Use of boiling point–Lee retention index correlation for rapid review of gas chromatography-mass spectrometry data

William P. Eckel [a,*], Tobias Kind [b]

[a] *US Environmental Protection Agency, Office of Pesticide Programs, Environmental Fate and Effects Division (7507C),
1921 Jefferson Davis Highway, Arlington, VA 22202, USA*
[b] *Department of Chemical Ecotoxicology, UFZ Centre for Environmental Research Leipzig-Halle,
Permoserstraße 15, D-04318 Leipzig, Germany*

Abstract

Gas chromatographic (GC) retention indices are useful for reviewing mass spectral library searches for the tentative identification of organic compounds. If the known GC retention index (RI) of a compound that ranks high on the (gas chromatography-mass spectrometry) GC-MS library search "hit list" matches the RI of the unknown, there is a strong presumption that a correct identification has been made. If the Lee retention index of the "hit list" compound is not known from published data, and cannot be estimated by RI–boiling point correlation, its boiling point (°C) may be directly compared to the unknown's Lee retention index. In general, an unknown compound with a particular RI will have a boiling point of $RI - 10$ (°C) to $RI + 50$ (°C). Some compound classes deviate predictably from this rule. Library matches for unknowns whose boiling points fall outside the $RI - 10$ to $RI + 50$ range may safely be rejected. Conversely, library matches whose boiling points fall within the range of $RI - 10$ to $RI + 50$ should be further considered as possibly correct identifications.
© 2003 Elsevier B.V. All rights reserved.

Keywords: Lee retention index; Boiling points; Correlation; Gas chromatography; Mass spectral library

1. Introduction

Gas chromatography-mass spectrometry (GC-MS) is a widely used method for identification of organic compounds in environmental studies. US Environmental Protection Agency (EPA) methods for GC-MS analysis [1,2] have target lists of compounds, for which the instrument is specifically calibrated (e.g. the priority pollutants). Other, non-target compounds are tentatively identified by comparing their mass spectra to large, computerized mass spectral libraries [1]. Usually, three library search "best matches" are returned to the analyst, who uses them to either assign an identity to the unknown (making it a "tentatively identified compound"), or to decide that no reasonable identification can be made.

It is well recognized that the reliability of non-target compound identification, in the process of GC-MS data review, could be improved [3]. One technique that has been successfully applied [4–8] to the review and confirmation of tentatively identified compounds is the GC retention index (RI). If both the mass spectrum and RI of the unknown are well-matched to known data, a reliable identification has been made. This technique is limited only by the availability of RI data for the compounds in the MS

* Corresponding author. Fax: +1-703-308-6181.
E-mail addresses: eckel.william@epamail.epa.gov (W.P. Eckel),
publish@amdis.net (T. Kind).

libraries. To date, no large RI database has been compiled.

Because of the great effort required to measure the RI of a large number of compounds, a variety of methods have been proposed [9] to estimate or predict the RI, either directly from physical properties or from quantitative structure–property relationship (QSPR) models. Complex equations with up to 20 molecular descriptors (including physicochemical, geometrical, and electronic parameters) are used for such QSPR and QSRR (quantitative structure–retention relationship) calculations [21,22]. Even if applicable to certain molecule classes, it is safe to say that no universal system for predicted RIs has yet been developed. Therefore, such methods require specialized knowledge and a large calculation effort. This is impractical for a fast review of GC-MS data. One of the physical properties commonly used [5] to predict the RI for non-polar stationary phases is the normal boiling point (nbp), since the two correlate very well for chemical classes that are sufficiently narrowly defined. This correlation is in accord with gas chromatographic theory. A compound's RI may be predicted from its nbp (the boiling point at 760 mmHg pressure) with a small error if the regression line for its particular chemical class has been established.

In most cases, the relationship between RI and nbp for a chemical class of interest is not known, and therefore the RI of a particular member of that class cannot be estimated from a prediction equation and its nbp. The purpose of this paper is to establish the range of reasonable boiling point values for an unknown compound with a particular Lee-RI, so that library matches with nbp outside the range may safely be rejected. This should enable the data reviewer to determine if a particular tentative compound identification is reasonable (assuming a good spectral match), by comparing the boiling point of that compound directly to the experimental RI of the unknown. Because the Lee retention index [10] is used, the magnitudes of the RI and the boiling point ($^\circ$C) are very similar. Lee-RI data are available for hundreds of compounds for the commonly used non-polar DB-5 GC column or equivalent (HP-5 or Rtx-5 with 95% methyl, 5% phenyl siloxane) under linear-ramp temperature programming conditions. The Lee retention index is based on the PAH standard compounds: naphthalene (RI = 200), phenanthrene (RI = 300), chrysene (RI = 400) and

picene or benzo(*g,h,i*)perylene (RI = 500) [10]. Other RI values are calculated with the following formula.

$$\mathrm{RI} = 100 \left(z + \frac{t_{\mathrm{R}x} - t_{\mathrm{R}z}}{t_{\mathrm{R}z+1} - t_{\mathrm{R}z}} \right) \qquad (1)$$

where $t_{\mathrm{R}x}$ is the retention time of the substance of interest and $t_{\mathrm{R}z}$ (before $t_{\mathrm{R}x}$) and $t_{\mathrm{R}z+1}$ (after $t_{\mathrm{R}x}$) are the retention times of the standards, which bracket the substance of interest. The factor z contains the number of rings in the PAH standard that elutes prior the substance of interest. Retention indices outside the interval of the standard compounds are linear extrapolated from the nearest interval.

2. Experimental

Lee retention index data were gathered from references [5,7,10–14]. Boiling point data were gathered from [15–18,20]. Boiling points of 56 compounds were corrected to atmospheric pressure using a utility in the NIST Structures and Properties Database [19]. These data were entered into a Microsoft Excel database for analysis. Three hundred and seventy data pairs, representing over 20 chemical classes, were entered. Table 1 summarizes the number of compounds from each chemical class that were used. For the calculation of boiling points we recommend the use of the freely available EPA-EPI-Suite [23] with the PHYSPROP database (containing 6381 experimental boiling points) and the MPBPWIN program using the Stein and Brown [24] method. MPBPWIN can estimate unknown boiling points directly from the molecular structure. The structure is very easily obtained via entering the Chemical Abstracts Service Registry (CAS) number of the substance or alternatively the Simplified Molecular Input Line Entry Specification (SMILES) structure code. The method used in MPBPWIN was validated on a dataset of 6584 compounds (not used in deriving the method) and gave the following statistical accuracy: average absolute error = 20.4 K; standard deviation = 38.1 K; average error = 4.3% [24]. Further studies have shown that certain compound classes like organometallics and fluoro-organic compounds have a larger standard error of estimation and should be avoided. Retention indices easily can be obtained using the freely available Automated Mass Spectral Deconvolution

Table 1
Compound-class specific regressions between Lee-RI and boiling point

Compound class	Standard error RI vs. nbp (exp)	Standard error RI vs. nbp (calc)	R^2 (exp nbp)	R^2 (calc nbp)	Slope	Intercept	Number of points	Reference
Cyclic alkenes	–	–	–	–	–	–	5	[5]
Primary alkenes	–	–	–	–	–	–	3	[5]
PASH	–	–	–	–	–	–	4	[7,10–12]
Alkyl benzenes	2.67	4.26	0.9977	0.9997	1.13	1.19	6	[5]
Aromatic acid esters	2.94	2.82	0.9933	0.9958	0.93	29.80	8	[12]
Nitriles	3.12	2.09	0.9971	0.9986	0.94	36.13	9	[12]
Dicarboxylic acid esters	4.10	0.84	0.9913	0.9998	0.87	51.96	8	[12]
Fatty acid esters	5.31	4.69	0.9945	0.9959	0.94	15.81	17	[12]
Phenols	5.46	13.87	0.9707	0.8257	0.97	24.13	26	[5,12,14]
Fatty acids	6.01	2.94	0.9923	0.9982	0.90	60.77	14	[5]
Oxy-PAHs	6.10	11.27	0.9954	0.9756	1.21	−24.37	6	[10–12]
Alcohols	6.17	6.82	0.9923	0.9916	0.95	24.05	15	[12]
Ketones	7.34	6.75	0.9923	0.9935	1.00	10.85	10	[12]
Aromatic amines	8.54	6.99	0.9600	0.9765	1.02	20.08	16	[12]
Pyridines, quinolines, nitro-compounds	9.61	12.34	0.9828	0.9666	1.13	−6.28	57	[10–13]
Aldehydes	10.32	7.37	0.9684	0.9834	0.93	29.84	7	[5]
PAHs	11.16	16.80	0.9828	0.9506	1.05	6.86	70	[10–12]
Alkanes	12.76	6.81	0.9870	0.9967	0.87	43.04	29	[12]
Priority pollutants-CLP	14.26	16.54	0.9606	0.9416	1.03	18.16	30	[5]
EPA Method 1625C and 525	16.69	18.32	0.9063	0.8705	0.98	25.03	30	[2]
All compounds/exp boiling points	14.80	–	0.9674	–	0.98	24.36	370	
All compounds/calc boiling points	–	16.84	–	0.9552	0.94	31.12	370	

Linear regression parameters for every compound class under investigation. Standard errors are for linear regressions using experimental boiling points and for calculated boiling points (EPI-Suite). Slope and intercept are obtained from experimental values and can be used for more accurate estimations like: nbp = (Lee-RI) × slope + intercept; exp, experimental, calc, calculated.

and Identification System (AMDIS) program [25]. The RI-values are calculated from internal or external standards with special calibration scans.

3. Results and discussion

Fig. 1 is a plot of Lee retention index versus nbp (°C) for the 370 compounds. The regression for these data yield an $R^2 = 0.967$ with a standard error of 14.80 and $P < 0.0001$, which is highly significant. The lower and upper line on the plot show a 95% prediction band. Only a small number of compounds lie outside these bands. The series of points that fall below the lower 95% prediction band (above RI = 480) are the normal alkanes above hentriacontane.

The frequency distribution for this series (on the basis of the difference between the boiling point and the RI, expressed as temperature in degree centigrade) is shown in Fig. 2. It should be mentioned that the dataset used for prediction is not uniformly distributed. But the regression data show a high significance. Clearly, most of the data points (95.1%) fall between the lines (nbp = RI − 10) and (nbp = RI + 50). This means that the boiling point of a compound with a given RI will likely fall somewhere between RI − 10 and RI + 50 (°C), regardless of its compound class. The major exception to this rule is the *n*-alkanes. Eleven normal alkanes above *n*-hexadecane (RI ∼ 400) fall below the linear regression line, meaning that nbp < RI. Other aliphatic compound classes whose nbp fall below their RI are *n*-aldehydes (C − 10 and above) and fatty acid methyl
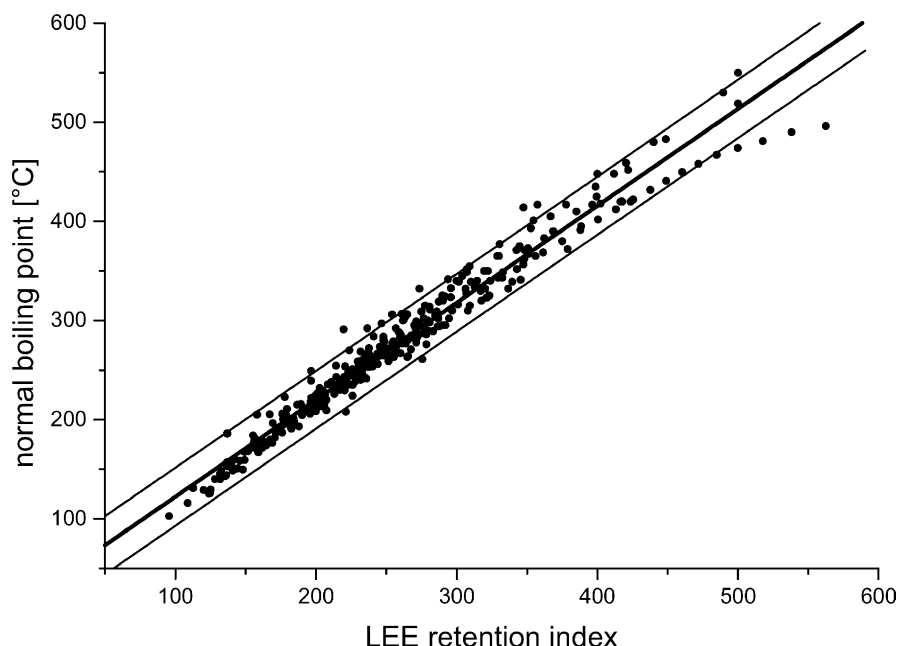
Fig. 1. Relationship between boiling points and Lee retention index. The Linear regression follows the formula nbp $= 0.98 \times$ RI $+ 24.36$. Only 19 out of 370 compounds (5%) fall outside the 95% prediction bands.

esters (C $-$ 11 and above) as seen in Table 2. Of the 12 compounds in Table 2, only 3 fall below our proposed boiling point range lower limit of RI $-$ 10. For two of these compounds (decanal and methyl tride-

canoate) this may be due to boiling point correction errors.

Table 3 is a list of seven compounds whose retention indices are extremely low compared to the nbp

Table 2
Compounds with retention index greater than normal boiling point

Compound	bp data (°C)	p (mmHg)	bp est.[c]	RI	RI–bp[d]
Decanal[a]	208	760	–	221.34	13.34
Undecanal[a]	117	18	225	239.08	14.08
Tetradecanal[a]	166	24	276	278.43	2.43
Methyl decanoate[b]	224	760	–	226.04	2.04
Methyl tridecanoate[b]	92	1	261	275.62	14.62
Methyl heptadecanoate[b]	185	9	332	336.74	4.74
Methyl eicosanoate[b]	215	10	372	378.87	6.87
1-Octadecanol	210.5	15	339	345.48	6.48
	203	10	341	345.48	4.48
2,3,6-Trimethylnaphthalene	263	760	–	264.99	1.99
Ortho-quaterphenyl	420	760	–	423.63	3.63
Methapyrilene	173	3	339	339.60	0.6
2,3,4,6-Tetrabromophenol	150	15	264	265.58	1.58

Boiling points (bp) were obtained at pressures (P); est. estimated.
[a] Aldehydes above C9 appear generally to have RI < nbp.
[b] Fatty acid esters appear generally to have RI values less than 10 points below their nbp.
[c] Boiling point corrected to 760 mmHg pressure by NIST Structures and Properties Database program utility.
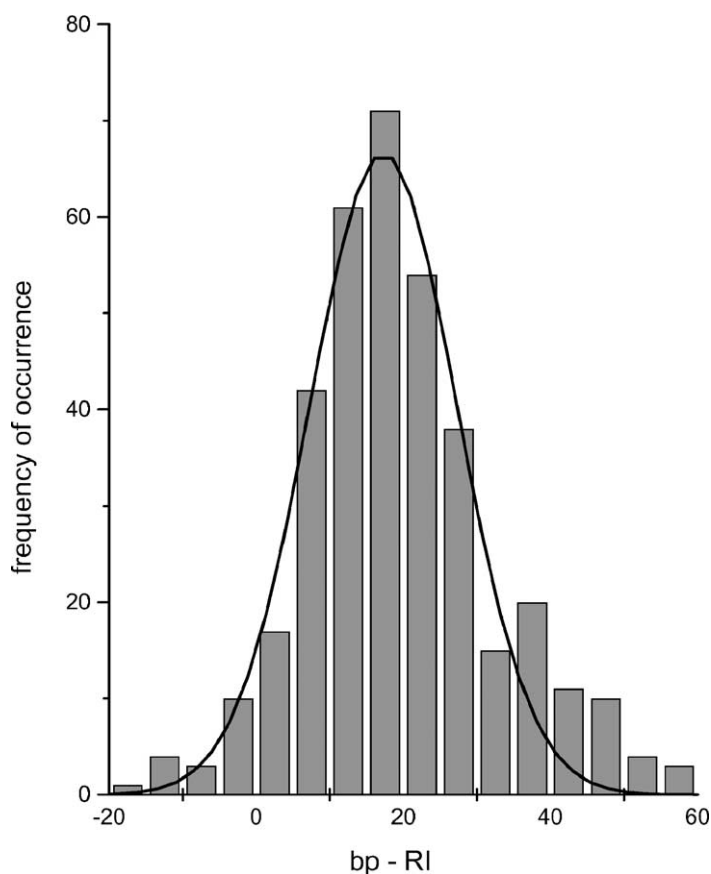[d] RI minus bp at 760 mmHg or bp corrected to 760 mmHg.

Fig. 2. A frequency distribution of the current data set shows that compounds with boiling points outside $(RI - 10)$ and $(RI + 50)$ may be safely rejected during GC-MS data evaluation.

Table 3
Compounds with retention index 50 points lower than normal boiling point

Compound	bp data (°C)	p (mmHg)	bp est.[a]	RI	RI–bp[b]
Benzoic acid	249.2	760	–	196.52	−52.68
2-Phenylindole	250	10	414	347.47	−66.53
1,3-Benzenediol	178	16	291	219.64	−71.36
	Range 277–281	–	–	–	−57.36
	–	–	–	–	−61.36
9-Nitroanthracene	275	17	417	357.42	−59.58
	>360	–	–	–	–
1,4-Dinitrobenzene	297	760	–	246.75	−50.25
4-Nitroaniline	332	760	–	273.55	–58.45
3-Nitroaniline	306 decomposed	760	–	254.17	−51.83

Boiling points (bp) were obtained at pressures (p); est., estimated.

[a] Boiling point corrected to 760 mmHg pressure by NIST Structures and Properties Database program utility.

[b] RI minus bp at 760 mmHg or bp corrected to 760 mmHg.

(more than 50 points lower). This may be due to polar groups (–OH, –NO2, –COOH) causing intermolecular forces (e.g. hydrogen bonding) in the condensed phase. Certain outliers may also be due to boiling point estimation errors.

It is important to point out, that for congeneric molecules (like PCBs or PCNs) and substituted chemicals with the same molecular mass (like different methyl-anthracenes) this approach will only identify an unknown as one of a set of isomers. Important parameters for the retention behavior (like molecular size and shape) are not considered in this approach. But for complex chromatograms with up to 200 peaks, it is very useful to have a fast and reliable true/false criterion. The following example demonstrates this method in a comprehensible way.

To demonstrate the suggested method, a sample chromatogram (see Fig. 3) (HP5-MS column; length 30 m, 0.25 mm i.d., 0.25 μm film thickness; temperature program: initial value 70 °C; initial time: 4.00 min; heat rate 7 °C/min to 280 °C/min and a hold time of 20 min at 280 °C; full-scan mode 50−300 m/z) containing more than 40 compounds was taken. The PAH-class compounds were used for calibration of the Lee retention indices, whereas most other substances were not amongst the compounds of the prediction set used for this RI–bp method.

A subsequent search for a peak with a characteristic m/z of 227 with the NIST98 mass spectral library gave a hit list, part of which is given in Table 4, case 1. The CAS-numbers were taken from the hit list and boiling points were calculated with MPBPWIN from the EPA-EPI-Suite. Three methoxychlor-related compounds and two S-hydrindacene compounds where obtained from the hit list. The compound itself had a Lee-RI of 403. Strictly following the rule, all five compounds should be rejected, but as they all contain more than one polar group, certain deviations in both nbp estimation and nbp/RI correlation may occur that make it necessary to relax the rule somewhat (as seen in Table 3). Because the two S-hydrindacene compounds show a deviation of nearly 100 RI points they should be discarded and only the methoxylchor derivatives should be taken into further consideration.

In a second case a compound with characteristic chlorine patterns in the mass spectra was found, amongst them m/z values of 214, 249, and 286. A search in the NIST98 mass spectral library gave the hit list in Table 4, case 2. The calculated Lee-RI of the compound under investigation was 452. So following the rule of thumb, compounds with nbps ranging between RI − 10 and RI + 50 (442–502 °C) may be safely retained. It is strongly suggested by this analysis that decachlorobiphenyl is the correct compound. But why are characteristic peaks with m/z at 356, 428, and 489 absent in the mass spectrum? This GC-MS run was taken at 50–300 m/z to increase the sensitivity of the full scan method. So this was leading to misinterpretations, because certain peaks in the mass spectrum, including the molecular ion, are absent. But with the help of the suggested rule, this unexpected error can be avoided. In the end, both compounds could be confirmed, the first as methoxychlor and the second as decachlorobiphenyl.

Table 4
Compounds found during mass spectral evaluation. Normal boiling points were obtained from the EPI-Suite

| Case number	Lee-RI of unknown	Name	CAS	nbp (°C)
1	403	Methoxychlor	72-43-5	377
		2,2-Dichloro-1,1-bis(4-methoxyphenyl)ethane	7388-31-0	376
		2,4′-Methoxychlor	30667-99-3	377
		1,1,4,5,5,8-Hexamethyl-S-hydrindacene	17465-59-7	314
		1,2,3,5,6,7-Hexahydro-1,1,4,7,7,8-hexamethyl-S-Indacene	17465-58-6	315
2	452	Decachlorobiphenyl	2051-24-3	471
		1,2-Dichloro-3,4-bis(dichloromethylene-cyclobutane	55044-46-7	276
		Hexachloro-benzene	118-74-1	291
		1,2,3,4-Tetrachloro-5-(dichloromethylene)-1,3-cyclopentadiene	6317-25-5	267
3	310	Chlorphenamine	132-22-9	357
		4-Chlorodiphenylamine	1205-71-6	335
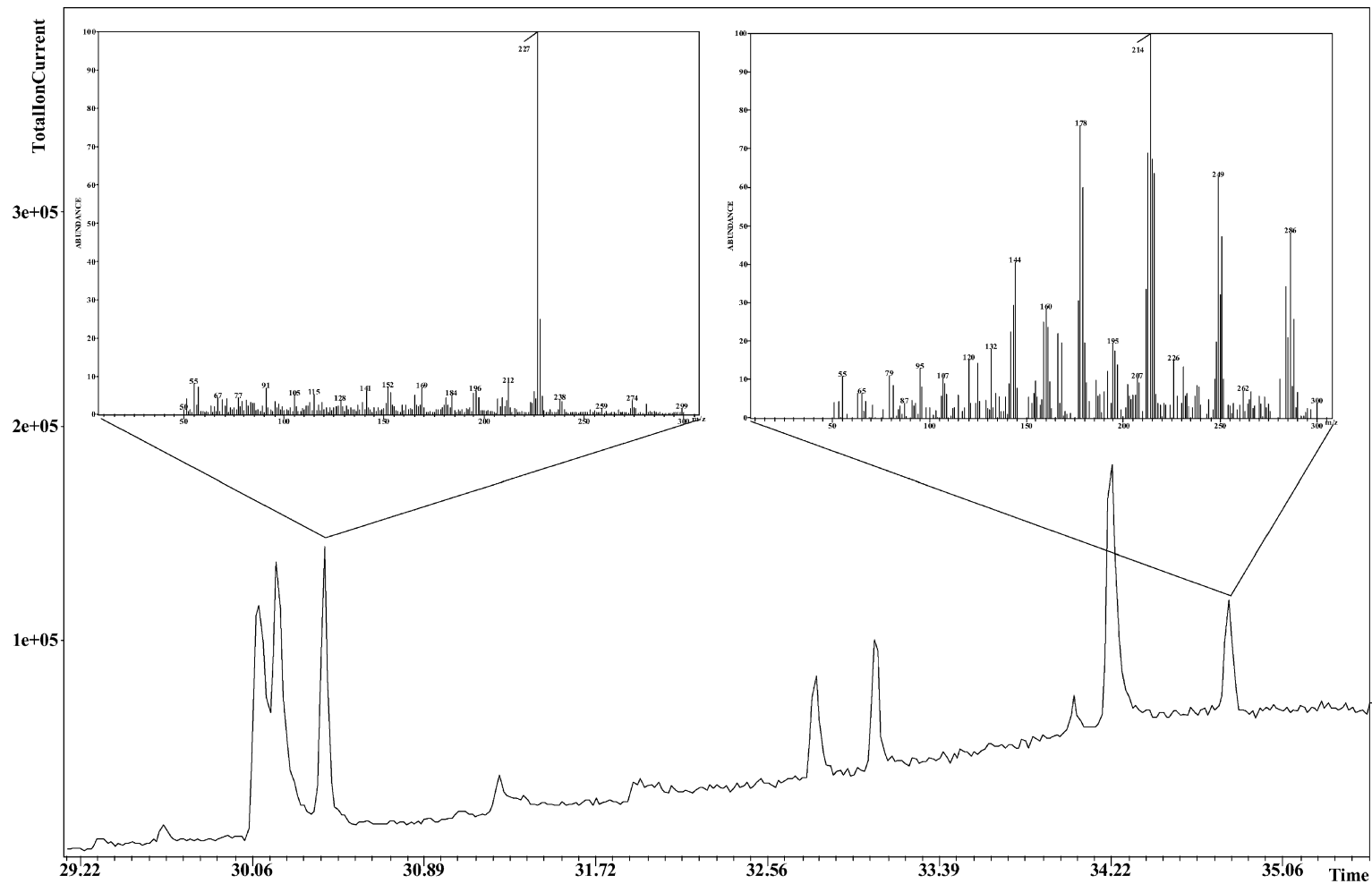
Fig. 3. Chromatogram with two selected compounds (methoxychlor, decachlorobiphenyl) and related mass spectra.

The third example (Table 4, case 3) discusses the evaluation of a river sediment extract, where a substance with a Lee retention index of 310 was found. A mass spectral library search resulted in two similar spectra, both with high matching scores. One compound was chlorphenamine and the second compound was 4–chlorodiphenylamine. For 4-chlorodiphenylamine an experimental nbp of 335 °C was obtained from the PHYSPROP database. According to the regression formula for aromatic amines from Table 1 and the formula nbp = (Lee-RI) × slope + intercept, it was easily calculated: nbp = $1.02 \times 310 + 20.08 = 336$ °C. That means 4-chlorodiphenylamine (or isomers) could be accepted and chlorphenamine can be rejected.

It should be mentioned again, that this method combining Lee retention indices and boiling points cannot be used to differentiate among isomers or structurally very similar compounds. Also compounds with retention indices outside the RI-calibration range may have higher errors. The MPBPWIN program itself produces errors during the prediction of boiling points. Certain highly polar compounds with may show a strong deviation from this rule (see first example in Table 4).

4. Conclusions

It has been demonstrated that 351 of 370 compounds (95%) in over 20 chemical classes have nbps (boiling point at 760 mmHg pressure) lying between their Lee retention index minus ten and their Lee retention index plus fifty, expressed as temperature (°C). Many of the exceptions are for compounds whose boiling points were extrapolated to atmospheric pressure. *N*-Alkanes, *n*-aldehydes and nitro compounds are compound classes that do not strictly obey the rule.

When reviewing tentatively identified compound data, library matches whose boiling points fall outside the range of RI − 10 to RI + 50 may be safely rejected, with the exception of *n*-alkanes with RI > 480. Compounds with nbp only slightly less than the RI (difference less than 10 points) may also be retained as possibly correct, particularly for aliphatic compound classes. A better accuracy can be obtained using the linear equations obtained from regression calculations in Table 1. These parameters are highly specific for every single compound class. Library spectra with high spectral matching scores and boiling points falling within the RI − 10 to RI + 50 range should be further considered as possible correct identifications of the unknown compound.

Great care must be taken in estimation of the boiling point at atmospheric pressure (nbp) from reduced pressure data, and in the selection of nbp data from secondary sources. More than one secondary source should be consulted to avoid errors in transcription of nbp data from the primary source. Plots of nbp versus carbon number or RI can reveal errors in nbp data, for homologous series. Using calculated boiling points from the EPI-Suite, additional deviations outside the range RI − 10 to RI + 50 (°C) may occur due to inaccuracy in boiling point estimation.

This nbp–RI range method may be of great help during the preliminary evaluation and identification of a large number of diverse compounds.

Acknowledgements

References

[1] US Environmental Protection Agency, Contract Laboratory Program,
Statement of Work for Organics Analysis Multi-Media Multi-Concentration, Document Number OLM01.0, including revisions to OLM01.9, July 1993.

[2] US Environmental Protection Agency, Office of Water Regulations and Standards, Method 1625: Semivolatile Organic Compounds by Isotope Dilution GCMS, June 1989.

[3] K.C. Swallow, N.S. Shifrin, P. J. Doherty, Environ Sci. Technol. 22 (1988) 2.

[4] W.P. Eckel, B. Ross, R.K. Isensee, Ground Water 31 (1993) 801.

[5] J.R. Donnelly, M.S. Abdel-Hamid, J.L.JeterD.F. Gurka, J. Chromatogr. 642 (1993) 409.

[6] C. Bichi, A.D. Amato, M. Orlandin, J. High Resolut. Chromatogr. 17 (1994) 335.

[7] C. Willey, M. Iwao, R.N. Castle, M.L. Lee, Anal. Chem. 53 (1981) 400.

[8] S.A. Wise, B.A. Benner, G.D. Boyd, S.N. Chesler, R.E. Rebbert, M.M. Schantz, Anal. Chem. 60 (1988) 887.

[9] R. Kaliszan, Anal. Chem. 64 (1992) 619A.

[10] M.L. Lee, D.L. Vassilaros, C.M. White, M. Novotny, Anal. Chem. 51 (1979) 768.

[11] D.L. Vassilaros, R.C. Kong, D.W. Later, M.L. Lee, J. Chromatogr. 252 (1982) 1.

[12] C.E. Rostad, W.E. Pereira, J. High Resolut. Chromatogr. Chromatogr. Commun. 9 (1986) 328.

[13] C.A. Krone, D.G. Burrows, D.W. Brown, P.A. Robisch, A.J. Friedman, D.C. Malins, Environ. Sci. Technol. 20 (1986) 1144.

[14] W.P. Eckel, T.A. Jacob, J.F. Fisk, in: Superfund 89 Proceedings of the 10th National Conference, Washington, DC, November, 1989, Hazardous Materials Control Research Institute, Silver Spring, MD, 1989, p. 86.

[15] D.R. Lide (Ed.), CRC Handbook of Chemistry and Physics, 75th ed., CRC Press, Boca Raton, FL, 1995.

[16] D.R. Lide (Ed.), CRCnetBase 1999 (CD-ROM), 1999.

[17] J.A. Dean (Ed.), Lange's Handbook of Chemistry, 13th ed., McGraw-Hill, New York, 1985.

[18] S. Budavari (Ed.), The Merck Index, 11th ed., Merck & Co., Rahway, NJ, 1989.

[19] NIST Structures & Properties Database and Estimation Program Version 1.1, US Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MD, 1991.

[20] NIST Chemistry WebBook, http://webbook.nist.gov/chemistry/, verified 4/2003.

[21] J. Olivero, K. Kannan, J. Chromatogr. A. 849 (1999) 621–627.

[22] M. de Freitas Soares, F.D. Monache, V.E.F. Heinzen, R. A. Yunes J. Braz. Chem. Soc. 10 (1999) 189–196.

[23] EPA Estimation Program Interface (EPI) Suite US Environmental Protection Agency Office of Pollution Prevention Toxics, http://www.epa.gov/opptintr/exposure/docs/episuite.htm, verified 4/2003.

[24] S.E. Stein, R.L. Brown, J. Chem. Inf. Comput. Sci. 34 (1994) 581–587.

[25] Automated Mass Spectral Deconvolution and Identification System (AMDIS), http://chemdata.nist.gov/mass-spc/amdis/, verified 4/2003.