on molecular code (i.e., sequence of the number of graph paths of different length).

Finally, an index that would be highly selective and could at the same time be successful in correlating many different properties is not yet discovered. Because of this, one must choose for correlation studies an index according to the property and, perhaps, use also other statistical methods besides the standard linear regression.

ACKNOWLEDGMENT

M.R. acknowledges financial support of the Research Community of Slovenia, Ljubljana, and of Ministère des Affaires Etrangères (CIES), Paris.

REFERENCES AND NOTES

(1) Lynch, M. F.; Harrison, J. M.; Town, W. G.; Ash, J. E. "Computer Handling of Chemical Structure Information"; Macdonald: London; American Elsevier: New York, 1971; p 12.
(2) Morgan, H. L. "Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at CAS". *J. Chem. Doc.* **1965**, *5*, 107–113.
(3) Dubois, J. E. In "Chemical Applications of Graph Theory"; Balaban, A. T., Ed.; Academic Press: New York, 1976; Chapter 11, p 333.
(4) Balaban, A. T.; Chiriac, A.; Motoc, I.; Simon, Z. "Steric Fit in QSAR". *Lect. Notes Chem.* **1980**, *15*, 22–39.
(5) Kier, L. B.; Hall, L. H. "Molecular Connectivity in Chemistry and Drug Research"; Academic Press: New York, 1976.
(6) Bonchev, D.; Mekenyan, Ov.; Trinajstić, N. "Isomer Discrimination by Topological Information Approach". *J. Comput. Chem.* **1981**, *2*, 127–148.
(7) Bonchev, D.; Knop, J. V.; Trinajstić, N. "Mathematical Models of Branching". *Math. Chem.* **1979**, *No. 7*, 21–42.
(8) Balaban, A. T. "Chemical Graphs. XXXIV. Five New Topological Indices for the Branching of Tree-Like Graphs". *Theor. Chim. Acta* **1979**, *53*, 355–375.
(9) Razinger, M.; Chrétien, J. R.; Dubois, J. E. "Graphes Chimiques: Génération Automatique des Descripteurs Topologiques". *Vestn. Slov. Kem. Drus.* **1984**, *31*, 211–227.
(10) Wiener, H. "Structural Determination of Paraffin Boiling Points". *J. Am. Chem. Soc.* **1947**, *69*, 17–20.
(11) Gutman, I.; Ruščić, B.; Trinajstić, N.; Wilcox, C. F., Jr. "Graph Theory and Molecular Orbitals. XII. Acyclic Polyenes". *J. Chem. Phys.* **1975**, *62*, 3399–3405.
(12) Hosoya, H. "Graphical Enumeration of the Coefficients of the Secular Polynomials of the Hueckel Molecular Orbitals". *Theor. Chim. Acta* **1972**, *25*, 215–222.
(13) Randić, M. "On Characterization of Molecular Branching". *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.
(14) Bonchev, D.; Trinajstić, N. "Information Theory, Distance Matrix, and Molecular Branching". *J. Chem. Phys.* **1977**, *67*, 4517–4533.
(15) Balaban, A. T. "Highly Discriminating Distance-Based Topological Index". *Chem. Phys. Lett.* **1982**, *89*, 399–404.
(16) Hermann, R. B. "Theory of Hydrophobic Bonding. II. The Correlation of Hydrocarbon Solubility in Water with Solvent Cavity Surface Area". *J. Phys. Chem.* **1972**, *76*, 2754–2759.
(17) Kier, L. B.; Hall, L. H.; Murray, W. J.; Randić, M. "Molecular Connectivity I-III". *J. Pharm. Sci.* **1975**, *64*, 1971–1981; "Molecular Connectivity V". *J. Pharm. Sci.* **1976**, *65*, 1226–1230.
(18) Razinger, M.; Mićović, S. "Structural Selectivity of Topological Codes in Alkane Series". *Vestn. Slov. Kem. Drus.* **1983**, *30*, 199–212.
(19) Randić, M.; Wilkins, C. L. "Graph Theoretical Analysis of Molecular Properties. Isomeric Variations in Nonanes". *Int. J. Quant. Chem.* **1980**, *18*, 1005–1027.
(20) Wilkins, C. L.; Randić, M. "A Graph Theoretical Approach to Structure-Property and Structure-Activity Correlations". *Theor. Chim. Acta* **1980**, *58*, 45–68.

Optimization of a Similarity Metric for Library Searching of Highly Compressed Vapor-Phase Infrared Spectra

MICHAEL F. DELANEY,* JOHN R. HALLOWELL, JR., and F. VINCENT WARREN, JR.†

Department of Chemistry, Boston University, Boston, Massachusetts 02215

Received March 29, 1984

Compound identification by library searching of experimental spectra using instruments based on small computer systems is becoming increasingly common. For successful searching performance, large libraries are needed. Compressed spectra are typically used to increase both the number of spectra that can be stored on a small computer and the search speed. In this study, a similarity metric for matching highly compressed binary intensity vapor-phase infrared spectra is optimized with three distinct approaches. Two of the approaches are generally applicable for library searching performance evaluation. The results of all three approaches are in excellent mutual agreement.

INTRODUCTION

Library searching (LS) remains the method of choice for rapid, on-line computerized compound identification by chemical spectra.[1] Libraries of various types of spectra have been employed, most notably mass,[2] infrared,[3] and nuclear magnetic resonance[4] spectra. In many cases, the only way to accomodate the large numbers of reference compounds needed for a useful library is to greatly compress the library spectra. The size of the library, and its effect on searching speed, can be especially critical when microcomputers are used for LS.[5]

Among highly compressed spectral representations for LS, the most popular has been the use of binary intensity spectra.[6] In this situation, a peak in a spectrum is encoded as a "1", and the absence of a peak as a "0". This approach has been used to advantage in mass spectrometry (MS). For vapor-phase infrared (VPIR) spectrometry, we demonstrated[7] how the information content of the library and the LS performance could be improved by incorporating a measured amount of peak-width information into a binary spectrum.

The comparison metric by which similarity between an unknown spectrum and the reference library members is quantitated is also an important criterion in designing a spectral search system. Various metrics have been used with binary intensity representations of different types of spectrometry. The metrics studied, which of necessity are based on Boolean functions and combinations thereof, have included AND,[8] XOR (exclusive OR),[6,9] AND/OR,[10] EXNOR (exclusive NOR),[11] and XOR/OR.[7] An additional metric, originally introduced by Grotch[8,12,13] and the subject of this paper, is a composite metric: XOR $-\mu$ AND. In this metric, μ is a variable factor that controls the relative contribution of the XOR and AND functions.

We recently developed a procedure for the quantitative evaluation of library searching performance.[14] In this approach, any type of spectrometry, library representation, or

† Present address: Waters Chromatography Division of Millipore Corp., Milford, MA 01757.

0095-2338/85/1625-0027$01.50/0

Table I. Logical Truth Table for the Function XOR − μ AND

	x_j	
x_i	0	1
0	0	1
1	1	$-\mu$

comparison metric can be studied. We applied this procedure to optimize the representation of width-enhanced vapor-phase infrared spectra.[7] In this paper we will use this general performance evaluation technique to optimize the flexible parameter in the composite Grotch metric, XOR − μ AND. We will also demonstrate the optimization of this metric with an independent approach based on a statistical examination of the library searching process.[6]

THEORY

Definition of a Spectral Representation and a Comparison Metric. A spectrum can be treated as a d-dimensional vector:

$$X_i = (x_1, x_2, ..., x_i, ..., x_d) \quad (1)$$

where each x_i is the intensity in spectral (wavelength) channel i. In this case, there are d abscissa resolution elements. We can then envision each spectrum to be a point in d-dimensional space. That is, there are d orthogonal coordinate axes, each one corresponding to a particular wavelength channel. In the case of a binary intensity spectral representation, each spectral element is either "0" or "1".

A comparison metric is a function of the two spectra being compared:

$$D_{i,j} = f(X_i, X_j) \quad (2)$$

where the result, D, is usually a measure of the dissimilarity between the two spectra. If the two spectra are identical, then $D = D_{min}$ and D increases with increasing dissimilarity between the spectra. For binary intensity spectra, the simplest dissimilarity metric uses the exclusive OR function (XOR). The dissimilarlity between two spectra is measured by summing, channel by channel, the result of XOR comparison:

$$D_{XOR} = \sum_{k=1}^{k=d} (x_{i,k} \text{ XOR } x_{j,k}) \quad (3)$$

This is the binary analog of Euclidean distance (least-squares) comparison.

The Composite Grotch Metric. The composite Grotch metric is formulated as

$$D_{i,j} = \sum_{k=1}^{k=d} [(x_{i,k} \text{ XOR } x_{j,k}) - \mu(x_{i,k} \text{ AND } x_{j,k})] \quad (4)$$

This composite dissimilarity metric weights together XOR and AND characteristics by using the weighting parameter μ. When $\mu = 0$, the criterion is solely XOR. As μ increases, the matching tends toward AND character. The effect of this composite metric is to distinguish between the two cases in which a spectral channel for a pair of spectra can match: $x_{i,k} = x_{j,k}$ = "0" and $x_{i,k} = x_{j,k}$ = "1". In the composite metric, these two occurrences of a match are weighted differently, as seen in the logical truth table (Table I).

Statistical Description of Library Searching. A useful representation for the study of LS results is a histogram formed from the dissimilarity values obtained when a test spectrum is compared to each of the N reference spectra. As seen in Figure 1, this histogram tends to be at least approximately Gaussian. Histograms of this type have been employed by Grotch[6] to describe LS for mass spectra and by Delaney and Uden[15] and Warren and Delaney[7] for vapor-phase infrared spectra. This model is presented below in general terms, suitable for any comparison metric or spectral library, and then used to estimate an optimal value of μ.

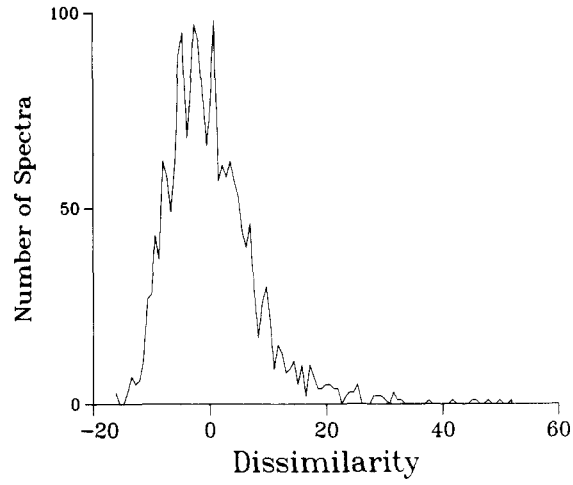

Figure 1. Typical dissimilarity histogram.

For a given dissimilarity metric there will be a minimum value of the metric, D_{min}, obtained when the test spectrum is identical with a library spectrum. Conversely, the largest possible value, D_{max}, is obtained when the two spectra are as dissimilar as is possible. For a given unknown spectrum, the best match from the library will have a metric value of D_1, the Mth closest match will have a value of D_M, and the worst match will have a value of D_N. Each of the N dissimilarity values obtained from the library can be presented in the form of a histogram (Figure 1) in which the number of library members that displayed each dissimilarity value is shown. For this histogram, we can calculate a mean dissimilarity value, $\bar{D}$, and a standard deviation s_D. In many practical cases, this distribution is observed to be monovariate and at least approximately Gaussian.[6,7,14,15]

In the development of an LS system, a choice has to be made of the representation of the library reference spectra and of the specific comparison metric to be used. These would be selected according to the constraints of search speed and storage requirements, but the most fundamental consideration is the degree of success at placing the correct compound on the "hit list". One way of describing a desirable LS system is that such a system would make it as difficult as possible for spectra that do not correspond to the correct compound to occur on the hit list.

A mathematical statement for making it difficult for a random spectrum to appear on the hit list is that the dissimilarity value for the Mth closest match should be as large as possible. However, this needs to be expressed relative to the distribution of metric values observed for the entire library. To optimize an LS system, then, is analogous to maximizing the criterion

$$F = (D_M - D_1)/s_D \quad (5)$$

In the case where the "unknown" test spectrum is drawn directly from the reference library, D_1 will be the metric value for an exact match between two spectra (D_{min}).

For the composite Grotch metric (eq 2), each of the dissimilarity (D) values is a function of the adjustable parameter μ. The optimal value of μ will be the one that maximizes the criterion F.

Quantitative Evaluation of Library Searching Performance (QELS). The quantitative approach for LS performance evaluation, developed in this laboratory, is described in detail elsewhere.[14] In QELS, the performance of any combination of library spectra representation and comparison metric is compared against an LS standard. In the present case, we employ full-resolution spectra and the Euclidean distance

comparison metric as our standard.

This evaluation technique quantiatively compares the LS performance for a compressed library to the performance for the full spectra by using a set of test spectra, which need not be members of the reference library. The approach measures how similar the compressed library hit lists are to the full-spectra hit lists.

In this study a representative test set of compounds has been selected from the standard reference library of N full-resolution spectra. Each test spectrum is searched through the N spectra library to yield hit lists of the M best matching spectra. (N is usually large, and $M = 10$ is typical.) Each test compound is searched through the library of reduced spectra, by use of the comparison metric under consideration, to produce a long search list, of length N, containing each library member.

For each spectrum on the standard search lists, the list index position for the same spectrum on the corresponding long search list is found. Low list index positions indicate a high degree of similarity between the standard search lists and the compressed-spectra search lists. The list index positions are summed to form a raw score. From the best and worst possible raw scores, a normalized figure of merit (FOM) is calculated. The FOM allows different LS alternatives to be compared on a quantitative basis.

EXPERIMENTAL SECTION

Library Reference Spectra. Full-intensity vapor-phase infrared spectra, covering a wide range of compound types, were used. Each raw spectrum consisted of 1842 data points, sampled at 2 cm^{-1} from 4000 to 450 cm^{-1}. Each of the 2000 library spectra was reduced to a 231-dimensional spectrum by a combination of moving and boxcar averaging.[7] The intensities were normalized to be between 0 and 1000 at unit resolution.

Width-Enhanced VPIR Spectra. The width-enhanced representation[7] was derived by using a "peak-picking" algorithm to locate all discernible peaks with intensities greater than 2% of the spectral maximum. For each peak, the base width was determined. A given width-enhanced representation was formed by encoding as "1" a given amount of the peak width. Our 100% library contains no peak width (only the peak center is encoded), while the 50% library encodes half of the peak width. In this study, the 70% library was used, since this was found previously[7] to be the optimal amount of width.

A test set of 15 compounds that span the range of functional group types present in the library was selected. This is similar to approaches employed earlier.[7,14,15]

Computational Details. The 231-dimensional binary intensity compressed library spectra were bit packed into 32-bit words and stored in a random access file. All computations were conducted with FORTRAN-77 programs on a Digital Equipment Corp. VAX-11/730 superminicomputer. The RS/1-PLUS software package (Bolt Beranek and Newman, Inc., Cambridge, MA) was used to examine and present the results.

RESULTS AND DISCUSSION

The value of μ to be used in eq 4 should be chosen to maximize the separation between the correct (best matching) spectrum and the rest of the library. Since the composite metric includes both similarity and dissimilarity character, it seems intuitively reasonable to weight both aspects equally. When two binary intensity spectra are compared, there are four possible combinations. The "no-peaks" combination (0–0) makes no contribution to the composite metric. Since there are two dissimilarity combinations (0–1 and 1–0) and one similarity combination (1–1), an initial estimate of the optimum value of μ would be 2.

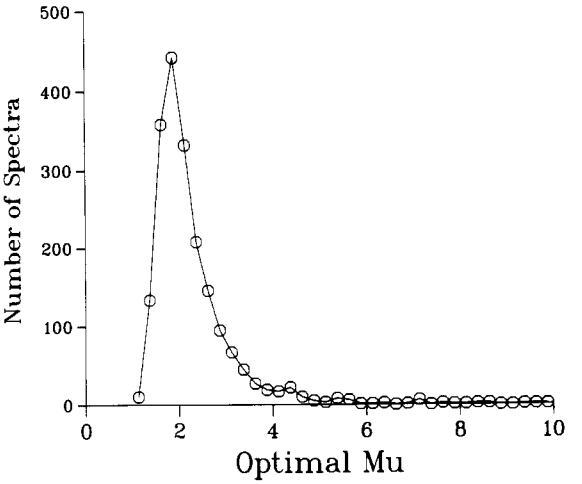

Figure 2. Distribution of optimal μ values (μ^*) generated for the 2000-spectra library with eq 6.

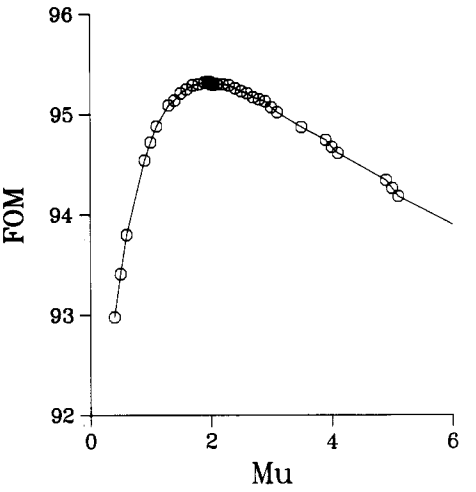

Figure 3. Figure of merit vs. the value of the composite Grotch metric flexible parameter, μ.

A more refined method for selection of μ has been considered for mass spectrometry[8] on the basis of a given test spectrum and the probability of peaks being found in each wavelength channel. The optimum μ for a given spectrum X and a given library was estimated by Grotch to be

$$\mu^* = 1 + \sum_{k=1}^{k=d} p_i(1 - x_i) / \sum_{k=1}^{k=d} x_i(1 - p_i) \qquad (6)$$

where x_i is the intensity, "0" or "1", for the unknown spectrum and p_i is the probability of a "1", in each of the d spectral channels.[6] Using a library of low-resolution mass spectra, Grotch calculated the average μ^* to be in the range of 2.1–2.4, depending on the threshold used to encode the mass spectral peaks into binary form.

By use of eq 6 the μ^* was computed for each of the 2000 spectra in the vapor-phase infrared library. The distribution of μ^* is shown in Figure 2. The median of the μ^* values is computed to be 2.02.

QELS. A test set of 15 compounds and a library of 2000 spectra were used to evaluate the effect of the Grotch composite metric weighting value μ. A total of 39 values of μ in the range of 0–1000 was studied. The figure of merits (FOMs) observed are displayed in Figure 3. When $\mu = 0$, the FOM obtained is, of course, the same as that for a pure XOR metric (88.4%). Also, the FOM reaches a limiting value, equal to that observed for a pure AND metric, for sufficiently large values of μ (88.3%). In this case, the FOM remained constant for $\mu = 100$–1000.

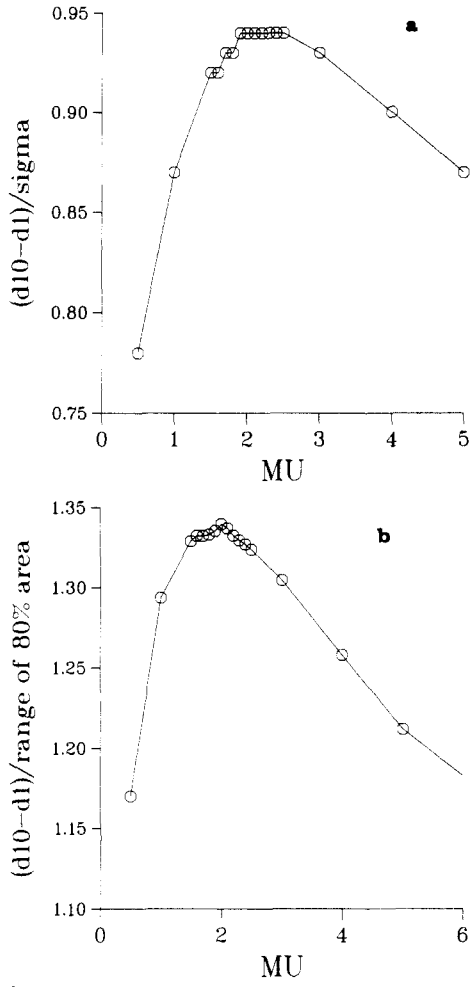

Figure 4. Effect of the value of the composite metric flexible parameter μ on the statistically based F metric: (a) using standard deviation of the histogram (eq 5); (b) using eq 7 with $w_p = 0.80$.

The FOM is seen to change dramatically though systematically with μ. Of particular significance is the observation that the FOM peaked at an optimal μ of 2.0, in agreement with the reasoning and results presented above.

Statistical Modeling. The F metric (eq 5) was also used to monitor the effect of μ on the relative width of the best 10 hit list. As seen in Figure 4, the F metric behaves similarly to FOM and attains a maximum at $\mu = 2.0$. The F metric is still observed to peak at $\mu = 2.0$ when hit lists based on the best two, three, five, or ten spectra are considered.

Since the histograms (Figure 1) are not guaranteed to be Gaussian, the definition of the F metric was modified to use a nonparametric measure of the width of the histograms:

$$F_w = (D_M - D_1)/w_p \quad (7)$$

where w_p is the width of the histogram, centered around the mean, that contains the fraction p of the total histogram's area. The term w_p is in the same units as the dissimilarity scores, D_M and D_1. The effect of μ on the modified statistical metric F_w was determined for various values of w_p. Examples are shown in Figure 4. In all cases for w_p in the range of p = 0.30–0.80, the peak of the curve was found to be at $\mu = 2.0$, once again in agreement with all results presented above.

CONCLUSIONS

Three independent approaches for LS optimization have been used to select the best value for the comparision metric flexible parameter μ. The desire to weight the contributions of similarity and dissimilarity equally and the mathematical form of the comparision metric led to a demonstration that on average this can be accomplished with $\mu = 2$. The second approach was based on selecting the value of μ that gives hit lists that are most like the hit lists obtained for full-resolution spectra compared by a least-squares distance. This also occurred for $\mu = 2$. The third approach used a statistical description of the dissimilarity histograms. The best value of μ was chosen, which made it most difficult for spectra to get onto the hit list. A value of $\mu = 2$ was also found to provide this optimal behavior.

The consequence of $\mu = 2$ being optimal for this composite metric is significant for library searching on small computer systems. The XOR and AND scores for comparing two spectra would be represented in the computer as integers. Multiplying the AND score by exactly 2 can be rapidly accomplished by shifting the AND score left by 1 bit. This would be much faster than computing the composite metric result by floating-point arithmetic—which would be necessary for values of μ that are not even powers of 2.

The exact agreement between the statistically based model and the quantitative evaluation of library searching technique (QELS) is taken as a further indication of the validity of the QELS approach. In all cases to date, the results from QELS have been in agreement with other techniques or with an intuitive understanding. We expect that QELS will continue to be a valuable tool for a variety of LS studies.

ACKNOWLEDGMENT

We thank Bolt Berenek and Newman, Inc. (Cambridge, MA), for providing the RS/1-PLUS software that was used for displaying and analyzing many of the results of this study. This material is based upon work supported by the National Science Foundation under Grant IST-8120255.

REFERENCES AND NOTES

(1) Delaney, M. F. *Anal. Chem.* **1984**, *56*, 261R.
(2) Rasmussen, G. T.; Isenhour, T. L. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 179.
(3) de Haseth, J. A.; Azarraga, L. V. *Anal. Chem.* **1981**, *53*, 2292.
(4) Shelley, C. A.; Munk, M. E. *Anal. Chem.* **1982**, *54*, 516.
(5) Uthman, A. P.; Koontz, J. P.; Hinderliter-Smith, J.; Woodward, W. S.; Reilley, C. N. *Anal. Chem.* **1982**, *54*, 1772.
(6) Grotch, S. L. *Anal. Chem. 42*, **1970**, 1214.
(7) Warren, F. V.; Delaney, M. F. *Appl. Spectrosc.* **1983**, *37*, 172.
(8) Grotch, S. L. *Anal. Chem.* **1971**, *43*, 1362.
(9) Grotch, S. L. *Anal. Chem.* **1974**, *46*, 526.
(10) Rogers, D. J.; Tanimoto, T. T. *Science (Washington, D.C.)* **1960**, *132*, 1115.
(11) Lam, R. B.; Foulk, S. J.; Isenhour, T. L. *Anal. Chem.* **1981**, *53*, 1679.
(12) Grotch, S. L. *Anal. Chem.* **1973**, *45*, 2.
(13) Grotch, S. L. *Anal. Chem.* **1975**, *47*, 1285.
(14) Delaney, M. F.; Warren, F. V., Jr.; Hallowell, J. R., Jr. *Anal. Chem.* **1983**, *55*, 1925.
(15) Delaney, M. F.; Uden, P. C. *Anal. Chem.* **1979**, *51*, 1242.