Vol. 12: 131–137, 2020
https://doi.org/10.3354/aei00353

AQUACULTURE ENVIRONMENT INTERACTIONS
Aquacult Environ Interact

Published April 9



# Machine learning analyses of bacterial oligonucleotide frequencies to assess the benthic impact of aquaculture

**Ethan G. Armstrong*, Joost T. P. Verhoeven*,****

**Department of Biology, Memorial University of Newfoundland, St John's, Newfoundland A1B 3X9, Canada**

ABSTRACT: Aquaculture is a rapidly expanding industry and is now one of the primary sources of all consumed seafood. Intensive aquaculture production is associated with organic enrichment, which occurs as organic material settles onto the seafloor, creating anoxic conditions which disrupt ecological processes. Bacteria are sensitive bioindicators of organic enrichment, and supervised classifiers using features derived from 16s rRNA gene sequences have shown potential to become useful in aquaculture environmental monitoring. Current taxonomy-based approaches, however, are time intensive and built upon emergent features which cannot easily be condensed into a monitoring pipeline. Here, we used a taxonomy-free approach to examine 16s rRNA gene sequences derived from flocculent matter underneath and in proximity to hard-bottom salmon aquaculture sites in Newfoundland, Canada. Tetranucleotide frequencies (k = 4) were tabulated from sample sequences and included as features in a machine learning pipeline using the random forest algorithm to predict 4 levels of benthic disturbance; resulting classifications were compared to those obtained using a published taxonomy-based approach. Our results show that k-mer count features can effectively be used to create highly accurate predictions of benthic disturbance and can resolve intermediate changes in seafloor condition. In addition, we present a robust assessment of model performance which accounts for the effect of randomness in model creation. This work outlines a flexible framework for environmental assessments at aquaculture sites that is highly reproducible and free of taxonomy-assignment bias.

KEY WORDS: Aquaculture · Machine learning · Environmental monitoring · Organic enrichment · Bacterial eDNA · Random forest · Supervised classification

## 1. INTRODUCTION

Aquaculture is a significant global industry producing over 80 million t of food fish annually (FAO 2018). Over the last 3 decades, the industry has seen continued growth in production and now contributes up to 46% of the total global fish output, including capture fisheries and aquaculture fisheries combined (FAO 2018). However, concerns exist about the sustainability of aquaculture operations, in part due to the potential for negative environmental modification of associated ecosystems (Keeley et al. 2014, Salvo et al. 2017, Verhoeven et al. 2018). Effluent and particulate matter from aquaculture operations released into the environment can drive significant benthic community changes, the detrimental effects of which have been widely studied, and extended aquaculture activities typically lead to changes in macrofaunal succession, decline in species diversity and in some cases, complete elimination of native infauna (Keeley et al. 2014, Stoeck et al. 2018). Furthermore, studies have shown that these effects are long-lasting, as they persist years after aquaculture operations have ceased (Verhoeven et al. 2018).

*These authors contributed equally to this work
**Corresponding author: verhoevenjtp@googlemail.com



Publisher: Inter-Research · www.int-res.com

The surveillance of aquaculture environmental impact has therefore become a priority. Tracking aquaculture-associated detrimental changes is often performed through environmental monitoring and impact assessment programs, and typical approaches include the characterization of macrofaunal biodiversity, as well as the detection of the presence/absence of specific indicator species associated with both ecosystem health and disturbance (Keeley et al. 2014, Salvo et al. 2017, Hamoutene et al. 2018). While effective, these methods are comparatively labor-intensive as extensive imaging data or a large number of environmental samples need to be collected, and taxonomic expertise and labor are required to obtain, analyze and interpret results (Maurer 2000, Cordier et al. 2019).

More recently, the use of high-throughput sequencing to characterize microbial communities has been explored as a more streamlined and automatable method for detecting ecosystem change (Cordier et al. 2019, He et al. 2019). Often, these methods involve the amplification and sequencing of a marker gene (for bacteria, typically a portion of the 16S rRNA gene) and combining closely related sequences into operational taxonomic units (OTUs), which can then subsequently be used for elucidating the taxonomic composition of a community and gather information on the relative abundance of occurring OTUs (Pollock et al. 2018).

Since microbial communities are sensitive to environmental stimuli (Logue et al. 2015), previous work has highlighted the potential of using shifts in their taxonomic and OTU composition, as well as detecting the presence of specific biomarker taxa in microbial communities, to infer aquaculture environmental impact and organic enrichment at fish farms (Verhoeven et al. 2016, 2018, Stoeck et al. 2018). Nevertheless, several limitations can make the use of sequence- and taxonomic-based approaches suboptimal. Taxonomic classification is inherently limited to classifying sequences for microorganisms that are identical or highly similar to those present within the reference databases used, which can lead to a large proportion of sequences unclassified or classified at a less informative taxonomic level (Youssef et al. 2015), and thus unusable for biomarker studies. In addition, typical amplicon sequencing experiments produce high dimensional and sparse OTU datasets representing the complete genotypic diversity present in each investigated sample, from which extracting specific or co-occurring features significantly related to an ecosystem status can be challenging and computationally expensive (Gloor et al. 2017).

Such challenges can in part be addressed by combining marker-gene analysis with supervised machine learning (SML). SML algorithms generate predictive models based on user-supplied training datasets, from which specific features (or combination of features) correlating to the known classification are autonomously detected. Once established, this model can subsequently be used to predict a classification for future, unknown, samples.

Within the context of biomonitoring, the integration of SML has enabled new approaches in analyzing amplicon data, including the possibility of employing a taxonomy- and reference database-free approach, using OTU sequences directly as inherent features of investigated environments. Recent work has shown that not only are OTU SML-based approaches capable of accurately predicting environmental biotic index values, but they also outperform the traditional, taxonomy-based assessment of these indices (Cordier et al. 2018). However, grouping sequences into OTUs has several undesirable properties, including sensitivity to the used bioinformatic pipeline and associated settings causing variations in OTU composition, biases due to the possibility of combining closely related sequences into phylogenetically incoherent OTUs and the inherent inability to compare OTUs from different datasets, as the boundaries and membership of OTUs are dependent on, and invalid outside, the dataset in which they are defined (Callahan et al. 2017).

As an alternative, the distribution of oligonucleotides of specific length (k-mers), calculated from biological sequences, can be used as input features for performing machine learning (Asgari et al. 2018). Oligonucleotide distributions are a well-defined representation of 16S rRNA amplicon sequence data, in which sequence similarities are naturally incorporated, and are robust to bioinformatic pipeline and parameter variations, making them a particularly well-suited feature set for downstream machine learning (Asgari et al. 2018). Indeed, recent studies have shown that k-mer representations of 16S rRNA gene sequencing experiments contain sufficient information for SML to accurately predict the phenotypical and environmental characteristics of biological samples in a variety of applications (Asgari et al. 2018). As such, the usage of oligonucleotide distributions, coupled with SML, can potentially be a valuable tool in assessing changes to specific environmental niches in response to external stimuli, such as anthropogenic impacts.

During previous studies, we used 16S rRNA gene sequencing to demonstrate that salmon aquaculture

operations can create long-lasting significant benthic disturbances that in turn drive large-scale specific shifts in benthic bacterial populations (Verhoeven et al. 2016, 2018). Here, we reanalyzed 16S rRNA gene sequencing data from our previous study and investigated the potential and benefits of utilizing oligonucleotide distribution representations, specifically tetranucleotide frequencies (TNFs, k = 4), over conventional OTU counts in an SML setting, as a possible automated method for predicting benthic disturbance levels.

## 2. MATERIALS AND METHODS

### 2.1. Sample and data description

This study examined a previously investigated microbiome dataset (NCBI BioProject PRJNA503189) containing Illumina-based sequencing data of the V6–V8 16S rRNA gene region performed on 108 flocculent matter samples collected below and near salmon aquaculture operations in Newfoundland, Canada (Verhoeven et al. 2018). An Eckman grab was used to collect samples, comprised of either naturally occurring sedimentation or aquaculture-associated flocculent matter. In order to better capture the existing bacterial diversity, up to 3 sub-samples were collected from each successful grab (Verhoeven et al. 2018). Samples were previously assigned an environmental impact interpretation based on sample metadata, percentage of total organic carbon (%TOC) measurements, as well as bacterial taxonomic composition and diversity, and subsequently categorized as low impact (N = 34, similar to sites with no aquaculture, low %TOC), recently disturbed (N = 13, deviating from the low-impact group, elevated %TOC), intermediate impact (N = 19, drop in biodiversity, significantly increased %TOC) or high impact (N = 42, lowest biodiversity, highest levels of %TOC) (Verhoeven et al. 2018).

### 2.2. TNF and OTU calculation

TNFs were calculated per sample by using a sliding window (k = 4) across all sequences for each sample, summing the occurrences of tetranucleotides in a matrix. TNF occurrence count data were then subsequently normalized using the centered log ratio (clr) transform available in the 'codaSeq' R package (Gloor & Reid 2016). Similarly, sample-stratified OTU count data (generated as in Verhoeven et al. 2018) were imported, tabulated and clr-transformed. Both of these sets were used independently as input for the SML analyses to compare the efficiency and accuracy of the developed methods on different dataset types.

### 2.3. SML workflow

Model creation and statistical analysis (code available upon request) were performed in R (v3.5.2) using the RStudio v1.1.463 IDE (R Core Team 2015). The 'caret' package (v6.0-81) was used for data partitioning, cross validation, hyperparameter optimization and model fitting (Kuhn, 2008). Stratified random sampling was performed with 'caret::createData Partition' to maintain class ratios between training and test sets. We included 75% of observations (N = 83) for model training, with 25% (N = 25) withheld to evaluate model performance on unseen data. Predictive models were trained with 'ranger' (v0.11.1), a multithreaded implementation of the random forest algorithm (Wright & Ziegler 2017). All visualizations were created with 'ggplot2' (Wickham 2009).

The 'caret::trainControl' function was used to specify resampling and hyperparameter search methods. We used repeated, stratified 10-fold cross validation (CV) to search the hyperparameter space for settings minimizing classification error. Ten folds were selected to reduce variance and to ensure that at least 1 of each class was present in each partition. We performed 100 repetitions to account for covariate drift during division of training examples into folds and to ensure that performance estimates had stabilized (Moreno-Torres et al. 2012). Hyperparameter tuning was done via grid search over CV folds with the best performing hyperparameter tuple fit to the entirety of training data. Tuned hyperparameters included: (1) 'mtry': number of features randomly selected as candidates for each split, (2) 'splitrule': the split quality evaluation function and (3) 'num.trees:' the number of trees in a forest. The number of trees was set to 2001, with an odd number specified to ensure no ties could occur during generation of class predictions. The number of trees was set to a large, computationally feasible number which balances reductions in model variance with diminishing returns as tree count increases (Oshiro et. al. 2012, Probst & Boulesteix 2018). All 256 TNF combinations were included as features with values corresponding to their frequencies after clr transformation. 'Caret::confusionMatrix.train' and 'caret::confusionMatrix' were used to create confusion matrices and summary statistics for results from CV and test set predictions.

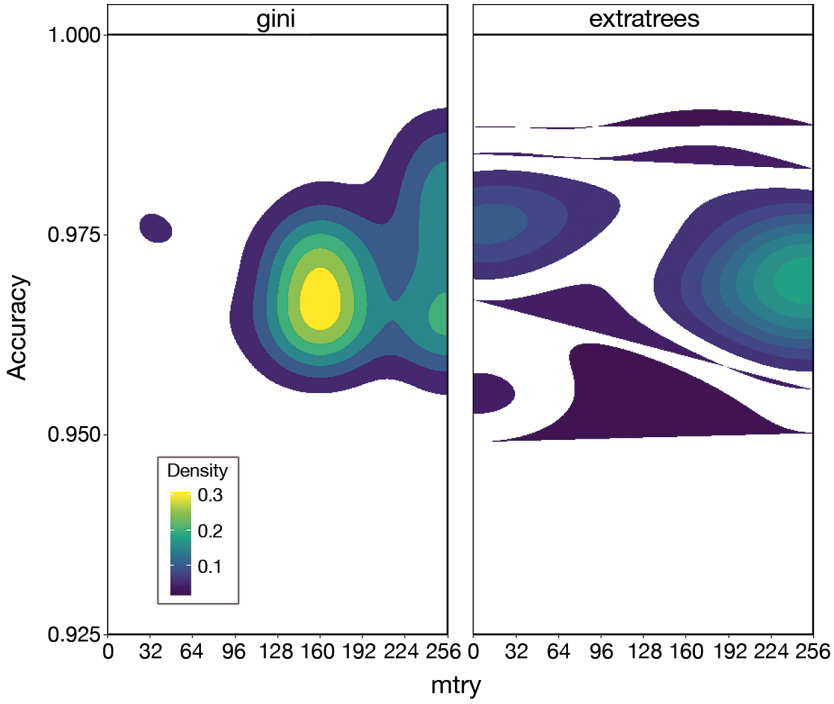


Fig. 1. Cross-validation performance of predictive models with varying hyperparameters. Shown are the accuracy of aggregated validation scores created through repeated 10-fold cross validation on the vertical axis, with the number of randomly selected features as split candidates ('mtry') indicated on the horizontal axis. The density of models reporting similar performance and hyperparameters is indicated by warmer colors. The left and right plots correspond to gini and extratrees ('splitrule'), respectively, which determine how the algorithm creates decision tree splits

To account for the effect of randomness on model performance, a for-loop generating 100 random seeds was created, with each iteration resulting in unique train/test splits, hyperparameter optimizations, predictions and patterns of tree growth. The results of these 100 models were aggregated into a single confusion matrix to demonstrate overall model performance and stability despite randomness inherent in the model-building process.

## 3. RESULTS

### 3.1. Average accuracy of resampling folds

Hyperparameters resulting in the lowest classification error were fit to the entirety of training data by evaluating differing hyperparameter settings on folds in our repeated CV procedure. Predictions made during CV were aggregated to provide initial estimates of model performance and stability. CV and hyperparameter tuning were performed independently for each random seed iteration. While computations based on the full OTU-based dataset did not finish due to excessive memory usage, TNF-based predictions generated on resampling folds resulted in an average accuracy of 0.9704, with the lowest performance reported at 0.9506 (Fig. 1). Performance was comparable between the gini and extratrees 'splitrule' hyperparameter, with the most frequent hyperparameters being gini with an 'mtry' of 160 features (Fig. 1).

Low-impact samples were most consistently predicted, with 0.991 of cases accurately reported (Table 1). Conversely, the intermediate disturbance level was the least accurate category (0.904), with incorrect classifications being labeled as low and recently impacted for 6.8 and 2.7% of cases, respectively (Table 1). Similarly, 2.3% of high-impact predictions were misclassified as intermediate impact (Table 1). Overall, only 1.5% of observations were misclassified by >1 level of impact (Table 1).

### 3.2. Model evaluation on withheld test data

Predicted labels on withheld TNF-based test data showed a high level of agreement with known cases for all disturbance levels (Table 2, Fig. 2). All model predictions significantly outperformed (median p =

Table 1. Confusion matrix of aggregated counts from 100 rounds of 10-fold, 100 repetition cross validation created during a hyperparameter search (N = 83) with randomly sampled seed states. Column and row values correspond to known and predicted cases of seafloor disturbance, respectively, with 4 levels of seafloor disturbance ranging from low to high. Numbers represent the result of 100 seed iterations of independent train/test splits

| Predicted impact | Actual impact | | | |
|---|---|---|---|---|
| | Low | Recent | Intermediate | High |
| Low | 257765 | 55 | 10274 | 0 |
| Recent | 0 | 99653 | 4044 | 63 |
| Intermediate | 2226 | 133 | 135638 | 7565 |
| High | 9 | 159 | 44 | 312372 |

Table 2. Confusion matrix demonstrating model performance on withheld test data (N = 25) when predicting levels of seafloor disturbance ranging from low to high. Column and row values correspond to known and predicted cases of seafloor disturbance, respectively, with 4 levels of seafloor disturbance ranging from low to high. Numbers represent the result of 100 seed iterations of independent train/test splits

| Predicted impact | Actual impact | | | |
|---|---|---|---|---|
| | Low | Recent | Intermediate | High |
| Low | 782 | 0 | 40 | 0 |
| Recent | 0 | 300 | 16 | 0 |
| Intermediate | 16 | 0 | 334 | 20 |
| High | 2 | 0 | 10 | 980 |

$4.335 \times 10^{-9}$, see Table S1 in the Supplement at www.int-res.com/articles/suppl/q012p131_supp.pdf) the 'no information rate', which represents a naïve prediction of all observations belonging to the majority class. In addition, random seed iteration testing indicates that the random seed state did not significantly impact accuracy scores (Table S1), with mean and median model accuracies of 0.9584 and 0.96 being detected, respectively. In addition, models fell within the 95% CI created with 'caret::confusionMatrix,' which performs an exact binomial test to determine the probability of success in a Bernoulli experiment. OTU-based prediction did not successfully complete due to memory constraints.

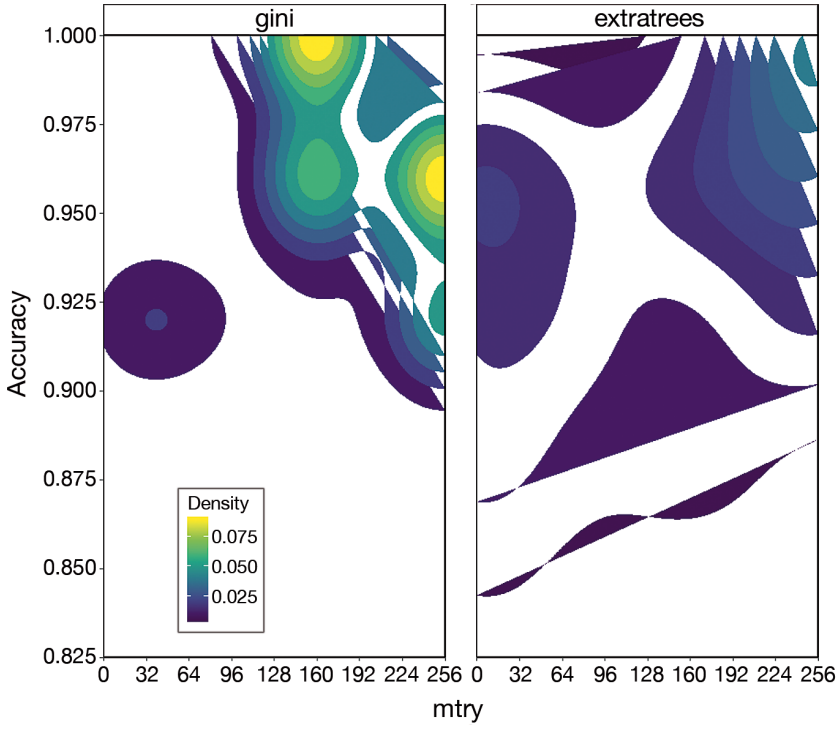


Fig. 2. Test set performance of predictive models with randomized seed states. Shown are the prediction accuracies of trained models on test set data, with the number of randomly selected features as split candidates ('mtry') indicated on the horizontal axis. The density of models reporting similar performance and hyperparameters is indicated by warmer colors. The left and right plots correspond to gini and extratrees ('splitrule'), respectively, which determine how the algorithm creates decision tree splits

## 4. DISCUSSION

The classification of benthic disturbances near aquaculture sites has received increased attention, but its practical application and the potential for real-time assessment have yet to be introduced. Here, we examined the use of k-mer count features (k = 4) in a model tasked with classifying levels of benthic disturbance at aquaculture sites and demonstrate that highly accurate predictions of seafloor condition can be generated using a defined feature set.

Traditionally, biotic indices which examine macroinvertebrate richness and diversity have been used to assess ecological quality and disturbance level (Borja & Dauer 2008, Rygg & Norling 2013). However, more recently, bacterial eDNA metabarcoding has shown promise by associating specific bacterial community compositions with environmental disturbances, demonstrating the potential of bacteria as highly responsive bioindicators (Lejzerowicz et al. 2015, Stoeck et al. 2018). However, sequence-analysis pipelines are not standardized, and most of them rely on sequence taxonomy assignment, a process that is heavily affected by the available knowledge of sequenced microbial taxa and may fail to provide accurate data when in the presence of a high amount of highly divergent taxa. Machine learning, in conjunction with features derived from 16s rRNA sequencing (such as OTUs) as model inputs, can effectively predict biotic indices (Cordier et al. 2017), outperforming taxonomy-based assessments and providing faster evaluation of seafloor conditions (Cordier et al. 2018). In this context, k-mer count features (such as TNFs) have been used to accurately classify distinct ecological environments (Asgari et al.

2018), but their ability to resolve within-environmental change has not been previously addressed.

In this study, we demonstrate that TNFs can effectively distinguish levels of benthic disturbance with performance maintained across 100 randomized iterations of cross-validation (Table 1) and holdout test data (Table 2). While longer k-mer count features have been found to improve classification performance (Alsop & Raymond 2013, Vervier et al. 2016, Asgari et al. 2018), the discriminatory power of tetranucleotides is well documented (Teeling et al. 2004, Yoon et al. 2017), and their use in the current context balances performance with computation time by limiting the number of features, which increase quartically with k-mer length. Furthermore, the use of TNF features in a supervised classifier circumvents taxonomic assumptions associated with seafloor condition and simplifies data processing by restricting the feature set to 256 tetranucleotide combinations (Asgari et al. 2018). This is a desirable quality in developing monitoring pipelines as it standardizes predictive model inputs. TNF-based classifications do not require sequence alignments and reference databases to identify bacterial groups nor the construction of OTUs, which can vary depending on settings used in diverse bioinformatics pipelines and may not reflect genuine taxonomic relationships, introducing multiple levels of biases. Furthermore, OTU construction and taxonomy-based approaches have emergent and location-dependent features which are difficult, if not impossible, to standardize, and comparisons between sample sites are not possible with OTUs constructed from different datasets. The taxonomy-independent nature of the TNF-based method presented here makes it intrinsically less sensitive to these variables and more suitable for comparisons across locations and datasets (Callahan et al. 2017). Additionally, we were unable to successfully use SML approaches in combination with the full dataset of OTU-based count data due to the high-dimensional size that led to unsustainable computational requirements. While filtering OTUs to reduce the dimensionality of the feature space could have been performed to reduce computational demands, valuable information on potentially important rare taxa would have been lost (Wang et al. 2017), a problem which is not observed within the TNF-based method, as all available data are compressed into the respective TNFs.

While the utility of machine learning approaches, like those used in this paper, is indisputable, concerns regarding reproducibility of machine learning algorithms and the reporting of model performance have been raised (Drummond 2009, Henderson et al. 2017, Colas et al. 2018). Setting seed states allows random events, such as partitioning data into training and test sets to be replicated and compared, but replication may not be sufficient to arrive at genuine performance estimates (Drummond, 2009). In several cases, the best or average of n-best performing seeds is selected for publication (Henderson et al. 2017, Colas et al. 2018). This behavior of seed optimization is problematic as it allows investigators to report seeds resulting in good performance without disclosing trials with poorer outcomes or those which do not improve upon currently existing benchmarks. By including results from numerous seed states we demonstrated that our model is stable over random iterations accounting for differences in train/test splits and patterns of tree growth. Model stability over 100 train/test splits accounts for variance associated with the small (but representative) sample retained for model evaluation (N = 25). When computationally feasible, we recommend that these statistics be reported.

In conclusion, k-mer features such as TNF are a valuable addition to the benthic assessment toolkit, reducing computation costs associated with sequence alignment and reference database comparison while outperforming OTU and taxonomic features when predicting environment types (Asgari et al. 2018). Future studies should examine larger collections over wider geographic areas as to better characterize the robustness of seafloor condition boundaries and assess the generalizability of predictions over larger spatial scales. Additionally, the establishment of an open-source database of sequenced samples near aquaculture sites and the inclusion of different 16s rRNA hypervariable regions could provide increased flexibility to seafloor condition classifications and allow investigators to detect changes at high resolution with a variety of eDNA-sequencing pipelines.

*Acknowledgements*. This research was undertaken thanks in part to funding from the Canada First Research Excellence Fund, through the Ocean Frontier Institute. We thank Dr. Suzanne Dufour, Dr. Flora Salvo and Dr. Dounia Hamoutene for their help in realizing this study and manuscript.

## LITERATURE CITED

Alsop EB, Raymond J (2013) Resolving prokaryotic taxonomy without rRNA: longer oligonucleotide word lengths improve genome and metagenome taxonomic classification. PLOS ONE 8:e67337

Asgari E, Garakani K, McHardy AC, Mofrad MRK (2018) MicroPheno: predicting environments and host phenotypes from 16S rRNA gene sequencing using a k-mer

based representation of shallow sub-samples. Bioinformatics 34:i32–i42

Borja A, Dauer DM (2008) Assessing the environmental quality status in estuarine and coastal systems: comparing methodologies and indices. Ecol Indic 8:331–337

Callahan BJ, McMurdie PJ, Holmes SP (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J 11: 2639–2643

Colas C, Sigaud O, Oudeyer PY (2018) How many random seeds? Statistical power analysis in deep reinforcement learning experiments. arXiv:1806.08295

Cordier T, Esling P, Lejzerowicz F, Visco J and others (2017) Predicting the ecological quality status of marine environments from eDNA metabarcoding data using supervised machine learning. Environ Sci Technol 51: 9118–9126

Cordier T, Forster D, Dufresne Y, Martins CIM, Stoeck T, Pawlowski J (2018) Supervised machine learning outperforms taxonomy-based environmental DNA metabarcoding applied to biomonitoring. Mol Ecol Resour 18: 1381–1391

Cordier T, Lanzén A, Apothéloz-Perret-Gentil L, Stoeck T, Pawlowski J (2019) Embracing environmental genomics and machine learning for routine biomonitoring. Trends Microbiol 27:387–397

Drummond C (2009) Replicability is not reproducibility: nor is it good science. Proc Evaluation Methods for Machine Learning Workshop at the 26th ICML, Montreal, Canada, 2009. National Research Council Canada, Ottawa

FAO (2018) The state of world fisheries and aquaculture 2018—meeting the sustainable development goals. FAO, Rome

Gloor GB, Reid G (2016) Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. Can J Microbiol 62:692–703

Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ (2017) Microbiome datasets are compositional: and this is not optional. Front Microbiol 8:2224

Hamoutene D, Salvo F, Cross S, Dufour SC, Donnet S (2018) Linking the presence of visual indicators of aquaculture deposition to changes in epibenthic richness at finfish sites installed over hard bottom substrates. Environ Monit Assess 190:750

He X, Sutherland TF, Pawlowski J, Abbott CL (2019) Responses of foraminifera communities to aquaculture-derived organic enrichment as revealed by environmental DNA metabarcoding. Mol Ecol 28:1138–1153

Henderson P, Islam R, Bachman P, Pineau J, Precup D, Meger D (2017) Deep reinforcement learning that matters. arXiv:1709.06560

Keeley NB, Macleod CK, Hopkins GA, Forrest BM (2014) Spatial and temporal dynamics in macrobenthos during recovery from salmon farm induced organic enrichment: When is recovery complete? Mar Pollut Bull 80: 250–262

Kuhn M (2008) Building predictive models in R using the caret package. J Stat Softw 28:1–26

Lejzerowicz F, Esling P, Pillet L, Wilding TA, Black KD, Pawlowski J (2015) High-throughput sequencing and morphology perform equally well for benthic monitoring of marine ecosystems. Sci Rep 5:13932

Logue JB, Findlay SEG, Comte J (2015) Editorial: Microbial responses to environmental changes. Front Microbiol 6:1364

Maurer D (2000) The dark side of taxonomic sufficiency (TS). Mar Pollut Bull 40:98–101

Moreno-Torres JG, Saez JA, Herrera F (2012) Study on the impact of partition-induced dataset shift on *k*-fold cross-validation. IEEE Trans Neural Netw Learn Syst 23: 1304–1312

Oshiro TM, Perez PS, Baranauskas JA (2012) How many trees in a random forest? In: Perner P (ed) Machine learning and data mining in pattern recognition. Lecture notes in computer science, Springer, Berlin, p 154–168

Pollock J, Glendinning L, Wisedchanwet T, Watson M (2018) The madness of microbiome: attempting to find consensus 'best practice' for 16S microbiome studies. Appl Environ Microbiol 84:e02627–17

Probst P, Boulesteix AL (2018) To tune or not to tune the number of trees in random forest. J Mach Learn Res 18(181):1–18

R Core Team (2015) R: a language and environment for statistical computing. www.R-project.org/

Rygg B, Norling K (2013) Norwegian Sensitivity Index (NSI) for marine macroinvertebrates, and an update of Indicator Species Index (ISI). NIVA Rapport 6475. Norwegian Institute for Water Research, Oslo

Salvo F, Mersereau J, Hamoutene D, Belley R, Dufour SC (2017) Spatial and temporal changes in epibenthic communities at deep, hard bottom aquaculture sites in Newfoundland. Ecol Indic 76:207–218

Stoeck T, Frühe L, Forster D, Cordier T, Martins CIM, Pawlowski J (2018) Environmental DNA metabarcoding of benthic bacterial communities indicates the benthic footprint of salmon aquaculture. Mar Pollut Bull 127:139–149

Teeling H, Meyerdierks A, Bauer M, Amann R, Glöckner FO (2004) Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ Microbiol 6: 938–947

Verhoeven JTP, Salvo F, Hamoutene D, Dufour SC (2016) Bacterial community composition of flocculent matter under a salmonid aquaculture site in Newfoundland, Canada. Aquacult Environ Interact 8:637–646

Verhoeven JTP, Salvo F, Knight R, Hamoutene D, Dufour S (2018) Temporal bacterial surveillance of salmon aquaculture sites indicates a long lasting benthic impact with minimal recovery. Front Microbiol 9:3054

Vervier K, Mahé P, Tournoud M, Veyrieras JB, Vert JP (2016) Large-scale machine learning for metagenomics sequence classification. Bioinformatics 32:1023–1032

Wang Y, Hatt JK, Tsementzi D, Rodriguez-R LM and others (2017) Quantifying the importance of the rare biosphere for microbial community response to organic pollutants in a freshwater ecosystem. Appl Environ Microbiol 83: e03321-16

Wickham H (2009) ggplot2: Elegant graphics for data analysis. Springer-Verlag, New York, NY

Wright MN, Ziegler A (2017) ranger: A fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw 77:1–17

Yoon SH, Ha SM, Kwon S, Lim J, Kim Y, Seo H, Chun J (2017) Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. Int J Syst Evol Microbiol 67: 1613–1617

Youssef NH, Couger MB, McCully AL, Criado AEG, Elshahed MS (2015) Assessing the global phylum level diversity within the bacterial domain: a review. J Adv Res 6: 269–282

*Editorial responsibility: Peter Cranford, Dartmouth, Nova Scotia, Canada*

*Submitted: June 6, 2019; Accepted: February 14, 2020*
*Proofs received from author(s): March 25, 2020*