*Journal of Chromatography*, 364 (1986) 63–77
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

CHROMSYMP. 791

# COMPUTER ALGORITHM FOR THE IDENTIFICATION OF SULPHUR-CONTAINING COMPOUNDS IN COMPLEX MIXTURES WITH THE USE OF GAS CHROMATOGRAPHIC DATA

D. N. GRIGORYEVA*, R. V. GOLOVNYA, T. A. MISHARINA and A. F. AEROV
*A. N. Nesmeyanov Institute of Organo-Element Compounds, USSR Academy of Sciences, Ul. Vavilova 28, Moscow (U.S.S.R.)*

## SUMMARY

Algorithms for the computer-aided gas chromatographic identification of sulphur-containing compounds are suggested. The computer program provides an identification according to three algorithms. The reliability of the identification according to these algorithms was determined by the analysis of a 23-component mixture as an example. An equation for calculating the retention parameters of homologues is proposed. The coefficients of the equation were determined for 28 homologous series for three stationary phases of different polarity.

## INTRODUCTION

The investigation of complex natural mixtures containing several hundred components is extremely complicated. Of particular importance is the design of multi-purpose systems for the chemical analysis of trace organic substances which necessarily include a chromatographic separation of the components and the use of a computer for identification purposes[1]. The substances may be identified with the help of stationary phases of different polarity. This method is particularly useful in the analysis of aliphatic compounds[2] and has been successfully applied to the design of computer software algorithms for gas chromatographic (GC) data.

The technique most frequently used in computer-aided GC analysis involves the comparison of retention parameters (usually retention indices) with their values for standard substances obtained under the same GC conditions[3–6]. Information retrieval systems employing this technique allow the presence of those components whose retention parameters are stored in the computer memory to be established.

A different approach was suggested earlier[7]. The algorithm is based on linear-fractional equations correlating the retention indices of compounds with their structures, boiling points and retention index differences ($\Delta I$) on stationary phases of different polarity. The applicability of the algorithm is demonstrated by the analysis of a mixture of nitrogen-containing compounds characterized by strong adsorption activity and donor–acceptor interaction capability. It should be noted that $\Delta I$ values are unacceptable for the identification of compounds with non-specific functional

groups, which limits the applicability of the program. In addition, the technique for identifying the components of a mixture, daccording to the literature[8], requires the use of *a priori* data on the boiling points of the substances.

This study was aimed at the development and verification of a more universal identification algorithm which could be used to identify substances in complex mixtures.

## EXPERIMENTAL AND RESULTS

### *Sample preparation and analysis*

The sample was prepared by dissolving 1–10 μl of each of the 23 test substances listed in Table I in 10 ml of absolute diethyl ether. The mixture was analysed using a Pye Unicam Series 104 chromatograph equipped with a flame ionization detector on 3.2 m × 2 mm I.D. glass columns packed with Chromosorb W AW DMCS (80–100 mesh) coated with 5% of either Apiezon M (Ap), OV-17 polymethylphenylsilicone (OV) or Triton X-305 (Tr) as the stationary phase. The analysis of 0.5–

TABLE I

RETENTION INDICES OF TEST SUBSTANCES ON APIEZON M AND GC COMPUTER IDENTIFICATION RESULTS WITH THE USE OF ALGORITHMS *B* AND *C*

| *Standard* | *I (standard)** | *I (mixture)** | *θ during identification* | *No. of fraction* |
|---|---|---|---|---|
| 2-Thiapropane | 523(13) | – | – | – |
| 1-Propanethiol | 623(13) | 630(4) | 16 | 1 |
| 3-Thiapentane | 698(13) | 705(2) | 11 | 1 |
| 2,3-Dithiabutane | 765(6) | 767(2) | 3 | 1 |
| 2-Thiahexane | 815(6) | 819(1) | 16 | 2 |
| 1,2-Ethanedithiol | 844(6) | 846(2) | 4 | 2 |
| Furfurylmercaptan | 903(5) | 905(1) | 4 | 2 |
| 2,4-Dithiapentane | 903(5) | 905(1) | 3 | 2 |
| 2-Thiaheptane | 916(5) | 918(1) | 5 | 3 |
| 3,4-Dithiahexane | 935(5) | 936(1) | 5 | 3 |
| 2,4-Dithiahexane | 979(5) | 983(1) | 7 | 3 |
| 3,3-Dimethyl-2,4-dithiapentane | 986(5) | 983(1) | 5 | 3,4 |
| 2-Thiophenaldehyde | 995(5) | 1000(1) | 5 | 4 |
| 1,3-Dithiolane | 1013(5) | 1011(1) | 4 | 4,5 |
| 3,5-Dithiaheptane | 1054(5) | 1056(1) | 4 | 5,6 |
| 2-Acetylthiophene | 1084(5) | 1084(1) | 5 | 6 |
| 3,3,5-Trimethyl-2,4-dithiahexane | 1089(5) | 1084(1) | 5 | 6 |
| 1,4-Dithiane | 1095(5) | 1093(1) | 5 | 6 |
| Phenylthiamethane | 1106(5) | 1109(1) | 3 | 6 |
| 2,6-Dimethyl-1,4-dithiane | 1179(5) | 1179(1) | 5 | 6 |
| 2,4,4,6-Tetramethyl-3,5-dithiaheptane | 1191(5) | 1191(1) | 7 | 6 |
| 2,5-Dimethyl-1,4-dithiane | 1191(5) | 1191(1) | 7 | 6 |
| 2-Propyl-1,3-dithiolane | 1222(5) | 1222(1) | 3 | 7 |

* Confidence intervals given in parentheses.

TABLE II

RETENTION INDICES OF TEST SUBSTANCES ON OV-17 AND TRITON X-305 AT 130°C

| *No.* | *Stationary phase* | | | |
|---|---|---|---|---|
| | *OV-17* | | *Triton X-305* | |
| | *Retention index* | $\rho^*$ | *Retention index* | $\rho^*$ |
| 1 | 711 | 5 | 817 | 7 |
| 2 | 798 | 4 | 882 | 3 |
| 3 | 891 | 3 | 997 | 2 |
| 4 | 977 | 1 | 1036 | 2 |
| 5 | 1010 | 1 | 1095 | 1 |
| 6 | 1059 | 1 | 1177 | 1 |
| 7 | 1077 | 1 | 1208 | 1 |
| 8 | 1111 | 1 | 1236 | 1 |
| 9 | 1135 | 1 | 1291 | 4 |
| 10 | 1147 | 2 | 1306 | 2 |
| 11 | 1199 | 2 | 1332 | 3 |
| 12 | 1206 | 1 | 1353 | 1 |
| 13 | 1230 | 2 | 1380 | 1 |
| 14 | 1269 | 2 | 1412 | 1 |
| 15 | 1283 | 2 | 1489 | 1 |
| 16 | 1296 | 1 | 1523 | 1 |
| 17 | 1316 | 2 | 1554 | 1 |
| 18 | 1343 | 2 | 1564 | 3 |
| 19 | | | 1627 | 1 |

* $\rho$ = confidence interval corresponding to 0.95 probability of the retention index reproducibility.

2-$\mu$l samples of the ether solution was carried out at 130°C, the flow-rate of the carrier gas (extra-pure nitrogen) being 20–60 ml/min. The temperature of both the injector and detector was 180°C. The amount of vapour of the standard *n*-alkane mixture $C_5$–$C_{15}$ injected was 0.3 $\mu$l.

Retention indices averaged over seven to nine measurements, confidence intervals calculated with the use of Student's *t*-test and the corresponding 0.950 probabilities, together with retention index reproducibilities, are presented in Table I for the stationary phase Ap and in Table II for OV and Tr.

*Fractionation*

To obtain a reliable correlation of retention indices of components showing inseparable chromatographic zones on any of the columns, the solution of the test substances was separated into seven fractions by GC. Volumes of 40 $\mu$l of solution were fractioned on an Ap column, the outlet of which was equipped with a stream splitter heated to 200°C, intended to direct 10% of the effluent to the flame ionization detector and 90% to a trap containing a sorbent in order to collect the fractions. The traps employed were 80 × 2 mm I.D. glass tubes filled with 50 mg of Porapak R. Fig. 1 shows the splitter and traps. The ends of the splitter and the trap were butted with a thin silicone-rubber tube. The traps were replaced manually at the termination of the elution of the substance zones.

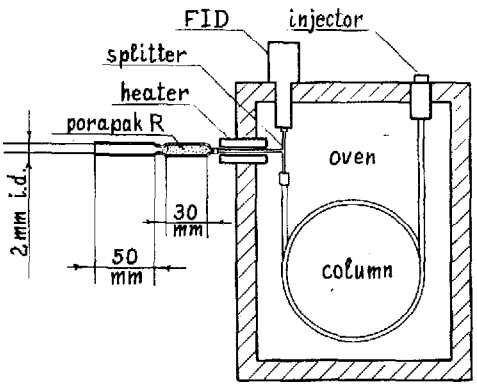


Fig. 1. Splitter and traps.

Desorption of the substances from the traps was carried out by passing through diethyl ether in the direction opposite to adsorption. For this purpose the traps were disconnected from the chromatograph and butted with a silicone-rubber tube to the extruded end of a 100 × 2 mm I.D. cylindrical glass receiver. The trap was filled with 35 μl of diethyl ether and the receiver was placed in liquid nitrogen. The elution lasted 1–3 min until a 1–2 mm height of the diethyl ether layer remained above the top of the Porapak layer. Subsequently the trap was disconnected and the receiver was sealed and heated to room temperature. The GC fractions thus obtained were analysed on each of the Ap, OV and Tr columns under the above conditions, the sample being 2 μl of solution. Retention indices of the components determined for each fraction (see Table III) were used in the design and testing of the B algorithm of the computer identification program for complex mixtures.

*Identification algorithm*

The identification algorithm was developed with the use of an experimental array of retention indices[9]. The overall data bank contained more than 2000 retention indices for 564 sulphur-containing compounds (SCCs) on four columns of different polarity.

The algorithm is shown in Fig. 2. It requires the following data to be stored in the computer memory:

$P$ = the number of groups or homologous series ($k$) of the substances sought. For instance, $k_1 = 1$ for $HSR_m$; $k = 2$ for $CH_3SR_m$; ..., $k = P$;

$Q$ = the number of stationary phases ($q$) with which GC analysis of the mixture was carried out ($q = 1, 2, ..., Q$);

$Z_j$ = the array of retention index values (numbered $j = 1, 2, ..., M_q$) of the sample mixture calculated from chromatograms for different sorbents;

$Z_i^0$ = experimental or calculated (eqn. 1) values of retention parameters of standards $i$ not referring to homologous series;

$S_{init.}$, $S_{max}$ = initial and maximal values, respectively, of permissible deviation ($\theta$) of calculated retention parameters $Z_j$ from their standard values $Z_j^0$;

$\Delta$ = a variation step of the permissible error $\theta$;

$F_1$ = number of columns allowing for the increase of the error;

TABLE III

RETENTION INDICES OF SULPHUR-CONTAINING COMPONENTS OF EACH OF THE ISOLATED FRACTIONS ON THREE COLUMNS AND IDENTIFICATION RESULTS OBTAINED WITH THE USE OF ALGORITHM *C*

| *Fraction No.* | *Stationary phase* | | | *Identification results* | |
|---|---|---|---|---|---|
| | *Apiezon M* | *OV-17* | *Triton X-305* | *No. identified* | *No. misidentified* |
| 1 | 628 | 711 | 817 | | |
| | 705 | 796 | 882 | 3 | 1 |
| | 768 | 891 | 1035 | | |
| 2 | 820 | 890 | 997 | 4 | 0 |
| | 846 | 1012 | 1049 | | |
| | 905 | 1059 | 1081 | | |
| | | 1077 | 1236 | | |
| | | | 1332 | | |
| 3 | 918 | 1012 | 1095 | 4 | 0 |
| | 936 | 1059 | 1176 | | |
| | 982 | 1111 | 1236 | | |
| | | 1133 | 1291 | | |
| 4 | 983 | 1111 | 1235 | 3 | 4 |
| | 1000 | 1147 | 1291 | | |
| | 1011 | 1199 | 1412 | | |
| | | 1228 | 1554 | | |
| 5 | 1011 | 1135 | 1353 | 2 | 0 |
| | 1056 | 1199 | 1412 | | |
| | | 1206 | | | |
| | | 1230 | | | |
| 6 | 1056 | 1147 | 1290 | 7 | 4 |
| | 1084 | 1206 | 1306 | | |
| | 1093 | 1269 | 1353 | | |
| | 1109 | 1283 | 1380 | | |
| | 1179 | 1296 | 1488 | | |
| | 1191 | 1316 | 1523 | | |
| | | 1342 | 1627 | | |
| 7 | 1222 | 1316 | 1553 | 1 | 0 |
| | | 1385 | | | |

$F_2$ = number of columns where no GC peak of the compound sought for is observed;

$\alpha, \beta, \gamma, \xi, A, B, C$ = coefficients of eqns. 1 and 2 for calculating $Z_i^0$ and boiling points ($T_b$) of the compounds:

$$Z_{m,q} = \alpha_{kq} + \beta_{kq}m + 1/m \cdot \gamma_{kq} \cdot \log m + \xi_{kq}/(m-2)^2 + 0.1 \quad (1)$$

$$T_{i,q} = A_{kq} + B_{kq}/(C_{kq} + Z_i) \quad (2)$$

where $m$ = the homologue number of the $R_mX$ homologous series. The program operation may be governed by any of the three algorithms *A*, *B* and *C*.

The form of eqn. 1 was obtained from a thermodynamic study of the GC

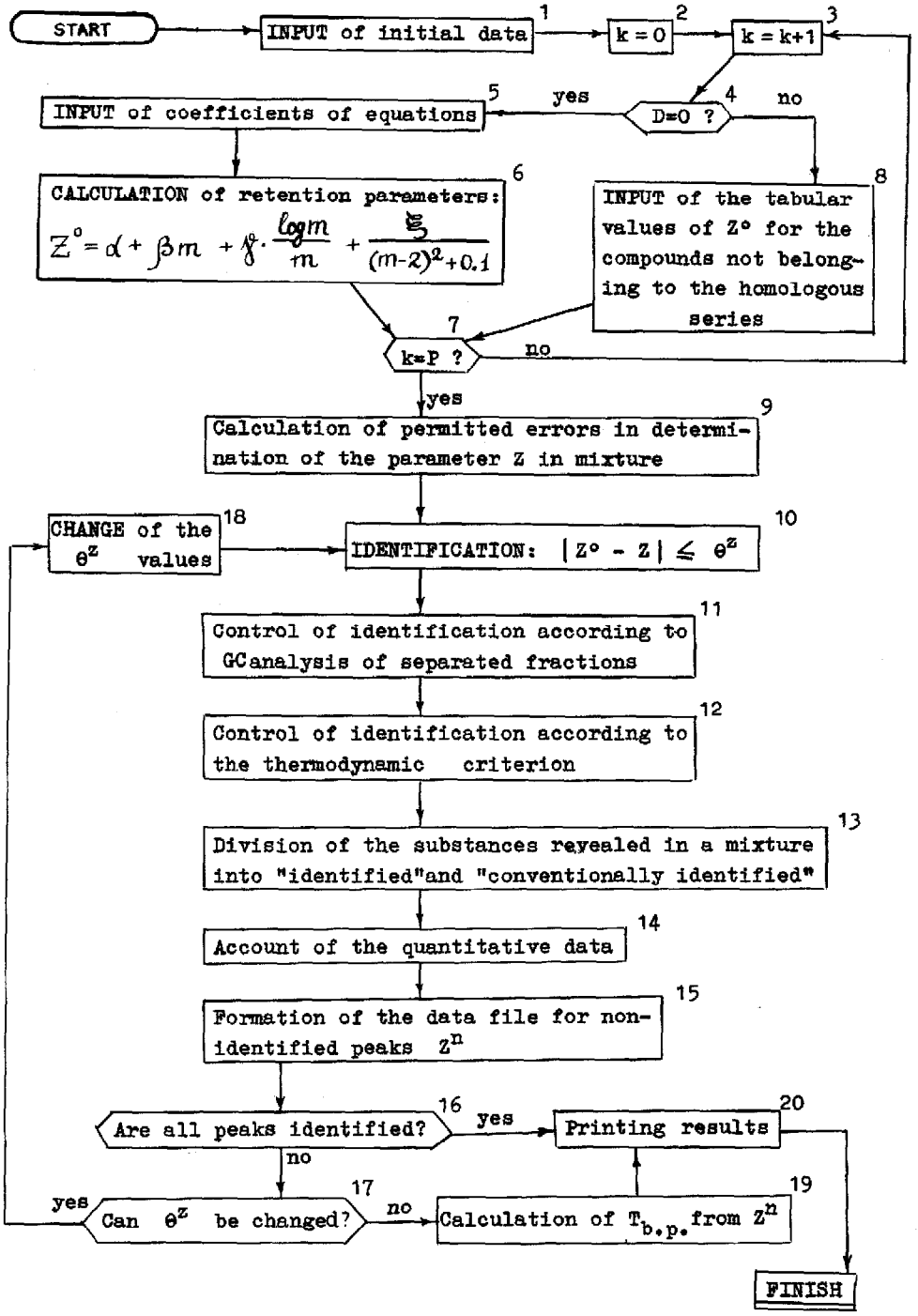


Fig. 2. Block diagram of the GC computer identification algorithm.

TABLE IV

COEFFICIENTS OF EQN. 1 FOR CALCULATING THE RETENTION INDICES OF SULPHUR-CONTAINING COMPOUNDS AND STANDARD DEVIATIONS OF RETENTION INDICES CALCULATED FROM EXPERIMENTAL DATA FOR FOUR STATIONARY PHASES AT 130°C

The experimental values of retention indices for calculation of the coefficients were taken from ref. 9.

| *No.* | *Homologous series* | *Stationary phase* | *Coefficients of eqn. 1* | | | | $\sigma$ |
|---|---|---|---|---|---|---|---|
| | | | $\alpha$ | $\beta$ | $\gamma$ | $-\xi$ | |
| 1 | $HSR_m$ | Apiezon M | 331.9 | 100.88 | 61.63 | −0.48 | 0.8 |
| | ($m$ = 2–12) | OV-17 | 419.87 | 100.51 | 105.76 | −0.61 | 0.5 |
| | | Triton X-305 | 523.28 | 100.87 | 150.69 | −0.64 | 0.4 |
| | | PEG 1000 | 587.01 | 100.33 | 186.66 | −0.66 | 0.6 |
| 2 | $CH_3SR_m$ | Apiezon M | 422.58 | 100.18 | 55.60 | 0.15 | 0.4 |
| | ($m$ = 1–12) | OV-17 | 510.40 | 100.42 | 57.30 | 0.23 | 0.5 |
| | | Triton X-305 | 609.76 | 100.78 | 86.00 | 0.38 | 0.6 |
| | | PEG 1000 | 673.19 | 99.74 | 149.12 | 0.90 | 1.1 |
| 3 | $C_2H_5SR_m$ | Apiezon M | 516.83 | 98.87 | 131.10 | 0.32 | 0.6 |
| | ($m$ = 1–10) | OV-17 | 605.48 | 98.95 | 143.81 | 0.64 | 0.8 |
| | | Triton X-305 | 700.75 | 98.25 | 224.18 | 1.18 | 1.1 |
| | | PEG 1000 | 761.04 | 97.16 | 243.38 | 0.85 | 1.2 |
| 4 | $n$-$C_3H_7SR_m$ | Apiezon M | 616.88 | 97.70 | 151.69 | 0.47 | 1.0 |
| | ($m$ = 1–9) | OV-17 | 703.52 | 97.82 | 161.56 | 0.73 | 1.0 |
| | | Triton X-305 | 799.59 | 96.53 | 264.39 | 0.94 | 1.8 |
| | | PEG 1000 | 853.44 | 96.30 | 283.79 | 1.39 | 1.4 |
| 5 | $n$-$C_4H_9SR_m$ | Apiezon M | 717.36 | 97.15 | 164.07 | 0.61 | 1.0 |
| | ($m$ = 1–9) | OV-17 | 805.18 | 97.71 | 177.97 | 0.61 | 0.4 |
| | | Triton X-305 | 899.37 | 95.64 | 273.36 | 1.18 | 1.5 |
| | | PEG 1000 | 954.64 | 95.20 | 303.37 | 1.38 | 1.6 |
| 6 | $n$-$C_5H_{11}SR_m$ | Apiezon M | 818.66 | 96.80 | 179.01 | 0.68 | 1.1 |
| | ($m$ = 1–8) | OV-17 | 908.0 | 96.49 | 189.82 | 0.67 | 1.0 |
| | | Triton X-305 | 999.67 | 94.98 | 274.28 | 1.08 | 1.3 |
| | | PEG 1000 | 1054.84 | 94.89 | 314.63 | 1.50 | 2.0 |
| 7 | $n$-$C_6H_{13}SR_m$ | Apiezon M | 920.04 | 95.66 | 163.95 | 0.53 | 0.4 |
| | ($m$ = 1–7) | OV-17 | 1010.53 | 95.16 | 175.99 | 0.57 | 0.4 |
| | | Triton X-305 | 1102.5 | 93.73 | 274.49 | 1.04 | 0.7 |
| | | PEG 1000 | 1157.48 | 92.77 | 284.07 | 1.08 | 0.9 |
| 8 | *iso*-$C_3H_7SR_m$ | Apiezon M | 578.86 | 96.05 | 179.1 | 0.20 | 0.2 |
| | ($m$ = 1–8) | OV-17 | 665.46 | 95.90 | 204.82 | 0.67 | 1.2 |
| | | Triton X-305 | 752.56 | 93.89 | 282.24 | 0.72 | 0.7 |
| | | PEG 1000 | 814.75 | 93.44 | 336.53 | 0.83 | 2.2 |
| 9 | *iso*-$C_4H_9SR_m$ | Apiezon M | 679.54 | 95.21 | 155.01 | 0.34 | 0.1 |
| | ($m$ = 1–7) | OV-17 | 764.06 | 94.58 | 179.9 | 0.49 | 0.3 |
| | | Triton X-305 | 843.19 | 93.31 | 223.31 | 0.68 | 0.7 |
| | | PEG 1000 | 914.12 | 91.50 | 314.56 | 0.63 | 0.6 |
| 10 | *sec.*-$C_4H_9SR_m$ | Apiezon M | 686.28 | 93.64 | 171.18 | 0.22 | 0.7 |
| | ($m$ = 1–7) | OV-17 | 773.46 | 93.34 | 211.85 | 0.38 | 0.7 |
| | | Triton X-305 | 856.31 | 91.32 | 270.64 | 0.58 | 0.8 |
| | | PEG 1000 | 919.30 | 90.06 | 322.91 | 0.92 | 0.8 |
| 11 | *tert.*-$C_4H_9SR_m$ | Apiezon M | 629.88 | 93.94 | 193.31 | 0.32 | 0.4 |
| | ($m$ = 1–8) | OV-17 | 709.86 | 92.89 | 204.68 | 0.42 | 0.5 |
| | | Triton X-305 | 789.94 | 89.87 | 253.27 | 0.34 | 0.3 |
| | | PEG 1000 | 849.27 | 89.44 | 313.75 | 0.51 | 0.6 |
| 12 | *iso*-$C_5H_{11}SR_m$ | Apiezon M | 784.50 | 95.22 | 168.66 | 0.45 | 0.2 |
| | ($m$ = 1–7) | OV-17 | 871.08 | 94.64 | 179.36 | 0.46 | 0.2 |
| | | Triton X-305 | 958.12 | 92.53 | 259.74 | 0.60 | 0.9 |
| | | PEG 1000 | 1021.11 | 91.29 | 294.03 | 0.86 | 0.8 |

*(Continued on p. 70)*

TABLE IV (*continued*)

| *No.* | *Homologous series* | *Stationary phase* | *Coefficients of eqn. 1* | | | | *σ* |
|---|---|---|---|---|---|---|---|
| | | | *α* | *β* | *γ* | $-\xi$ | |
| 13 | *sec.-iso*-$C_5H_{11}SR_m$ | Apiezon M | 768.80 | 92.84 | 209.68 | 0.51 | 0.5 |
| | ($m$ = 1–6) | OV-17 | 850.94 | 92.73 | 241.05 | 0.61 | 0.9 |
| | | Triton X-305 | 935.85 | 88.48 | 279.85 | 0.94 | 1.2 |
| | | PEG 1000 | 995.90 | 88.58 | 374.16 | 0.83 | 0.8 |
| 14 | *tert.*-$C_5H_{11}SR_m$ | Apiezon M | 754.75 | 92.86 | 212.35 | 0.55 | 0.9 |
| | ($m$ = 1–7) | OV-17 | 833.91 | 92.61 | 245.16 | 0.68 | 0.5 |
| | | Triton X-305 | 913.24 | 90.19 | 305.64 | 0.84 | 0.8 |
| | | PEG 1000 | 976.31 | 87.97 | 333.95 | 1.01 | 0.8 |
| 15 | $CH_3SSR_m$ | Apiezon M | 667.28 | 97.41 | 88.01 | 0.41 | 0.1 |
| | ($m$ = 1–6) | OV-17 | 790.83 | 96.83 | 100.34 | 0.46 | 0.1 |
| | | Triton X-305 | 941.95 | 93.49 | 164.49 | 0.79 | 0.9 |
| | | PEG 1000 | 1025.13 | 93.20 | 205.23 | 0.94 | 0.9 |
| 16 | $C_2H_5SSR_m$ | Apiezon M | 757.44 | 95.25 | 115.86 | 0.45 | 0.7 |
| | ($m$ = 1–6) | OV-17 | 877.56 | 95.02 | 142.06 | 0.58 | 0.3 |
| | | Triton X-305 | 1019.05 | 92.45 | 224.11 | 0.78 | 0.3 |
| | | PEG 1000 | 10.98 | 90.68 | 215.18 | 0.94 | 1.5 |
| 17 | *n*-$C_3H_7SSR_m$ | Apiezon M | 851.91 | 93.73 | 123.14 | 0.52 | 0.9 |
| | ($m$ = 1–6) | OV-17 | 970.99 | 94.50 | 153.72 | 0.71 | 0.2 |
| | | Triton X-305 | 1106.32 | 90.89 | 245.44 | 1.09 | 0.8 |
| | | PEG 1000 | 1184.61 | 88.66 | 221.22 | 1.04 | 1.4 |
| 18 | *n*-$C_4H_9SSR_m$ | Apiezon M | 949.60 | 93.93 | 160.20 | 0.67 | 0.3 |
| | ($m$ = 1–6) | OV-17 | 1019.1 | 97.2 | 47.00 | 0.84 | 1.7 |
| | | Triton X-305 | 1276.3 | 88.87 | 258.4 | 1.19 | 0.8 |
| | | PEG 1000 | 1069.4 | 94.09 | 152.33 | 0.74 | 0.9 |
| 19 | *n*-$C_5H_{11}SSR_m$ | Apiezon M | 1048.52 | 93.02 | 155.39 | 0.68 | 0.3 |
| | ($m$ = 1–5) | OV-17 | 1167.77 | 92.62 | 158.08 | 0.88 | 0.5 |
| | | Triton X-305 | 1296.63 | 89.46 | 254.33 | 1.28 | 1.0 |
| | | PEG 1000 | 1373.66 | 88.35 | 272.18 | 1.36 | 1.0 |
| 20 | *iso*-$C_3H_7SSR_m$ | Apiezon M | 811.87 | 93.51 | 139.60 | 0.71 | 0.1 |
| | ($m$ = 1–6) | OV-17 | 939.09 | 92.44 | 164.99 | 0.59 | 0.2 |
| | | Triton X-305 | 1064.00 | 92.87 | 386.76 | 0.35 | 0.4 |
| | | PEG 1000 | 1144.09 | 92.73 | 416.75 | 0.42 | 0.4 |
| 21 | *tert.*-$C_4H_9SSR_m$ | Apiezon M | 861.73 | 88.41 | 87.90 | 0.21 | 0.3 |
| | ($m$ = 1–5) | OV-17 | 967.03 | 87.13 | 81.28 | 0.41 | 0.4 |
| | | Triton X-305 | 1082.92 | 87.46 | 243.28 | 0.88 | 0.4 |
| | | PEG 1000 | 1150.12 | 86.76 | 259.92 | 1.45 | 0.4 |
| 22 | $CH_2=CHSR_m$ | Apiezon M | 509.88 | 98.27 | 65.56 | −0.09 | 0.2 |
| | ($m$ = 1–7) | OV-17 | 618.15 | 97.68 | 132.89 | 0.31 | 0.3 |
| | | Triton X-305 | 737.90 | 96.82 | 183.54 | 0.46 | 0.5 |
| | | PEG 1000 | 816.20 | 95.32 | 220.92 | 0.7 | 1.3 |
| 23 | $CH\equiv CCH_2SR_m$ | Apiezon M | 607.84 | 97.16 | 144.42 | 0.12 | 0.6 |
| | ($m$ = 1–7) | OV-17 | 761.60 | 97.23 | 138.54 | 0.33 | 0.5 |
| | | Triton X-305 | 994.12 | 95.37 | 253.76 | 0.77 | 0.6 |
| | | PEG 1000 | 1108.35 | 94.81 | 267.32 | 1.16 | 0.8 |
| 24 | $CH_2=CHCH_2SR_m$ | Apiezon M | 599.83 | 96.98 | 135.66 | 0.33 | 0.5 |
| | ($m$ = 1–7) | OV-17 | 700.94 | 97.53 | 136.73 | 0.72 | 0.6 |
| | | Triton X-305 | 824.62 | 95.6 | 209.09 | 1.02 | 1.5 |
| | | PEG 1000 | 905.58 | 94.75 | 251.08 | 0.92 | 1.0 |
| 25 | $C_6H_5SR_m$ | Apiezon M | 1013.59 | 92.67 | 251.90 | −0.10 | 0.4 |
| | ($m$ = 1–5) | OV-17 | 1173.36 | 93.50 | 326.01 | 0.37 | 0.4 |
| | | Triton X-305 | 1402.49 | 86.67 | 423.33 | 0.09 | 0.5 |
| | | PEG 1000 | 1532.22 | 85.46 | 509.50 | 0.66 | 0.6 |

TABLE IV *(continued)*

| *No.* | *Homologous series* | *Stationary phase* | *Coefficients of eqn. 1* | | | | $\sigma$ |
|---|---|---|---|---|---|---|---|
| | | | $\alpha$ | $\beta$ | $\gamma$ | $-\xi$ | |
| 26 | $CH_3C(O)SR_m$ | Apiezon M | 592.71 | 96.59 | 195.80 | −0.24 | 0.3 |
| | ($m$ = 1–5) | OV-17 | 737.01 | 97.07 | 216.00 | −0.04 | 0.4 |
| | | Triton X-305 | 903.25 | 94.65 | 329.62 | 0.30 | 0.3 |
| | | PEG 1000 | 995.44 | 94.25 | 390.19 | 0.58 | 0.2 |
| 27 | $C_2H_5C(O)SR_m$ | Apiezon M | 686.18 | 98.65 | 249.99 | 0.32 | 0.2 |
| | ($m$ = 1–5) | OV-17 | 833.12 | 95.88 | 228.63 | 0.15 | 0.4 |
| | | Triton X-305 | 987.04 | 93.74 | 397.21 | 0.43 | 0.4 |
| | | PEG 1000 | 1067.96 | 92.83 | 298.72 | 0.44 | 0.4 |
| 28 | $C_3H_7C(O)SR_m$ | Apiezon M | 778.06 | 95.93 | 216.20 | 0.16 | 0.3 |
| | ($m$ = 1–5) | OV-17 | 911.79 | 103.55 | 331.51 | 0.9 | 0.3 |
| | | Triton X-305 | 1064.00 | 92.87 | 286.76 | 0.35 | 0.4 |
| | | PEG 1000 | 1144.09 | 92.73 | 416.75 | 0.42 | 0.4 |

behaviour of the homologues of aliphatic compounds[10]. It has certain advantages over eight other equations in the literature. The coefficients of eqn. 1 and standard deviations from experimental values of the retention indices calculated for 28 SCCs homologous series are listed in Table IV. A knowledge of four or five retention indices of homologues of each homologous series $R_mX$ which include the parameters of the first three terms is sufficient for calculating the equation coefficients. The equations obtained were then used to seek the retention parameters of all homologues of the series starting with the first one. In this sense the method of identification suggested in ref. 2, which does not require the use of a large number of standard substances, is referred to as standardless.

The algorithm developed involves two identification techniques: the standardless technique (parameter $D = 0$ in Fig. 2) with which the retention indices are calculated from the universal eqn. 1 and the technique of determining substances by table-specified retention indices ($D = 1$). Identification of substances by direct comparison of their sorption characteristics with the corresponding table-specified retention parameters of standard compounds (block 5 in Fig. 2) is particularly important in the determination of heterocyclics and compounds with several functional groups such as thioalkanethiols, thiophenes, sulphur-containing furans and compounds with non-symmetric branched radicals not belonging to homologous series.

The program is designed for the simultaneous analysis of data obtained from four columns and recognizes that each chromatogram may contain up to 200 peaks. These parameters may be automatically replaced with any other parameters. In this respect, three stationary phases were considered ($Q = 3$).

In contrast with ref. 7, the algorithm (see Fig. 2) automatically changes the substance determination limits ($\theta$) in identifying mixture ingredients. The need to extend the range of error $\theta$ in the course of the identification is dictated by a number of reasons, of which the major ones are as follows:

(1) the $Z_{iq}^0$ values of standards are determined for amounts from 0.01 to 10 $\mu$g, whereas the concentrations of substances in natural mixtures may be even less and, consequently, the errors in determining $Z_j$ may be higher;

TABLE V

PERMISSIBLE ERROR RANGE IN IDENTIFICATION BY RETENTION INDICES ON COLUMNS WITH STATIONARY PHASES OF DIFFERENT POLARITY (ALGORITHMS *B* AND *C*)

| *Apiezon M* | | *OV-17* | | *Triton X-305* | |
|---|---|---|---|---|---|
| *Range of retention index values* | $\theta_{max}$* | *Range of retention index values* | $\theta_{max}$* | *Range of retention index values* | $\theta_{max}$* |
| 500–750 | 13 | 600– 900 | 16 | 700– 900 | 16 |
| 750–900 | 6 | 900–1100 | 5 | 900–1200 | 5 |
| >900 | 5 | 1100–1300 | 6 | >1200 | 7 |
| | | >1300 | 7 | | |

* $\theta_{max}$ = Maximal error value.

(2) the actual GC conditions for the anlysis of a mixture may differ slightly from those under which the $Z_{iq}^0$ values of standard compounds were obtained because of differences in the chromatographic columns used and the GC conditions;

(3) incomplete separation of chromatographic zones of the mixture under analysis increases (or decreases) $Z_j$ compared with the standard values.

To improve the reliability of GC computer identification, theoretically based limits of the range of variation in $\theta$ (see Table V) have been entered into the program. A comparison of the data in Tables I, II and V shows that the range of variation in $\theta$ permissible in GC identification exceeds the confidence interval for the retention indices of a mixture (see Tables I and II). For this effect, we have established the permissible range of variation in $\theta$ in identification ($S_{min} \leqslant \theta \leqslant S_{max}$) on the basis of actual identification situations where the reproducibility of the retention indices of standards depends not only on the accuracy of the determination of the retention indices in a given set of serial measurements (as depicted in Tables I and II), but also on the reproducibility of column preparation techniques and the stability of the sorbent with time[11]. $S_{min}$ and $S_{max}$ were obtained (see Table V) with the use of experimental retention indices for a large number of standard SCCs obtained by different workers with different times of operation of the columns, from several months to 4 years starting from their preparation[9]. As can be seen from Table V, the value of $\theta$ depends on the type of stationary phase and the retention index of the substance under analysis. Scanning the standard compounds, the computer calculates from the retention indices on each stationary phase the range of $\theta$ permissible in the identification of a given compound.

Every step of the increase in $\theta$ permits automatic control of identification results based on chromatographic data for individual fractions and the thermodynamic identification criterion $\Delta Q$ (block 12 in Fig. 2)[12]. Introducing the thermodynamic criterion $\Delta Q$, one can avoid the need for values of $\Delta I$. As identification from $\Delta I$ values might result in an erroneous determination of the functional group owing to identical $\Delta I$ values for substances of different classes and non-constancy of these values within a homologous series[8,12], the replacement of $\Delta I$ with $\Delta Q$ in the algorithm improves the reliability of identification.

Following the identification with a given value of $\theta$, the algorithm provides for subdivision of the substances detected in the mixture into "identified" and "conventionally identified" (block 13 in Fig. 2). A compound displayed on at least one chromatographic column as an individual zone is regarded as identified, whereas that displayed on all columns as a mixture with other substances is regarded as conventionally identified[8,12].

In compliance with the algorithm presented in Fig. 2, mixture components may be identified in two different modes. In mode 1 the computer changes the values of $\theta$ for all the columns simultaneously and the process of identification is multiply repeated with a value of $\theta$ chosen automatically. The initial values $\theta = S_{min}$ are identical with the accuracy of experimental determination, and for complex mixtures these are normally 1–3 retention index units (i.u.). In this mode the computer operates until a conventionally identified substance appears.

The major difference between modes 1 and 2 is that in each step of mode 2 only those compounds are identified which are displayed, at least on one of the columns, by a GC zone not identified in the previous steps. The computer stores the obtained results in its memory and makes full use of them.

In mode 2, $\theta$ does not change simultaneously for all columns as in mode 1, but gradually. The number of columns allowing for the error increase is set by parameter $F_1$, (block 18 in Fig. 2). This approach, never employed before, permits the number of misidentified compounds to be sharply decreased.

When identifying natural mixtures of trace organic substances, their GC zones may often be displayed on the chromatogram from only some of the columns or may be eluted from some stationary phases with a very significant deviation from standard values. This is often the case when a chromatogram displays a peak of an admixture immediately following a peak of the substance present in the mixture in large amounts[13]. Large errors (5–20 i.u.) are obtained in the analysis of compounds whose GC zones are at the beginning of chromatograms (see Tables I, II and V). The algorithm suggested permits the identification of such compounds. In such a case, parameter $F_2$ (block 18 in Fig. 2) determines the number of columns that may display no peaks of the compound under analysis.

On termination of its operation, the computer prints out the results (block 20 in Fig. 2). First to be printed is a data array (names of substances and their retention indices in the mixture on all stationary phases) on identified compounds, and then on conventionally identified compounds. For each stationary phase retention indices of unidentified GC peaks $Z_{jq}^{n}$ are also printed out. For these components, boiling points may be calculated from the $Z_{jq}^{n}$ values by eqn. 2 and printed out (block 19 in Fig. 2).

The program is written in ALGOL-60. The identification of a 60-component mixture whose chromatograms were obtained on three stationary phases requires 15 min of BESM-6 computer time if the presence of 300 compounds is to be detected.

## DISCUSSION

SCCs play an important role in creating flavours in various foodstuffs[14] and are key compounds in natural boiled and fried meat odours. The complication of SCC identification arises from their low concentrations in natural products ($10^{-4}$–

$10^{-12}$%)[15]. Several hundred SCCs may participate in the formation of an aroma[14], which also hinders their identification. A particular feature of this class of substances is the low specific sorption activity of the sulphur-containing functional groups. The observed sorbate–sorbent donor–acceptor interactions are weak. The sorption properties of HS–, –S– and –SS– functional groups, for instance, are similar, which makes it difficult to separate and identify such substances in a mixture. $\Delta I$ values cannot be used for establishment of the composition of complex SCC mixtures as in the homologous series we studied the $\Delta I$ values are either close to each other or identical[12] and, consequently, do not provide the information on the structures of functional groups required for the identification of such compounds. Various problems in SCC identification have dictated the choice of the particular class of organic compounds that can be subjected to the GC computer identification algorithm designed for multi-component mixture analysis.

An artificial 23-component SCC mixture was used to investigate the problems arising in establishment of the composition of a complex mixture that may not be completely resolved on a particular column. The mixture contained compounds often encountered in natural products: mercaptans, aliphatic and aromatic sulphides, ethers of thioacids, thioacetals, thiophenes and sulphur-containing heterocyclics (dithianes, dithiolanes) at concentrations from $0.1 \cdot 10^{-5}$ to $4 \cdot 10^{-5}$%. The components were chosen so that the number of coincidences of their retention parameters with parameters of other SCCs stored in the computer data bank was maximal. For example, the retention indices of 2-formylthiophene on Ap M coincides within ± 5 i.u. with those of ten other compounds: $CH_2SC_5H_{11}$, *tert.*-$C_5H_{11}SC_3H_7$, *iso*-$C_3H_7$S-*iso*-$C_5H_{11}$, (*iso*-$C_4H_9)_2S$, $CH_2=CHSC_5H_{11}$, $CH_2=CHCH_2$S-*sec.*,*iso*-$C_5H_{11}$, $CH_2=CHCH_2$S-*tert.*-$C_5H_{11}$, $CH\equiv$CS-*sec.*,*iso*-$C_5H_{11}$, $C_6H_5SH$ and $CH_3S_2$-*iso*-$C_4H_9$. With the same accuracy the retention indices of 2-formylthiophene on Tr X-305 and OV-17 coincide with those of ten and fourteen other substances, respectively. Similar results were obtained for a number of other components of the mixture being analysed. Hence the task set for the computer deliberately complicated the problems encountered.

The suggested algorithm is intended for identification in different modes. To choose the best algorithm for GC computer analysis of the composition of a complex mixture, we studied the effect of the type of identification algorithm on the adequacy of the results. Three types of the identification algorithm executed by the program, *A*, *B* and *C*, were considered.

(A) Algorithm *A* simulates the manual identification process. Its underlying principle is that used in manual identification: a comparison with data for standard compounds. The identification is carried out in a single step. The permissible deviation of the retention parameters of the components under analysis from the corresponding standards are the same for all the substances considered and are unchanged in the course of the process. This algorithm is normally employed in the manual identification of the components of mixtures.

Table VI presents the results obtained for an artificial mixture of test substances, and provides evidence of the particular importance of an appropriate choice of the permissible error range $\theta$. The effect of this choice on the result is clearly exemplified with the use of algorithm *A*. Thus, for instance, with $\theta = 3$ i.u. the number of compounds identified in the mixture was five (four which were identified

TABLE VI

EFFECT OF THE PERMISSIBLE ERROR VALUE $\theta$ ON THE GC COMPUTER IDENTIFICATION RESULTS FOR A 23-COMPONENT TEST MIXTURE USING ALGORITHMS *A*, *B* AND *C*

| $\theta$ | *No. of identified compounds* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | *Correct algorithm* | | | *Incorrect algorithm* | | | *Conventional algorithm* | | |
| | *A* | *B* | *C* | *A* | *B* | *C* | *A* | *B* | *C* |
| 3 | 4 | 4 | 4 | 1 | 0 | 0 | – | 1 | 1 |
| 4 | 8 | 8 | 8 | 4 | 3 | 1 | – | 1 | 1 |
| 5 | 16 | 16 | 16 | 13 | 6 | 0 | – | 6 | 5 |
| 7 | 19 | 18 | 19 | 26 | 4 | 1 | – | 10 | 7 |
| 9 | 19 | 18 | 19 | 50 | 4 | 1 | – | 10 | 7 |
| 11 | 20 | 19 | 20 | 81 | 4 | 1 | – | 10 | 67 |
| 16 | 22 | 21 | 22 | 122 | 4 | 1 | – | 11 | 8 |

correctly and the other misidentified owing to a fortuitous coincidence of retention indices), and with $\theta = 7$ i.u. the number identified was 45. The number of misidentified substances, *i.e.*, those not present in the mixture and "determined" only owing to coincidence of their retention indices, sharply increases. The applicability of algorithm *A* is obviously efficient only with $\theta \leqslant 5$ i.u. The results of using algorithm *A* for identification with $\theta = 5$ i.u. were as follows: 16 compounds out of the total of 23 were determined correctly and almost as many (13) were misidentified (see Table VI). This proves that manual identification techniques are inapplicable to the analysis of such complex mixtures.

(B) In comparison with *A*, algorithm *B* implies automatic variation of the permissible range $\theta$ and employs the universal eqn. 1 for calculating standard retention indices. The identification is executed in several steps. At every step $\theta$s are automatically increased with regard to their limiting values (see Table V) for each substance analysed and each stationary phase. Both identified and conventionally identified substances are listed out by the printer. The efficiency of the algorithm has been checked in the analysis of nitrogen-containing bases[16] and fatty acid methyl ethers[17].

According to algorithm *B*, identification is executed in several steps, for instance seven steps as in Table VI. Table VI shows that the use of algorithm *B* (see Fig. 2) decreases the number of misidentified substances considerably. With $\theta \geqslant 7$ i.u. their number remains the same, in contrast with the results obtained with the use of algorithm *A*. A test mixture of 23 compounds was used to determine the efficiency of the algorithm. As a result of identification, the computer listed out 37 compounds; 22 were those actually present in the mixture and 15 were displayed owing to coincidence of the retention indices stored in the computer with those of the components of the mixture under analysis. It is particularly important that eleven out of fifteen the substances were determined as conventionally identified, *i.e.*, as those whose actual presence in the mixture must be verified by some other technique, which reduces the probability of incorrect identification.

Some 11% of all chromatographic peaks remain unidentified. As a rule, these are the peaks of poorly separated substances and the peaks of the initial sections of the chromatograms.

In the analysis of complex mixtures, the probability of overlapping of chromatographic zones is known to increase hyperbolically with increase in the number of components[18]. The use of multi-dimensional chromatography is a means of improving the reliability of SCC identification in complex mixtures. Therefore, we considered the third type of the identification algorithm (*C*).

(c) GC identification proceeds as in algorithm *B* but the results of the GC separation of individual fractions on different stationary phases are additionally considered.

Table I lists the test substances included in the test mixture. For each of the substances, the minimal value of $\theta$ with which the substance was determined is presented. It can be seen that algorithm *C*, like *B*, could identify 22 compounds out of 23; $(CH_3)_2S$ (No. 1) could not be detected on the chromatogram recorded at 130°C as its boiling point is too low (70°C). Table I also shows that the automatic choice of the deviation range provides identification of most compounds with $\theta \leqslant 7$ i.u. With significant errors (9–16 i.u.) only peaks in the initial sections of the chromatograms and compounds whose concentrations in the mixture were low enough were identified. A more preciese assessment of the results with the help of the GC separation of the fractions (algorithm *C*), as shown in Table VI, allows the number of misidentified substances to be reduced to nine, with eight substances regarded by the computer as "conventionally identified". These are printed in a separate list and their presence in the mixture has to be checked by some other technique. One should expect such a situation in the analysis of natural products as it is impossible to separate completely all components in practical situations.

The retention indices of the substances in the test mixture in Table III are divided into groups according to the results of the analysis of the fractions. The number of misidentified compounds is given for each retention index group. When the number of peaks in the chromatogram of a fraction doubles (see Table III) the number of errors increases four-fold (fractions 1 and 4). Hence the identification results depend on the number of substances in the fraction. Obviously there is no need to separate the mixture under analysis into components of fractions 3–4 as they may actually contain a greater number of compounds. However, one should bear in mind that the increase in the complexity of the composition of the fractions results in a sharp increase in the probability of misidentification. It is evident that the prior chemical separation of components into classes with identical functional groups[23] is also necessary in order to reduce the number of overlaps of the chromatographic zones.

The above results indicate that the use of algorithms *B* and *C* makes this method comparable, in terms of its potential, to techniques such as gas chromatography–mass spectrometry (GC–MS). For instance, the application of GC–MS to the analysis of an artificial mixture of 45 components allowed the correct identification of only 35 substances[19]. The major reason for misidentification in GC–MS is incomplete GC separation. Owing to overlapping of chromatographic zones in the analysis of multi-component mixtures, the GC computer identification technique suggested here gives results comparable to those obtained with the use of the most advanced techniques (GC–MS, GC–Fourier transform infrared spectrometry, etc.). In the analysis of mixtures with trace components, the proposed computerized GC identification method combined with preliminary chemical separation

of substances into classes is the only reliable way to establish the composition of the mixture.

The results obtained for the artificial SCC complex mixture allow algorithms *B* and *C* to be recommended for the establishment of the composition of unknown multi-component mixtures. The use of GC computer identification considerably reduces the number of misidentified compounds as compared with manual identification and improves the reliability of the results.

## REFERENCES


1 L. A. Gribov, Yu. A. Zolotov, V. I. Kalmanovskii, L. L. Kunin, Yu. M. Luzhkov, A. A. Popov and V. S. Toroncev, *Zh. Anal. Khim.*, 37 (1982) 1104.
2 R. V. Golovnya, *Usp. Khim.*, 14 (1976) 1825.
3 K. R. Betty and F. W. Karasek, *J. Chromatogr.*, 166 (1978) 111.
4 B. E. Blaisdell, S. C. Gates, F. E. Martin and C. C. Sweeley, *Anal. Chim. Acta*, 117 (1980) 35.
5 J. M. Owen and A. D. Dale, *J. Chromatogr.*, 47 (1975) 2385.
6 L. Kampman, *J. Chromatogr.*, 150 (1978) 367.
7 D. N. Grigoryeva, R. V. Golovnya, I. L. Zhuravleva, N. I. Svetlova and G. B. Cergeev, *Zh. Anal. Khim.*, 34 (1979) 2434.
8 V. D. Shatz, A. A. Avots, A. M. Kophman and Ya. Ya. Silis, *Zh. Anal. Khim.*, 30 (1975) 2306.
9 R. V. Golovnya, V. G. Garbuzov, T. A. Misharina and A. F. Aerov, *Zh. Anal. Khim.*, 40 (1985).
10 R. V. Golovnya and D. N. Grigoryeva, *Chromatographia*, 17 (1983) 613.
11 N. I. Svetlova, D. N. Grigoryeva, I. L. Zhuravleva and R. V. Golovnya, *Zh. Anal. Khim.*, 39 (1984) 1282.
12 R. V. Golovnya and D. N. Grigoryeva, *J. Chromatogr.*, 290 (1983) 275.
13 I. Novak, *Kolichestvennii Analiz Methodom Gazovoi Khromatografii (Quantitative Analysis by Method of Gas Chromatography)*, Mir, Moscow, 1978, p. 179.
14 S. Van Straten, Fl. Vrijer and J. C. Beauveser, *Volatile Compounds in Food*, Krips Repro, Zeist, 1977, p. 162.
15 H. Maarse and R. Bels, *Isolation, Separation and Identification of Volatile Compounds in Aroma Research*, Akademie-Verlag, Berlin, 1981.
16 R. V. Golovnya, N. I. Svetlova, I. L. Zhuravleva and D. N. Grigoryeva, *Prik. Biokhim. Mikrobiol.*, 19 (1983) 277.
17 R. V. Golovnya, T. E. Kuzmenko, A. L. Samusenko and D. N. Grigoryeva, *Prikl. Biokhim. Mikrobiol.*, 17 (1981) 60.
18 D. A. Leathard and B. C. Shurlock, *Identification Techniques in Gas Chromatography*, Wiley-Interscience, New York, 1970, p. 44.
19 D. Rosenthal, *Anal. Chem.*, 54 (1982) 63.