Available online at www.sciencedirect.com
ScienceDirect
International Journal for Parasitology 38 (2008) 1425–1434
www.elsevier.com/locate/ijpara

Random forests, a novel approach for discrimination of fish populations using parasites as biological tags

Diana Perdiguero-Alonso [a,*], Francisco E. Montero [b], Aneta Kostadinova [a,c], Juan Antonio Raga [a], John Barrett [d]

[a] *Marine Zoology Unit, Cavanilles Institute of Biodiversity and Evolutionary Biology, University of Valencia, P.O. Box 22085, 46071 Valencia, Spain*
[b] *Department of Animal Biology, Plant Biology and Ecology, Autonomous University of Barcelona, Campus Universitari, 08193 Bellaterra, Barcelona, Spain*
[c] *Central Laboratory of General Ecology, Bulgarian Academy of Sciences, 2 Gagarin Street, 1113 Sofia, Bulgaria*
[d] *Institute of Biological Sciences, University of Wales, Aberystwyth, Ceredigion, SY23 3DA, UK*

Received 13 February 2008; received in revised form 24 March 2008; accepted 1 April 2008

Abstract

Due to the complexity of host–parasite relationships, discrimination between fish populations using parasites as biological tags is difficult. This study introduces, to our knowledge for the first time, random forests (RF) as a new modelling technique in the application of parasite community data as biological markers for population assignment of fish. This novel approach is applied to a dataset with a complex structure comprising 763 parasite infracommunities in population samples of Atlantic cod, *Gadus morhua*, from the spawning/feeding areas in five regions in the North East Atlantic (Baltic, Celtic, Irish and North seas and Icelandic waters). The learning behaviour of RF is evaluated in comparison with two other algorithms applied to class assignment problems, the linear discriminant function analysis (LDA) and artificial neural networks (ANN). The three algorithms are used to develop predictive models applying three cross-validation procedures in a series of experiments (252 models in total). The comparative approach to RF, LDA and ANN algorithms applied to the same datasets demonstrates the competitive potential of RF for developing predictive models since RF exhibited better accuracy of prediction and outperformed LDA and ANN in the assignment of fish to their regions of sampling using parasite community data. The comparative analyses and the validation experiment with a 'blind' sample confirmed that RF models performed more effectively with a large and diverse training set and a large number of variables. The discrimination results obtained for a migratory fish species with largely overlapping parasite communities reflects the high potential of RF for developing predictive models using data that are both complex and noisy, and indicates that it is a promising tool for parasite tag studies. Our results suggest that parasite community data can be used successfully to discriminate individual cod from the five different regions of the North East Atlantic studied using RF.

Keywords: Random forests; Classification algorithms; Fish population discrimination; Parasite communities; Atlantic cod; *Gadus morhua*; North East Atlantic

1. Introduction

The "stock" concept is controversial due to the marked difference in its perception and application between fishery biologists and fisheries managers, perceived as biological entities by the former and as management units by the latter with both not necessarily matching (Hammer and Zimmermann, 2005; Waldman, 2005). In spite of the problems with operational definitions, stock identification (i.e. defining stock characteristics and boundaries) and discrimination (or separation, i.e. identification of members of different stocks in the catches of mixed aggregations) studies have developed rapidly in the last decade due to their importance for the development of sustainable harvest and monitoring strategies by fisheries management.

* Corresponding author. Tel.: +34 963543685; fax: +34 963543733.
E-mail address: diana.perdiguero@uv.es (D. Perdiguero-Alonso).

0020-7519/$34.00
doi:10.1016/j.ijpara.2008.04.007

Fish parasites have been used as biological markers of fish populations/stocks since the work of Herrington et al. (1939). MacKenzie and Abaunza (1998, 2005) reviewed key literature on the use of parasites as biological tags in fish population studies, provided the criteria for the selection of parasites and recognised two main approaches to the use of parasites as tags, one including examination of the distributions in large host samples of a small number of parasite species which fulfil the criteria for a biological tag, and the other based on entire parasite assemblages utilising multivariate techniques. Timi (2007) provided a recent review on studies in the South West Atlantic justifying the advantages of the latter approach. The increased application of more sophisticated statistical techniques has allowed consideration of almost the entire parasite assemblages in stock identification methods and a subsequent selection of the most discriminating parasite species.

One important problem in class assignment tasks is that there are often many weak input variables with each one containing only a small amount of information so that no single input or small group of inputs can distinguish between classes (Breiman, 2001). The types of data are difficult to interpret using methods of classification such as linear discriminant function analysis (LDA) and artificial neural networks (ANN).

Here, we demonstrate, to our knowledge for the first time, an ensemble classification approach using random forests (RF) (Breiman, 2001) to the application of parasite community data as biological tags for fish class assignment. Random forests are ensembles of tree-type classifiers that use the method of bagging, an improved method of bootstrapping. The following advantages of the RF approach for multisource classification may prove to be useful in solving the problem of population assignment: it is non-parametric (i.e. no assumptions of normality or independence are required concerning the data), it handles data with many zeros well and provides a means of estimating the importance of individual variables in classification. In contrast to other machine learning methods, RFs are not prone to overtraining and develop models from data which are both noisy and complex. Misclassified cases are easily identified and the evolved models can be readily applied to new data.

The aim of this study is 2-fold. Firstly, using the same version of parasite community data derived from sampling of cod populations over a wide geographical range, we compare the learning behaviour of RF with two other algorithms, one traditionally applied to class assignment problems (LDA) and one used more recently (ANN). Second, we scrutinise the RF analysis by reducing the noise in the data in order to evaluate the effect of variable/data reduction on efficiency and consistency in class assignment and to examine the importance of annual and seasonal variation in parasite community composition and structure for discrimination of individual fish with respect to the region of sampling.

2. Materials and methods

2.1. Parasite community dataset

Parasites in cod populations sampled at five Spring (spawning) and Autumn (feeding) areas in the North East Atlantic (Baltic, Celtic, Irish and North seas, and Icelandic waters, see Fig. 1 for sampling locations and Tables 1 and 2 for sample sizes and parasites recovered, respectively) served as material for the present study. Analyses were carried out on the data on parasite communities in individual fish (i.e. infracommunities, see Bush et al., 1997). The entire dataset comprised parasite infracommunities in 763 fish collected in the five regions of study. Altogether 31 different parasite forms were considered (subsequently referred to as species). Of these, 26 were identified to species level, three were identified to generic level and two larval forms were identified to family/order (Table 2). Four taxa are currently recognised as complexes of sibling species (Väinölä et al., 1994; Anderson, 2000 and references therein). Since the distinction of these species was inferred from electrophoretic/molecular evidence, we were not able to discriminate between them (further referred to as *Anisakis simplex sensu lato* (*s. l.*), *Contracaecum osculatum s. l.*, *Pseudoterranova decipiens s. l.* and *Echinorhynchus gadi s. l.*). Species with occurrences of <1% in the total sample were excluded from analyses. Thus, the distributions of 31 parasite species in the total sample of 763 fish were used as independent variables and the sampling region was used as the dependent variable (see Table 3 for size of the sampling series and experimental design). A 'blind' mixed sample of 50 fish used in model validation was collected in Spring 2003 and comprised 10 (20%) individuals from each region.

2.2. Classification algorithms

2.2.1. Random forests

RF is an ensemble learning technique where the individual decisions of a large set of random classifiers (trees) are combined by majority voting (i.e. 'votes' for the most popular class) in order to obtain more accurate predictions than any individual classifier. Seemingly, the RF algorithm is reminiscent of the computerised minimum-length-tree approaches such as in PAUP* 4.0 (Swofford D.L., 2002. PAUP*. Phylogenetic Analysis Using Parsimony (* and other methods) beta. Sinauer Associates, Sunderland, Massachusetts, USA) perhaps due to the implementation of bootstrap analysis (Felsenstein, 1985) in the latter, i.e. sampling the original dataset with replacements to construct a series of bootstrap replicates of the same size as the original dataset in order to assign statistical confidence to hypotheses of relationship via construction of a majority-rule consensus. However, in addition to the different hypothesis in Felenstein's approach, the taxa (samples) are held constant and the characters (variables) are sampled with replacement, whereas RF uses both random inputs and random variable selection. Further, each decision tree built by RF