Taxonomic identification of environmental DNA with informatic sequence classification trees.

Shaun P. Wilkinson[1*], Simon K. Davy[1], Michael Bunce[2] & Michael Stat[2]

[1]School of Biological Sciences, Victoria University of Wellington, New Zealand

[2]Department of Environment and Agriculture, Curtin University, Perth, Western Australia

*Contact e-mail: shaunpwilkinson@gmail.com

Short description: Introducing the 'insect' bioinformatics pipeline for probabilistic taxon identification of environmental DNA sequences.

Abstract:

High-throughput sequencing of environmental DNA (eDNA) offers a simple and cost-effective solution for marine biodiversity assessments. Yet several analytical challenges remain, including the incorporation of statistical inference in the assignment of taxonomic identities. We developed a probabilistic method for DNA barcode classification that can be used for both eDNA and traditional single-source sampling. The pipeline involves: (1) compiling a primer-specific database of barcode sequences to be used as training data (obtained from GenBank and other sequence repositories), (2) generating a classification tree using an iterative learning algorithm that divisively sorts the training data into hierarchical clusters based on profile hidden Markov models, (3) assignment of each query sequence to a cluster using a recursive series of model-comparison tests, and (4) taxonomic identification of the query sequences based on the lowest common taxonomic rank of the training sequences within the cluster. This method compares favorably to other DNA classification methods when tested on benchmark datasets, and offers the added features of classifying at higher taxonomic ranks and returning interpretable confidence values in the form of the Akaike weight statistic. This bioinformatics pipeline is available as an open source R package called 'insect' (informatic sequence classification trees).

Keywords: Bioinformatics, eDNA, machine-learning, meta-barcoding

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26812v1 | CC BY 4.0 Open Access | rec: 30 Mar 2018, publ: 30 Mar 2018