Comparisons of Likelihood and Machine Learning Methods of Individual Classification

B. Guinand, A. Topchy, K. S. Page, M. K. Burnham-Curtis, W. F. Punch, and K. T. Scribner

Classification methods used in machine learning (e.g., artificial neural networks, decision trees, and *k*-nearest neighbor clustering) are rarely used with population genetic data. We compare different nonparametric machine learning techniques with parametric likelihood estimations commonly employed in population genetics for purposes of assigning individuals to their population of origin ("assignment tests"). Classifier accuracy was compared across simulated data sets representing different levels of population differentiation (low and high F_{ST}), number of loci surveyed (5 and 10), and allelic diversity (average of three or eight alleles per locus). Empirical data for the lake trout (*Salvelinus namaycush*) exhibiting levels of population differentiation comparable to those used in simulations were examined to further evaluate and compare classification methods. Classification error rates associated with artificial neural networks and likelihood estimators were lower for simulated data sets compared to *k*-nearest neighbor and decision tree classifiers over the entire range of parameters considered. Artificial neural networks only marginally outperformed the likelihood method for simulated data (0–2.8% lower error rates). The relative performance of each machine learning classifier improved relative likelihood estimators for empirical data sets, suggesting an ability to "learn" and utilize properties of empirical genotypic arrays intrinsic to each population. Likelihood-based estimation methods provide a more accessible option for reliable assignment of individuals to the population of origin due to the intricacies in development and evaluation of artificial neural networks.

In recent years, characterization of highly polymorphic molecular markers such as mini- and microsatellites and development of novel methods of analysis have enabled researchers to extend investigations of ecological and evolutionary processes below the population level to the level of individuals (e.g., Bowcock et al. 1994; Estoup and Angers 1998; Jarne and Lagoda 1996). Analyses of individual-based genotypic information could substantially improve our understanding of evolutionary phenomena and contribute to effective management of natural populations (review in Bernatchez and Duchesne 2000). The use of individual-based methods remained largely unexplored in animal populations until recently due to a lack of highly polymorphic markers (Bernatchez and Duchesne 2000; Smouse and Chevillon 1998). Traditional analytical methods in population genetics rely almost exclusively on descriptors of genetic characterizations of populations (Bernatchez and Duchesne 2000) and not on individual genotypes.

"Assignment tests" are designed to determine population membership for individuals. One particular application based on a likelihood estimate (LE) was introduced by Paetkau et al. (1995; see also Vásquez-Domínguez et al. 2001) to assign an individual to the population of origin on the basis of multilocus genotype and expectations of observing this genotype in each potential source population. The LE approach can be implemented statistically in a Bayesian framework as a convenient way to evaluate hypotheses of plausible genealogical relationships (e.g., that an individual possesses an ancestor in another population) (Dawson and Belkhir 2001; Pritchard et al. 2000; Rannala and Mountain 1997). Other studies have evaluated the confidence of the assignment (Almudevar 2000) and characteristics of genotypic data (e.g., degree of population divergence, number of loci, number of individuals, number of alleles) that lead to greater population assignment (Bernatchez and Duchesne 2000; Cornuet et al. 1999; Haig et al. 1997; Shriver et al. 1997; Smouse

From the Departments of Fisheries and Wildlife (Guinand, Page, and Scribner) and Computer Science and Engineering (Topchy and Punch), Michigan State University, East Lansing, MI 48824; and the USGS Great Lakes Science Center, 1451 Green Rd., Ann Arbor, MI 48105 (Burnham-Curtis). Bruno Guinand is currently at UMR CNRS 5000 Génome, Populations, Interactions, Station Méditerranéenne de l'Environnement Littoral, 1, Quai de la Daurade, F34200 Sète, France. Kevin S. Page is currently at the Minnesota Department of Natural Resources, Division of Fisheries, 1601 Minnesota Dr., Brainerd, MN 56401. Mary K. Burnham-Curtis is currently at the U.S. Fish and Wildlife Service, National Fish and Wildlife Forensics Laboratory, 1490 East Main St., Ashland, OR 97520. We are grateful to the reviewers, who improved the quality of the results presented here. We thank the U.S. Fish and Wildlife Service for providing samples of the lake trout strains used in this study. This study was supported by the Great Lakes Fishery Trust, Great Lakes Protection Fund, and Partnership for Ecosystem Research and Management (PERM) between the Michigan Department of Natural Resources and Michigan State University. Address correspondence to Kim T. Scribner at the address above, or e-mail: scribne3@pilot.msu.edu.

 93:260–269

and Chevillon 1998). Main statistical and conceptual differences between methods leading to the use of an assignment test are given in, for example, Cornuet et al. (1999) and Rosenberg et al. (2001). However, the relative power of those tests has certainly not been fully appreciated and empirical comparisons are scarce (Eldridge et al. 2001). Assignment tests can also be considered as surrogates at the individual level (sensu Hansen et al. 2001a) for other statistical tools developed earlier, such as mixed-stock analysis (e.g., Pella and Masuda 2001; Pella and Milner 1987). Detailed theoretical comparison of the interests and limitations of both methods are still lacking, but empirical studies have revealed correlations between outputs of methods (Knutsen et al. 2001; Potvin and Bernatchez 2001).

Assignment tests have been widely used in different applications, including determination of degree of population differentiation or to establish the relationship among individuals within and among various taxonomic groupings (e.g., Bogdanowicz et al. 1997; Koskinen et al. 2001; Marshall et al. 2000; Müller 2000; Neraas and Spruell 2001; Nielsen et al. 2001b; Polzhien et al. 2000; Primmer et al. 1999; Roeder et al. 2001; Roques et al. 1999; Schulte-Hostedde et al. 2001; Sefc et al. 2000; Spidle et al. 2001; Vásquez-Domínguez et al. 2001), including hybrids (e.g., Beaumont et al. 2001; Congiu et al. 2001; Randi et al. 2001), introgressed individuals (e.g., Martinez et al. 2001; Randi and Lucchini 2002), and ecotypes (e.g., Taylor et al. 2000). Applications of assignment tests also include [human] forensics (e.g., Evett and Weir 1998; Primmer et al. 2000), identification and/or source of dispersers (e.g., Davies et al. 1999; Eldridge et al. 2001; Galbusera et al. 2000; Petersson et al. 2001; Tsutsui et al. 2001; Vasemägi et al. 2001), phylogeographical analyses (e.g., King et al. 2001; Zeisset and Beebee 2001), and the evaluation of the contribution of stocked individuals to natural populations (e.g., Fritzner et al. 2001; Hansen et al. 2000, 2001b) and of supportive breeding programs (Nielsen et al. 2001a; Olsen et al. 2000). Fish are among the organisms that have received considerable attention using such tools (see Hansen et al. [2001a] for a review). Moreover, these techniques are now used for profiles of traits outside the limited scope of population genetics (Thorrold et al. 2001).

Methods of classification vary widely based on several criteria (e.g., Jain et al. 2000) (Figure 1). Two basic classification

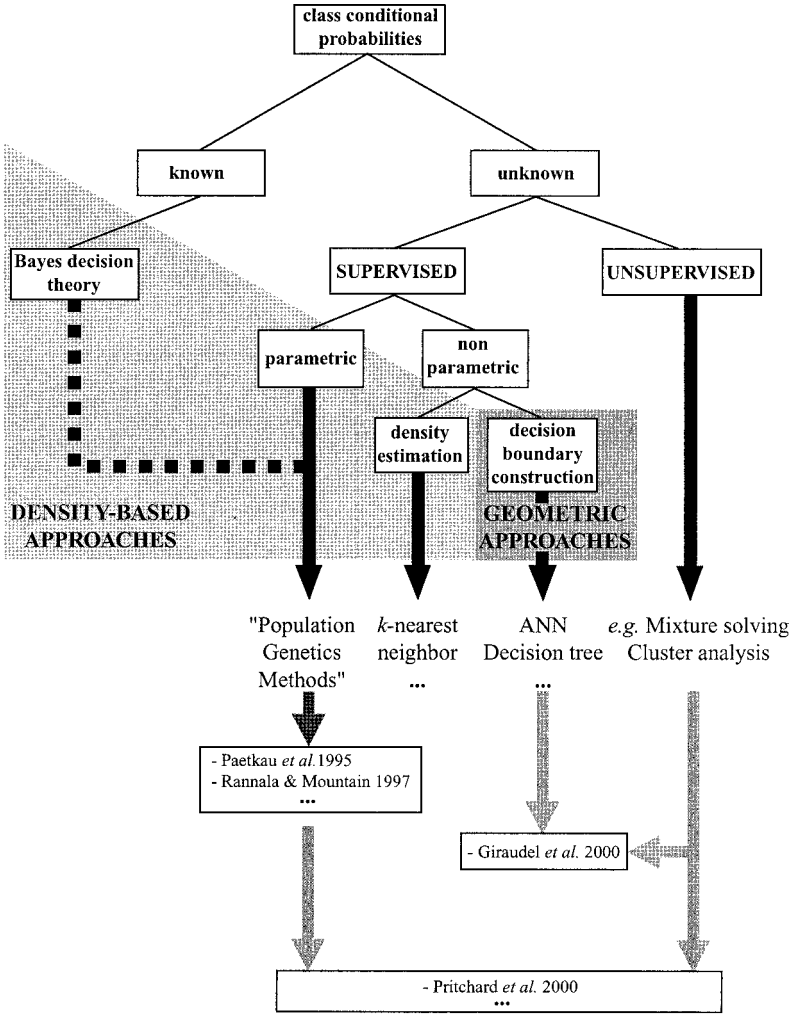

Figure 1. Diagrammatic representation of different methods used in statistical pattern recognition to build classifiers (adapted from Jain et al. 2000). The distinction between supervised and unsupervised learning is shown. Rectangles connected by two arrows indicate methods that can be implemented as supervised or unsupervised.

processes are traditionally recognized in machine learning: *supervised* classifiers and *unsupervised* classifiers (Figure 1; e.g., Duda et al. 2000; Jain et al. 2000). Supervised classifiers represent a group of methods whereby individual assignment is made to predefined classes (i.e., populations of origin). Unsupervised classification classes are unknown and are defined a posteriori on the basis of the degree of difference or similarity in attributes characterized from sampled individuals. Clustering methods (e.g., multidimensional scaling, principal component analysis) are examples of unsupervised classification.

Applications of assignment testing in population genetics first used supervised parametric likelihood-based approaches (Figure 1). Other machine learning classification methods are widely used in the physical and social sciences and in other biological disciplines (e.g. Boddy et al. 2000; Leung and Tran 2000; Manel et al. 1999; Raymer et al. 1997). Artificial neural networks (ANNs) are a popular technique used in machine learning (e.g., Boddy and Morris 1999; Duda et al. 2000; Lek and Guégan 2000; Ripley 1996). However, while recognized (Hansen et al. 2001a), ANN methods rarely have been employed for population genetics applications (Aurelle 1999; Aurelle et al. 1999; Cornuet et al. 1996; Curtis et al. 2001; Giraudel et al. 2000; Grigull et al. 2001; Taylor et al. 1994; Whitler et al. 1994). Other popular classification methods in machine learning, such as decision trees (e.g., Bell 1996, 1999; Duda et al. 2000; Mitchell 1997) and *k*-nearest neighbor analysis (*k*-NN; e.g., Dasarathy 1991; Duda et al. 2000) have yet to be applied in population genetics (Figure 1). Moreover, there has not been a directed effort to compare machine learning methodologies with the likelihood-based procedures widely used in population genetics. Cornuet et al. (1996) compared the

relative merits of ANNs to discriminant analysis in an empirical study involving different populations and subspecies of honeybee (*Apis mellifera*). However, they did not compare LE and ANN supervised classifiers. Aurelle (1999) used the approach of Rannala and Mountain (1997) (Figure 1) and ANN analysis using brown trout (*Salmo trutta*) microsatellite data; however, he did not provide a direct comparison of classification results or accuracies. Hansen et al. (2001a) briefly presented ANNs, but rejected their use without really testing their ability to classify individuals.

The objective of this article is to describe several of the more widely used machine learning classifiers that may have utility when used with empirical population genetics data. We compare likelihood-based "assignment tests" (Paetkau et al. 1995) with supervised machine learning classifiers including ANN, decision tree, and a *k*-NN clustering. Simulations were conducted which estimated and compared the assignment accuracy associated with different classifiers using ranges of parameter values (number of loci, allelic diversity, and interpopulation variance in allele frequency) typically encountered in natural populations. Comparative analyses were extended to empirical examples using lake trout (*Salvelinus namaycush*; Salmonidae).

Background of Machine Learning Classifiers

Classification is a fundamental activity in systematic biology (e.g., Wiens 1999), population genetics, and many other disciplines. Classification is performed by measuring traits or features that occur in different states for different individuals. Supervised classifiers (Figure 1) are able to evaluate patterns in different features and place individuals into one population or another. Development of assignment tests with supervised classifiers involves collection and evaluation of a baseline (training) data set and a testing data set (e.g., Duda et al. 2000). The success of a classifier is influenced by the properties of the baseline data set (i.e., in the context of genotypic or haplotypic data arrays, the number of individuals, number of loci, allelic diversity, and levels of population differentiation reflecting the degree of overlap in data distribution). Classification accuracy can be measured by the total number of unknown individuals correctly classified to their population of origin.

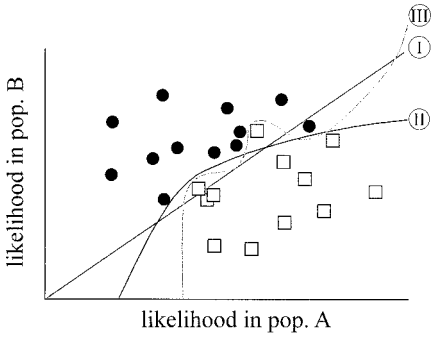

Figure 2. Diagrammatic representation of three decision boundaries in a two-dimensional space. Decision boundary I is purely deterministic, based on rules (e.g., equation 1) and classifying unknown individuals to population *A* (open squares) or to individuals of population *B* (closed circles). Misclassified individuals from population *A* are above the line of equal probability (see also Waser and Strobeck 1998). No learning occurs in this case. Decision boundary II exemplifies results of a well-designed machine learning classifier. Classification error is reduced compared to case I, representing the trade-off between classifier complexity (as shown by the relatively simple shape of its decision boundary) and classification accuracy. Decision boundary III represents the classifier with no classification error. The complex shape of the associated decision boundary indicates overfitting of the data and likely poor generalization to other data sets.

Once trained, each classifier considers a decision rule to determine if test individuals of unknown origin are more likely to have originated from one population over another. The general analytical framework partitions the *n*-dimensional feature space into *T* regions, where *T* is the number of populations/classes considered during the supervised classification (also often called "output" classes). Each feature represents a characteristic measured for each individual, such as genetic locus or phenotypic trait. For a hypothetical two-dimensional (e.g., two loci) and two-population case, the decision theory states, the individual is assigned to class A if the individual's vector falls above a given decision boundary, and is placed into class B otherwise (Figure 2). Classifier accuracy is assessed as the proportion of test individuals belonging to A (or B) correctly classified as A (or B).

One potentially useful aspect of machine learning methodologies is that they have the ability to learn features of baseline population data, allowing for adjustments of decision boundaries (Figure 2) independent of input (and possible arbitrary decisions) from biologists. Decision boundaries can thus be adjusted for different classifiers to improve assignment accuracy and precision.

Error rates associated with each classifier are easily computed. A baseline of sampled individuals is used to develop the classifier (training data set), and a separate set of individuals (testing data set) is used to estimate the assignment accuracy, representing the error rate. Unfortunately obtaining accurate decision boundaries with training datasets of reasonable size can be difficult (e.g., Duda et al. 2000; Fielding 1999). Due to limitations of sample size, it is often difficult to partition data sets into a set of individuals used to test model accuracy and those used to generalize decision boundaries for the classifier. As far as we know, only Cornuet et al. (1996) and Nielsen et al. (2001b) strictly applied this rule and considered a testing data set. In principle, classifiers should be built and trained with large sample sizes. Test data should be independent from training data. Various statistical resampling procedures such as resubstitution (Cornuet et al. 1999), *m*-fold cross-validation (e.g., Duda et al. 2000; Taylor et al. 2000), and "leave-one-out" (e.g., Efron 1983) are currently employed to overcome this problem. Resampling approaches are useful, but only asymptotically converge on the "true" error rate, and only with large sample sizes which are seldom realized for empirical studies of natural populations.

Materials and Methods

Four supervised classifiers will be considered in this study: the parametric likelihood-based approach of Paetkau et al. (1995) and three nonparametric, distribution-free machine learning classifiers—a back-propagation (multilayer perceptron) artificial neural network (ANN), a decision tree, and one *k*-NN algorithm (Figure 1). In this study we were not interested in comparing classifiers ("assignment tests") now commonly used in population genetics.

Likelihood-Based Estimator (LE)

The assignment test (Paetkau et al. 1995) considered here is based on *deterministic* likelihood estimation (LE) that assumes a multinomial distribution of allelic values. An individual is assigned to population A if the ratio

$$\frac{L_A}{L_B} = \frac{\Pr(\text{genotype}\,|\,\hat{\theta}_A)}{\Pr(\text{genotype}\,|\,\hat{\theta}_B)} \qquad (1)$$

is greater than one, and to population B if the ratio is less than one. In the above expression, L_A is the probability of the individual's multilocus genotype (assuming Hardy–Weinberg and linkage equilibrium within the source population), conditional

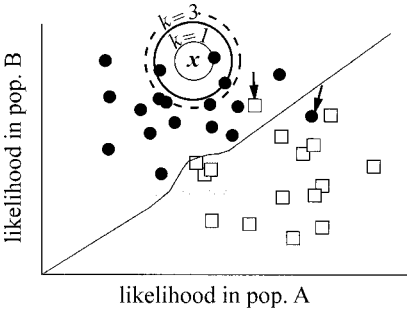

Figure 3. Diagrammatic representation of a *k*-NN classifier where interindividual relationships are defined in likelihood estimation space. Classification of each unknown individual *x* is based on individuals of known origin in the training sample (here, circles or squares). Unknown individual *x* will be assigned to the population of its most frequent *k* neighbors of known origin. Black arrows indicate misclassified individuals for which consideration of nearest neighbors cannot improve classification.

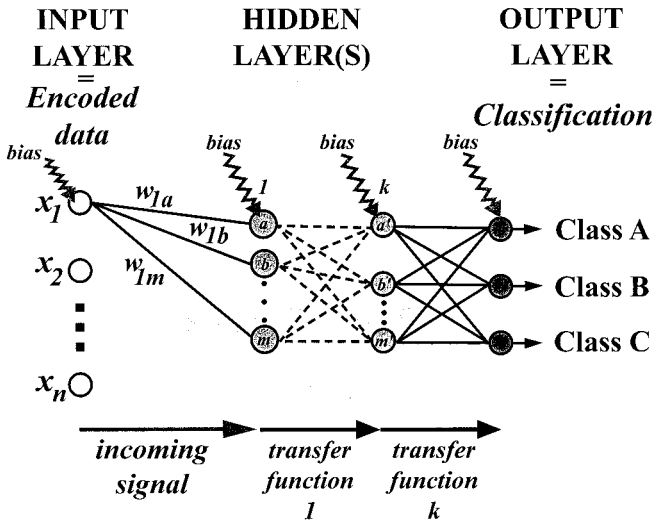

Figure 4. Schematic representation of a three-layer feed-forward artificial neural network (ANN). The processing elements in the network, "neurons" (circles in each layer), are discrete. Signals and/or transfer functions are sent through connections in only one direction, from input layer to output layer and no feedback connections are permitted. Connections are given a weight *w* that modulates the intensity of the signal through the network. The input layer contains encoded data (Table 1). Each training individual (x_1, x_2, . . ., x_n) is represented by a neuron linked to each neuron of the first hidden layers by vectors of weights *w* representing the incoming signal of the networks. The connection weights *w,* initially taken at random, are adjusted by the gradient descent method, based on the difference between the observed and expected outgoing signals (Duda et al. 2000).

on its origin being in a population with allele frequencies $\hat{\theta}_A$.

Using this representation, the definition of a classification decision boundary is deterministically imposed by the data (i.e., by the probability of an observed genotype given the expected frequencies in each putative population of origin) (Figure 2, case I) (see also Waser and Strobeck 1998). Likelihood estimators assume a multinomial distribution of data, a reliable estimate of allele frequency, Hardy–Weinberg equilibrium, and locus independence (i.e., no linkage).

Authors (Cornuet et al. 1999; Paetkau et al. 1995) noted that when alleles comprising an individual's genotype are absent in a potential source, this leads to a likelihood of zero and eliminates de facto this population as the population of likely origin. In assignment test studies, null frequencies can be accounted for in several ways: null allele frequencies can be replaced by a small constant value, by the inverse number of gene copies sampled in each population, or using a combination of these two procedures. In this study, in simulations as well as for empirical data sets, null frequencies were replaced by $0.1/2n$, where n is the sampled (or simulated) number of individuals in each population. This criterion is ad hoc, generally used in applied and theoretical studies (e.g., Cornuet et al. 1999). Other criteria should be defined on a more rigorous statistical basis (Huang and Weir 2001), but need further evaluation.

k-NN Analysis

The *k*-NN classifier employs a nonparametric algorithm that does not assume an underlying distribution of data. The *k*-NN classifier is based on measures of interindividual distance defined on the basis of user-defined metrics. A variety of *k*-NN classifications have been developed in several disciplines (e.g., Dasarathy 1991; Devijver and Kittler 1982; Duda et al. 2000). They are attractive because of their simplicity in assumptions of distributional properties of data and computational requirements.

The *k*-NN decision rule is based on individuals of known origin whose relationships in the appropriate feature space are based on their multilocus genotype. Individuals that share common properties are likely to belong to the same population of origin (Figure 3). For example, unknown individual *x* will be assigned to the population most frequently identified among *k*-NN individuals of known origin (Figure 3). The proximity of individuals must then be defined according to a metric. A large number of potential metrics have been proposed for *k*-NN classification (e.g., Dasarathy 1991). For simplicity we have chosen to consider only the Euclidean distance in LE score space, as presented in the previous section. In this case the results of parametric likelihood computations are used as a preprocessing step for nonparametric *k*-NN classification (Figure 3). Transformation of *k*-NN rules to the LE space could potentially reveal deviations of the LE decision boundary. Semagn et al. (2000) used a similar approach using discriminant analysis as the preprocessing step. The *k*-NN method based on LE preprocessing is potentially valuable in cases where individuals are misclassified by LE, but are close to the boundary decision line (i.e. $\Pr(\text{genotype}|\hat{\theta}_A) \approx \Pr(\text{genotype}|\hat{\theta}_B)$; Figure 3, gray box).

An odd number of 1, 3, and 5 nearest neighbors were considered in this study. Results considering more neighbors are not reported, as the use of larger numbers of known nearest neighbors leads to lower or identical assignment accuracy. Correction for null allele frequency was made as for LE, substituting $0.1/2n$ to missing values.

Artificial Neural Networks

Artificial neural networks (ANNs) were inspired by the structure and process of biological cognition and learning. ANNs learn from experience and can potentially rapidly solve complex computational problems. ANNs are programmed as multilayered structures possessing "neurons" composed of an input layer (e.g., data for each of *x* individuals), one or several hidden layers (representing or functioning as neurons), and an output layer representing the output classes (e.g., populations) in which individuals are assigned (Figure 4). We used a feed-forward ANN with a supervised back-propagation algorithm (Boddy and Morris 1999; Duda et al. 2000) which has been applied previously to population genetics questions (Aurelle et al. 1999; Cornuet et al. 1996; Taylor et al.

Table 1. Summary of data coding and testing procedures for simulated and empirical data sets used for each classification method

Classifier	Coding of data	Testing in simulated data sets	Testing in empirical data sets
Parametric			
Likelihood estimation	Use the observed probability of the individual multilocus genotypes	Testing made on 1000 independent individuals	Leave-one-out
Nonparametric			
k-NN in likelihood space	Likelihood method preprocessing	Same as for likelihood estimation	Leave-one-out
Artificial neural network	0, 1, or 2 (0 if the allele is not represented in the individual, 1 if the individual is heterozygous at a given locus, 2 if it is homozygous)	Same as for likelihood estimation	Cross-validation Five iterations Validation file used Training: 80% Testing: 20%
Decision tree	Same as for ANN	Same as for likelihood estimation	Cross-validation Five iterations Training: 80% Testing: 20%

1994). Each neuron in the network is linked to other neurons of adjacent layers, and neurons receive information through these links. Inner hidden layers define vectors of weight w for connections of each input neuron with each hidden neuron to process the incoming signal (Figure 4). The input signal is passed through a "transfer function" to produce the outgoing signal (classification) (Figure 4) (e.g., Duda et al. 2000; Ripley 1996). Classification of each individual is made to the population where the outgoing signal is the largest. The ANN is really a machine learning approach that facilitates searches for decision boundaries (i.e., combinations of weights w), minimizing the error rate. Using a combination of training and testing datasets, different networks can be built, each represented by different combinations of weights resulting from a search process (Figure 4) in a multidimensional space. Precautions must be taken to avoid overfitting (see below).

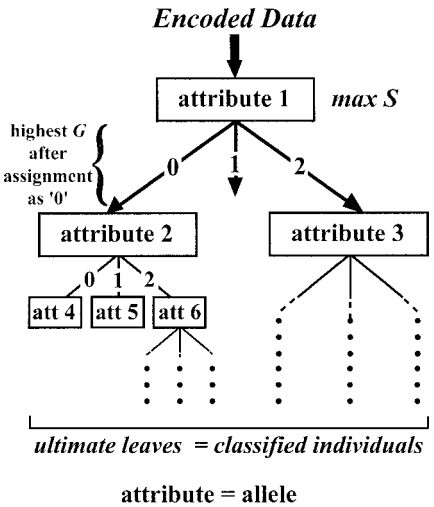

Figure 5. Diagrammatic representation of a decision tree (DT). Classification is performed by a tree structure that partitions sample data. Starting from the root (i.e., all unclassified individuals), each node in the tree repartitions individuals on the basis of feature values (e.g., allele states) that result in maximal information gain, quantified in terms of reduction in entropy within groups (see text for details).

In ANNs, each allele is coded following Cornuet et al. (1996). For each locus, the alleles are coded as 0 if no copy was present, as 1 if one copy is present (heterozygous individual), and as 2 if two copies are present (homozygous individual). Hence each individual of the input layer is characterized by a vector of [0,1,2] values specifying the multilocus genotype (Figure 4). Four hidden layers were considered and the two classes (potential population of origins) representing the output layer were used for both simulated and empirical data sets.

ANNs can "learn" features of a training data set. However, weights established for training data may not be generalizable to other data sets, even from the same populations. This can lead to overfitting (Figure 2, case III), when a testing data set is used to validate the network (Table 1). For optimal generalization of the results (i.e., the combinations of weights w issued for training), overfitting can be prevented by using additional validation data sets. Classification errors of individuals contained in a validation data set are used to stop training and to select an optimal combination of weights (e.g., Duda et al. 2000; Karystinos and Pados 2000). In this study the network having the lowest error rate in the validation data set was used. The testing data set was then used to estimate error rates and the accuracy of the selected ANN. Such a "stopping rule" has previously been used to prevent overfitting in an ecological application (Karul et al. 2000). Other means of avoiding overfitting include selection of a minimum number of "neurons" (i.e., minimizing the number of hidden layers) (Figure 4) (Cornuet et al. 1996). Too many neurons can lead to overfitting of the training data set, while employing too few may impede problem solving (e.g., Duda et al. 2000; Ripley 1996). Adjustments are time-consuming, but necessary. The ANN classifier used in this study was built and performed with Propagator software (ARD Corp.).

Decision Tree

Decision trees (DTs) are very popular machine learning classification methods because they create knowledge models that are easily comprehensible (Mitchell 1997). Mitchell (1997) and Quinlain (1993) provide detailed discussions of the methods for building decision trees. The distinction is often made between methods using continuous rather than discrete, categorical variables (e.g., Bell 1999). The classification is performed using a tree structure that partitions the input space based on sample data (Figure 5), classifying individuals by sorting through the "tree" from the base in "root" to "leaf" nodes (e.g., Bell 1999; Duda et al. 2000; Mitchell 1997). Starting at the root (e.g., a mixture of individuals originating from different populations), each node in the tree contains a test question about a single variable (or feature) (Figure 5). The critical step in the decision tree is to specify which feature A (e.g., alleles) to test at each node. This decision is made considering first the entropy S of each variable with entropy defined as

$$S = \sum_{i=1}^{c} -p_i \log_2 p_i, \qquad (2)$$

with p representing the proportion of individuals sharing a particular state for a given allele and c representing the number of states (0, 1, 2 coding for the copy of each allele; see Table 1). Information gain is measured by reduction in entropy with-

Table 2. Mean error rates (with standard deviation in parentheses; 10 replicates) of each classifier for 5 and 10 loci over the range allelic diversity and levels of inter-population variance considered in this study

	Low-diversity case (mean 3 alleles per locus)				High-diversity case (mean 7 alleles per locus)			
	Low F_{ST}		High F_{ST}		Low F_{ST}		High F_{ST}	
	5 loci	10 loci	5 loci	10 loci	5 loci	10 loci	5 loci	10 loci
Likelihood estimation	33.5 (1.8)	25.6 (1.6)	8.4 (0.9)	3.0 (0.4)	28.4 (1.4)	21.4 (1.7)	4.3 (0.3)	0.45 (0.1)
k-NN 1	40.2 (2.2)	34.7 (2.0)	15.2 (1.9)	5.8 (0.9)	39.6 (4.1)	34.5 (3.6)	4.8 (0.8)	0.9 (0.3)
3	45.5 (1.9)	36.3 (2.3)	14.4 (1.7)	5.5 (0.8)	36 (3.6)	31.2 (3.0)	5.6 (0.5)	0.9 (0.3)
5	48.0 (2.1)	31.6 (2.1)	12.5 (1.7)	6.2 (0.6)	33.7 (3.3)	28.5 (3.1)	5.6 (0.6)	1.2 (0.3)
Decision tree	39.5 (1.7)	38.6 (1.6)	18.3 (1.1)	19.7 (1.0)	40.3 (1.9)	38.1 (1.8)	10.5 (1.0)	8.6 (0.9)
Neural networks	30.7 (1.3)	25.0 (1.4)	6.2 (0.9)	2.3 (0.3)	27.5 (1.1)	20 (0.5)	3.1 (0.3)	0.9 (0.2)

in the subsets of individuals before and after splitting. Thus at each split, all the attributes are evaluated to select the most discriminating feature resulting in a split. Information gain $G(S,A)$ of a feature A is estimated by incorporating entropy as

$$G(S, A) = S - \sum_{v \in |0,1,2|} \frac{|S_v|}{|S|} S_v. \qquad (3)$$

S_v represents the subset of S for which attribute A has value v (Mitchell 1997). Decision tree analyses were performed using the C4.5 program (Quinlain 1993).

Simulated Data

It is important to assess how classifier assignment accuracy depends on properties of data given the now widespread use of, for example, SNPs in human genetics. Factors such as the degree of population differentiation, allelic diversity, and total number of loci have previously been evaluated for likelihood-based estimators (e.g., Bernatchez and Duchesne 2000; Smouse and Chevillon 1998) and will be examined here for each classifier. Levels of population differentiation are known to influence assignment accuracy (Cornuet et al. 1999; Smouse et al. 1982). We considered two levels of interpopulation variance (Wright 1965) in allele frequency representing a plausible range of population differentiation ($F_{ST} = 0.01$ and $F_{ST} = 0.1$ per locus; hereafter a low and high F_{ST} case, respectively). Five and 10 loci were considered to represent the range of loci used in many applied studies. Finally, two different levels of allelic diversity—low (three alleles per locus) and high (eight alleles per locus)—were considered. Individuals were drawn independently by generating two random alleles with prescribed probabilities at every locus. For each experiment we randomly drew 50 training individuals for each population from allelic distributions. Fifty individuals were considered here in baseline data sets as reflecting sample sizes per population generally used in applied studies. Smaller sample sizes can introduce bias because of inaccuracies in the estimation of allele frequencies that can influence assignment results. Ten replications of this experiment were made with 10 different baseline (training) data. Hence 10 different classifiers were constructed and compared in each case.

Empirical Data

Broodstocks for each of three lake trout hatchery strains (Marquette [SMD], Isle Royale [SIW], and Seneca Lake [SLW]) used in this study are maintained at U.S. Fish and Wildlife Service hatcheries. Details concerning data collections, as well as DNA extraction and polymerase chain reaction (PCR) protocols for the eight polymorphic microsatellite loci (*Ogo1a, Oneμ9, Oneμ10, Scoμ19, Sfo1, Sfo12, Sfo18,* and *Ssa85*) used are given in Page (2001). The SMD and SIW strains originate from native Lake Superior lake trout populations. The SLW strain was derived from a native population from Seneca Lake, a small lake in New York state. Preliminary works (Guinand B, Page KS, and Scribner KT, unpublished results) have shown that such a difference in the origin of the samples was reflected in the level of population differentiation as measured by F_{ST}. The level of population differentiation between the SMD and SIW strains was ≈0.01 across all loci and will be used as the "low" F_{ST} case. The level of population differentiation of both the SMD and SIW strains with SLW was ≈0.1 across all loci. SIW-SLW will be used as the "high" F_{ST} case.

Estimating Error Rates for Simulated and Empirical Data Sets

Classification error rates associated with each method for simulated data sets (10 randomly drawn baseline data sets for each method; see above) were established using 1000 independent simulated test individuals derived from the same allele frequency distributions. For each method, standard deviations of classification error rates were thus established for those 10 replicates. The number of replicates was kept low, because results indicated low standard deviations and consistency of mean error rates for each method when independent tests individuals were used. Extending the number of replicates was not required and was computationally intensive. For empirical data sets, procedures were specific to each classifier (Table 1). Error rates for the likelihood estimation and *k*-NN were computed with a leave-one-out procedure, as is frequently done in empirical studies. The accuracy of the ANN was assessed using a cross-validation procedure after drawing individuals at random for a validation data file. Individuals where randomly selected in a fivefold cross-validation procedure (80% of individuals used for training and 20% for testing). Different cross-validation procedures allocating more or fewer individuals to the training and testing data sets have been tested (e.g., Huberty 1994). Classification errors were shown to be almost constant (±2.5% for reported error rates) and did not affect the relative accuracies of the ANN and DT classifiers or the results of this study for empirical data. The leave-one-out procedure is possible with ANN (e.g., Aurelle et al. 1999), but it is computationally very intensive because a different classifier must be constructed for each individual tested.

Results

Assignment accuracies associated with each classifier varied across simulated data sets (Table 2). Classification error rates were generally lower when loci with high allelic diversity were used, when large numbers of loci were considered, and when population differentiation was high.

Likelihood and ANN classifiers outperformed other methods across all case studies considered (Table 2). The *k*-NN

Table 3. Summary of classification error rates (%) reported by each classifier for *Salvelinus namaycush* empirical data sets

Classifier		Low F_{ST} case SMD-SIW	High F_{ST} case SIW-SLW
Likelihood estimation		39.2	10.2
k-NN	1	37.3	12.0
	3	50.0	10.2
	5	48.0	11.1
Decision tree[a]		35.1	20.5
Neural networks[a]		17.9	4.0

[a] Average over five iterations (Table 1). For decision tree and neural networks, reported results represent averages over five iterations using various test (and validation) data sets (see Table 1). For neural networks, ranges of classification error rates never encompassed the classification error rates reported for the likelihood estimation case.

and DT classifiers were comparatively less accurate over the range of conditions surveyed. Based on classification error rate, ANNs marginally outperform the likelihood estimation. The range of differences in error rate between ANN and LE classifications was 0.6–2.8%, except in one case (high diversity, high F_{ST}, 10 loci case; Table 2). However, standard deviations around mean estimates of assignment accuracy were overlapping between LE and ANN classifiers, reflecting similar classification performance (Table 2). Generally the error rate decreases slightly more from the 5 locus case to the 10 locus case for LE than for ANN, indicating that the number of loci is likely more important for the LE classifier than for ANN.

The relative performance of each classifier when applied to empirical data differed from results observed for simulated data sets. The ANN outperforms the likelihood estimate in both cases considered (Table 3). Estimated error rates for the DT and *k*-NN classifiers were slightly lower than the error rate associated to the likelihood estimation in the low F_{ST} case (Table 3).

Discussion

Classification of individuals to populations of likely origin based on multilocus genotypes is a growing area of interest in population genetics (e.g., Bernatchez and Duchesne 2000; Hansen et al. 2001a; Waser and Strobeck 1998). Among the numerous assignment methodologies proposed, the Paetkau et al. (1995) likelihood estimation is computationally simple and is the most frequently employed (e.g., Hansen et al. 2000, 2001; Kyle and Strobeck 2001; Polzhien et al. 2000; Pope et al. 2000). Supervised machine learning classifiers have been used less frequently in population genetics. Only ANN methods have been used effectively (Aurelle et al. 1999; Cornuet et al. 1996; Giraudel et al. 2000; Taylor et al. 1994). We compare for the first time machine learning classifiers with likelihood-based estimation using common data sets representing a wide range of conditions (Tables 2 and 3).

Low classification error rates have previously been reported for ANN, particularly when the degree of population differentiation is low (Aurelle et al. 1999; Cornuet et al. 1996). This finding is confirmed in this study based on simulated and empirical data sets. The ANN classifier outperformed all other classification techniques evaluated for most of the parameter combinations and underlying data distributions (Tables 2 and 3). Only one exception occurred: where LE outperformed ANN (high allelic diversity, high F_{ST}, 10 loci case; Table 2). When different classifiers previously were used (Cornuet et al. 1996; Taylor et al. 1994), the authors did not report results from simulations. The results reported indicated that ANN reduced error rates by more than 5% over other methods in cases where population differentiation was low. This range in estimation of error rates is typically observed in the empirical data between ANN and other classification techniques (Table 3), but not in simulated data sets, where the range between ANN and LE classifications was lower (0–2.8%; Table 2). The error rates based on likelihood classification (assignment test) and ANN are nearly equivalent for simulated data sets (Table 2), but not for empirical examples (Table 3). Of interest is that error rates associated with likelihood classification were similar for both the empirical examples and simulated data sets (Tables 2 and 3). Error rates of DT and *k*-NN classifiers were generally several percent higher in simulated data sets compared to empirical data sets, regardless of the number of loci, allelic diversity, or the degree of population differentiation (Table 2). DT and *k*-NN classifications provide better results in the empirical examples, where their error rates are very close or slightly outperform the assignment accuracy associated with LE (e.g., empirical low F_{ST} case; Table 3).

Based on the results for the simulated and empirical data sets, no single classification method appears more adapted to multilocus genotype data. Machine learning classifiers generally exhibited large differences in assignment accuracy between simulated and empirical data sets (Tables 2 and 3), suggesting a lack of general application or, minimally, that assignment accuracy should be examined on a case-by-case basis (Duin 1996). Disparities in results could be due to particular features of the empirical data that are not present in the simulated data or to differences in the manner in which error rates are estimated in the simulated and empirical data sets.

Empirical data sets always deviate from underlying model assumptions. Classical assignment tests such as LE (Paetkau et al. 1995) and others (e.g., Pritchard et al. 2000) assumed random mating and independence of alleles and loci (i.e., expectations of observing multilocus genotypes in populations can be estimated from allele frequencies). Hatchery strains are characterized by departures from Hardy–Weinberg equilibrium at several loci and low levels of linkage disequilibrium between loci (Guinand B, Scribner KT, Page KS, and Burnham-Curtis MK, unpublished data). The influence of departures of underlying assumptions is largely unknown. We may hypothesize that machine learning classifiers can learn patterns present in data that likelihood estimation cannot. This represents a profitable area of future research.

Error rates are not reported in the same way in simulated and empirical data sets (Table 1). In simulations, error rates represent "true" error rates because testing individuals were independent from the training data set used for classifier design (Table 2). This is not the case for empirical data because of the cross-validation scheme (Table 1). Reported error rates are likely biased (Table 3), despite precautions such as using a validation data set to avoid overfitting in ANN. The use of independent samples, as is done in the simulated data sets, is believed to minimize bias in estimation of error rates (Duda et al. 2000; Fielding 1999; Fielding and Bell 1997; Salzberg 1997). Bias in estimating accurate error rates with ANN was recently reported by Manel et al. (1999) for an ecological application. Flexer (1996) reviewed experimental studies using ANNs in the machine learning literature. He reported that only 3 of 43 studies (7%) used a separate data set for parameter investigation, leaving open the possibility that many of the reported error rates of such ANN classifiers were overly optimistic. For the present study, the relative merits of classifiers are thus more appropriately assessed using simulated data sets. Previous popu-

lation genetics studies considering ANNs (Aurelle et al. 1999; Cornuet et al. 1996; Taylor et al. 1994) dealt only with empirical data sets and cross-validation (or leave-one-out) testing. When comparisons with other classifiers were made on empirical data (discriminant analysis) (Cornuet et al. 1996; Taylor et al. 1994), error rates were reduced by more than 7% with ANNs. Similar improvements in assignment accuracy were seen in analyses of our empirical data (Table 3).

An additional source of assignment error from empirical data is caused by sampling error. Only a small number of individuals are typically sampled in each population. Classifiers are thus frequently built only on these limited data, and are incapable of capturing all potential information present in the entire population (e.g., poor estimation of allele frequency and presence/absence of rare alleles). If intuitively important (e.g., Smouse and Chevillon 1998), the role of the number of baseline individuals used for population assignment has rarely been considered. To our knowledge, only Rosenberg et al. (2001) investigated the paramount importance of this parameter in an empirical study of chicken breeds. Moreover, samples can be affected by gene correlations attributed to behavioral or ecological characteristics of the species studied or high levels of coancestry in domestic hatchery strains (e.g., in fishes, relatedness between individuals or family structure) (Hansen et al. 1997).

The performance of all classifiers is affected by the nature of the problem and the data (Fielding 1999:chap. 8). Decisions to use one classifier over another should be based on criteria and techniques for ranking performances (Duin 1996; Fielding 1999; Salzberg 1997). However, this criterion alone is not sufficient (Duin 1996; see also Hansen et al. 2001a). Classifier simplicity (i.e., a classifier that does not require excessive parameter adjustment) is another valuable criterion. As shown in this study, ANNs, while consistently the most accurate method, are conceptually and computationally more difficult to use. A classifier such as likelihood estimation (Paetkau et al. 1995) or discriminant analyses (e.g., Beacham et al. 1999; Douglas and Brunner 2002; Schmidt 1999) may thus be preferred. Similar sentiments have been forwarded based on studies comparing classifiers in ecological (e.g., Leung and Tran 2000; Manel et al. 1999) and conservation-oriented (Riordan 1998) applications, where complex classifiers including ANNs performed only slightly better than other classifiers.

Conclusion

Results from simulated and empirical data sets do not categorically demonstrate that one classification method is superior to another (Tables 2 and 3). ANNs and other machine learning techniques employed in this study seem to be able to capture additional information present in empirical data to improve classification. The comparatively simple, deterministic likelihood estimation proposed by Paetkau et al. (1995) classifies individuals with accuracies comparable to ANNs (Table 2). Other comparisons of these classifiers are needed to understand how classification accuracy may be affected (or biased) by model assumptions (linked loci, heterozygote excess or deficit). For empirical cases where even low levels of linkage and departures from Hardy–Weinberg are observed, the use of machine learning classifiers could lead to a better understanding of the factors that determine classification accuracy for the discrete, multistate characters commonly used in population ecology and genetics.

References

Almudevar A, 2000. Exact confidence regions for species assignment based on DNA markers. Can J Stat 28: 81–95.

Aurelle D, 1999. Contacts secondaires naturels et artificiels chez la truite commune (*Salmo trutta,* L.) des Pyrénées Occidentales françaises: utilisation de marqueurs microsatellites pour la distinction de taxons faiblement différenciés (PhD dissertation). France: University Montpellier II.

Aurelle D, Lek S, Giraudel JL, and Berrebi P, 1999. Microsatellites and artificial neural networks: tools for the discrimination between natural and hatchery brown trout (*Salmo trutta,* L.) in Atlantic populations. Ecol Model 120:313–324.

Beacham TD, Pollard S, and Le KD, 1999. Population structure and stock identification of steelhead in southern British Columbia, Washington, and the Columbia River based on microsatellite DNA variation. Trans Am Fish Soc 128:1068–1084.

Beaumont MA, Barratt EM, Gotelli D, Kitchener AC, Daniels MJ, Pritchard JK, and Bruford MW, 2001. Genetic diversity and introgression in the Scottish wildcat. Mol Ecol 10:319–336.

Bell JF, 1996. Application of classification trees to habitat preference of upland birds. J Appl Stat 23:349–359.

Bell JF, 1999. Tree-based methods. In: Machine learning methods for ecological applications (Fielding AH, ed). Boston: Kluwer Academic.

Bernatchez L and Duchesne P, 2000. Individual-based genotype analysis in studies of parentage and population assignment: how many loci, how many alleles? Can J Fish Aquat Sci 57:1–12.

Boddy L and Morris CW, 1999. Artificial neural networks for pattern recognition. In: Machine learning methods for ecological applications (Fielding AH, ed). Boston: Kluwer Academic.

Boddy L, Morris CW, Wilkins MF, Al-Haddad L, Tarran GA, Jonker RR, and Burkill PH, 2000. Identification of 72 phytoplankton species by radial basis function neural networks analysis of flow cytometry data. Mar Ecol Prog Ser 195:47–59.

Bogdanowitcz SM, Mastro VC, Prasher DC, and Harrison RG, 1997. Microsatellite DNA variation among Asian and North American gypsy moths (Lepidoptera: Lymantridae). Ann Entomol Soc Am 90:768–775.

Bowcock AM, Ruiz Linares A, Tomfohrde E, Minch E, Kidd JR, and Cavalli-Sforza R, 1994. High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455–457.

Congiu L, Dupanloup I, Patarnello T, Fontana F, Rossi R, Arlati G, and Zane L, 2001. Identification of interspecific hybrids by amplified fragment length polymorphism: the case of sturgeon. Mol Ecol 10:2355–2359.

Cornuet JM, Aulagnier S, Lek S, Franck P, and Solignac M, 1996. Classifying individuals among infra-specific taxa using microsatellite data and neural networks. C R Acad Sci Paris Life Sci 319:1167–1177.

Cornuet JM, Piry S, Luikart G, Estoup A, and Solignac M, 1999. New methods employing multilocus genotypes to select or exclude populations as origins of individuals. Genetics 153:1989–2000.

Curtis D, North BV, and Sham PC, 2001. Use of an artificial neural network to detect association between a disease and multiple marker genotypes. Ann Hum Genet 65:95–107.

Davies N, Villablanca FX, and Roderick GK, 1999. Bioinvasions of the medfly *Ceratitis capitata*: source estimation using DNA sequences at multiple intron loci. Genetics 153:351–360.

Dasarathy BV, 1991. Nearest neighbor (NN) norms: NN pattern classification techniques. Washington, DC: IEEE Computer Society.

Dawson KJ and Belkhir K, 2001. A Bayesian approach to the identification of panmictic populations and the assignment of individuals. Genet Res 78:59–77.

Devijver PA and Kittler J, 1982. Pattern recognition: a statistical approach. Englewood Cliffs, NJ: Prentice-Hall International.

Douglas MR and Brunner PC, 2002. Biodiversity of central alpine *Coregonus* (Salmoniformes): impact of one-hundred years of management. Ecol Appl 12:154–172.

Duda RO, Hart PE, and Stork DG, 2000. Pattern classification, 2nd ed. New York: John Wiley & Sons.

Duin RPW, 1996. A note on comparing classifiers. Pattern Recogn 17:529–536.

Efron B, 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. J Am Stat Assoc 78:103–130.

Eldridge MDB, Kinnear JE, and Onus ML, 2001. Source population of dispersing rock-wallabies (*Petrogale lateralis*) identified by assignment tests on multilocus genotypic data. Mol Ecol 10:2867–2876.

Estoup A and Angers B, 1998. Microsatellites and minisatellites for molecular ecology: theoretical and empirical considerations. In: Advances in molecular ecology (Carvalho GR, ed). Amsterdam: IOS Press.

Evett IW and Weir BS, 1998. Interpreting DNA evidence. Boston: Sinauer.

Fielding AH, ed. 1999. Machine learning methods for ecological applications. Boston: Kluwer Academic.

Fielding AH and Bell JF, 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ Conserv 24:38–49.

Fritzner NG, Hansen MM, Madsen SS, and Kristiansen K, 2001. Use of microsatellite markers for identification of indigenous brown trout in a geographical region heavily influenced by stocked domesticated trout. J Fish Biol 58:1197–1210.

Galbusera P, Lens L, Schenck T, Waiyaki E, and Matthysen E, 2000. Genetic variability and gene flow in the

globally, critically-endangered taita brush. Conserv Genet 1:45–55.

Giraudel JL, Aurelle D, Lek S, and Berrebi P, 2000. Application of the self-organizing map method to microsatellite data: how to detect genetic structure in brown trout (*Salmo trutta*) populations. In: Artificial neuronal networks: application to ecology and evolution (Lek S and Guégan JF, eds). Berlin: Springer-Verlag; 187–202.

Grigull J, Alexandrova R, and Paterson AD, 2001. Clustering of pedigrees using marker allele frequencies: impact on linkage analysis. Genet Epidemiol 21(suppl 1): 61–66.

Haig SM, Gratto-Trevor CL, Mullins TD, and Colwell MA, 1997. Population identification of western hemisphere shorebirds throughout the annual cycle. Mol Ecol 6: 412–427.

Hansen MM, Kenchington E, and Nielsen EE, 2001a. Assigning individual fish to populations using microsatellite DNA markers. Fish Fish Ser 2:93–112.

Hansen MM, Nielsen EE, and Mensberg KLD, 1997. The problem of sampling families rather than populations: relatedness among individuals in samples of juvenile brown trout *Salmo trutta* L. Mol Ecol 6:469–474.

Hansen MM, Ruzzante DE, Nielsen EE, and Mensberg KLD, 2000. Microsatellite and mitochondrial DNA polymorphism reveals life-history dependent interbreeding between hatchery and wild brown trout (*Salmo trutta* L.). Mol Ecol 9:583–594.

Hansen MM, Ruzzante DE, Nielsen EE, and Mensberg KLD, 2001b. Brown trout (*Salmo trutta*) stocking impact assessment using microsatellite DNA markers. Ecol Appl 11:148–160.

Huang S-P and Weir BS, 2001. Estimating the total number of alleles using sample coverage method. Genetics 159:1365–1373.

Huberty CJ, 1994. Applied discriminant analysis. New York: Wiley Interscience.

Jain AK, Duin RPW, and Mao J, 2000. Statistical pattern recognition: a review. IEEE Trans Patt Anal Mach Intell 22:4–37.

Jarne P and Lagoda PJL, 1996. Microsatellites, from molecules to populations and back. Trends Ecol Evol 11:424–429.

Karul C, Soyupak S, Cilesiz AF, Akbay N, and Germen E, 2000. Case studies on the use of neural networks in eutrophication modeling. Ecol Model 134:145–152.

Karystinos GN and Pados DA, 2000. On overfitting, generalization, and randomly expanded training sets. IEEE Trans Neur Network 11:1050–1057.

King TL, Kalinowski ST, Schill WB, Spidle AP, and Lubinski BA, 2001. Population structure of Atlantic salmon (*Salmo salar* L.): a range-wide perspective from microsatellite DNA variation. Mol Ecol 10:807–821.

Knutsen H, Knutsen JA, and Jorde PE, 2001. Genetic evidence for mixed origin of recolonized sea trout populations. Heredity 87:207–214.

Koskinen MT, Piironen J, and Primmer CR, 2001. Interpopulation genetic divergence in European grayling (*Thymallus thymallus,* Salmonidae) at a microgeographic scale: implications for conservation. Conserv Genet 2:133–143.

Kyle CJ and Strobeck C, 2001. Genetic structure of North American wolverine (*Gulo gulo*) populations. Mol Ecol 10:337–347.

Lek S and Guégan JF, eds. 2000. Artificial neuronal networks: application to ecology and evolution. Berlin: Springer-Verlag.

Leung PS and Tran LT, 2000. Predicting shrimp disease occurrence: artificial neural networks vs. logistic regression. Aquaculture 187:35–49.

Manel S, Dias JM, Buckton ST, and Ormerod SJ, 1999. Alternative methods for predicting species distribution: an illustration with Himalayan river birds. J Appl Ecol 36:734–747.

Marshall AR, Blankenship HL, and Connor WP, 2000. Genetic characterization of naturally spawned Snake River fall-run chinook salmon. Trans Am Fish Soc 129: 680–698.

Martinez JL, Dumas J, Beall E, and Garcia-Vazquez E, 2001. Assessing introgression of foreign strains in wild Atlantic salmon populations: variation in microsatellites assessed in historic scale collections. Freshwater Biol 46:835–844.

Mitchell TM, 1997. Machine learning. Boston: McGraw-Hill.

Müller J, 2000. Mitochondrial DNA variation and the evolutionary history of cryptic *Gammarus fossarum* types. Mol Phylogenet Evol 15:260–268.

National Research Council, 1996. The evaluation of forensic DNA evidence. Washington, DC: National Academy Press.

Neraas LP and Spruell P, 2001. Fragmentation of riverine systems: the genetic effects of dams on bull trout (*Salvelinus confluentus*) in the Clark Fork River system. Mol Ecol 10:1153–1164.

Nielsen EE, Hansen MM, and Bach LA, 2001a. Looking for a needle in a haystack: discovery of indigenous Atlantic salmon (*Salmo salar* L.) in stocked populations. Conserv Genet 2:219–232.

Nielsen EE, Hansen MM, Schmidt C, Meldrup D, and Grønkjaer P, 2001b. Population of origin of Atlantic cod. Nature 413:272.

Olsen JB, Bentzen P, Banks MA, Shaklee JB, and Young S, 2000. Microsatellites reveal population identity of individual pink salmon to allow supportive breeding of a population at risk of extinction. Trans Am Fish Soc 129: 232–242.

Paetkau D, Calvert W, Stirling I, and Strobeck C, 1995. Microsatellite analysis of population structure in Canadian polar bears. Mol Ecol 4:347–354.

Page KS, 2001 Genetic diversity and interrelationships of wild and hatchery lake trout in the upper Great Lakes: inferences for broodstock management and development of restoration strategies (MSc thesis). East Lansing: Michigan State University.

Pella JJ and Masuda M, 2001. Bayesian methods for analysis of stock mixtures from genetic characters. Fish Bull 99:151–167.

Pella JJ and Milner GB, 1987. Use of genetic marks in stock composition analysis. In: Population genetics and fishery management (Ryman N and Utter F, eds). Seattle: University of Washington Press; 247–276.

Pettersson JCE, Hansen MM, and Bohlin T, 2001. Does dispersal from landlocked trout explain the coexistence of resident and migratory trout females in a small stream? J Fish Biol 58:487–495.

Polzhien RO, Hamr J, Mallory FF, and Strobeck C, 2000. Microsatellite analysis of North American wapiti (*Cervus elaphus*) populations. Mol Ecol 9:1561–1576.

Pope LC, Estoup A, and Moritz C, 2000. Phylogeography and population structure of an ecotonal marsupial, *Bettongia tropica,* determined using mtDNA and microsatellites. Mol Ecol 9:2040–2053.

Potvin C and Bernatchez L, 2001. Lacustrine spatial distribution of landlocked Atlantic salmon populations assessed across generations by multilocus individual assignment and mixed-stock analysis. Mol Ecol 10:2375–2388.

Primmer CR, Aho T, Piironen J, Estoup A, Cornuet JM, and Ranta E, 1999. Microsatellite analysis of hatchery stocks and natural populations of Arctic charr, *Salvelinus alpinus,* from the Nordic region: implications for conservation. Hereditas 130:277–289.

Primmer CR, Koskinen MT, and Piironen J, 2000. The one that did not get away: individual assignment using microsatellite data detects a case of fishing competition fraud. Proc R Soc Lond B 267:1699–1704.

Pritchard JK, Stephens M, and Donnelly P, 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Quinlain JR, 1993. C4.5: programs for machine learning. San Mateo, CA: Morgan Kaufmann.

Randi E and Lucchini V, 2002. Detecting rare introgression of domestic dog genes into wild wolf (*Canis lupus*) populations by Bayesian admixture analyses of microsatellite variation. Conserv Genet 3:31–45.

Randi E, Pierpaoli M, Beaumont M, Ragni B, and Sforzi A, 2001. Genetic identification of wild and domestic cats (*Felis sylvestris*) and their hybrids using Bayesian clustering methods. Mol Biol Evol 18:1679–1693.

Rannala B and Mountain JL, 1997. Detecting immigration using multilocus genotypes. Proc Natl Acad Sci USA 94:9197–9202.

Raymer ML, Punch WS III, Venkataraman S, Sanschagrin PC, Goodman ED, and Kuhn LA, 1997. Predicting conserved water-mediated and polar ligand interactions in proteins using a K-nearest neighbor genetic algorithm. J Mol Biol 265:445–464.

Riordan P, 1998. Unsupervised recognition of individual tigers and snow leopards from their footprints. Anim Conserv 1:253–262.

Ripley BD, 1996. Pattern recognition and neural networks. Cambridge: Cambridge University Press.

Roeder AD, Marshall RK, Mitchelson AJ, Visagathilagar T, Ritchie PA, Love DR, Pakai TJ, McPartlan HC, Murray ND, Robinson KR, Kerry KR, and Lambert DM, 2001. Gene flow on the ice: genetic differentiation among Adélie penguin colonies around Antarctica. Mol Ecol 10:1645–1656.

Roques S, Duchesne P, and Bernatchez L, 1999. Potential of microsatellites for individual assignment: the North Atlantic redfish (genus *Sebastes*) species complex as a case study. Mol Ecol 8:1703–1717.

Rosenberg NA, Burke T, Feldman MW, Friedlin P, Groenen MA, Hillel J, Maki-Tanila A, Tixier-Boichard M, Vignal A, Wimmers K, and Weigend S, 2001. Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. Genetics 159:699–713.

Salzberg SL, 1997. On comparing classifiers: pitfalls to avoid and a recommended approach. Data Min Knowl Disc 1:317–328.

Schmidt CA, 1999. Variation and congruence of microsatellite markers for *Peromyscus leucopus.* J Mammal 80:522–529.

Schulte-Hostedde AI, Gibbs HL, and Millar JS, 2001. Microgeographic genetic structure in the yellow-pine chipmunk (*Tamias amoenus*). Mol Ecol 10:1625–1631.

Sefc KM, Lopes MS, Lefort L, Botta R, Roubelakis-Angelakis K, Ibanez J, Pejic I, Wagner H, Glössl J, and Steinkellner H, 2000. Microsatellite variability in grapevine cultivars from different European regions and evaluation of assignment testing to assess the geographic origin of cultivars. Theor Appl Genet 100:498–505.

Semagn K, Bjornstadt A, Stedje B, and Bekele E, 2000. Comparison of multivariate methods for the analysis of genetic resources and adaptation in *Phytolacca dodecandra* using RAPD. Theor Appl Genet 101:1145–1154.

Smouse PE and Chevillon C, 1998. Analytical aspects of population-specific DNA fingerprinting for individuals. J Hered 89:143–150.

Smouse PE, Spielman RS, and Park MH, 1982. Multiplelocus allocation of individuals to groups as a function of the genetic variation within and differences among human populations. Am Nat 119:445–463.

Spidle AP, Schill WB, Lubinski BA, and King TL, 2001. Fine-scale population structure in Atlantic salmon from Maine's Penobscot River drainage. Conserv Genet 2:11–24.

Taylor EB, Beacham TD, and Kaeriyama M, 1994. Population structure and identification of North Pacific Ocean chum salmon (*Oncorhynchus keta*) revealed by an analysis of minisatellite DNA variation. Can J Fish Aquat Sci 51:1430–1442.

Taylor EB, Kuiper A, Troffe PM, Hoysak DJ, and Pollard

S, 2000. Variation in developmental biology and microsatellite DNA in reproductive ecotypes of kokanee, *Oncorhynchus nerka*: implications for declining populations in a large British Columbia lake. Conserv Genet 1:213–249.

Thorrold SR, Latkoczy C, Swart PK, and Jones CM, 2001. Natal homing in a marine fish metapopulation. Science 291:297–299.

Tsutsui ND, Suarez AV, Holway DA, and Case TJ, 2001. Relationships among native and introduced populations of the Argentine ant (*Linepithema humile*) and the source of introduced populations. Mol Ecol 10:2051–2161.

Vázquez-Domínguez E, Paetkau D, Tucker N, Hinten G, and Moritz C, 2001. Resolution of natural groups using iterative assignment tests: an example from two species of Australian rats (*Rattus*). Mol Ecol 10:2069–2078.

Waser PM and Strobeck C, 1998. Genetic signatures of interpopulation dispersal. Trends Ecol Evol 13:43–44.

Whitler RE, Beacham TD, Watkins RF, and Stevens TA, 1994. Identification of farm-reared and native chinook salmon (*Oncorhynchus tshawytscha*) on the west coast of Vancouver Island, British Columbia, including the nuclear DNA probe B2–2. Can J Fish Aquat Sci 51(suppl 1):267–276.

Wiens JJ, 1999. Polymorphisms in systematics and comparative biology. Ann Rev Ecol Syst 30:327–362.

Wright S, 1965. The interpretation of population structure by *F*-statistics with special regards to system of mating. Evolution 19:395–420.

Zeisset I and Beebee TJC, 2001. Determination of biogeographical range: an application of molecular phylogeography to European pool frog *Rana lessonae*. Proc R Soc Lond B 268:933–938.

Received July 12, 2001
Accepted June 11, 2002

Corresponding Editor: Bruce S. Weir