(16) Dyson, G. M. "Some New Concepts in Organic Chemical Nomenclature". "Subcommittee Report of the Chemical Structure Association", 1983, p D 4/21.

(17) By a "topological transformation" is meant the mapping of one point set (figure) onto another so that the mapping is (1) biunique (each point on the object corresponds to exactly one point of the image) and (2) continuous in both directions (if the distance between two points in the object approach zero, then the corresponding distance in the image will also approach zero).[18] Physically, this is equivalent to saying that an object can be stretched or twisted but not torn or overlapped and joined.

(18) Courant, R.; Robins, H. "What Is Mathematics?"; Oxford University Press: New York, 1941; p 241.

(19) Elk, S. B. "Topologically Different Models To Be Used as the Basis for Ring Compound Taxonomy". submitted for publication in *J. Chem. Inf. Comput. Sci.*

(20) The single adjective "simple" is generally not used in geometry and topology textbooks when referring to figures—as it does not suggest a means of dividing the class of all figures into two disjoint sets: "simple" vs. "nonsimple". However, many geometrical properties, such as closure, connectivity, etc., do allow such a line of demarcation. Thus, any intuitive concept of simple would include simply closed, simply connected, etc. Nevertheless, one important property usually associated with simple figures in geometry—having the boundary set exactly one dimension less than the content set—is not applicable to the chemical model of molecules.

(21) Balaban, A. T. "Valence-Isomerisms of Cyclopolyenes". *Rev. Roum. Chim.* **1966**, *11*, 1097-1116.

(22) Walba, D. M.; Richards, R. M.; Haltiwanger, R. C. "Total Synthesis of the First Molecular Mobius Strip". *J. Am. Chem. Soc.* **1982**, *104*, 3219-3221.

A Priori Estimates of the Elution Profiles of the Pure Components in Overlapped Liquid Chromatography Peaks Using Target Factor Analysis

PAUL J. GEMPERLINE

Department of Chemistry, East Carolina University, Greenville, North Carolina 27834

Received April 5, 1984

Factor analysis is used to detect the presence of overlapping peaks in simulated high-performance liquid chromatography (HPLC) data produced by an ultraviolet/visible (UV/vis) photodiode array detector. The abstract solutions produced by factor analysis are rotated via target tests to produce estimates of the deconvolved elution profiles and the pure spectra of the overlapped peaks. An a priori method of selecting test vectors for target testing is reported here that does not force the elution profiles to fit any predefined functions. Various examples are given to demonstrate the effects of peak separation and random noise on the results. Examples are also given that illustrate the ability of the technique to resolve mixtures of three and four overlapping peaks.

INTRODUCTION

The recent commercial availability of photodiode array detectors for high-performance liquid chromatography (HPLC) has made new and powerful techniques available to detect the presence of overlapping HPLC peaks. One method involves plotting the ratio of two absorbance channels that have been slightly offset in time from each other.[1] This technique works well under some circumstances, but overlapping peaks usually have very similar chemical properties and thus similar spectral properties. In the instance where the spectra are nearly identical, one expects the technique to fail. A more sensitive method involves recording the complete spectrum on the upslope and downslope of the eluting peak.[2] The two spectra are then normalized and superimposed. Any slight mismatch in the two spectra indicates the presence of overlapping or contaminated peaks. Both methods require operator inspection and are not likely to be used routinely. Second, neither technique yields any information concerning the number of contaminating species present.

A form of factor analysis (FA) called principal component analysis (PCA) was first applied to the analysis of chromatographic data by Macnaughtan et al. to deconvolve overlapped mixtures of two components.[3] PCA has since been successfully applied to the detection of overlapping peaks in gas chromatography/mass spectrometry (GC/MS) data.[4,5] Two systems have been described that use PCA to automatically detect overlapping peaks in GC/MS data.[6,7] The technique should be amenable to automation for routine use in conjunction with photodiode array detectors for LC.

Knorr and Futrell were able to separate mass spectra of mixtures by factor analysis when one "pure mass" was available for every component present.[8] Kowalski et al. were able to separate overlapped peaks from GC/MS data for binary mixtures.[9] Malinowski and McCue were able to identify and quantitate mixtures by target transformation factor analysis of the mass spectra.[10] Malinowski and McCue recently reported the adaptation of target transformation factor analysis to detect the presence of overlapped HPLC peaks.[11] The number of components were calculated from analysis of the ultraviolet/visible (UV/vis) spectra of successive fractions collected under overlapped LC peaks. Qualitative analysis was then performed by target testing the spectra of the components suspected to be in the mixture.

This paper describes the application of FA to detect the presence of overlapped LC peaks by using simulated data. After determining the number of overlapping components present, we apply a new method of selecting test factors that model the elution profiles of the unknown mixtures. Unlike previous work, this technique does not require pure spectra of suspected components or wavelengths that exhibit a unique response for one component. Estimates of chromatographic peak shapes, peak areas, and spectra of the unknown components are produced a priori. The technique is not limited to binary systems, and examples will be given involving three and four overlapping components.

THEORY

Consider the data produced by simultaneously measuring the absorption of light at many wavelengths during the elution of several chromatographic peaks. The total absorbance, $A_{i,j}$, at the jth wavelength during th ith scan is the sum of the absorbance of the n absorbing components present. In eq 1,

$$A_{i,j} = b\sum_{k=1}^{n}(c_{i,k}e_{k,j}) \quad (1)$$

b is the cell path length, $c_{i,k}$ is the concentration of the kth

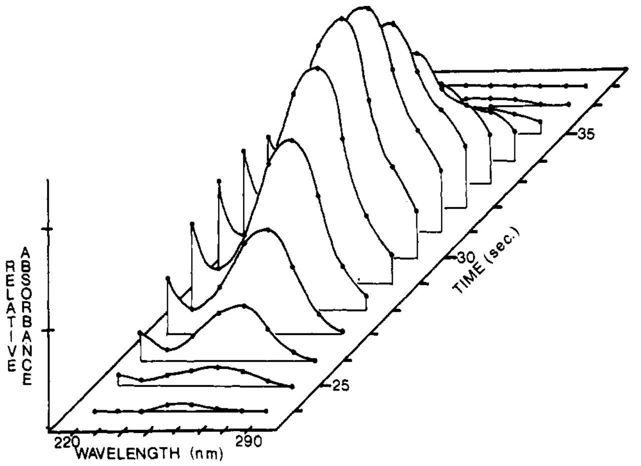


Figure 1. HPLC-absorbance matrix [A] for two-component mixture of adenylic and guanylic acids at resolution = 0.34.

component during the ith scan, and $e_{k,j}$ is the molar absorptivity of the kth component at wavelength j. Equation 1 can be rewritten in matrix form, where [A] is an $i \times j$ data matrix with i absorption spectra in each row at j wavelengths, b is constant, [C] is an $i \times n$ matrix with its columns representing the concentration profiles of the n components, and [E] is a $n \times j$ matrix with the pure absorption spectra of the n components in its rows:

$$[\mathbf{A}] = b[\mathbf{C}][\mathbf{E}] \qquad (2)$$

As an example, a raw data matrix is illustrated in Figure 1 that was synthetically generated from two Gaussian curves (σ = 2.0 s) sampled at 1.0-s intervals. The curves are separated by 2.72 s, corresponding to a chromatographic resolution of 0.34. The UV spectrum of adenylic acid was used for the first peak (λ_{max} = 255 nm), and the spectrum of guanylic acid was used for the second (λ_{max} = 260 nm).[12]

In the first step of our procedure, a raw data matrix [A] is selected from the set of scan data to represent a "time window" during which a cluster of peaks is observed to elute. Principal component analysis (PCA) is then used to detect the presence of overlapping peaks and the number of components. Analysis is performed with the covariance matrix [Z] rather than the original data matrix:

$$[\mathbf{Z}] = [\mathbf{A}]^{\mathrm{T}}[\mathbf{A}] \qquad (3)$$

PCA is a method of eigenanalysis that is performed to find the matrix of eigenvectors, [Q], that satisfies eq 4. The

$$[\mathbf{Z}][\mathbf{Q}] = [\lambda][\mathbf{Q}] \qquad (4)$$

columns of matrix [Q] are the eigenvectors of [Z], and [λ] is the matrix with eigenvalues on the diagonal. When random error is present in the data, c eigenvectors are required to satisfy eq 4, where c is the number of columns in [A]. In PCA, only n principal eigenvectors are sought to give the reduced eigenvector matrix, $[\mathbf{Q}]_n$. The n eigenvectors selected correspond to the n largest eigenvalues. The raw data can be represented by linear combinations of these vectors. The remaining vectors account only for random error. The determination of n yields the rank of the raw data matrix and the minimum number of components in [A]. It is possible for more components to be present if the differences between the elution profiles or spectra of the individual components are masked by random error.

The empirical Malinowski indicator function (IND) is used to find the number of principal components present:[13]

$$\mathrm{IND} = \mathrm{RE}/(x - n)^2 \qquad (5)$$

A local minimum is often observed in IND vs. n, where n is the number of principal components. In eq 5, x is the number of rows or the number of columns in [A], whichever is smaller, and RE is the "real error" in the raw data. The real error is easily calculated from eq 6 during factor analysis. Here, y

$$\mathrm{RE} = \sum_{j=n+1}^{c} (\lambda_j/[y(x-n)])^{1/2} \qquad (6)$$

is the number of rows or the number of columns in [A], whichever is larger. The variables x and n have the same values as in eq 5.

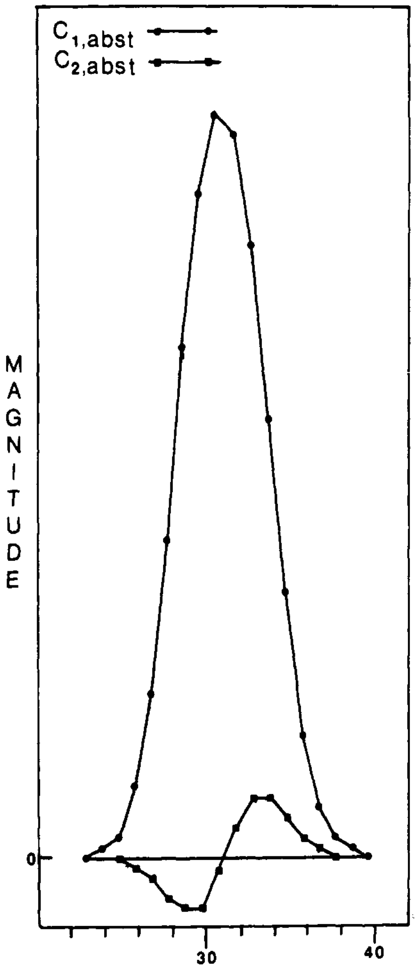


Figure 2. Abstract concentration vectors for adenylic and guanylic acids at resolution = 0.34.

An abstract solution is formulated from the reduced eigenvector matrix $[\mathbf{Q}]_n$, which reproduces the raw data within experimental error according to eq 7. $[\mathbf{E}]_{abst}$ is the set of

$$[\mathbf{A}] = [\mathbf{C}]_{abst}[\mathbf{E}]_{abst} \qquad (7)$$

abstract molar absorptivity vectors (eigenspectra) and is calculated from eq 8. $[\mathbf{C}]_{abst}$ is the set of abstract concen-

$$[\mathbf{Q}]_n^{\mathrm{T}} = [\mathbf{E}]_{abst} \qquad (8)$$

tration vectors calculated by eq 9. Figure 2 shows a plot of

$$[\mathbf{A}][\mathbf{Q}] = [\mathbf{C}]_{abst} \qquad (9)$$

the abstract concentration vectors produced by decomposition of the raw data matrix [A] illustrated in Figure 1. The term "abstract" is used here to indicate that neither the abstract concentration vectors nor the abstract eigenspectra represent physically meaningful quantities. Unfortunately, the abstract matrices are related to the real matrices through an unknown, nonorthogonal rotation. The remaining problem is to find the best transformation matrix [T], which rotates the abstract absorption and concentration matrices to an estimate of their real form according to

$$[\mathbf{C}]_{\text{real}} = [\mathbf{C}]_{\text{abst}}[\mathbf{T}] \qquad (10)$$

$$[\mathbf{E}]_{\text{real}} = [\mathbf{T}]^{-1}[\mathbf{E}]_{\text{abst}} \qquad (11)$$

TARGET TESTS

Target testing is performed by selecting test vectors that are good approximations of suspected real factors. The test vectors may either be approximations of elution profiles or spectra. In this work, we use approximations of elution profiles, denoted by $\mathbf{C}_{\text{test}}$. Once selected, a transform vector, $\mathbf{T}_i$, is calculated according to eq 12. The transform vector is then

$$\mathbf{T}_i = [\lambda]^{-1}[\mathbf{C}]_{\text{abst}}^{\mathrm{T}}\mathbf{C}_{\text{test}} \qquad (12)$$

used to calculate a predicted vector, $\mathbf{C}_{\text{pred}}$, according to eq 13.

$$\mathbf{C}_{\text{pred}} = [\mathbf{C}]_{\text{abst}}\mathbf{T}_i \qquad (13)$$

A thorough derivation is offered in the monograph of Malinowski and Howery, which shows that $\mathbf{T}_i$ is a least-squares result that minimizes the sum of the squares of the difference between $\mathbf{C}_{\text{test}}$ and $\mathbf{C}_{\text{pred}}$.[14] Good agreement is observed between the test vector and the predicted vector when valid test vectors are chosen. It is possible, however, to choose test vectors that are poor approximations of the elution profiles. In this case, the sum of the squares of the difference between $\mathbf{C}_{\text{test}}$ and $\mathbf{C}_{\text{pred}}$ will be greater than the error in the original data. The appropriate combination of n transform vectors from the set of valid target tests yields a transform matrix, [T].

To select test vectors for target testing, the uniqueness test is performed by constructing a vector of zeros with a single element set to a value of 1. The test is performed for each row (scan) in the original data matrix:

$$\begin{aligned} C_{t1,\text{test}} &= (1, 0, 0, \ldots, 0, 0, 0) \\ C_{t2,\text{test}} &= (0, 1, 0, \ldots, 0, 0, 0) \\ &\cdot \\ &\cdot \\ &\cdot \\ C_{ti,\text{test}} &= (0, 0, 0, \ldots, 0, 0, 1) \end{aligned}$$

Each test vector approximates a very narrow Gaussian or skewed Guassian distribution at one particular retention time. When the retention time represented by $\mathbf{C}_{\text{test}}$ corresponds to the retention time of a real component, a local minimum is observed in the sum of the squares of the difference between the test vector, $\mathbf{C}_{\text{test}}$, and the predicted vector, $\mathbf{C}_{\text{pred}}$. The local minimum indicates that the very narrow Gaussian test peak is a better approximation of the real elution profile at the selected retention time. The test is performed at each of the retention times represented in the raw data matrix so that n local minima may found, each one corresponding to the retention time of one of the n real components. When more than n minima are found, only the n smallest are selected. This is the so-called "needle search".

The n predicted vectors selected by the above method are least-squares results that are better approximations of the elution profiles than the original test vectors. Since the negative regions in the predicted vectors are physically meaningless, refined test vectors are generated by truncating all data beyond the boundary(ies) marked by the first negative regions encountered as one moves left or right from the peak maxima in the predicted vectors. The refined test vectors are used to produce new predicted vectors in an iterative fashion, each time having less and less negative area. As an example, the abstract solution shown in Figure 2 was subjected to the iterative target test procedure. Figure 3 shows the predicted vector for peak 1 after 1, 5, and 20 iterations.

Two criteria are used to terminate the iterative process. The first method uses the theory of error in target testing developed by Malinowski.[15] The apparent error in the target (AET) is defined as the sum of the squares of the difference between the test vector and the predicted vector, where r is the number of rows (scans) in the raw data matrix:

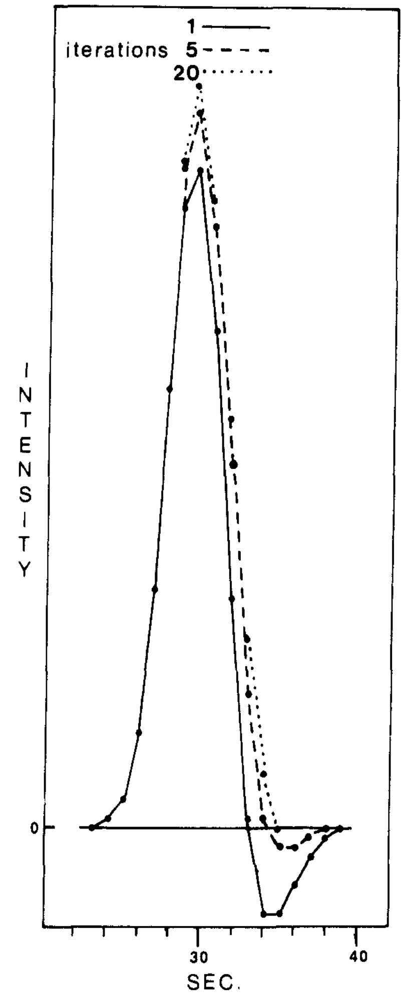


Figure 3. Predicted concentration vector for adenylic acid after 1, 5, and 20 iterations.

$$\text{AET} = [\sum_{i=1}^{r}(c_{\text{test},i} - c_{\text{pred},i})^2/r]^{1/2} \qquad (14)$$

Two other error functions called the real error in the predicted vector (REP) and the real error in the test vector (RET) are defined as

$$\text{REP} = (\text{RE})_n(\mathbf{T}_i{\cdot}\mathbf{T}_i)^{1/2} \qquad (15)$$

$$\text{RET} = [(\text{AET})^2 - (\text{REP})^2]^{1/2} \qquad (16)$$

$(\text{RE})_n$ is the real error for n components from eq 6 and $\mathbf{T}_i{\cdot}\mathbf{T}_i$ is the dot product of the transform vector, $\mathbf{T}_i$. The iteration is terminated when RET is less than RE. This error criterion stops the iterative process when the estimated random error in the test vector (RET) is less than the random error in the original data.

For the second criterion, we calculate the adjusted error in the target (ADJ):

$$\text{ADJ} = [\sum_{i=1}^{r}(c_{\text{test},i})_{\text{trunc}}{}^2/r]^{1/2} \qquad (17)$$

In eq 17, $(c_{\text{test},i})_{\text{trunc}}$ are the truncated data points. The iteration is stopped when the change in ADJ between successive iterations becomes less than RE:

$$\text{RE} > |\text{ADJ}_{k-1} - \text{ADJ}_k| \qquad (18)$$

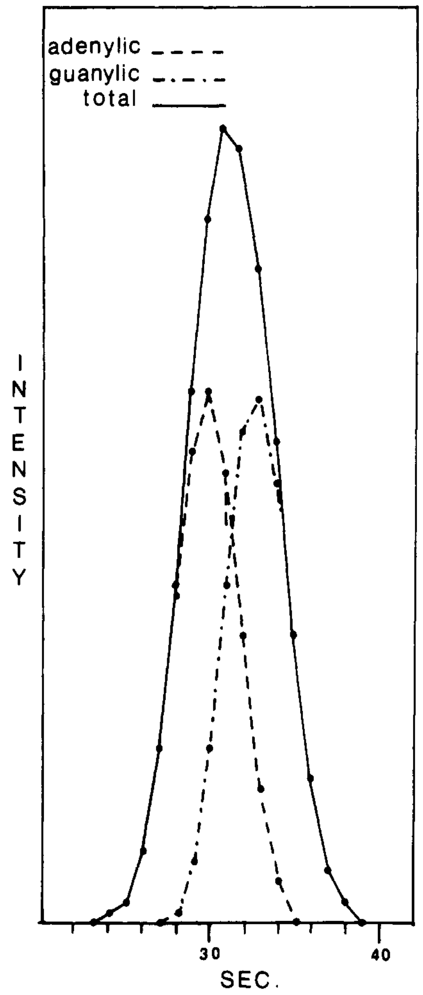


Figure 4. Predicted concentration vectors for adenylic and guanylic acids at resolution = 0.34.

In this case, the change in the target vector is considered slight enough that further improvements in the predicted vector are not significant. Figure 4 shows the results from the target tests of the abstract vectors in Figure 2. The iteration was stopped by the above criteria after 21 and 30 steps for peak 1 and peak 2, respectively.

The predicted spectra of the pure components can be easily calculated from eq 11. If the predicted spectra are normalized, the normalization constants can be used to scale the predicted peak profiles and, to a first approximation, give an estimate of the relative peak areas. This requires that one makes the assumption that the molar absorptivities of the overlapped peaks are nearly identical, which may not necessarily be valid.

The self-modeling capability of the technique is easily demonstrated by substituting overlapping acute triangles for Guassian curves. The predicted results accurately reproduce the original triangles.

EXPERIMENTAL

Several FORTRAN programs have been written that run on a Z80 microcomputer to generate simulated data, select a raw data matrix, and perform the principal component analysis and target tests. In the interest of speed and economy of memory, all calculations are carried out in single precision. As a result, round-off errors accumulate and control the ultimate accuracy of some of the results. Figure 5 shows a flow chart of how the FORTRAN programs are connected through intermediate disk files. The program called PEAKGEN generates Guassian chromatographic elution profiles specified by peak width, peak height, and retention time and then saves the results on disk. SPECGEN inputs absorbance data for user-specified compounds and saves them on disk for future use. SCANGEN inputs disk files generated by PEAKGEN and SPECGEN and then performs the matrix multiplication in eq 2 to generate a raw data file. SCANGEN also allows the user to add random noise. WINDOW selects a small matrix for factor analysis where peak clusters are observed to elute. Peaks are detected by the slope-threshold method. FACTOR performs the principal component analysis on raw data matrices having dimensions of up to 50 × 50. Up to five principal factors can be found. TARGET uses the abstract results from FACTOR and rotates them by the deconvolution method described previously. TARGET plots and compares the predicted results with the original elution profiles and spectra produced by PEAKGEN and SCANGEN.

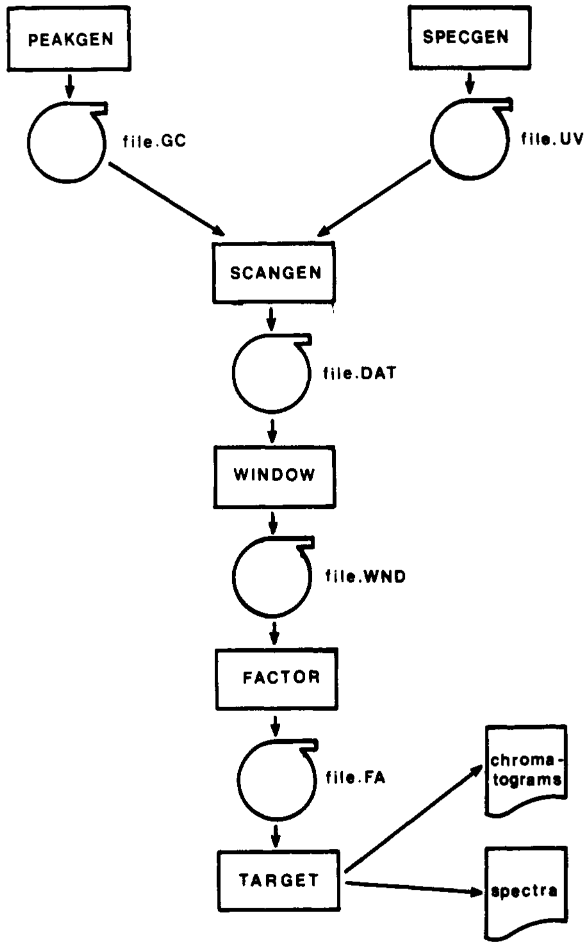


Figure 5. Program sequence and intermediate disk files.

Table I. UV Spectra

wavelength (nm)	molar absorptivities ($\times 10^{-3}$)			
	adenylic	cytidylic	guanylic	uridylic
220	7.00	8.70	4.50	4.30
230	3.30	3.96	2.65	2.10
240	6.00	1.90	6.05	4.00
250	11.40	3.20	10.45	7.60
260	13.40	6.90	11.00	9.70
270	9.10	11.12	8.40	7.58
280	2.95	12.59	7.60	2.85
290	0.50	8.65	5.38	0.18

RESULTS AND DISCUSSION

We have tested the deconvolution technique for two overlapped components with simulated data at various degrees of chromatographic resolution. The simulated data were pro-

Table II. Effect of Peak Separation on Accuracy

		adenylic						guanylic					
resolution	RE[a]	iter	AET[a]	REP[a]	RET[a]	ADJ[a]	% error in peak shape	iter	AET[a]	REP[a]	RET[a]	ADJ[a]	% error in peak shape
0.64	1.7	16	11.0	6.4	8.5	9.9	0.4	16	11.0	6.7	8.7	10.2	0.4
0.42	1.7	29	29.6	6.6	28.8	28.8	2.8	27	37.8	6.7	37.2	37.0	3.1
0.34	2.7	21	80.2	11.7	79.4	78.9	7.1	30	55.8	12.0	54.5	54.4	6.0
0.26	1.2	36	60.6	5.2	60.4	60.1	10.2	39	54.5	5.4	54.3	54.0	9.9
0.18	1.0	57	51.1	5.4	50.8	50.6	14.9	61	64.2	5.9	64.0	63.8	15.7

[a] ×10^{-5}.

Table III. Effect of Random Error on Accuracy at Resolution = 0.34

		adenylic						guanylic					
% noise	RE[a]	iter	AET[a]	REP[a]	RET[a]	ADJ[a]	% error in peak shape	iter	AET[a]	REP[a]	RET[a]	ADJ[a]	% error in peak shape
0.1	3.8	19	86.9	16.2	85.4	85.4	7.4	25	70.5	16.8	68.5	68.7	6.9
1.0	39.0	8	252.0	177.0	180.0	235.0	11.3	9	257.0	188.0	175.0	244.0	14.3
5.0	213.0	4	927.0	939.0	152.0	890.0	24.9	4	949.0	962.0	160.0	912.0	28.0

[a] ×10^{-5}.

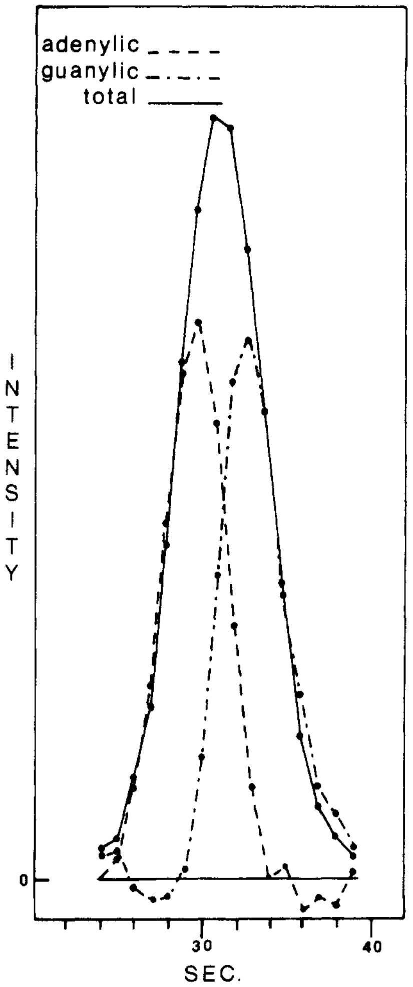


Figure 6. Predicted concentration vectors for adenylic and guanylic acids at resolution = 0.34 and 1.0% noise.

duced from Guassian elution profiles of equal area (σ = 2.0 s) sampled at 1.00-s intervals. The UV spectra of adenylic acid, guanylic acid, uridylic acid, and cytidylic acid are listed in Table I.[12] The spectra of adenylic acid and guanylic acid were used for the first and second peaks, respectively, in the two-component test. The results of these tests are summarized in Table II. The relative error in the peak shape is the sum of the absolute value of the difference between the normalized predicted and original concentration profiles divided by the total peak area. At a resolution of 0.64, the predicted concentration vectors faithfully reproduce the original elution profiles. The error increases as chromatographic resolution becomes worse. This is to be expected. As the uniqueness of peaks (the nonoverlapped areas) decreases, the ability to separate them also decreases. At a resolution of 0.18, the predicted peak shapes are less reliable.

Table IV. Accuracy of Results Using Three Overlapped Peaks at Resolution = 0.42

compd	iter	AET[a]	REP[a]	RET[a]	ADJ[a]	% error in peak shape
adenylic	17	71.5	50.6	50.5	69.9	6.1
guanylic	18	58.4	17.8	55.6	56.7	5.7
uridylic	18	80.0	75.0	27.7	78.1	6.4

[a] ×10^{-5}.

Table V. Accuracy of Results Using Four Overlapped Peaks at Resolution = 0.42

compd	iter	AET[a]	REP[a]	RET[a]	ADJ[a]	% error in peak shape
adenylic	19	62.0	46.0	41.6	60.8	6.5
guanylic	24	42.3	20.3	37.2	40.6	6.2
uridylic	20	63.4	67.0	21.6	61.2	7.3
cytidylic	21	49.1	11.2	47.8	47.5	5.9

[a] ×10^{-5}.

Scaled random numbers were added to the raw data to test the effect of random error on the accuracy of the deconvolution technique. The results for a two-component mixture of adenylic acid and guanylic acid at a resolution of 0.34 and various noise levels are summarized in Table III. The percent noise levels listed in Table III were calculated relative to the maximum simulated absorbance. Figure 6 shows a plot of the predicted vectors at a 1.0% noise level.

We have also tested the deconvolution technique for three and four overlapped components with simulated data. Three or four overlapped Guassian elution profiles (σ = 2.00 s) of equal area were used. Adjacent peaks are separated by 3.4 s, corresponding to chromatographic resolution of 0.42. A sampling interval of 1.00 s was used. For the three-component mixture, the UV spectra of adenylic acid, guanylic acid, and uridylic acid were used for peaks 1, 2, and 3, respectively. The spectra of adenylic acid, guanylic acid, uridylic acid, and cytidylic acid were used for peaks 1, 2, 3, and 4 in the four-component mixture. The results for the three-component

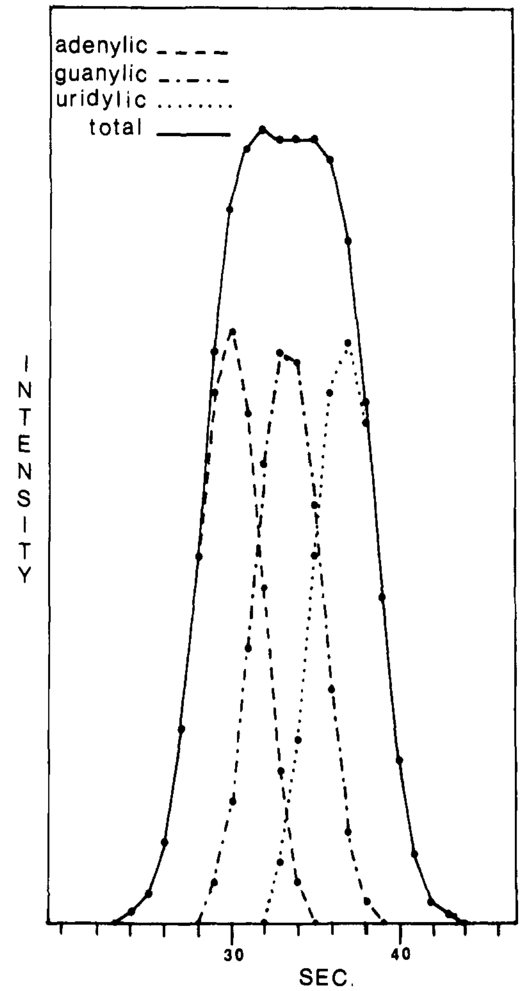


Figure 7. Predicted concentration vectors for mixture of adenylic, guanylic, and uridylic acids at resolution = 0.42.

mixture are summarized in Table IV, and the predicted peak profiles are plotted in Figure 7. The results for the four-component mixture are shown in Table V and Figure 8. The predicted peak shape in both the three- and four-component examples are in good agreement with original Guassian peak profiles.

CONCLUSIONS

We have described a self-modeling technique for selecting test vectors for use in target transformation of chromatographic data. This is the first technique we know of that can give a priori estimates of the transformation matrix to rotate the abstract factor analysis solutions to good approximations of real peak profiles when there are no data points unique to the individual components. The examples used to demonstrate our deconvolution technique represent a severe test of the method. For Guassian curves of equal area at a chromatographic resolution of 0.34, separate shoulders are not observable in the combined signal, and approximately 73% of the combined peak area is overlapped. At a resolution of 0.18, about 96% of the combined peak area is overlapped. In both cases, the presence of two overlaped peaks is easily detected. Acceptable elution profiles of the individual components were produced from the target transformation. The Guassian peaks and spectra used in the examples were sparsely sampled. Preliminary trials show that smaller sampling intervals do not significantly improve the accuracy of the technique. This unexpected result is being investigated further.

Finally, the calculations can be performed rapidly. Principal component analysis performed by FACTOR on a 20 × 8 raw data matrix requires about 10 s running on a 4-MHz Z80 microcomputer to find two principal factors. The target tests performed by TARGET require about 1 s per iteration. The estimated elution profiles and deconvolved spectra produced by our technique can provide analysts with valuable insights and could potentially increase productivity when insufficient chromatographic resolution is available.

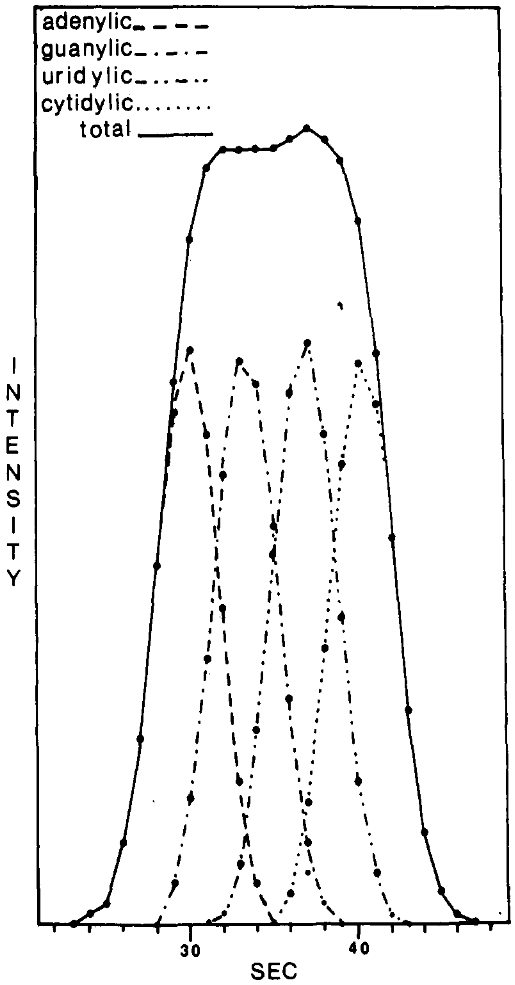


Figure 8. Predicted concentration vectors for mixture of adenylic, guanylic, uridylic, and cytidylic acids at resolution = 0.42.

Registry No. Adenylic acid, 61-19-8; guanylic acid, 85-32-5; uridylic acid, 58-97-9; cytidylic acid, 63-37-6.

REFERENCES AND NOTES

(1) Scoble, H. A.; Brown, P. R. In "High-Performance Liquid Chromatography, Advances and Perspectives"; Academic Press: New York, 1983; pp 22-32.

(2) Jaffe, H. "Separation of Two Invertebrate Peptides by HPLC with a Multichannel High-Speed Spectrophotometric Detector". *Liq. Chromatogr. HPLC* **1984**, *1*, 418-426.

(3) Macnaughtan, D.; Rogers, L. B.; Wernimont, G. "Principal-Component Analysis Applied to Chromatographic Data". *Anal. Chem.* **1972**, *44*, 1421-1427.

(4) Davis, J. E.; Shepard, A.; Stanford, N.; Rogers, R. B. "Principal-Component Analysis Applied to Combined Gas Chromatographic-Mass Spectrometric Data". *Anal. Chem.* **1974**, *46*, 821-825.

(5) Ritter, G. L.; Lowry, S. R.; Isenhour, T. L.; Wilkins, C. L. "Factor Analysis of the Mass Spectra of Mixtures". *Anal. Chem.* **1976**, *48*, 591-595.

(6) Halket, J. M. "Factor Analysis of Repetitively Scanned Spectra in Gas Chromatography-Mass Spectrometry". *J. Chromatogr.* **1979**, *175*, 229-241.

(7) Tway, P. C.; Cline Love, L. J. "A Totally Automated Data Aquisition Reduction System for Routine Treatment of Mass Spectroscopic Data by Factor Analysis". *Anal. Chim. Acta* **1980**, *117*, 45-52.

(8) Knorr, F. J.; Futrell, J. H. "Separation of Mass Spectra of Mixtures by Factor Analysis". *Anal. Chem.* **1979**, *51*, 1236-1241.

(9) Sharaf, M. H.; Kowalski, B. R. "Quantitative Resolution of Fused Chromatographic Peaks in Gas Chromatography/Mass Spectrometry". *Anal. Chem.* **1982**, *54*, 1291-1296.
(10) Malinowski, E. R.; McCue, M. "Qualitative and Quantitative Determination of Suspected Components in Mixtures by Target Transformation Factor Analysis of Their Mass Spectra". *Anal. Chem.* **1977**, *49*, 284-287.
(11) McCue, M.; Malinowski, E. R. "Target Factor Analysis of the Ultraviolet Spectra of Unresolved Liquid Chromatographic Fractions". *Anal. Chem.* **1983**, *37*, 463-469.
(12) Kalivas, J. H. "Precision and Stability for the Generalized Standard Addition Method". *Anal. Chem.* **1983**, *55*, 565-567.
(13) Malinowski, E. R. "Determination of the Number of Factors and the Experimental Error in a Data Matrix". *Anal. Chem.* **1977**, *49*, 612-617.
(14) Malinowski, E. R.; Howery, D. G. In "Factor Analysis in Chemistry"; Wiley: New York, 1980; pp 50-52.
(15) Malinowski, E. R. "Theory of Error for Target Factor Analysis with Applications to Mass Spectrometry and Nuclear Magnetic Resonance Spectrometry". *Anal. Chim. Acta* **1978**, *103*, 339-354.

A Convenient Notation System for Organic Structure on the Basis of Connectivity Stack

HIDETSUGU ABE, YOSHIHIRO KUDO,[†] TOHRU YAMASAKI,[‡] KAZUO TANAKA,[§]
MASAHIRO SASAKI, and SHIN-ICHI SASAKI*

Laboratory for Chemical Information Science, Toyohashi University of Technology, Toyohashi, Aichi,
Japan 440

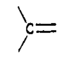
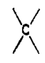
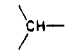
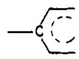
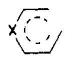
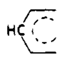
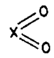
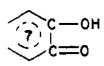
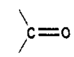
Received August 8, 1983

A convenient notation system for organic structures has been developed for the application of the connectivity stack. A notation arbitrarily encoded for a structure by a user through a rather simple procedure using 35 codes, which have been previously prepared, is automatically canonicalized in a computer. The notation given by the user is standardized according to the rules for rearranging the codes into a dictionary order. The connectivity stack is estimated for each of the standard notations and its permuted derivatives. The notation whose stack is the largest amount is decided to be canonical. This notation method will be widely applicable in the field of structure manipulation because of its extreme simplicity.

Several methods for the representation of organic structures have been investigated for computer-aided storage and retrieval of the structures.[1] Linear notations and connection table methods are two major techniques for the topologically unambiguous and unique representation of chemical structures.[2] The connection table descriptions specify all the atoms of a molecule (hydrogen is often suppressed) and may explicitly describe the connectivity of each atom. On the other hand, one of the features of the linear notation method is that chemical structure can be expressed more compactly by the use of letters, numerals, and some symbols. The number of letters, numerals, and symbols used to represent a structure is, in general, much fewer than the number of atoms included in the structure. According to such compactness, the linear notation method seems to be more preferable than the connection table method for compilation of a vast number of structures to be treated in a computer. However, the procedure for canonicalization of a linear notation is generally so tedious and complicated that users hesitate to adopt the methods.

In this paper, we present a new notation system on the basis of a "connectivity stack", which has been published by Y.K. and S.S.[3] The notation system, CANOST (autoCANOnicalization system for organic STructures), has the following features. (1) The notation given arbitrarily by the user through rather simple procedures described later is automatically canonicalized in a computer. (2) Thirty-five symbols expressing atoms, atomic groups, ionic charges, and others as listed in Table I are used to make the arbitrary notation. Two or three hours is normally sufficient to learn how to encode chemical structures for even a beginner in chemistry. (3) Though most of structures are expressed with the 35 items, any other symbols consisting of up to four letters may be added if necessary. (4) The notation can be easily converted into a corresponding connection table. The latter, in some cases, is more usable and convenient than the linear notation for computer-aided manipulation of structures.

Table I. Code of Substructure in CANOST[a]

no.	substructure	code	no.	substructure	code
1	—C≡	T	16	=C=O	VD
2	HC≡	T1	17	—O—	Q
3	=C=	DD	18	—OH	Q1
4		DS	19	=O	QD
			20	—F	LF
5	—CH=	D1	21	—Cl	LC
6	H_2C=	D2	22	—Br	LB
			23	—I	LJ
7		C	24[d]	single bond	SG
			25	cation	+
8		C1	26	anion	–
			27	radical	·
9	—CH_2—	C2	28	chelation	/
10	—CH_3	C3	29	other atom	X
11[a]		Y	30[e]		XR
12[b]		Y1	31		XW
13[c]		YT	32	=X	XD
			33	=X=	XX
			34	≡X	XT
14		V	35	XH_P	XP
15	—CHO	V1			

[a] Aromatic carbon without hydrogen. [b] Aromatic carbon with hydrogen. [c] –C(OH)–C(O)– in troponoid. [d] Prepared for connecting D1 to clearly express conjugated double bond (see Figure 3). [e] Non-carbon atom in aromatic structure.

GENERAL ENCODING PROCEDURES

The following describes how to encode a chemical structure into CANOST notation.

Step 1. Select proper symbols from Table I for the atoms and atomic groups in a structure concerned. If two or more alternative encodings are possible for the structure, the one

[†] Present address: Faculty of Engineering, Yamagata University, Yonezawa, Yamagata, Japan 992.

[‡] Present address: Mitsui Petrochemical Industry Ltd. Co., Iwakuni, Yamaguchi, Japan 740.

[§] Present address: Asahi Research Center Co. Ltd., Uchisaiwai-cho, Chiyoda, Tokyo, Japan 100.

0095-2338/84/1624-0212$01.50/0 © 1984 American Chemical Society