Paper V

S.A. Mjøs, O. Grahl-Nielsen

Prediction of gas chromatographic retention of polyunsaturated fatty acid methyl esters

J. Chromatogr. A 1110 (2006) 171-180

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Chromatography A, 1110 (2006) 171–180
JOURNAL OF CHROMATOGRAPHY A
www.elsevier.com/locate/chroma

Prediction of gas chromatographic retention of polyunsaturated fatty acid methyl esters

Svein A. Mjøs [a,*], Otto Grahl-Nielsen [b]

[a] *Fiskeriforskning, Kjerreidviken 16, N-5141 Fyllingsdalen, Norway*
[b] *Department of Chemistry, University of Bergen, Bergen, Norway*

Received 30 November 2005; received in revised form 17 January 2006; accepted 19 January 2006
Available online 7 February 2006

Abstract

Multivariate regression models were applied to predict retention indices as equivalent chain lengths (ECL) for methylene-interrupted polyunsaturated fatty acids. Simple molecular descriptors, the chain length, the number of double bonds and the position of the double bond system, were used as predictors. The merits of different variable combinations were evaluated. For general models, it was necessary to include the distance from the double bond system to both the carbonyl group (Δ-position) and the methyl end of the fatty acid (*n*-position). The best accuracy was found for models including higher order terms of Δ and *n*. For models restricted to *n*-3 and *n*-6 isomers, it was not necessary to include the *n*-position among the variables. The highest residuals for the most accurate models were below 0.06 ECL units, and root mean square error of prediction was below 0.030. The ECL data was achieved by three different temperature programs on a cyanopropyl column.

Keywords: Fatty acid methyl esters; FAME; Polyunsaturated fatty acids; Equivalent chain lengths; Partial least squares regression

1. Introduction

Retention indices based on homologous series of reference compounds are often applied for tentative identifications of analytes in gas chromatography. The advantages of using retention indices instead of retention times for this purpose are obvious, since retention indices are relatively invariant to analytical conditions, such as column dimensions and carrier gas flow. Column temperature will also be of minor importance on most stationary phases. Thus, compounds can be tentatively identified from historical and tabulated data achieved on similar stationary phases. While Kovats' indices [1] based on the *n*-alkanes are well established as a general-purpose retention index system, several other indices have been developed for specific purposes [2]. In analysis of fatty acid methyl esters (FAME), equivalent chain lengths (ECL) [3,4] are the dominating system. Since the introduction of the ECL concept, numerous lists of these values for common and uncommon fatty acids have been published for a large variety of stationary phases.

The ECL system uses the saturated straight chain FAMEs as reference compounds and the ECL-values of the references are by definition equal to the number of carbons in the saturated fatty acid chain. Like Kovats' indices, the ECL concept was originally developed for isothermal analysis where there exist a linear relationship between log t_R' and the number of carbons in members of a homologous series. Today FAMEs are usually analysed using temperature programs where the linear relationship between log t_R' and ECL is not valid. With temperature programming, the relationships between the retention times and ECLs can be established using the van den Dool and Kratz method [5] or by non-linear regressions [6–8].

The fractional chain length is defined as the difference between the ECL-value and the number of carbons in the fatty acid chain of the FAME molecule and is calculated by the following formula:

$$\mathrm{FCL}_{(x)} = \mathrm{ECL}_{(x)} - \mathrm{NC}_{(x)} \quad (1)$$

$\mathrm{NC}_{(x)}$ is the number of carbons in the fatty acid chain. It follows from the definition of ECL that FCLs of the saturated unbranched FAMEs are zero. The unsaturated FAMEs, which on polar columns elute after the saturated FAME with the same

* Corresponding author. Tel.: +47 55 50 12 30; fax: +47 55 50 12 99.
E-mail address: svein.mjos@kj.uib.no (S.A. Mjøs).

0021-9673/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2006.01.092

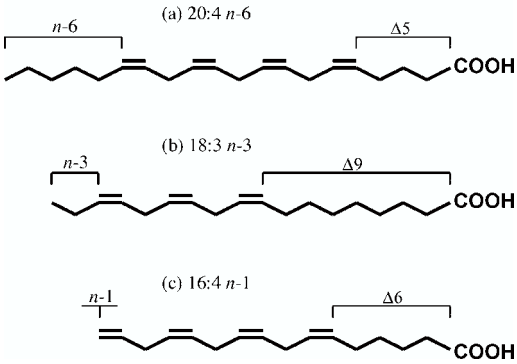

Fig. 1. Fatty acid structure. Δ-Positions and *n*-positions shown for three methylene-interrupted polyunsaturated fatty acids.

number of carbons, have positive FCL values and FCL is used as an indication of the polarity of a compound.

A large number of naturally occurring polyunsaturated fatty acids (PUFA) have been reported. Although there exists numerous exceptions, the double bonds in the majority of PUFAs have *cis* geometry and are separated by a single methylene unit. Examples of fatty acids with methylene-interrupted (MI) double bonds are shown in Fig. 1. Because of the regularity in double bond positions in MI PUFAs, the complete structure is defined if the number of carbons, number of double bonds and the position for the first double bond from the end of the carbon chain is given. It is therefore common to use the notation A:B *n*-C for MI PUFAs, where A is the number of carbons, B is the number of double bonds and C is the position of the first double bond counted from the methyl end of the fatty acid chain. Alternatively, the position of the double bond system may be specified from the carbonyl group by the Δ-position (Fig. 1). The majority of naturally occurring PUFAs have the double bond system in either *n*-3 or *n*-6 positions, and these groups are often referred to as the *n*-3 and *n*-6 "families". However, other families also exist and both *n*-1 and *n*-4 PUFAs are common in marine lipids, which show a very large diversity in fatty acid structure.

It is of interest to be able to predict the chromatographic properties of the large number of possible MI PUFAs that may be found in marine samples, not only for identification purposes, but also for the prediction of possible chromatographic overlaps. Several different strategies have been applied for the prediction of chromatographic properties. One strategy has been to assume that the influence of double bonds is additive, and that FCL values of PUFAs can be predicted by summing the FCL values of monounsaturated fatty acids with double bonds in the corresponding positions [9–12] or by adding FCLs for monoenes to FCLs of other PUFAs [11,12], e.g. FCL for 18:3 *n*-3 is predicted from FCL for 18:2 *n*-6 and FCL for 18:1 *n*-3. The accuracies of these calculations are low because methylene-interrupted double bonds behave different than the sum of the corresponding isolated double bonds and additional correction factors must be introduced [9–12]. The availability of relevant FCL data of the monoenes is also limited.

A much used strategy is to assume that members in the same homologous series will have similar behaviour relative to the saturated analogues, i.e. members in a homologous series have equal FCL values. These relationships have been widely applied in isothermal chromatography, where fatty acids are identified from parallel lines drawn between members of the same homologous series in plots of $\log t_R$ against the number of carbons in the molecule [13,14]. The assumption of similar FCL values for members in the same series can be expected to be accurate as long as interactions between the carbonyl group and the double bond system can be neglected, but may give inaccurate predictions for molecules with double bonds close to the carbonyl group.

Accurate prediction of ECL-values is more challenging with temperature-programmed chromatography than with isothermal chromatography, especially with the highly polar cyanopropyl stationary phases. The properties of these phases have been shown to be more dependent on temperature than properties of other common coatings [15] and the ECL-values of unsaturated fatty acids increase with increasing temperature [7,8,16,17]. Because the highest members in homologous series elute at higher temperatures in temperature-programmed GC, the FCL values increase with chain length within the series. Another effect that limits the use of historical ECL data on these columns is that the polarity of these columns tends to decrease with time, leading to significant drift in the ECL-values [8].

A large number of works have been published on the prediction of chromatographic retention based on various molecular descriptors. These descriptors are often electronic or topological parameters derived from molecular modelling, or experimentally determined physical parameters like boiling points, solubility, etc. Several hundred different parameters are often evaluated [18]. However, if the models are restricted to classes of compounds with limited variation in structure and functional groups, retention indices may often be predicted with high accuracy from a few simple descriptors describing the number and positions of certain atoms or functional groups [19–21].

The purpose of this work has been to investigate whether ECL and FCL values of MI PUFAs analysed by temperature-programmed GC can be accurately predicted by multivariate calibration from simple molecular descriptors, i.e. number of carbons, number of double bonds and the position of the double bond system. The merits of models with different means of describing the position of the double bond system are evaluated.

Three different multivariate regression methods were applied in the study. Multiple linear regression (MLR) permits the estimation of a response variable (y) from several predictors (x-variables). In principal component regression (PCR) and partial least squares regression (PLSR), the original predictors are transformed to orthogonal (uncorrelated) latent variables and the regression is performed with the latent variables as x-variables. Regression on latent variables is known to produce more stable solutions than MLR when there is correlation among the predictors. See refs. [22,23] for further details on the regression methods.

2. Methods

2.1. Instrumentation

All analyses were performed on a HP-5890 GC equipped with split/splitless injector, electronic pressure control, HP-7673A autosampler and HP-5972 MS detector. The system was equipped with G1034C MS Chemstation software. BPX-70, $L = 60$ m, I.D. = 0.25 mm, $d_f = 0.25$ μm (SGE, Ringwood Australia) was used as analytical column. Helium, 99.996% was used as carrier gas.

2.2. GC–MS parameters

Three programs with linear temperature gradients were applied. The samples were injected at an oven temperature of 60 °C that was held for 4 min. The temperature was increased by 30 °C/min to start temperature of 160 (Prg. 1), 175 (Prg. 2) or 190 °C (Prg. 3), followed by a gradient of 2 (Prg. 1), 3 (Prg. 2) or 4 °C/min (Prg. 3) until the final compound had eluted. The injector pressure was increased with oven temperature to give a constant velocity of 26 (Prg. 1), 22 (Prg. 2) or 18 cm/s (Prg. 3). The samples (0.5 μL) were injected in splitless mode. The split valve was opened after 4 min. Injector temperature was 250 °C and MS transfer line temperature 270 °C. The MS detector was used in selected ion monitoring mode, and the ions *m/z* 55, 74, 79, 80, 91 and 93 were recorded at a frequency of 3.5 scans per second. The combination of these ions has proved to be suitable for fatty acid identification [24].

2.3. Samples

The calibration sample was GLC-461 FAME reference mixture (Nu-Chek Prep, Elysian, MN, USA) spiked with additional saturated FAMEs: 19:0, 21:0, 25:0, 26:0, 27:0, 28:0 and 22:3 *n*-3. Other samples were silver ion HPLC fractions of FAMEs from various marine sources (salmon, blue whiting, mussels) that were isolated by silver ion chromatography and identified by mass spectrometry in scan and selected ion monitoring mode [24] and by matching with mass spectra and ECL data from previous investigations [8]. In cases where the same fatty acid appeared in several samples, the median of the calculated ECL-values was applied in the final dataset.

2.4. ECL-regressions

The peak apex was used to determine the retention time and the unbranched saturated fatty acids from C8 to C28 (not including 23:0) were used as references. The relation between retention time and ECL-value was determined by a stepwise procedure using local second order regressions as explained elsewhere [8]. The ECL-regressions were performed in an in-house written program, 'Q (9-04)', programmed in Matlab 6.5 (Mathworks, Natick, MA, USA).

2.5. Multivariate regressions.

Multiple linear regression (MLR), principal component regression (PCR) and partial least squares regression (PLSR) were performed in Unscrambler 7.5 (CAMO, Oslo, Norway). Prior to PLSR and PCR, the *x*-variables were standardized (each variable was divided by the standard deviation) and centred (the mean value was subtracted). The rank of the *x*-matrix, the number of linearly independent factors, was evaluated by the Matlab 'rank' command, and was always equal to the number of predictors.

Root mean squared error of prediction (RMSEP), the square root of the average squared residual, was used as error estimate and calculated according to the formula below:

$$\mathrm{RMSEP} = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(y_{\mathrm{p},i} - y_{\mathrm{t},i})^2} \quad (2)$$

I is the number of objects (fatty acids) in the dataset, y_p and y_t are predicted and experimental values of the response variable. Since there is no correction for bias in the RMSEP formula, RMSEP is an error estimate that includes both accuracy and precision, i.e. both systematic deviations and random errors increase RMSEP. For simplicity, the term 'accuracy' is applied in this work.

RMSEP was calculated either on test set residuals, where none of the objects in the test set were present in the calibration set or on full cross-validation residuals. In full cross-validation, the prediction sample (i) is left out of the dataset and a calibration model is made from the remaining objects. This model is used to predict $y_{p,i}$, and the residual, $y_{p,i} - y_{t,i}$, is calculated. The process is repeated for all objects. All objects in the dataset were unique, i.e. each fatty acid is only represented once.

3. Results and discussion

3.1. Data and structure

The analysed PUFAs are listed in Table 1 together with equivalent chain lengths achieved with the three programs. The double bond system is positioned in four different *n*-positions, *n*-1, *n*-3, *n*-4 and *n*-6 and there exist several homologous series in the dataset. It is worth noting that the FCL values increase with increasing chain lengths, especially for the most unsaturated series. This increase is caused by the temperature effect on cyanopropyl phases, as explained in Section 1. The ECLs also show a marked increase from Program 1 to 3, especially for the most unsaturated compounds. The increase in ECLs is caused by increased temperature gradients and reduced column flow from A to C, which make the fatty acids elute at higher temperatures. Further details can be found elsewhere [7,8,25].

3.2. Molecular descriptors

The object of this work was to evaluate the merits of simple molecular descriptors for prediction of ECL and FCL by multivariate regression methods. The applied descriptors, designated

Table 1
Equivalent chain lengths (ECL) values for polyunsaturated fatty acids applied in the study

Compound No.	Fatty acid	ECL-values			n[a]	In GLC-461
		Prg. 1	Prg. 2	Prg. 3		
1	16:3 *n*-3	17.832	17.949	18.075	3	
2	16:3 *n*-4	17.741	17.860	17.978	3	
3	16:4 *n*-1	18.419	18.571	18.735	7	
4	16:4 *n*-3	18.092	18.237	18.391	4	
5	18:2 *n*-6	19.061	19.151	19.244	3	x
6	18:3 *n*-3	19.864	19.987	20.116	7	x
7	18:3 *n*-4	19.754	19.881	20.019	3	
8	18:3 *n*-6	19.498	19.623	19.751	6	x
9	18:4 *n*-1	20.439	20.600	20.772	7	
10	18:4 *n*-3	20.314	20.476	20.655	7	
11	18:5 *n*-1	20.761	20.954	21.163	2	
12	19:2 *n*-6	20.067	20.163	20.253	1	
13	20:2 *n*-6	21.076	21.169	21.263	3	x
14	20:3 *n*-3	21.881	22.006	22.135	7	x
15	20:3 *n*-6	21.509	21.644	21.780	7	x
16	20:4 *n*-1	22.454	22.623	22.793	3	
17	20:4 *n*-3	22.331	22.501	22.680	7	
18	20:4 *n*-6	21.809	21.979	22.152	8	x
19	20:5 *n*-3	22.654	22.853	23.055	7	x
20	21:4 *n*-3	23.341	23.510	23.691	1	
21	21:5 *n*-3	23.794	23.991	24.206	2	
22	22:2 *n*-6	23.089	23.182	23.274	3	x
23	22:3 *n*-3	23.910	24.036	24.162	3	
24	22:4 *n*-3	24.354	24.526	24.710	5	
25	22:4 *n*-6	23.934	24.107	24.285	5	x
26	22:5 *n*-3	24.794	25.000	25.209	7	x
27	22:5 *n*-6	24.185	24.376	24.583	4	
28	22:6 *n*-3	25.060	25.286	25.515	4	x
29	24:4 *n*-3	26.383	26.562	26.739	1	
30	24:5 *n*-3	26.833	27.048	27.261	2	
	Precision[b]	0.007	0.010	0.016		

Experimental conditions for the three GC-programs (Prg. 1, 2 and 3) are described in Section 2.2.
[a] Number of each compound analysed. ECL-values in the table are the median values of the analysed compounds.
[b] Estimated precision for single peaks; pooled standard deviations for compounds with $n > 4$. The ECL-values given in the table are median values of several peaks (except compound no. 12, 20 and 29) and can therefore be expected to be more precise than this estimate.

A–L, are given in Table 2. The background for the selection of these variables is briefly summarised below.

Normal methylene-interrupted polyunsaturated fatty acids vary only in the length of the carbon chain, the number of double bonds and the position of the double bond system, usually described by the *n*-position or Δ-position. Even though MI PUFA molecules may be completely described by only three variables, these may not be suitable as predictors if applied

Table 2
Molecular descriptors applied in ECL/FLC regressions

Variable	Description	Note	Examples			
			18:4 *n*-1	22:6 *n*-3	16:3 *n*-4	20:4 *n*-6
A	*n*. Carbons		18	22	16	20
B	*n*. Double bonds		4	6	3	4
C	Δ-Position		8	4	6	5
D	Δ-Position^2	C^2	64	16	36	25
E	Δ-Position^3	C^3	512	64	216	125
F	Δ-Position^4	C^4	4096	256	1296	625
G	*n*-Position		1	3	4	6
H	*n*-Position^2	G^2	1	9	16	36
I	*n*-Position^3	G^3	1	27	64	216
J	*n*-1	Category	1	0	0	0
K	*n*-3	Category	0	1	0	0
L	*n*-4	Category	0	0	1	0

Table 3
Cross-validation RMSEP for models predicting retention indices for all PUFA in the dataset, showing variations in accuracy and prediction for different combinations of the fatty acid molecular descriptors given in Table 2

Model	Position descriptors	Vars. incl.	FCL				ECL				Mean[d]
			a Prg.1	b Prg.2	c Prg.3	NC[a]	d Prg.1	e Prg.2	f Prg.3	NC[a]	
Models with no Δ position											
M1	no *n*-position	AB	0.209	0.211	0.213	2	0.212	0.211	0.213	2	0.212
M2	*n*-pos: categories	AB JKL	0.060[b]	0.064	0.068	5	0.060	0.064	0.068	5	0.064
M3	*n*-pos: n	AB G	0.072	0.075	0.078	3	0.072	0.075	0.078	3	0.075
M4	*n*-pos: $n+n^2$	AB GH	0.059	0.063	0.067[c]	3	0.057	0.062	0.066	3	0.062
M5	*n*-pos: $n+n^2+n^3$	AB GHI	0.057[c]	0.061	0.067	4	0.059	0.063	0.067	4	0.062
Models with Δ-position: Δ											
M6	no *n*-position	ABC	0.072	0.075	0.078[c]	3	0.072	0.075	0.078	3	0.075
M7	*n*-pos: categories	ABC JKL	0.060	0.064[c]	0.068	6	0.057[b]	0.063	0.067	5	0.063
M8	*n*-pos: n	ABC G	0.072[b]	0.075	0.077	2	0.072	0.075	0.079	3	0.075
M8	*n*-pos: $n+n^2$	ABC GH	0.056	0.059	0.064	2	0.055[b]	0.063	0.067	4	0.061
M10	*n*-pos: $n+n^2+n^3$	ABC GHI	0.059	0.063	0.067[b]	3	0.058	0.063	0.067	4	0.063
Models with Δ-position: $\Delta+\Delta^2$											
M11	no *n*-position	ABCD	0.059	0.059	0.059	4	0.059	0.059	0.059	4	0.059
M12	*n*-pos: categories	ABCD JKL	0.038	0.036	0.041	7	0.039	0.039	0.038	7	0.039
M13	*n*-pos: n	ABCD G	0.059	0.058[b]	0.059	4	0.059	0.058[b]	0.059	4	0.059
M14	*n*-pos: $n+n^2$	ABCD GH	0.035	0.034	0.039	6	0.037	0.040[c]	0.034	6	0.037
M15	*n*-pos: $n+n^2+n^3$	ABCD GHI	0.038	0.035	0.039	7	0.039	0.037	0.037	7	0.038
Models with Δ-position: $\Delta+\Delta^2+\Delta^3$											
M16	no *n*-position	ABCDE	0.054	0.054	0.053	5	0.054	0.054	0.053	5	0.054
M17	*n*-pos: categories	ABCDE JKL	0.029	0.030	0.028[c]	8	0.030	0.029	0.026[b]	7	0.029
M18	*n*-pos: n	ABCDE G	0.054	0.054	0.053	5	0.054	0.054	0.053	5	0.054
M19	*n*-pos: $n+n^2$	ABCDE GH	0.028	0.028	0.027	6	0.028	0.027[b]	0.027	6	0.028
M20	*n*-pos: $n+n^2+n^3$	ABCDE GHI	0.029	0.029	0.028	6	0.029	0.029	0.028	6	0.029
Models with Δ-position: $\Delta+\Delta^2+\Delta^3+\Delta^4$											
M21	no *n*-position	ABCDEF	0.055	0.054	0.053	5	0.055	0.054	0.053	5	0.054
M22	*n*-pos: categories	ABCDEF JKL	0.027	0.028	0.026[b]	8	0.028	0.024	0.024[b]	8	0.026
M23	*n*-pos: n	ABCDEF G	0.055	0.054	0.053	5	0.055	0.054	0.053	5	0.054
M24	*n*-pos: $n+n^2$	ABCDEF GH	0.027	0.026	0.026[b]	6	0.028	0.027	0.025[b]	7	0.027
M25	*n*-pos: $n+n^2+n^3$	ABCDEF GHI	0.029	0.027	0.026[b]	6	0.029[c]	0.028	0.026	8	0.028

The dependent variables in the models are FCL (a–c) or ECL-values (d–f) from three different temperature/pressure programs described in Section 2.
[a] Number of PLS-components, median of the three models (a–c or d–f).
[b] Number of PLS-components in one (two for models 24c and 25c) higher than the median for the three models.
[c] Number of PLS-components in one (two for models 7b and 25d) lower than the median for the three models.
[d] Mean RMSEP of models a–f.

directly in the models. Linear models like MLR, PCR and PLSR performs poorly when there is non-linear dependence between the predictors and the response, or when the response depends on complex interactions between several predictors, i.e. the models cannot be expected to handle interactions between the carbonyl group and the double bond system if the positions of the double bonds are given only as the *n*-positions.

It is well known that there is no linear dependence between the positions of the double bonds and the effect on ECL/FCL values. At least for monoenes, shifts in positions have nearly no effect near the centre of the carbon chain, while the effects increase substantially as the double bonds approaches either of the ends. A common trick to handle such non-linearities by linear methods is to include higher order terms as variables [23]. For this reason, higher order terms of both Δ- and *n*-positions were included as separate variables, variable C–F and G–I, respectively.

While the PUFAs in the dataset have Δ-positions covering every number from 4 to 13, the objects may be divided into four classes (*n*-1, *n*-3, *n*-4 and *n*-6) based on the *n*-positions. Since there are only four classes, there is no point in using higher terms than the cubic function of the *n*-position. An alternative is to use category variables for *n*-1, *n*-3, *n*-4 and *n*-6, where the variable is one for fatty acids belonging to the class or zero otherwise. Since all objects in the dataset belong to one of the classes, three category variables are sufficient to describe the class memberships. No category variable was defined for *n*-6 and the variables J–L, therefore, describe the effect of the double bond system in *n*-1, *n*-3 and *n*-4 relative to *n*-6.

3.3. Regression models on all objects

The merits of different regression models are given in Tables 3–6. An initial study showed that MLR was less accurate than PCR and PLS for models including higher order terms of *n*- and Δ-positions. There where no difference in RMSEP between PCR and PLSR; all results given are therefore based solely on

Table 4
Cross-validation RMSEP for models predicting retention indices for *n*-3 and *n*-6 PUFA

Model	Position descriptors	Vars. incl.	FCL a Prg.1	FCL b Prg.2	FCL c Prg.3	FCL NC[a]	ECL d Prg.1	ECL e Prg.2	ECL f Prg.3	ECL NC[a]	Mean[d]
Models including *n*-position											
M26	no Δ-position	AB G	0.064[c]	0.069	0.075	3	0.064	0.069	0.075	3	0.069
M27	Δ-pos: Δ	ABC G	0.064[b]	0.068	0.073[c]	3	0.063	0.068[c]	0.073	4	0.068
M28	Δ-pos: $\Delta + \Delta^2$	ABCD G	0.039	0.039[b]	0.041	4	0.039[c]	0.038	0.040	5	0.039
M29	Δ-pos: $\Delta + \Delta^2 + \Delta^3$	ABCDE G	0.029[c]	0.029	0.028	6	0.029	0.029[c]	0.028	6	0.029
M30	Δ-pos: $\Delta + \Delta^2 + \Delta^3 + \Delta^4$	ABGDEF G	0.028[c]	0.027[b]	0.028	6	0.029[b]	0.028	0.027	6	0.028
Models without *n*-position											
M31	no Δ-position	AB	0.216	0.217	0.219	2	0.216	0.217	0.220	2	0.218
M32	Δ-pos: Δ	ABC	0.064	0.069[b]	0.075	3	0.064[b]	0.069	0.075	2	0.069
M33	Δ-pos: $\Delta + \Delta^2$	ABCD	0.039	0.040	0.041	4	0.039	0.040	0.041	4	0.040
M34	Δ-pos: $\Delta + \Delta^2 + \Delta^3$	ABCDE	0.029	0.029	0.030	5	0.029	0.029	0.030	5	0.029
M35	Δ-pos: $\Delta + \Delta^2 + \Delta^3 + \Delta^4$	ABGDEF	0.029	0.028	0.028	6	0.029[c]	0.028	0.028	6	0.028

See Table 3 for additional information.
[a] Number of PLS-components, median of the three models (a–c or d–f).
[b] Number of PLS-components in one higher than the median for the three models.
[c] Number of PLS-components in one lower than the median for the three models.
[d] Mean RMSEP of models a–f.

Table 5
Cross-validation RMSEP for models predicting retention indices for PUFA in the GLC-461 reference mixture

Model	Position descriptors	Vars. incl.	FCL a Prg.1	FCL b Prg.2	FCL c Prg.3	FCL NC[a]	ECL d Prg.1	ECL e Prg.2	ECL f Prg.3	ECL NC[a]	Mean[c]
Models including *n*-position											
M36	no Δ-position	AB G	0.060	0.063	0.068	3	0.060	0.063	0.068	3	0.064
M37	Δ-pos: Δ	ABC G	0.060	0.063	0.068	3	0.060	0.063	0.068	3	0.064
M38	Δ-pos: $\Delta + \Delta^2$	ABCD G	0.047	0.048	0.049	3	0.051	0.051	0.052	4	0.050
M39	Δ-pos: $\Delta + \Delta^2 + \Delta^3$	ABCDE G	0.042	0.042	0.041[b]	3	0.048	0.044	0.041	5	0.043
M40	Δ-pos: $\Delta + \Delta^2 + \Delta^3 + \Delta^4$	ABGDEF G	0.045	0.045	0.045	3	0.048	0.047	0.048	5	0.046
Models without *n*-position											
M41	no Δ-position	AB	0.199	0.198	0.197	2	0.199	0.198	0.197	2	0.198
M42	Δ-pos: Δ	ABC	0.060	0.063	0.068	3	0.060	0.063	0.068	3	0.064
M43	Δ-pos: $\Delta + \Delta^2$	ABCD	0.051	0.051	0.052	4	0.051	0.051	0.052	4	0.051
M44	Δ-pos: $\Delta + \Delta^2 + \Delta^3$	ABCDE	0.048	0.044	0.041	5	0.048	0.044	0.041	5	0.044
M45	Δ-pos: $\Delta + \Delta^2 + \Delta^3 + \Delta^4$	ABGDEF	0.048	0.047	0.048	5	0.048	0.047	0.048	5	0.048

See Table 3 for additional information.
[a] Number of PLS-components, median of the three models (a–c or d–f).
[b] Number of PLS-components in two higher than the median for the three models.
[c] Mean RMSEP of models a–f.

PLSR. In PLSR, the optimal numbers of PLS-components in the models must be determined by the validation procedures. The number of components was varied from zero to the rank of the *x*-matrix and the lowest number of PLS-components that gave the minimum RMSEP value was applied. Differences in RMSEP below 0.001 ECL units were regarded as negligible. The numbers of applied PLS-components in the models are given in Tables 3–6.

Since all models include information about chain length and the number of double bonds, the differences between the models are how information about the double bond system is given. The models listed in Table 3 are grouped after the order of variables describing the Δ-position. Within each group, there are five combinations of variables describing the *n*-position. For each combination of variables, regression models with both ECLs and FCLs as response variable were calculated for all three GC-programs.

Although the different GC-programs gave relatively large differences in ECL-values (Table 1), the RMSEP for the models are approximately identical for the three programs. It can also be seen that the choice of ECLs or FCLs as response variable has no influence on the results. The following discussion will therefore be focused on the choice of *x*-variables.

Models 1–5 (M1–5) are models where no information about the Δ-position has been included. The best models (M4 and M5) include quadratic and cubic terms of *n*-position with

Table 6
RMSEP for prediction of retention indices for *n*-3 and *n*-6 PUFA not present in the GLC-461 reference mixture

Model	Position descriptors	Vars. incl.	FCL				ECL				Mean[d]
			a Prg.1	b Prg.2	c Prg.3	NC[a]	d Prg.1	e Prg.2	f Prg.3	NC[a]	
Models including *n*-position											
M36	no Δ-position	AB G	0.071	0.080	0.087	3	0.070	0.077	0.086	3	0.079
M37	Δ-pos: Δ	ABC G	0.070	0.077	0.085	2	0.063	0.068	0.075	2	0.073
M38	Δ-pos: $\Delta+\Delta^2$	ABCD G	0.041[b]	0.044	0.049	4	0.041	0.044	0.049	4	0.045
M39	Δ-pos: $\Delta+\Delta^2+\Delta^3$	ABCDE G	0.029	0.031	0.036	5	0.029	0.031	0.036	5	0.032
M40	Δ-pos: $\Delta+\Delta^2+\Delta^3+\Delta^4$	ABGDEF G	0.031[c]	0.032	0.034[b]	6	0.031[c]	0.032	0.035	6	0.033
Models without *n*-position											
M41	no Δ-position	AB	0.250	0.253	0.259	2	0.250	0.219	0.259	2	0.248
M42	Δ-pos: Δ	ABC	0.070	0.077	0.085[c]	3	0.070	0.078	0.086	3	0.078
M43	Δ-pos: $\Delta+\Delta^2$	ABCD	0.042	0.044	0.048	4	0.042	0.044	0.049	4	0.045
M44	Δ-pos: $\Delta+\Delta^2+\Delta^3$	ABCDE	0.029	0.031	0.036	5	0.029	0.031	0.036	5	0.032
M45	Δ-pos: $\Delta+\Delta^2+\Delta^3+\Delta^4$	ABGDEF	0.031[c]	0.032	0.035	6	0.030[c]	0.032	0.035	6	0.033

Predictions are based on the models in Table 5.
[a] Number of PLS-components, median of the three models (a–c or d–f).
[b] Number of PLS-components in one higher than the median for the three models.
[c] Number of PLS-components in one lower than the median for the three models.
[d] Mean RMSEP of models a–f.

average RMSEP of 0.062. However, the models where *n*-positions are described as categories (M2) are nearly as good. The inclusion of the Δ-position described by variable C had no effect (M6–10), except for the models without *n*-position (M6). However, RMSEP falls as higher order terms of Δ are included and all models including Δ^3 and Δ^4 have RMSEP below 0.030 for models including n^2 and n^3 (M19–20 and M24–25), or for models where the *n*-position is described as category variables (M17 and M22). Including the cubic term of *n*-position has no positive effect (M15, M20 and M25), and the effect of including Δ^4 is also negligible (M21–25). Thus, it can be concluded that the models should include n, n^2, Δ, Δ^2, Δ^3 in addition to the chain length and the number of double bonds. Alternatively, *n*-position can be described by category variables. The effect of including higher order terms of the Δ-position corresponds well with a rather complex elution pattern of monoenes with double bonds between Δ8 and the carbonyl group [17,26].

The peak widths on the ECL scale are approximately equal for all peaks in the chromatograms and can be used to illustrate the merits of the models in practical situations. The prediction errors of M19a–c are compared to the peak widths in Fig. 2a–c. The peak width (in ECL units) increases from Program 1 to 3 because of poorer resolution caused by the steeper temperature gradients. The prediction errors are approximately equal for all three programs and range from −0.06 to +0.04 ECL units. Even though the curves for the normal distribution of the residuals are slightly wider than the chromatographic peaks, the majority of predicted ECL-values will appear where they are covered by the real chromatographic peaks.

Fig. 2 also shows that the objects with the largest residuals are the same for all programs. Even though the three programs show large differences in ECL-values, and also in elution order of the PUFAs, there were high correlations between the residuals for the models based on the different programs, $R^2 = 0.98$ for Prg. 1 and 2, $R^2 = 0.91$ for Prg. 1 and 3 and $R^2 = 0.95$ for Prg. 2 and 3. The corresponding R^2 values for M20 were 0.95, 0.73 and 0.77. The correlations between the residuals show that the major source of error is systematic. Even for the models with lowest RMSEP, there are still relationships between the fatty acid structure and the ECL-values that are not explained by the models. Inclusion of higher terms of the Δ-position did not improve the results; neither did inclusion of cross terms of the main variables (A × B, A × C, A × G, B × C and B × G) or squared terms of the chain length and number of double bonds (A^2 and B^2).

3.4. Regression models for n-3 and n-6 fatty acids

Since the majority of natural PUFAs and commercially available references are either *n*-3 or *n*-6, models for only these classes were evaluated. It was also tested whether the compounds in the reference mixture GLC-461 could be used as a basis for prediction of the ECLs of other *n*-3 and *n*-6 PUFAs found in the marine samples. Since there are only two *n*-positions in this dataset, variable G behaves as a category variable, and only the presence/absence of this variable was evaluated together with the various orders of the Δ-position.

The errors for the models including all *n*-3 and *n*-6 PUFA are summarised in Table 4. The results were similar to those seen in Table 3. The best models have RMSEP below 0.030 and include Δ^3; marginal improvements are achieved with inclusion of Δ^4. It is worth noting that models without *n*-position (M34 and M35) are as accurate as those including *n*-position (M29 and M30).

The models based on only the PUFAs in the reference mixture (Table 5) gave significantly higher RMSEP compared with the models based on all *n*-3 and *n*-6. There were only marginal improvements going from the models with Δ^2 (M38 and M43) to models with Δ^3 (M39 and M44). However, when these models were applied for the prediction of ECL/FCL of the remaining

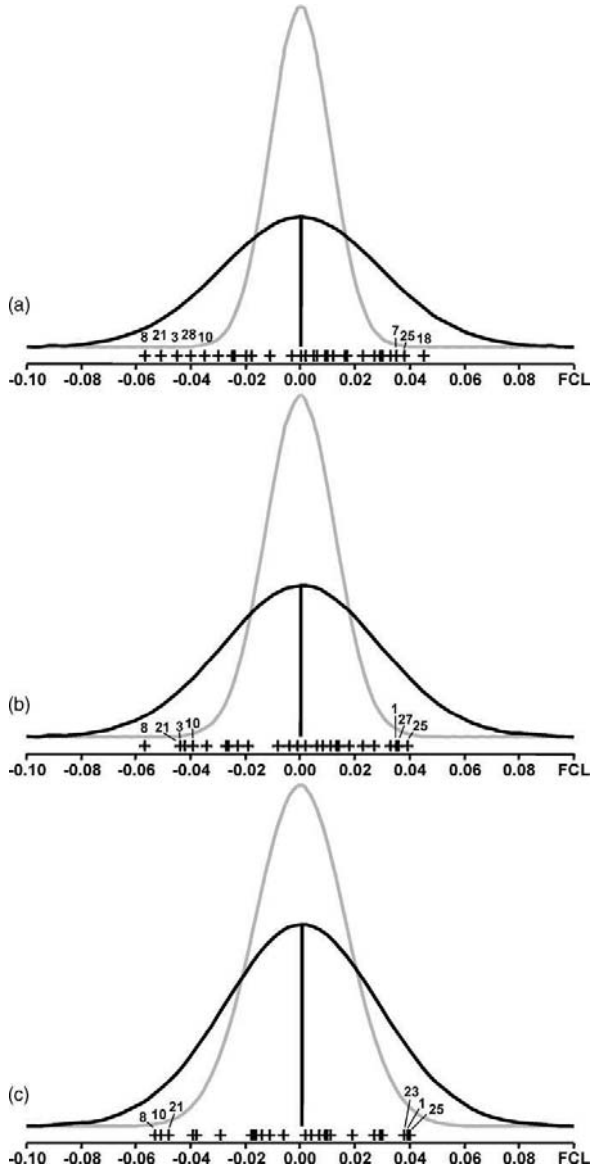

Fig. 2. Cross-validation residuals (+) for models 19a–c based on GC-program 1 (a), 2 (b) and 3 (c). The black curve is the normal distribution and mean of the residuals. The grey curve is the normal distribution representing the peak widths estimated from 20:5 *n*-3. The numbers on the largest residuals corresponds to the fatty acids listed in Table 1.

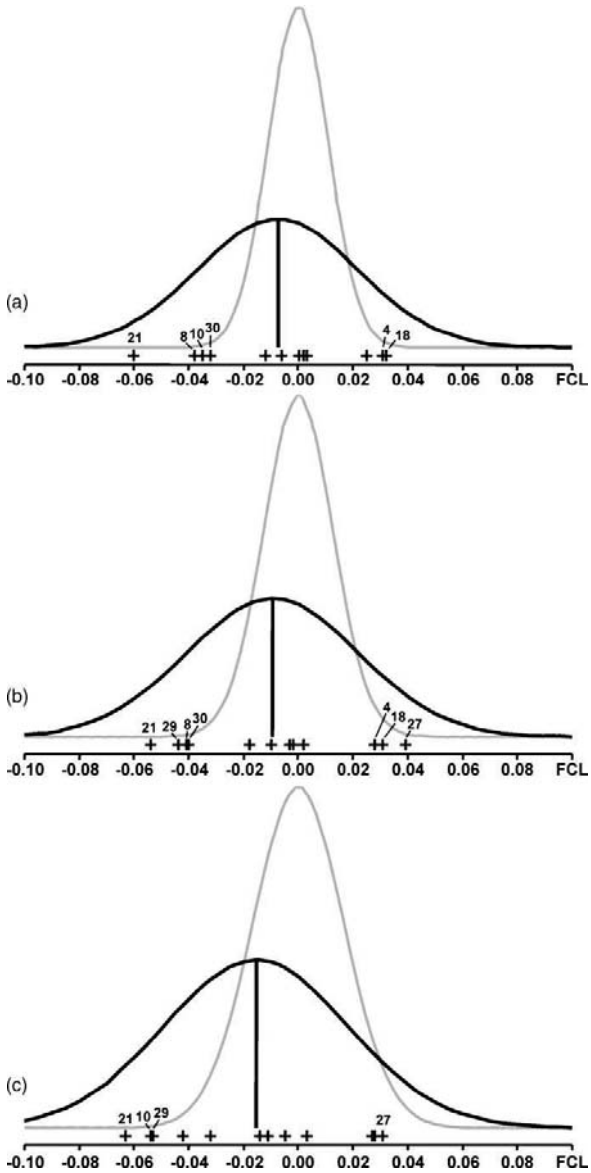

Fig. 3. Testset residuals (+) for models 39a–c based on GC-program 1 (a), 2 (b) and 3 (c). The black graph is the normal distribution curve and mean of the residuals. The grey curve is the normal distribution curve representing the peak widths estimated from 20:5 *n*-3. The numbers on the largest residuals corresponds to the fatty acids listed in Table 1.

n-3 and *n*-6 in the dataset (Table 6), RMSEP was significantly lower than the cross-validated RMSEP of the prediction set, and only marginally higher than those seen in Table 4. The predictions of the PUFAs in the test set are shown in Fig. 3. The predictions are more biased than in Fig. 2. Both the difference in RMSEP between the calibration set and the prediction set, and the increased bias may be explained by the low number of objects in the models, which give results that are more dependent on the behaviour of single objects.

3.5. General discussion

It has been shown that all ECLs given in Table 1 could be predicted with residuals lower than 0.06 ECL units (Fig. 2) and RMSEP for the models lower than 0.030. The alternative classical approach, using linear regressions between ECL-values of members of homologous series, can only be applied for series with three or more members; two compounds are necessary for the regression line used to predict the third. This method can only

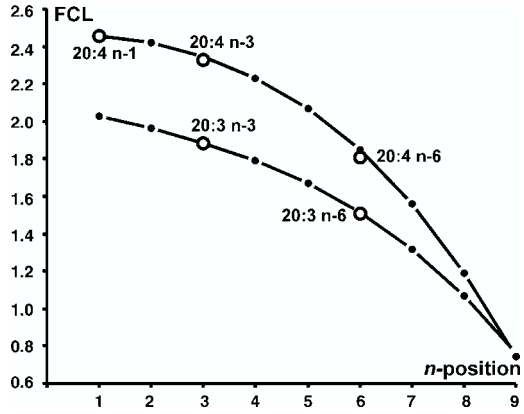

Fig. 4. Predicted fractional chain lengths (FCL) of 20:3 and 20:4 with *n*-positions from 1 to 9. Real values for *n*-1, *n*-3 and *n*-6 isomers are shown. The predictions are based on model 19a, including the variables ABCDEGH (Table 2).

be applied for some of the compounds in Table 1, the members of the series *x*:2 *n*-6, *x*:3 *n*-3, *x*:4 *n*-3 and *x*:5 *n*-3. When this method was applied on the ECL data from Prg. 1, RMSEP for the four series was 0.073. The *x*:2 *n*-6 and *x*:3 *n*-3 series were predicted with high accuracy, the largest deviations were 0.002 units for the *x*:2 *n*-6 and 0.007 units for *x*:3 *n*-3. However, the deviations for the x:4 *n*-3 and x:5 *n*-3 series were as large as 0.195 and 0.124 ECL units, respectively. The larger deviations may be explained by the low Δ-positions of 4 (16:4 *n*-3) and 5 (20:5 *n*-3) found in these series.

Accurate prediction of ECL and FCL values may be applied both for tentative identification of fatty acids and to foresee possible chromatographic overlaps. It should be emphasised that different fatty acids may have similar retention times, and matching retention indices is never a proof of the identity of a compound. However, prediction-models may be an efficient tool in excluding alternatives. With similar accuracy as in the most precise models, it is highly unlikely that a compound will appear more than 0.06 ECL-values from the predicted value. More refined statistical tests may be based on the error distribution of the residuals.

Another application is the prediction of possible chromatographic overlaps. There are a limited number of commercially available PUFA references, and suitable samples including all compounds of interest may not be available for the optimisation of temperature programs. A suitable model based on the compounds available may be used to predict if other compounds of interest can be hidden under larger peaks.

The models can only be expected to be valid for MI PUFA in the range represented by the fatty acids in Table 1. Thus, accurate predictions for PUFA with other *n*-positions than *n*-1, *n*-3, *n*-4 and *n*-6 cannot be expected. Predicted FCLs (M19a) of hypothetical 20:3 and 20:4 with *n*-positions from 1 to 9 are shown in Fig. 4. The FCLs for *n*-2 monoenes [26] and dienes [27] are known to be remarkably larger than FCLs of the corresponding *n*-3 and *n*-1 isomers. Although it is a rough approximation, summation of the values given ref. [26] indicates that FCLs for MI *n*-2 PUFA should be significantly higher than FCLs for *n*-1 and *n*-3 isomers. Fig. 4 shows that predicted values for 20:4 *n*-2 and 20:3 *n*-2 falls between the corresponding *n*-1 and *n*-3 isomers. Thus, it can be concluded that the models will not accurately predict *n*-2 PUFA. The *n*-5 isomers should be expected to appear roughly midway between *n*-4 and *n*-6 isomers, which are approximately where they is predicted. The effect of double bond positions usually levels off with increasing *n*-positions. The models in Fig. 4 show the opposite trends and accurate predictions cannot be expected for double bond position higher than *n*-6. If fatty acids with other *n*-positions, e.g. the *n*-9 fatty acids (not common in marine lipids), are included in similar models, higher order terms of the *n*-position (e.g. n^3 and n^4) can be expected to be significant. There also exist MI PUFAs that are not covered by the models because the Δ-position is lower than four. These are rare, but 18:5 *n*-3 is occasionally reported [28,29].

4. Conclusions

ECL and FCL values of methylene-interrupted PUFA could be predicted from the molecular structures by partial least squares regression models. The distance between the double bond system and the carbonyl group (Δ-position) should be included in the models and the highest accuracy was found for models including Δ^2, Δ^3 and Δ^4 among the variables. For models including *n*-1, *n*-3, *n*-4 and *n*-6 PUFA, it was necessary to include the first and second order terms of the distance between the double bonds and the methyl end of the carbon chain, *n* and n^2. It was not necessary to include the *n*-position in models restricted to *n*-3 and *n*-6 PUFA.

The highest residuals for the most accurate models were below 0.06 ECL units, and RMSEP was below 0.030. Correlation among the residuals of different models indicated that there is still systematic variance that is not explained by the most accurate models.

The multivariate regression on molecular descriptors can be applied to a wider range of compounds than the traditional approaches dependent on the presence of homologous series. For compounds with the double bond system close to the carbonyl group, the multivariate regression also gave higher accuracy than regression based on homologues.

References

[1] E. Kováts, Helv. Chim. Acta 41 (1958) 1915.
[2] G. Castello, J. Chromatogr. A 842 (1999) 51.
[3] T.K. Miwa, K.L. Mikolajczak, F.R. Earle, I.A. Wolff, Anal. Chem. 32 (1960) 1739.
[4] F.P. Woodford, C.M. van Gent, J. Lipid Res. 1 (1960) 188.
[5] H. van den Dool, P.D. Kratz, J. Chromatogr. 11 (1963) 463.
[6] F.T. Gillan, J. Chromatogr. 21 (1983) 293.
[7] S.A. Mjøs, J. Chromatogr. A 1015 (2003) 151.
[8] S.A. Mjøs, J. Chromatogr. A 1061 (2004) 201.
[9] J.A. Barve, F.D. Gunstone, F.R. Jacobsberg, P. Winlow, Chem. Phys. Lipids 8 (1972) 117.
[10] R.G. Ackman, S.N. Hooper, J. Chromatogr. 86 (1973) 73.
[11] R.G. Ackman, A. Manzer, T. Joseph, Chromatographia 7 (1974) 107.
[12] J.-L. Sebedio, R.G. Ackman, J. Chromatogr. Sci. 20 (1982) 231.

[13] A.T. James, J. Chromatogr. 2 (1959) 552.
[14] R.G. Ackman, J. Gas Chromatogr. 1 (1963) 11.
[15] G. Castello, S. Vezzani, G. D'Amato, J. Chromatogr. A 779 (1997) 275.
[16] J. Krupcik, P. Bohov, J. Chromatogr. 346 (1985) 33.
[17] R.H. Thompson, J. Chromatogr. Sci. 35 (1997) 536.
[18] Z. Garkani-Nejad, M. Karlovits, W. Demuth, T. Stimpfl, W. Vycudilik, M. Jalali-Heravi, K. Varmuza, J. Chromatogr. A 1028 (2004) 287.
[19] X. Liang, W. Wang, W. Wu, K.W. Schramm, B. Henkelmann, A. Kettrup, Chemosphere 41 (2000) 923.
[20] S. Rayne, M.G. Ikomou, J. Chromatogr. A 1016 (2003) 235.
[21] C. Yin, W. Liu, Z. Li, Z. Pan, T. Lin, M. Zhang, J. Sep. Sci. 24 (2001) 213.
[22] E.N. Malinowski, Factor Analysis in Chemistry, third ed., Wiley, New York, 2002.
[23] H. Martens, T. Næs, Multivariate Calibration, Wiley, New York, 1991.
[24] S.A. Mjøs, Eur. J. Lipid Sci. Technol. 106 (2004) 550.
[25] S.A. Mjøs, J. Chromatogr. A 1100 (2005) 185.
[26] C.D. Bannon, J.D. Craske, L.M. Norman, J. Chromatogr. 447 (1988) 43.
[27] W.W. Christie, J. Chromatogr. 37 (1968) 27.
[28] G.E. Napolitano, W.M.N. Ratanayake, R.G. Ackman, Phytochemistry 27 (1988) 1751.
[29] W.W. Christie, E.Y. Brechany, K. Stefanov, Chem. Phys. Lipids 46 (1988) 127.