

On the Minimum Information Sufficient for Identifying Individual Organic Properties from Coincident Lines in Mass Spectra

V. L. Tal'roze, V. V. Raznikov, G. D. Tantsyrev

1964

SovietRxiv

View the original and related papers at <https://sovietrxiv.org/items/ru-196401.65569>

Source: Math-Net.Ru and CyberLeninka. Machine translation. Verify with the original.

Abstract

Full Text

Reports of the Academy of Sciences of the USSR
1964. Volume 159, No. 1

PHYSICAL CHEMISTRY

V. L. TAL' ROZE, V. V. RAZNIKOV, G. D. TANTSYREV

ON THE MINIMUM INFORMATION SUFFICIENT FOR THE IDENTIFICATION OF INDIVIDUAL ORGANIC SUBSTANCES FROM COINCIDENT LINES OF MASS SPECTRA

(Presented by Academician N. N. Semenov, 23 VII 1964)

For the mass-spectrometric identification of individual substances, one usually tries to make use of the most characteristic lines of their spectra. This generally justified approach requires recording the whole, or almost the whole, mass spectrum. To obtain such a spectrum one must either spend a considerable amount of time or use instruments with a very rapid mass scan, in particular time-of-flight mass spectrometers, sacrificing, if necessary, sensitivity and accuracy in determining the amount of substance. In this connection the question arises whether identification cannot be based on a few lines that are not characteristic but, on the contrary, are present in the mass spectra of most substances. This question is especially relevant when a mass spectrometer is used as the detector of a chromatograph that separates the mixture being analyzed into individual components, since, as will be seen below, an affirmative answer leads to the simplest constructive solutions (with high sensitivity and a sufficiently short recording time).

Consideration of the molecular mass spectra of organic substances shows that, in the overwhelming majority of them, for a number of structural reasons, certain lines in the range of mass numbers 36–45 are present and more or less intense. Ions heavier than 48 a.m.u. are already absent from the spectra of many low-molecular substances. On the other hand, lines of ions with mass less than 36 a.m.u. are already very low in intensity in the spectra of many organic substances.

These circumstances are the reason why we have concentrated attention on the region 36–45 a.m.u. The problem was formulated as follows: the ratios of the intensities of how many lines in this region, and with what accuracy, must be known in order to identify an individual substance? Obviously, such a formulation of the question is meaningful if the substances are distributed more or less uniformly with respect to the values of the relative intensities of the lines.

Fig. 1. Distribution function of the relative identification accuracy

Figure 1: Fig. 1. Distribution function of the relative identification accuracy

On the basis of material available to the authors on approximately 900 mass spectra of various organic compounds ⁽¹⁾, the distribution function of these substances with respect to the value of the ratio η of the intensities of the lines 39 and 41 a.m.u.* was constructed. Unexpectedly, a rather smooth function was obtained, on average close to logarithmic. Owing to this it proved possible to introduce a certain parameter G , which characterizes this function and is independent of the type of substance and the value of η :

$$G = \frac{\nu}{\ln \eta_\nu / \eta_1}, \quad (1)$$

where ν is the number of the substance if the substances are arranged in order of increasing η ; η_ν is the value of the ratio for the ν -th substance; η_1 is the value of the ratio for the first substance.

* 24 substances did not have the lines 39 and 41 and were not included in the consideration.

In Fig. 1 a graph is presented in which the quantity ν is plotted along the abscissa, and the parameter G along the ordinate. It is seen that this parameter is constant over the entire range of values of ν , to an accuracy of $\pm 15\%$. It can be shown that the existence of such a distribution leads us directly to the concept of the existence of a certain, approximately identical in all cases—if a sufficiently large number of substances is tested—identification accuracy ψ . The meaning of the quantity ψ is most easily explained by saying that its reciprocal $1/\psi$ is equal to the number of substances having, within the measurement accuracy, one and the same value of η .

Fig. 1. Distribution function of the relative identification accuracy

Indeed, let the relative error of determination be equal to $\pm\delta$, and let the values of η for the set of substances under consideration lie in a range whose limiting values are η_1 and η_ν . It is easy to see that no more than n groups of substances can differ in the value of the ratios η within the given measurement accuracy, where n is determined from the equation*

$$(1 + 2\delta)^{n-1} = \eta_\nu / \eta_1, \quad (2)$$

whence

$$n = \frac{\ln \eta_\nu / \eta_1}{\ln(1 + 2\delta)} + 1. \quad (3)$$

If the number of all possible substances is equal to ν , with $\nu > n$, and each group contains the same number of substances, equal to ν/n , then, knowing the value of η , the given substance can be assigned to a definite group of ν/n substances. The quantity n/ν will then be the identification accuracy ψ .

From (3)

$$\psi = \frac{1}{\nu} \left[\frac{\ln \eta_\nu / \eta_1}{\ln(1 + 2\delta)} + 1 \right]. \quad (4)$$

Since $\eta_\nu / \eta_1 > 1$ and $\delta \ll 1$,

$$\psi \simeq \frac{\ln \eta_\nu / \eta_1}{2\nu\delta}. \quad (5)$$

Comparing (5) and (1), we obtain

$$\psi = \frac{1}{2\delta G}. \quad (6)$$

Thus, the existence of an approximately constant parameter G , following from the analysis found for spectra of the logarithmic distribution, does indeed lead to the idea of an identification accuracy ψ that is approximately constant for the entire class of substances under consideration. Since the parameter G is constant only approximately, it is reasonable to speak of average $\overline{G}$ and $\overline{\psi}$. Thus, for a known $\overline{G}$, the quantity ψ is determined by the measurement error δ . For the class of “common” substances under consideration, $\overline{G}$, as can be determined from Fig. 1, proved to be equal to $2.7 \cdot 10^2$. Taking the measurement error $\delta = 0.02$, we obtain $\overline{\psi} = 8.5 \cdot 10^{-2}$. Thus, by measuring the quantity η , i.e., the ratio of the intensities of just two lines of the mass spectrum, we can establish that the given substance is one of 12 (of just such!) substances (out of 900 *a priori* possible).

* For a similar simple approach to calculating the amount of information for another problem, see in ⁽²⁾.

The next question is to what extent the accuracy of identification will increase as a result of measuring the intensity ratio of another pair of spectral lines, if one assumes that for the distribution with respect to the second ratio the logarithmic law is also valid and, consequently, the assumption of approximately constant identification accuracy is valid. Finally, one may pose the question of the resulting accuracy of identification if the intensity ratios are known for l pairs of lines.

Let us liken each of our substances to a ball and say that there are n_1 boxes, each of which is characterized by a value η different from that of the preceding one, i.e., in each box there are $1/\psi$ balls. To each row of boxes there corresponds

the complete set of values η for the corresponding pair of mass-spectral lines. Suppose that in the j -th row there are n_j boxes and that the balls in this row are again distributed uniformly among the boxes— ν/n_j balls in each.

First let us carry out the calculations for two rows of boxes (two types of ratios). Mathematically the problem is formulated as follows: the balls from the boxes of the first row are taken out and distributed among the boxes of the second row according to the law of chance, but in such a way that each box contains the same number of balls; it is necessary to find the probability that exactly k balls from a specified box of the first row fall into one and the same specified box of the second row (obviously, $k \leq \nu/n_2 = 1/\psi$). We shall call such an event a k -fold coincidence in the second row of boxes, and denote its probability by ${}_2P^{(k)}$.

According to the scheme of independent trials ⁽³⁾,

$${}_2P^{(k)} = C_{\nu/n_1}^k \left(\frac{1}{n_2} \right)^k \left(1 - \frac{1}{n_2} \right)^{\nu/n_1 - k}. \quad (7)$$

The total number of k -fold coincidences is

$$Y = \sum_{ij} I_k^{ij}, \quad (8)$$

where I_k^{ij} is the indicator of the occurrence of exactly k balls from the i -th box of the first row in the j -th box of the second row. By virtue of the additive property of mathematical expectation, we have

$${}_2E^{(k)} = E_Y = C_{\nu/n_1}^k \left(\frac{1}{n_2} \right)^k \left(1 - \frac{1}{n_2} \right)^{\nu/n_1 - k} n_1 n_2, \quad (9)$$

where ${}_2E^{(k)}$ is the mathematical expectation of the number of k -fold coincidences in the second row of boxes. By analogy with the preceding, let us call a k -fold coincidence in the l -th row the event consisting in the fact that exactly x of the same balls belong to a selected set of l boxes (one from each row).

The probability of this event is given by the formula

$${}_lP^{(x)} = \frac{{}_lE^{(x)}}{n_1 n_2 \dots n_l}, \quad (10)$$

where ${}_lE^{(x)}$ is the mathematical expectation of the number of x -fold coincidences in the l -th row of boxes.

After transformations we obtain

Fig. 2

Figure 2: Fig. 2

$${}_l P^{(x)} = \sum_{i_1 i_2 \dots i_{l-1}} C_{\nu/n_1}^{i_1} C_{i_1}^{i_2} \dots C_{i_{l-1}}^x \left(\frac{1}{n_2}\right)^{i_1} \left(\frac{1}{n_3}\right)^{i_2} \dots \left(\frac{1}{n_l}\right)^x \times \\ \times \left(1 - \frac{1}{n_2}\right)^{\nu/n-i_1} \left(1 - \frac{1}{n_3}\right)^{i_1-i_2} \dots \left(\frac{1}{n_l}\right)^{i_{l-1}-x}. \quad (11)$$

With the aid of formulas of type (11), for a sufficiently large number of substances one can predict in the spectra of how many of these substances there will coincide

by a factor of 2, 3, or more, the number of ratios. It is obvious that, when planning an analytical experiment, we must proceed from the necessity of recording such a number of pairs of lines that the majority of substances will have at least one noncoinciding ratio.

It is useful to introduce the concept of the resulting mean identification accuracy $\bar{\psi}_l$ for the case of recording l ratios. It is reasonable to define $\bar{\psi}_l$ as

$$\bar{\psi}_l = \frac{1}{\varepsilon_l} = \frac{1}{\sum_i \frac{l_i^2 E^{(i)}}{v}}, \quad (12)$$

where ε_l is the mathematical expectation of the “size” of a group of substances with indistinguishable l ratios. If the resulting mean identification accuracy proves, for example, to be equal, when three pairs of lines are recorded, to 1/1.02, this means that only 2% of all substances cannot be determined unambiguously.

A comparison was made of the available material on 900 mass spectra with the predictions of this probability theory for two ratios of lines, $m/e = 39$ and 41, and $m/e = 41$ and 43. Figure 2 presents a histogram showing what fraction of substances participates in 2-, 3-, 4-, and 5-fold coincidences. It may be considered that probability theory predicts the number of coincidences with accuracy sufficient for practical purposes. The observed deviations from theory are directed in the direction “favorable” for identification.

Fig. 2. Relative fraction of coincidences in identification by two pairs of lines 39 and 41, and 41 and 43 a.m.u. a —experiment, $\delta = 0.02$; b —experiment, $\delta = 0.05$; v —theory, $\delta = 0.02$; g —theory, $\delta = 0.05$. On the abscissa axis—multiplicity of coincidence (coincidence of “multiplicity 1” = absence of coincidence).

Thus, for unambiguous identification of the majority of the most common substances for which mass spectra are known, 3 ratios (with an accuracy of determination of the ratios $\delta = 0.05$), i.e., 4 lines of the spectrum, are sufficient. However, as our experience has shown, in practice it is often sufficient to measure 2 or even 1 ratio, i.e., only 2-3 lines of the spectrum, since the experimenter always obtains or has *a priori* additional information of a chromatographic or other nature. It is essential that the discussion concerns ions of small mass, i.e., in the analysis of substances of large molecular weight one can manage with a mass spectrometer for a small mass range and with low resolving power.

These conclusions made it possible to propose using, as a sufficiently universal detector for a chromatograph, a very simple mass spectrometer with a resolving power of about 50, continuously recording, during acquisition of the chromatogram, only 2 lines of the mass spectrum^(4,5,6). Repeated acquisition of the chromatogram "in the light" of two other lines makes it possible, if necessary, to obtain not 1, but 2 or 3 ratios.

The authors express their gratitude to S. T. Kozlov, V. I. Gorshkov, and N. G. Ivannikov for assistance in the computations.

Institute of Chemical Physics
Academy of Sciences of the USSR

Received
15 VII 1964

REFERENCES CITED

1. *Mass spectral data*, API, Res. Pr. 44.
2. Yu. N. Lyubotov, Dissertation, Institute of Metallurgy, Moscow, 1963.
3. B. V. Gnedenko, *Course of Probability Theory*, Moscow, 1961.
4. V. L. Tal' roze, G. D. Tantsyrev, Author' s Certificate Application No. 699247, 1961.
5. V. L. Tal' roze, V. D. Grishin, G. D. Tantsyrev, Author' s Certificate Application No. 753509, 1962.
6. V. L. Tal' roze, G. D. Tantsyrev et al., French Patent No. 1367641, 1964.

Note: Figure translations are in progress. See original paper for figures.

Source: Math-Net.Ru and CyberLeninka. Machine translation. Verify with the original.