Available online at www.sciencedirect.com

SCIENCE DIRECT®

Catalysis Today 98 (2004) 217–225

CATALYSIS TODAY

www.elsevier.com/locate/cattod

Prediction of gas chromatographic retention times and indices of sulfur compounds in light cycle oil

Hongbin Du*, Zbigniew Ring, Yevgenia Briker, Patricia Arboleda

National Center for Upgrading Technology, 1 Oil Patch Drive, Devon, Alta., Canada T9G 1A8

Available online 13 September 2004

Abstract

Ninety sulfur compounds consisting of mercaptans, sulfides and thiophenes, were identified in a fluid-catalytic-cracking light cycle oil using gas chromatography with atomic emission detection. Their retention times and indices were correlated with molecular descriptors generated from their molecular structures. The best seven- and eight-parameter multi-linear regression models showed good predictive ability. The descriptors involved in the models reflect the geometrical, topological, and electronic properties of the molecules, related to the interactions between the solute and the stationary phase. For the 34 thiophenic sulfur compounds (benzothiophenes and dibenzothiophenes) of most interest in petroleum processing, another two five-parameter multi-linear models were developed for retention times and indices with standard errors $s = 0.61$ and 1.63, respectively. Such models for retention times and/or indices can be used for identification of unknown chromatographic peaks by matching their retention times or indices with those of sulfur compounds of known molecular structure when the corresponding chemical standards are unavailable.

© 2004 Elsevier B.V. All rights reserved.

Keywords: QSPR; Chromatographic retention; Retention time; Retention index; Sulfur compounds; LCO

1. Introduction

As environmental regulations become more stringent, desulfurization of gasoline and diesel blending components remain important areas of research. There are continuing efforts to develop new and improve old technologies suitable for deep desulfurization. These efforts frequently involve kinetics modeling of desulfurization that requires detailed knowledge of the types of sulfur compounds or even individual sulfur species in the feedstock [1–4]. Information on the sulfur species present in a crude or refinery stream to be purchased by a refiner determines the value of such streams. All this calls for improved identification and quantification methods for individual sulfur species.

While the number of sulfur-containing compounds in naphtha is more manageable, the number of potential sulfur compounds in the middle distillate range is excessively large. Recent advances in analytical procedures, particularly sequenced liquid chromatographic separations with gas chromatography–mass spectroscopy (GC–MS) and sulfur-specific GC detectors, have allowed for the identification of many individual sulfur species [3–6]. However, full identification of individual components, particularly in virgin oil fractions, is still far from a reality, as numerous compounds are unavailable as standards. Although the molecule structure can be deduced from mass spectral fragmentation patterns, the large number of compounds of very similar structures present in oil fractions make the assignments very difficult. A method for accurately predicting retention times/indices would be of help for the identification of individual sulfur compounds. Interestingly, there exist certain general rules for chromatographic elution orders of isomers. Depauw and Froment [5] developed a linear regression for substituted benzothiophene component retention time based on the positions of substituents. Their model predicted elution times well for di-, tri, and tetra-methyl substituted benzothiophenes, providing a useful way of identifying some sulfur compounds in a cracked middle distillate. In this paper, we report on the development of more generally applicable correlations between GC retention times/indices of various

* Corresponding author.
E-mail address: hodu@nrcan.gc.ca (H. Du).

0920-5861/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.cattod.2004.07.035

sulfur compounds and their molecular structures. This work was motivated by the desire to identify unknown sulfur chromatographic peaks in oil fractions by matching them with sulfur compounds of known molecular structure. This may be the only way to identify peaks in the absence of corresponding chemical standards.

Quantitative structure–property relationship (QSPR) is a powerful approach for predicting the chromatographic behavior of various classes of compounds [7–9]. The QSPR modeling of a given set of molecules first involves the generation, from their molecular structures, of a large number of descriptors which are thought to reflect solute-stationary-phase interactions responsible for chromatographic retention. These molecular descriptors may contain constitutional, topological, geometrical, electronic and, in some cases, physical properties of the solute molecule in question. The statistical analysis that follows reduces the number of these descriptors and gives rise to a QSPR model. Studies have shown that the descriptors in predictive QSPR correlations usually differ depending on the sets of solute molecules and the stationary phases [10,11]. As the stationary phase changes, the influence of the distribution of the eluted molecules changes and different descriptors become important. On the other hand, even if a QSPR correlation could be developed for broad sets of diverse molecules, it is usually necessary to model individual classes of compounds separately to improve the accuracy of predictions.

2. Experimental

2.1. Data sets

The database of retention times used in this study was generated using an Agilent Technologies HP-6890 gas chromatography system equipped with a RESTEK 10526 (Crossbond 50% methyl polysiloxane capillary column with dimensions 60 m × 250 μm × 0.25 μm) and a G2350A Atomic Emission Detector (AED). Helium was used as the carrier gas at a constant pressure of 32.90 psi, initial flow of 2.1 mL/min and average velocity of 32 cm/min. The initial oven temperature was 40 °C; and it was heated at a rate of 2 °C/min to a final temperature of 285 °C, which was maintained for 20 min. All the data were generated using samples containing 2–20 ppm sulfur, prepared in a dichloromethane to give an acceptable retention time deviation of 0.047 min per peak. The standard compounds were obtained commercially from various sources. The peak assignments in the chromatogram of a catalytic-cracker light cycle oil (LCO) were established based on retention times that matched previously analyzed standards. The 90 sulfur compounds considered in this study are listed in Table 1. They include 19 mercaptans and thiols, 11 sulfides and disulfides, 10 thiophenes, 21 benzothiophenes, 13 dibenzothiophenes and 16 other polycyclic aromatic sulfur compounds.

Table 1
A comparison of observed retention times (min) and indices, calculated using the best multi-linear correlations (see Table 2) for 90 sulfur-containing compounds

No.	Structure	RT	Calculated RT	Difference	%	RI	Calculated RI	Difference	%
1	*iso*-Propyl mercaptan	5.042	2.95	−2.09	−41.4	48.08	56.15	8.07	16.8
2	*n*-Propyl mercaptan	5.05	3.20	−1.85	−36.6	48.29	58.28	9.99	20.7
3	Ethyl methyl sulfide	5.195	2.90	−2.29	−44.2	52.16	60.98	8.82	16.9
4	*iso*-Butyl mercaptan	6.125	7.12	0.99	16.2	76.97	79.98	3.00	3.9
5	*tert*-Amyl mercaptan	6.852	6.98	0.13	1.9	96.37	91.79	−4.58	−4.8
6	Ethyl sulfide	6.951	6.57	−0.38	−5.5	99.01	84.02	−14.99	−15.1
7	Thiophene	6.988	6.07	−0.91	−13.1	100.00	93.34	−6.66	−6.7
8	*n*-Butyl mercaptan	7.328	7.73	0.40	5.4	100.86	84.67	−16.19	−16.0
9	Dimethyl disulfide	10.21	8.71	−1.50	−14.7	108.11	95.88	−12.23	−11.3
10	3-Methylthiophene	11.222	12.14	0.92	8.2	110.66	109.43	−1.23	−1.1
11	2-Methylthiophene	11.428	12.18	0.76	6.6	111.18	110.41	−0.77	−0.7
12	*n*-Amyl mercaptan	11.553	12.33	0.77	6.7	111.50	107.46	−4.04	−3.6
13	Allyl sulfide	14.733	15.98	1.25	8.5	119.50	132.39	12.89	10.8
14	Di-*n*-propyl sulfide	15.42	17.25	1.83	11.9	121.23	122.38	1.15	1.0
15	2,5-Dimethylthiophene	15.875	17.52	1.64	10.3	122.38	129.72	7.34	6.0
16	2,4-Dimethylthiophene	15.927	18.12	2.19	13.8	122.51	128.69	6.18	5.0
17	2-Ethylthiophene	16.237	19.82	3.59	22.1	123.29	126.35	3.06	2.5
18	*n*-Hexyl mercaptan	17.489	17.93	0.44	2.5	126.44	128.31	1.87	1.5
19	*n*-Heptyl mercaptan	24.47	23.65	−0.82	−3.3	144.02	150.59	6.57	4.6
20	Benzene thiol	27.008	32.23	5.22	19.3	150.41	144.86	−5.55	−3.7
21	*n*-Butyl sulfide	28.687	28.12	−0.57	−2.0	154.64	157.07	2.43	1.6
22	*tert*-Butyl disulfide	31.622	35.12	3.50	11.1	162.03	177.38	15.35	9.5
23	2-Pentylthiophene	36.901	41.79	4.89	13.3	175.32	169.57	−5.75	−3.3
24	2-Ethyl benzenethiol	41.472	43.51	2.04	4.9	186.83	189.38	2.55	1.4
25	4-Ethyl thiophenol	41.741	40.98	−0.77	−1.8	187.51	182.73	−4.78	−2.5
26	2-*n*-Butyl-5-ethyl thiophene	42.494	43.58	1.09	2.6	189.40	179.80	−9.61	−5.1
27	2,4-Dimethyl benzenethiol	42.639	40.96	−1.68	−3.9	189.77	188.09	−1.68	−0.9
28	3,5-Dimethyl benzenethiol	42.714	42.90	0.19	0.4	189.96	193.61	3.65	1.9
29	2,5-Dimethyl benzenethiol	42.842	41.60	−1.24	−2.9	190.28	191.12	0.84	0.4

Table 1 (*Continued*)

No.	Structure	RT	Calculated RT	Difference	%	RI	Calculated RI	Difference	%
30	2,6-Dimethyl benzenethiol	43.459	42.81	−0.65	−1.5	191.83	193.45	1.62	0.8
31	2-Isopropyl benzenethiol	45.016	47.67	2.66	5.9	195.76	200.54	4.79	2.4
32	Benzothiophene	46.702	49.47	2.77	5.9	200.00	203.41	3.41	1.7
33	Thieno[2,3-*b*]thiophene	48.415	48.43	0.02	0.0	204.31	205.51	1.20	0.6
34	2,4,6-Trimethyl thiophenol	50.416	47.10	−3.32	−6.6	209.35	212.43	3.08	1.5
35	7-Methyl-benzo[*b*]thiophene	53.101	55.08	1.98	3.7	216.11	221.22	5.11	2.4
36	2-Methyl-benzo[*b*]thiophene	53.385	53.46	0.08	0.1	216.83	217.26	0.43	0.2
37	5-Methyl-benzo[*b*]thiophene	53.949	53.42	−0.53	−1.0	218.25	216.95	−1.30	−0.6
38	6-Methyl-benzo[*b*]thiophene	54.361	53.09	−1.27	−2.3	219.28	216.56	−2.72	−1.2
39	3-Methyl-benzo[*b*]thiophene	55.13	56.49	1.36	2.5	221.22	219.33	−1.89	−0.9
40	4-Methyl-benzo[*b*]thiophene	55.26	55.85	0.59	1.1	221.55	224.58	3.03	1.4
41	2,7-Dimethyl-benzo[*b*]thiophene	58.847	59.84	0.99	1.7	230.58	235.85	5.27	2.3
42	1-Dodecane thiol	59.706	52.63	−7.08	−11.9	232.74	232.36	−0.38	−0.2
43	2,5-Dimethyl-benzo[*b*]thiophene	60.264	57.71	−2.55	−4.2	234.15	231.66	−2.49	−1.1
44	2,6-Dimethyl-benzo[*b*]thiophene	60.579	57.82	−2.76	−4.5	234.94	231.60	−3.34	−1.4
45	4,7-Dimethyl-benzo[*b*]thiophene	60.757	61.78	1.02	1.7	235.39	242.86	7.47	3.2
46	2,4-Dimethyl-benzo[*b*]thiophene	61.029	59.91	−1.12	−1.8	236.07	237.60	1.53	0.6
47	3,7-Dimethyl-benzo[*b*]thiophene	61.06	61.76	0.70	1.2	236.15	236.14	−0.01	0.0
48	3,5-Dimethyl-benzo[*b*]thiophene	61.832	61.48	−0.35	−0.6	238.09	236.22	−1.87	−0.8
49	3,6-Dimethyl-benzo[*b*]thiophene	61.913	61.16	−0.75	−1.2	238.30	235.51	−2.79	−1.2
50	2-Phenyl thiophene	62.016	63.53	1.51	2.4	238.56	247.24	8.68	3.6
51	2,3-Dimethyl-benzo[*b*]thiophene	62.408	61.79	−0.62	−1.0	239.54	233.63	−5.91	−2.5
52	6,7-Dimethyl-benzo[*b*]thiophene	62.851	59.89	−2.96	−4.7	240.66	235.74	−4.92	−2.0
53	3-Phenyl thiophene	63.529	61.33	−2.20	−3.5	242.37	235.10	−7.27	−3.0
54	3,4-Dimethyl-benzo[*b*]thiophene	66.615	64.57	−2.05	−3.1	250.14	244.47	−5.67	−2.3
55	2,3,7-Trimethyl-benzo[*b*]thiophene	68.114	68.43	0.32	0.5	253.91	253.93	0.02	0.0
56	2,3,5-Trimethyl-benzo[*b*]thiophene	69.013	66.61	−2.40	−3.5	256.17	250.08	−6.10	−2.4
57	2-Naphthalenethiol	71.799	68.74	−3.06	−4.3	263.19	262.18	−1.01	−0.4
58	Phenyl sulfide	74.802	80.47	5.66	7.6	270.75	267.70	−3.05	−1.1
59	2-Phenylthio thiophene	75.232	75.13	−0.10	−0.1	271.83	275.79	3.95	1.5
60	2,3,4,7-Tetramethyl-benzo[*b*]thiophene	78.516	75.45	−3.07	−3.9	280.10	272.95	−7.15	−2.6
61	1,2,3,4-Tetrahydro dibenzothiophene	84.009	78.53	−5.48	−6.5	293.93	275.88	−18.05	−6.1
62	Dibenzothiophene	86.42	87.38	0.96	1.1	300.00	300.51	0.51	0.2
63	Naphtho[1,2-*b*]thiophene	87.153	85.58	−1.57	−1.8	302.13	303.89	1.76	0.6
64	Benzyl sulfide	89.099	96.36	7.26	8.2	307.77	305.97	−1.80	−0.6
65	Naphtha[2,1,*b*]thiophene	89.229	85.58	−3.65	−4.1	308.15	308.70	0.55	0.2
66	Naphtha[2,3-*b*]thiophene	90.602	86.74	−3.86	−4.3	312.13	318.14	6.00	1.9
67	Phenyl disulfide	90.919	89.44	−1.48	−1.6	313.05	321.32	8.27	2.6
68	4-Methyl-dibenzothiophene	91.164	93.24	2.07	2.3	313.76	316.97	3.21	1.0
69	3-Methyl-dibenzothiophene	92.579	91.11	−1.46	−1.6	317.87	313.93	−3.94	−1.2
70	1-Methyl-dibenzothiophene	94.479	92.96	−1.52	−1.6	323.38	316.59	−6.79	−2.1
71	4,6-Dimethyl-dibenzothiophene	95.946	99.58	3.63	3.8	327.64	333.58	5.94	1.8
72	1,4-Dimethyl-dibenzothiophene	99.241	99.61	0.37	0.4	337.20	335.56	−1.64	−0.5
73	Thianthrene	100.037	100.97	0.94	0.9	339.51	341.77	2.26	0.7
74	4-Ethyl-6-methyl-dibenzothiophene	100.356	104.28	3.92	3.9	340.43	343.46	3.03	0.9
75	2,3-Dimethyl-dibenzothiophene	100.923	98.02	−2.90	−2.9	342.08	333.25	−8.83	−2.6
76	2,4,6-Trimethyl-dibenzothiophene	101.312	104.88	3.57	3.5	343.21	349.66	6.45	1.9
77	2,4,8-Trimethyl-dibenzothiophene	101.949	105.66	3.71	3.6	345.05	354.27	9.22	2.7
78	2,4,7-Trimethyl-dibenzothiophene	102.426	103.45	1.03	1.0	346.44	349.41	2.97	0.9
79	Benzyl disulfide	103.287	100.13	−3.15	−3.1	348.94	354.20	5.26	1.5
80	1,4,7-Trimethyl-dibenzothiophene	104.655	104.36	−0.29	−0.3	352.90	351.01	−1.89	−0.5
81	Acenaphtho[1,2-*b*]thiophene	104.695	103.51	−1.19	−1.1	353.02	369.12	16.10	4.6
82	1,3,7-Trimethyl-dibenzothiophene	105.674	102.17	−3.51	−3.3	355.86	347.44	−8.42	−2.4
83	Phenanthro[4,5-*bcd*]thiophene	107.42	108.07	0.65	0.6	360.93	360.74	−0.19	−0.1
84	Benzo(*b*)naphtho[2,1-*d*]thiophene	120.888	122.40	1.52	1.3	400.00	405.51	5.51	1.4
85	Benzo(*b*)naphtha(1,2-*d*)thiophene	121.954	123.13	1.18	1.0	403.69	406.13	2.44	0.6
86	3-[1-Naphthalenyl]benzo[*b*]thiophene	122.016	127.99	5.98	4.9	405.89	405.83	−0.06	0.0
87	Phenanthro(9,10-*b*) thiophene	123.814	124.37	0.55	0.4	415.28	413.68	−1.60	−0.4
88	Phenanthro(4,3-*b*) thiophene	124.042	124.10	0.05	0.0	416.47	413.89	−2.58	−0.6
89	Phenanthro[1,2-*b*] thiophene	124.26	120.93	−3.33	−2.7	417.61	408.71	−8.90	−2.1
90	Phenanthro(2,1-*b*) thiophene	125.977	121.43	−4.55	−3.6	426.57	414.34	−12.23	−2.9

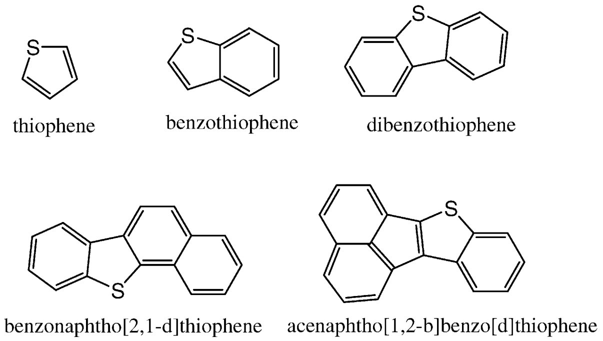
thiophene
benzothiophene
dibenzothiophene
benzonaphtho[2,1-d]thiophene
acenaphtho[1,2-b]benzo[d]thiophene

Fig. 1. Sulfur compounds used as standards in the retention index calculation.

Kovats retention indices (RIs) were calculated for the retention-time database in order to account for retention time shift due to a pressure change or deterioration of the column [12]. Instead of using *n*-alkanes as retention index standards, as defined by the original Kovats retention index scale, hydrogen sulfide, thiophene, benzothiophene, dibenzothiophene, benzonaphtho[2,1-*d*]thiophene and acenaphtho[1,2-*b*]benzo[*d*]thiophene were arbitrarily assigned the exact RI values of 0, 100, 200, 300, 400 and 500, respectively, somewhat relative to the number of aromatic rings in each of these compounds (see Fig. 1). At temperature-programmed analyses, the Kovats index equation is simplified utilizing direct numbers instead of their logarithm as follows:

$$\mathrm{RI} = 100y + 100\left[\frac{t_\mathrm{r}(x) - t_\mathrm{r}(y)}{t_\mathrm{r}(z) - t_\mathrm{r}(y)}\right],$$

where t_r is the retention time, x the compound of interest, y and z correspond to the aromatic sulfur standards eluting immediately prior to and after the compound of interest, respectively.

2.2. Descriptor generation

A large number of descriptors were generated using the CODESSA program (CODESSA 2.64. by SemiChem, Inc., 2002) in this study. These included the conventional sets of constitutional, geometrical, topological, electronic and quantum-chemical descriptors. The constitutional descriptors are derived from the molecular composition of the compounds, including the atomic and molecular weights, counts of atoms, bonds, special groups, etc. The geometrical descriptors are calculated from 3D-coordinates of the atoms in the given molecule after minimizing its internal energy. These include moments of inertia, shadow indices, molecular volume, surface area and gravitational indices. Topological descriptors, derived from molecular graph invariants, describe the atomic connectivity in the molecule (e.g. indices of Wiener [13], Randic [14], Kier and Hall [15], Balaban [16], information content and their derivatives [17,18]). The electronic descriptors reflect characteristics of the partial charge distribution of the molecule. The quantum-chemical descriptors are supplementary to the conventional descriptors and related to the charges, bonding, energies, reactivities and thermodynamic properties of molecules [19]. They were calculated using quantum semi-empirical PM3 methodology of the AMPAC program (AMPAC 7.0, Semichem Inc., 2003). Charged partial surface area (CPSA) descriptors were also calculated using both empirical (Zefirov's partial charge) and quantum-chemical methods to describe polar intermolecular interactions. These descriptors were invented by Jurs and coworkers [20,21] and they account for the partial atomic charge distribution and solvent-accessible surface area of the molecule. To this end, the surface of the molecule is represented by the surface of overlapped hard spheres defined by the van der Waals radii of the atoms. The surface area is traced out by a sphere representing a solvent molecule, termed the solvent-accessible surface area.

2.3. Statistical treatment

The QSPR models were built using the best multi-linear regression method and the following methodology implemented in the CODESSA program. A total of 467 descriptors were initially calculated for the entire data set of 90 compounds. The total number of descriptors was reduced by eliminating the descriptors that were deemed insignificant (i.e. one-parameter correlation coefficient is less than 0.01). From these leftover descriptors, all orthogonal pairs were found and treated by using the two-parameter regression with the property. The top pairs ($\leq$400) with highest regression correlation coefficients were chosen to carry out higher-order regression analysis that involved successively adding (from the pool of still available descriptors) the third (and then higher) descriptor to the best two-descriptor correlation until the Fisher criterion at a given probability level was smaller than that for the previous correlation, or until all the available descriptors had been tried.

3. Results and discussion

For the set of 90 compounds, the best possible seven- and eight-parameter multi-linear models were found for retention times and indices, respectively. The ratios of the number of data points to the number of descriptors for the two models (15 and 13, respectively) were kept greater than 10 to prevent overfitting. Ratios greater than 5 are usually considered sufficient [11]. A summary of the two models is listed in Table 2. The correlation coefficient (R) is an indicator of the amount of variation in the dependent variable accounted for by the model. The Fisher F-criterion can be interpreted as a ratio of the variance explained by the model and the variance not explained by the model, characteristics of the degree of statistical credibility of the model. The Student's t-value characterizes the relative

Table 2
Best multi-linear regression models of retention times and retention indices for 90 sulfur-containing compounds

Descriptors	RT	*t*	RI	*t*
R	0.9974		0.9978	
F	2654		2637	
s	2.66		6.70	
R_{cv}	0.9963		0.9967	
*d*0	37.8 ± 9.2	4.1	−0.953 ± 8.117	−0.1
*d*1	−0.0523 ± 0.0012	−43.7		
*d*2	−11.033 ± 2.064	−5.4		
*d*3	−0.0229 ± 0.0057	−4.0		
*d*4	139.7 ± 19.6	7.1		
*d*5	−9.096 ± 0.861	−10.6		
*d*6	−0.558 ± 0.053	−10.4	−1.618 ± 0.153	−10.6
*d*7			5.609 ± 0.822	6.8
*d*8			381.7 ± 29.8	12.8
*d*9			132.0 ± 42.8	3.1
*d*10			2.324 ± 0.049	42.3
*d*11			2444.5 ± 288.3	8.5
*d*12			−360.7 ± 54.2	−6.6

Note: The parameters of the descriptors are listed with 95% confidence; *t*, Student's *t*-factor; *R*, the correlation coefficient; *F*, Fisher criterion; *s*, the standard error of estimate; R_{cv}, the leave-one-out cross-validated correlation coefficient; *d*0, intercept; *d*1, total molecular 1-center electron–nuclear attraction; *d*2, minimum valency of a C atom; *d*3, DPSA-1 difference in the positively and negatively charged solvent-accessible surface area (Zefirov's charges); *d*4, relative number of benzene rings; *d*5, Balaban index; *d*6, ZX Shadow; *d*7, Kier shape index (order 2); *d*8, relative number of rings; *d*9, relative number of S atoms; *d*10, α-polarizability; *d*11, max. partial charge for a H atom (Zefirov's charges); *d*12, RPCG relative positive charge (Zefirov's charges).

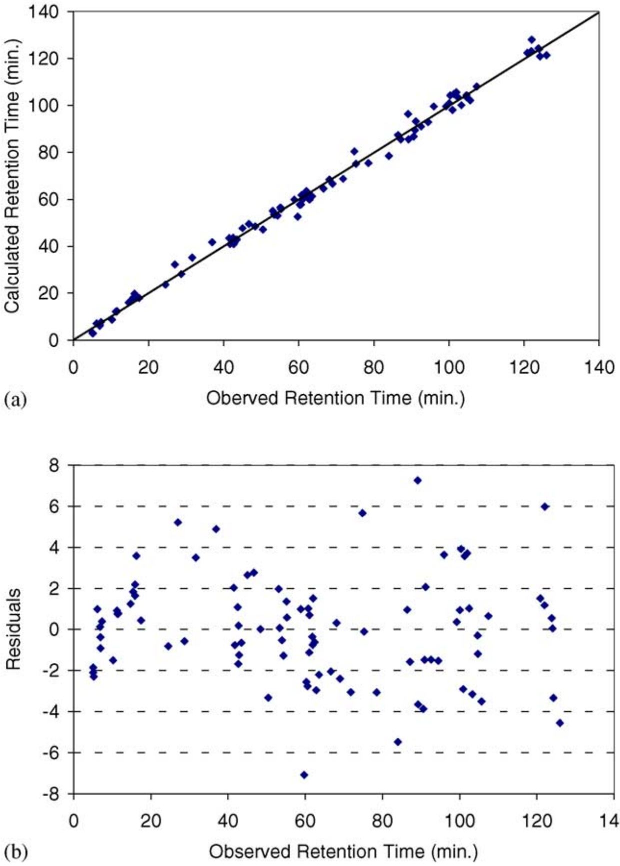

Fig. 2. (a) Calculated vs. observed retention times for 90 sulfur compounds. (b) Plot of residuals vs. observed retention times for 90 sulfur compounds.

significance of the parameter in a particular correlation, i.e. part of the variance explained by the parameter. The calculated retention times and indices are compared with the corresponding experimental values in Table 1. Figs. 2 and 3 show the relationships between the calculated and observed retention times and indices, respectively.

Chromatographic retention results from the solvation and partition of individual compounds in a stationary phase. The solubility is affected by intermolecular forces which include hydrogen bonding, ion–dipole, dipole–dipole, dipole–induced dipole interactions, etc. Solubility is related to the descriptors accounting for the counts of atoms, functional groups or charged moieties, surface areas of donor and acceptor groups, hydrophobicity, polarizability and conformation flexibility. Chromatographic partition, in turn, results from adsorption–desorption of compounds on the stationary phase, during which the change in degree of freedom of the molecule occurs. The entropic effects also play important roles. As a result, the chromatographic retention depends mainly on the electronic properties, and the size and shape of the solute molecule.

The seven-parameter model for retention times contains two electronic (*d*1, *d*2), one CPSA (*d*3), one constitutional (*d*4), one topological (*d*5), and one geometrical descriptor (*d*6) (Table 2). The total molecular 1-center electron–nuclear attraction (*d*1) is the most important descriptor in the correlation, which explains about 44% of the variances as indicated by the Student's *t*-factor. This descriptor represents a sum of the attraction of electrons and atoms that constitute a molecule and is related to the polarity and size of the molecule. The more polar functional groups and atoms, the larger absolute value of the total electron–nucleus attraction energy. More polar compounds have stronger interactions with the stationary phase and, thus, longer retention time, as indicated by the negative value of the corresponding model parameter. Descriptor *d*2 (minimum valency of a C atom) characterizes the hydrophobicity of the molecule. The negative value of the parameter for this descriptor suggests that the compounds with a higher value for minimum valency of a C atom have higher hydrophobicity (e.g. more H atoms attached), less hydrophobic/hydrophilic interaction between the molecule and the GC stationary phase and, thus, smaller retention time. The CPSA descriptor *d*3 is the difference between the values of the positively and negatively charged partial surface areas of a molecule, describing conformational flexibility of the molecule. It provides information related to strong polar intermolecular interactions (e.g. electron pair donor/acceptor interaction) in chromatographic solvation and partition. The count of benzene rings (*d*4) represents the size and electronic distribution of a molecule, accounting for the polarizability and hydrophobicity of the molecule. The topological

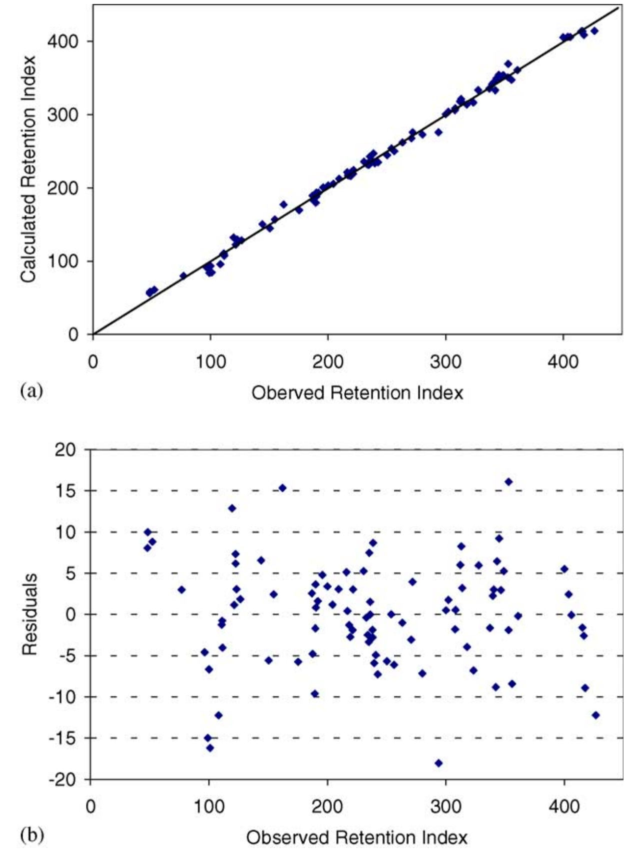

Fig. 3. (a) Calculated vs. observed retention indices for 90 sulfur compounds. (b) Plot of residuals vs. observed retention indices for 90 sulfur compounds.

descriptor Balaban index (*d*5) characterizes the branching pattern within a molecule [16]. This type of descriptor is evaluated from various combinations and weightings of the vertices (atoms) and edges (bonds) of the molecular graph (chemical structure). The geometrical descriptor *d*6 (molecular shadow area in the *z*–*x* plane) characterizes the three-dimensional shape of the molecule. These last two descriptors underscore the importance of factors such as size, degree of branching and steric interaction on the retention behavior. Small and unbranched molecules are expected to have shorter retention times, consistent with the negative parameters of the two descriptors.

For the retention indices, the best eight-parameter model was obtained (Table 2). Of the seven descriptors involved, six were new (i.e. not in the corresponding model for retention times). The topological 'Kier shape index' (*d*7) describes the shape, branching and the hydrophobic tail structure of the molecule [22]. It increases with the size of the hydrophobe. It is lower for branched hydrophobes and even lower for the cyclic structures of the same carbon number. The 'relative number of S atoms' (*d*9) accounts for the electronic effects that the electronegative atoms have on the molecule. The 'maximum partial charge for an H atom' (*d*11) and 'relative positive charge' (*d*12) reflect the ability of a molecule to participate in hydrogen bonding and electrostatic interaction between the molecule and the stationary phase. The α-polarizability (*d*10) is the most contributing descriptor in the best seven-parameter retention index model, accounting for about 42% of the variance. This descriptor gave the best one-descriptor correlation for both retention times (R^2 = 0.9746) and retention indices (R^2 = 0.9765). The α-polarizability characterizes the effectiveness of intermolecular induction and dispersion interaction between the compound and the GC stationary phase. The higher the polarizability, the stronger the interactions and, thus, the larger the retention indices.

It is important to note that CODESSA's semi-automatic variable selection generated different sets of descriptors for the retention times and retention indices, which clearly resulted from the rescaling of retention times relative to those of 'standard' thiophenic compounds. This finding should serve as a reminder that the mathematical structures of the models developed here are far from being theoretically correct, most likely nonlinear, structures and that the large number of partially correlated descriptors selected by CODESSA may lead to different sets of descriptors describing essentially the same property. Indeed, the descriptors involved in the retention index model are similar to those in the retention time model. Both models involve the geometrical descriptor, 'ZX shadow' (*d*6). Among the six new descriptors involved in the retention index model, the α-polarizability (*d*10) is highly correlated with the descriptors *d*1 (total molecular 1-center electron–nuclear attraction) (R^2 = 0.9842) and *d*4 (relative number of benzene rings, R^2 = 0.8128) used in the retention time model. The descriptor 'relative number of rings' *d*8 in the retention index model corresponds to the descriptor 'relative number of benzene rings' (*d*4, R^2 = 0.7309) of the retention time model. The 'Kier shape index' (*d*7) consists of information also used by 'Balaban index' (*d*5) and 'DPSA-1' (*d*3) of the retention time correlation. The intercorrelated descriptors *d*9 (relative number of S atoms), *d*11 (maximum partial charge for a H atom) and *d*12 (relative positive charge) are to some extent accounted for by *d*1 (total molecular 1-center electron–nuclear attraction) and *d*2 (minimum valency of a C atom) in the retention time model.

As shown in Table 2 and Figs. 2 and 3, the multi-linear correlations for retention time and indices are of excellent quality. Both models explain more than 99% of the variances in the observed values of the retention times and indices for the 90 sulfur compounds. Although the standard errors are still too large to predict small values (see Table 1), in general, the correlations predict retention times and indices surprisingly well within standard errors of 2.66 and 6.70, respectively. The leave-one-out cross-validation correlation coefficients (0.9963 for retention times, and 0.9967 for retention indices) indicate good stability of the obtained models.

The errors of prediction make it difficult for these models to be applied for peak identification in a chromatogram, particularly for small retention times. Since the thiophenic

Table 3
Best multi-linear regression models of retention times and retention indices for the thiophenic subset of 34 sulfur-containing compounds

Descriptors	RT	RI
R	0.9996	0.9995
F	8669	8400
R_{cv}	0.9994	0.9994
s	0.6098	1.6337
$d0$	−529.4 ± 39.8	206.4 ± 16.4
$d3$		−0.226 ± 0.029
$d8$		−485.1 ± 63.4
$d10$		2.560 ± 0.041
$d13$		−0.444 ± 0.047
$d14$	347.6 ± 23.9	
$d15$	7.425 ± 0.300	
$d16$	68.11 ± 6.24	
$d17$	−366.9 ± 44.0	

Note: The parameters of the descriptors are listed with 95% confidence; $d13$, total enthalpy (300 K)/no. of atoms; $d14$, relative number of C atoms; $d15$, Kier and Hall index (order 0); $d16$, zero point vibrational energy/no. of atoms; $d17$, FNSA-2 fractional total charge weighted partial negative surface area.

sulfur compounds such as benzothiophenes and dibenzothiophenes are of more interest in petroleum processing [1,5,23–25], a subset of data containing only these compounds was selected for further study to improve the correlation. This new database consisted of 21 benzothiophenes and 13 dibenzothiophenes. Five-parameter models for both retention times and indices were found using the best multi-linear regression analysis (Table 3). The descriptors involved in the two models are similar to those from the study using the full database. The retention time model contains one constitutional ($d14$), one topological ($d15$), one electronic ($d16$) and one CPSA ($d17$) descriptor, while the retention index model contains one constitutional ($d3$), two electronic ($d8$, $d10$) and one CPSA ($d13$) descriptors. The 'relative number of C atoms' ($d14$) describes the constituent atoms of the molecules. Kier and Hall index ($d15$) is a molecular connectivity index containing information about the size and the degree of branching in the molecule [15]. The CPSA descriptor PNSA-2 ($d17$) is the sum of the surface area of the negative portion of the molecule divided by its negative charge. The 'zero point vibrational energy/number of atoms' ($d16$) represents the conformational flexibility. Higher values correspond to higher conformational flexibility, higher possibility for strong interaction with the stationary phase and, consequently, longer retention times. The descriptor 'total enthalpy/number of atoms' ($d13$) in the retention index model is related to the hydrophobicity, solvation, conformation and binding affinity of the molecule.

As shown in Table 3 and Figs. 4 and 5, the new models show much improved performance for prediction of retention times (s = 0.6098) and indices (s = 1.6337) for 34 thiophenic sulfur compounds. The retention time predictions vary within ±1.0 min (maximum error). Only 3-methyl-benzo[b]thiophene and 2,3-dimethyl-dibenzothiophene have larger prediction errors of −1.28 min (or −2.3%) and −1.21 min (or −1.2%), respectively. The relative errors of the predicted 34 retention indices are under 1.0%, with the exception of 5-methyl-benzo[b]thiophene (2.33 or 1.1%), 3,7-dimethyl-benzo[b]thiophene (−3.07 or −1.3%) and 2,3,5-trimethyl-benzo[b]thiophene (3.45 or 1.3%). The balanced distributions of the residuals suggest no systematic error generated by the two models. The cross-validation correlation coefficients, in comparison with the coefficient of determination, indicate good stability of both models. Unlike the previous models, these models can be used for identification of unknown sulfur peaks in a chromatogram.

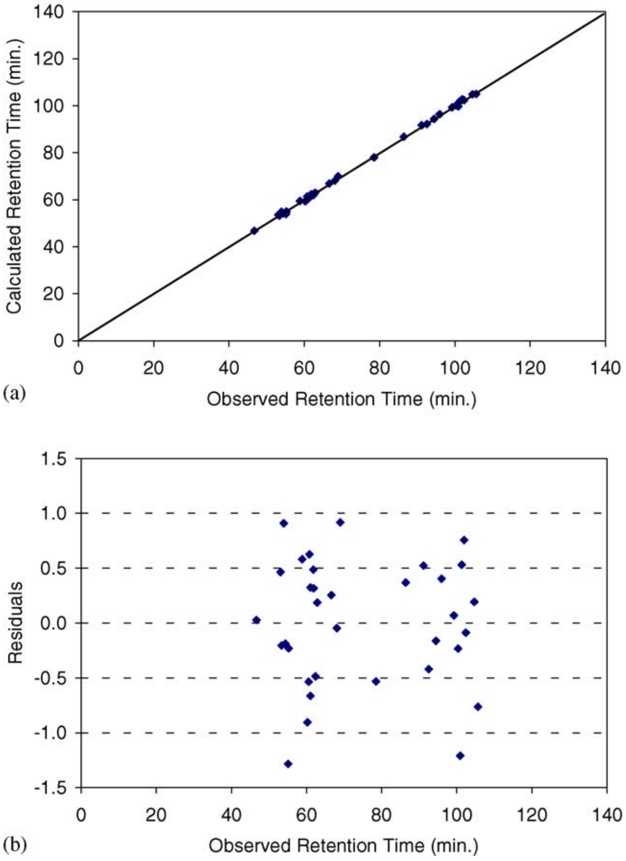

Fig. 4. (a) Calculated vs. observed retention times for 34 sulfur compounds. (b) Plot of residuals vs. observed retention times for 34 sulfur compounds.

It is worth noting that the relative prediction errors of the QSPR correlations developed here are similar to the errors of experimental determination of retention times and indices, indicating limited opportunity for model improvement even with fundamental models. Furthermore, the retention times strongly depend on the condition of the GC column, operating conditions employed (e.g. pressure), and even solute concentrations. The concept of the retention index was invented to account for such variations. In this regard, it is more reliable to use the retention index model for the identification of sulfur compounds. Even in cracked middle distillates with relatively short substituents on the relatively

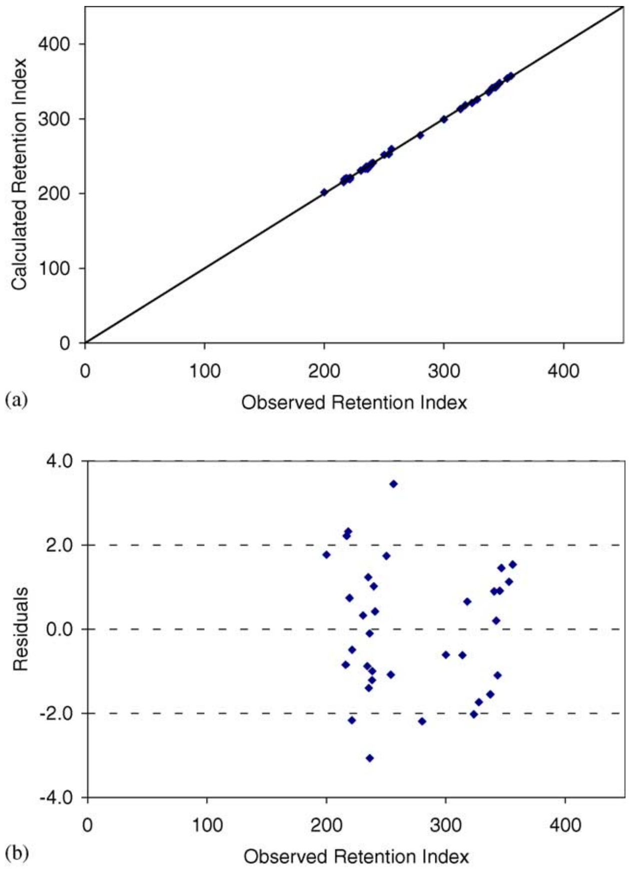

Fig. 5. (a) Calculated vs. observed retention indices for 34 sulfur compounds. (b) Plot of residuals vs. observed retention indices for 34 sulfur compounds.

a few key structures, there exist a large number of sulfur compounds with closely related structures. Co-elution makes it extremely difficult to identify and quantify various isomers (an example of chromatographic separation of sulfur compounds in LCO is shown in Fig. 6). This calls for better separation (e.g. longer columns) and even more accurate prediction models for retention times/indices to enable positive identification of sulfur compounds.

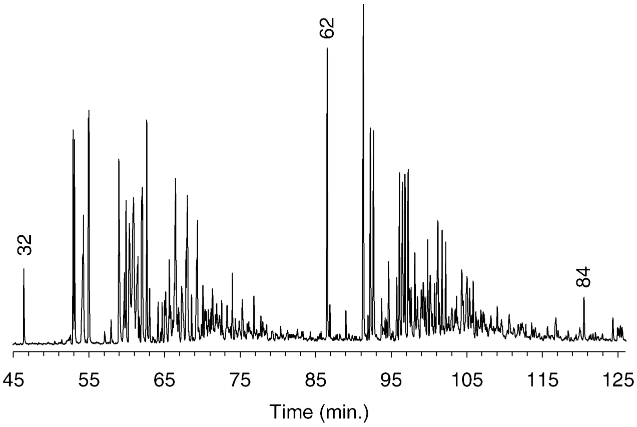

Fig. 6. Sulfur chromatogram of the LCO using GC-AED (the numbers correspond to the compounds in Table 1).

It is probably possible to further improve our ability to identify unknown sulfur compounds. For example, the subset of data containing only 21 benzothiophenes gives rise to each four-descriptor linear model for retention times ($R^2 = 0.9972$, $F = 1430$, $s = 0.4050$, $R^2_{\mathrm{cv}} = 0.9952$) and indices ($R^2 = 0.9972$, $F = 1429$, $s = 1.020$, $R^2_{\mathrm{cv}} = 0.9952$), which can predict retention times and indices within ±0.6 (minutes) and ±1.5, respectively, with the exception of 2,3,5-trimethyl-benzo[*b*]thiophene (1.04 and 2.61). The subset of data of 13 dibenzothiophenes produces three-descriptor linear models for retention times ($R^2 = 0.9983$, $F = 1739$, $s = 0.2735$, $R^2_{\mathrm{cv}} = 0.9962$) and indices ($R^2 = 0.9983$, $F = 1741$, $s = 0.7931$, $R^2_{\mathrm{cv}} = 0.9962$) that can also predict retention times and indices within ±0.5 (minutes) and ±1.5, respectively. Unfortunately, in these cases, the small data sets limit the ability to predict retention data due to statistical irregularities. Another opportunity to identify unknown sulfur peaks in a chromatogram is to use a similar QSPR approach to correlate the reactivity of various sulfur compounds. The retention time/index and reactivity correlations used in tandem will offer further significant help in this regard.

4. Conclusions

This study demonstrates that QSPR models can be used to predict the retention times and indices without the need for chemical standards. The retention times and indices of 90 sulfur-containing compounds identified in LCO were modeled using multi-linear models based on calculated molecular descriptors. Both models show good correlation and predictive ability. The models were further improved over a subset of data consisting of 34 thiophenic sulfur compounds of most concern in petroleum processing. The resulting five-parameter models for retention times and indices of benzothiophenes and dibenzothiophenes can be used to predict the retention times and indices of unknown but structurally similar compounds with a considerable degree of confidence.

Acknowledgement

Partial funding for NCUT has been provided by the Canadian Program for Energy Research and Development (PERD), the Alberta Research Council (ARC) and the Alberta Energy Research Institute (AERI).

References

[1] X. Ma, K. Sakanishi, I. Mochida, Ind. Eng. Chem. Res. 35 (1996) 2487.

[2] G.F. Froment, G.A. Depauw, V. Vanrysselberghe, Ind. Eng. Chem. Res. 33 (1994) 2979.

[3] H. Schulz, W. Bohringer, P. Waller, F. Ousmanov, Catal. Today 49 (1999) 87.
[4] M. Te, C. Fairbridge, Z. Ring, Petrol. Sci. Technol. 21 (2003) 157.
[5] G.A. Depauw, G.F. Froment, J. Chromatogr. A 761 (1997) 231.
[6] T. Kabe, A. Ishihara, H. Tajima, Ind. Eng. Chem. Res. 31 (1992) 1577.
[7] A.R. Katritzky, U. Maran, V.S. Lobanov, M. Karelson, J. Chem. Inf. Comput. Sci. 40 (2000) 1.
[8] R. Kaliszan, Quantitative Structure–Chromatographic Retention Relationships, Wiley, New York, 1987.
[9] A.R. Katritzky, R. Petrukhin, D. Tatham, S. Basak, E. Benfenati, M. Karelson, U. Maran, J. Chem. Inf. Comput. Sci. 41 (2001) 679.
[10] W. Guo, Y. Lu, X.M. Zheng, Talanta 51 (2000) 479.
[11] T.F. Woloszyn, P.C. Jurs, Anal. Chem. 64 (1992) 3059.
[12] J.T. Andersson, J. Chromatogr. 354 (1986) 83.
[13] H. Wiener, J. Am. Chem. Soc. 69 (1947) 17.
[14] M. Randic, J. Am. Chem. Soc. 97 (1975) 6609.
[15] L.B. Kier, L.H. Hall, Eur. J. Med. Chem. 12 (1977) 307.
[16] A.T. Balaban, Chem. Phys. Lett. 89 (1981) 399.
[17] S.C. Basak, D.K. Harriss, V.R. Magnuson, J. Pharm. Sci. 73 (1984) 429.
[18] L.B. Kier, J. Pharm. Sci. 69 (1980) 807.
[19] M. Karelson, V.S. Lobanov, A.R. Katritzky, Chem. Rev. 96 (1996) 1027.
[20] D.T. Stanton, P.C. Jurs, Anal. Chem. 62 (1990) 2323.
[21] D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci. 32 (1992) 306.
[22] L.B. Kier, Computational Chemical Graph Theory, Nova Science Publishers, New York, 1990.
[23] X. Ma, K. Sakanishi, I. Mochida, Ind. Eng. Chem. Res. 33 (1994) 218.
[24] X. Ma, K. Sakanishi, T. Isoda, I. Mochida, Ind. Eng. Chem. Res. 34 (1995) 748.
[25] S.G. Mossner, S.A. Wise, Anal. Chem. 71 (1999) 58.