Fatty acid signatures and classification trees: new tools for investigating the foraging ecology of seals

Stephen J. Smith, Sara J. Iverson, and W.D. Bowen

Abstract: Analysis of the fatty acid composition of milk lipids in marine mammals offers a potential means of determining changes in diet and lactation stage. However, the large number of fatty acids routinely identified (over 60) relative to the number of animals usually sampled can limit the usefulness of standard multivariate statistical models for characterizing these patterns. Classification trees or tree-based models, which are not limited by the number of variables, were used here to study the fatty acid patterns in the milk of female harbour seals (*Phoca vitulina*) at parturition and during lactation. Tree analyses correctly classified 44 of 51 seals based on milk fatty acid composition to four stages of lactation, which corresponded to states of fasting versus increasingly intensive feeding. The fatty acid 16:2*n*–6 was quite effective in differentiating between seals at parturition and those 4 days or more later. Seals were grouped into early and late lactation by fatty acid 24:1*n*–9. A comparison between classification rules derived from classification trees and discriminant analysis showed that each gave similar rates of misclassification but that the latter required a method for the a priori choice of which fatty acids to analyze.

Résumé : L'analyse de la composition en acides gras des lipides du lait chez les mammifères marins constitue une méthode possible pour déterminer les changements dans le régime alimentaire et le stade de lactation. Cependant, le grand nombre d'acides gras (plus de 60) identifiés de manière habituelle par rapport au nombre d'animaux habituellement échantillonnés peut limiter l'utilité des modèles statistiques standard à variables multiples pour la caractérisation de ces compositions. Des arbres de classification ou des modèles arborescents, qui ne sont pas limités par le nombre de variables, ont été utilisés pour étudier les profils d'acides aminés dans le lait de phoques communs (*Phoca vitulina*) femelles au moment de la parturition et de la lactation. Les analyses par arborescence ont permis de classer correctement 44 des 51 phoques à partir de la composition du lait en acides gras pour quatre étages de lactation, qui correspondaient à des états de jeûne par opposition à des périodes d'alimentation de plus en plus intensives. L'acide gras 16:2*n*–6 était relativement efficace pour distinguer les phoques à la parturition des phoques qui en sont rendus à quatre jours ou plus après la parturition. Les phoques ont été regroupés en deux groupes (début et fin de la lactation) par l'acide gras 24:1*n*–9. Une comparaison entre les règles de classification dérivées des arbres de classification et de l'analyse de discrimination a révélé que chacune de ces méthodes donnaient des taux semblables d'erreurs de classification, mais que la dernière exige une méthode pour choisir a priori les acides aminés à analyser. [Traduit par la Rédaction]

Introduction

It has been recognized for some time that dietary fatty acids affected the fatty acid composition of the blubber lipids of baleen whales and phocid seals (e.g., Klem 1935; Ackman and Eaton 1966; Ackman et al. 1971). Building on these early observations, Iverson (1988, 1993) proposed that the pattern of fatty acids ("fatty acid signatures") in the blubber and milk of marine mammals might be used to determine changes in and the components of their diets. A recent study of the transfer of milk fatty acids from hooded seal (*Cystophora cristata*) mothers to their pups illustrates this approach (Iverson et al. 1995*a*). In this study, the female's milk was the "prey" and the pup the "predator". Milk fatty acid signatures can also be useful in studies of the foraging ecology of terrestrial carnivores, such as the black bear (*Ursus americanus*) (Iverson and Oftedal 1992).

Lipids in marine organisms are characterized by their great diversity and high levels of long-chain and polyunsaturated fatty acids that originate in various unicellular phytoplankton and seaweeds (Ackman 1980). Fatty acids are the most abundant constituent of common lipids. Unlike other nutrients, such as proteins and carbohydrates which are readily broken down during digestion, in monogastric animals, dietary fatty acids pass into the circulation intact and those of carbon chain length >14 can be deposited in animal tissue with little or no modification. Thus, it is possible to consider the pattern of fatty acids in the blubber of a pinniped as a signature that should reflect an integration of the fatty acid signatures of the major prey items in the diet (Iverson 1993).

The diversity of marine fatty acids (over 60 can be routinely identified) presents a problem for their statistical analysis. In most multivariate methods used (e.g., discriminant analysis, principal component analysis) to characterize patterns in the fatty acids and identify groups of prey, or the nutritional state of the animals, the number of animals sampled must exceed the number of variables. With over 60 variables, this puts a

Received May 14, 1996. Accepted November 29, 1996.
J13472

S.J. Smith[1] **and W.D. Bowen.** Department of Fisheries and Oceans, P.O. Box 1006, Bedford Institute of Oceanography, Dartmouth, NS B2Y 4A2, Canada.
S.J. Iverson. Department of Biology, Dalhousie University, Halifax, NS B3H 4J1, Canada.

[1] Author to whom correspondence should be addressed. e-mail: S_Smith@bionet.bio.dfo.ca

considerable burden on investigators to obtain extremely large numbers of samples. One approach to this problem has been to select a subset of major fatty acids (usually about 12–18) for multivariate analysis (e.g., Grahl-Nielsen and Mjaavatten 1991). The disadvantages of this approach are that fatty acid selection is subjective and information may be lost in the process. Others have attempted to avoid this problem by using univariate methods, but the probability of type 1 error increases rapidly given the large number of comparisons that must be made.

In contrast with these multivariate methods, classification trees or tree-based models (Breiman et al. 1984; Clark and Pregibon 1992) offer a means of characterizing patterns and identifying groups with as many fatty acids as available, even if this number exceeds the number of animals sampled. This method proceeds by recursively partitioning the subjects into two or more groups based on a series of dichotomous splits from a set of explanatory variables — fatty acids in this case. The entire set of fatty acids can be screened by this procedure to choose a subset that can be used to classify subjects into relatively homogeneous groups based on similarities in patterns of fatty acid proportions. The tree-based models are unaffected by spurious correlations, do not require a statistical distribution assumption for the observations, use computer power to test all of the fatty acids, and use a statistical criterion (change in deviance which is directly related to the log-likelihood) to choose a subset of fatty acids. In addition, the visual nature of the trees makes them easy to present and understand.

The harbour seal (*Phoca vitulina*), a member of the family Phocidae, is found in coastal marine habitats throughout the Northern Hemisphere. Pregnant harbour seal females haul out on Sable Island and give birth between mid-May and June after a period of intensive fattening in preparation for lactation. In contrast with other phocid species, the small body size of adult female harbour seals apparently precludes their ability to support all of lactation from stored energy (Bowen et al. 1992). Recent studies using time–depth recorders have shown that, after about 6 days postpartum, most females exhibit intermittent foraging bouts, which after about 11 days become increasingly intensive in both depth and duration as lactation progresses (Boness et al. 1994).

We collected milk samples from adult females at parturition and over the course of lactation and used the analysis of fatty acid signatures to investigate their foraging ecology. We hypothesized that milk samples taken at parturition and during initial fasting should reflect blubber stores and hence represent an integrated view of the diet prior to giving birth whereas samples taken after the onset of feeding trips should reflect current feeding (e.g., Iverson 1993; Iverson et al. 1995*a*, 1995*b*, 1997). However, the purpose of this paper is not to explicitly examine the foraging ecology of these females, but to illustrate the use of classification trees in the analysis of milk fatty acids, which together offer some promise as useful tools in such studies.

Methods

Milk samples were collected from female harbour seals during the May–June pupping seasons in 1990 and 1993 on Sable Island, Nova Scotia, Canada (43°55′N, 60°00′W). Each day, all newly born pups were individually tagged in the hind flipper such that the age of pups and hence stage of lactation of the females was known to within 24 h.

In 1990, females were captured on the day of parturition (Day 0) and at 4–7, 12–14, and 19–21 days postpartum. In 1993, females were captured on the day of parturition and 19–21 days postpartum. At each capture, females were sedated with valium (0.2 mg/kg body mass, intravenous injection into the extradural vein) and milked using syringe suction (Iverson et al. 1993). To facilitate milking, females were given an intramuscular injection of oxytocin (15–30 IU). Approximately 60-mL samples were obtained from each female. At the time of collection, a 0.5-mL aliquot of milk was placed in a Kimax tube containing 3 mL of 2:1 (v/v) chloroform–methanol containing 0.01% BHT (w/v) and stored frozen.

Fatty acid analyses

Lipids were extracted into chloroform according to the method of Folch et al. (1957) as modified by Iverson (1988) using the ratios of 18 parts 2:1 chloroform–methanol to 1 part sample and 3 parts solvent to 1 part aqueous salt (i.e., milk plus 0.7% NaCl). Fatty acid methyl esters were prepared from 100 mg of the pure extracted lipid (filtered and dried over anhydrous sodium sulfate) using 1.5 mL of 8% boron trifluoride in methanol (v/v) and 1.5 mL of hexane, capped under nitrogen, and heated at 100°C for 1 h. Fatty acid methyl esters were extracted into hexane, concentrated, and brought up to volume (50 mg/mL) with high-purity hexane. This method produced results identical to those using 0.5 N H_2SO_4 in methanol as transesterfying reagent.

Duplicate analyses of fatty acid methyl esters and their identifications were performed using temperature-programmed gas–liquid chromatography basically according to Iverson (1988) and Iverson et al. (1995*a*) on a Perkin Elmer Autosystem II Capillary FID gas chromatograph fitted with a 30-m column (0.25-mm inside diameter) coated with 50% cyanopropyl polysiloxane (0.25-μm film thickness; J&W DB-23; Folsom, Calif.) and linked to a computerized integration system (Turbochrom 4.0 software, PE Nelson). Individual fatty acids are expressed as mass percentage of total fatty acids and designated by shorthand IUPAC nomenclature of carbon chain length:number of double bonds, and location (*n*–*x*) of the double bond nearest the terminal methyl group. For example, a fatty acid with a carbon chain length of 16, one double bond, and the location of this double bond 7 carbons back from the terminal methyl group would be designated as 16:1*n*–7.

Classification and regression tree analysis

A total of 51 milk samples were collected from females sampled in 1990. These samples were classified into four categories based on the days postpartum when they were collected: 0 days ($n = 15$), 4–7 days ($n = 15$), 12–14 days ($n = 12$), and 19–21 days ($n = 9$). We called this our "training sample". We wanted to be able to predict which of these four categories a seal belonged to based only on the fatty acid composition of the milk sample from the animal.

Next, each of our predictor variables (i.e., fatty acids) was screened by the classification tree algorithm (see below) to see which would best classify observations into a particular category. For example, the proportion of one particular fatty acid in the milk might separate the majority of the seals in the first two day-categories from those in the latter two day-categories. Additionally, another fatty acid might be instrumental in separating seals in the Day 0 group from those in the Days 4–7 group whereas a different fatty acid may separate seals in the Days 12–14 and 19–21 categories. Thus, the procedure is best described as creating an inverted tree structure with the root node at the top and with the original observations represented as vector $y_0 =$ (15,15,12,9), where each position represents its respective day-category. For a specified fatty acid, observations in this vector will be classified as "travelling" down either the left branch or right branch leading away from the root node to other intermediate nodes. The decision on whether a particular seal is assigned to the right branch or

Table 1. Fatty acids in harbour seal milk samples from Sable Island, Nova Scotia (1990).

	Day-category			
Fatty acid	0 (n = 15)	4–7 (n = 15)	12–14 (n = 12)	19–21 (n = 9)
12:0	0.17	0.17	0.21	0.21
13:0	0.00	0.00	0.00	0.00
14:0	4.57	4.28	3.95	3.48
14:1*n*–9	0.17	0.17	0.20	0.20
14:1*n*–7	0.04	0.02	0.01	0.03
14:1*n*–5	0.32	0.43	0.45	0.55
Iso 15	0.24	0.18	0.23	0.21
Anti 15	0.20	0.12	0.14	0.13
15:0	0.38	0.36	0.35	0.31
15:1	0.02	0.01	0.01	0.03
Iso 16	0.18	0.11	0.10	0.11
16:0	15.65	14.43	13.06	11.92
16:1*n*–11	0.61	0.61	0.59	0.60
16:1*n*–9	0.45	0.39	0.39	0.44
16:1*n*–7	9.69	10.53	9.56	9.41
7Me16:0	0.34	0.31	0.28	0.24
16:1*n*–5	0.33	0.30	0.36	0.32
16:2*n*–6	0.14	0.10	0.06	0.07
16:2*n*–4	0.36	0.39	0.45	0.44
16:3*n*–6	0.47	0.45	0.38	0.31
16:3*n*–4	0.29	0.27	0.23	0.19
16:3*n*–1	0.22	0.19	0.21	0.21
16:4*n*–1	0.57	0.58	0.43	0.31
17:0	0.23	0.22	0.24	0.25
17:1	0.34	0.36	0.38	0.42
18:0	2.50	2.22	2.40	2.34
18:1*n*–13	0.06	0.05	0.03	0.00
18:1*n*–11	3.37	2.97	3.04	3.89
18:1*n*–9	13.49	13.71	14.56	16.94
18:1*n*–7	4.11	3.87	3.76	3.78
18:1*n*–5	0.71	0.64	0.62	0.57
18:2Δ5.7	0.25	0.22	0.21	0.25
18:2*n*–7	0.08	0.04	0.02	0.01
18:2*n*–6	1.36	1.35	1.48	1.62
18:2*n*–4	0.28	0.22	0.18	0.15
18:3*n*–6	0.26	0.20	0.20	0.20
18:3*n*–4	0.23	0.18	0.16	0.16
18:3*n*–3	0.47	0.58	0.67	0.72
18:3*n*–1	0.06	0.04	0.03	0.03
18:4*n*–3	1.33	1.39	1.40	1.23
18:4*n*–1	0.45	0.34	0.25	0.20
20:1*n*–11	1.21	1.24	1.38	1.72
20:1*n*–9	3.90	4.77	5.52	6.34
20:1*n*–7	0.34	0.34	0.44	0.46
20:1*n*–5	0.06	0.08	0.09	0.12
20:2*n*–9	0.02	0.03	0.02	0.00
20:2*n*–6	0.22	0.21	0.25	0.24
20:3NMIT	0.01	0.02	0.02	0.02
20:3*n*–6	0.09	0.05	0.05	0.06
20:4*n*–6	0.66	0.64	0.60	0.54
20:3*n*–3	0.04	0.06	0.08	0.08
20:4*n*–3	0.57	0.57	0.64	0.62
20:5*n*–3	9.95	8.56	7.30	5.97
21:5*n*–3	0.54	0.52	0.66	0.42
22:1*n*–11	1.32	1.48	2.06	2.01
22:1*n*–9	0.31	0.32	0.44	0.51
22:1*n*–7	0.01	0.00	0.03	0.02
22:2*n*–6	0.10	0.11	0.12	0.16
22:4*n*–6	0.07	0.07	0.09	0.08
22:5*n*–6	0.12	0.14	0.18	0.22
22:4*n*–3	0.07	0.10	0.10	0.14
22:5*n*–3	4.52	4.68	4.75	4.83
22:6*n*–3	10.27	12.40	13.11	12.08
24:1*n*–11	0.18	0.18	0.22	0.23
24:1*n*–9	0.22	0.21	0.32	0.33

Table 1 (*concluded*).

Note: Mean percentage composition by mass for each fatty acid according to day-category. Sample size (n) is given for each category.

Table 2. Results from setting the first node of the classification tree in different ways: (*a*) standard method of choosing explanatory variable (e.g., fatty acid) that gives the maximum change in deviance and (*b*–*d*) deliberately setting the first node to fatty acid 16:0, 7Me16:0, or 18:3*n*–4, respectively.

	No. of terminal nodes	Residual mean deviance	Misclassification error rate	Variables used in tree construction
(*a*)	8	0.6504	7/51	16:2*n*–6, 24:1*n*–9, 17:1, 16:3*n*–6, 18:2Δ5,7, 12:0, 16:0
(*b*)	8	0.6346	7/51	16:0[a], 18:2*n*–6, 16:3*n*–6, 18:2Δ5,7, 16:2*n*–6, 20:3*n*–6
(*c*)	7	0.7176	7/51	7Me16:0, 16:3*n*–6, 16:1*n*–7, 16:2*n*–6, 16:0, 14:1*n*–5
(*d*)	6	0.7907	8/51	18:3*n*–4, 7Me16:0, 16:3*n*–6, 12:0, 16:2*n*–6

[a]Used in two different locations on tree.

left branch will depend on whether the observed level for the specified fatty acid is above or below some cutoff value. At each of the intermediate nodes, another fatty acid is used to assign observations to even more left and right branches until a terminal node is reached. Ideally, a terminal node will have observations belonging to one group only, e.g., y_k = (15,0,0,0). The decisions and branches taken to get to this node will define, in this case, seals in the Day 0 category.

Which predictor variable (fatty acid) and what cutoff value is used is determined for each node by calculating the maximum change in deviance:

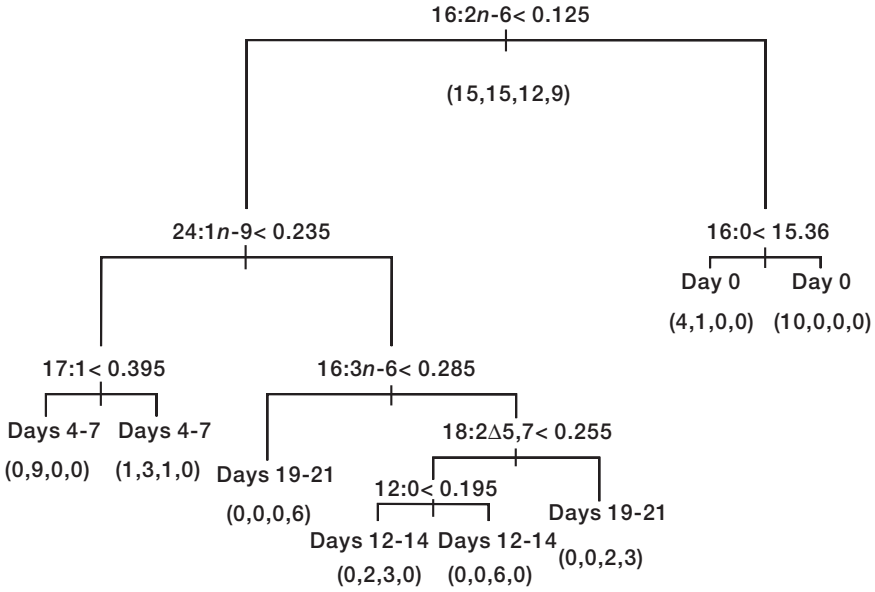
$$(1) \quad D(\mu; y) - D(\hat{\mu}_L, \hat{\mu}_R; y).$$

The term on the left corresponds to the deviance at the root node and the term on the right is the sum of the deviances for the left ($\hat{\mu}_L$) and right branches ($\hat{\mu}_R$) or splits. The deviance function is defined as minus twice the log-likelihood function (Clark and Pregibon 1992):

$$D(\mu; y) = -2 \sum_{i=1}^{n'} \sum_{k=1}^{K} y_{ik} \log(p_{ik});$$

K groups, n' observations in the node.

Fig. 1. Initial classification tree for determining days postpartum from fatty acids measured in harbour seal milk samples collected in 1990. The length of descending branches is proportional to change in deviance (see text). The fatty acid and the cutpoint are given for each node in the tree, with the less than sign referring to the left-hand decision. The right-hand decision was made for all values of the fatty acid greater than or equal to the cutpoint. Entries under the first node refer to the original number of seals in each of the four categories (Day 0, Days 4–7, Days 12–14, Days 19–21). Entries under each of the terminal nodes refer to the original categories of the seals assigned to the category given as the name of the terminal node.



We used the tree functions in S-PLUS (version 3.3 for Windows (1995), Statsci Division, MathSoft Inc., Seattle, Wash.) to conduct our analysis (see also Venables and Ripley 1994). The default criteria in S-PLUS for terminating branching are a change in deviance of <1% of the root node deviance or when the minimum number of observations at a node was <10. These stopping rules tend to be conservative and thus, "pruning" back of the tree is often necessary (see Results) to prevent overfitting of variables. The stopping rules used here are not the only ones being used and alternatives are discussed in Breiman et al. (1984) and Venables and Ripley (1994).

Misclassification rates for the final tree are generally calculated by predicting the categories for the same data that were used to construct the tree. Such estimates, referred to as resubstitution error rates, usually underestimate the true error rate. Therefore, we also used the *k*-fold cross-validation and leave-one-out methods. Misclassification rates using *k*-fold cross-validation are estimated by dividing the *n* observations randomly into *k* roughly equal-sized groups and using the data from $k - 1$ of the groups to construct the tree. This tree is then used to predict the categories for the one group of data held back. In the related leave-one-out method, the tree is constructed with $n - 1$ observations and used to predict the category of the one observation that was held back. The difference between the two methods is that the cross-validation estimate is usually applied to one random realization of *k* groups whereas the leave-one-out method constructs *n* trees leaving each observation out in turn.

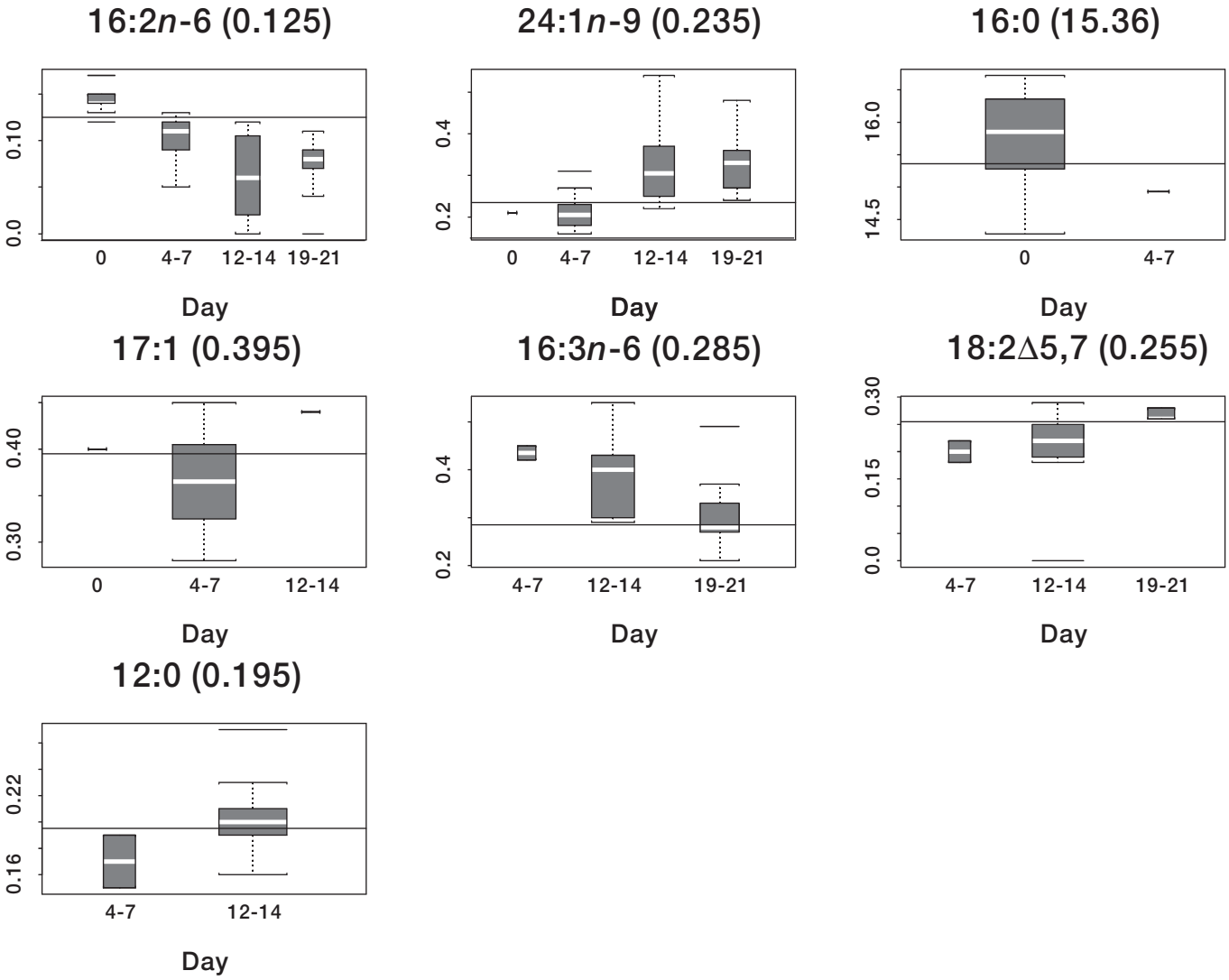
Results

In all of the 1990 samples, 50 of the 65 milk fatty acids quantified each accounted for <1% by mass of the total fatty acids in the milk (Table 1). The dominant fatty acids (>2%) were 14:0, 16:0, 16:1*n*–7, 18:0, 18:1*n*–11, 18:1*n*–9, 18:1*n*–7, 20:1*n*–9, 20:5*n*–3, 22:1*n*–11, 22:5*n*–3, and 22:6*n*–3. These 12 fatty acids in total accounted for 83.3, 84.0, 83.1, and 83.0%, respectively, of the total fatty acids for all four day-categories.

Classification of the milk samples from 1990 into their day-categories required seven fatty acids, resulting in eight terminal nodes on the tree (Fig. 1). The residual mean deviance for this tree was 0.6504. Fatty acid 16:2*n*–6 was chosen by the tree algorithm for the first node based on maximum change in deviance (47.096). All milks with proportions of 16:2*n*–6 ≥0.125% were eventually classified as Day 0 categories. Fatty acid 16:0 was selected to differentiate between the 14 Day 0 milks and the one Days 4–7 sample that were directed down the right branch from the root node. On the main left branch of the tree, 24:1*n*–9 separated Days 4–7 samples from Days 12–14 and 19–21 samples. Fatty acid 17:1 was chosen to break out the one misclassified Day 0 and three Days 12–14 samples. Finally, three fatty acids, 16:3*n*–6, 18:2Δ5,7, and 12:0, were used to further separate the remaining Days 12–14 and 19–21 samples. This initial tree misclassified 7 out of the 51 cases.

The choices of cutpoints for the fatty acids used to construct the tree in Fig. 1 are illustrated in Fig. 2. The boxplots in the panels of this figure summarize the distributions of the observed proportions by day-category for each of the fatty acids used to construct the tree. In the top left panel the distribution of fatty acid 16:2*n*–6 for all 51 cases is given. The cutpoint of 0.125 was best for differentiating between the Day 0 samples and the rest of the day-categories with the minimal amount of misclassification. The middle top panel presents the distribution for fatty acid 24:1*n*–9 for those samples not assigned to the right branch and eventually classed as Day 0 (top right panel). Terminal nodes are named according to which of the categories assigned to the node were the most abundant. Similarly, the rest of the panels present the distributions only for those samples that were remaining to be tested for that specific fatty acid. With the exception of 16:0, the fatty acids used to

Fig. 2. Boxplots demonstrating how the cutpoints (full horizontal lines) were used to classify the seals according to the different categories of days postpartum for each of the fatty acids used in the initial tree plot in Fig. 1. The positions of the medians are represented as white horizontal bars within a box, and the upper and lower boundaries of the box indicate the upper (75th) and lower (25th) quartiles, respectively. Vertical lines from the boxes extend to 1.5 times the interquartile range or to the closest observed value. Disconnected horizontal lines are used to indicate extreme values.



construct the tree were minor components (<1%) of the total mass of all of the fatty acids.

The changes in deviance for all of the fatty acids as root node contenders are presented in Fig. 3. Whereas 16:2*n*–6 gave the largest change, three other fatty acids, 16:0, 7Me16:0, and 18:3*n*–4, also resulted in changes of deviance in excess of 40. Using any one of these three fatty acids at the root node resulted in similar classification and misclassification rates (Table 2).

The tree using 16:0 at the root node is the closest to the original tree (Fig. 1), having the same misclassification rate and a slightly lower residual mean deviance. Whereas the residual mean deviance reflects the goodness of fit of the whole tree, the trees are constructed one node at a time. Therefore, it is likely that the tree constructed recursively by having the largest change in deviance at each node will not explain the largest amount of deviance overall. However, the difference in residual mean deviance between the original tree and that with 16:0 as the first node is quite small and there is no theory yet to statistically test the difference in deviance between the two trees. The difference of 0.0158 is well within the 1% change used for the stopping rules. In comparing misclassification rates, the two trees are equivalent.

The stopping rules used to construct the initial tree were largely arbitrary, but commonly used, and may have resulted in overfitting of variables. Starting with the tree in Fig. 1, fatty acids were successively removed or "pruned" and both the residual mean deviance and the misclassification rate of the resultant tree were calculated. From our point of view, misclassification rate is a more important criterion than change in deviance for judging alternative pruned trees (Fig. 4). The error rate of 7/51 remained constant for trees with five, six, seven, and eight nodes and then increased for fewer nodes than these (Fig. 4, left panel). The "best" five-node tree (Fig. 4,

Fig. 3. Dot chart of the change in deviance for the optimal cutpoint for each of the fatty acids used in this study to determine the fatty acid used for the first node. Fatty acid 16:2*n*–6 had the largest change at 47.096 (upper left panel) and thus was algorithmically chosen for the first node in Fig. 1.

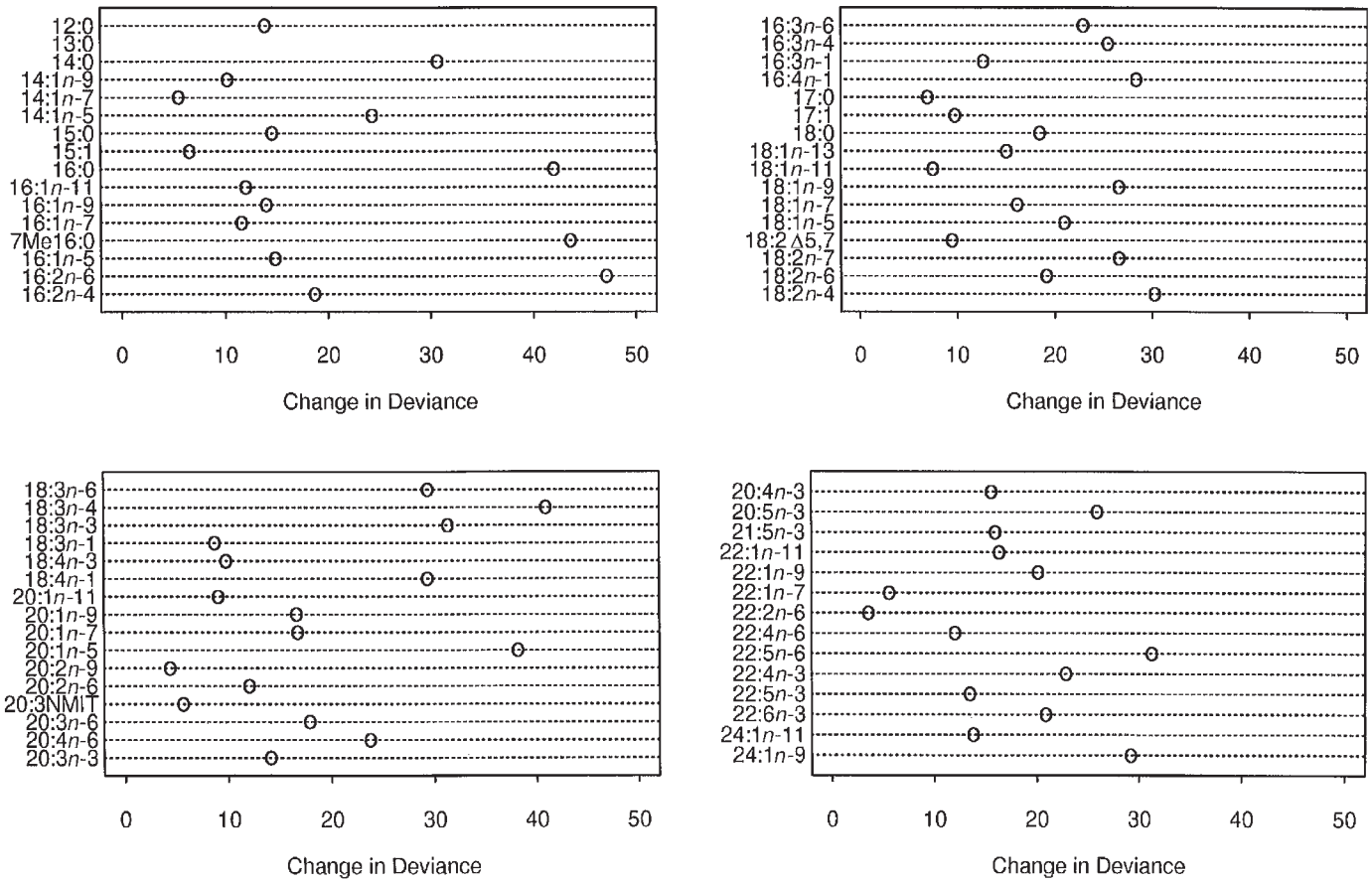


Table 3. Prediction results from applying the pruned tree from the 1990 data to the harbour seal sample collected in 1993.

			Fatty acids used in pruned tree			
Case			16:2*n*–6	24:1*n*–9	16:3*n*–6	18:2Δ5,7
No.	Observed	Predicted	0.125	0.235	0.285	0.255
1	Days 19–21	Days 12–14	0.11	0.26	0.37	0.20
2	Days 19–21	Days 12–14	0.11	0.49	0.68	0.14
3	Days 19–21	Days 12–14	0.11	0.45	0.55	0.18
4	Days 19–21	Days 4–7	0.09	0.07	0.38	0.14
5	Days 19–21	Days 12–14	0.12	0.28	0.37	0.19
6	Day 0	Day 0	0.18	0.19	0.46	0.41
7	Day 0	Day 0	0.19	0.21	0.37	0.31
8	Day 0	Day 0	0.18	0.2	0.48	0.26
9	Day 0	Day 0	0.17	0.17	0.46	0.24
10	Day 0	Day 0	0.24	0.18	0.60	0.32
11	Day 0	Day 0	0.15	0.15	0.47	0.21
12	Day 0	Day 0	0.21	0.2	0.46	0.31
13	Day 0	Day 0	0.15	0.16	0.42	0.24
14	Day 0	Day 0	0.19	0.21	0.54	0.27
15	Day 0	Day 0	0.15	0.11	0.43	0.23

right panel) required only four fatty acids to discriminate between the four day-categories. The mean residual deviance of this tree was 0.8427. Dropping 17:1, 12:0, and 16:0 from the original tree is easily understood when misclassification rates are considered. For example, 16:0 was used to separate the one Days 4–7 sample from the 14 Day 0 samples classified to the right branch from the first node. However, because node names reflect the major category assigned, the Days 4–7 sample was

Fig. 4. Results of pruning the tree in Fig. 1 by the number misclassified criterion. The left panel gives the number misclassified as a function of the number of terminal nodes in the tree. The right panel presents the optimal pruned tree, which has the same misclassification rate as the original tree but with only five terminal nodes instead of the eight in the original tree. Labeling of the tree as per Fig. 1.

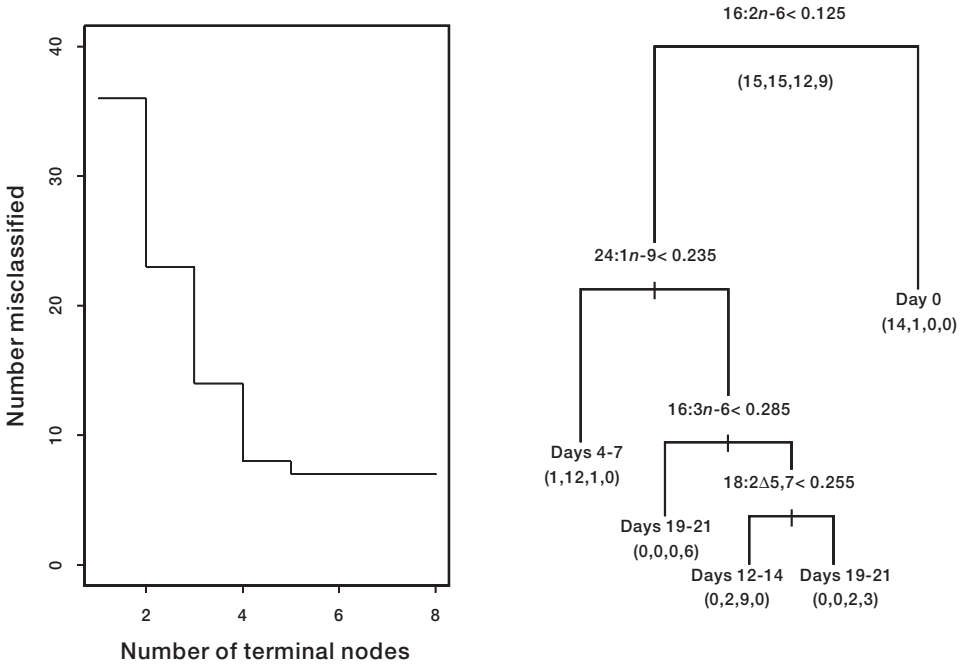
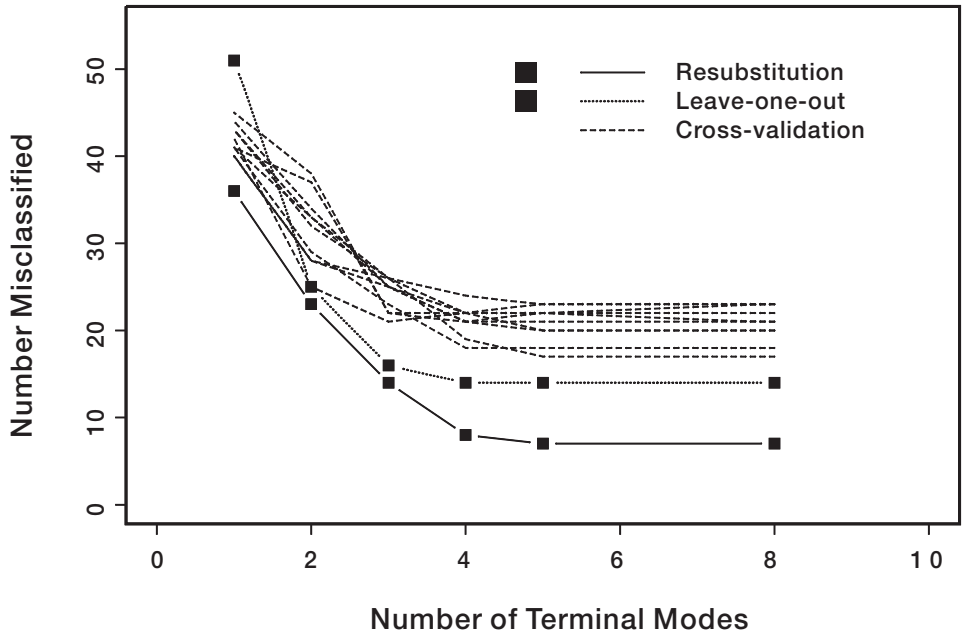


Fig. 5. Comparison of misclassification rates estimated for different-sized optimal pruned trees by the methods of resubstitution, leave-one-out, and 10-fold cross-validation.



still misclassified. Thus, these pruned fatty acids did not reduce the misclassification error rate and therefore were superfluous.

Misclassification rates estimated from cross-validation and leave-one-out methods were compared with resubstitution rates for a series of pruned trees in Fig. 5. The cross-validation rates represent 10 random realizations for the data divided into 10 groups. For the five-node tree, the leave-one-out estimate of 14/51 (0.275) misclassified is twice the resubstitution rate of 7/51 (0.137). However, the cross-validation rates range from 16 to 27 misclassified out of 51 (0.314–0.529).

All of these misclassification rates are meant to evaluate the future performance of the tree on a sample independent of that used to construct the tree. We applied the pruned tree given in Fig. 4 to fatty acid observations from 15 lactating harbour seals sampled on Sable Island in 1993. The misclassification rate

Table 4. Discriminant analysis results for three methods of choosing variables (fatty acids) to be included: (*a*) subset of fatty acids used by Grahl-Nielsen and Mjaavatten (1991) (stepwise discriminant analysis using Wilks Λ criterion for entering and removing variables; proportion of total variation explained by each discriminant axis = 0.87, 0.12, 0.01), (*b*) fatty acids from original tree in Table 2 and Fig. 1 (proportion of total variation explained by each discriminant axis = 0.91, 0.075, 0.015), and (*c*) fatty acids chosen from tree constructed on fatty acids used by Grahl-Nielsen and Mjaavatten (1991) (proportion of total variation explained by each discriminant axis = 0.88, 0.07, 0.05).

	Predicted			
Observed	Day 0	Days 4–7	Days 12–14	Days 19–21
(*a*) Misclassification rate = 6/51 (no. of fatty acids = 5)				
Day 0	15	0	0	0
Days 4–7	0	15	0	0
Days 12–14	0	1	9	2
Days 19–21	0	0	3	6
(*b*) Misclassification rate = 8/51 (no. of fatty acids = 7)				
Day 0	14	1	0	0
Days 4–7	2	13	0	0
Days 12–14	0	1	9	2
Days 19–21	0	0	2	7
(*c*) Misclassification rate = 12/51 (no. of fatty acids = 5)				
Day 0	13	2	0	0
Days 4–7	2	12	1	0
Days 12–14	0	2	9	1
Days 19–21	0	0	4	5

from applying the tree constructed from 1990 data to these 1993 data was 5/15 or 0.333, in the low end of the range predicted by the 10-fold cross-validation (Table 3). All of the Day 0 observations were correctly and consistently classified using 16:2*n*–6; however, the Days 19–21 samples were misclassed as either Days 4–7 or 12–14.

Linear discriminant analysis has also been used to classify tissue fatty acid samples (Grahl-Nielsen and Mjaavatten 1991). However, with fatty acid signatures, there are usually too many variables relative to the number of observations for discriminant analysis to be applied to the entire data set. Even stepwise variable selection for discriminant analysis requires the entire data set be used (Lachenbruch 1975). One approach has been to choose a subset of the most abundant fatty acids (Grahl-Nielsen and Mjaavatten 1991). We compared the results of stepwise linear discriminant analysis using Grahl-Nielsen and Mjaavatten's (1991) choice of the 18 most abundant fatty acids with those from the classification tree analysis. In addition, we investigated the use of the classification tree for choosing a subset of fatty acids for linear discriminant analysis by using either the seven fatty acids selected by the original tree (Table 2) or the five chosen by fitting a tree to the group of the 18 most abundant fatty acids.

The linear discriminant function based on Grahl-Nielsen and Mjaavatten's (1991) subset of fatty acids had the lowest misclassification error rate (resubstitution estimate = 6/51) of the three applications (Table 4*a*). The first of the five fatty acids chosen by the stepping algorithm (Wilks Λ) was 16:0, which was also chosen by the classification tree (Fig. 1) but was removed from the pruned tree (Fig. 4). Recall that deliberately choosing this fatty acid for the first node of a classification tree resulted in a tree with the same misclassification error as the trees in Figs. 1 and 4. None of the other fatty acids chosen by the stepwise linear discriminant analysis matched those chosen by the tree algorithm.

Table 5. Predictions from discriminant functions in Table 4 for the 1993 sample: (*a*) fatty acids chosen from subset given in Grahl-Nielsen and Mjaavatten (1991), (*b*) fatty acids from original tree in Table 2 and Fig. 1, and (*c*) fatty acids chosen from tree constructed on fatty acids from subset given in Grahl-Nielsen and Mjaavatten (1991).

	Predicted			
Observed	Day 0	Days 4–7	Days 12–14	Days 19–21
(*a*) Misclassification rate = 3/15				
Day 0	9	1	0	0
Days 19–21	1	0	1	3
(*b*) Misclassification rate = 2/15				
Day 0	10	0	0	0
Days 19–21	0	0	2	3
(*c*) Misclassification rate = 9/15				
Day 0	4	0	6	0
Days 19–21	0	1	2	2

The discriminant function for the seven fatty acids chosen by the original tree had a slightly higher misclassification rate of 8/51 (Table 4*b*). However, the discriminant analysis based on the five fatty acids chosen by the classification tree from 18 major fatty acids (Table 4*c*) had a higher misclassification error rate (12/51) than the classification tree (9/51) based on these same five fatty acids.

Application of all three discriminant functions to the 1993 sample showed that the first two had similar misclassification error rates whereas the third function had a considerably higher rate (Table 5). The first two discriminant functions had lower misclassification rates for the 1993 sample than the pruned tree (Table 3). On the other hand, the classification tree used to choose the fatty acids for the third discriminant function actually had a better misclassification rate (4/15) for the 1993 data than did the discriminant function based on these same fatty acids (9/15).

Discussion

Our results indicate that classification trees of milk fatty acids can be a useful tool in evaluating differences or changes in foraging patterns of individuals within a population, as has also been demonstrated in another species of seal (Iverson et al. 1997). In fact, it appears quite feasible to use milk fatty acids to estimate lactation stages in harbour seal populations where it may be difficult to tag pups at birth. The change from initial fasting to feeding in lactating harbour seal females has direct implications for differences in the uptake and secretion of milk fatty acids by the mammary gland, and these appear to be readily distinguished using classification trees.

In the lactating harbour seal, milk secreted during the initial 6 days postpartum should be derived almost entirely from blubber and thus represent an integration of diet during the fattening period prior to parturition. Milk secreted subsequently during foraging trip intervals should include fatty acid influx from the immediate diet (Iverson and Oftedal 1992,

1995; Iverson 1993; Iverson et al. 1995*a*, 1995*b*). If diet differs before and during lactation, then we would expect to detect changes in milk fatty acid composition. Indeed, in the females sampled in 1990, the tree analysis successfully classified most milk samples based on fatty acid composition into four lactation stages assigned by day of collection (Figs. 1 and 4). These results are consistent with our current understanding of the behaviour of female harbour seals during lactation, indicating that females begin foraging after about 6 days postpartum. We know that individual females may begin to feed as early as 2 days or as late as 14 days after parturition (Boness et al. 1994). Thus, we should expect to find variation among females, particularly during the Days 4–7 period, which may reflect interannual variation of the onset of foraging. Similarly, given that most females begin increasingly deeper and longer dive bouts after 11 days postpartum, indicating more intensive feeding, we would expect changes in fatty acids to be somewhat cumulative such that Days 12–14 and 19–21 milks would be more difficult to distinguish from one another, but to be readily distinguished from the fasting period. This was clearly the case in Fig. 4.

Based on lavage data obtained from Sable Island, harbour seal females are believed to feed predominantly on northern sand lance (*Ammodytes dubius*) throughout the 24-day lactation period (W. D. Bowen, unpublished data). Although little is known about the distribution of female harbour seals at other times of the year, tag recoveries suggest a rather wide distribution throughout the Scotian Shelf including the waters near Sable Island. Thus, there is reason to expect that the species composition of female diets may differ prior to and during the breeding season. Support for these differences is suggested by the observed changes in fatty acid composition of female milks. For instance, steady increases over lactation in components such as 20:1*n*–9, 22:1*n*–11, and 24:1*n*–9 may reflect the generally high levels of these fatty acids found in northern sand lance (unpublished observations).

Misclassification error rates calculated by the resubstitution method were more optimistic than the estimated rates produced by the 10-fold cross-validation and leave-one-out methods. However, the latter rates were closer to the observed rate of misclassification for the 1993 sample. Given the range of misclassification rates that can be obtained from the cross-validation method (Fig. 5), it may be more useful to calculate a number of realizations when evaluating the potential future performance of a tree.

A more relevant test of value of classification trees to the ecological analysis of milk fatty acid signatures was our ability to predict lactation stage (i.e., fasting versus feeding) of females in another year (3 years later). Despite potential interannual variation in diet, most milks from the 1993 season were correctly separated into fasting versus feeding based on the 1990 classification rules. The clear distinction of Day 0 milks from later foraging milks and the classification of Days 19–21 milks into "foraging milks" (mostly Days 12–14) is consistent with the foraging patterns of lactating females.

Several points are evident from our comparison of discriminant analysis with classification trees. First, when confined to the same group of fatty acids chosen by the original tree, the resubstitution misclassification rate for the discriminant function was not much better than that for the classification tree. However, this discriminant function did outperform the classification tree when predicting the categories for the 1993 tree. If this advantage was to hold in general, then classification trees may offer a means for choosing fatty acids for the linear discriminant analysis. However, our results from using the classification tree to choose fatty acids from the 18 given by Grahl-Nielsen and Mjaavatten (1991) suggest that discriminant analysis will not always outperform classification trees. Lynn and Brook (1991) compared the cross-validation misclassification error rates from classification trees and discriminant analysis for 12 data sets. They tentatively concluded that classification trees performed as well or better for the cases where there were large data sets with complex structure or heterogeneous covariance structures. In such cases, the assumptions of linear discriminant analysis may often be violated. Discriminant analysis outperformed classification trees for the smaller data sets studied by Lynn and Brook (1991).

The discriminant analysis based on the fatty acids used in Grahl-Nielsen and Mjaavatten (1991) provided a resubstitution misclassification error rate similar to that of the pruned tree. Only one of the 18 major fatty acids, 16:0, given by Grahl-Nielsen and Mjaavatten (1991) was chosen by the algorithm for the original tree. In fact, it was one of the less abundant fatty acids, 16:2*n*–6, that by itself was so effective in classifying the Day 0 samples for both the 1990 and the 1993 data sets. This serves to underscore the potential loss of information that could result from analysing only the most abundant fatty acids.

Acknowledgements

The authors thank Jeff Smith for computing assistance and other contributions. B.D. Ripley (Oxford) offered useful comments on the tree functions. The S-PLUS discriminant analysis function **lda( )** and the misclassification-based **prune.tree( )** were from B. D. Ripley's public domain S-PLUS library. R.G. Ackman provided general laboratory support as well as identifications of some of the unusual fatty acids. R. Stewart, Anne York, and an anonymous referee provided useful comments on the manuscript. This study was supported by NSERC Strategic Grant STR0133825, an NSERC international postdoctoral fellowship to S.J. Iverson, and the Department of Fisheries and Oceans.

References

Ackman, R.G. 1980. Fish lipids, part 1. *In* Advances in fish science and technology. *Edited by* J.J. Connell. Fishing News Books, Ltd., Surrey, U.K. pp. 86–103.

Ackman, R.G., and Eaton, C.A. 1966. Lipids of the fin whale (*Balaenoptera physalus*) from north Atlantic waters. III. Occurrence of eicosenoic and docosenoic fatty acids in the zooplankter *Meganyctiphanes norvegica* (M. Sars) and their effect on whale oil composition. Can. J. Biochem. **44**: 1561–1566.

Ackman, R.G., Epstein, S., and Eaton, C.A. 1971. Differences in the fatty acid compositions of blubber fats from northwestern Atlantic fin whales (*Balaenoptera physalus*) and harp seals (*Pagophilus groenlandica*). Comp. Biochem. Physiol. **408**: 683–697.

Boness, D.J., Bowen, W.D., and Oftedal, O.T. 1994. Evidence of a maternal foraging cycle resembling that of otariid seals in a small phocid, the harbor seal. Sociobiology, **34**: 95–104.

Bowen, W.D., Oftedal, O.T., and Boness, D.J. 1992. Mass and energy transfer during lactation in a small phocid, the harbor seal (*Phoca vitulina*). Physiol. Zool. **65**: 844–846.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. 1984. Classification and regression trees. Wadsworth International, Belmont, Calif.

Clark, L.A., and Pregibon, D. 1992. Tree-based models. *In* Statistical models in S. *Edited by* J.M. Chambers and T.J. Hastie. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, Calif. pp. 377–419.

Folch, J., Lees, M., and Sloane-Stanly, G.H. 1957. A simple method for the isolation and purification of total lipids from animal tissues. J. Biol. Chem. **226**: 497–509.

Grahl-Nielsen, O., and Mjaavatten, O. 1991. Dietary influence on fatty acid composition of blubber fat of seals as determined by biopsy: a multivariate approach. Mar. Biol. **110**: 59–64.

Iverson, S.J. 1988. Composition, intake and gastric digestion of milk lipids in pinnipeds. Ph.D. thesis, University of Maryland, College Park, Md.

Iverson, S.J. 1993. Milk secretion in marine mammals in relation to foraging: can milk fatty acids predict diet? Symp. Zool. Soc. Lond. **66**: 263–291.

Iverson, S.J., and Oftedal, O.T. 1992. Fatty acid composition of black bear (*Ursus americanus*) milk during and after the period of winter dormancy. Lipid, **27**: 940–943.

Iverson, S.J., and Oftedal, O.T. 1995. Phylogenetic and ecological variation in the fatty acid composition of milks. *In* Handbook of milk composition. *Edited by* R.G. Jensen. Academic Press, San Diego, Calif. pp. 789–827.

Iverson, S.J., Bowen, W.D., Boness, D.J., and Oftedal, O.T. 1993. The effect of maternal size and milk output on pup growth in grey seals (*Halichoerus grypus*). Physiol. Zool. **66**: 61–88.

Iverson, S.J., Oftedal, O.T., Bowen, W.D., Boness, D.J., and Sampugna, J. 1995*a*. Prenatal and postnatal transfer of fatty acids from mother to pup in the hooded seal. J. Comp. Physiol. B Biochem. Syst. Environ. Physiol. **165**: 1–12.

Iverson, S.J., Hamosh, M., and Bowen, W.D. 1995*b*. Lipoprotein lipase activity and its relationship to high milk fat transfer during lactation in grey seals. J. Comp. Physiol. B Biochem. Syst. Environ. Physiol. **165**: 384–395.

Iverson, S.J., Arnould, J.P.Y., and Boyd, I.L. 1997. Milk fatty acid signatures indicate both major and minor shifts in the diet of lactating Antarctic fur seals. Can. J. Zool. **75**: 188–197.

Klem, A. 1935. Studies in the biochemistry of whale oils. Hvalradets Skr. Nr. **11**: 49–108.

Lachenbruch, P.J. 1975. Discriminant analysis. Hafner Press, New York.

Lynn, R.D., and Brook, R.J. 1991. Classification by decision trees and discriminant analysis. N.Z. Stat. **26**: 18–26.

Venables, W., and Ripley, B. 1994. Modern applied statistics with S-Plus. Springer-Verlag, New York.