Food Chemistry 116 (2009) 766–773

Contents lists available at ScienceDirect

Food Chemistry

journal homepage: www.elsevier.com/locate/foodchem

Analytical Methods

# Identification of the farm origin of salmon by fatty acid and HR $^{13}$C NMR profiling

I. Martinez [a,b,*], I.B. Standal [a,c], D.E. Axelson [a], B. Finstad [d], M. Aursand [a,c]

[a] SINTEF Fisheries and Aquaculture Ltd., Processing Technology, N-7465 Trondheim, Norway
[b] Department of Marine Biotechnology, Norwegian College of Fishery Science, University of Tromsø, N-9037 Tromsø, Norway
[c] Department of Biotechnology, University of Trondheim, N-7465 Trondheim, Norway
[d] Norwegian Institute for Nature Research, N-7047 Trondheim, Norway

ARTICLE INFO

*Article history:*
Received 2 September 2008
Received in revised form 27 January 2009
Accepted 3 March 2009

*Keywords:*
Fish
Salmon
NMR
Lipids
Chemometrics
Geographical origin
Authenticity

ABSTRACT

Lipid analyses by gas chromatography (GC) and by high resolution (HR) $^{13}$C NMR combined with chemometrics were used to identify wild and farmed Atlantic salmon and the farm origin of farmed salmon. Reference samples were 59 specimens from four different farms in the Hardangerfjord (Norway) and the test fish were 17 free-living fish, caught in the same fjord. Four free-living fish were identified as wild by their fatty acids profile, n3/n6 ratio and by principal component analysis. To identify the farm of origin of farmed salmon, Bayesian belief networks (BBN) and support vector machines (SVM) were the best methods classifying correctly 58 (BBN and GC) and 56 (SVM and $^{13}$C NMR) of the 59 reference samples. Of the 12 free-living fish identified as farmed, four seemed to originate from farm 2 and 3 from farm 4. The rest could not be clearly attributed to any of the four farms and may originate from any of the other 26 farms located in the fjord.



## 1. Introduction

Suitable analytical techniques are required by the organisations responsible for consumer protection in order to ensure the authenticity of foodstuff. Similarly, fisheries and aquaculture management require the same type of techniques to document the origin of fish: wild, farmed and the farm of origin. In addition, there has been an increasing awareness and concern among consumers about the origin and the conditions under which their foodstuffs are produced and consumers demand correct information about the species, method of production and geographical origin (Frewer & Kher, 2008, personal communication). Moreover, this information is regulated in the EU (CR 2065/2001) and other countries, such as USA (The Fair Packaging and Labeling Act) and Japan (Japanese Agricultural Standards Law). Regarding fisheries and aquaculture management, the authorities call for reliable producers to identify the origin of the products, going back to the fishing ground (Primmer, Koskinen, & Piironen, 2000) or the farm. In the case of aquaculture products, markers need to be developed to identify escaped from wild fish and also the farm origin of the escaped fish: 600,000 cultivated salmon escaped from farms in Norway in 2007 (www.statistics.no). The true figures of escaped fish are considered to be higher than the official ones, sometimes because farmers are not aware of their fish escaping and others due to under-reporting, since they have to pay severe fines both if they delay reporting a suspected incident and also if the escape is due to negligence.

The discrimination of wild from farmed Atlantic salmon has been successfully achieved by a variety of methods, from genetic analysis, often using microsatellite polymorphisms (Coughlan et al., 1998; Glover, Skilbrei, & Skaala, 2008; Skaala, Taggart, & Gunnes, 2005), to stable isotope ratio analysis ($\delta^{13}$C, $\delta^{15}$N, $\delta^{18}O_{glycerol}$, $\delta^{18}O_{oil}$, $\delta^{18}O_{water}$; Aursand, Mabon, & Martin, 2000; Thomas et al., 2008), gas chromatography analysis of the fatty acids from the triglyceride fraction and $^{13}$C and $^{1}$H HR NMR spectroscopic analyses (Aursand et al., 2000; Thomas et al., 2008). The assignment of individuals to their population of origin by using multilocus genotyping, in particular using microsatellite data, has been proven possible provided that there are enough genetic markers differing among the potential populations of origin (Skaala, Høyheim, Glover, & Dahle, 2004). The analysis of the fatty acid composition of the triglyceride fraction has often been proven to be sufficient to discriminate farmed from wild salmon (Aursand et al., 2000; Martinez, in press; Molkentin, Meisel, Lehmann, & Rehbein, 2007) due to the fact that this lipid fraction reflects the

* Corresponding author. Address: SINTEF Fisheries and Aquaculture Ltd., Processing Technology, N-7465 Trondheim, Norway. Tel.: +47 957 09 772; fax: +47 93 27 07 01.
*E-mail address:* iciar.martinez@sintef.no (I. Martinez).

0308-8146/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.foodchem.2009.03.026

composition of the diet, and the commercial diets have variable – and sometimes high – amounts of vegetable oils which are normally absent from the natural feed of wild salmonids (Martinez, 2006; Thomas et al., 2008). Similarly, the HR NMR spectra contain, in addition to the information regarding the fatty acid composition, other relevant parameters such as the positional distribution of the fatty acids and other biochemical compounds, and it has also been proven suitable to discriminate the production method (Aursand, Jørgensen, & Grasdalen, 1995; Martinez, 2006).

Optimal application of these analytical techniques, in particular of the complex information contained in the HR NMR spectra, requires the use of advanced chemometrics in order to obtain correct classifications, such as principal component analysis (PCA), probabilistic neural networks (PNN) (Aursand, Standal, & Axelson, 2007; Specht, 1990), Bayesian belief networks (BBN) (Glover et al., 2008; Heckerman, Geiger, & Chickering, 1995) or support vector machines (SVM) (Masoum et al., 2007).

In 2003, the Norwegian Ministry of Fisheries took the initiative to set up a national committee to investigate some questions regarding the tagging of farmed fish. In the same year, the Director of Fisheries set up the Tagging Committee with representatives of the aquaculture industry, the research community and the authorities, with the mandate to present a concrete range of tagging/tracing systems for farmed salmon. Several techniques were selected for testing, including genetic analysis (Glover et al., 2008) and the lipid analyses presented here. The aim of this work was to examine the suitability of fatty acid profiling by gas chromatography and $^{13}C$ HR NMR of lipids extracted from the muscle fraction of Atlantic salmon in combination with chemometric analyses, in order to identify: (1) the farm where cultivated fish were reared and (2) in the case of fish captured outside pens in the fjord, whether it was possible to classify them as originating from a given farm. From each of the four selected farms from the Hardangerfjord (Norway), 15 Atlantic salmon and their feeds were collected in July 2006. Additionally, 17 salmon caught in the same fjord during the period October 2005–October 2006, were examined in this work.

## 2. Materials and methods

### 2.1. Fish and feed samples

Salmon and samples of their feeds were collected from four of the about 30 farms located in the Hardangerfjord (Norway) in July 2006. In addition, 17 free-living salmon (individually numbered 501–517) were caught by net bags in the same fjord during October 2005–October 2006 by fishermen. The free-living fish number 512 was visually identified as a rainbow trout. Microscopic examination of the scales of the free-living fish, carried out routinely by the Norwegian Institute for Nature Research according to Lund and Hansen (1991) indicated that only fish 505 and 510 had the sharp and clearly defined transition zones from freshwater to saltwater characteristic for wild salmon. All the other free-living fish (501–504; 506–509 and 511–517), lacking the transition zone, were considered farmed according to this information. See Table 1 for additional data.

The fish was frozen at −20 °C and transported by boat to our laboratory in August 2006 (farmed fish) and in January 2007 (free-living fish). The fish were thawed, gutted and muscle samples were cut and further stored at −80 °C until the lipid was extracted. The samples of feed were transported and stored at room temperature until the lipid was extracted.

**Table 1**
Data on the fish samples. Weight and fat content are given as average ± std except for the % fat content of the free-living fish, where it is given the lowest and highest % fat values measured in that group.

| *n* | Weight (kg) | Fat (%) | Farm number | Feed |
|---|---|---|---|---|
| 15 | 3.15 ± 0.77 | 13.12 ± 3.47 | 1 | Skretting optiline |
| 15 | 3.69 ± 1.62 | 11.89 ± 5.28 | 2 | Ewos |
| 15[a] | 4.82 ± 1.11 | 17.24 ± 1.84 | 3 | Skretting optiline |
| 15[b] | 4.38 ± 1.47 | 16.67 ± 3.01 | 4 | Ewos (Bremnes Seawash) |
| 17 | 3.52 ± 0.94 | 2.21–21.52[c] | Free-living | Unknown |

[a] One fish not available for NMR.
[b] One fish not available for GC.
[c] Lowest and highest % fat recorded in the group.

### 2.2. Fat extraction

Lipid extraction was performed according to a modified Bligh and Dyer (1959) procedure: 10 g of fish muscle or 10 g of feed were homogenised for 2 min with a mixture of 16 ml of $H_2O$, 40 ml of methanol and 20 ml of chloroform. Then 20 ml of chloroform were added to the mixture and homogenised for another 40 s prior the addition of 20 ml of $H_2O$ and a final homogenisation for 40 s. The homogenate was centrifuged for 10 min at 4100*g* and the chloroform phase containing the lipids was recuperated for subsequent GC and NMR analyses.

### 2.3. Fatty acid (FA) analysis by gas chromatography (GC)

The lipids were first transesterified with boron trifluoride-methanol and 0.5 M methanolic sodium hydroxide and the fatty acid methyl esters (FAMEs) were extracted into hexane (AOCS Method CE 2-66). An internal standard (21:0 methyl ester) was added to the extract prior to methylation. FAMEs were analysed on a Fison 8160 (Fisons Instruments S.pA. Milan, Italy) capillary gas chromatograph equipped with capillary cold on-column injector, a fused silica capillary column, Omegawax 320 (30 m, 0.32 mm id, 0.25 μm film thickness; Supelco Inc., Bellefonte, PA) connected to a flame ionisation detector (FID). The FID was connected to a computer implemented with Chrom-card for Windows 1.21 software. The gas chromatograph was provided with AS800 auto-sampler. The oven temperature was increased from 80 to 180 °C at 25 °C $min^{-1}$ and held for 2 min. Then the temperature was further increased by 2.5 °C $min^{-1}$ to 205 °C and held for 8 min, and up to 215 °C $min^{-1}$ and held for 3 min. The temperature of the detector was 250 °C. Hydrogen was used as carrier gas at a flow rate of 1.6 ml $min^{-1}$. The instrument was calibrated using the reference standards mixture GLC-68-D, (Nu-Chek-Prep, Elysian, MN). The samples were run under the same conditions than the reference standards and the FAMEs were identified by comparison of their retention times with those of the reference standards used to calibrate the instrument. In addition, each fatty acid had been identified in our laboratory in previous works using the individual FAMEs.

The fatty acids analyzed were: C12:0, C14:0, C14:1n5, C16:0, C16:1n7, C18:0, C18:1n9, C18:1n7, C18:2n6, C18:3n6, C18:3n3, C18:4n3, C20:0, C20:1n9, C20:1n7, C20:2n6, C20:3n6, C20:4n6, C20:3n3, C20:4n3, C20:5n3, C22:0, C22:1n11, C22:1n9, C22:5n3, C24:0, C22:6n3 and C24:1n9.

### 2.4. $^{13}C$ NMR analysis

Before analysing the lipid extract by NMR, most of the chloroform phase was removed by evaporation. The acquisition of NMR spectra was carried out as follows: about 70 mg of fish oil were mixed with 0.5 ml of $CDCl_3$ (99.8% purity, Isotec Inc., Matheson) and placed in 5-mm NMR tubes. Proton-decoupled $^{13}C$ spectra were recorded on a Bruker Avance instrument at 125.75 MHz (Bruker BioSpin GmbH, Rheinstetten, Germany). A semiquantitative approach was chosen due to the fact that quantitative measurements

require considerably longer experimental time. Although the signal intensities within each spectrum are not quantitative, the relative intensities for corresponding signals across different spectra are comparable. When the same oil was run at different concentrations, we could detect minor variations in signal intensities for some resonances as a consequence of differential signal overlap and minor solvent effects. Minor perturbations of selected resonances do not significantly impact on the classification accuracy since it is the total relationship among all peaks that is being modelled. In this respect, the PNN, SVM and BBN approaches are robust (including routine and unavoidable variations in such basic factors as signal-to-noise ratios). For pattern recognition and classification studies, it is the overall relationship among resonances that determine class assignment. These patterns exist clearly even for spectra acquired under non-equilibrium conditions, the key factor being that the spectra must be acquired under conditions as similar as possible: optimally, under identical conditions.

The following experimental conditions were applied: spectral width 200.78 ppm, pulse angle 30°, relaxation delay 2.5 s, dwell time 19.80 μs, acquisition time 2.0 s, and time domain 101,006 data points. The number of scans was set to 1k for the spectra used in the multivariate data analysis. Prior to Fourier transformation, a line-broadening factor of 0.1 Hz was applied to minimise noise, but not at the expense of resolution among closely spaced resonances. Chemical shifts were referenced to the $CDCl_3$ peak at 77.0 ppm. The resulting Bruker 1r files were converted to ASCII files and preprocessed prior to multivariate data analysis treatment.

### 2.5. Chemometric analyses

For data treatment of the GC analysis, the input data were the relative amounts of each of the 28 analyzed FA in % of the area under the peak, when 100% was the sum of all the areas for the peaks corresponding to the 28 FAs. From the NMR full spectra, we obtained first the position of the peaks (shift resonance values in ppm) in each spectrum and the corresponding intensities for all resonances with relative intensities greater than 1.5% of the maximum peak intensity (excluding any solvent resonances). Spectra were peak aligned (Lee & Woodruff, 2004), and then normalised, with the maximum non-solvent peak in each spectrum scaled to a value of 100. This produced over 10,000 chemical shift values from each HR $^{13}C$ NMR spectrum, which were reduced to the best (most informative) 249 chemical shift values by Uniformative Variable Elimination Partial Least Square Regression (PLS-UVE) for chemometric analyses.

The results of the GC analyses and the NMR spectra were submitted to four classification techniques: (1) principal component analysis (PCA) using the Unscrambler®, v9.7, CAMO software AS, Norway; (2) probabilistic neural network analysis (PNN), using the software AI Trilogy, NeuroShell Classifier, Ward Systems Group Inc., Frederick, MD, USA; (3) support vector machines (SVM) (Cristiani & Shawe-Taylor, 2000) with the programme Tiberius v6.13, Tiberius Data Mining, Melbourne, Australia and (4) Bayesian belief network (BBN) classifications employing Netica v4.02, Norsys Software Corporation, Vancouver BC, Canada.

Principal component analysis (PCA) (Jolliffe, 1986), is a vector space transform often used to reduce multidimensional data sets to lower dimensions for analysis. It is mostly used as a tool in exploratory data analysis and for making predictive models. In PCA, the original variables are transformed into new, uncorrelated variables called principal components, which retain as much as possible of the information present in the original data. Each principal component (PC) is a linear combination of the original variables. The scores of a subset of the principal components, can be used in subsequent multivariate analysis. Validation procedures are used to indicate how well a model will perform for future samples taken from the same population as the calibration samples and to selected the suitable number of components (to avoid using too many components that would be trying to explain the noise in the data). There are several available validation methods. The method used here, called full cross-validation, is a validation method where some samples are kept out of the calibration and used for prediction. This is repeated until all samples have been kept out once. In full cross-validation, only one sample at a time is kept out of the calibration. For interpretation of the data, samples that cluster together in the scores plot have more features in common than samples that do not cluster. In the present case, this statistical analysis may be used to attribute a given origin to an unknown sample, but it does not need to be so: samples originating from the same farm may be scattered rather than form compact clusters but also there may be additional variables, not included in the model, containing information essential to identify the origin. If that is the case, the clustering will not be sufficient to identify the origin, although it will indicate similarity regarding the variables used as input for the model.

The probabilistic neural network was developed by Specht (1990). It uses a supervised training set to develop, from the known input reference data, distribution functions within a pattern layer. PNNs have input, pattern and summation layers. PNN operates by defining a probability density function (pdf) for each class based on the training set data and an optimised kernel width parameter. 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, while the kernel width determines the amount of interpolation that occurs between adjacent kernels. 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 whose elements 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, and produces a one for that class and a zero for the other classes. In the present case, we defined a model with four output classes (one for each farm) for classification when only the reference samples were used as input, since we knew a priori that there were only four classes in the model. For the second part of the work, the classification of the free-living fish (whose origin was unknown) we defined a model with five possible output classes: one for each farm and a fifth one for "any other origin" (which would include wild or another farm not sampled). Two types of classification test were used: the first used leave-one-out cross-validation and the second involved training the classification model on a randomly chosen set of samples ("training" set) and then applying it to the a "test" (or validation) set of samples that had not been used in creating the model, in order to simulate the prediction of unknowns.

Support vector machines (SVMs) (Cortes & Vapnik, 1995) are a set of related supervised learning methods used for classification and regression that belong to a family of generalised linear classifiers (see also http://en.wikipedia.org/wiki/Support_vector_machine). In class separation by SVM, the optimal separating hyperplane between the two classes are searched for by maximising the margin between the classes' closest points. Those training points lying on one of the hyperplanes and whose removal would change the solution found are called support vectors, and the middle of the margin is the optimal separating hyperplane. 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. Using SVM algorithms, the model will classify the samples analyzed as "belonging" or as "not belonging" to a given class. The classification test used involved, as above, training the classification model on a "training" set and them applying it to a "test" set of samples.

A Bayesian network, also called a belief network, is a probabilistic graphical model that represents a set of variables and their probabilistic independencies. A Bayesian net (Heckerman et al., 1995; Pearl, 1988) is a graph-based model for representing probabilistic relationships between random variables. The random variables are modelled as graph nodes, probabilistic relationships are captured by directed edges between the nodes and conditional probability distributions associated with the nodes. Nodes can represent any kind of variable, be it a measured parameter, a latent variable or a hypothesis. A Bayesian net asserts that each node is statistically independent of all its nondescendants, once the values of its parents (immediate ancestors) in the graph are known. Like in the case of PNN, the BBN algorithm was tested first to classify the known farmed fish to their farm of origin, i.e., an output of four classes for the model. The BBN algorithm was then applied to all farmed and free-living farmed fish with an output of five classes: the four farms and an extra class of "any other origin". As before, the classification test used involved training the classification model on a "training" set and them applying it to a "test" set of samples.

In Section 3, the training and test data sets appear combined in only one instead of reporting the validation and classification tests separately. This has been done because for most methods the training accuracy is almost always close to 100%.

## 3. Results and discussion

### 3.1. FA profiles

The feed used in farms 2 and 4 contained higher levels of C16:0 (palmitic acid), C20:1n9 (eicosenoic acid), C20:5n3 (eicosapentaenoic acid, EPA), C22:1n11 (cetoleic acid) and C22:6n3 (docosahexaenoic acid, DHA), while the feed used in farms 1 and 3 had higher contents of C18:1n9 (oleic acid), C18:2n6 (linoleic acid, LA) and C18:3n3 (α-linolenic acid, ALA). The latter three are abundant in plant oils, in particular C18:2n6, high in soybean oil and present only in minor amounts in some wild fatty fish species (Martinez, 2006). Salmonids, sardines and peruvian PUFA (polyunsaturated fatty acids mixture used as ingredient in feed formulations) are usually rich in C16:0; C18:1n9; C20:5n3 and C22:6n3, and C22:1 (coho salmon), while herring has high contents of C20:1n9 and C22:1n11. This indicates that the feed used in farms 1 and 3 had a higher amount of vegetable oils than the feed used in farms 2 and 4.

Principal component analysis was performed using 29 variables in the model: the 28 FA and the n3/n6 ratio. The results of the model for all the farmed fish and feeds are shown in Fig. 1: the first component explained 68% and the second 15% of the total variability of the model and the most relevant FA were C18:1n9, C18:2n6 and C18:3n3 with the highest positive loadings on PC1 and C22:6n3, C20:5n3, C16:0 and n3/n6 with the highest negative loadings for the same PC1, i.e., the distribution of samples along the PC1 axis would indicate the amount of vegetable (positive loadings) versus fish oils (negative loadings).

The second factor was mostly due to the FA C16:0. The scores plot clearly separated the feed used in farms 1 and 3 versus the feeds used in farms 2 and 4. The distribution of the individual fish did not follow the same pattern as that of their feeds: while most salmons from farm 2 did resemble their diet, with low score values on PC1 (low content in vegetable oils and a high n3/n6 ratio), fish

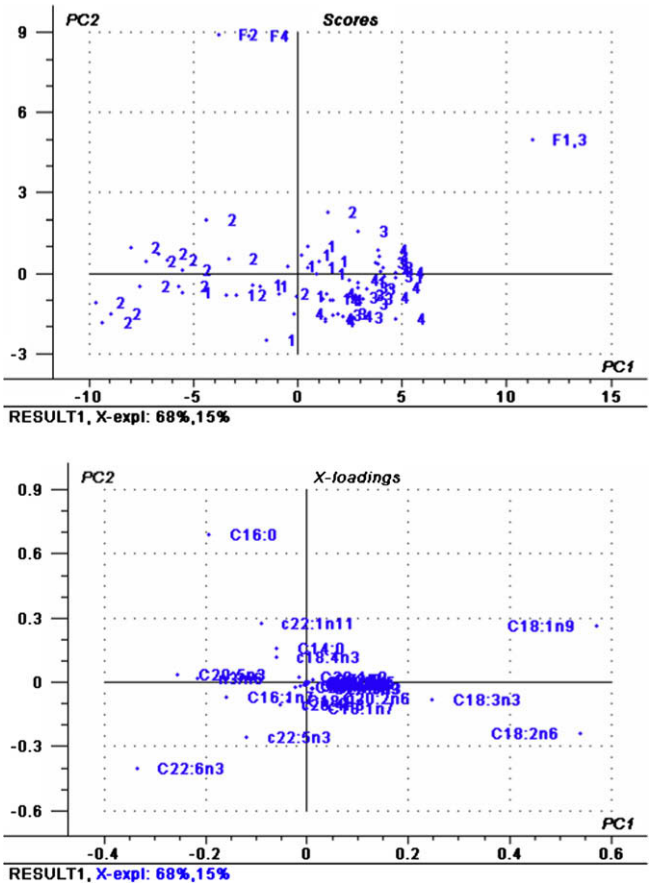


**Fig. 1.** Principal component analysis of the farmed fish and their feeds using 28 FA and the n3/n6 value as variables. Top, scores plot, bottom, loadings plot. The model was cross validated and centered.

from farm 4 had in fact the highest positive scores in PC1 and clustered closer to fish from farm 3, while fish from farm 1 were between those from farm 2 and the cluster of farms 3 and 4. Moreover, unlike fish from farms 3 and 4 that formed compact clusters, fish from farm 1 were more stretched along the PC1 axis and fish from farm 2 extended along both axis. This is shown in Fig. 2 (top), where only the farmed fish have been included in the model.

Atlantic salmon selectively retains or metabolises different fatty acids: C22:5n6 is selectively deposited, so that the concentration in the flesh is usually higher than in the diet, while C22:1n11, C18:2n6 and C18:3n3 are selectively metabolised (Bell et al., 2001). Also, the fatty acids C18:1n9 and C18:2n6 can be considered as markers for vegetable oils and the latter seems to be the most persistent after a dietary switch (Bell et al., 2001). This affects the restoration of the n3/n6 ratio which is usually higher in wild than in farmed salmon (Martinez, 2006; Thomas et al., 2008). Thus, the difference in the FA profile of farms 2 and 4 can be explained if fish from farm 4 have been receiving the analyzed feed only for the last period of time so that the washing out period from a hypothetical previous feed richer in vegetable oils, had not been completed. That would explain their proximity to fish receiving a diet richer in vegetable oils, such as those from farm 3. An alternative explanation may be that fish from farms 3 and 4 were larger and had a higher fat content than fish from farms 1 and 2, which would explain a higher content of the most persistent FA, namely C18:1n9 and C18:2n6, in their higher triglyceride fraction.

Fig. 2 (middle) shows the results of analysing all the fish, both farmed and the free-living. Four of the free-living fish, specimens

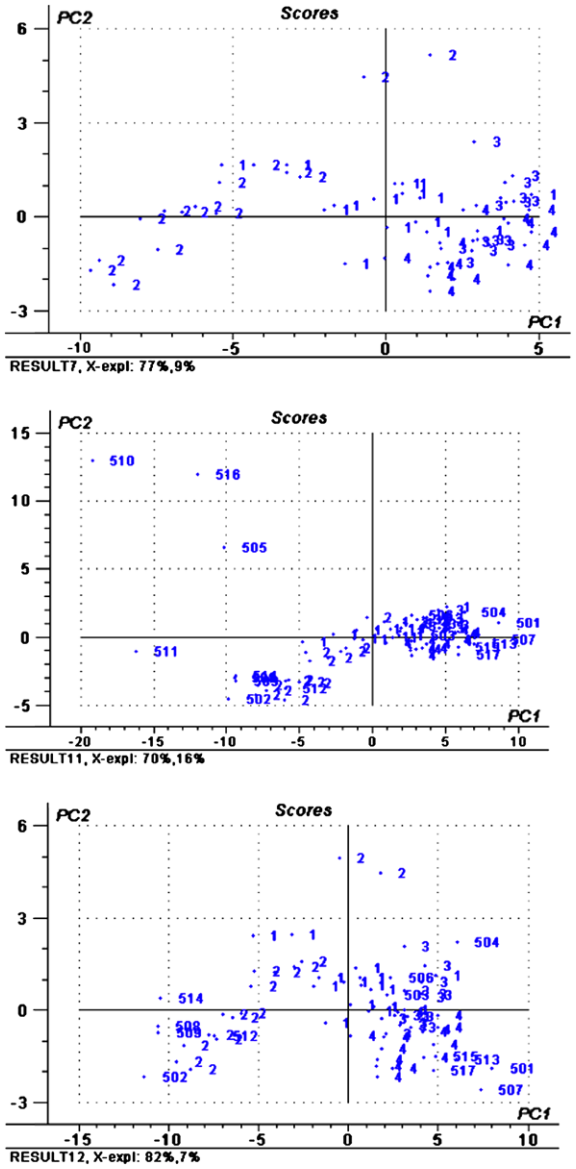


**Fig. 2.** Scores plot of the principal component analysis of only the farmed fish (top) the farmed and all free-living farmed fish (middle) and of the farmed and the free-living escaped fish (bottom), using 28 FA and the n3/n6 value as variables. The model was cross validated and centered.

505, 510, 511 and 516 displayed the FA composition (high DHA and EPA and low C18:1n9 and C18:2n6) and high n3/n6 ratio characteristic of wild fish (Aursand et al., 2000; Martinez, 2006; Thomas et al., 2008). Of the rest, individuals 502, 508, 509, 512 and 514 clustered close to fish from farm 2; fish numbers 513, 515 and 517 were close to fish from farm 4; and fish 501 and 507 were also close to farm 4 but scored higher than farm 4 on PC1. Free-living fish 503, 504 and 506 clustered close to farms 3 and 1, which were, as mentioned above, difficult to separate by this analysis (Fig. 2, middle and bottom). Fish 512, identified morphologically as a rainbow trout, was according to this analysis undistinguishable from fish from farm 2, apparently indicating that it had received the same feed used in farm 2. These results are summarised in Table 2.

The contradiction between the results of the scale morphology and FA composition for fish numbers 511 and 516 (that lacked the transition zone characteristic for wild salmon but did not classify with farmed salmon according to the FA profile) can be explained because according to scale morphology, about 5% of wild salmon would be identified as farmed for displaying a diffusing transition between the freshwater and the saltwater zones (Lund & Hansen, 1991). However, a more likely explanation may be that these two fish either escaped or were freed for the purpose of repopulation, early in their life. Indeed, there is a stocking station in the Hardangerfjord that liberates about 30,000 individuals a year. In any case, whether escaped of freed, these two fish had been long enough in the wild to achieve a FA profile that made them indistinguishable from wild individuals and they were clearly different from all the other farmed and free-living salmons.

PCA is only capable of identifying gross variability, it is not capable of distinguishing 'among-groups' and 'within-groups' variability, and it is frequently successful since the among-groups variability dominates the within-groups variability. When PCA locates a direction of 'maximum gross variability' it has in fact found a direction that is consistent with group separation. However, if this is not the case, and the group-to-group differences do not dominate the total variability as measured by the variance–covariance matrix, PCA fails as a classification tool. That may be the reason for the improvement in the classification of the reference samples when we applied any of the additional multivariate data analyses. Very good classifications were obtained, with values of between 93.3% and 100% correct assignments (Tables 3 and 4). The data shown in Table 3 include the added results of the training and testing data sets for the PNN and BBM models. The PNN model was unable to classify one fish from farm 2, but all the other ones were correctly ascribed to their farm of origin, while BBN misclassified only one fish from farm 1 to farm 3 (Table 3). The SVM model misclassified one fish from farm 2, one from farm 3, and misclassified to farm 3 a fish of unknown origin (Table 4).

The 10 most relevant FA on the BBN model were C18:2n6, C18:3n3, C20:4n6, C22:1n9, C18:4n3, C20:3n3, C22:6n3, C16:1n7, C20:5n3 and C18:0, with mutual information values varying from 1.18 to 0.74. For feature selection, the mutual information (Peng, Long, & Ding, 2005) between each attribute and the class attribute is measured. Mutual information measures the strength of the correlation between the values of the attribute and the values of the class; it quantifies the distance between the joint distribution of two discrete random variables $X$ and $Y$ and what the joint distribution would be if $X$ and $Y$ were truly independent. The influence of each FA on the BBN model was different from that exerted on the PCA: for example C18:2n6 was one of the most relevant FA in both cases, but then C18:3n3 was the second most relevant for the BBN, while it had only a modest positive loading of PC1 and practically no influence on PC2.

In order to identify the origin of the free-living fish, the models for PNN and BBN had five output classes, one for each farm and another one for "any other origin". Table 2 shows that all the wild fish were correctly classified as not belonging to any of the farms. Of the three multivariate data analyses used on the GC data, PNN seemed to be the worst when it came to identifying the origin of the free-living individuals: while both SVM and BBN were able to allocate eight individuals to farms 2, 3 and 4, (with full agreement in these assignments), PNN was only able to allocate two fish. One of them agreed with the other statistics but interestingly, the other one, number 513, was allocated to farm 4 by PCA of GC and by all the NMR analyses (see below), but not by either SVM or BBN of GC data. In total, GC analysis classified 4 specimens as belonging to farm 2 (including the rainbow trout), 3 or 4 individuals to farm 4 and 1 to farm 3.

### 3.2. $^{13}C$ NMR data

PCA analysis of the $^{13}C$ NMR spectra showed the same pattern of the PCA analyses of the fatty acid profiles (not shown): the samples

**Table 2**
Classification of free-living fish to their farm of origin according to the different methods tested. PCA, principal component analysis; PNN, probabilistic neural networks; SVM, support vector machines; BBN, Bayesian belief network; GC, gas chromatography; NMR, HR $^{13}C$ NMR, the number in parenthesis indicates the farm to which the specimen was close to, but not clustering with; –, individual classified as being from "any other origin"; *X*, unknown origin; LC, lack of consensus in the origin of the fish.

| Fish number | Possible origin | GC | | | | NMR | | |
|---|---|---|---|---|---|---|---|---|
| | | PCA | PNN | SVM | BBN | PNN | SVM | BBN |
| 501 | *X* | – | – | – | – | – | – | – |
| 502 | *X* | (2) | – | – | – | – | – | – |
| 503 | LC | 3/4 | – | – | – | 4 | 4 | 3 |
| 504 | *X* | – | – | – | – | – | – | – |
| 505 | *W* | Wild | – | – | – | – | – | – |
| 506 | LC | 1/3 | – | 3 | 3 | 1 | – | 1 |
| 507 | LC | – | – | 4 | 4 | 3 | 1 | 4 |
| 508 | 2 | (2) | – | 2 | 2 | 2 | – | 2 |
| 509 | 2 | (2) | – | 2 | 2 | – | 2 | 2 |
| 510 | *W* | Wild | – | – | – | – | – | – |
| 511 | *W* | Wild | – | – | – | – | – | – |
| 512 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 4 |
| 513 | 4 | 4 | 4 | – | – | 4 | 4 | 4 |
| 514 | 2 | (2) | – | 2 | 2 | – | – | 2 |
| 515 | 4 | 4 | – | 4 | 4 | 4 | 4 | 4 |
| 516 | *W* | Wild | – | – | – | – | – | – |
| 517 | 4 | 4 | – | 4 | 4 | 4 | 4 | 4 |

**Table 3**
Results of the classification of farmed fish according to the GC and NMR data and using PNN and Bayesian analyses: number of fish allocated to each farm.

| Analysis | | | Actual farm of origin | | | | Total | Correct classification (%) |
|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | | |
| PNN of GC data[a] (*n* = 59 fish) | Classified as from farm | 1 | 15 | 0 | 0 | 0 | 15 | 100 |
| | | 2 | 0 | 14 | 0 | 0 | 14 | 93.3 |
| | | 3 | 0 | 0 | 15 | 0 | 15 | 100 |
| | | 4 | 0 | 0 | 0 | 14 | 14 | 100 |
| Bayesian of GC data (*n* = 59 fish) | Classified as from farm | 1 | 14 | 0 | 1 | 0 | 15 | 93.3 |
| | | 2 | 0 | 15 | 0 | 0 | 15 | 100 |
| | | 3 | 0 | 0 | 15 | 0 | 15 | 100 |
| | | 4 | 0 | 0 | 0 | 14 | 14 | 100 |
| PNN of NMR data (*n* = 59 fish) | Classified as from farm | 1 | 13 | 1 | 0 | 1 | 15 | 86.7 |
| | | 2 | 0 | 14 | 0 | 0 | 14 | 100 |
| | | 3 | 1 | 0 | 14 | 0 | 15 | 93.3 |
| | | 4 | 1 | 0 | 0 | 14 | 15 | 93.3 |
| Bayesian of NMR data (*n* = 59 fish) | Classified as from farm | 1 | 14 | 0 | 1 | 1 | 15 | 93.3 |
| | | 2 | 0 | 15 | 0 | 0 | 15 | 100 |
| | | 3 | 1 | 0 | 10 | 3 | 14 | 71.4 |
| | | 4 | 1 | 0 | 0 | 14 | 15 | 93.3 |

[a] One individual from farm 2 could not be classified according to PNN analysis.

**Table 4**
Results of the classification of farmed fish according to the GC and NMR data and using SVM analysis: number of fish allocated to each farm.

| | Actual farm of origin | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | Not 1 | 2 | Not 2 | 3 | Not 3 | 4 | Not 4 |
| *GC data* | | | | | | | | |
| Correctly classified | 15 | 44 | 14 | 44 | 14 | 43 | 14 | 45 |
| Wrongly classified | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 |
| Correctly classified (%) | 100 | | 98.31 | | 96.61 | | 100 | |
| *NMR data* | | | | | | | | |
| Correctly classified | 14 | 43 | 14 | 44 | 14 | 45 | 15 | 44 |
| Wrongly classified | 1 | 1 | 1 | | 0 | 0 | 0 | 0 |
| Correctly classified (%) | 96.6 | | 98.3 | | 100 | | 100 | |

from farm 2 were quite different from the others, showing a high degree of variability; samples from farm 1 were the most similar to farm 2 and the most scattered in the map, and samples from farms 3 and 4 clustered together at the other end of PC1.

Application of the chemometric techniques to the reference farmed fish showed very good results with the NMR data, although not as good as with the GC data, with correct classifications varying between 71.4% and 100% (Tables 3 and 4). Thus, PNN analysis misclassified a total of four fish: two fish from farm 1 (allocating them to farms 3 and 4); one specimen from farm 2 (misclassified in farm 1) and one from farm 4 (misclassified in farm 1). The BBN calculations gave the worst results with the reference fish: it misclassified about 10% of the individuals and only classified correctly all fish from farm 2 (Table 3). SVM classified correctly all fish from farms 3 and 4, but misclassified one fish to farm 2, another fish to farm 1 and one fish that belonged to farm 1 was classified as of other origin (Table 4).

The application of these models to the whole set of NMR data, both reference farmed and free-living fish, is shown in Table 2. As expected, none of the fish shown to be wild by GC analysis were attributed to any of the farms. There was consensus among all the analysis regarding fish 513, 515 and 517 as originating from farm 4; fish 508, 509 and 514 as escapes from farm 2, and 512 was assigned to farm 2 by all the analyses except BBN of NMR data that placed this fish in farm 4. Fish 503, 506 and 507 might have belonged to any of farms 3 or 4; 1 or 3 or either 1, 3 or 4, respectively. Alternatively, taking into account that there are approximately 30 farms in this fjord, it is possible that these fish may have escaped from some farm(s) that may use similar feeds to the ones tested here. Fish 501, 502 and 504 were identified as farmed, but not belonging to any of the four reference farms examined in this work.

In general, the ability to correctly classify the reference samples according to their farm of origin can be considered as excellent by most of these methods (see Tables 3 and 4), mainly SVM on NMR data, with almost 99% correct classification, followed by any of the statistical treatments on GC data (96.8–98.3% success) while NMR data in combination with either PNN or BBN gave the lowest number of correctly allocated reference samples (89.5–93.3% correct assignments). These figures represent a much higher number of correct assignments than the ones reported by Glover et al. (2008) on a related study on samples in the Romsdalfjord. Using microsatellite markers, these authors were able to classify correctly on average 62.5% of their reference samples, with correct assignments varying from 28% to 100% using microsatellite markers. Fortunately, the escaped fish in that work belonged to a tank with a characteristic genotype that allowed a 96% correct classifications for the reference individuals.

Salmon farmers may obtain their smolt from different producers although, in most instances, they do not mix smolt from different producers in the same cage. Different producers may sell smolts that are genetically almost identical while the genetic make up of smolts sold from a given producer may vary substantially (Glover et al., 2008). The feed received may also vary: although farmers may use only one feed producer, they may order different feeds or mixes depending on the size of the fish, the desired growth rate, price and time to slaughter. Similarly, feed producers may optimise their formulations according to the customer's requirements and price and availability of ingredients in the international market. The FA profile of the salmon is a reflection of its diet, and therefore a farmed fish that escapes will start changing its profile from the moment it escapes, but wild fish feeding on the rests of feed around cages will also change its characteristic wild profile and show phenotypic characteristics similar to farmed fish (Skog, Hylland, Torstensen, & Berntssen, 2003). Thus, during a period of time of about six months (Torstensen, Frøyland, Ørnsrud, & Lie, 2004) these two types of fish, escaped and opportunistic-wild, will present an intermediate FA profile in their triglyceride fraction and finally they will adopt the one reflecting its final diet. This is illustrated here by fish 511 and 516, that may have been farmed in origin but that had been long enough in the wild to completely change their profiles. GC analysis only provides information regarding the FA composition but the information contained in the HR $^{13}C$ NMR spectrum of the same sample is more complex and includes other lipidic compounds as well as the positional distribution of the FA in the glycerol molecule (Aursand et al., 1995). The additional information has probably a genetic component since it has been shown to be vary according to the stock (Grahl-Nielsen, 1997) and FA profiling of tryglcerides has been shown to be of value for species identification species (Medina, Auborg, & Martin, 1997).

The present work showed an apparently extraordinary ability of fatty acid profiling combined with chemometrics to perform correct assignments of farmed fish to their farm of origin, even for fish that had apparently very similar FA profiles, as those from farms 1 and 3. It is more difficult to explain the reason why NMR analysis was not on average as successful as GC but it may be that NMR provides additional information that confused the model. For example, genetic analysis gave on average a lower degree of correct classification than GC analysis, if the NMR spectra contain part of the genetic information that made fish from different farms similar, then this additional information will contribute to create a noise that will dilute the amount of discriminant information.

It must be kept in mind that there are approximately 30 farms in the fjord. Therefore, allocation of escapes to their origin would demand the analysis of reference fish from all the farms from the region. An increase in the number of samples contained in the reference database usually provides better classifications for unknown samples, at least when the samples are from very different origins (Aursand et al., 2007). Alternatively, if it turns out that the fish cultivated by different breeders in the same fjord (i.e., under almost identical environmental conditions) are genetically similar and the breeders also use the same feed, then increasing the number of fish analyzed may show an overlap of samples. The only way to elucidate this issue would be to carry out the work. Insufficient sampling of reference farms may be the reason for the inability to classify free-living specimens 501, 502 and 504, while the reason for the lack of consensus in the classification of the free-living fish 503, 506, 507 may be insufficient sampling of reference farms together with the change in the FA profile from the moment the fish changes the diet when escaping.

The accusation of being responsible for fish escaping is a very serious one and it may have severe economical consequences for the farmers. Taking into account that different analyses and different statistical treatment on the very same data gave occasionally different assignments (Table 2) and that there may always be some true wild fish feeding around farms that end up displaying an intermediate phenotype, we would recommend the application of several analytical techniques and chemometrics to ensure the reliability of the results. A clear example here is illustrated by sample 512: identified as a trout by visual examination, genetic analysis would also have identified correctly the species, but both GC and NMR analyses identified it as farmed, most probably from farm 2 or from another farm using the same diet. The opposite may also be true: if the escaping of fish can be kept hidden for a long enough period of time (about six months, depending on the water temperature), the escaped fish may end up acquiring the FA profile of wild fish and the escape will remain undetected by these analyses, as seems to be here the case for fish 511 and 516.

Finally, since both the genetic make up of the farmed fish and its feed varies with time (as already mentioned, breeders may receive different smolts and they also purchase feeds of variable composition), the correct classification of escaped fish would require that the authorities construct and continuously update a database containing the genetic and phenotypic (FA) profiles of all the fish cultivated in the area, so that if a escape takes place, it would be possible to compare the escaped fish to all the possible donors (in the given time-space window) and so identify their most likely origin.

In conclusion, we believe that the analysis presented here will be of great value to identify farmed and wild fish and also to trace back farmed fish to their farm of origin. Feed and veterinary treatments have been identified in a parallel study as the most critical steps where undesirable substances can enter the production chain of farmed salmon (*ΣChain*, EU Strep Project FP6-FOOD-5184). The analytical methods described here together with genetic analysis as described by Glover et al. (2008) will not only aid the authorities to ensure correct consumer information and identify escapes; they will also aid the stakeholders of the farmed salmon chain to trace back the farm and tank of provenance of fish involved in incidents.

## Acknowledgments

The financial support of the Norwegian Research Council/Ministry of Fisheries (TRACES: Tracing escaped salmon by means of naturally occurring DNA markers, fatty acid profiles, trace elements and stable isotopes, Project Number 172628/S40) and of the European Community (Developing a Stakeholders' Guide on the vulnerability of food and feed chains to dangerous agents and substances: *ΣChain*. EU Strep Project FP6-FOOD-518451) are gratefully acknowledged. Marthe Schei and Merete Selnes are also gratefully acknowledged for the technical assistance. We are indebted to our colleagues from the Institute of Marine Research (Bergen) and in particular Dr. Øystein Skaala for coordinating the project and Dr. Kevin Glover for the collection of free-living fish. Hardanger Fish Health Network is also acknowledged for their help in organising the collection of fish from the four fish farms tested.

## References

Aursand, M., Jørgensen, L., & Grasdalen, H. (1995). Positional distribution of ω3 fatty acids in marine lipid triacylglycerols by high-resolution $^{13}C$ nuclear magnetic resonance spectroscopy. *Journal of the American Oil Chemists Society, 72*, 293–297.

Aursand, M., Mabon, F., & Martin, G. J. (2000). Characterization of farmed and wild salmon (*Salmo salar*). *Journal of the American Oil Chemist Society, 77*, 659–666.

Aursand, M., Standal, I. B., & Axelson, D. E. (2007). High-resolution C-13 nuclear magnetic resonance spectroscopy pattern recognition of fish oil capsules. *Journal of Agricultural and Food Chemistry, 55*, 38–47.

Bell, J. G., McEvoy, J., Tocher, D. R., McGhee, F., Campbell, P. J., & Sargent, J. R. (2001). Replacement of fish oil with rapeseed oil in diets of Atlantic salmon (*Salmo salar*) affects tissue lipid compositions and hepatocyte fatty acid metabolism. *Journal of Nutrition, 131*, 1535–1543.

Bligh, E. G., & Dyer, W. J. (1959). A rapid method for total lipid extraction and purification. *Canadian Journal of Biochemistry and Physiology, 37*, 911–917.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297.

Coughlan, J. P., Imsland, A. K., Galvin, P. T., Fitzgerald, R. D., Naevdal, G., & Cross, T. F. (1998). Microsatellite DNA variation in wild populations and farmed strains of turbot from Ireland and Norway: A preliminary study. *Journal of Fish Biology, 52*, 916–922.

Cristiani, N., & Shawe-Taylor, J. (2000). *An introduction to machine learning*. Cambridge: Cambridge University Press.

Glover, K. A., Skilbrei, O. T., & Skaala, Ø. (2008). Genetic assignment identifies farm of origin for Atlantic salmon *Salmo salar* escapees in a Norwegian fjord. *ICES Journal of Marine Science, 65*, 912–920.

Grahl-Nielsen, O. (1997). Fatty acid profiles as natural markers for stock identification. *ICES CM1997/M:2; Annex 3*.

Heckerman, D., Geiger, D., & Chickering, D. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning, 20*, 197–243.

Jolliffe, I. T. (1986). *Principal component analysis*. New York: Springer-Verlag.

Lee, G. C., & Woodruff, D. L. (2004). Beam search for peak alignment of NMR signals. *Analytica Chimica Acta, 513*, 413–416.

Lund, R. A., & Hansen, L. P. (1991). Identification of wild and reared Atlantic salmon, *Salmo salar* L., using scale characters. *Aquaculture and Fisheries Management, 22*, 499–508.

Martinez, I. (2006). *Evaluation of the profile of lipids as a tool to discriminate wild from farmed salmon*. SEAFOODplus publication series report, 6.3.8. ISBN 978-87-7075-001-1.

Martinez, I. (in press) Authenticity assessment based on other principles: Analysis of lipids, stable isotopes and trace elements. In J. Oehlenschläger & H. Rehbein (Eds.), *Fishery products: Quality, safety and authenticity*. Blackwell Publishing.

Masoum, S., Malabat, C., Jalali-Heravi, M., Guillou, C., Rezzi, S., & Rutledge, D. N. (2007). Application of support vector machines to $^{1}H$ NMR data of fish oils: Methodology for the confirmation of wild and farmed salmon and their origins. *Analytical and Bioanalytica Chemistry, 387*, 1499–1510.

Medina, I., Auborg, S. P., & Martin, R. P. (1997). Species differentiation by multivariate analysis of phospholipids from canned Atlantic tuna. *Journal of Agricultural and Food Chemistry, 45*, 2495–2499.

Molkentin, J., Meisel, H., Lehmann, I., & Rehbein, H. (2007). Identification of organically farmed Atlantic salmon by analysis of stable isotopes and fatty acids. *European Food Research and Technology, 224*, 535–543.

Pearl, J. (1988). *Probabilistic reasoning for intelligent systems*. San Francisco: Morgan Kaufmann.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*, 1226–1238.

Primmer, C. R., Koskinen, K. T., & Piironen, J. (2000). The one that did not get away: Individual assignment using microsatellite data detects a case of fishing competition fraud. *Proceedings of the Royal Society of London, Series B, 267*, 1699–1704.

Skaala, Ø., Høyheim, B., Glover, K., & Dahle, G. (2004). Microsatellite analysis in domesticated and wild Atlantic salmon (*Salmo salar* L.): Allelic diversity and identification of individuals. *Aquaculture, 240*, 131–143.

Skaala, Ø., Taggart, J. B., & Gunnes, K. (2005). Genetic differences between five major domesticated strains of Atlantic salmon and wild salmon. *Journal of Fish Biology, 67*, 118–128.

Skog, T. E., Hylland, K., Torstensen, B. E., & Berntssen, M. H. G. (2003). Salmon farming affects the fatty acid composition and taste of wild saithe *Pollachius virens* L. *Aquaculture Research, 34*, 999–1007.

Specht, D. F. (1990). Probabilistic neural networks and the polynomial adaline as complementary techniques for classification. *IEEE Transactions on Neural Networks, 1*, 111–121.

Thomas, F., Jamin, E., Wietzerbin, K., Guerin, R., Lees, M., Morvan, E., et al. (2008). 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. *Journal of Agricultural and Food Chemistry, 56*, 989–997.

Torstensen, B., Frøyland, L., Ørnsrud, R., & Lie, Ø. (2004). Tailoring of a cardioprotective muscle fatty acid composition of Altantic salmon (*Salmo salar*) fed vegetable oils. *Food Chemistry, 87*, 567–580.