3444 J. Agric. Food Chem. 2009, 57, 3444–3451
DOI:10.1021/jf8039268

JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY
ARTICLE

^{13}C NMR Pattern Recognition Techniques for the Classification of Atlantic Salmon (*Salmo salar* L.) According to Their Wild, Farmed, and Geographical Origin

Marit Aursand,*[,†,§] Inger B. Standal,[†,§] Angelika Praël,[#] Lesley McEvoy,[#] Joe Irvine,[#] and David E. Axelson[†,⊥]

[†]SINTEF Fisheries and Aquaculture, N-7465 Trondheim, Norway, [§]Department of Biotechnology, Norwegian University of Natural Science and Technology, N-7491 Trondheim, Norway, [#]North Atlantic Fisheries College, Port Arthur, Scalloway, Shetland ZE1 0UN, United Kingdom, and [⊥]MRi_Consulting, 8 Wilmot Street, Kingston, Ontario, Canada K7L 4V1

^{13}C nuclear magnetic resonance (NMR) in combination with multivariate data analysis was used to (1) discriminate between farmed and wild Atlantic salmon (*Salmo salar* L.), (2) discriminate between different geographical origins, and (3) verify the origin of market samples. Muscle lipids from 195 Atlantic salmon of known origin (wild and farmed salmon from Norway, Scotland, Canada, Iceland, Ireland, the Faroes, and Tasmania) in addition to market samples were analyzed by ^{13}C NMR spectroscopy and multivariate analysis. Both probabilistic neural networks (PNN) and support vector machines (SVM) provided excellent discrimination (98.5 and 100.0%, respectively) between wild and farmed salmon. Discrimination with respect to geographical origin was somewhat more difficult, with correct classification rates ranging from 82.2 to 99.3% by PNN and SVM, respectively. In the analysis of market samples, five fish labeled and purchased as wild salmon were classified as farmed salmon (indicating mislabeling), and there were also some discrepancies between the classification and the product declaration with regard to geographical origin.

KEYWORDS: Atlantic salmon (*Salmo salar* L.) lipid extracts; fish muscle; ^{13}C NMR spectroscopy; authentication; multivariate data analysis; wild; farmed; SVM; PNN

INTRODUCTION

In 1970, only about 4% of the world's seafood came from fish farms. Today, aquaculture accounts for 32% (*1*). World production of salmon was 1.24 million tons in 2006. Production in Europe constituted 48% of the global production of salmon; Norway produced 78% of the total European production volume in 2006 (*2*). In 1985, 6% of all salmon consumed around the world was farmed. In 2000, the amount of farmed salmon consumed had risen to 58% (*3*). Progress in aquaculture techniques has led to year-round availability of farmed salmon and lower prices for the consumer. Once considered an expensive delicacy, salmon is now quite commonplace in North American, European, and Japanese diets.

In the European Union, the common organization of the markets in fishery and aquaculture products comes under Council Regulation (EC) 104/2000. In October 2002, Commission Regulation (EC) 2065/2001 was adopted that details the labeling, packaging, and traceability requirements for fishery and aquaculture products. According to this regulation, fish shall be labeled with specification of the commercial designation and scientific name, method of production, and the area in which it was caught. There is a clear trend in the international market to labeling products with information about composition and quality. This, together with the increasing production and consumption of fish products including salmon, both farmed and wild fish, has led to an increasing demand for standardized analytical methods efficient in the authentication of fish products. At present, few reliable methods exist for the unequivocal determination of the geographical origin, wild versus farmed specimens, ecological production, or the life history of the product, data that are necessary to confirm the traceability documentation of the products. Recently, relevant methods to study the production method of fish (wild/farmed) and geographical origin have been reviewed (*4*, *5*). Potential methods to identify the production method of fish (wild/farmed) include morphological analyses, individual tagging of fish, genetic analyses, carotenoid content (natural vs synthetic), and analysis of protein/enzyme profiles of some tissues (*4*).

In the lipids of fish muscle, about 20 fatty acids appear in relative amounts of $>1\%$, and different species of fish have characteristic fatty acid profiles (*6*). The variability in the composition of the tissue fatty acids of fish is very large. In each type of tissue the fatty acids are bound in different lipid classes, phospholipids, triacylglycerols, cholesterol esters, etc., all with different profiles.

*Corresponding author (telephone +47 48200158; fax +47 93270701; e-mail Marit.Aursand@sintef.no).

 Published on Web 4/6/2009

The composition of muscle lipids may be influenced by factors such as diet, age, maturity, condition, and reproductive cycle of the fish, in addition to water, temperature, and salinity (*7*). The fatty acid profiles examined as methyl esters by gas−liquid chromatography (GC) have been used as natural marks for stock identification analyzing heart, muscle, or brain tissue for different fish species (*8*). The fatty acid profile of triacylglycerols of depot fat in fish muscle is more influenced by the diet than membrane phospholipids. Because the composition of storage lipids in fish reflects the diet (*9*), farmed fish have a fatty acid profile that differs from wild fish (*4*). High-resolution (HR) nuclear magnetic resonance (NMR) spectroscopy is used increasingly to provide insight, both qualitatively and quantitatively, into the nature of lipid mixtures and offers the opportunity to study hetereogenous lipid mixtures (*10*−*21*). Consequently, when the hetereogeneous lipid mixture is studied, a ^{13}C NMR spectrum of lipid extracted from fish muscle contains information about the lipid classes (*10*, *15*, *18*, *19*), the fatty acid profile (*11*), phospholipids (*15*, *16*, *20*), the positional distribution of fatty acids in both triacylglycerides and phospholipids (*12*, *17*), and cholesterol/cholesteryl content (*16*). In previous studies, when lipid extracts from muscles of different fish species and origin were examined, the ^{13}C NMR profiles allowed discrimination between fish species and wild and farmed salmon (*18*, *21*). Stable isotope analysis, in combination with fatty acid analysis, has recently been applied to identify organic farmed salmon (*22*) and to detect Atlantic salmon from different sources (*23*, *24*).

For authentication purposes, both when using fatty acid profile examined with GC and when the total lipid composition (lipid classes, content of phospholipids, positional distribution of fatty acids in triacylglycerides/phospholipids content) was examined by HR ^{13}C NMR, multivariate treatment of data is necessary to distinguish among variations. Pattern recognition techniques have been frequently and successfully applied to a variety of applications related to food composition and authentication (*18*), and species differentiation has been reported by multivariate analysis of phospholipids from canned Atlantic tuna (*25*).

Since September 2001, a European consortium of five partners from France, Italy, the United Kingdom, and Norway has been working to develop a validated method to enable official laboratories to discriminate between wild and farmed salmon and geographical origin. The analytical methodologies involved in the project (COFAWS, Confirmation of the Origin of Farmed and Wild Salmon and Other Fish) include stable isotope analysis by SNIF-NMR (site-specific natural isotope fractionation studied by nuclear magnetic resonance spectroscopy) and IRMS (isotope ratio mass spectrometry) of the fish oil, water from the fish, and other parts of the fish; ^{1}H and ^{13}C NMR profiling; and determination of fatty acid content by GC. The results from the ^{1}H NMR studies showed that the ^{1}H NMR profiles of the muscle lipids in combination with multivariate data analyses allowed discrimination between wild and farmed salmon (*26*). The ^{1}H NMR spectrum of lipid extracted from the muscle of salmon gives information about the lipid classes, level of unsaturation, and molar fractions of specific fatty acids, such as total n-3 and 22:6n-3 fatty acids (*11*, *27*). Compared to ^{13}C spectra, the ^{1}H NMR spectra show small chemical shift dispersion and extensive multiplicity, which often result in several overlaps of signals. In addition, ^{1}H spectra may contain broad resonances from phospholipids, and the spectra lack information about the positional distribution of fatty acids and the total fatty acid profile. Results from isotopic analysis combined with fatty acid composition have also been published recently (*28*). The aim of this study was to test the possibility of using ^{13}C NMR in combination with multivariate data analysis as a validated method to enable discrimination between farmed and wild salmon and geographical origin and to verify the origin of market samples.

MATERIALS AND METHODS

Fish Samples. Wild Atlantic salmon (*Salmo salar* L.) (n = 52) were obtained from Norway, Scotland, Canada, Iceland, and Ireland. Farmed Atlantic salmon (n = 143) were obtained from two different Norwegian, Scottish, Irish, Faroes, and Canadian sea farms and also from farms in Iceland and Tasmania. Fish from feeding trials run at North Atlantic Fisheries College, Port Arthur, Scalloway, Shetland, U.K., are included in the data set. Market samples (n = 43) were collected from supermarkets in Italy, the United Kingdom, and Norway. Because not all market samples were labeled with both production method and geographical origin, two different subsets of market samples were used in the wild/farmed and geographical origin predictions. In total, 238 samples were analyzed, a sufficient number to establish feasibility of classification.

Lipid Extraction. Lipids were extracted from white muscle of the fish using a modified Bligh and Dyer procedure (*29*) as evaluated by Thomas et al. (*28*). Homogenized salmon muscle (400 g) was extracted in chloroform/methanol (1:2, 1200 mL). The homogenate was filtered, and the residue was rinsed with chloroform (400 mL). This wash was added to the original filtrate, to which KCl (0.88%, 400 mL) was added. The mixture was shaken for 1 min and then left to separate into phases. The lipid phase was obtained, and the chloroform was removed by evaporation.

NMR Parameters. Proton-decoupled ^{13}C NMR spectra were recorded on a Bruker Avance DRX500 (Bruker BioSpin GmbH, Rheinstetten, Germany) instrument at 125.75 MHz. Approximately 70 mg of the lipid extracts was transferred to 5 mm NMR tubes and diluted with 0.5 mL of deuteriated chloroform ($CDCl_3$, 99.8% purity, Isotec Inc., Matheson). The NMR experimental parameters were as follows: spectral width, 200.78 ppm; pulse angle, 30°; dwell time, 19.8 μs; acquisition time, 2.0 s; number of data points, 101006; recycle delay, 2.5 s; number of acquisitions, 512 (2048). The NMR spectra were obtained by using an autosampler. The 1D ^{13}C spectra were run in a semiquantitative manner, because quantitative measurements require a significantly longer experimental time. Prior to Fourier transformation, a line-broadening factor of 0.1 Hz was applied to minimize noise, but not at the expense of resolution among significant closely spaced resonances. The chemical shift scale is referred indirectly to tetramethylsilane (TMS) by the triplet of $CDCl_3$ at 77.00 ppm. Maximum peak height (except for the solvent peak) was set to 100 for each spectrum. Peak positions and intensities were obtained for resonances >1% of the maximum peak intensity within each spectrum. The resulting peak list was exported for manual alignment (necessary because of small variations in chemical shift between samples) and multivariate data analyses. The resulting data matrix consisted of 187 variables for the 238 samples investigated. The corresponding aligned chemical shift intensities or principal component scores were used as input for the multivariate data analysis.

Multivariate Data Analysis. Probabilistic neural networks (PNN) (*30*) and support vector machines (SVM) (*31*, *32*) were applied to provide quantitative, supervised classification.

PNN calculations were performed with AI Trilogy (Ward Systems Group Inc., Frederick, MD). A PNN has three layers: input, pattern, and summation (*30*). The input layer has as many elements as there are individual parameters needed to describe the samples to be classified. In the present case the input parameters were the selected peak intensities. Although the PNN classification initially used 187 spectral intensities, the number of inputs was systematically decreased via genetic analysis (*33*) and determination of the

relative importance of each variable, such that the top 10 shifts could be identified. The PNN operates by defining a probability density function (pdf) for each class based on the training set data and an optimized kernel width parameter, also optimized by a genetic analysis. Each pdf is estimated by placing a Gaussian-shaped kernel at the location of each pattern in the training set such that the pdf defines the boundaries for each data class, whereas the kernel width determines the amount of interpolation that occurs between adjacent kernels. The probability that a pattern vector will be classified as a member of a given output data class increases the closer it is to the center of the pdf for that class. When an input test vector is presented, the first layer computes distances from the input vector to the training input vectors and produces a vector the elements of which indicate how close the input is to a training input. The second layer sums these contributions for each class of inputs to produce as its net output a vector of probabilities. Finally, a complete transfer function on the output of the second layer picks the maximum of these probabilities. This is then presented as the predicted class; however, the actual classification probabilities may be examined. This approach also allows the user to set a minimum probability threshold below which the algorithm will not attempt a classification prediction. Instead, the sample in question would be presented as not classified. This situation occurs when a new sample presented for classification is too different from any of the samples used in the training set.

Both leave-one-out (LOO) cross-validation (CV) and the use of separate training and validation data sets were used to estimate the accuracy of the predictions. To evaluate the statistical significance of the predictions and to ensure that most samples are represented at some time in the validation data set, multiple runs of different randomly chosen subjects, split into training and validation sets, were performed. Model selection with PNN involved LOO CV of the training data set. However, the separate validation data set (not used in any way in the model development) was used to assess classification accuracy. Although the most important variables in the classification were determined, there are limitations in quantifying the relative importance of an input variable. When one is dealing with nonlinear models, the concept of the contribution of a variable is an imprecise concept, because the effect of a variable on a model depends heavily on the values of all other variables. This uncertainty exists in any analysis of complex collections of variables. Running the PNN calculation many times may frequently lead to many slightly different solutions with different combinations of variables, each of which may yield equally accurate classifications. The relative importance of variables only has meaning when compared to other values from the same network. This issue is also of relevance in the attempt to identify specific components or fatty acids that relate to classification accuracy. There may be numerous combinations of different chemical shifts that lead to equally accurate classification; one should be careful, therefore, to avoid overinterpretation of specific results. Here we report the most frequently observed significant variables.

SVM were developed for binary classification (*31*, *32*). SVM calculations were performed with Tiberius v6.02 (Tiberius Data Mining, Melbourne, Australia). In class separation by SVM, the optimal separating hyperplane between the two classes is searched for by maximizing the margin between the classes' closest points. Those training points lying on one of the hyperplanes and the removal of which would change the solution found are called support vectors, and the middle of the margin is the optimal separating hyperplane. An SVM classifier depends only on the support vectors, and the classifier function is not influenced by the whole data set, as may be the case for many neural network systems as well as partial least-squares analyses. For overlapping classes, data points on the "wrong" side of the discriminant margin are weighted down to reduce their influence. When a linear separator cannot be found, data points are projected (via kernel techniques involving Gaussian radial basis functions or polynomials) into a higher dimensional space where the data points effectively become linearly separable. SVMs have many favorable properties. They are

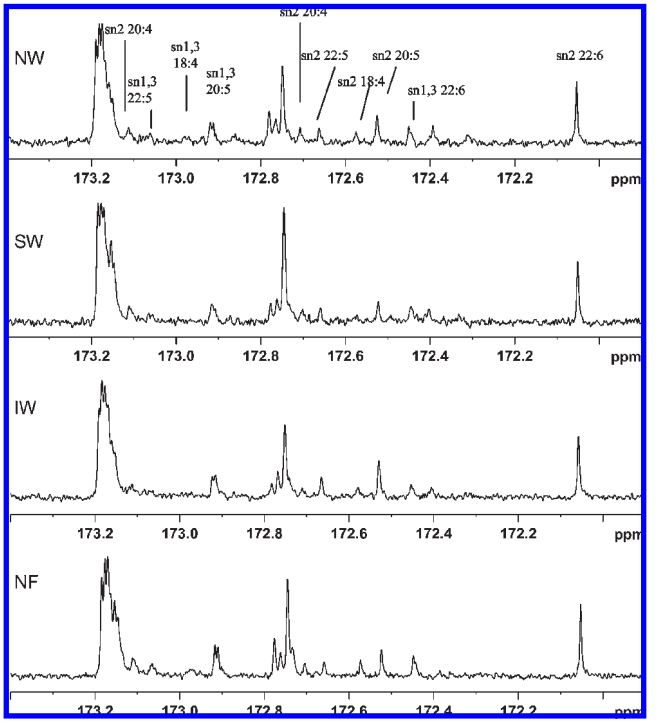

Figure 1. ^{13}C NMR carbonyl region (173.4−172.0 ppm) of lipids extracted from salmon muscle of four different origins: (from top) wild salmon from Norway (NW), Scotland (SW), and Ireland (IW) and farmed salmon from Norway (NF). The position of fatty acids in triacylglycerols is designated (sn1,3 or sn2).

robust against high dimensionalities (a large number of variables) and ill-behaved distributions and generally exhibit good performance without any feature selection. Most significantly, generalization ability is also robust. Whereas most learning techniques do not perform well on data sets where the number of features is large compared to the number of samples, SVMs are believed to be an exception. Traditional neural network (and partial least-squares) approaches are based on the empirical risk minimization (ERM) principle. ERM does not necessarily produce a good model that generalizes well to unseen data due to overfitting phenomena. The foundation of SVM embodies the structural risk minimization (SRM) principle, which has been shown to be superior to the ERM principle. SRM minimizes an upper bound on the expected risk, as opposed to ERM that minimizes the error on the training data. To guarantee an upper bound on generalization error, the capacity of the learned functions must be controlled, that is, by the Vapnik−Chervonenkis (VC) dimension. According to the SRM principle, a function that describes the training data well (minimizes the empirical risk) and belongs to a set of functions with lowest VC dimension will generalize well regardless of the dimensionality of the input space. It has been shown that maximizing the margin distance between the classes in the SVM method is equivalent to minimizing the VC dimension. Therefore, SVM embodies excellent generalization in its theory.

The largest peaks in MR spectra are not necessarily the most informative. In the absence of scaling, variation in these regions can dominate and obscure systematic variation of interest in low-intensity regions. Thus, variable stability scaling (VAST) (*34*) was used before further analysis. The stability parameter used is the ratio of the standard deviation and mean of each variable. VAST can be applied in a supervised manner, in that the coefficient of variation within each prior class can be calculated separately and then the mean of the class coefficients of variation used as the stability scale weight:

$$\text{supervised VAST scale weight} = \frac{1}{n}\sum_{j=1}^{n}\frac{\overline{x_j}}{\sigma_j}$$

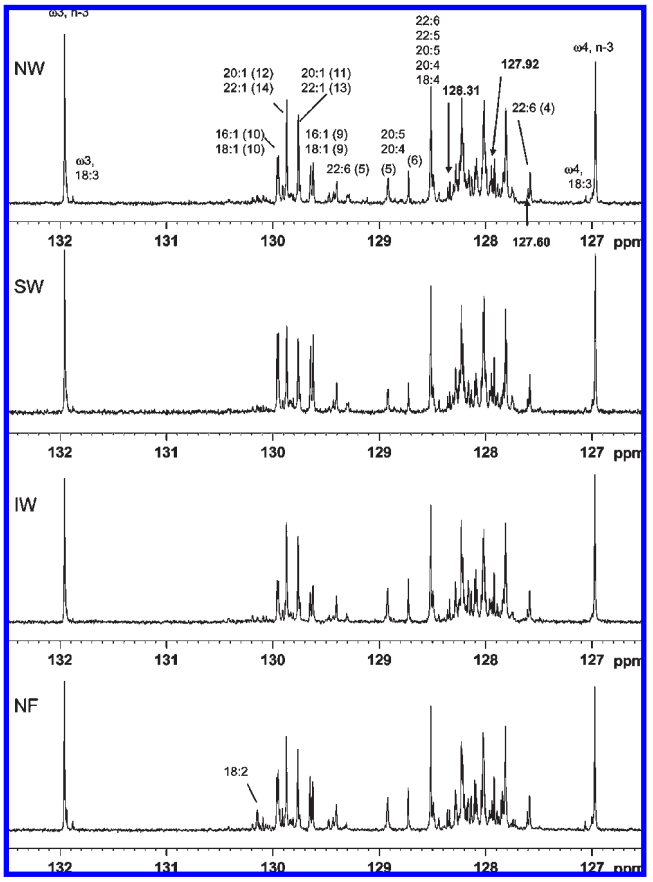

Figure 2. ^{13}C NMR olefinic region (132.5−126.5 ppm) of lipids extracted from salmon of four different origins: (from top) wild salmon from Norway (NW), Scotland (SW), and Ireland (IW) and farmed salmon from Norway (NF). Also, the three most significant chemical shifts in the PNN wild/farmed classification are highlighted. ω, carbon number from the methyl end; FA, fatty acid.

x_j and σ_j denote the mean and standard deviation of a variable x for the j^{th} class, respectively, and n is the total number of classes. This method down-weights variables that are the least stable, especially within each prior class of interest, and may improve distinction between the classes in subsequent multivariate analysis.

RESULTS AND DISCUSSION

Interpretation of ^{13}C NMR Spectra. ^{13}C NMR spectroscopy provides a fingerprint of the specific lipids extracted from salmon muscle. The interpretation of spectra of lipid extracted from muscle of farmed and wild salmon is given in **Figures 1−3** and is based on published data (*11*, *12*). In general, the chemical shift in each region depends on factors such as the type of glycerol ester (i.e., triacylglycerols, diacylglycerols, or monoacylglycerols), stereospecific conformation (fatty acids in sn1,3 or sn2 in acylglycerols), and for unsaturated fatty acids; the number and position of double bonds (*11*).

Carbonyl carbons in triacylglycerols appear in the region of 173.4−172.0 ppm. **Figure 1** shows the carbonyl region of lipids extracted from salmon of four different origins (wild salmon from Norway (NW), Scotland (SW), and Ireland (IW) and farmed salmon from Norway (NF)). Peaks from the main n-3 fatty acids in sn1,3 and sn2 positions of the triacylglycerols can be identified in this region. Even though some differences between the four spectra can be seen, it is difficult to conclude whether there are differences among groups on the basis of this spectral region.

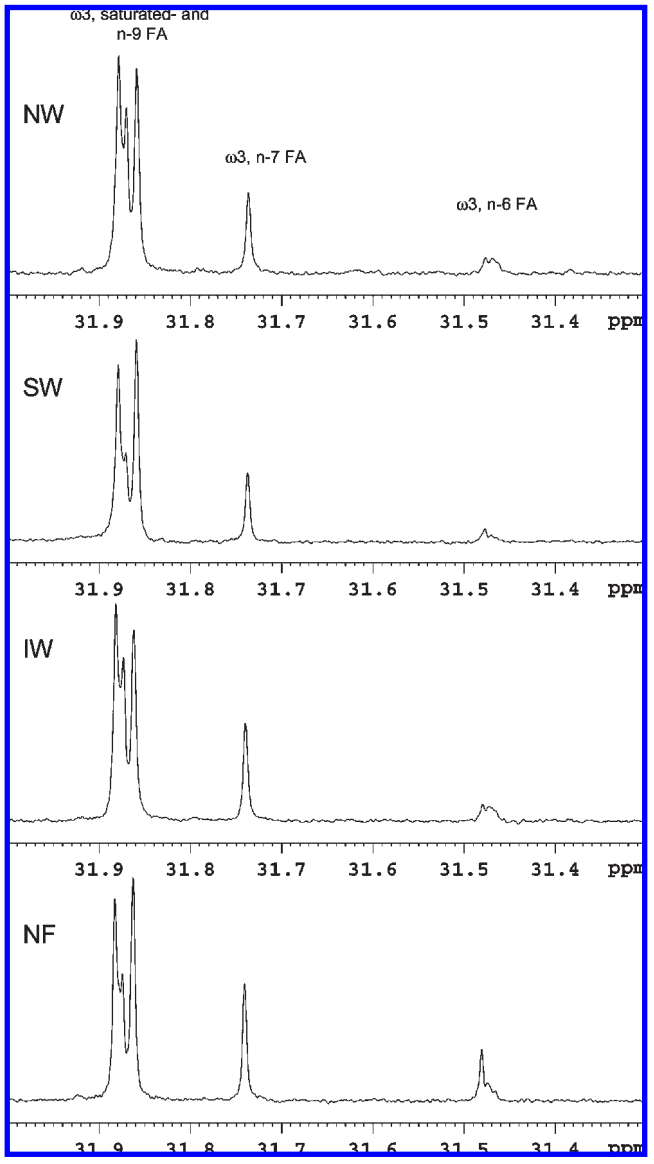

Figure 3. ^{13}C NMR aliphatic region (32.0−31.3 ppm) of lipids extracted from salmon of four different origins: (from top) wild salmon from Norway (NW), Scotland (SW), and Ireland (IW) and farmed salmon from Norway (NF). ω, carbon number from the methyl end; FA, fatty acid.

Olefinic carbon atoms appear in the ^{13}C NMR spectrum between 132.5 and 126.5 ppm. **Figure 2** shows this region of salmon lipids from the four selected origins (wild salmon from Norway (NW), Scotland (SW), and Ireland (IW) and farmed salmon from Norway (NF)). Information about unsaturated fatty acids can be obtained in this region. Examples of intensities that show relatively large variation among the samples are peaks assigned to 20:1 and 22:1 fatty acids (129.75 and 129.88 ppm), in addition to the peak arising from 18:2n-6 (at 130.2 ppm) in the spectrum of farmed fish. This fatty acid is abundant in vegetable oil, and the spectrum of the farmed fish (**Figure 2** NF) illustrates that the feed contained raw material of vegetable origin.

Aliphatic carbon atoms (ω3 carbon atoms) give rise to peaks in the region 32.0−31.3 ppm (**Figure 3**). Also in this region, it is clear that the farmed fish (**Figure 3**, NF) has a relatively high level of n-6 fatty acids, abundant in vegetable oils, compared to the wild fish.

However, although characteristic resonances may be observed for various classes or groupings of samples, such

Table 1. Country of Origin PNN Analysis[a]

	1	2	3	4	5	6	7	total
classified as 1	**43**	1	0	0	0	0	1	45
classified as 2	1	**43**	0	0	0	0	0	44
classified as 3	0	0	**5**	0	0	0	0	5
classified as 4	2	0	0	**19**	0	0	0	21
classified as 5	0	0	0	0	**18**	0	0	18
classified as 6	0	1	0	0	0	**5**	0	6
classified as 7	0	0	0	0	0	0	**4**	4
total	**46**	**45**	**5**	**19**	**18**	**5**	**5**	**143**[b]
true-positive ratio	0.934	0.955	1.0	1.0	1.0	1.0	0.8	
false-positive ratio	0.020	0.010	0.0	0.016	0.0	0.007	0.0	
true-negative ratio	0.979	0.989	1.0	0.983	1.0	0.992	1.0	
false-negative ratio	0.065	0.044	0.0	0.0	0.0	0.0	0.2	
sensitivity (%)	93.5	95.6	100	100	100	100	80	
specificity (%)	97.9	98.9	100	98.4	100	99.3	100	

[a] Country of origin: 1, Norway; 2, Scotland; 3, Iceland; 4, Ireland; 5, Canada; 6, Faroes; 7, Tasmania. All variables (146 × 143) (leave-one-out cross-validation). [b] **93.8%** correct (137), 4.1% incorrect (6), and 2.1% not classified (3).

Table 2. Country PNN

	1	2	3	4	5	6	7	total
				Train				
classified as 1	**26**	2	0	0	0	0	1	29
classified as 2	3	**30**	0	2	0	0	0	35
classified as 3	0	0	**4**	0	0	0	0	4
classified as 4	0	0	0	**10**	0	0	0	10
classified as 5	0	0	0	0	**13**	0	0	13
classified as 6	0	0	0	0	0	**3**	0	3
classified as 7	0	0	0	0	0	0	**2**	2
total	**29**	**32**	**4**	**12**	**13**	**3**	**3**	**96**[a]
true-positive ratio	0.896	0.937	1.0	0.833	1.0	1.0	0.666	
false-postive ratio	0.044	0.078	0.0	0.0	0.0	0.0	0.0	
true-negative ratio	0.955	0.921	1.0	1.0	1.0	1.0	1.0	
false-negative ratio	0.103	0.062	0.0	0.166	0.0	0.0	0.333	
sensitivity (%)	89.7	93.8	100	83.3	100	100	66.7	
specificity (%)	95.5	92.2	100	100	100	100	100	
				Validation (Every Third Validation)				
classified as 1	**13**	0	0	1	0	0	0	14
classified as 2	1	10	0	2	0	0	0	13
classified as 3	0	0	0	0	0	0	0	0
classified as 4	0	0	0	**4**	0	0	0	4
classified as 5	0	0	0	0	**6**	0	0	6
classified as 6	0	2	1	0	0	**2**	0	5
classified as 7	0	0	0	0	0	0	**2**	2
total	**14**	**12**	**1**	**7**	**6**	**2**	**2**	**44**[b]
true-positive ratio	0.928	0.833	0.0	0.571	1.0	1.0	1.0	
false-postive ratio	0.033	0.093	0.0	0.0	0.0	0.071	0.0	
true-negative ratio	0.966	0.906	1.0	1.0	1.0	0.928	1.0	
false-negative ratio	0.071	0.166	1.0	0.428	0.0	0.0	0.0	
sensitivity (%)	92.9	83.3	0.0	57.1	100	100	100	
specificity (%)	96.7	90.6	100	100	100	92.9	100	

[a] **87.1%** correct (88), 7.9% incorrect (8), and 5.0% not classified (5). [b] **82.2%** correct (37), 15.6% incorrect (7), and 2.2% not classified (1).

visual inspection is generally insufficient to unambiguously identify any given class without multivariate methods that consider all of the relevant peaks and combinations thereof to discriminate among classes.

Discrimination between Wild and Farmed Fish. Several different calculations were undertaken, involving predictions using all of the samples, as well as creating classification models from training and validation data sets of different numbers of samples chosen randomly. Both PNN and SVM analyses were implemented.

With the PNN analysis, and using all 187 chemical shifts, LOO CV analysis gave correct classifications for 47 of 52 wild salmon and for 136 of 143 farmed salmon.

Systematically reducing the number of chemical shifts to the most significant 12 shifts (128.31, 127.92, 127.60, 61.97, 29.43, 29.04, 29.00, 28.94, 27.15, 26.46, 22.64 ppm), resulted

Table 3. Market Samples: Wild Versus Farmed Predictions by SVM

sample	label	predicted	
1	farmed	farmed	
2	farmed	farmed	
3	farmed	farmed	
4	farmed	farmed	
5	farmed	farmed	
6	farmed	farmed	
7	farmed	farmed	
8	farmed	farmed	
9	farmed	farmed	
10	farmed	farmed	
11	farmed	farmed	
12	farmed	farmed	
13	farmed	farmed	
14	farmed	farmed	
15	farmed	farmed	
16	farmed	farmed	
17	farmed	farmed	
18	farmed	farmed	
19	farmed	farmed	
20	farmed	farmed	
21	farmed	farmed	
22	farmed	farmed	
23	farmed	farmed	
24	wild	farmed	***error***
25	wild	farmed	***error***
26	wild	farmed	***error***
27	wild	farmed	***error***
28	farmed	farmed	
29	farmed	farmed	
30	farmed	farmed	
31	farmed	farmed	
32	farmed	farmed	
33	farmed	farmed	
34	wild	farmed	***error***

Table 4. Market Samples: Country of Origin Predictions by SVM[a]

sample	predicted origin	labeled origin	
1	Norway	Norway	
2	Norway	Norway	
3	Norway	Norway	
4	Norway	Norway	
5	Norway	Norway	
6	Norway	Norway	
7	Norway	Norway	
8	Norway	Norway	
9	Norway	Norway	
10	Norway	Norway	
11	Norway	Norway	
12	Ireland	Norway	***error***
13	Norway	Norway	
14	Norway	Norway	
15	Faroes	Norway	***error***
16	Norway	Norway	
17	Scotland	Norway	***error***
18	Norway	Norway	
19	Ireland	Norway	***error***
20	Norway	Norway	
21	Ireland	Norway	***error***
22	Ireland	Norway	***error***
23	Norway	Norway	
24	Norway	Norway	
25	Norway	Norway	
26	Ireland	Norway	***error***
27	Norway	Norway	
28	Norway	Norway	
29	Scotland	Scotland	
30	Norway	Norway	
31	Norway	Norway	
32	Norway	France	***Error***
33	Scotland	Scotland	
34	Scotland	Scotland	
35	Scotland	Scotland	

[a] Sample number does not correspond to **Table 3** (see Materials and Methods).

in 47 of 52 wild salmon being predicted correctly (sensitivity = 90.38% and specificity = 97.9%) as were 140 of 142 farmed samples (1 sample was not classified). Overall, the predictions were 95.9% correct (187) and 4.10% incorrect (*8*). The three most significant chemical shifts are shown in bold font in **Figure 2**; these were all in a crowded area of the olefinic region, which made unambiguous assignment of peaks difficult.

VAST scaling was then applied, and the samples were divided into training data (used to create the model) and a validation data set that was not used in any way during the model development. For LOO CV, using all samples for the wild and farmed classifications, 52 of 52 (wild) and 143 of 143 (farmed) were correctly predicted. With a training set of 130 samples, 34 of 34 (wild) and 96 of 96 (farmed) were correctly predicted, whereas in the corresponding validation set 18 of 18 (wild) and 47 of 47 (farmed) were correctly classified. We further reduced the training set to 100 samples and 28 of 28 (wild) and 72 of 72 (farmed) were accurately classified, whereas the validation predictions resulted in 20 of 24 (wild) and 71 of 71 (farmed) correct predictions. Finally, using only the top 10 chemical shifts, all samples, and LOO CV, all samples were predicted correctly; similarly, using a training set of 129 samples also resulted in correct classifications for both wild (37 of 37) and farmed (92 of 92); the corresponding validation set of 66 samples was almost as successful, correctly predicting 14 of 15 (wild) and 51 of 51 (farmed), giving a correct classification rate of 98.5% (65 of 66).

For the SVM analysis, and wild versus farmed calculations, 135 samples were used in the training data set (36 wild/99 farmed) and 59 samples (15 wild/44 farmed) in the validation set, with all samples being correctly classified. Ten randomly chosen wild and farmed training and validation sets were created, and all samples were correctly predicted using this method (100%).

The observation that such methods can so accurately model the wild or farmed status of salmon is consistent with related studies within the COFAWS consortium project. Multiprobe/multielement isoptopic analyses in combination with GC fatty acid composition also allowed for complete discrimination between authentic samples of wild and farmed salmon (*28*). This discrimination occurred despite variations in season, location (geographical origin), farming practice, year of capture, and diet, although these factors contribute to the overall intragroup variability observed for each class. Similarly, ^{1}H NMR analyses combined with SVM also resulted in complete discrimination of wild and farmed salmon (*26*).

Discrimination Regarding Geographical Origin of Farmed Salmon. Predictions of geographical origins are somewhat more difficult, with country of origin predictions using PNN, LOO CV, and all variables being shown in **Table 1**. Overall, 93.8% (137) were correctly classified, 4.1% (*6*) incorrectly classified, and 2.1% (*3*) not classified. Creating training data from two-thirds of the samples and validation sets from the remaining data and recalculating still generated reasonably good predictive results (**Table 2**). In particular, for the training data, 87.1% (88) were correct, 7.9% (*8*) incorrect,

and 5.0%(*5*) not classified, whereas for the validation test set, 82.2% (37) were correctly classified, 15.6% (*7*) incorrect, and 2.2% (*1*) not classified. It is recognized that there are limitations with respect to the number of samples representing some classes in this analysis.

Country of origin was also investigated using the SVM approach. LOO CV on all 146 samples resulted in 1 error (99.3% correct). Similarly, ^{1}H NMR and SVM analysis (*25*) demonstrated that approximately 93% of the samples could be accurately predicted.

The classification accuracy by SVM is similar to PNN in the training set. However, the results by SVM are slightly better than those yielded by the PNN approach in the validation set. SVMs have been shown to be robust with respect to a low ratio of training samples to the dimensionality of the input data and ill-behaved distributions. They are able to automatically learn difficult nonlinear boundaries and deliver a reproducible solution for a given training set and parameter settings. Differences in the generalization ability of SVM and PNN are related to differences in optimization strategies for the two methods, which may favor the SVM approach.

Market Samples. One of the major objectives of these studies is to be able to reliably detect labeling fraud. Therefore, two sets of data were tested with respect to prediction of wild versus farmed (**Table 3**) and geographical origin (**Table 4**). **Table 3** shows the results for 34 market samples in the wild/farmed prediction by SVM. Although the majority appear to be correctly labeled, we note that 5 samples marketed and labeled as "wild" appear to have been "farmed" in origin. Related discrepancies in labels were also demonstrated through isotopic analyses (*28*) and ^{1}H NMR (*26*). **Table 4** shows the predictions for country of origin for a set of 35 samples. Although many appear to be consistent with the labels and available information, it is clear that discrepancies also exist with respect to geographical origins.

In conclusion, ^{13}C NMR spectra of heterogeneous lipid extracts from muscle of farmed and wild salmon contain enough information (fatty acid profile, lipid classes, and positional distribution) to discriminate between wild and farmed salmon. Discrimination with respect to geographical origin was somewhat more difficult, but the geographical origin classification may suffer from limitations with respect to the number of samples representing some classes. The five market samples apparently mislabeled as wild Atlantic salmon, as confirmed by other analysis (*26*, *28*), were also readily detected by the method described here, which shows the potential of this technique for verification of production method of salmon.

ABBREVIATIONS USED

HR NMR, high-resolution nuclear magnetic resonance; GC, gas−liquid chromatography; PNN, probabilistic neural networks, SVM, support vector machines; SNIF-NMR, site-specific natural isotope fractionation NMR; IRMS, isotope ratio mass spectrometry; TMS, tetramethylsilane; pdf, probability density function; LOO CV, leave-one-out cross-validation; ERM; empirical risk minimization, SRM; structural risk minimization; VC, Vapnic−Chervonenkis; VAST, variable stability scaling.

ACKNOWLEDGMENT

We thank Pheroze Jungalwalla and Tassal Ltd., Tasmania, for supplying samples of cultivated salmon.

LITERATURE CITED

(1) Food and Agriculture Organization of the United Nations. *The State of World Fisheries and Aquaculture*; FAORome, Italy, 2007

(2) Jensen, B.-A. *Akvakultur i EU og Europa, Industry Report*; IntraFish Media: Bodo, Norway, 2007.

(3) *World Fisheries and Aquaculture Atlas*; Food and Agriculture Organization of the United Nations: Rome, Italy, 2001.

(4) Martinez, I. Revision of analytical methodologies to verify the production method of fish. In *Seafood Reserach from Fish to Dish*; Luten, J. B., Jacobsen, C., Bekaert, K., Sæbø, A., Oehlenschläger, J., Eds.; Wageningen Academic Publishers: Wageningen, The Netherlands, 2006; pp 541−550.

(5) Martinez, I. Authenticity assessment based on other principles: analysis of lipids, stable isotopes and trace elements. In *Fishery Products: Quality, Safety and Authenticity*; Oehlenschläger, J., Rehbein, H., Eds.; Blackwell Publishing: Oxford, U.K., 2009.

(6) Ackman, R. G. Fish lipids. In *Advances in Fish Science and Technology*; Connell, J. J., Ed.; Fishing New Books: Farnham, U.K., 1980; pp 86−103.

(7) Sargent, J. R.; Tocher, D. R.; Bell, J. D. The lipids. In *Fish Nutrition*, 3rd ed.; Academic Press: San Diego, CA, 2002; pp 181−257.

(8) Grahl-Nielsen, O. Fatty acid profiles as natural marks for stock identification. In *Stock Identification Methods*; Cadrin, S. X., Friedland, K. D., Waldman, J. R., Eds.; Academic Press: New York, 2004; pp 239−261.

(9) Torstensen, B. E.; Lie, O.; Froyland, L. Lipid metabolism and tissue composition in Atlantic salmon (*Salmo salar* L.)—effects of capelin oil, palm oil, and oleic acid-enriched sunflower oil as dietary lipid sources. *Lipids* **2000**, *35*, 653–664.

(10) Gunstone, F. D. High-resolution C-13 NMR—a technique for the study of lipid structure and composition. *Prog. Lipid Res.* **1994**, *33*, 19–28.

(11) Aursand, M.; Rainuzzo, J.; Gradalen, H. Quantitative high resolution ^{13}C and ^{1}H nuclear magnetic resonance of w-3 fatty acids from white muscle of Atlantic salmon (*Salmo salar*). *J. Am. Oil Chem. Soc.* **1993**, *70*, 971–981.

(12) Aursand, M.; Jorgensen, L.; Grasdalen, H. Positional distribution of omega-3-fatty-acids in marine lipid triacylglycerols by high-resolution C-13 nuclear-magnetic-resonance spectroscopy. *J. Am. Oil Chem. Soc.* **1995**, *72*, 293–297.

(13) Aursand, M., Gribbestad, I. S.; Martinez, I. Omega-3 fatty acid content of intact muscle of farmed Atlantic salmon (*Salmo salar*) examined by ^{1}H MAS NMR spectroscopy. In *Modern Magnetic Resonance, Part 1, Application in Chemistry, Biological and Marine Science*; Webb, G., Ed.; Springer: The Netherlands, 2006; pp 931−935.

(14) Gribbestad, I. S.; Aursand, M.; Martinez, I. High-resolution ^{1}H magnetic resonance spectroscopy of whole fish, fillets and extracts of farmed Atlantic salmon (*Salmo salar*) for quality assessment and compositional analyses. *Aquaculture* **2005**, *250*, 445–457.

(15) Falch, E.; Storseth, T. R.; Aursand, A. Multi-component analysis of marine lipids in fish gonads with emphasis on phospholipids using high resolution NMR spectroscopy. *Chem. Phys. Lipids* **2006**, *144*, 4–16.

(16) Falch, E.; Storseth, T. R.; Aursand, M. High resolution NMR for studying lipid hydrolysis and esterification in cod (*Gadus morhua*) gonads. *Chem. Phys. Lipids* **2007**, *147*, 46–57.

(17) Aursand, M.; Mabon, F.; Martin, G. J. High-resolution H-1 and H-2 NMR spectroscopy of pure essential fatty acids for plants and animals. *Magn. Reson. Chem.* **1997**, *35* S91–S100.

(18) Aursand, M.; Standal, I. B.; Axelson, D. E. High-resolution C-13 nuclear magnetic resonance spectroscopy pattern recognition of fish oil capsules. *J. Agric. Food Chem.* **2007**, *55*, 38–47.

(19) Siddiqui, N.; Sim, J.; Silwood, C. J. L.; Toms, H.; Iles, R. A.; Grootveld, M. Multicomponent analysis of encapsulated marine oil supplements using high-resolution H-1 and C-13 NMR techniques. *J. Lipid Res.* **2003**, *44*, 2406–2427.

(20) Medina, I.; Sacchi, R. Acyl stereospecific analysis of tuna phospholipids via high resolution ^{13}C NMR spectroscopy. *Chem. Phys. Lipids* **1994**, *70*, 53–61.
(21) Aursand, M.; Axelson, D. Origin recognition of wild and farmed salmon (Norway and Scotland) using ^{13}C NMR spectroscopy in combination with pattern recognition techniques. In *Magnetic Resonance in Food Science: A View to the Future*; Webb, G. A., Belton, P. S., Gil, A. M., Delgadillo, I. Eds.; RSC Books: London, U.K., 2001; pp 227−231
(22) Molkentin, J.; Meisel, H.; Lehmann, I.; Rehbein, H. Identification of organically farmed Atlantic salmon by analysis of stable isotopes and fatty acids. *Eur. Food Res. Technol*. **2007**, *224* (5), 535–543.
(23) Aparicio, R.; McIntyre, P.; Aursand, M.; Eveleigh, L; Marighetto, N.; Rossell, B.; Sacchi, R.; Wilson, R.; Woolfe, M. Fish oil products. In *Food Authenticity Issues and Methodologies*. *Final Report of the Concerted Action AIR3-CT94-2452*; Lees, M., Ed.; Eurofins Scientific: Nantes, France, 1998; pp 213−218.
(24) Aursand, M.; Mabon, F.; Martin, G. J. Characterization of farmed and wild salmon (*Salmo salar*) by a combined use of compositional and isotopic analyses. *J. Am. Oil Chem. Soc*. **2000**, *77*, 659–666.
(25) Medina, I.; Aubourg, S. P.; Martin, R. P. Species differentiation by multivariate analysis of phospholipids from canned Atlantic tuna. *J. Agric. Food Chem*. **1997**, *45*, 2495–2499.
(26) Masoum, S.; Malabat, C.; Jalali-Heravi, M.; Guillou, C.; Rezzi, S.; Rutledge, D. N. Application of support vector machines to H-1 NMR data of fish oils: methodology for the confirmation of wild and farmed salmon and their origins. *Anal. Bioanal. Chem*. **2007**, *387*, 1499–1510.
(27) Igarashi, T.; Aursand, M.; Hirata, Y.; Gribbestad, I. S.; Wada, S.; Nonaka, M. Nondestructive quantitative determination of docosahexaenoic acid and n-3 fatty acids in fish oils by high-resolution H-1 nuclear magnetic resonance spectroscopy. *J. Am. Oil Chem. Soc*. **2000**, *77*, 737–748.
(28) Thomas, F.; Jamin, E.; Wietzerbin, K.; Guerin, R.; Lees, M.; Morvan, E.; Billault, I.; Derrien, S.; Rojas, J. M.; Serra, F.; Guillou, C.; Aursand, M.; McEvoy, L.; Prael, A.; Robins, R. J. Determination of origin of Atlantic salmon (*Salmo salar*): the use of multiprobe and multielement isotopic analyses in combination with fatty acid composition to assess wild or farmed origin. *J. Agric. Food Chem*. **2008**, *56*, 989–997.
(29) Bligh, E. G.; Dyer, W. J. A rapid method of totoal lipid extraction and purification. *Can. J. Biochem. Physiol*. **1959**, *37*, 911–917.
(30) Specht, D. F. Probabilistic neural networks and the polynomial Adaline as complementary techniques for classification. *Neural Networks, IEEE Trans*. **1990**, *1*, 111–121.
(31) Cortes, C.; Vapnik, V. Support-vector network. *Machine Learning* **1995**, *20*, 273–297.
(32) Anguita, D.; Boni, A.; Ridella, S.; Rivieccio, F.; Sterpi, D. Theoretical and practical model selection methods for support vector classifers. In *Support Vector Machines: Theory and Applications*; Wang, L., Ed.; CSVM, v3.1.8; 2005.
(33) Davis, L. In *Handbook of Genetic Algorithms*; Davis, L., Ed.; Van Norstrand Reinhold: New York, 1991; Chapter 4.
(34) Keun, H. C.; Ebbels, T. M. D.; Antti, H.; Bollard, M. E.; Beckonert, O.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Improved analysis of multivariate data by variable stability, scaling: application to NMR-based metabolic profiling. *Anal. Chim. Acta* **2003**, *490*, 265–276.

Received for review December 18, 2008. Revised manuscript received March 6, 2009. Accepted March 13, 2009. This work received financial support through the shared-cost RTD Project COFAWS (Confirmation of the origin of farmed and wild salmon and other fish, Contract G6RD-CT-2001-00512) funded by the European Community under the Competitive and Sustainable Growth Programme (1998−2002), for which the other partners were Eurofins Scientific (Nantes, France), Universit de Nantes (Nantes, France), and Joint Research Centre (Ispra, Italy), and a project financed by the Norwegian Research Council (NFR Project 146932/130).