Identification of Lower Terpenoids from Gas-Chromatography-Mass Spectral Data by On-Line Computer Method*

R.P. Adams, M. Granat, L.R. Hogge, and E. von Rudloff,** Prarie Regional Laboratory, National Research Council of Canada, Saskatoon, Saskatchewan, Canada S7N 0W9

Abstract

It was found that the less intense ion peaks of the mass spectra of lower terpenes can play an important role in distinguishing compounds that have similar mass spectra. Two computer programs were developed that generate, from combined gas chromatographic—mass spectra data, F-1 weights (ANOVA) of mass ion peaks between 41 and 300. The library program FIXLIB contains mass spectral and retention data from known mono- and sesquiterpenes, as well as some phenyl propanoid ethers, which are then used for identification of unknowns. The identification program GETID allows rapid comparison of mass spectra and retentions of compounds that fall within specified retention windows. Weighted similarity ratios are calculated which establish the relative similarities of an unknown with the reference compounds falling into a given retention window. An absolute similarity is then calculated for the unknown and that reference compound which has the closest fit, which provides the measure of confidence for the identification. The method was applied to the identification of lower terpenes and phenylpropanoid ethers found in the volatile leaf oils of three true fir and three juniper species, as well as two Douglas-fir varieties. All compounds of previously known identity were correctly identified by the GETID program.

Introduction

Unequivocal identification of the many different terpenes found in essential oils presents a formidable task (1,2). Many prefractionation and isolation steps are required to obtain the individual components in sufficient purity and amounts, and even then many minor or trace components may be missed or misidentified. Also, there is the danger of rearrangement, autoxidation, or polymerization (8) during such isolation procedures. In chemosystematic studies of conifer leaf oils (3), studies of different chemical races (4,5), or seasonal changes (6,7), identification of individual components may prove to be the most time consuming aspect. Many constituents of such oils are well known mono- and sesquiterpenes; thus, a fast and reliable identification method would be of great benefit. Since identification by relative retention characteristics on gas-liquid chromatographic (GC) columns of different polarity (2,8) is inadequate, the most advantageous method would be an on-line spectrometric one. Both combined GC-infrared (IR) and GC-mass spectrometry (MS) have seen much improvement in recent years, yet there are inherent problems. In GC-IR the spectra are recorded in the vapour phase, resulting in less well defined spectra than those recorded as liquid films or solid (K Br) disks (9). Also, certain terpenes with similar retention characteristics have rather similar infrared spectra (e.g., α-pinene and α-thujene; β-pinene and sabinene). The same problem may be encountered in GC-MS work. Comparison of published mass spectra (10) and those recorded in our chemosystematic studies (11,12) indicates that such structurally different terpenes as tricyclene, α-pinene, and car-3-ene show differences only in ion intensities of the principle ion peaks with identical mass. One suspects that ring opening upon electron impact to the same intermediate ion is the cause of such similar mass spectra. These minor differences are often within the normal operational variations and hence, make unequivocal identification by mass spectra alone impossible. This may also be the reason why in our own experience the library search techniques and data available to GC-MS users fail to give satisfactory results for mono- and sesquiterpenes. Use of the chemical ionization technique does not solve the problem because of even greater difficulties in obtaining reproducible mass ion intensities for a given compound.

To solve the problem of quick and reliable on-line identification we have studied various conditions of GC-MS operation and computerized treatment of mass spectra, as well as retention data. A reference library of known monoterpenes was established and the overall method was tested with the various components (known and unknown) of true fir, Douglas-fir, and juniper leaf oils.

Basic Concept and Requirements

At the start of this study it appeared that the use of relative retention times (RRT) coupled with mass spectral information

---

*NRCC No. 17144

**Present address: Science Research Center 8555 S. Escalante, Sandy, UT 84092

could be of considerable use with the terpenoids. This approach was also taken by Blaisdell (13) and Smith, et al. (14) who have reviewed the literature.

Since many terpenoids differ only quantitatively in their mass spectra (10) (see also Figure 1 comparing tricyclene and α-pinene), simple qualitative matching of a few intense mass ions is not sufficient to resolve the identities. For example, the matching of Blaisdell (13) uses only the most intense ion per 14 ions with a simple presence-absence matching coefficient (Sokal and Sneath 15, 16). Smith, et al. (14) use the correlation coefficient on 2 ions per 14 ions scored into qualitative states [i.e. 0,1,2,3]. Unfortunately, the correlation coefficient has at least three significant problems when used in classification of this manner.

First there is an *a priori* bias toward more abundant ions being more valuable for classification. Of course, this may or may not be true in the discrimination of a given set of compounds as we shall show. Secondly, when data are discontinuous, the correlation coefficient behaves statistically rather poorly (17). Thirdly, the correlation coefficient is not sensitive enough and is not linear in measuring similarities (18).

In order to obtain similarities that are sensitive enough to discriminate between the common terpenoids, we have used all 260 ions between mass 41 and 300, quantitatively varying between 0 and 100 percent of the base ion. GC-MS analysis shows that with a given set of chemically related compounds, some ions vary little within replicate runs (good fidelity) of a given compound and yet these ions discriminate well among the compounds in the group. This aspect is not new and is well known in biological classification. Adams (19) has shown that analysis of variance (ANOVA) of each character used in numerical taxonomy generates a set of weights (F-values) which do extremely well in low-weighting those characters (ions) that are mostly noise and/or fail to discriminate between items (compounds). At the same time these F-values are very high for characters (ions) that discriminate well between items (compounds).

We found, just as Blaisdell (13) points out, that comparison of unknowns with a library of known compounds seems to be the most successful way to identify compounds. Our programs are written in FORTRAN and all the work was done on the Finnigan 3300 GC-MS coupled with an INCOS Data System. Two significant problems in using quantitative data for all ions are storage and retrieval time. In order to reduce library search times and minimize storage, we have ordered entries in the library by each compound's relative retention time to an internal standard (IS), heptanyl acetate. This internal standard was chosen because it runs in a region of the chromatogram where the terpenes of conifer leaf oil samples are usually absent, and its fragmentation pattern is very sensitive to calibration conditions. The latter is an important point because in our experience, spectra of terpenoids, such as α-pinene, are very sensitive to MS conditions (Figure 2).

Building the library is the most critical portion of the work. Extreme precautions must be taken to insure that the GC-MS is calibrated as with previous runs.

The set of compounds to be searched when identifying an unknown is determined by the RRT of the unknown, the RRT of the compounds in the library (Table I), and the variances of the RRTs of the library. The library may be searched by means of the program called GETID (see below) for all compounds, or those compounds in the library between RRTs specified by the user (interactively), or automatically. Each compound in the library has 3 to 5 replicate runs stored for all ions (41 to 300) as well as replicates of the RRT. Since one needs the mean and variance to do ANOVA, only these two values per ion need to be stored. We actually store the sum, sum of squared values, number of replicates for all ions, and the RRT for all compounds in the library (Table II). This enables one to add additional replicates to the library and also compute ANOVA rather easily.

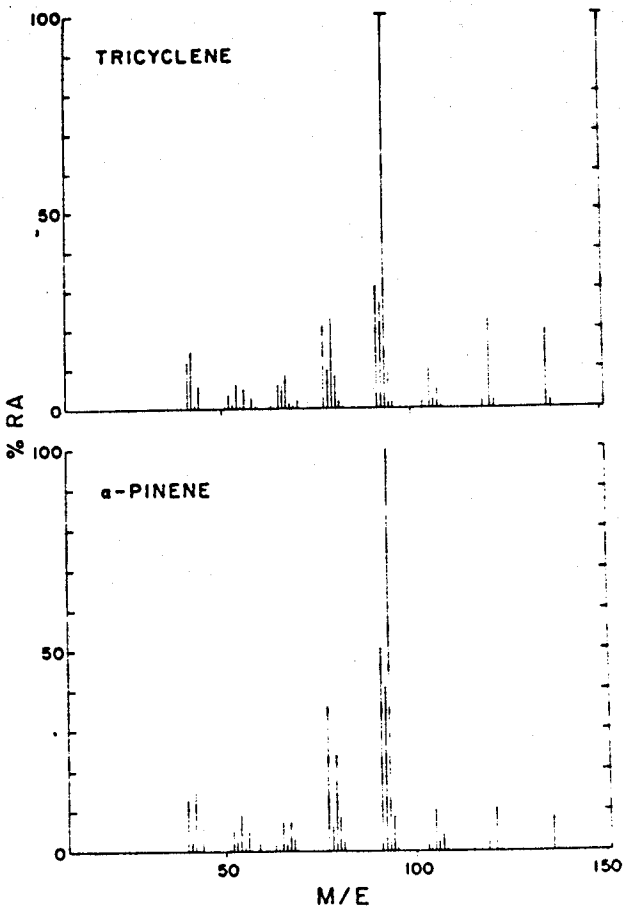

Figure 1. Comparison mass spectra and α-pinene.

The library is divided into 4 sections (Table II): general information, information concerning the IS, a directory, and compounds with their spectral data (sum and sums of squares). The directory enables the library curation and construction program (FIXLIB) and the identification program (GETID) to quickly determine the region of the library to be searched. A 5 times standard deviation value (Table II) is stored for the RRT of each compound in the library. The probability of a RRT being outside ± 5 standard deviations is about 5.7×10^{-5} (20). In practice this amounts to about the width of the base of a medium sized peak on an 1/8 inch packed column. If this window is thought to be too small, the number of standard deviations can be changed, interactively, to search a wider window, or one could take the option that allows one to specify what portion of the library is to be searched. In the automatic mode, GETID compares the unknown with only those library compounds that are within ± 5 standard deviations of the unknown's RRT.

After a window containing at least 2 library compounds is determined, GETID calculates the ANOVA for each of the ions (41 to 300) by: Let: CTERM = correction term; TOTSS = total sum of squares; TRTSS = treatment sum of squares;

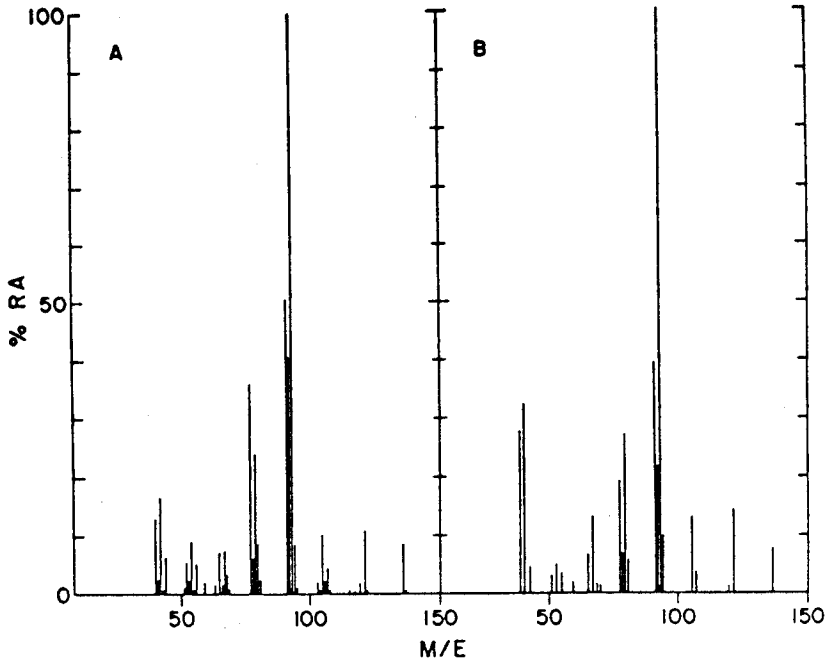

Figure 2. Mass spectra of 2 different calibrations of a-pinene to show how different the spectrum of the same compound can be.

Table I. Relative Retention Times (RRT) and the 5-Standard Deviation Windows of Some Conifer Leaf Oil Components (Internal Standard: *n*-Heptyl Acetate = 1.00)

Monoterpene Hydrocarbons	
Santene	0.221 ± 0.039
Tricyclene	0.268 ± 0.039
α-Pinene	0.298 ± 0.039
Camphene	0.364 ± 0.039
β-Pinene	0.446 ± 0.039
Sabinene	0.456 ± 0.049
Myrcene	0.525 ± 0.039
Car-3-ene	0.548 ± 0.039
α-Phellandrene	0.549 ± 0.039
α-Terpinene	0.592 ± 0.039
Limonene	0.613 ± 0.039
β-Phellandrene	0.637 ± 0.039
1:8-Cineole	0.658 ± 0.039
Cis-Ocimene	0.686 ± 0.015
γ-Terpinene	0.717 ± 0.044
Trans-Ocimene	0.718 ± 0.038
p-Cymene	0.736 ± 0.039
Terpinolene	0.815 ± 0.039
Oxygenated Monoterpenes	
Fenchone	1.034 ± 0.044
Thujone	1.084 ± 0.043
Isothujone	1.124 ± 0.011
Citronellal	1.210 ± 0.021
Fenchyl Acet.	1.270 ± 0.069
Linalool	1.274 ± 0.026
Camphor	1.303 ± 0.079
Linalyl Acet.	1.421 ± 0.029
Terpinen-4-ol	1.439 ± 0.099
Methyl Thymol	1.480 ± 0.063
Bornyl Acet.	1.528 ± 0.109
Pulegone	1.583 ± 0.040
Borneol	1.614 ± 0.053
α-Terpineol	1.626 ± 0.129
Verbenone	1.656 ± 0.129
Citronellyl Acet.	1.695 ± 0.149
α-Terpinyl Acet.	1.731 ± 0.054
Carvone	1.742 ± 0.057
Piperitone	1.755 ± 0.083
Citronellol	1.776 ± 0.169
Myrtenol	1.796 ± 0.199
Geranyl Acet.	1.885 ± 0.199
Thymol	2.583 ± 0.151
Sesquiterpenes	
Caryophyllene	1.615 ± 0.049
Elemol	2.45 ± 0.082
γ-Eudesmol	2.662 ± 0.115
α-Eudesmol	2.774 ± 0.132
β-Eudesmol	2.774 ± 0.132
Non-Terpenoids	
n-Decane	0.334 ± 0.047
1-Decene	0.375 ± 0.049
2-Hexenal	0.554 ± 0.043
Estragole	1.598 ± 0.059
n-Decanol	1.774 ± 0.058
Safrole	2.028 ± 0.098
Methyl Eugenol	2.347 ± 0.120
Chavicol	2.518 ± 0.135
Eugenol	2.576 ± 0.135

•As obtained on the 10' × 2 mm. 1% PEG 20M + 1% OV-17 column under GC-MS conditions. Actual retention time of *n*-Heptyl acetate = 10.5 min.

Table II. File Structure of Library

I. Information about the library
   A. number of compounds currently in the library
   B. maximum number of mass intensities allowed (presently 260)
   C. time and date when library was created

II. Internal Standard data
   A. RT of IS (in mins)
   B. number of replicates of IS in library
   C. 5 standard deviation value of RT of IS
   D. lower mass number of the lower range of the IS
   E. upper mass number of the lower range of the IS
   F. lower mass number of the upper range of the IS
   G. upper mass number of the upper range of the IS
   H. average mass intensity values for the IS (41-300)

III. Library directory
   A. For each compound in the library the following record exists
   B. These records are arranged by RRT
      1. RRT, 5 SD of RRT
      2. 4 character i.d. for compound
      3. no. of reps. in library

IV. Data records for compounds arranged as in the directory (by RRT) for each compound
   A. RRT
   B. number of replicates
   C. identification code
   D. sum of RRTs
   E. sum of mass ion values, (41-300)
   F. sum squared mass ion values, (41-300)

File structure of the library of known compounds.

ERRSS = error sum of squares; SUM_{ij} = sum of values for $mass_j$, cpd_i; $SSUM_{ij}$ = sum of squared values for $mass_j$, cpd_i; n = number of cpds in this comparison; and r_i = number of reps of cpd_i.

Then: for $mass_j$

$$CTERM = \left(\sum_{i=1}^{n} SUM_{ij}\right)^2 / \sum_{i=1}^{n} r_i$$

$$TOTSS = \sum_{i=1}^{n} SSUM_{ij} - CTERM$$

$$TRTSS = \left[\sum_{i=1}^{n} (SUM_{ij}^2)\right] / n - CTERM$$

$$ERRSS = TOTSS - TRTSS$$

$$F_j = [TRTSS/(n-1)] \,/\, \left[ERRSS/\left(\sum_{i=1}^{n} r_i - n\right)\right] \qquad \text{(Eq. 1)}$$

These F ratios are then used as character weights in calculating similarities as:

for unknown x and library compound y:

$$D = \sum_{j=1}^{n} (F_j - 1) \frac{|m_{jx} - \bar{m}_{jy}|}{\text{Range } m_j} \qquad \text{(Eq. 2)}$$

Where m_{jx} = value of $mass_j$ in unknown x; $\bar{m}_{jy}$ = average value of $mass_j$ in compound y from the library; F_j = F ratio for $mass_j$ from ANOVA; n = number of valid comparisons; and Range m_j = range of observed values of $mass_j$ for those compounds in the library in the window plus the unknown.

Skip comparison if: $m_{jx} < 2\%$ *and* $m_{jy} < 2\%$

$$S_r = 1.0 - D \Big/ \sum_{j=1}^{n} (F_j - 1) \qquad \text{(Eq. 3)}$$

After these are printed and displayed on the CRT screen, GETID iterates by casting out the compound with the lowest similarity, and recalculating new weights (ANOVA) on the reduced set of compounds from the library. Note that the unknown is not used except in the calculation of the similarities. When only 1 compound is left, no weights can be calculated by ANOVA, and GETID proceeds to calculate an unweighted, absolute similarity between the unknown and the most similar compound of the library as:

$$Sr = \frac{1}{n} \sum_{i=41}^{300} \min(m_i, \bar{m}_i) / \max(m_i, \bar{m}_i) \qquad \text{(Eq. 4)}$$

Where: m_i = intensity of the i'th mass ion of the unknown compound; $\bar{m}_i$ = average intensity of the i'th mass ion of the most similar compound from the library; $\min(m_i, \bar{m}_i)$ = minimum value of m_i or $\bar{m}_i$; $\max(m_i, \bar{m}_i)$ = maximum value of m_i or $\bar{m}_i$; and n = number of comparisons where $m_i > 2\%$ *or* $\bar{m}_i > 2\%$.

Comparisons in which both m_i and $\bar{m}_i$ are less than 2% are skipped to eliminate mismatches involving very small mass intensities which may contain considerable noise.

Experimental

All GC-MS analyses were done with a Finnigan Model 3300 Quadrupole Mass Spectrometer (Finnigan Instruments, Sunnyvale, California) controlled by a Finnigan Incos Data System with 32 K words of core memory, and a 5 megabyte disk. Mass spectral scans were taken repetitively from mass 40 to mass 520 every 2 seconds. A 10 ft × 2 mm i.d. Pyrex glass column containing 1% PEG + 0.5% OV-17 on Chromosorb G (High performance, 80/100 mesh) was used. By adjusting the flow rate (approx. 20 ml/min He) so that the retention time of the IS is within ± 2.5% of the library internal standard (LIS), the elution of library-stored reference compounds will occur within the assigned retention windows.

The column temperature was programmed from 60°C at 4°/min to a final temperature of 190°C. The injector, separator oven, and transfer lines were maintained at 190°C.

The Quadrupole mass filter and E.I. ionizer must be clean in order to produce symmetrical mass peaks necessary for accurate mass assignment (10-30 mmu) and quantitation of mass peaks. With these precautions and careful tuning of the ionizer voltages, we have found it possible to measure accurately differences in mass intensities of compounds that have similar spectra. We use α-pinene, bled through a variable leak, as a calibration compound for setting the ionizer voltages. For our work we arbitrarily adjusted the α-pinene spectra so that mass 41 and mass 136 were approximately 20% and 10% respectively of the base ion, mass 93. It is important that the constraining magnet focus at least 70% of the ion beam on the collector and the ion program voltage is adjusted to the maximum number of ions are pulled from the ionizer throughout the scan.

The internal standard (IS) is present in every analysis and its mass spectrum as well as retention time are compared to those in the library as a calibration check to assure that the tuning of the MS and the flow rate of the GC column are within acceptable limits. Table III shows a typical comparison of the internal standard (IS) versus the library internal standard (LIS). If the fit of the run IS is not close enough to that of the library, one must recalibrate and repeat the run.

Finally, background was removed from all spectra by subtracting appropriate spectra either just before the leading, or just after the trailing edge of the GC peak.

The known compounds entered into the library for comparisons in this study are listed in Table I. These compounds were either commercial terpenes purified by fractional distillation and preparative GC (2) or those available by isolation from conifer leaf oils (3,11,12). *Cis*- and *trans*-ocimene were kindly donated by International Fragrances Limited, Union Beach, New Jersey. The conifer leaf oils were steam-distilled from the appropriate species shortly before use; that of *Abies sibirica* Ledeb. was of commercial origin (Fritzsche Bros., New York, New York). The purity of each compound was checked by analytical GC (2,8,11,12) and identities were reconfirmed by comparison of infrared and proton magnetic resonance spectra. Some of the oils were prefractionated into hydrocarbons and oxygenated compounds on a column of silicic acid deactivated with polyethylene glycol 20M (8,21).

The use of the programs in the identification of the individual component of a conifer oil is demonstrated in the following example with the leaf oil of *Abies sibirica*. After recalibration of the mass spectrometer and equilibration of the gas chromatograph oven at 60°C an aliquot [1.0 μl of the oil, in analytical grade ethyl ether (1:20 v/v)], was injected. The column-to-spectrometer line was vented until the solvent had eluted (50 sec). Temperature programming was started immediately after injection, and the mass-spectra (1 scan every 2 sec) were accumulated continuously until the end of the run (40 min). The reconstructed gas chromatogram, as derived from the total ion count, was copied and, after subtraction of the background, a mass spectrum representive of each peak was placed into a disc file for recall. For each peak the mass spectrum was chosen preferentially at the upslope (where possible) to give a spectrum with 15,000-100,000 total ion counts. Using the program GETID, each spectrum was then compared with those compounds in the reference library (Table II, derived from FIXLIB) that fell within the given retention window. Initially, the relative similarities of the unknown spectrum in comparison with the known ones are listed; if more than two compounds are compared, that with the lowest relative similarity is dropped and a new set of relative similarities is listed. This is repeated until only one compound is left. The absolute similarity of the unknown with the reference compound is then listed (see also Table IV). For 12 previously identified (3) peaks of the leaf oil of *Abies sibirica* the following identities were confirmed (for relative retentions see Table I). Only the initial relative similarities (first iteration) and the absolute similarities are shown below, except for peaks 3 and 6, which demonstrate how the relative similarities change as the compound of least relative similarity is omitted from the set and the weights recalculated.

Table III. Calibration of IS (Output from GETID)

SAMPLE: 161. SCANS 311 to 311. DATE: 11/16/77. TIME 10. J. ASHEI

	INTR, SUM(N)	LIB, SUM(N)	AVG DIFF	CV DIFF	% LOW/ HIGH
M/E 41-63	288.88(18)	295.21(20)	-0.31	-695.46	-2.14% LOW
M/E 91-116	33.03(19)	27.99(23)	0.20	284.97	17.98% HIGH

Similarity of INTR To Library = 0.8361;

RT = 10.59 Compared to 10.69 of INTR of Library;

95% Range of LIB RT = 10.49 - 10.90.

Based on these values, do you want to continue (CO), or exit (EX) so you can recalibrate the MS?

Computer output from program GETID showing a comparison of the internal standard (heptanyl acetate) with the average values for this internal standard as stored in the library. INTR = current values of the IS; LIB = average values in the library for the IS; AVG DIFF = the algebraic average of the differences between IS and library IS; CV DIFF = the coefficient of variation of the differences; % LOW/HIGH = the average difference (INTR-LIB) divided by the sum of the LIB over a particular set of mass ions. If the INTR is smaller than the LIB, the IS is judged to be LOW (skewed) and conversely if the INTR is larger than the LIB the IS is called HIGH (also skewed). Similarity is calculated by formula 4.

Table IV. Output of Search from GETID

SIMILARITY OF S92 WITH RRT = 0.27 To:

APNN = 0.7593 RRT = 0.29;

TRCY = 0.0703 RRT = 0.26;

SIMILARITY OF S92 WITH RRT = 0.27 TO:

APNN = 0.8128 RRT = 0.29.

Computer output of search comparing an unknown (α-pinene from *Juniperus ashei* Buch.) to α-pinene (APNN) and tricyclene (TRCY) of the library. Mass ion weighting factors (F-1) are shown in Table V. The final similarity was computed as the absolute unweighted similarity (formula 4). S92 is a code for the unknown compound, entered interactively by the operator and can be any combination of 4 letters and/or numbers (we used the scan number in this case).

Peak 1 (RRT 0.22) = Santene; relative similarity RS to (a) santene, 0.9593; (b) tricyclene, 0.0179; absolute similarity, AS 0.6984.

Peak 2 (RRT 0.26) = Tricyclene; RS (a) tricyclene, 0.6044; (b) α-pinene, 0.3520; AS 0.7643.

Peak 3 (RRT 0.30) = α-pinene; RS (a) α-pinene, 0.9895; (b) tricyclene, 0.9649; (c) *n*-decane, 0.0071. *n*-Decane omitted from the set, weights recalculated using the remaining 2 compounds. RS (a) α-pinene, 0.6645; (b) tricyclene, 0.0063; AS 0.7296.

Peak 4 (RRT 0.36) = Camphene; RS (a) camphene, 0.9564; (b) 1-decene, 0.1812; (c) *n*-decane, 0.1697; AS 0.7769.

Peak 5 (RRT 0.44) = β-pinene; RS (a) β-pinene, 0.6111; (b) sabinene, 0.3075; AS 0.7627.

Peak 6 (RRT 0.52) = Myrcene; RS (a) myrcene, 0.9826; (b) Car-3-ene, 0.931; (c) α-phellandrene, 0.9657; (d) hexanal, 0.0296. Compound hexanol omitted from the set, weights recalculated using the remaining three compounds. RS (a) myrcene, 0.7257; (b) Car-3-ene, 0.3840; (c) α-phellandrene, 0.2097; AS 0.7064.

Peak 7 (RRT 0.61) = Limonene; RS (a) limonene, 0.9633; (b) β-phellandrene, 0.2115; α-terpinene, 0.0586, AS 0.7891.

Peak 8 (RRT 0.63) = β-Phellandrene; RS (a) β-phellandrene, 0.9386; (b) limonene, 0.8144; (c) 1:8 cineole, 0.2021; AS 0.6802.

Peak 9 (RRT 0.72) = *p*-Cymene; RS (a) *p*-cymene, 0.9374; (b) γ-terpinene, 0.0870; (c) *trans*-ocimene, 0.0354; AS 0.9333.

Peak 10 (RRT 1.24) = Fenchyl acetate; RS (a) fenchyl acetate, 0.9124; (b) camphor, 0.4830; (c) Douglas-fir, unknown II 0.2123; AS 0.7006.

Peak 11 (RRT 1.28) = Camphor; RS (a) camphor, 0.9302; (b) Douglas-fir, unknown II 0.7311; (c) fenchyl acetate, 0.7128; (d) linalool, 0.2535; AS 0.6248.

Peak 12 (RRT 1.49) = Bornyl acetate; RS (a) bornyl acetate, 0.9645; (b) terpinen-4-ol, 0.5926; (c) methyl thymol, 0.2591; AS 0.6642.

Results and Discussion

Although most of the search and matching techniques in use today use only the largest ions, we have found that discrimination between the terpenoids shows that there is no *a priori* reason the assume that the more intense ions are necessarily more useful. Table V shows the F-1 weights generated from ANOVA of tricyclene and α-pinene. Only 15 ions had F ratios larger than 1.0, indicating that they could be used to discriminate between these two compounds. The largest weight was for the parent ion (mass 136; 8-20% of base peak) in this example, although we have run many compounds where the parent ion did not discriminate. The next highest weights were for mass ions 121, 57, 68, 70, and 66. Several of these ions were present in quantities less than 5%. Two of the most intense ions (90 and 91) have only moderate weights. Of course the base ion is missing because its value is 100% in both compounds. Even though the base ion was eliminated in the comparison involving these particular compounds, it would certainly come into play when the retention time window includes compounds with different base ions.

Visual examination of the mass spectra of the "unknown" tricyclene and α-pinene (Figure 1) shows that these two compounds have similar spectra. Relative retention data also does not help much to differentiate these two terpenes (RRT = .26 and .29). Comparison of the mass spectra of the "unknown" α-pinene with the compounds listed in the appropriate retention time window by the GETID program (last iteration in searching the library on a relative basis; Table IV) shows relative, weighted similarities much closer to α-pinene (0.7593) than to tricyclene (0.0703). This merely indicates that it may be α-pinene. More particularly, it appears that the most similar compound in the library, α-pinene, is very closely related (structurally) to the unknown. Absolute similarity of the "unknown" α-pinene to the library α-pinene is 0.8128 (Table IV), which may be considered to be a good match of similarities (see below).

In order to determine which level of similarities are significant for matches and mismatches, statistics were compiled for 47 matches versus 28 mismatches (Table VI). In the cases of correct identification (matches), the most similar weighted similarity (last iteration) averaged 0.8482 compared to 0.0820 for the second most similar compound. This is considerably different from the case of the mismatches (unknown not in the library) which had a most similar average of 0.6192 and a second most similar average of 0.2912. Although the weighted similarity might lead one to believe the unknown has been identified, the primary purpose of the weighted similarities is to find the most similar compound in the library. It remains the task of the absolute similarity calculation to determine the confidence of identification. Notice that absolute similarities show considerably larger values (Table VI) with matches than mismatches. We have generally found that an absolute similarity of less than 0.5 is too small to indicate proper identification.

We have done some experiments with various threshold levels for skipping character matches in formula 4 for the absolute similarity and presently use a cutoff of 2% (ion intensity less than 2% in *both* unknown and compounds from the library). Additional research is needed to maximize the absolute similarity for matches and minimize it for mismatches.

Perhaps the largest single source of error we have encountered to date has been in obtaining pure compounds for the library and getting a pure spectrum of the unknown.

Several areas of improvement need to be made. One improvement would be to incorporate several internal standards as suggested by Smith, et al. (14). However, we do not have a significant problem when dealing with volatile oils, which contain hundreds of compounds, in finding standards that do not run on top of any one of these compounds. Comparisons of elution of single ion scans (e.g., mass 136, 121, 91, etc.) for suspect chromatographic peaks promise to be an aid in the determination of purity of a compound. Obtaining pure spectra may be aided by further computer processing of the raw data file (14) and conversion to high resolution capillary columns. Future expansion of the system will probably include storage of some prose regarding the structure of a compound, its common name, and its chemical name.

Table V. F-1 Weights for Tricyclene and α-Pinene

Mass	F-1 wt.	avg. ion value (% of base ion) Tricyclene	α-Pinene
43	0.3	7.1	4.3
51	1.1	3.4	4.3
53	1.9	4.3	6.5
57	3.4	2.6	0.4
65	0.4	5.2	6.0
66	6.7	5.2	2.4
68	8.4	1.6	3.8
70	7.9	2.5	0.4
77	4.6	19.7	29.1
78	5.0	8.9	6.1
91	5.5	80.7	43.9
92	5.9	24.0	36.2
94	0.5	12.2	11.0
121	9.0	21.5	14.6
136	13.4	15.4	7.7

A comparison of character weights (F-1) generated in ANOVA of α-pinene and tricyclene along with average mass intensities for these two compounds. Only those mass values that had an F greater than 1.0 (F-ℙ0.0) are shown since all other masses are not effective in discriminating between these two compounds.

Table VI. Summary of Statistics Relating to Matched versus Mismatched Compounds and Relative versus Absolute Similarities. sd = Standard Deviation of the Average

	Matches (compd. found) Avg. Sr. (# obs.) ± sd	Mismatches (compd. not found) Avg. Sr (# obs.) ± sd
Last iteration weighted Sr		
most similar compd. in lib.	0.8482(47) ± 0.1638	0.6192(28) ± 0.1806
2nd most sim. compd. in lib.	0.0820(47) ± 0.0890	0.2912(28) ± 0.1843
Absolute similarity	0.7179(47) ± 0.0878	0.2634(28) ± 0.1303

In summary, we conclude that although the unequivocal identification of the terpenoids by MS was initially thought to be an unattractive technique, our experiments have shown that even spectra that may appear to be visually identical, can be distinguished by the weighted similarities method. The present library of known terpenes and non-terpenoid conifer leaf oil components has about 70 compounds and is structured as shown in Table II. Application of the GETID program to GC-MS analysis of the leaf oils of *Abies sibirica* (3), *A. amabilis* (22), *A. lasiocarpa* (23), *Pseudotsuga menziesii* (2. varieties) (24), and *Juniperus ashei* (25), *J. virginiana* (26) and *J. pinchotii* (27) resulted in the correct identification of all the terpenes and phenyl propanoid ethers identified previously. An example of the mode of operation and results obtained with the commercial leaf oil of *Abies sibirica* is given in the experimental section.

References

1. B.M. Lawrence. *Can. Inst. Food Technol. J.* 4: A44-A48 (1971).
2. E. von Rudloff. Gas-liquid Chromatography of Terpenes, in *Advances Chromatography*. J.C. Giddings and R.A. Keller eds. Vol. 10, pp. 173-226 (1974).
3. E. von Rudloff. *Biochem. Syst. and Ecol.* 2: 131-67 (1975).
4. E. Zavarin and K. Snajberk. *Phytochem.* 11: 1407-21 (1972).
5. R.P. Adams. *J. Mol. Evol.* 5: 177-85 (1975).
6. E.von Rudloff. *Can. J. Bot.* 50: 1595-1603 (1972) and 53: 2978-82 (1975).
7. R.A. Powell and R.P. Adams. *Am. J. Bot.* 60: 1041-50 (1973).
8. E. von Rudloff. Scope and Limitations of Gas Chromatography of Terpenes in Chemosystematic Studies, in *Recent Advances in Phytochemistry*. M.K. Seikel and V.C. Runeckles, eds. Appleton-Century-Crofts, Vol. 2, pp. 127-62 (1969).
9. D. Welti. *Infrared Vapour Spectra*. Heyden and Sons Ltd. (1970).
10. E. von Sydow, K. Anjau, and G. Karlsson. *Archives of Mass Spectral Data SIK-Rapport Nr279* (Goteborg, Sweden). Vol. 1, 1970.
11. E. von Rudloff. *Can. J. Bot.* 54: 1926-31 (1976).
12. E. von Rudloff. *Phytochem.* 17: 127-30 (1976).
13. B.E. Blaisdell. *Anal. Chem.* 49: 180-86 (1977).
14. D.H. Smith, M.A. Achenback, W.J. Yeager, P.J. Anderson, W.L. Fitch, and T.C. Rindfleisch. *Anal. Chem.* 49: 1623-32 (1977).
15. R.R. Sokal and P.N.A. Sneath. *Principles of Numerical Taxonomy*. W.H. Freeman and Co., San Francisco, (1963).
16. R.R. Sokal and P.N.A. Sneath. *Numerical Taxonomy*. W.H. Freeman and Co., San Francisco, (1973).
17. R.R. Sokal and F.J. Rohlf. *Biometry, the Principles and Practice of Statistics in Biological Research*. W.H. Freeman and Co., San Francisco, (1969).
18. D.C. Eades, *Syst. Zool.* 14: 98-100 (1965).
19. R.P. Adams. *Brittonia* 27: 305-16 (1975).
20. C.D. Hodgeman. *C.R.C. Standard Mathematical Tables*. 11th Ed. Chem. Rubber Publ. Co., Cleveland, Ohio, (1957).
21. E. Kugler and E. Kovàts. *Helv. Chim. Acta.* 46: 1480-1513 (1963).
22. E. von Rudloff and R.S. Hunt. *Can J. Bot.* 55: 3087-92 (1977).
23. R.S. Hunt and E. von Rudloff. *Can. J. Bot.* 52: 477-87 (1974).
24. E. von Rudloff. *Can. J. Bot.* 50: 1025-40 (1972).
25. E. von Rudloff. *Can. J. Chem.* 46: 679-83 (1978).
26. A.R. Vinutha and E. von Rudloff. *Can. J. Chem.* 46: 3743-50 (1968).
27. R.P. Adams. *Biochem. Syst. and Ecol.* 3: 71-74 (1975).