike those discussed above, the distance indices for optimum matching are not properly related in magnitude.

TABLE 1

Comparative results for an "unknown" matched with three similar spectra scaled with respect to the "unknown", and with the "unknown" scaled to each reference spectrum in turn

Reference spectrum	Optimum mass-dependent match	Optimum constant match	TIC equal match	Largest of both spectra (1000)
Each spectrum scaled to the "unknown"				
1	40.06	42.71	47.83	48.20
3	43.27	172.83	174.68	198.72
2	134.83	155.86	181.31	390.44
"Unknown" scaled to each reference spectrum				
1	35.02	36.57	37.56	48.20
3	18.34	198.61	244.77	198.72
2	250.23	230.72	489.97	390.44

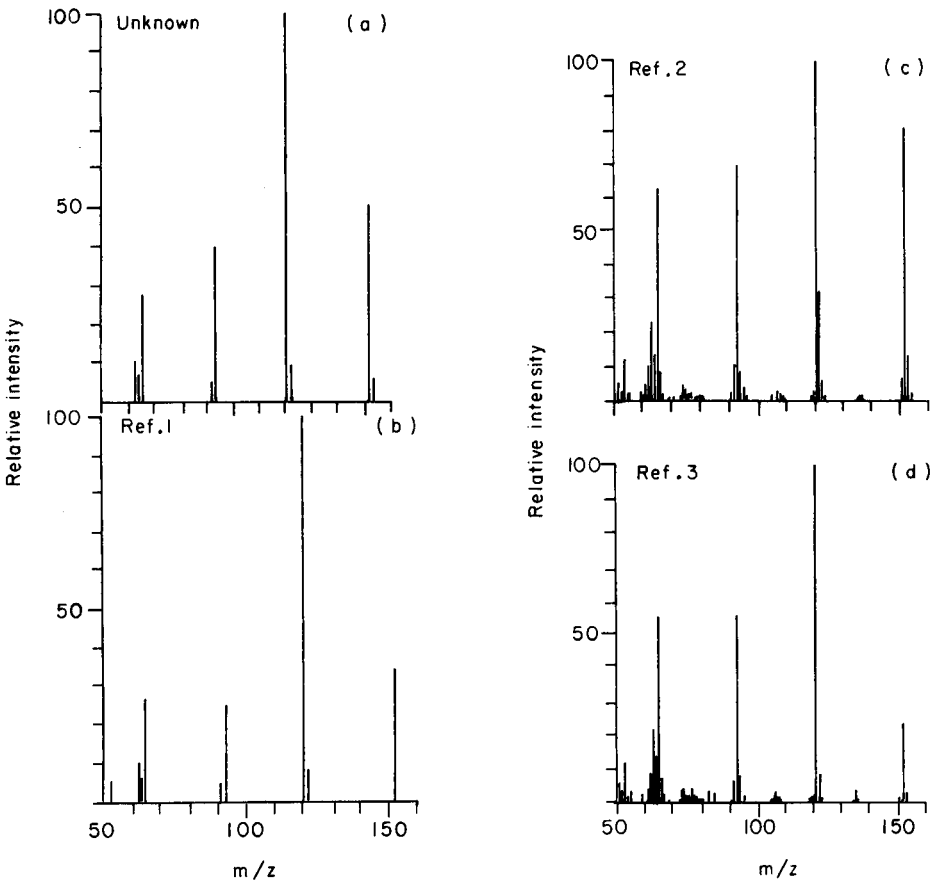


Fig. 3. (a) "Unknown" mass spectrum (3-hydroxybenzoic acid methyl ester) used in matching against a set of reference compounds. (b) Reference spectrum (1) of 4-hydroxybenzoic acid methyl ester. (c) Reference spectrum (2) of 4-hydroxybenzoic acid methyl ester. (d) Reference spectrum (3) of 3-hydroxybenzoic acid methyl ester. All spectra were taken from the EPA/NIH source library.

Comparison of the results in Table 1 clearly illustrates the difference between scaling spectrum A to spectrum B and scaling spectrum B to spectrum A. The magnitude of the differences between the results serves to emphasize the need for careful consideration of relative scaling to obtain reliable performance in computer-matching of mass spectra.

To establish further the usefulness of optimum mass-dependent matching of mass spectra, a set of 27 duplicate spectra from the EPA/NIH library was examined. For this group of spectra, the average percentage reduction in distance between the duplicates was 40.1%. This represents a substantial distance reduction and so underlines the effectiveness of optimum mass-dependent scaling.

EXISTING MATCHING TECHNIQUES AND OPTIMUM SCALING

In the light of the results on optimum scaling, it was considered necessary to examine its relevance to other currently available computer-matching methods. For this purpose, two closely similar spectra were selected, and the effects of several perturbations were studied under conditions of optimum constant scaling. The three matching methods considered were the mean-squares difference method described above (MSD), the absolute difference method (ADIF), and the divergence method (DIV) of Reed and co-workers [5]. The ADIF method involves computing the sum of absolute intensity differences: $D_{\mathrm{ADIF}} = \Sigma |U_m - R_m|$. The divergence method involves computing the distance given by

$$D_{\mathrm{DIV}} = \sum [(U_m - R_m)^2/(U_m + R_m)]$$

For the purposes of this experiment it was decided to make the following three perturbations to the closely similar spectra (Fig. 4): (a) a 10% peak was removed from one spectrum; (b) an 82% peak was changed to a 92% peak; (c) a 33.4% peak was changed to a 43.4% peak. To make an absolute comparison among the matching methods, the minimum distances separating the two spectra (prior to the three perturbations) were calculated and normalized to 100. Only optimum constant scaling was used for the test, because of the

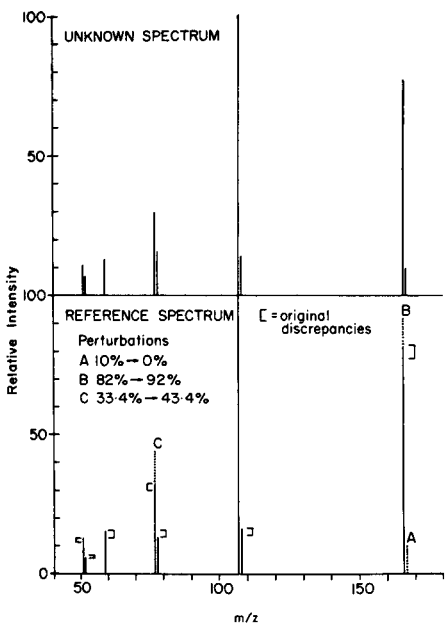


Fig. 4. Two similar spectra used for comparison of matching methods are shown. The original differences between the spectra are bracketed and the three perturbations are marked by dotted lines.

difficulty of obtaining it for the absolute difference method (unlike the mean-squares difference method it does not yield an analytical solution to the scaling problem).

The results obtained for the three perturbations described above are given in Table 2. The most prominent feature is that only the divergence method distinguishes between peak absence and a variation in peak intensity of the same magnitude. For example, the missing 10% peak causes a change in the distance index from 100 to 1210 whereas a change of 10 from 82% to 92% causes a shift from 100 to 164. Clearly, of the three methods, only the divergence distance focusses on the percentage difference in magnitude between peaks. The small discrepancies that do occur for the three perturbations for the MSD and ADIF methods are caused by differences in optimum constant scaling.

The large emphasis that the DIV method places on peak absences (and large percentage differences between peaks at the same mass) is at once an advantage and a disadvantage in that the absence of even very minor peaks can cause considerable changes in the similarity index. It is certainly desirable to weight peak absences more heavily than intensity variations of the same magnitude. However, it would seem more fruitful to reduce this weighting, so that the absence of very minor peaks does not induce significant changes in the similarity index. One way to do this is to reduce the squared term in the numerator of the divergence equation to an absolute intensity difference. In this method (PDIF), percentage differences are independent of intensity and missing peaks are weighted equally. The distance formulated for this method is

$$D_{\mathrm{PDIF}} = \sum [|U_m - R_m|/(U_m + R_m)]$$

This method focusses on the overall profiles of the spectra being compared. Results for this method are included in Table 2. Like the absolute difference method it has the drawback of requiring that optimum scaling be done algorithmically rather than analytically, although scaling by means of the equation given for D_{M} and then evaluating PDIF would be possible.

This peak perturbation study emphasizes that for the divergence method and any other sensitive matching methods, spectra must be optimally scaled. If this precaution is not taken, very large discrepancies in similarity indices

TABLE 2

Response of different matching methods to peak absence and intensity variations. In all cases, the minimum distance before perturbation is 100

Method	10% Peak missing	Peak 82% to 92%	Peak 33.4% to 43.4%
MSD	186.8	187.8	205.9
ADIF	156.8	147.0	156.8
DIV	1210.0	164.0	281.5
PDIF	313.4	107.7	128.6

may be caused by mass discrimination and other distortions. To summarize, the more sensitive the matching method and the closer the spectra are together, the more critical it becomes to apply optimum scaling before similarity indices are evaluated.

Conclusions

The results described above suggest that the method is quite capable of adequately compensating for instrumental intensity distortions. Furthermore, the idea of normalization of spectra for comparison purposes is placed on a sounder formal basis. The scaling method, however, introduces additional computations before the distance indices can be evaluated, and so should be applied only to the small subset of spectra most like the unknown. Methods exist [6—8] for extracting such subsets from a large library, and so the scaling method when used in conjunction with these techniques, does not introduce any serious efficiency problems. The performance of computer-matching methods should be improved significantly by the appropriate consideration of spectrum scaling.

The author thanks Miss Ann Titus for typing the manuscript.

REFERENCES

1 L. R. Crawford and J. D. Morrison, Anal. Chem., 40 (1968) 1464.
2 S. R. Heller, Anal. Chem., 44 (1972) 1951.
3 S. L. Grotch, Anal. Chem., 42 (1970) 1214.
4 S. Abrahamsson, Sci. Tools, 14 (1967) 29.
5 S. Farbman, R. I. Reed, D. H. Robertson and M. E. Silva, Int. J. Mass Spectrom. Ion Phys., 12 (1973) 123.
6 J. T. Clerc and P. R. Nageli, Anal. Chem., 46 (1974) 739A.
7 R. G. Dromey, Anal. Chem., 48 (1976) 1464.
8 R. G. Dromey, Anal. Chem., (1979).

Analytica Chimica Acta, 112 (1979) 143—150
Computer Techniques and Optimization

SEARCH STRATEGY AND DATA COMPRESSION FOR A RETRIEVAL SYSTEM WITH BINARY-CODED MASS SPECTRA

GEERT VAN MARLEN* and JAN H. VAN DEN HENDE

Department of Analytical Chemistry, Delft University of Technology, Jaffalaan 9, Delft (The Netherlands)

(Received 14th December 1978)

SUMMARY

A retrieval system for binary-coded mass spectra is described. The data base used contains 9628 low-resolution mass spectra from the Aldermaston Mass Spectra Data Collection. These spectra are reduced to 106 preselected binary-coded m/z values each. Storage of the compound names and formulae is minimized by using a special set of characters and file organization. The search strategy permits fast generation of the N-nearest neighbours. Depending on the number of best matches generated, an average search requires access to only 24—33% of the spectra contained in the data base. Because of its limited storage requirements, this search system can be used even on microcomputers.

The minicomputer plays an increasingly important role in the functioning of modern laboratories. As a result the available mass spectral data bases have grown to such an extent that their sheer size is becoming a handicap to their in-house applicability for routine mass-spectral retrieval systems. The storage of compound names and empirical formulae for a data base of 10 000 spectra would require at least 1 million bytes and the spectral information a commensurate amount or even more. The use of data compression techniques has therefore become inevitable.

This paper describes the organization of a mass-spectral retrieval system based on the optimized use of storage combined with a feature selection technique. In addition, the problem of reducing search time, also affected by the size of the data base, is addressed.

EXPERIMENTAL

Data base

A library of 9628 low-resolution mass spectra, originating from the Mass Spectra Data Collection [1], was used as a reference file for the retrieval system. These spectra were reduced by binary coding of the intensity values with an intensity threshold of 1% of the base peak. Further reduction was obtained by selecting 106 binary-coded peak positions. The selection, based on the information content of a peak position corrected for the correlation between

peak positions, has been described previously [2]. With this method only those peak positions significant for the entire reference file were coded, requiring a storage capacity of 106 bits per spectrum.

File organization

The data base consists of three files with random access organization. The first file contains, beside the binary-coded spectrum, a unique identification number ID and a presearch parameter, the distance d_R. This distance parameter is defined as the number of peak positions coded "present" in the spectrum or the number of "bit mismatches" between the spectrum and the "empty' spectrum with no peak positions coded present. The file is pre-arranged in order of increasing values of the distances d_R. All reference spectra with the same d_R are combined into a "cluster" of contiguous records in the file, each record containing up to 24 spectra. Figure 1 shows a frequency plot of the number of spectra for all clusters in the file versus the d_R value.

The second file is used to store pointers for each spectrum to the empirical formula and name of the compound, stored in the third file. This method of indirect reference was chosen to eliminate any duplication of empirical formulae and name of compounds in the data base, thereby reducing the total storage requirements. As a result, only 3300 different empirical formulae and 8100 different compound names are stored. The relationship between these files is illustrated in Fig. 2. The storage requirements are summarized in Table 1.

Keywords

To avoid lengthy and variable storage of compound names a special 8-bit "character" set was generated. This character set contains, in addition to the normal numerical and alphanumerical ASCII characters (a total of 64), a set of 160 "keywords" representing character combinations which occur frequently such as ACETYL, METHYL PHENYL, etc. The keywords generated and their frequencies of occurrence in the data base are given in Table 2. With this compression, most of the compound names occupy less than 12 bytes of storage. The names requiring more than 12 characters are split up in chemically signi-

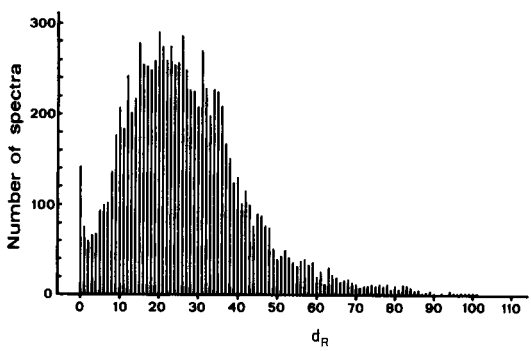


Fig. 1. Distribution of reference spectra clusters.

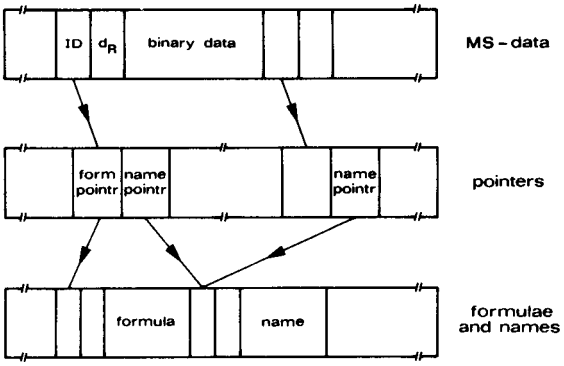


Fig. 2. File organization.

TABLE 1

Storage requirements for 9628 binary-coded mass spectra

File	Contents	No. of entries	Record length (byte)	Storage (Kbyte)
1	Binary data	462	480	223
2	Pointers	9628	4	39
3	Compound names	8106	16	216
	and formulae	3279		

ficant segments, separately stored in the file, and shared by all compound names. These segments are referred to within the 12-byte name-space by means of a 1-byte indicator followed by a 2-byte record number. The segments can also contain one or more references to other segments. The process of reconstructing a compound name is therefore recursive.

Configuration

The retrieval system is based on a PDP11/45 minicomputer, which is used for various laboratory applications, under RSX-11D in a multi-user environment. The data base is stored on a RK05 disk with a capacity of 2.4 Mbytes and an average transfer time to memory of 2 ms per spectrum. According to Table 1, the storage of the data base requires about 20% of the capacity of a disk cartridge. The acquisition of the g.c.—m.s. or m.s. data is carried out on a PDP11/45 preprocessor.

Search program

The retrieval program is written in PDP11 FORTRAN IV-PLUS and requires a storage capacity of 5.5K words exclusive of system functions. The general structure is given in the flow sheet presented in Fig. 3. The most important aspects and modules of the program are described in the following sections.

TABLE 2

Keywords and their frequencies of occurrence in the data base

Char.[a]	Keyword	Freq.	Char.	Keyword	Freq.	Char.	Keyword	Freq.	Char.	Keyword	Fre
140	ACETYL	117	210	ETHYL	863	260	OXY	978	330	OLE	45
141	ANTHRA	60	211	ETHER	159	261	HEPT	40	331	INE	186
142	BENZYL	175	212	ENONE	95	262	STYR	14	332	BUT	248
143	BORANE	43	213	FLUOR	77	263	PHEN	250	333	OCT	17
144	CARBON	279	214	HEXYL	300	264	IMINO	48	334	TRI	688
145	CHLORO	964	215	DESOXY	29	265	IMID	42	335	FUR	90
146	CROTON	31	216	HEPTA	184	266	MERC	56	336	PYR	209
147	DEHYDE	97	217	IDINE	149	267	3,	557	337	7-	229
150	DEUTER	90	220	NOATE	160	270	NATE	55	340	IOD	22
151	EPOXY-	23	221	NITRO	86	271	NITR	75	341	PER	74
152	FLUORO	430	222	ORTHO	17	272	NON	81	342	SIL	52
153	HEPTYL	43	223	OXIDE	52	273	6,	130	343	IDE	103
154	HEXANE	244	224	OCTYL	73	274	HYDRO	527	344	ONE	305
155	METHYL	3053	225	PENTA	469	275	OCTA	142	345	OXO	129
156	NAPHTH	352	226	PHENE	154	276	OATE	26	346	BI	139
157	PENTYL	132	227	SULPH	88	277	PENT	91	347	ACR	47
160	PHENYL	934	230	TETRA	634	300	PROP	392	350	ETH	563
161	PROPIO	129	231	UNDEC	24	301	SPIRO	33	351	OL	386
162	PROPYL	451	232	VINYL	45	302	QUIN	24	252	PI-	140
163	THIOL	119	233	AMYL	36	303	UREA	40	353	6-	354
164	TRANS-	119	234	ANOL	259	304	THIO	212	354	DI	1742
165	AMINE	97	235	OXYL	41	305	8-	138	355	AN	301
166	AMIDE	67	236	ANTH	20	306	THIA	153	356	NE	241
167	AMINO	172	237	ACET	368	307	AZINE	42	357	YL	613
170	ALPHA	244	240	ACID	148	310	SEC-	36	360	IC	88
171	ANTHR	105	241	BENZ	415	311	ISO	275	361	T-	116
172	ANOIC	32	242	BETA	182	312	PHO	120	362	1,	858
173	BENZO	286	243	1,2,3,	160	313	YNE	26	363	ENO	29
174	BUTYL	584	244	IDENE	67	314	ANE	737	364	AL	186
175	BUTYR	102	245	CHOL	43	315	ATE	281	365	EN	204
176	BROMO	278	246	CARB	107	316	5,	204	366	O-	328
177	ALLYL	68	247	CYAN	64	317	IND	126	367	2-	997
200	CYCLO	966	250	CIS-	104	320	ALENE	134	370	3-	814
201	CHRYS	16	251	DECA	278	321	HYDR	270	371	LE	62
202	CHLOR	70	252	DIOL	63	322	4,	321	372	N-	343
203	COHOL	52	253	ERIN	14	323	PH	158	373	4-	552
204	ANONE	122	254	ENYL	184	324	ENE	929	374	2,	569
205	CHROM	39	255	FORM	33	325	AZO	123	375	P-	64
206	DECYL	172	256	GLYC	49	326	BIS	137	376	1-	514
207	EICOS	60	257	HEXA	186	327	HEX	125	377	5-	332

[a]8-bit representation in octal notation.

Input. The preprocessed m.s. or g.c.—m.s. data file or a manually entered mass spectrum is converted to a binary code with the same format as the reference spectra in the data base. For the "unknown" spectrum, the distance parameter d_U and a distance limit value are also calculated.

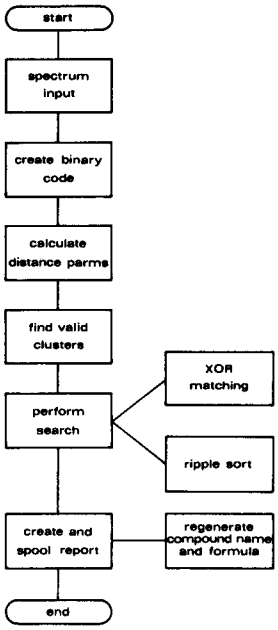


Fig. 3. General structure of the search program.

Matching algorithm. This most frequently required module of the program was written in the PDP11 MACRO assembler. The number of "bit mismatches" between the unknown and reference spectrum is taken as a matching criterion and is calculated with an "exclusive or" (XOR) logic operator. The XOR matching was selected to allow fast comparison of the two spectra. A previous study [4] has indicated that other matching criteria did not give better retrieval results when adopted in the system. To decrease execution time, the comparison is terminated when the calculated distance limit is exceeded. The CPU time needed for one comparison varies from 150 to 800 μs, the execution time of 50—250 instructions. This variation is caused by the location and number of bit mismatches between the two spectra.

Generation of the N*-nearest neighbours.* In order to generate the N-nearest neighbours, a fixed list of results is constantly sorted by a "ripple sort" method. A high sorting speed is obtained by using pointers (tags) as indirect references so as to avoid unnecessary rearrangements of the list.

Search strategy. Only clusters of reference spectra for which the preset distance condition is valid are selected for comparison with the unknown spectrum. The sequence in which these reference spectra are compared is as follows. When the distance parameters d_R and d_U are used for the reference and the unknown spectrum, respectively, the a priori minimum distance D between the two spectra is calculated from $D = |d_R - d_U|$ before the comparison is actually started. The reference spectra are compared with the unknown spectrum in sequence of

increasing of D starting with 0. The search is terminated when D exceeds the number of bit mismatches found for the Nth nearest neighbour generated. The a priori distance between unknown and reference spectrum then exceeds the generated distance for the N-nearest neighbours found so far, and continuation of the search becomes fruitless. This search strategy generates the same N-nearest neighbours compared with a normal sequential search, but matches only those reference spectra which most probably belong to the desired set.

The number of spectra searched

In terms of information theory concepts, the introduction of the presearch parameter d_R implies use of additional data with its own information content, I_H, in addition to the binary-coded peak positions. I_H may be calculated from the formula of Shannon and Weaver [3]:

$$I_H = -\sum_d p_d \log_2 p_d - \log_2 \Delta d \qquad (1)$$

The probabilities p_d of measuring a spectrum at a distance d_R are derived from the data presented in Fig. 1. If a criterion of "perfect matching" is used the distance D between the unknown and the reference spectrum equals 0 in the event of a match and the "search window" Δd equals 1. For the data set under consideration, I_H then amounts to 5.9 bit. Use of this parameter therefore reduces the average number of spectra to be searched to 161 (9628 divided by $2^{5.9}$) assuming that the unknown spectra exhibit the same distribution as the reference spectra.

In a previous paper [4], the distance D between an unknown and a reference spectrum of the same compound was investigated. For a large set of compounds, a skew distance distribution was found with an average distance of 6.9 and a median distance of 3.2. When eqn. (1) was applied with search windows of 13.8 and 6.4, the average and median number of spectra searched before the correct reference spectrum was found, became approximately 2200 and 1050, respectively.

In practice however, the identity of the unknown spectrum is not known beforehand. Generally, the search is terminated after a certain number of best matches, the N-nearest neighbours, have been produced. To estimate the number of spectra to be searched for 3 or 10 nearest neighbours, a set of 459 "unknown" spectra was extracted from the file and used as input to the search system. The average number of spectra that had to be searched was about 2300 and 3200 for these two examples. The influence of this approach to the search strategy on the number of spectra to be searched is illustrated in Table 3 for a typical compound extracted from the file. It is obvious that the saving in search time is worthwhile.

DISCUSSION

It would be expected that for other general data bases, a similar picture would emerge with regard to the storage optimization described. Smaller

TABLE 3

Number of spectra searched and search time for the spectra of phenylacetylene under different conditions (d_U = 21), with a simulation of the report generated for the spectrum of phenylacetylene

No. of spectra searched	*N*-nearest neighbours	Search time (s)
7501[a]	10	13.4
3896[b]	10	8.2
2381[b]	3	5.0

MS–RETRIEVAL DATE: 19-APR-78 TIME: 14:06:09 PAGE: 1

BINARY V08 MODE: MANUAL
TITLE: UNKNOWN
PEAKS IN SPECTRUM: 21
SPECTRA SEARCHED: 2381

ID#	MATCH	BRUTOFORMULA	NAME
180	100	C8.H6	PHENYLACETYLENE
2168	97	C8.H6	PHENYLACETYLENE
4380	95	C7.H5.N	BENZONITRILE

PRESEARCH TIME: 0.4 S
PRINT TIME: 3.3 S
SEARCH TIME: 5.0 S
MATCHING AVERAGE: 2.1 MS

[a]Search with a preset distance limit between unknown and reference spectrum. ($4 < d_R < 38$). [b]Search strategy with a priori distance.

specific data sets, e.g. a set of alkane spectra, will result in the use of fewer keywords. The size of the program and the data base described do not require a large computer system. A microcomputer, e.g. a PDP11/03 combined with a dual floppy-disk drive, would suffice for this type of retrieval. However, because the search time is primarily dependent on the time needed for the transfer of data from disk to memory, the total search time for such a system would become approximately 1 min. For the described system with a transfer time of 2 ms per spectrum, the generation of the 10 nearest neighbours takes about 6 s on the average. The same retrieval system run on the IBM370/158 at the University computer centre requires an average of about 3 s to generate similar results. The file organization and search strategy outlined are generally applicable for other search systems, provided that a parameter can be found to pre-order the data base. This is in agreement with the observations made by Grotch [5]. The effects of the search strategy on other retrieval systems, such as the Biemann search method [6], will be explored further.

The authors are indebted to A. Dijkstra and H. A. van 't Klooster for helpful discussions.

REFERENCES

1 Mass Spectral Data Centre, AWRE, Aldermaston, U.K., library purchased in 1971.
2 G. van Marlen and A. Dijkstra, Anal. Chem., 48 (1976) 595.
3 C. E. Shannon and W. Weaver, The Mathematical Theory of Communication, The University of Illinois Press, Urbana, Ill., 1949.
4 G. van Marlen, A. Dijkstra and H. A. van 't Klooster, Anal. Chem., 51 (1979) 420.
5 S. L. Grotch, 25th Annual Conference on Mass Spectrometry and Allied Topics, Washington, D. C., 1977.
6 H. S. Hertz, R. A. Hites and K. Biemann, Anal. Chem., 43 (1971) 681.

Analytica Chimica Acta, 112 (1979) 151—164
Computer Techniques and Optimization

IDENTIFICATION OF COMPONENTS IN MIXTURES BY A MATHEMATICAL ANALYSIS OF MASS SPECTRAL DATA

G. T. RASMUSSEN, B. A. HOHNE, R. C. WIEBOLDT and T. L. ISENHOUR

Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27514 (U.S.A.)

(Received 23rd February 1979)

SUMMARY

Mathematical techniques for the identification of components in mixtures from the mass spectra of a series of related mixtures are described. The approach is analogous to library search methods in that spectra from a reference collection are compared with a multi-dimensional unknown. Searches are conducted with a library file containing approximately 17000 mass spectra. Results for the analyses of several mixtures are reported, to illustrate the effectiveness of the method.

Of the various pattern recognition techniques available for the identification of organic compounds from mass spectral data, perhaps the most widely used are library search methods. Computerized search programs can identify the mass spectra of pure compounds rapidly and efficiently. However, library search methods can encounter problems when the unknown spectrum pertains to a mixture of compounds. Because more than one compound may contribute to the observed intensities at some mass positions, the relative intensity information may seem to be distorted. Apparently extraneous peaks will be present, and intensities at some mass positions will be computed relative to an incorrect base peak. Various techniques have been applied in efforts to deal with the spectra of mixtures. One response has been to use a reverse search strategy, which considers only information common to both the unknown and reference spectra, while ignoring "extraneous" information in the unknown [1, 2]. Another approach has been to define an algorithm that selects spectra from a reference library and computes concentration coefficients for these spectra in an effort to fit several known spectra to the spectrum of the unknown mixture. Such algorithms, which may be quite sophisticated, rely on least-squares or other criteria to find the best set of concentration coefficients [3, 4]. A recent paper has investigated the effects on such methods of departures from linear behavior in the physical system [5]. The problem of identifying the compounds present in a series of related mixtures has also been examined and is the experimental situation considered in this paper.

Related mixtures are those in which the same components are present, but the concentrations of the components vary from mixture to mixture. Spectra

of related mixtures arise in several common experimental circumstances. The mass spectra acquired during the elution of poorly resolved chromatographic peaks are such mixtures, as are the spectra produced by sampling a reaction mixture at various times during the course of a reaction. The required concentration changes might be caused by the varying rates of diffusion through a molecular leak into a mass spectrometer source for compounds of different molecular weight or by the differing rates of volatilization observed for different compounds during the heating of the direct probe of the spectrometer.

Early efforts at the mathematical analysis of spectra of related mixtures required the identification of peaks unique to each component [6, 7]. If two peaks unique to each component in the set of mixtures could be found, the spectra of the pure components could be extracted from the spectra of the mixtures. More recently, mass chromatograms and mass fragmentograms have been used to locate components in mixtures, and sophisticated background subtraction methods have been used to resolve the component mass spectra [8—12]. These regenerated spectra are then identified by conventional interpretation techniques. Principal component analysis has been applied to mass spectrometric data to determine the number of components in series of related mixtures [13—15]. In attempting to identify the components in a set of related mixtures without requiring assumptions about unique peaks, two general approaches are possible. As with principal component analysis, these techniques rely on the fact that the mass spectra of mixtures can be treated as linear combinations of the component spectra each multiplied by a concentration coefficient. One approach, described by Lawton and Sylvestre [16], permits estimates of pure component spectra to be derived from the principal components of a set of mixture spectra [17]. Although originally used with absorption spectra of dye mixtures, the method, which utilizes non-negativity constraints for spectral intensities and concentration coefficients, is also applicable to mass spectral mixtures. The other approach, which is examined in this paper, is analogous to library search techniques and, therefore, can take advantage of the vast amount of information available in large reference collections of mass spectra.

In general, mass spectra are considered as points or vectors in a multidimensional space with each dimension indicating the observed intensity at a specific mass position. In a conventional library search, the unknown spectrum is compared with entries in a library file by finding the distance between the point representing the unknown spectrum and the points representing the library entries. The compounds represented by the points nearest the unknown are reported by the search program as possible matches. Instead of a single point, the spectra of a set of related mixtures will define a subspace in the larger multidimensional space. To identify the components of the mixture set, a comparison between this subspace and the points representing library entries can be made. Compounds corresponding to points lying near the subspace are tentatively identified as components of the mixture. In implementing a search strategy along these lines, two factors need to be considered: (1) the

definition of the subspace representing the mixture spectra, and (2) the comparison between this subspace and the library entries. Several related methods drawn from linear algebra are available as possible alternative ways of performing a search of mixed spectra.

One method of defining the subspace which characterizes the mixed spectra and of comparing the reference spectra with this subspace relies on the Gram—Schmidt vector orthogonalization process. This technique has previously been used to calculate reconstructed gas chromatograms directly from interferograms produced in g.c.—F.t.i.r. experiments [18]. In that application, interferogram segments are treated as multidimensional vectors, and a set of these vectors from interferograms collected when no sample was present in the light path of the instrument are used to define a basis subspace characteristic of the instrumental background or "baseline" conditions. Segments of interferograms collected during a g.c. run are treated individually as vectors, and the orthogonal distance between the basis subspace and each vector is computed. Plotting these distances for the vectors drawn from a series of sequential interferograms produces a reconstructed gas chromatogram. In an analogous manner, vectors representing the mass spectra of mixtures can define a basis subspace, and the distance between this subspace and vectors representing reference spectra can be computed. With the Gram—Schmidt vector orthogonalization method, a basis vector is formed for each mixture spectrum. Each basis vector is orthogonal to the others and together they define the basis subspace. Those spectra which lie nearest this subspace, as measured by the orthogonal distance, are identified as components of the set of related mixtures. This method gives no information about the number of components present in the mixture; however, a basis subspace will span as many dimensions as there are mixed spectra.

Previous work cited above has shown that a principal component analysis of the vectors representing the mixture spectra can provide an estimate of the number of components present, if several conditions are met. In addition to the constraint that mixture spectra must be linear combinations of pure component spectra, the component spectra must be linearly independent. The concentrations of the components must vary independently, and at least as many mass positions (dimensions) and mixture spectra as there are components must be used in the analysis. Under these circumstances, a principal component analysis can give a correct estimate of the number of components present in the mixtures. This estimate of the number of components is the dimensionality of the subspace spanned by real, non-noise principal components of the mixture data. If the "true" dimensionality of the basis space is known (as n) the eigenvectors computed during the principal component analysis can be used to define an n-dimensional basis space. Instead of a basis vector for each mixture spectrum, there is a basis vector for each component in the mixture. Because the number of mixture spectra used is generally an overestimate of the number of components, this approach has the advantage over the orthogonalization method that fewer basis vectors are necessary, and therefore fewer computations are required for each comparison of a library entry with

the basis subspace. Also, by defining a basis subspace in terms of the n principal components, residual components which correspond to error in the measurement of the mixed spectra will be omitted from the basis subspace [19]. However, if the number of components is underestimated, significant information will be omitted. The effectiveness of this approach demands an accurate estimate of the number of components present.

Several related methods are available for the comparison of vectors representing library mass spectra with the basis subspace. Figure 1 shows a hypothetical example which illustrates the comparison of a library reference vector, **R**, with a two-dimensional subspace defined by orthogonal vectors, $\mathbf{B}_1$ and $\mathbf{B}_2$. With the Gram—Schmidt vector orthogonalization method, the orthogonal distance, d, between the reference vector and the basis plane is the quantity of interest. The target transformation method described by Malinowski and McCue relies on a comparison between the library vector, which is referred to as a target, and the vector produced by projecting the reference vector into the basis subspace, which is called the transformed target [20, 21]. In Fig. 1, the vector **R** represents the target and **P** represents the transformed target which lies in the plane formed by $\mathbf{B}_1$ and $\mathbf{B}_2$. If the transformed target matches the target within experimental error, the target is considered to be a possible component of the mixture. A third method for comparing reference vectors with the basis subspace is to compute the length of the projection into the basis space for a reference vector normalized to unit length. Ho et al. [22] have pointed out the utility of this method, which is related to Bessel's inequality, for determining the fit between a normalized vector and an orthonormal basis set. The length of this projected vector is denoted by l in Fig. 1.

These alternative methods of defining a basis subspace which contains the information from a set of mass spectra of related mixtures and of comparing

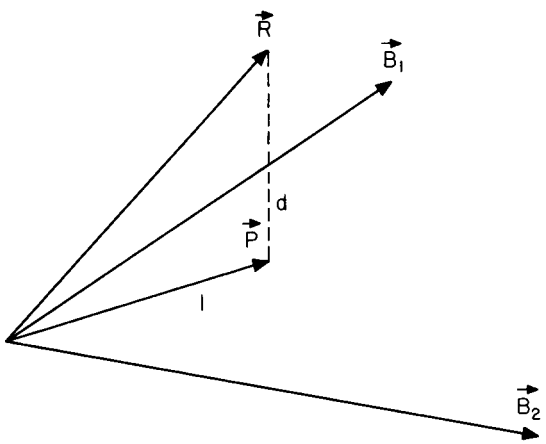


Fig. 1. Hypothetical vectors illustrating the comparison of a library entry with a two-dimensional basis subspace.

reference spectra with this space are tested with a large library of mass spectra and data for several sets of mixtures. A mathematical description of each method is presented, and the computational efficiency of alternative methods is assessed. A specific search strategy is described and tested with data from several sources.

THEORY

As previously mentioned, the key assumption in the application of these methods is that the mass spectra of mixtures behave as linear combinations of the pure component spectra. Ideally, the mass spectral data matrix $\hat{D}$, which is an $m \times l$ matrix of observed intensities, where m is the number of mass positions and l is the number of mixed spectra, can be expressed as the product of a matrix $\hat{S}$ of the spectra of the pure components and a matrix $\hat{C}$ of concentration coefficients. The $\hat{S}$ matrix is an $m \times n$ matrix, where n is the number of components in the mixture, and the $\hat{C}$ matrix is an $n \times l$ matrix. Thus $\hat{D} = \hat{S}\hat{C}$. The columns of $\hat{S}$ are the mass spectra of the pure components, and the rows of $\hat{C}$ are concentration profiles for the pure components. In these terms the problem is to identify the matrix, $\hat{S}$, given $\hat{D}$ and a reference library containing possible columns of $\hat{S}$.

Gram—Schmidt orthogonalization

In applying the Gram—Schmidt orthogonalization technique, the set of basis vectors is first computed. The first basis vector is formed by simply normalizing the first mass spectral data vector to unit length as expressed in eqn. (1), where $\mathbf{D}_1$ is the vector formed from the first column of $\hat{D}$ and $\mathbf{B}_1$ is the first basis vector:

$$\mathbf{B}_1 = \mathbf{D}_1/(\mathbf{D}_1{}^T\mathbf{D}_1)^{1/2} \tag{1}$$

To form the second basis vector, the component of $\mathbf{D}_2$ orthogonal to $\mathbf{D}_1$ is computed from

$$\mathbf{O}_2 = \mathbf{D}_2 - (\mathbf{B}_1{}^T\mathbf{D}_2)\mathbf{B}_1; \quad \mathbf{B}_2 = \mathbf{O}_2/(\mathbf{O}_2{}^T\mathbf{O}_2)^{1/2} \tag{2}$$

where $\mathbf{O}_2$ denotes the orthogonal component, and this vector is normalized to unit length.

The orthogonal component of each mass spectral data vector to all previously computed basis vectors is found and normalized until all columns of $\hat{D}$ have been used to form the $m \times l$ matrix $\hat{B}$ of orthonormal basis vectors. In a similar manner, a given m-dimensional column vector representing a library entry $\mathbf{L}$ is compared with the basis vectors by finding the component of $\mathbf{L}$ orthogonal to all basis vectors from the equation $\mathbf{O}_\mathbf{L} = \mathbf{L} - \hat{B}\hat{B}^T\mathbf{L}$, where $\mathbf{O}_\mathbf{L}$ is the orthogonal component of the $\mathbf{L}$ vector. The distance $d_\mathbf{O}$ between the library vector and the basis subspace is simply the length of $\mathbf{O}_\mathbf{L}$: $d_\mathbf{O} = (\mathbf{O}^T\mathbf{O})^{1/2}$.

Target transformation

The initial step in applying the target transformation method is to perform a principal component analysis [22]. The $l \times l$ covariance matrix $\hat{X}$ is computed from $\hat{X} = \hat{D}^T\hat{D}$, and an eigenanalysis is carried out to determine the eigenvalues λ and eigenvectors **e** which satisfy the equation

$$\hat{X}\mathbf{e}_k = \lambda_k \mathbf{e}_k \tag{3}$$

The eigenvectors, which when ordered by decreasing magnitude of their respective eigenvalues to form a matrix $\hat{e}$, are used to rotate the original data matrix $\hat{D}$ to produce an $m \times l$ working matrix $\hat{W}$. The number of components, n', is estimated by some suitable criterion, and the first n' columns of $\hat{W}$ define the basis subspace. To test a given library entry (or target) represented by L, an n'-dimensional rotor vector **R** is formed, with the elements of **R** being calculated from

$$R_i = 1/\lambda_i \, \mathbf{W}_i^T \mathbf{L} \tag{4}$$

where λ_i is the ith eigenvalue and $\mathbf{W}_i$ is the ith column of $\hat{W}$ for $\hat{W} = \hat{D}\hat{e}$.

Finally, the transformed target **T** is computed by multiplying the first n' columns of $\hat{W}$ by **R**: $\mathbf{T} = \hat{W}_{n'}\mathbf{R}$. Here $\hat{W}_{n'}$ denotes the reduced matrix composed of the first n' column of $\hat{W}$.

The vectors **T** and **L** are vectors of mass spectral intensities and for purposes of a search can be compared by any conventional distance metric. In addition to this qualitative information, quantitative data can also be derived with the target transformation. If the n rotor vectors, **R**, of the mixture components are taken together to form an $n \times n$ rotor matrix, $\hat{R}$, an estimate of the concentration matrix, $\hat{C}'$, can be computed from the equation, $\hat{C}' = \hat{R}^{-1}\,\hat{e}'^T$, where $\hat{e}'$ consists of the first n' columns of $\hat{e}$. Although precise quantitative information requires knowledge of the concentrations for each component in one mixed spectrum, rough estimates can be obtained without a standard known mixture spectrum.

Bessel's inequality test

The approach which makes use of Bessel's inequality test also begins with a principal component analysis. Another way to define the basis subspace is to perform an eigenanalysis of the $m \times m$ covariance matrix $\hat{Y}$, as summarized by the equations

$$\hat{Y} = \hat{D}\hat{D}^T \text{ and } \hat{Y}\mathbf{f}_k = \lambda_k \mathbf{f}_k$$

where $\mathbf{f}_k$ represents the eigenvectors of this covariance matrix.

As with the target transformation, the number of components is estimated, and is again denoted n'. To test a vector **L** representing a library entry, the vector is first normalized to unit length to form the vector **L***. A coefficient of fit, b, is computed as the sum of the squares of the dot products between **L*** and the first n' eigenvectors:

$$\mathbf{L}^* = \mathbf{L}/(\mathbf{L}^T\mathbf{L})^{1/2} \text{ and } b = \sum_{k=1}^{n'} (\mathbf{f}_k \mathbf{L}^*)^2$$

This coefficient of fit reflects the nearness of the library vector to the basis subspace and will have a value between 0 and 1 according to Bessel's inequality [16]. In geometric terms it represents the square of the length of the projection of L^* onto the basis space. The similarity index can be converted to a distance metric for search purposes: $d_B = 1 - b$. The square root of d_B is the orthogonal distance between L^* and the basis space. In the context of a library search, omitting the square root is acceptable because the order of matches will be unchanged.

In practice, more mass positions than mixture spectra are generally considered so that m is typically greater than l. For computational convenience it is preferable to perform the eigenanalysis on the smaller $l \times l$ covariance matrix. If this is done, the first n' of the eigenvectors, f_k, used with the Bessel test can be computed from

$$f_k = W_k/\lambda_k^{1/2} \tag{5}$$

The matrix $\hat{f}$ is analogous to $\hat{W}$ with the difference that the columns of $\hat{f}$ are normalized to unit length by dividing the elements in each column of $\hat{W}$ by the square root of the respective eigenvalue.

EXPERIMENTAL

All computer programs were written in FORTRAN IV and executed on IBM 360/75 or 370/155 computer systems available at the University of North Carolina Computation Center. The reference library used for the searches consisted of 16924 mass spectra of organic compounds drawn from the Registry of Mass Spectra Data [23]. These spectra included all peaks at integral mass positions with intensities above 1% relative to the base peak. Intensity information was recorded with a 1% resolution. The mass spectra of series of related mixtures were obtained from several sources. In computing the mathematically generated spectra used initially, the spectra of pure components were taken from a source independent of the library file in order to provide a more rigorous test of the methods [24]. Sets of mass spectra of real mixtures were taken from the literature [15] or were prepared and run locally. Those mass spectra collected locally were obtained on a DuPont Model 21-490B mass spectrometer equipped with an oscillographic recorder and digitized by hand. Like the library spectra, all mixture spectra were base peak normalized, and only peaks equal to or greater than a 1% relative intensity were retained.

RESULTS AND DISCUSSION

For an initial evaluation of this general approach to the identification of components in mixtures, three separate search strategies were implemented and tested with sets of spectra of mixtures generated mathematically. One search program used the Gram—Schmidt orthogonalization to define the basis space and compared reference spectra with the basis space by computing the

orthogonal distance. A second search program used the target transformation. The number of components was estimated with the indicator function described by Malinowski [25], and reference spectra were compared with transformed targets by a Euclidean distance metric. The third search strategy used the first n' eigenvectors of the $\hat{D}^T\hat{D}$ covariance matrix to define the basis space, and n' was determined by a 99% variance criterion. By this criterion, eigenvalues are ranked in descending order, and beginning with the first, successive eigenvalues are considered until the cumulative variance spanned by the associated eigenvectors equals or exceeds 99% of the total variance. The fit between normalized reference spectra and the basis space was determined with the distance metric based on Bessel's inequality. Sets of mixtures containing three and five components were searched, and for an n-component mixture, the correct compounds were generally found as the n nearest matches. As an example, Table 1 reports the results for searches based on seven mixed spectra of a

TABLE 1

Search results for a five-component mixture
(Asterisks denote correct components)

Search strategy	Estimated number of components	Distance	Nearest matches
Orthogonal distance	Not estimated	9.2	*Isopropyl ether
		14.9	*Phenyl acetate
		21.4	*2,9-Dimethyl-5,6-dithiadecane
		22.7	*Acetophenone
		23.3	Phenyl-*N*-methylcarbamate
		23.8	Phenol
		25.6	*Benzyl acetate
		26.8	1-Phenyl-1,2-propanedione
		27.2	Phenyl n-propyl ether
Target transformation	5	9.0	*Isopropyl ether
		15.1	*Phenyl acetate
		18.6	*2,9-Dimethyl-5,6-dithiadecane
		22.8	*Acetophenone
		23.0	Phenyl-*N*-methylcarbamate
		23.1	Phenol
		26.0	*Benzyl acetate
		27.3	*s*-Methyl thiobenzoate
		51.0	1-Phenyl-1,2-propanedione
Bessel's inequality	5	0.006	*Isopropyl ether
		0.021	*Phenyl acetate
		0.024	*2,9-Dimethyl-5,6-dithiadecane
		0.028	*Acetophenone
		0.037	*Benzyl acetate
		0.044	1-Phenyl-1,2-propanedione
		0.049	Benzoyl chloride
		0.050	Phenyl-*N*-methylcarbamate
		0.050	Phenol

five-component mixture. The method based on Bessel's inequality found the five correct compounds as the five nearest matches, while the other two searches found four of the five compounds as the four nearest matches and the fifth compound as the seventh nearest match. The similarity in distances computed for specific compounds by the target transformation and by the vector orthogonalization suggest that the two methods have defined very similar basis spaces. This example typifies the results obtained with mathematically generated mixture spectra and illustrates the feasibility of the general approach.

Further tests on mass spectra of real mixtures were performed with a search strategy designed to combine the most useful features of the three earlier programs. A method in which principal component analysis was used to define the basis subspace was preferred because this approach provides an estimate of the number of components present and therefore generally permits use of fewer basis vectors than does the vector orthogonalization method. Because the number of computations required for comparison of a reference vector with the basis space is a roughly linear function of the number of basis vectors, it is desirable to use only as many basis vectors as necessary. The target transformation method requires computation of a rotor vector, calculation of the transformed target, and then comparison of the transformed target with the library entry. By contrast, the test based on Bessel's inequality requires only the computation of squares of dot products of each basis vector with a normalized reference vector, and this method was therefore used for distance calculations. To limit the size of the covariance matrix diagonalized in the eigenanalysis, the $\hat{D}^T\hat{D}$ covariance matrix was used with a subsequent rotation and scaling of the data matrix according to eqn. (5) and the equation $\hat{W} = \hat{D}\hat{e}$ to produce the normalized basis vectors. Because it was believed to be more reliable than the 99% variance criterion, Malinowski's indicator function was used to estimate the number of components present. Results from a search were reported by listing the compound name, molecular weight, molecular formula, and computed distance for a fixed number of nearest matches specified by the operator. Because a comparison between a library spectrum and the transformed target can be useful in the evaluation of the results of a search, an option to list or plot library spectra and transformed targets of selected nearest matches is included. Table 2 summarizes the main steps of this search program.

This search strategy was used to identify the components in three sets of two-component mixtures. Table 3 reports the five nearest matches selected for each series of related mixtures. In each case, the two components were correctly identified by the search. Results for the two searches of the cyclohexane—hexane mixtures are particularly interesting. In the original paper, Ritter et al. [15] were initially surprised to find that three components were present in this mixture set, but were able to attribute the third component to nitrogen. When data for the peaks at 28 a.m.u. were excluded from the analysis, the mixture set was found to contain only two components. When

TABLE 2

Outline of search program

I. Set up the data matrix
   A. Read the mixture spectra
   B. Generate the data matrix retaining data for all mass positions with peak intensities above a threshold
   C. List the data matrix (optional)
   D. Plot the mixture spectra (optional)
II. Perform the principal component analysis
   A. Calculate the covariance matrix
   B. Perform the eigenanalysis
   C. Estimate the number of components
   D. Compute the basis vectors
III. Perform the search
   A. Compute a distance for each library entry and maintain a running list of nearest matches
   B. List the distance, name, and molecular formula and weight for the nearest matches
IV. Perform the target transformations (optional)
   A. Retrieve targets from library file
   B. Compute transformed targets
   C. List and/or plot targets and transformed targets
   D. Estimate approximate relative concentrations

data for the same peaks were included here, the search found spectra with prominent 28 a.m.u. peaks as the nearest matches after cyclohexane and hexane. However, when data from the peaks at 28 a.m.u. were omitted from the mixtures searched, saturated hydrocarbons appeared as nearest matches after the correct compounds. Considering that the data for the mixtures reported in Table 3 were collected by different operators on three different mass spectrometers, the results indicate that components in real mixtures can be successfully identified by this search strategy.

Of the nearly 17000 mass spectra in the library file, many will be quite dissimilar from the basis space defined by a given set of mixture spectra. To save the computation time that would be spent calculating the distances for library spectra which are quite different from the components of the mixture, two optional prefilters were added to the search program. Prefilters in conventional library searches are simple preliminary comparisons between library and unknown spectra which are used to eliminate from consideration those library entries which are significantly different from the unknown. When the distance calculation must consider a multidimensional rather than a one-dimensional unknown, effective prefilters are even more important. One prefilter applied here is referred to as an "overlap" requirement. As a library spectrum is being normalized, the intensities of all peaks in it and those at mass positions included in the mixture subspace are summed separately. The overlap prefilter requires that a minimum fraction of a total peak intensity in a library spectrum occur at mass positions included in the definition of the basis space. Generally, true components of a mixture set show an overlap in excess of

TABLE 3

Results for searches of two-component mixtures

	Mixture 1		Mixture 3	
True components	Cyclohexane[a] Cyclohexene		Cyclohexane[a] Hexane	
Number of mixtures	4		5[b]	
Number of mass positions	20		19	
Estimated number of components	2		3	
	Distance	Compound	Distance	Compound
Nearest matches	0.008	Cyclohexane	0.008	Cyclohexane
	0.015	Cyclohexene	0.012	Hexane
	0.028	Bicyclo(3.1.0)hexane	0.030	Carbon monoxide
	0.144	Bicyclopropyl	0.032	Nitrogen
	0.148	Fluorocyclohexane	0.058	Dideuteroacetylene
	Mixture 2		Mixture 3A	
True components	Acetophenone Phenyl-2-propanone		Cyclohexane[a] Hexane	
Number of mixtures	5		5[b]	
Number of mass positions	41		18[c]	
Estimated number of components	2		2	
	Distance	Compound	Distance	Compound
Nearest matches	0.010	Phenyl-2-propanone	0.011	Cyclohexane
	0.017	Acetophenone	0.012	Hexane
	0.081	1-Phenyl-1,2-propanedione	0.101	2,2,3-Trimethylbutane
	0.091	Benzoic acid hydrazide	0.122	2,5-Dimethylheptane
	0.104	3-Hydroxy-3-phenyl-cyclohexanone	0.134	2,3-Dimethylpentane

[a]Data taken from ref. 9.
[b]Spectra of pure components reported in ref. 9 are omitted.
[c]Data for peaks at 28 a.m.u. omitted.

TABLE 4

Prefilter statistics as percentage of library spectra passing the prefilter

Prefilter	Cyclohexane Cyclohexene	Cyclohexane Hexane	Acetophenone Phenyl-2-propanone
0.1% overlap	89.6	86.7	96.0
80% overlap	2.3	3.5	3.5
Molecular weight	4.0	4.0	21.1
and 80% overlap	0.9	0.9	1.9

90%, although overlap requirements set lower than this can be quite effective. The second optional prefilter relies on the assumption that the highest mass position appearing in the mixture spectra represents approximately the molecular weight of the heaviest component in the mixture. This prefilter takes

advantage of the fact that the library file is ordered on molecular weight. This prefilter causes the search to stop considering library entries when the molecular weight of a library entry exceeds the highest observed mass in the mixture spectra rounded off to the next higher multiple of 10 a.m.u. Although it must be avoided if molecular ions may be absent, this prefilter can eliminate from consideration higher molecular weight homologues of true mixture components and save considerable time that would otherwise be spent searching high molecular weight compounds not likely to be mixture components. Table 4 reports statistics showing the effectiveness of these prefilters. By reducing the number of compounds considered in detail by the distance algorithm, these prefilters can substantially improve the efficiency of the search.

For a more rigorous test of the search program, mass spectra of nine related mixtures containing four components were collected and analyzed. The results of this search are reported in Table 5. The mixture is identified as containing four components but the third, fourth, and fifth nearest matches are long-chain saturated alkane spectra. One suspects that the similarity of the alkane spectra may have confused the search and that the terpene alcohol found as the sixth nearest match may be a component of the mixture in addition to one of the three alkanes. This suspicion is confirmed by an examination of the approximate concentrations calculated using the rotor matrix computed with a target transformation of the four nearest matches as listed in Table 6. The large negative concentrations are unacceptable, and similar results are obtained if the first, second, third, and fifth nearest matches are used. However, if the first, second, third, and sixth nearest matches are considered, the concentrations estimated seem reasonable within experimental error, confirming the presence of the terpene alcohol and one alkane instead of two alkanes. Although dodecane is the best match of the alkanes, it may not really be the best choice. The presence of a peak at 170 a.m.u. in the mixture spectra rules out n-undecane. In deciding between 5-methylundecane and dodecane an examination of the library spectra and the transformed targets is useful. Figure 2 shows the mass spectra and target transformations obtained with 5-methylundecane and n-dodecane. For the library entries, the major differences are that the intensity of the molecular ion is reduced in the branched alkane, and the peaks at 85, 86, 112, and 113 are enhanced as a result of

TABLE 5

Search results for a four-component mixture
(Estimated number of components 4. Number of mass positions 93)

Distance	Compound	Distance	Compound
0.010	ar-Methoxybenzaldehyde	0.070	5-Methylundecane
0.036	di-n-propylamine	0.071	3,7-Dimethyl-2,6-octadiene-1-ol
0.044	n-Dodecane	0.076	n-Decane
0.069	n-Undecane	0.086	20-Methylnonane

TABLE 6

Estimated and actual relative concentrations for nine spectra of the four-component mixture

Components	1	2	3	4	5	6	7	8	9
Estimated									
Methoxybenzaldehyde	0.83	0.68	0.39	0.12	−0.01	0.01	0.00	0.00	0.00
Di-n-propylamine	−0.02	−0.04	−0.07	−0.10	−0.09	0.10	0.69	0.77	0.56
n-Dodecane	−2.97	−5.74	−10.0	−15.2	−13.8	−1.77	2.24	2.79	5.53
n-Undecane	3.17	6.10	11.6	16.2	14.9	2.57	−1.92	−2.55	−5.09
Methoxybenzaldehyde	0.94	0.89	0.72	0.36	0.02	0.01	−0.01	0.00	0.00
Di-n-propylamine	0.00	0.00	0.00	0.00	−0.01	0.23	0.63	0.68	0.44
n-Dodecane	0.01	0.01	0.02	0.12	0.56	0.72	0.40	0.35	0.61
3,7-Dimethyl-2,6-octadiene-1-ol	0.04	0.10	0.26	0.52	0.43	0.04	−0.02	−0.03	−0.05
Actual (mole fraction)									
Methoxybenzaldehyde	0.91	0.84	0.64	0.29	0.05	0.00	0.00	0.00	0.00
Di-n-propylamine	0.00	0.00	0.00	0.00	0.01	0.25	0.63	0.70	0.46
n-Dodecane	0.00	0.00	0.01	0.08	0.36	0.54	0.34	0.29	0.53
3,7-Dimethyl-2,6-octadiene-1-ol	0.09	0.16	0.35	0.63	0.58	0.21	0.03	0.01	0.01

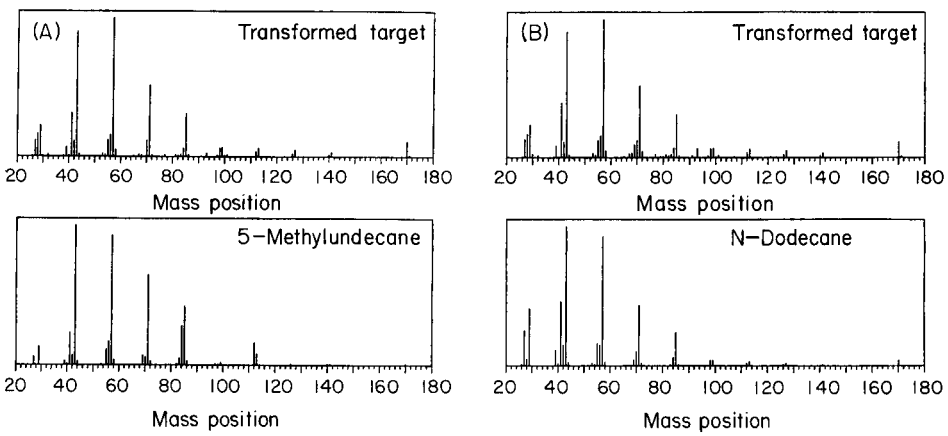


Fig. 2. Reference spectra and transformed targets of (A) 5-methylundecane and (B) n-dodecane.

bond cleavage and charge retention at the methyl-substituted carbon atom. These differences are clear in the reference spectra. The transformed target spectrum of 5-methylundecane lacks these features, and in fact, both transformed targets are most similar to the normal alkane, indicating that n-dodecane is a component of the mixture. This detailed examination of the nearest matches selected by the search program makes it possible to identify the correct components in this set of mixture spectra, even though the presence of very similar alkane spectra in the library had confused the search algorithm slightly.

In summary, these examples illustrate the use of a search method applied to mass spectra of mixtures. Mixtures of varying complexity from several sources are successfully analyzed. In principle, the method described here is not limited to mass spectral data, but may be applied to any type of data that obey the constraints described previously.

CONCLUSIONS

Components of mixtures can be effectively identified by a library search method utilizing a multidimensional basis space to characterize the information contained in the mass spectra of a series of related mixtures. Principal component analysis is used to extract the significant information from a data matrix. Prefilters increase the efficiency of this method by limiting the number of library entries compared completely by the distance algorithm. Target transformations of nearest matches provide a means of verifying the results of a search and enable the search performance to be evaluated in detail.

This work was supported by the National Science Foundation, Grant No. CHE78-00632.

REFERENCES

1 F. P. Abramson, Anal. Chem., 47 (1975) 45.
2 F. W. McLafferty, R. H. Hertel and R. D. Villwock, Org. Mass Spectrum., 9 (1974) 690.
3 D. P. Tunnicliff and P. A.Wadsworth, Anal. Chem., 37 (1965) 1083.
4 P. W. Fausett and J. H. Weber, Anal. Chem., 50 (1978) 722.
5 J. B. Gayle and H. D. Bennett, Anal. Chem., 50(14) (1976) 2085.
6 L. F. Monteiro and R. I. Reed, Int. J. Mass Spectrom. Ion Phys., 2 (1969) 265.
7 J. McK. Halket and R. I. Reed, Org. Mass Spectrom., 10 (1975) 370.
8 R. Reimendal and J. B. Sjovall, Anal. Chem., 45 (1973) 1083.
9 J. E. Billar and K. Biemann, Anal. Lett., 7 (1974) 515.
10 R. G. Dromey, M. J. Stefik, T. C. Rindfleisch and A. M. Duffield, Anal. Chem., 48 (1976) 1368.
11 D. Henneberg, H. Damen and B. Weimann, Adv. Mass Spectrom., 78 (1976) 975.
12 P. Powers, M. J. Wallington and J. A. V. Hopkinson, Adv. Mass Spectrom., 7B (1976) 1029.
13 J. E. Davis, A. Shepard, N. Stanford and L. B. Rogers, Anal. Chem., 46 (1974) 821.
14 J. McK. Halket and R. I. Reed, Org. Mass Spectrom., 10 (1975) 808.
15 G. L. Ritter, S. R. Lowry, T. L. Isenhour and C. L. Wilkins, Anal. Chem., 48 (1976) 591.
16 W. H. Lawton and E. A. Sylvestre, Technometrics, 13 (1971) 617.
17 N. Ohta, Anal. Chem., 45 (1973) 553.
18 J. A. de Haseth and T. L. Isenhour, Anal. Chem., 49 (1977) 1977.
19 E. R. Malinowski, Anal. Chem., 49 (1977) 606.
20 E. R. Malinowski and M. McCue, Anal. Chem., 49 (1977) 284.
21 P. H. Weiner, E. R. Malinowski and A. R. Levinstone, J. Phys. Chem., 74 (1970) 4537.
22 C.-N. Ho, G. D. Christian and E. R. Davidson, Anal. Chem., 50 (1978) 1108.
23 E. Stenhagen, S. Abrahamsson and F. W. McLafferty, Registry of Mass Spectral Data, Wiley—Interscience, New York, 1974.
24 R. M. Silverstein and G. C. Bassler, Spectrometric Identification of Organic Compounds 2nd edn., Wiley, New York 1967.
25 E. R. Malinowski, Anal. Chem., 49 (1977) 612.

Analytica Chimica Acta, 112 (1979) 165—173
Computer Techniques and Optimization

A COMPUTERIZED SYSTEM FOR THE DIGITAL IMAGE PROCESSING OF ION MICROSCOPE IMAGES

J. D. FASSETT, D. M. DRUMMER and G. H. MORRISON*
Department of Chemistry, Cornell University, Ithaca, N.Y. 14853 (U.S.A.)
(Received 8th December 1978)

SUMMARY

A general program for the digital image processing of ion microscope images is described. The program IONPIX and the accompanying aggregate of related subprograms comprise a basis for the process of extracting the large amount of compositional morphological information contained in ion images. Previous practical applications of the IONPIX software system are also described.

The ion microscope is a uniquely powerful analytical tool, combining high elemental sensitivity and sampling on a micrometer scale with the ability to provide information concerning the compositional morphology of solid samples [1]. The technique has previously demonstrated its utility in both the ion counting [2, 3] and depth profiling [4, 5] modes of operation. A less often used but potentially much greater source of information is the ion image. An ion microscope image contains a vast amount of information, both spatial and elemental. Previously it has been difficult to utilize ion images in anything other than a qualitative fashion, but the increased availability of high-speed computer systems and digital image-processing techniques has made it possible to extract copious information, both qualitative and quantitative, from ion images. This paper describes such a system which has been developed and is currently used in this laboratory.

The techniques of digital processing are now routinely used in microscopy, x-ray diffraction, astronomy and many other fields which generate large amounts of photographic data [6—8]. In most of these applications, the primary concern is the "quality" of the image, a subjective evaluation which utilizes a priori information about the image. While this is also a consideration in the treatment of ion images, processing methodology as developed for ion microscopy is more concerned with the correlation of structure and composition. Two major facets of this methodology are the identification of and ion intensity measurements of microfeatures within the image field. Other procedures of interest include film emulsion calibration, cross-correlation of ion images, manipulation of image orientation and size and finally, display of the processed image.

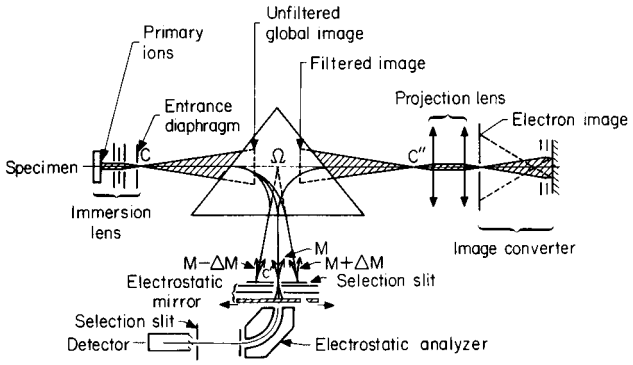


Fig. 1. Schematic representation of CAMECA IMS-300 ion microanalyzer [1].

HARDWARE

The CAMECA IMS-300 ion microscope used in this laboratory has been previously described [1], and is illustrated schematically in Fig. 1. Basically, ions leaving the sample surface are extracted and subjected to a double pass through the magnetic section, wherein the global ion image is momentum-filtered, producing a mass-resolved image of the species of interest. The ions then pass through a projection lens system, which determines image magnification, and strike a Cu-Be ion-to-electron converter. The electrons are accelerated to 21 keV and the image is recorded on electron-sensitive roll film (Agfa-Gevaert 37C50, 35 mm). A typical ion image has a diameter of 28 mm on the film, which corresponds to a sampled area diameter of approximately 250 μm with a point-to-point spatial resolution of about 1 μm. Standard photographic development procedures are used.

A Photomation Mark II scanning microphotodensitometer interfaced to a PDP 11/20 computer is used to digitize the images. The film is secured over an opening on a cylindrical drum which rotates at high speed, passing the film between an incoherent optical source and detector. The scanner is automatically stepped along each film axis, and the density of each point on the image is transferred to the computer as a digitized value. A density range of 0—2 D (100%—1% T) or 0—3 D (100%—0.1% T), determined by the operator, is divided uniformly into 256 discrete levels. Further details concerning the densitometer and digitization procedure have been previously published [9].

The computer facilities for image processing (Fig. 2) consist of a PDP 11/20 with a 24K word memory, a 1.2-M word cartridge disk, a dual fixed disk assembly, a 9-track magnetic tape unit, a high-speed line printer/plotter, an incremental plotter and a GT40 graphics display and processor.

The computer system at this time is used only as a processor of ion image data matrices. It would be possible, however, through an appropriate image detection/display system and computer-ion microscope interface, to collect ion images directly in digitized form, thus eliminating the film intermediate

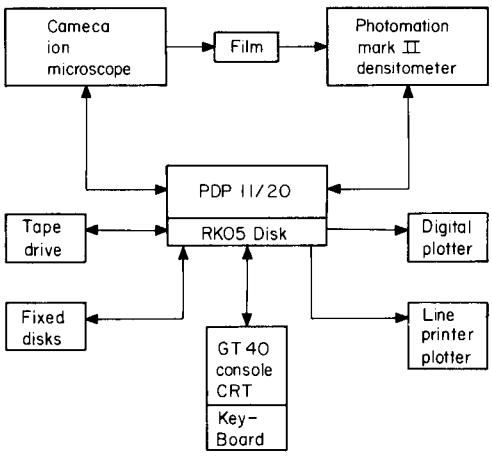


Fig. 2. Ion imaging system.

step. Another example of this case is the time-resolved scanning image normally obtained by an ion microprobe. With this in mind, the present software, with the exception of digitizing and emulsion calibration routines, should be applicable to any digitized ion image, irrespective of the method of acquisition.

IMAGE ACQUISITION SOFTWARE

The primary goal in the digitization of an ion image is that a minimum of information be lost. The acquisition software was designed to fulfil this requirement, the limitations imposed by the computer and the microphotodensitometer being taken into consideration. The computer is limited in the amount of data it can store, manipulate and transfer in a reasonable amount of time. The microphotodensitometer is limited by the specifications of the manufacturer.

The characteristics of the digitized ion image file (Fig. 3) result from what the authors consider the best possible combination of computer and microphotodensitometer parameters. The ion image is 28 mm in diameter and is digitized into a 256 × 256 square point matrix by using a 100-μm aperture setting on the microphotodensitometer. The actual film area scanned is therefore 25.6 mm × 25.6 mm. If the image is centered in the file, this results in a loss of about 6% of the total image, as well as inclusion of non-image parts at the file corners. These background points comprise about 10% of the file. Since there are 256 picture elements (pixels) spanning slightly less than an image diameter, and a typical image has a diameter of 250 μm with 1-μm spatial resolution, there is no loss of point-to-point resolution in the acquisition process.

The acquisition program ACQIRE, which creates the primary image data

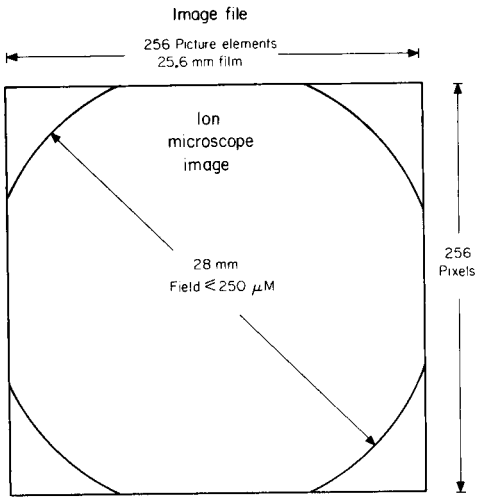


Fig. 3. Characteristics of digitized ion image.

file, is designed to store the data on cartridge disk in a minimum of space. Each image point has 8-bit resolution and is packed into a single byte of 8 bits in the image file. A 256 X 256 primary image file thus consists of 32K of 16-bit words. Also included in the file is a 512-word parameter line, which contains information concerning image size and mode, date and time of acquisition, and an optional description of the image. The total space required for storage of an image is therefore 32.5K, making it possible to store 36 images on a 1.2-M word cartridge disk.

IONPIX

The actual processing of image data is done by a general program, IONPIX, a keyboard command interpreter which supervises calls to the main FORTRAN routines. A comman line contains one instruction and consists of a letter code indicative of the called program followed by a series of operator-determined parameters necessary for the proper and desired execution of the instruction. Typically, the parameters include identification of the file or files being manipulated, code parameters which direct the flow of operation, and any numbers that may be required by a specific routine. Individual commands may also require subsequent information, which is handled in a normal operator—computer dialogue, but in most cases the command line will determine the entire flow of operation. The program also has provision for the creation of a command file, which consists of a series of commands that are executed sequentially. The system can then function unattended for extensive computations, permitting the operator to pursue alternative duties.

In normal IONPIX operation, five scratch files are created into which primary data can be transferred. The data loaded into the scratch files are converted from the packed bytes of the acquisition format to a real number format that

requires two words or 32 bits for each image datum. The size of the primary data file is therefore expanded by about a factor of four, with the 512-word header line unchanged.

The program IONPIX is much larger than the available core memory space in the computer, but is made operable by the overlay facility of the RT-11 monitor system. The command decoder of IONPIX is the root segment of the program and resides at all times in core. The subroutines to which the command decoder refers reside on disk. When a subroutine is called, it is loaded into the first overlay region, a specific run-time area of memory that is shared by all the subroutines. There also exists a second overlay region into which subroutines called from the first overlay region are loaded from the disk. The overlay structure is invisible in the actual operation of the program since the control monitor is extremely fast in its transfer of code. The overlay structure described is that which is required by the existing form of IONPIX and the presently available core memory space, but can easily be modified to be compatible with any change in software or hardware configurations.

All communication with image data files is by unformatted "read" or "write" statements that read or write an entire record, consisting of one image line, at a time. Since no calculation in any of the present programs requires more than two lines of data at a time, space is reserved in core for only two 256-point lines. The maximum space is set at 1K words of core. Data shuffling is the most time-consuming part of almost all the programs, and unformatted "reads" and "writes" offer the fastest transfers possible.

IONPIX subprograms

A list of the subprograms presently callable by IONPIX and their functions is found in Table 1. The subprograms are classified in several broad functional categories which include mathematical, single-file examining, image analysis, display, file shuffling, and ion-micrograph specific routines. These subprograms represent a basic set of useful file-handling routines expanded by the addition of subprograms developed specifically for images. The command interpreter structure of the main IONPIX program facilitates easy expansion by the incorporation of additional subprograms, which may be added as they are required.

The mathematical category of subprograms primarily contains standard image processing routines which perform simple mathematical manipulations. These subprograms have great utility in methods involving comparison or ratioing of images, as well as in various filtering, sharpening, and smoothing techniques.

Ion-micrograph specific routines were developed especially for the ion image as recorded on negative film, and apply primarily to the conversion from film-recorded density space to digitized ion-intensity space.

Single-file examining subprograms calculate various statistical values for single files, and allow immediate examination of current image data and file parameters. Most of the file-shuffling routines are standard image-processing

TABLE 1

Subprograms in IONPIX

Mathematical	
ADD	Adds two image files.
ARITH	Adds, subtracts, multiplies, or divides an image file by a constant. Takes the square root of an image file.
DERIV	Takes derivative of image file.
DETECT	Detects edges in image file.
DIVIDE	Divides one image file by another.
FOUREA	Takes fast Fourier transform of a file. Filters a Fourier-transformed image file.
MAGNIF	Magnifies image file.
MASK	Gradient edge masks image file; smoothes, defocuses or sharpens.
MULTIP	Multiplies two image files.
QUANT	Quantizes an image file at prescribed levels.
REDUCE	Demagnifies image file.
ROUNDO	Rounds off image data points to nearest integer.
SMOOTH	Smoothes image file by boxcar averaging.
XCORR	Cross correlates two complex image files.
Ion micrograph specific	
CALHIS	Analyzes characteristic curve standard images.
CIRCLE	Calculates center of ion image field.
DTOI	Converts density to ion intensity.
FILMCC	Calculates the characteristic curve of film.
KSUBT	Determines sampling constant from ion image.
Single file examining	
AVERGE	Calculates average value of an image file.
INSPEC	Examines a single line of data.
MINMAX	Determines maximum and minimum of file.
PRMLIN	Reads parameter line and allow changes.
File shuffling	
CHGSIZ	Reduces or expands size of image file.
DUPLIC	Duplicates an image file.
LOADA	Transfers primary file into scratch file.
ORIENT	Rotates an image file.
REDFIN	Redefines the size of a scratch f
SEQUNC	Creates a command file.
TRANSP	Transposes an image file.
UNLODE	Transfers scratch file into external storage.
Display	
DISPV	Displays image on electrostatic plotter.
HISTOG	Determines histogram of image file.
PLTCON	Contour plots image file.
XYZPLT	Plots an image in 3-dimensional perspective.
Image analysis	
FEATUR	Calculates statistical information about image features.
MAPMAS	Applies a feature map to an image file.
QUANTP	Determines features in image f

techniques, and serve to allow added flexibility within the existing IONPIX scratch file setup. The program UNLODE allows a processed file to be permanently stored on disk or magnetic tape for future use.

One subdivision of routines that deserves special mention is the display category. DISPV outputs a half-tone hard copy of an image or its negative to the high-speed electrostatic printer-plotter (Versatec 900A). The resulting image can contain up to 64 grey levels, and is particularly useful for inspection of an amended image file at any point between processing steps. It is also possible to display image information on an x—y digital plotter (Houston Instruments CØMPLØT) in the form of a histogram, a plot of density level

vs. frequency of occurrence, isodensity or isointensity contour plots, and pseudo-three-dimensional displays of image files, in which intensity information of an image is considered as a third dimension perpendicular to the spatial dimensions of the image. These functions are performed by subprograms HISTOG, PLTCON and XYZPLT, respectively. Although written for the specific display hardware mentioned above, similar software for different output devices should be easily realized.

Image analysis subprograms are implemented in the identification of microfeatures in images, and the calculation of information concerning the microfeatures.

APPLICATIONS

As an illustration of the potential power of digital image-processing techniques, the following treatments of selected ion images, which envelop many of the characteristics often encountered in the ion microscope images, are considered.

Characteristic curve correction

In the CAMECA IMS-300 used in this laboratory, electron-sensitive film is exposed to monoenergetic electrons produced by the collision of the mass-selected secondary ion beam on a Cu-Be ion-to-electron converter. Although the number of electrons impinging on the film will be proportional to the number of ions collected from a corresponding point on the sample, the same will not be true of the density produced on the film at that point, since the response of the film emulsion to electron exposure is not a linear function. The relationship between darkening of film and exposure to radiation has been well studied, the classical work in the field being that of Hurter and Driffield [10]. A procedure for conversion from film density to ion intensity for the ion detection system of the CAMECA instrument has been published by Fassett et al. [9], and is incorporated in the subroutine DTOI. The general philosophy of the conversion could easily be adapted to other methods of ion detection and, in the case of more direct imaging, may be unnecessary.

Microsampling

If the sample under study contains impurities of interest that are present as well-defined inclusions, the validity of a given analysis on a microscale may be questionable because of the problem of microsampling. Given a means of determining the degree of heterogeneity of the distribution of a given element within the sample, it is possible to estimate the precision to be expected for a microanalysis of a given sample area, or conversely, the number of replicate analyses necessary to achieve a desired precision. This approach has been demonstrated by Drummer et al. [11], and is accomplished by the use of the subroutine QUANTP, a routine in the image analysis category.

Multifeature multielement analysis

Much interest exists in the correlation of a series of ion images from the same sample area, as well as in the comparison of an ion image to an image produced by a complementary technique, such as light or electron microscopy. This type of approach would enable the analyst to extract from the image chemical or morphological data not obtainable from the analysis of a single ion image.

In order to be able to relate a series of images, it is most important to register each image relative to the image with which it is to be compared. Registration is required because of the slight transitional offset that results from manual positioning of the microphotodensitometer. It may also be necessary if the sample has been moved, or removed and then reinserted into the instrument, or for correlation with images from other techniques.

The translational and rotational offset can be found by applying the cross-correlation function, a traditional method of matching two images [12]. A cross-correlation is an iterative routine that converges to the offset required for best correlation. The cross-section routine of IONPIX, XCORR, uses a Cooley—Tukey fast Fourier transform [13] to calculate image offset.

Translational offset correction can be accomplished by a simple correction during any computational procedure involving the two files. Rotational offset is corrected by the routine ORIENT, which rotates an image file by a specified angle. When an ion image is compared to either another ion image or an image produced by another technique, any differences in image magnification can be corrected by application of the routines MAGNIF and REDUCE.

Identification of features is also an important facet of multifeature multielement analysis. If the features of interest are inclusions well-defined and isolated from the background, a simple thresholding method may suffice, as in the microsampling study mentioned previously. A more probable case, however, is that of homogeneous features that do not have a constant background. The differentiation of crystal grains in a polycrystalline sample is an example of this. Delineation of features in this instance would then require a more complex method.

This particular problem was encountered by Fassett and Morrison in the investigation of crystal grains in polycrystalline iron [14]. This study illustrates the use of boxcar averaging (SMOOTH) and edge detection (DETECT) routines. Additional applicable processing techniques include the ratioing of related images (DIVIDE), and the calculation of statistical information concerning the features (FEATUR).

The ultimate goal of obtaining point-to-point quantitative information has also undergone preliminary investigation by image-processing techniques. Rüdenauer [15] has used a modified one-parameter local thermal equilibrium model for quantitative correction of signals at each point of a time-resolved scanning image. A different approach was demonstrated by Fassett et al. [9] who used external standards to obtain quantitative results from images of a bimetallic composite system. Although simply designed, this example shows

the applicability of digital image processing to the eventual realization of concentration mapping.

Conclusion

The possible applications of digital image processing to ion microscopy are potentially very numerous. It should be emphasized that the flexibility of programming is such that the software can readily be adapted to specific as well as general problems of microanalysis. The production of artificial or amended images may prove to be extremely valuable in summarizing image information and displaying elemental interrelationships. The potential correction of image ion intensity to concentration is of special interest, but there are also many possible applications of image feature analysis to composition—structure problems in fields such as biology, geology and metallurgy.

Although the existing image-processing software system was designed for use with the particular hardware available in this laboratory, its general philosophy and open-ended structure should make it easily adaptable to systems which differ in one way or another.

Aphoristically, it has been many times stated that a picture is worth a thousand words. A scientific analogy to this statement reiterates the basic truth: the information content of pictures or images can be very great. Through the use of the techniques of digital image processing, it should at last be possible for the investigator to begin to extract from ion images some of the copious information which heretofore has been hidden from view.

Financial support was provided by the National Science Foundation under Grant No. CHE78-04405 and through the Cornell Materials Science Center, and the National Institutes of Health under Grant No. GM24314-02.

REFERENCES

1 G. H. Morrison and G. Slodzian, Anal. Chem., 47 (1975) 932A.
2 C. A. Andersen and J. R. Hinthorne, Science, 175 (1972) 853.
3 J. D. Ganjei, D. P. Leta and G. H. Morrison, Anal. Chem., 50 (1978) 285.
4 J. A. McHugh, NBS Special Publ., No. 427 (1975) 179.
5 W. J. Devlin, K. T. Ip, D. P. Leta, L. F. Eastman, J. Comas and G. H. Morrison, Institute of Physics Conference Series No. 45, Gallium Arsenide and Related Compounds, St. Louis, Missouri, 1978.
6 N. M. Short, P. D. Lowman, Jr., S. C. Freden and W. A. Finch, Jr., Mission to Earth: Landsat Views The World, NASA, Washington, D.C., 1976.
7 A. Rosenfeld (Ed.), Digital Picture Analysis, Springer-Verlag, New York, 1976.
8 R. C. Gonzalez and P. Wintz, Digital Image Processing, Addison-Wesley Publishing Company, Reading, Mass., 1977.
9 J. D. Fassett, J. R. Roth and G. H. Morrison, Anal. Chem., 49 (1977) 2322.
10 F. Hurter and V. C. Driffield, J. Soc. Chem. Ind., 9 (1890) 455.
11 D. M Drummer, J. D. Fassett and G. H. Morrison, Anal. Chim. Acta, 100 (1978) 15.
12 K. B. Welles, Ph.D. Thesis, Cornell University, Ithaca, N.Y., 1976.
13 J. W. Cooley and J. W. Tukey, Math. Comp., 19 (1965) 297.
14 J. D. Fassett and G. H. Morrison, Anal. Chem., 50 (1978) 1861.
15 F. G. Rüdenauer, U.S.—Japan Joint Seminar on Secondary Ion Mass Spectrometry: Fundamentals and Applications, October 23—27, 1978, Takarazuka, Japan.

Analytica Chimica Acta, 112 (1979) 175—184
Computer Techniques and Optimization

MICROPROCESSOR-BASED DATA PROCESSING AND QUALITY CONTROL IN HEMATOLOGY

N. J. VERHOEF[‡], P. A. MANTEL* and B. LEIJNSE

Department of Chemical Pathology, Erasmus University, P.O. Box 1738, 3000 DR Rotterdam (The Netherlands)

and

Department of Clinical Chemistry, University Hospital, Rotterdam (The Netherlands)

(Received 7th July 1978)

SUMMARY

Routine hematological determinations with Coulter Counter model S instruments benefit from electronic data processing, and appropriate quality control is essential. The system described here is based on an INTEL 8008 microprocessor provided with options for handling sample identification numbers, duplicate print-out of results on self-adhesive labels, and zero-level quality control functions which utilize both patient samples and quality-control samples. The arithmetical means of the results of ten patient samples are calculated in real time for four parameters in whole blood. The running mean values of two of these parameters (*MCV* and *MCHC*) are suitable for zero-level process control, together with signalling of extremes. Daily mean values of patient samples and quality-control samples are also calculated to check the stability from day to day. The microprocessor-based system is adequate for these functions and is relatively simple and inexpensive.

Hematology is concerned with the study of cells found in peripheral blood and bone marrow: their number and characteristics, the clotting properties of the blood and the temporal variations of the different parameters provide a sensitive indication of various diseases. In this laboratory routine hematological determinations are done with two Coulter Counter model S instruments, which determine the following parameters in whole blood: the number of leucocytes (white cells) per liter (*LEU*), the number of erythrocytes (red cells) per liter (*ERY*), the mean cell volume (*MCV*) and the hemoglobin concentration (*HB*). The indices which are calculated by the instrument itself are: the hematocrit ($HT = MCV \times ERY$; i.e., the relative amount of cells), the mean amount of hemoglobin per cell ($MCH = HB/ERY$), and the mean cell hemoglobin concentration ($MCHC = HB/HT$). With the introduction of devices of this kind, which produce a lot of data in a relatively short time, the aid of electronic data processing and adequate quality-control become essential.

[‡]Present address: Clinical Chemistry Laboratory, St-Joseph Hospital, Pasteurlaan 9, Oosterhout (NB), The Netherlands.

The results of the seven parameters are printed out by the Coulter Counter on standard cards available from the manufacturer, but in many cases the handling of these cards does not suit laboratory administration. Furthermore, it is desirable to base the processing of the results on a sample identification number instead of the sequence number generated by the instrument.

Quality control has two distinct aspects: accuracy and precision. Generally, data for quality control can be derived from two sources: standard control samples may be run along with patient samples, or the patient whole blood samples may themselves be used as the source of quality-control data. Certified quality-control samples for hematological analyses are available only for hemoglobin, because of problems with long-term stability and because a known number of particles cannot be weighed so that packed cell volume standards cannot be produced [1—4]. In the present paper, the method of checking the proper functioning of the Coulter Counter is based on patient samples as well as on quality-control samples, implemented in a small microprocessor-based system. The only parameters of patient samples that can be applied for quality control are those which have narrow physiological ranges. This is the case for the erythrocyte indices MCV and MCHC. The accuracy has to be checked periodically, because sudden minor defects in the mechanical, pneumatic or electronic parts of the Coulter Counter may cause deviations in one or more parameters. The precision of the instrument is good; measurements show (Table 1) that the repeatibility for most parameters is better than 1%.

The Coulter Counter has to be calibrated with samples for which the hematological parameters have been determined by manual and/or semi-automated methods. Although the manufacturer advises calibration of the instrument against 4C Hematology Control (Coulter Diagnostics Inc.), this has been challenged recently [5]. To control the accuracy, it is desirable to have a sensitive reliable method which detects any drift as early as possible (zero-

TABLE 1

Repeatibility measurements with two separate Coulter Counters

Parameter		Coulter Counter I			Coulter Counter II		
		Mean[a]	S.d.	C.v. (%)	Mean[b]	S.d.	C.v. (%)
LEU	($\times 10^9\ l^{-1}$)	7.19	0.07	1.0	6.81	0.15	2.2
ERY	($\times 10^{12}\ l^{-1}$)	4.04	0.02	0.6	4.02	0.03	0.7
HB	(mmol l^{-1})	7.51	0.04	0.5	7.48	0.05	0.7
HT	(l^{-1})	0.371	0.002	0.6	0.376	0.003	0.9
MCV	(fl)	91.6	0.60	0.7	92.4	0.73	0.8
MCH	(amol)	1864	11.6	0.7	1851	13.0	0.7
MCHC	(mmol l^{-1})	20.1	0.17	0.9	19.8	0.23	1.2

[a] $n = 21$. [b] $n = 16$.

level quality control). Either quality-control samples or patient-sample values can be used. A simple visual method has been described [6] but in a large laboratory a less time-consuming method is needed.

Although there are some electronic devices, these do not possess the necessary flexibility or the ability to synchronize identification numbers with the sample intake. Thus the application of microprocessors was considered, as in various other areas of clinical chemistry and hematology microprocessors have proved their value and economy in dedicated process-control systems. Some years ago, a research project was started in order to evaluate their theoretical and practical aspects. As part of this work a microcomputer system — HEMAsys — has been developed for data processing and quality control in the hematology laboratory.

As far as quality control is concerned, not all parameters of patient samples are suitable [7]; the constancy of the level of the various parameters of the samples analyzed during the day was first investigated. On the basis of the data obtained, several preliminary parameters applicable for quality control were chosen. The microcomputer checked these parameters in daily routine analysis and thus allowed them to be studied in more detail.

EXPERIMENTAL

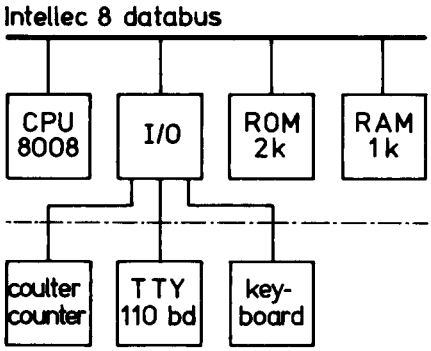
Calibration and samples

The Coulter Counter S was calibrated with samples obtained from blood donors as described by Coulter Diagnostics [5]. The hematological parameters of the samples were determined as follows: the leucocytes and erythrocytes were counted with a hemacytometer by three different technicians and with an electronic cell counter (Toa Microcell Counter, MCC-1002B, Toa Electric, Kobe, Japan). The hemoglobin content was determined by the hemicyanide method [8] with the use of a reference control from the Rijks Instituut voor de Volksgezondheid (RIV), Bilthoven, The Netherlands. The hematocrit (packed cell volume) was determined by a micro-method with two different centrifuges.

Control blood samples were obtained by dividing fresh donor bloods [6] into samples of approximately 25 ml in glass screw-topped flasks. The samples were stored at 4°C. The hematological parameters were determined twenty times in the Coulter Counter. The mean values of the parameters were taken as reference values and the range within about twice the standard deviation was regarded as the acceptable one. Each day, a flask was thoroughly mixed and measured after every twenty patient samples. Every fortnight a new series of samples was prepared.

System description

Hardware configuration. HEMAsys is based on the INTEL 8008 microprocessor and comprises several modular MCS-8 microcomputer modules (Fig. 1): a central processing unit (CPU) module, a read only memory (ROM)



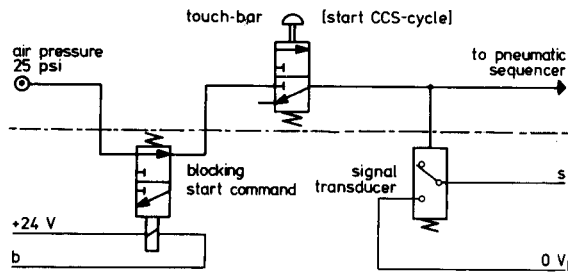


Fig. 1. Hardware configuration of the microcomputer system HEMAsys II.

Fig. 2. Start circuit modification for the Coulter Counter S.

module, a random access memory (RAM) module, and a special input/output (I/O) module which interfaces to the Coulter Counter, a teletype for data output and a specially designed console. The interface with the Coulter Counter takes care of signal conversions, but this I/O module also connects the microcomputer to a circuit added to the Coulter Counter by which the start signal can be controlled externally (Fig. 2). This modification enables the microcomputer to prohibit a start in certain circumstances and is essential for the synchronization of identification numbers with the respective results. For interfacing with the teletype, a universal asynchronous receiver transmitter (UART) was chosen. On the teletype, the Coulter Counter results are printed in duplicate on self-adhesive labels, one for the report form and one for the laboratory file; the lay-out of each label is similar to that shown in Fig. 4(b) (without the final TOT:012). Finally the module interfaces to a specially designed console, which is provided with a numeric keyboard for the input of the three-digit identification numbers indicated on Fig. 4(b) as KMG*. A LED (light emitting diode) display serves for visual check-out of the identification number (six-digit numbers are optional), and this console also provides LED displays for the presentation of the four quality-control parameters, *LEU*, *ERY*, *MCV* and *MCHC* (see below). A more detailed technical description of the microprocessor system will be published elsewhere [9].

Procedure. The functions of HEMAsys II are summarized in Fig. 3. To start a measuring cycle, a three-digit identification number is keyed-in at the console. The number is presented on the LED display for check purposes. As soon as the number is completed, the start key of the Coulter Counter is unlocked. A sample tube is then positioned at the sample intake and a start command given. Immediately after sample intake, the identification number is incremented by one automatically: usually the number of the next sample. If a different number is desired, the display can be cleared by the special key on the console. If the identification number is not complete, the key for the start is locked again. As soon as transmission of the result from the first sample

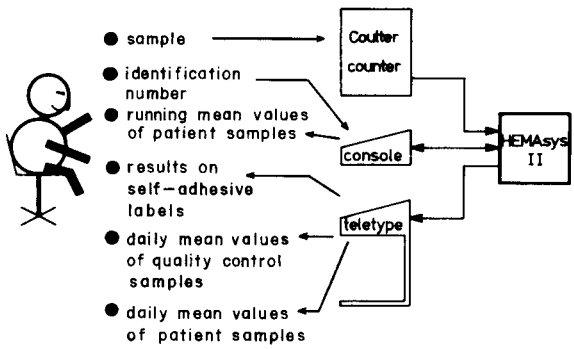


Fig. 3. The functions of HEMAsys II.

a.

G#	LEU	07.65	+175	-35	ERY	4.723	+202	-08
07	MCV	088.7	+199	-11	MCHC	20.45	+200	-10

b.

KMG#	LEU	02.45	ERY	4.216	HB	08.04	HT	.3745
07	MCV	088.2	MCH	1903	MCHC	21.26	TOT:012				

Fig. 4. Example of the presentation of: (a) the mean values of *LEU, ERY, MCV* and *MCHC* of patient samples, the total number of these samples within the truncation limits and the number of rejected values; (b) the mean values of *LEU, ERY, HB, HT, MCV, MCH* and *MCHC* of quality control samples measured during the day (in this case 12).

is completed, the data are printed out in duplicate on the self-adhesive labels on the teletype, along with the identification number. There are features for handling special samples, such as blank samples identified by number 000, quality-control samples identified by number 999, and wash solutions. There is a special switch on the console to allow for washing; all results are rejected if this switch is in the wash position, without loss of the identification of other samples.

Several aids for quality control are implemented. For zero-level process control on the *LEU, ERY, MCV* and *MCHC* results of the patient samples, the following operations are carried out. (a) The occurrence of extreme values is signalled at one of the LED lamps on the specially designed console (for truncation limits, see Table 2). (b) The arithmetic (running) mean of the ten last measured patient samples — except for extremes — is calculated in real time and presented on the LED displays of the console for each type of result. (c) Deviations of these mean values are signalled on the LED displays (for the action limits, see Table 2). In such cases, a quality-control sample has to be run to see if there is any malfunction of the Coulter Counter.

For day-to-day quality control, the system is expanded with options for calculation and printing of: (a) the daily mean of the *LEU, ERY, MCV* and *MCHC* of patient samples, ignoring the values outside the truncation limits, as well as the total number of patient samples included in these calculations and the number of extremes rejected; and (b) the daily mean of *ERY, LEU, HB, HT, MCV, MCH* and *MCHC* of the quality-control samples used during the day. Figure 4 shows the lay-out of the presentation of these parameters.

TABLE 2

Patient sample values with (fixed) truncation limits and the number of extremes and running mean values with (fixed) action limits and the number of signals (samples exceeding the action limits), as recorded on February 8, 1977

Parameters		Patient sample values				Truncation limits	No. of extremes
		n	Mean	S.d.	C.v. (%)		
LEU	($\times 10^9$ l^{-1})	292	8.22	3.88	47	3.0—18.0	23
ERY	($\times 10^{12}$ l^{-1})	292	4.66	0.72	15	2.50—6.70	6
HB	(mmol l^{-1})	292	8.62	1.22	14		
HT	(l^{-1})	292	0.410	0.057	13		
MCV	(fl)	292	88.1	5.8	7	72—106	10
MCH	(amol)	292	1850	138	8		
MCHC	(mmol l^{-1})	292	21.0	0.45	2.1	19.5—22.5	3
Parameters		Running mean values				Action limits	No. of signals
		n	Mean	S.d.	C.v. (%)		
LEU	($\times 10^9$ l^{-1})	269	8.01	1.18	15	5.6—9.8	17
ERY	($\times 10^{12}$ l^{-1})	286	4.67	0.31	7	3.90—5.20	13
HB	(mmol l^{-1})						
HT	(l^{-1})						
MCV	(fl)	282	88.0	1.9	2.2	83—93	21
MCH	(amol)						
MCHC	(mmol l^{-1})	289	20.9	0.25	1.2*	20.5—21.6	13

Finally, it is possible to obtain a punched tape of the results so that the data can be processed on a separate (micro)computer system, which has been done for several parameters (see below).

RESULTS

Variation of the parameters of patient samples

Table 2 indicates the behaviour of the parameters on an arbitrarily chosen day. The truncation limits are calculated from the mean (±3 s.d.) of all patient samples during a period of several weeks, except for the *LEU* value for which the limits are established more or less arbitrarily. The action limits for the running-mean values are calculated from the mean (±2 s.d.) of the accepted patient samples during a period of several weeks.

The variation, on the same day, of the running-mean values of the *LEU*, *ERY*, *MCV* and *MCHC* values is shown in Fig. 5. The *LEU* value tends to decrease during the day, which may probably be ascribed to the fact that most out-patient samples, with in general more normal values, are analysed in the afternoon. Moreover, there are relatively large variations in the mean

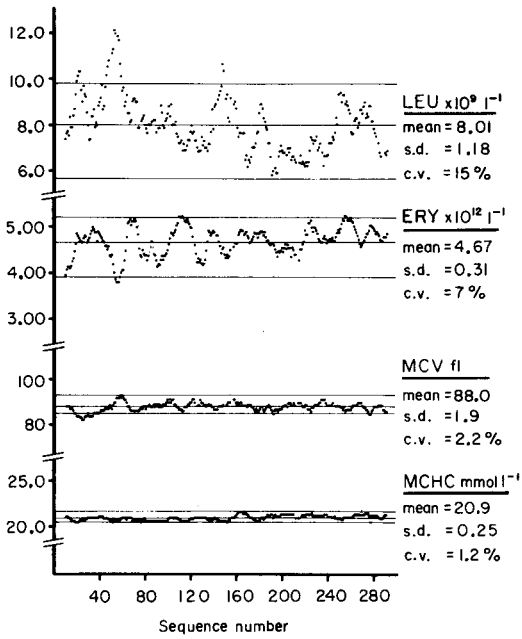


Fig. 5. The variation of the running mean values of the *LEU, ERY, MCV* and *MCHC* values relative to the action limits on February 8, 1977.

values for the erythrocytes and leucocytes during the day, so that small deviations in these parameters cannot be detected. Accordingly, neither parameter is suitable for checking accuracy.

However, the values for the *MCHC* are remarkably constant during the day. As the *MCHC* is calculated by the Coulter Counter from the erythrocyte count, the *MCV* and the hemoglobin concentration, variations in any of these parameters are easily detected from a change in the mean values of the *MCHC*. The variation in the mean values of the *MCV* for the last ten patient samples is usually larger than for those of the *MCHC*.

Long-term drifts of the instrumentation may be detected by comparing the daily mean of all patient samples analyzed each day. The *LEU*, *ERY*, *MCV* and *MCHC* values for a period of 75 days are shown in Fig. 6. The *LEU* and *ERY* values are fairly constant, but the *MCV* and *MCHC* values show better consistency. It can be concluded that the running mean values for the *MCV* and the *MCHC* are applicable for daily quality control and that long-term quality control can be based on the daily mean of the *LEU*, *ERY*, *MCV* and *MCHC* values in patient samples.

Variation of the daily mean of the control samples

In the currently applied quality-control samples, the leucocytes are not stable at room temperature for 8 h, hence this parameter cannot be tested properly with these samples. The other parameters are remarkably constant

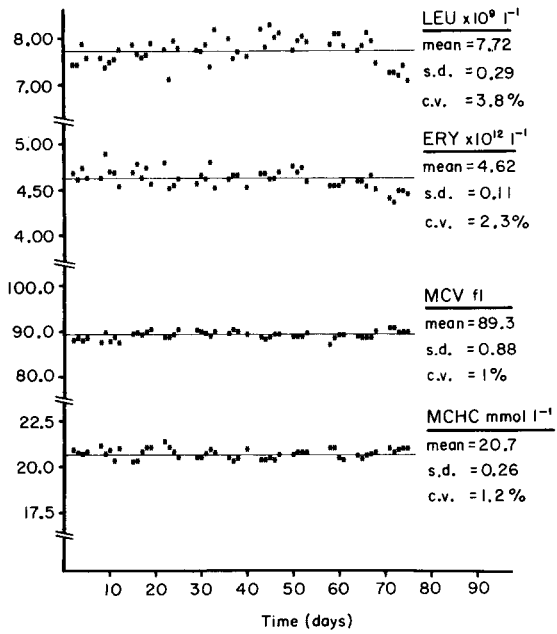


Fig. 6. The variation of the daily mean values (n = 50) of patient samples during February, March and April 1977.

if the control samples are stored at 4°C and thereafter at room temperature during the day [6]. As noted in the Experimental part, quality-control samples are measured after each batch of 20 patient samples. The results of these control samples must lie within the calculated range, otherwise the measurement must be repeated. The technician must act if the values are again outside the limits. From the daily mean of these parameters even small long-term drifts of the Coulter Counter can be detected; some examples are given in Fig. 7.

DISCUSSION

The two HEMAsys instruments were introduced in the laboratory in 1975 and 1976. Their flexibility has allowed several improvements to be introduced; both systems function very satisfactorily in the laboratory organization. The data-processing functions performed by the microcomputer system are helpful in the hematological sublaboratory, particularly the direct visual check-out of identification numbers. These functions will be indispensable when an on-line connection is made with the Hospital Information System.

The diurnal variation of the erythrocyte indices over the last ten patient samples is very small (Fig. 5). According to Rutten et al. [7], this is also true for the hemoglobin concentration if a distinction is made for samples from males and females. As the microprocessor has no patient file, the quality-

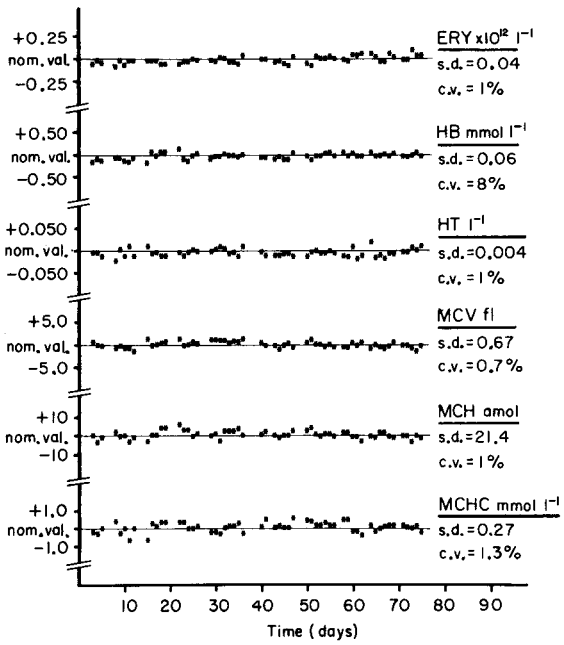


Fig. 7. The variation of the daily mean values (n = 56) of quality-control samples during February, March and April 1977.

control system was limited to the erythrocyte indices of patient samples. This is permissible because variations in the number of red blood cells, the hemoglobin concentration and the hematocrit are reflected in variations in one or more of the indices.

The use of quality-control samples makes it possible to check the other parameters except for the white blood cell count. These samples, obtained from fresh donor blood, were stable during the day at room temperature, although the coefficient of variation for the determinations during a particular day may vary from day to day. Therefore, if the values of one of the quality-control samples measured during a particular day are outside the limits, the measurement is repeated with a different bottle of donor blood. Only when both determinations are outside the limits and agree with each other, is malfunction of the Coulter Counter probable.

As has been demonstrated by others [10], the Coulter Counter S is very stable from day to day (cf. Figs. 5—7). The methods of detecting long-term drift described above are very sensitive. In conclusion, quality control both during a day and from day to day is possible with the aid of the relatively simple and inexpensive microprocessor-based system described, which has proved to be very helpful in routine work.

We are very grateful to H. A. G. Lagas-Smit and R. Brouwer, in cooperation with whom the specifications of HEMAsys were set up; to R. Veldkamp,

F. van den Dool and J. H. Meijerink, students at the Delft University of Technology, who each developed part of the program; to J. C. van Zwam and the Central Research Workshop of Erasmus University for construction of the system and to J. M. Pekelharing for help in preparing this paper.

REFERENCES

1 D. R. Prangnell and P. H. Johnson, J. Clin. Path., 29 (1976) 955.
2 S. M. Lewis, J. Clin. Path., 29 (1976) 955.
3 N. J. Verhoef and H. Daniels, in press.
4 E. J. van Kampen and W. G. Zijlstra, Clin. Chim. Acta, 6 (1961) 538.
5 Coulter Diagnostics, Inc., Information Sheet, June 10, 1976.
6 I. Cavill and A. Jacobs, Association of Clinical Pathologists, Broadsheet 75, 1973.
7 W. P. F. Rutten, R. J. H. Scholtis, N. A. Schmidt and R. J. M. van Oers, Z. Klin. Chem. Klin. Biochem., 13 (1975) 395.
8 P. R. Gilmer, L. J. Williams, J. A. Koepke and B. S. Bull, Am. J. Clin. Path., 68 (1977) 185.
9 P. A. Mantel, Euromicro Journal (1979), in preparation.
10 G. M. Brittin, G. Brechner and C. A. Johnson, Am. J. Clin. Path, 52 (1969) 679.

Analytica Chimica Acta, 112 (1979) 185—192
Computer Techniques and Optimization

COMPUTERIZED POTENTIOMETRIC ANALYSIS
Part I. Processing of Acid—Base Titration Curves without Inflexion Points

GUY NOWOGROCKI, JOËL CANONNE and MICHEL WOZNIAK*

Laboratoire de Physico-chimie des Solutions, Ecole Nationale Supérieure de Chimie de Lille, BP 40, 59650 Villeneuve D'Ascq (France)

(Received 20th October 1978)

SUMMARY

The automated titration system described consists of standard potentiometric apparatus and a laboratory medium-capacity computer. The computer controls the reagent additions, reads the delivered volume and the potentials, processes the data, and prints or displays the analytical results. Any parameter of the titration can be determined by multiparametric refinement. The system was tested on neutralization curves without inflexion points obtained in titrations of potassium sodium tartrate and of mixtures of fumaric and maleic acids.

Automatic potentiometric titrators have two functions: collection of experimental data and processing of these data to obtain the analytical results. Although the importance of the first task is well understood, the means used to carry out the second are often far from satisfactory. Acid—base titrations are usually founded on the existence of inflexion points on neutralization curves. The equivalence volumes are frequently obtained by titration to a predetermined potential, by plotting of first or second derivatives, or by approximate calculation of inflexion points. Several titrators, more or less automate, based on these techniques have been described [1] or are commercially available (Mettler, Radiometer, Tacussel—Solea). However, these methods are inefficient in the case of ill-defined or absent inflexion points. Unhappily, such cases are frequently encountered in the analysis of weak acid mixtures, salts and dilute solutions or in the presence of acidic impurities. Moreover, it is well known that the inflexion points do not necessarily coincide with the true equivalence points [2, 3]: this can lead to erroneous results, even in the determination of a single acid, unless standardization is done carefully.

An improvement can be brought about by several mathematical procedures which, by linearization of the titration curves (or parts of the curves) or by least-squares refinement, lead to the true concentrations or the actual equivalence points. Multiparametric curve-fitting is valuable [4, 7]. Most of these methods, reviewed elsewhere [6, 7], suffer from severe drawbacks; they can be applied only to relatively simple cases and none is sufficiently general for processing all titrations based on pH measurements. Furthermore, the calculations must

be performed independently of the collection of experimental data. In previou papers [6, 7], a very general program, MUPROT, which can be applied to a wide variety of acid—base titrations was described. Use of a laboratory computer of 16 Kbytes memory makes it possible to automate fully the collection of titration data, and application of a modified MUPROT program leads to the output of analytical results immediately the titration ends.

EXPERIMENTAL

Instrumentation

The instrumentation is shown schematically in Fig. 1. A Hewlett-Packard 9825A calculator (16 Kbytes memory) was used to control the system, from the collection to the processing of the experimental data. The read-only memories 98210A (string and advanced programming) and 98212A (general I/O and plotter) make the programming and editing of results easier. A 9862A plotter or a 9871A printer can be used as output peripheral.

Potential or pH values were measured by a Radiometer PHM pH 64 meter w a resolution of 10^{-4} V or 10^{-3} pH units. This meter was connected to the calculator via a 98033A BCD interface. A 98032A 16-bit interface provided the volume readings and the control of stepwise addition of reagent from a piston burette (Tacussel type EBX 1, with NUMEP option). The resolution for a 5-ml syringe is 10^{-3} ml.

The titrations were done in a box thermostated at 25 ± 1°C. The solution to be titrated was maintained at 25.00 ± 0.05°C in a Tacussel RMO6 titration vessel by means of a water bath; magnetic stirring was used. The electrodes used were a Jena type N glass electrode and an Ingold type 303 reference calomel electrode. All the measurements were made in a 1 M KCl medium and to ensure conservation of carbon dioxide, in a closed vessel [7].

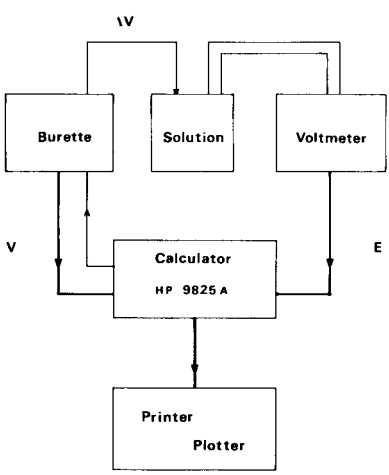


Fig. 1. Block diagram of the system.

Process control and data collection

The flow diagram shown in Fig. 2 summarizes the software used during a titration. The first step is the input of experimental parameters by means of the keyboard in conversational mode: essentially the maximum volume of added reactant v_m, the waiting time T, and the maximum voltage change allowed, D_m, in mV min^{-1}. The second step is the automatic titration. For any experimental point, the computer reads the potential (in fact, an average value of ten voltage readings taken 0.5 s apart). This value, E_i, is compared to the two preceding values, E_{i-1} and E_{i-2}, to evaluate the potential changes $d_1 = (E_i - E_{i-1})/T$ and $d_2 = (E_i - E_{i-2})/2T$. If d_1 or d_2 is greater than D_m, a new reading is taken after a waiting time T (usually about 15 s). When the stability criterion is fulfilled, the volume of titrant added is read: this volume and the last value of potential constitute an "experimental point".

If the volume read is smaller than v_m, the computer orders addition of a new increment Δv of titrant and starts the potential readings for the next point. After the maximum volume has been attained, the titration is ended and the experimental points (v, E) are stored on magnetic tape for subsequent processing. This mode of data collection ensures flexibility and the absence of anomalous potential values. Furthermore, waiting for a quasi-stability permits use of thermodynamic relations and so, a more rigorous mathematical treatment.

Obviously, this approach can be applied to any type of titration and any means of detection (amperometry, spectrophotometry, etc.).

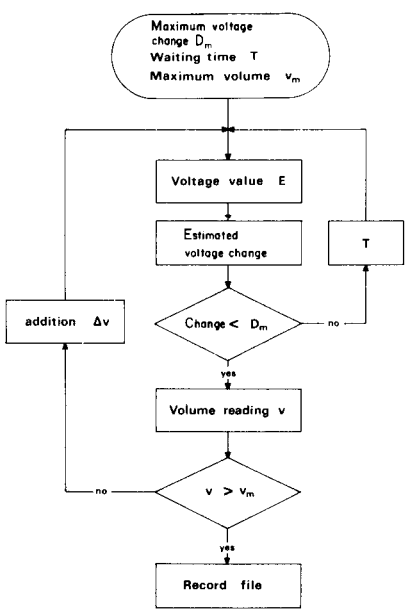


Fig. 2. Flow chart of the data collection program.

Data processing

Under fixed experimental conditions, i.e. at known constant temperature and ionic strength, the hydrogen ion concentration is related to the potential by $-\log [H] = E_0 + E_{ja} [H] + E_{jb} K_w [H]^{-1} + SE$. In the most general case, the added volume of reactant can be calculated from the equation

$$v_c = v_0 \{H_i^0 + \sum_{i=1}^{D} C_i^0 (N_i - \bar{n}_{Hi}) - [H] + K_w [H]^{-1}\} /$$

$$\{-H_x^0 - \sum_{i=1}^{B} C_x^0 (N_x - \bar{n}_{Hx}) + [H] - K_w [H]^{-1}\}$$

which is valid for all kinds of acid—base titrations. In these equations, D and B represent the initial and added species, and N_i and N_x the number of protons bound in the initial and added species; $\bar{n}_{Hi}$ and $\bar{n}_{Hx}$ relate to the acidity constants. Thus, there is an established relation between the added volumes of reactant v_c and the potentials E, in terms of strong acid concentrations H_i^0 and H_x^0, weak acid concentrations C_i^0 and C_x^0, ionic product of water K_w, and electrode characteristics (zero shift E_0, electrode slope S and liquid junction coefficients E_{ja} and E_{jb}). These formulae have already been proven [6, 7].

The unknown parameters appearing in these equations are determined by least-squares refinement by minimizing the sum $\Sigma_{n=1}^{N} W_n (v_n - v_{nc})^2$ obtained for n experimental points; here v_n and v_{nc} are the experimental and calculated values of the added volume for the nth point, and W_n is a weighting factor of the form $W_n = 1/[\sigma_{0v}^2 + \sigma_{0E}^2 (\partial v_{vc}/\partial E_n)^2]$, which takes into account the experimental errors in volume and potential. Here, the terms σ_{0v} and σ_{0E} were taken as the resolutions of the burette and of the voltmeter, respectively.

Starting from the approximate values for the parameters to be refined, the MUPROT program estimates the best values, i.e. the best fit to the experimental curves.

EXAMPLES OF APPLICATIONS

Salt impurity

This example describes the determination of the purity of potassium sodium tartrate by titration with hydrochloric acid. The neutralization curve (Fig. 3) shows no inflexion point, thus the concentration cannot be obtained directly. Figure 3 also shows the output obtained on the plotter: a table of values, the experimental points, the computed "best" curve and a distribution curve of the residuals. Results must be read as described previously [6, 7]. The total tartrate concentration found corresponds to a purity of 100.2% for $KNaO_6C_4H \cdot 4H_2O$ with a 0.6% confidence range (3σ); the high result is probably due to slight dehydration. In another refinement, searching for an excess of strong acid led to a strong initial concentration $H_i^0 = -3 \times 10^{-5}$ M ($3\sigma = -6 \times 10^{-5}$ M): this could be interpreted as a slight excess of alkali (0.2%), but the confidence range was too large for such a possibility to be accepted.

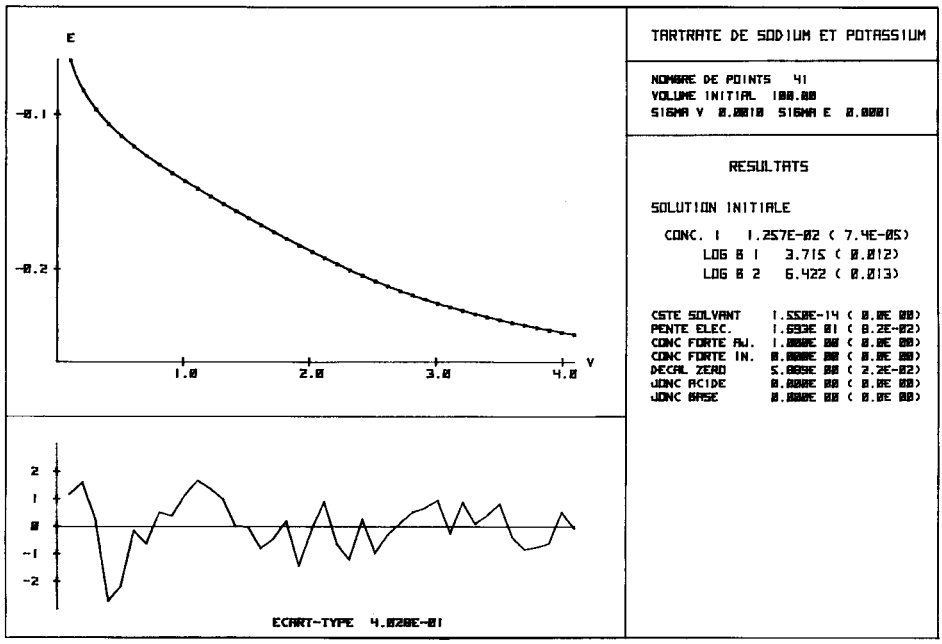


Fig. 3. Output on the plotter for salt analysis. Zero values in brackets indicate that the parameter was fixed in the refinement; refined values comport confidence ranges taken as 3σ.

As can be seen in Fig. 3, in such a simple case, it is not necessary to know precisely the acidity constants; they can be refined simultaneously with the other parameters. Such values of the constants might be useful in establishing the nature of the species titrated.

Refinement of the characteristics of the electrodes, i.e. the slope of the pH response curve and the zero shift (formally identical to a standard potential), makes it unnecessary to standardize the voltmeter previously, and thus allows complete automation of the process. The value of the standard deviation gives an estimate of the validity of the refinement. A weighted residual representing the distance from an experimental point to the calculated curve $v_c/\sigma_{0v} = f(E/\sigma_{0E})$, the standard deviation is usually close to unity in a correct refinement: the value here is 0.4, which indicates that the estimates of σ_{0v} and σ_{0E} (10^{-3} ml and 10^{-4} V, respectively) were too pessimistic. The distribution of the residuals is satisfactory, being randomly distributed about zero. That means that there are no systematic errors and that the refinement constitutes a correct interpretation of the experimental data. If the distribution is unsatisfactory, the adjusted parameters do not fit the experimental curve adequately, and the refinement must be resumed after correction of the initial hypothesis.

Figure 4 represents the change in potential, its first and second derivatives versus volume added, and the position of the true equivalence point determined by computation (v_e = 2.51 ml). The second derivative does not change sign and thus cannot be used to detect an equivalence point. The third deriva-

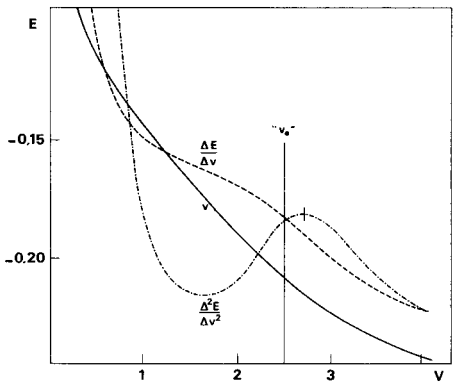


Fig. 4. Titration of potassium sodium tartrate: variation of the potential and of its first and second derivatives (arbitrary scale).

tive will obviously show a zero point but the fallacious use of such a curve would entail more than 10% errors in the analytical results.

The duration of a neutralization depends on the number of experimental points necessary which is a function of the difficulty of the analytical problem, i.e. of the number of parameters to be refined, and on the stabilization time of the glass electrode which is about 30 s for an electrode of low resistance (100 MΩ). In this example, collection of the 41 experimental points lasted about 20 min but could be abbreviated by diminishing the number of points. The refinement and output of results takes 2—4 min.

Mixture of dibasic acids

Another common and significant example is the analysis of a mixture of compounds with close acidity constants. Tests were made with two dibasic acids, fumaric acid (pK_1 = 2.801; pK_2 = 3.922) and maleic acid (pK_1 = 1.724; pK_2 = 5.544). Figure 5, curve a, represents the titration curve for a synthetic mixture containing 79.8% fumaric acid and 20.2% maleic acid (expressed in moles) with potassium hydroxide. There is a single inflexion point corresponding to neutralization of the four protons. Curves b and c represent the titration of the individual acids at a concentration of 10^{-2} M; these titrations were used to determine the acidity constants and the concentrations of the stock solution The refinement, in which the overall acidity constants of fumaric, maleic and carbonic acid (9.63 and 16.02) were fixed, was applied to the 34 experimental points.

A fumaric acid concentration of 7.997×10^{-3} M, i.e. 79.93%, was obtained with a relative confidence range of 0.4%. For maleic acid, the result was 2.008×10^{-3} M, i.e. 20.07% and the confidence range was 4×10^{-5} M, which is acceptable considering the concentration. The concentrations were thus correctly computed to within 0.2% and 0.6% of the expected values, respectively. In addition, the carbonate concentration in the added base (4.15×10^{-3} M) and the electrode characteristics were also obtained.

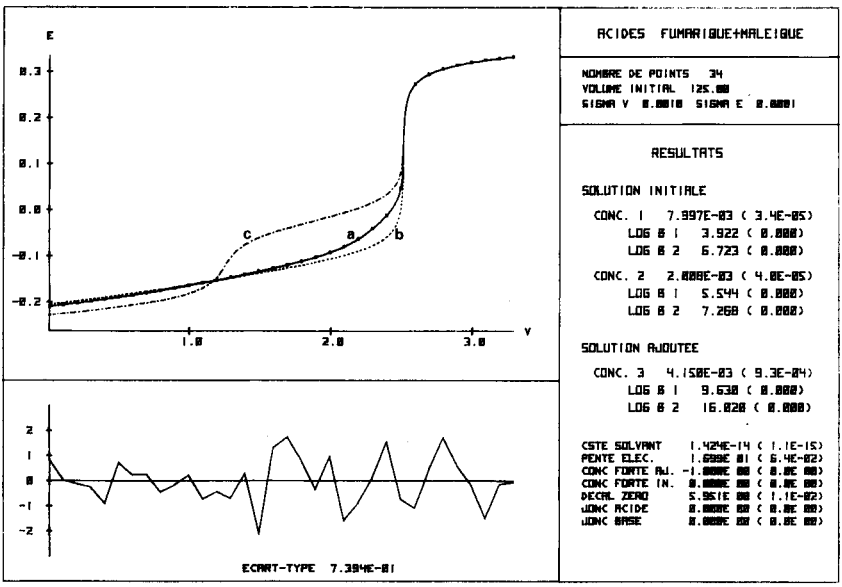


Fig. 5. Output on the plotter for acid mixtures. (a) Mixture; (b) fumaric acid; (c) maleic acid.

Conclusion

This method of acid—base potentiometric titration permits the simultaneous determination of all the unknown parameters and especially the concentrations of the constituents. It is the only method which can successfully handle analysis of protolyt mixtures, even the more tough ones. Besides classical potentiometric apparatus, the method requires only a medium-capacity computer (16 Kbytes or even less). For improvement of data presentation, several peripherals can be connected (plotter, printer, etc.); the software has been written for these different forms of output.

The two programs (data collection and parameter adjustment) are described separately, which is appropriate for research work or exploratory experiments. However, for routine analysis, a single program is available: all the useful parameters are introduced before the titration (or can be read from a magnetic tape) which then proceeds automatically. Such automated analysis with multiparametric refinement can solve many difficult cases with very little demand on working time, and should be generalized to cover other experimental methods to improve extraction of the information available in the measurements. The existence of low-priced personal computer systems with capacities up to 48 Kbytes makes the prospect even more attractive.

REFERENCES

1 S. Ebel and A. Seuring, Angew. Chem., 89 (1977) 129; Angew. Chem., Int. Ed. Engl., 16 (1977) 157.

2 L. Meites and J. A. Goldman, Anal. Chim. Acta, 29 (1963) 472.
3 W. Lund, Talanta, 23 (1976) 619.
4 D. M. Barry and L. Meites, Anal. Chim. Acta, 68 (1974) 435; 69 (1974) 143.
5 D. Murtlow and L. Meites, Anal. Chim. Acta, 92 (1977) 285.
6 G. Nowogrocki, J. Canonne and M. Wozniak, Bull. Soc. Chim. France, 5 (1976) 1369.
7 M. Wozniak and G. Nowogrocki, Talanta, 25 (1978) 633, 643.

Analytica Chimica Acta, 112 (1979) 193—198
Computer Techniques and Optimization

A QUANTITATIVE METHOD FOR FOLLOWING THE PRECIPITATION OF SLIGHTLY SOLUBLE SALTS OF POLYPROTIC WEAK ACIDS

B. PURGARIĆ*

Laboratory for Precipitation Processes, "Rudjer Bošković" Institute, Zagreb (Yugoslavia)

Z. TUTEK

Numerical Center of Mathematical Institute, University of Zagreb (Yugoslavia)

(Received 13th November 1978)

SUMMARY

A quantitative method of following the precipitation of slightly soluble salts of polyprotic weak acids is described; the method is based on a computer program designed to simulate the precipitation at constant pH. The program simulates the simultaneous formation of a number of precipitates of different composition which can be slightly soluble salts of the weak polyprotic acid H_nA ($n \leqslant 5$) and/or hydroxides of the metallic cations of groups I, II and III of the periodic table, provided that the molar ratio of the ions involved is constant during precipitation and that hydrogen or hydroxyl ions are liberated in the process. The program takes into account the probability of the formation of soluble complexes of the type $BH_{n-i}A$ as well as metal hydroxo complexes.

Many industrial and biological solutions are extremely complex, containing a large number of ionic species in metastable or stable equilibrium. It can be difficult to follow the kinetics of precipitation (or dissolution) of one or more salts from such solutions quantitatively. If, as a consequence of the precipitation (or dissolution) process, hydrogen or hydroxide ions are liberated, experiments at constant pH values may be the best choice because one of the most important parameters in the precipitation system is kept constant, and because the amount of hydrogen or hydroxide ions required to keep the pH constant is directly related to the amount of precipitate formed.

In this paper, a computer program is described which simulates the simultaneous precipitation (or dissolution) of a number of slightly soluble salts of different composition, e.g. salts of weak polyprotic acids H_nA ($n \leqslant 5$) and/or hydroxides of cations of groups I, II and III of the periodic system, at constant pH. The program calculates the distribution of all ionic species, including the amount of liberated hydrogen or hydroxyl ions, at any given time during precipitation (or dissolution). From the amount of hydroxide or hydrogen ions added in an experiment at constant pH, it is possible to calculate the quantity of precipitate(s) formed (or dissolved) at any given time.

---

*Present address: "Nikola Tesla - Telecommunications", Moskovska 45, Zagreb, Croatia, Yugoslavia.

The method and the program were tested with calcium monohydrogenphosphate dihydrate as a model.

THE COMPUTER PROGRAM

Determination of the distribution of ionic species

The first part of the program computes the distribution of ionic species in an electrolyte solution which consists of one polyprotic weak acid H_nA ($n \leqslant 5$) and any number of cationic components B of groups I, II and III of the periodic system. The program takes into account all possible relationships between the anions of the polyprotic acid ($H_{n-i}A^{i-}$) and free cations (B^{z+}), which result in the formation of soluble complexes of the type $BH_{n-i}A^{(i-z)-}$ (Fig. 1.). The possible participation of soluble hydroxo complexes of the type BOH, $B(OH)_2$ and $B(OH)_3$ is also taken into account.

The program is based on the mass balance equations shown below:

$$[H^+]\cdot[H_{n-i}A^{i-}]/[H_{n-(i-1)}A^{(i-1)-}] = K(i)$$

$$[B^{z+}]\cdot[H_{n-i}A^{i-}]/[BH_{n-i}A^{(i-z)-}] = BK(i)$$

$$[B^{z+}]\cdot[OH^-]/[BOH^{(z-1)+}] = KOH1$$

$$[B^{z+}]\cdot[OH^-]^2/[B(OH)_2^{(z-2)+}] = KOH2$$

$$[B^{z+}]\cdot[OH^-]^3/[B(OH)_3^{(z-3)+}] = KOH3$$

$$AT = \sum_{i=0}^{m}[H_{n-i}A^{i-}] + \sum_{j=1}^{k_1}\{\sum_{i=1}^{m}[B_jH_{n-i}A^{(i-1)-}]\}$$
$$+ \sum_{k=1}^{k_2}\{\sum_{i=1}^{m}[B_kH_{n-i}A^{(i-2)-}]\} + \sum_{l=1}^{k_3}\{\sum_{i=1}^{m}[B_lH_{n-i}A^{(i-3)-}]\} \quad (1)$$

$$BT(r) = [B_r^{z+}] + \sum_{i=1}^{m}[B_rH_{n-i}A^{(i-z)-}] + [B_rOH^{(z-1)+}]$$
$$+ [B_r(OH)_2^{(z-2)+}] + [B_r(OH)_3^{(z-3)+}] \quad (2)$$

Here $K(i)$, $BK(i)$, $KOH1$, $KOH2$ and $KOH3$ are the thermodynamic dissociation constants of the polyprotic acid and its anions, of soluble complexes of the type $BH_{n-i}A^{(i-z)-}$, and of soluble hydroxo complexes of the type BOH, $B(OH)_2$ and $B(OH)_3$, respectively. AT is the total molar concentration of the anionic component and $BT(r)$ is the total molar concentration of the r^{th} cationic component.

The total concentrations of the anionic and cationic components are equal to the sum of the concentrations of all ionic species which contain these components (eqns. 1 and 2). From the above equations, the total concentrations AT and $BT(r)$ can be expressed as functions of the concentrations of undissociated molecules of the polyprotic acid (H_nA) and the concentrations of free cations (B_r^{z+}). Thus

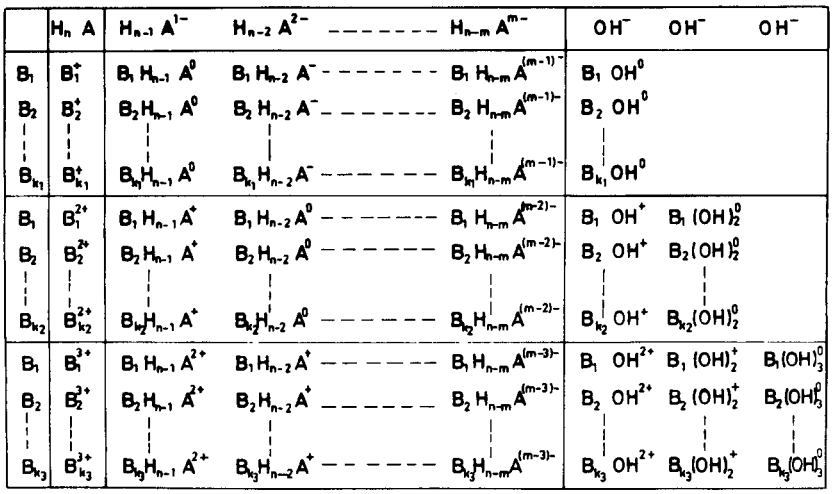


Fig. 1. The ionic distribution taken into account by the program. H_nA, polyprotic acid; $H_{n-i}A^{i-}$, anions of the acid; B_r^{z+}, free cations of the cationic components B_r of groups I ($B_1 - B_{k1}$), II ($B_1 - B_{k2}$) and III ($B_1 - B_{k3}$) of the periodic system; $BH_{n-i}A^{(i-z)-}$, $BOH^{(z-1)+}$, $B(OH)_2^{(z-2)+}$ and $B(OH)_3^{(z-3)+}$ are soluble complexes.

$$AT = [H_nA]\cdot\{1 + \sum_{i=1}^{m}\prod_{p=1}^{i} K_p/[H^+]^i + \sum_{r=1}^{NN}[B_r^{z+}]\cdot\sum_{i=1}^{m}\prod_{p=1}^{i} K_p/BK_{r,i}\cdot[H^+]^i\} \quad (3)$$

$$BT(r) = [B_r^{z+}]\cdot\{1 + [H_nA]\cdot\sum_{i=1}^{m}\prod_{p=1}^{i} K_p/BK_{r,i}\cdot[H^+]^i + [OH^-]/KOH1$$

$$+ [OH^-]^2/KOH2 + [OH^-]^3/KOH3 \quad (4)$$

Here K_p is equal to the dissociation constants of polyprotic acid ($K_p = K(i)$), and NN is the total number of cationic components ($NN = k_1 + k_2 + k_3$). Introducing the concentrations of free cations B_r^{z+} of all the cationic components from equations of the type of eqn. (4) into eqn. (3) gives an expression of the type

$$AT = X\left[P + \frac{BT(1)}{X + R(1)} + \frac{BT(2)}{X + R(2)} + \cdots \frac{BT(NN)}{X + R(NN)}\right] \quad (5)$$

where X is the concentration of H_n A species, and P and $R(r)$ are functions of the pH and the dissociation constants.

The degree of the polynomial developed from eqn. (5) is a function of the number of the components. For a pure polyprotic acid H_nA, the equation is linear; for each cationic component introduced, its degree increases by one.

Introduction of the pH values, the total concentration of polyprotic acid (AT), and the total concentrations of the cationic components — BT(1), BT(2) ... BT(NN) — into eqn. (5) makes it possible to calculate the concentration of the H_nA species (X). As a first approximation, the values of the constants used may be the thermodynamic constants. To enable the program

to use thermodynamic constants, the procedure for solving the equations was enclosed in an iterative loop so that the standard technique can be employed to adjust the values of the constants to those needed for the particular ionic strength of the solution under consideration. The modification of the Debye–Hückel equation described by Davies [1] was used for these adjustments. Iteration is stopped when the difference between two values for the ionic strength reaches a certain value, e.g. 10^{-5}.

Once the concentration of H_nA is known, the program calculates the ionic concentrations of all the other consitutents, ionic strength, activity coefficients, etc.

Simulation of precipitation

The second part of the program simulates the simultaneous formation of a number of precipitates of different composition, i.e. slightly soluble salts of the weak polyprotic acid H_nA and/or hydroxides of any cationic component, provided that their molar ratio is constant during precipitation and that hydrogen (or hydroxyl) ions are liberated in the process. The composition of the jth precipitate of the rth cationic component — $PB(r, j)$ — can be expressed in general as $(B_r)_{P_{r,j}} (H_{n-1}A)_{Q_{r,j,1}} (H_{n-2}A)_{Q_{r,j,2}} \ldots (H_{n-m}A)_{Q_{r,j,m}} (H)_{U_{r,j}} (OH)_{V_{r,j}}$, where $r \leqslant 5$ and $j \leqslant 3$.

The program diminishes, step by step, the total concentrations of all components of which ions are involved in formation of precipitates. The values of such diminutions are functions of precipitate increments in each step, $PBX(r, j)$, which are not limited in size and can be different for various precipitates. In each step, the program calculates the distribution of the ionic species and the total amount of hydrogen ion. From the difference in total hydrogen ion between steps, the amount of hydrogen ions liberated during precipitation can be calculated.

The information output after each step includes the concentrations of all ionic species, their activities, the ionic strength, the degree of precipitation for each precipitate, supersaturations, the amounts of precipitates formed, and the amount of hydrogen ions liberated during precipitation.

The computer program outlined can simulate the simultaneous precipitation (or dissolution after minor modification) of up to 15 precipitates of different composition, i.e. up to 3 precipitates containing up to 5 cationic components.

Description of the computer program

The program is written in FORTRAN V for a Univac 1100 Computer. It consists of a driver program (768 cards) and the subroutines POLM (83 cards) and RACUN (19 cards). The driver program contains all the input and most of the output instructions. It calls POLM whenever a polynomial must be solved. In the normal mode, all input data are read from punched cards. The data are divided into two groups: the first group contains the ten consecutive blocks necessary for the computation of the distribution of ionic species, and the second group contains the five blocks of data necessary for simulation of

precipitation. The program, including detailed descriptions of input and output data, is available on request.

TESTING OF THE METHOD

The method was applied in an examination of the precipitation of calcium monohydrogenphosphate dihydrate. The kinetics of spontaneous precipitation of $CaHPO_4 \cdot 2H_2O$, from aqueous solutions of calcium phosphate which were 0.15 M in sodium chloride, was followed quantitatively at constant pH (pH 5) by a pH-stat device. The chemical events leading to the formation of $CaHPO_4 \cdot 2H_2O$ can be summarized by

$$a\ Ca^{2+} + b\ CaHPO_4^0 + c\ CaH_2PO_4^+ + d\ CaPO_4^- + e\ H_3PO_4^0 + f\ H_2PO_4^- + g\ HPO_4^{2-} + h\ PO_4^{3-} \rightarrow CaHPO_4(s) + (b + 2c + 3e + 2f + g - 1)\ H^+ \quad (6)$$

for $a + b + c + d = 1$ and $b + c + d + e + f + g + h = 1$, where coefficients a—h depend on the pH, the total concentrations and concentration ratios of the reactants, the ionic strength, etc. According to eqn. (6), the pH is decreased by precipitate formation, the amount of liberated hydrogen ions depending on the amount of precipitate formed. Given the amount of sodium hydroxide solution added to keep the pH constant, the precipitation process can be followed quantitatively by the computer program described. The computer simulation was done by using the following thermodynamic dissociation constants: $K(1) = 7.11 \times 10^{-3}$ [2], $K(2) = 6.34 \times 10^{-8}$ [2], $K(3) = 4.3 \times 10^{-13}$ [2], $BK(1) = 3.91 \times 10^{-2}$ [3], $BK(2) = 1.82 \times 10^{-3}$ [3], $BK(3) = 3.45 \times 10^{-7}$ [3].

For comparison, the amounts of precipitate formed at given times in various kinetic experiments (Table 1, experiments 1—13, 15) and at consecutive time intervals (experiment 14) were also determined by chemical analyses. Discrepancies between these results (X_1) and the data calculated by the program (X_2) (from the amount of hydroxide added) were analysed by the Student test. The 95% confidence limits for the average relative discrepancy were $(X_1 - X_2)/X_1 = -0.005 \pm 0.017$ which shows that the discrepancies are not systematic.

DISCUSSION

Extreme care has to be taken in accepting the calculations from such programs. Two of the possible checks which can be applied have been described above. First, the calculated data must be consistent with the input data. A check on this requirement is incorporated into the program as a check on the mass balances; this tests whether the sum of the ion concentrations calculated is equal to the total concentrations of the components entered as the input data. A second check is the comparison of the amount of precipitate determined by chemical analysis with the amount calculated by the computer program.

Although many quantities, each with its range of possible error, are employed by the program, and Debye—Hückel corrections are applied to quite

TABLE 1

Comparison of observed and calculated data for the precipitation of $CaHPO_4 \cdot 2H_2O$ at constant pH. $X_1 = P_0 - 0.5\ (P_1 + Ca_1)$; $X_2 = P_0 - P_2$. P represents the phosphate ion and Ca the calcium ion

Exp. No.	Initial total reactant concentrations[a]		pH	Mother liquor composition[a] determined by				Amount of precipitate[a]		Molar ratio of precipitat (Ca/P)
	P_0	Ca_0		Chemical analyses		Simulation		X_1	X_2	
				P_1	Ca_1	P_2	Ca_2			
1	2.3	2.3	5.04	1.40	1.36	1.38	1.38	0.92	0.92	—
2	2.3	2.3	5.00	1.22	1.17	1.20	1.20	1.105	1.10	1.00
3	2.5	2.5	5.00	1.33	1.38	1.39	1.39	1.145	1.11	1.00
4	2.5	2.5	5.00	1.68	1.67	1.63	1.63	0.825	0.87	1.01
5	2.5	2.5	5.00	1.33	1.26	1.24	1.24	1.205	1.26	—
6	2.5	2.5	5.00	1.49	1.40	1.42	1.42	1.055	1.08	1.01
7	2.6	2.6	5.00	1.48	1.37	1.40	1.40	1.175	1.20	1.00
8	2.7	2.7	5.00	1.40	1.40	1.35	1.35	1.30	1.35	—
9	2.8	2.8	5.05	1.32	1.31	1.34	1.34	1.485	1.46	0.99
10	3.0	3.0	5.00	1.45	1.46	1.50	1.50	1.545	1.50	1.01
11	3.1	3.1	5.00	1.64	1.44	1.56	1.56	1.56	1.54	1.03
12	3.5	3.5	5.07	1.40	1.39	1.47	1.47	2.105	2.03	1.00
13	2.5	2.5	5.00	1.42	1.44	1.39	1.39	1.07	1.11	1.00
14	2.5	2.5	5.00	2.43	2.26	2.29	2.29	0.155	0.21	0.91
			5.00	1.96	1.99	2.00	2.00	0.525	0.50	1.01
			5.00	1.51	1.61	1.60	1.60	0.94	0.90	1.01
			5.00	1.37	1.46	1.44	1.44	1.085	1.06	1.00
			5.00	1.31	1.39	1.40	1.40	1.15	1.10	1.01
15[b]	6.0	1.2	5.01	5.25	0.44	5.27	0.47	0.76	0.73	1.02

[a] $\times 10^{-2}$ mol dm^{-3}.
[b] $X_1 = Ca_0 - Ca_1$; $X_2 = Ca_0 - Ca_2$.

high ionic strengths (0.25 mol dm^{-3}), discrepancies were within the expected experimental error. This indicates that the operation and assumptions of the program are satisfactory.

The authors thank Dr. H. Füredi-Milhofer for helpful discussions. This work was supported partly by N.I.H., Bethesda, Maryland, through the US—Yugoslav Joint Board for Scientific and Technological Cooperation (Grant No. 02-002-01) and partly by the Croatian Organization of Associated Interests for Scientific Research.

REFERENCES

1 C. W. Davies, Ion Association, Butterworth, London, 1962.
2 R. G. Bates and S. F. Acree, J. Res. Nat. Bur. Stand., 30 (1943) 129.
3 A. Chughtai, R. Marshall and G. H. Nancollas, J. Phys. Chem., 72 (1968) 208.

VOL. 112, No. 2 ANALYTICA CHIMICA ACTA June 15, 1979

(Computer Techniques and Optimization, Vol. 3, No. 2)

CONTENTS

Printed in The Netherlands