Biosensors and Bioelectronics 133 (2019) 64–71

Contents lists available at ScienceDirect

Biosensors and Bioelectronics

journal homepage: www.elsevier.com/locate/bios

Incorporating microbial community data with machine learning techniques to predict feed substrates in microbial fuel cells

Wenfang Cai[a,b], Keaton Larson Lesnik[b], Matthew J. Wade[c,d], Elizabeth S. Heidrich[c], Yunhai Wang[a,*], Hong Liu[b,*]

[a] Department of Environmental Science and Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[b] Department of Biological and Ecological Engineering, Oregon State University, Corvallis OR 97331, USA
[c] School of Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU, UK
[d] Department of Mathematics & Statistics, McMaster University, Hamilton, Canada L8S 4K1

ARTICLE INFO

Keywords:
Biosensors
Feed substrate
Microbial fuel cell
Microbial communities
Machine learning

ABSTRACT

The complicated interactions that occur in mixed-species biotechnologies, including biosensors, hinder chemical detection specificity. This lack of specificity limits applications in which biosensors may be deployed, such as those where an unknown feed substrate must be determined. The application of genomic data and well-developed data mining technologies can overcome these limitations and advance engineering development. In the present study, 69 samples with three different substrate types (acetate, carbohydrates and wastewater) collected from various laboratory environments were evaluated to determine the ability to identify feed substrates from the resultant microbial communities. Six machine learning algorithms with four different input variables were trained and evaluated on their ability to predict feed substrate from genomic datasets. The highest accuracies of 93 ± 6% and 92 ± 5% were obtained using NNET trained on datasets classified at the phylum and family taxonomic level, respectively. These accuracies corresponded to *kappa* values of 0.87 ± 0.10, 0.86 ± 0.09, respectively. Four out of six of the algorithms used maintained accuracies above 80% and *kappa* values higher than 0.66. Different sequencing method (Roche 454 or Illumina sequencing) did not affect the accuracies of all algorithms, except SVM at the phylum level. All algorithms trained on NMDS-compressed datasets obtained accuracies over 80%, while models trained on PCoA-compressed datasets presented a 10–30% reduction in accuracy. These results suggest that incorporating microbial community data with machine learning algorithms can be used for the prediction of feed substrate and for the potential improvement of MFC-based biosensor signal specificity, providing a new use of machine learning techniques that has substantial practical applications in biotechnological fields.

1. Introduction

Anthropogenic pollution has led to significant negative social, economic, and ecological problems impacting human health and the environment. These problems have acted as drivers towards the development of new techniques for detecting chemicals in water. Recent biotechnological advances have led to a wider application of specific quantitative or semi-quantitative methods for the detection of chemicals using biosensors (Chouler et al., 2018). Developing effective biosensors that are fast, easy to use, specific, and inexpensive remains a key challenge. Microbial fuel cell (MFC)-based biosensors continue to be explored as a long-term and cost-effective solution for environmental monitoring applications due to their capacity for self-regeneration and self- replication (Jiang et al., 2018; Wang et al., 2013).

Biosensor technology is founded on a specific biological recognition element in combination with a transducer for signal processing (Biswas et al., 2017). In MFC-based biosensors, the exoelectrogenic bacteria in the anodic biofilm or biocathode serve as a signal generator or biological recognition element, whilst the electrode acts as a transducer (Chang et al., 2004; Jiang et al., 2017a). Utilization of exoelectrogens as signal generators has been investigated for Biochemical Oxygen Demand (BOD) monitoring (Chang et al., 2004; Jiang et al., 2018) and chemical (Jiang et al., 2017b; Wang et al., 2018; Yang et al., 2017) and toxin detection (Di Lorenzo et al., 2014; Liu et al., 2014), where voltage change is used as the predictor variable. In chemical detection applications the electrical signals of the MFC-based biosensor using a mixed

[*] Corresponding authors.
E-mail addresses: wang.yunhai@mail.xjtu.edu.cn (Y. Wang), liuh@engr.orst.edu (H. Liu).

https://doi.org/10.1016/j.bios.2019.03.021
Received 31 December 2018; Received in revised form 11 March 2019; Accepted 12 March 2019
Available online 13 March 2019

microbial community are correlated to feed substrate. However, variation in anodic microbial activity and composition alters signal outputs (Kim et al., 2007; Sharma and Li, 2010). Electric signals in turn may lack specificity as influent components and the inoculum change (both in quantity and quality), thereby affecting both phenotypic expression in addition to variation in microbial abundance and species types (Kiely et al., 2011). Indirect factors in this system such as the operating conditions and the extracellular electron transfer rate add more complexity in determining correlation between chemical detection and microbial phenotypes (Ishii et al., 2014). Identifying the emergence of substrate-specific anodic communities will provide insight into the qualitative detection of chemical substrates, which will ultimately improve biosensor signal specificity. Furthermore, access to low-cost, high-throughput 16 S rRNA sequencing instruments allows for the incorporation of microbiome information into biosensing workflows to be increasingly feasible (Billard et al., 2012; Liu et al., 2012). However, no current research has directly connected chemical detection in biosensors to microbial population shifts. Developing methods capable of improving biosensor specificity for chemical detection under varying influent through incorporation of microbial population shifts poses a great challenge and is the focus of this work.

Well-developed physical-mathematical models describing MFC biofilm processes can accurately describe biofilm formation and some biochemical interactions in highly controlled settings (Cai et al., 2018; Ou et al., 2016; Stein et al., 2011). However, the sheer number of parameters required to accurately model more complex environments with mixed microbial communities leads to mathematical models that are impractical to use. Machine learning, as a data mining and model development tool, is being increasingly used in most branches of science and engineering, as methods develop and computing capacity improves (Krogh, 2008; Mjolsness and DeCoste, 2001). These techniques are used to discover latent interactions in large data matrices between the input features and output performance by deriving data-driven models. When predicting a specific process feature, data-mining techniques do not require detailed information or specifications of the system. These data-mining techniques have been applied in engineered biological systems for prediction, estimation, and process simulation. In MFCs, Artificial Neural Networks (ANN) were used to predict microbial community response to environmental perturbations, as well as functional ecosystem outcomes (Larsen et al., 2015; Lesnik and Liu, 2017). A Stacked Denoising Auto-Encoder (SDAE) deep learning network was used to predict the performance of a two-stage biofilm system based on traditional anaerobic/oxic processes (Shi and Xu, 2018). Supervised machine learning methods, such as discriminant analysis and Support Vector Machines were applied to measure organic acid concentration in digesters (Bongards et al., 2014). Generally, these machine learning techniques have been applied for microbial source tracking, wastewater treatment processes, and air quality monitoring (Dubinsky et al., 2016; Han and Qiao, 2013; Wu et al., 2017). Some studies have also attempted to use ANN to identify the presence of substrates using only the electrical signals from MFC-based biosensors, although accuracy was limited (Feng et al., 2013; Feng and Harper, 2013; King et al., 2014). Studies using genomic data as machine learning datasets input have been mainly focused on gene annotation, gene expression, pathway analysis, genetic association, and epistasis detection in biomedical field (Chen and Ishwaran, 2012; Libbrecht and Noble, 2015; Wang et al., 2019). According to published research, machine learning models that incorporate genomic data have not been applied in the biosensor field.

This study provides a proof-of-concept for the application of machine learning algorithms incorporating genomic data for the prediction of feed substrate and for the potential improvement of MFC-based biosensor signal specificity. We first collected the genomic data from 69 MFC samples from several source laboratories using different feed substrates, sequencing methods, and primers. We then used a number of machine learning algorithms to classify substrate type (wastewater (WW), carbohydrates (CARB, containing glucose, fructose, xylose, galactose and lactose) and acetate (AC)) from the genomic datasets. The algorithms, including logistic regression multiclass (GLMNET), random forest (RF), scalable tree boosting system (XGBOOST), neural network (NNET), K-nearest neighbor (KNN), and support vector machine with radial kernel (SVM), were compared to select the best algorithm for substrate type classification from the genomic data. The choice of suitable machine learning algorithms and identification of appropriate data inputs can provide a direct link between substrate groupings and genomic data without the need for further information, such as operation conditions and electric current, making this approach more widely applicable in systems with mixed microbial communities.

2. Materials and methods

2.1. Dataset details

A total of 69 samples were used in this study, with the genomic datasets provided from laboratory-scale experiments of Vilajeliu-Pons et al. (2016); Li et al. (2018); Heidrich et al. (2018); and Lesnik and Liu (2017). Sequencing of 36 samples was carried out by Roche 454 sequencing GS FLX Titanium Series and the remaining 33 samples were sequenced by Illumina sequencing primers for a 250-bp 221 paired-end run (v3) on the MiSeq platform. There are 36 samples for acetate feed, 27 samples for wastewater feed, and 6 samples for carbohydrates feed as detailed in Table S1.

2.2. Sequence analysis

Bio. SeqIO in BioPython (Chang et al., 2010) was used to convert the raw sequences obtained by 454 GS FLX from FASTA to FASTQ format, where the latter contains the original sequence data and quality score information. Raw sequence sample processing was performed using the DADA2 pipeline (Callahan et al., 2016) with the R-3.4.4 statistical software. Samples were demultiplexed and dual barcode/adapters sequences were removed. Initial quality pre-processing included removing sequence reads that were below an average phred quality of 20 using a 30 bp window, and trimming reads less than 75% of the original length, resulting in a 96% read retention. Further quality trimming involved removing sequences with three consecutive low-quality bases, ambiguous base calls, and setting a minimum sequence length of 200 after trimming. Duplicate samples were removed to transform the identical sequence reads into unique sequences with a corresponding abundance equal to the number of reads with that unique sequence. The divisive amplicon denoising algorithm (DADA) was used to remove PCR errors from the sequencing data (Callahan et al., 2016). Paired-end reads were merged together to obtain the full denoised sequences. Chimeras were also removed from the denoised sequences by identifying if they could be really reconstructed when a left-segment and a right-segment were combined from two or more abundant *parent* sequences. The taxonomy was assigned using the naïve Bayesian classifier method with the GreenGenes clustered at 97% identity and the Silva reference database (Wang et al., 2007). The taxonomic abundance analysis at family and phyla levels fed with different substrates were visualized by Krona (Ondov et al., 2011).

2.3. Statistical analysis

Multivariate analyses such as Principal Coordinate Analysis (PCoA) and Non-metric Multidimensional Scaling (NMDS) of the microbial communities were performed using the VEGAN package in R-3.4.4 (Dixon, 2003). Ordination distance metrics included Bray-Curtis (Beals, 1984), unweighted Unifrac, and weighted Unifrac (Lozupone and Knight, 2005). The alpha diversity of anode microbial communities operated with different feed substrates was calculated using the richness (the number of different species represented in one sample), Simpson (the probability of two randomly chosen individuals represent different species) (Lande, 1996) and Shannon (Hill et al., 2003) indices, which performed pair-wise ANOVA of diversity measured between groups with $p > 0.05$.

2.4. Machine learning algorithm development and evaluation

The capacity for machine learning algorithms to classify feed substrates in MFC based on genomic data was evaluated. Four input datasets were built at the family taxonomic level with a relative abundance higher than 5%, the phylum level with a relative abundance higher than 2%, and a dimensionally reduced dataset ordinated using PCoA and NMDS from the entire 32,952 amplicon sequence variants (ASVs) dataset. GLMNET, RF, XGBOOST, NNET, KNN, and SVM with radial kernel were tested with the four input datasets mentioned above for substrate classification. The GLMNET (generalized linear model via penalized maximum likelihood) algorithm uses lasso, ridge regression or an elastic net regularization penalty for estimation of generalized linear regression, binary classification and multinomial classification regression problems with a suitable penalty parameter lambda (λ) (Friedman et al., 2009). In this study, we used elastic net regularization with a λ value of 0.00023, the minimum cross-validation error. The RF algorithm for classification constructs a forest containing a multitude of decision tree classifiers. Each tree classifier is independent and is generated by a random vector sampled from input samples, and outputs a class of input vector that is ranked highest among all tree classifiers (Pal, 2007). The random forest classifier employed in this study uses randomly selected features (*mtry = 2*) to grow a tree with the minimum size of terminal nodes set as 1 (default for classification). The XGBOOST algorithm is an especially efficient implementation of gradient boosting decision trees and it can automatically perform parallel computation on a single machine, which is up to 10 times faster than gradient boosting (Chen and Guestrin, 2016). In this study, the maximum number of iterations was set to 200, the learning rate was 0.3, and the regularization value (*gamma*) was set to 5 according to loss function. The maximum depth of the tree is closely related to the algorithm fitting degree and is usually in the range of 3–10. In this study, its value was set to 6 after cross-validation. The minimum sum of instance weight of the leaf node was used to avoid overfitting, which was set to 1 by default. The proportion of random samples supplied to a tree based on columns usually ranges from 0.5 to 1 and was set to 0.8 in this study. The NNET algorithm used in this study was a feed-forward neural network with a single hidden layer (Ripley et al., 2016) and has 32 nodes in the hidden layer. The KNN algorithm is a nonlinear, supervised learning algorithm, where the *k* nearest (as Euclidean distance) training set vectors are found for each row of the test set, and the classification is decided by majority vote, with ties broken at random (Goldberger et al., 2005). If there are ties for the *k*th nearest vector, all candidates are included in the vote. The determining value of the nearest neighbor numbers (*k*) plays a significant role in determining the efficiency of the models. A large *k* value can reduce the variance due to noisy data but can introduce a bias by which the learner may ignore smaller, but significant patterns. Thus, the *k*-value in this study was set to 10 as a compromise for these effects. The support vector machine algorithm is a very strong classification technique for linear problems that does not use a probabilistic model like many other classifiers, but simply generates hyper-planes or line vectors to separate and classify the data into distinct feature spaces (Pal, 2007). However, for non-linear problems, this approach will cause linear separators and linear decision boundaries to fail. As the interaction between substrate and microbial populations is non-linear, then non-linear decision boundaries need to be generated, accordingly. The SVM with radial kernel was used in this study enabling separation of non-linear data by feature expansion and to decrease the variance (Scholkopf et al., 1997). For each algorithm, data was randomly separated into training ($n = 45$) and validation groups ($n = 24$). The models were trained, evaluated and tested using the Caret R package (Kuhn, 2008), and the performance of each algorithm was evaluated by accuracy and *kappa* (inter-rater agreement for categorical items, calculated by a confusion matrix) of the validation datasets. The validation was resampled 30 times to get an average accuracy and *kappa* values. Paired *t*-tests were used to determine the algorithm significance ($p > 0.05$) between the models and the input types.

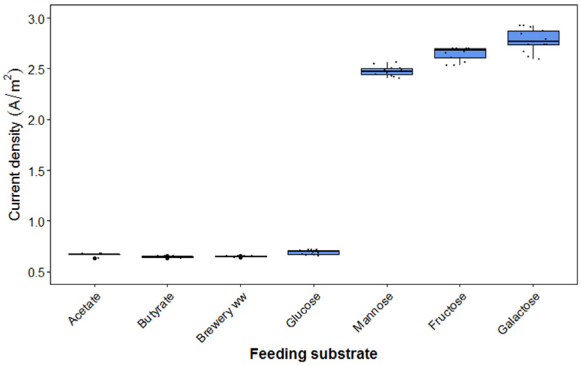

Fig. 1. Performance of single-chamber MFCs fed with different substrates.

3. Results and discussion

3.1. Specificity of MFC-based biosensors

In traditional MFC-based biosensors the sensing signals are generally electric current, electrical potential, electric conductance or impedance (Biswas et al., 2017; Quek et al., 2015). Current densities from single chamber air-cathode MFCs with different feed substrates were collected from previous studies Liu et al. (2005); Catal et al. (2008); Feng et al. (2008); Liu and Logan (2004) and then normalized to anode size (Fig. 1). MFC current density is typically affected by many factors, such as reactor configuration, electrode material, electrode distance, and operational condition. While these MFCs reactors all had a similar reactor configuration with the same type of electrode materials and were operated under similar environmental conditions, such as pH and temperature, some MFCs had an electrode distance of 4 cm and others had an electrode distance of 2 cm. Steady-state current densities of the MFCs with 4 cm electrode distance were all around 0.65–0.70 A/m^2 and did not show the specificity for the different substrates (acetate, brewery wastewater, or glucose) (Fig. 1). The MFCs with 2 cm electrode distance demonstrated higher current densities 2.5–2.78 A/m^2, but also did not show the specificity for the different substrates (galactose, mannose, or Fructose). In addition, Feng et al. (2013) found that the steady-state current densities of MFCs fed with butyrate, glucose, or corn starch were all around 0.01–0.02 A/m^2, further suggested the lack of specificity for substrates using traditional MFC-based biosensors. All these substrates could be degraded by the microbial communities in the MFC biofilms, with the exoelectrogens then responsible for transferring electrons to the anode. While acetate can be easily utilized by exoelectrogens for direct extracellular electron transfer to the anode, other substrates may require various different pathways involving many different organisms, eventually being converted to the substrate that can be utilized by exoelectrogens to generate the electric sensing signal (Gieg et al., 2014). Identifying the community that represents these complex pathways for each specific substrate would increase the specificity for the MFC-based biosensors for detecting these organic matters.

3.2. Characterization of dataset and sample diversity

The dataset used in this study contained 69 samples from different laboratory-scale experiments. Fig. S1 shows the sample numbers of the different types of feed substrates; acetate (52.2%), wastewater (39.1%), and carbohydrates (8.7%). Due to the skewed data distribution across the feed types, a combination of accuracy and *kappa* was used to evaluate the imbalanced dataset (Korotcov et al., 2017).

A variation in microbial community structure and alpha diversity based on all OTUs was observed between the samples fed on different substrates. MFCs fed with acetate had the lowest richness (224 ± 85) with coefficient of variation (C.V) of 38.20%, whereas carbohydrate fed MFCs showed the highest richness (799 ± 74) with C.V of 9.78%, with the wastewater samples having intermediate richness (411 ± 265) with C.V of 64.34%, shown in Fig. S2. This dissimilar distribution of richness indicated that divergent feed substrate resulted in large differences in microbial community diversity. The sample Shannon indices showed the same trend as for the richness, with significantly higher diversity in carbohydrate-fed MFCs (5.75 ± 0.40) than in acetate (4.52 ± 0.67), see Fig. S2a. The higher C.V of the richness and Shannon indices fed with acetate was possibly due to the different inoculum, while that higher C.V of the richness and Shannon indices fed with wastewater was likely due to the organic complexity of wastewater. The Simpson indices of these three substrates were heterogeneous as well, confirming that MFCs fed with different substrates promote distinct microbial community structure. The phylogenetic variation (Beta diversity), which was measured via weighted UniFrac distances (Lozupone and Knight, 2005), is displayed in Fig. 2 using the first two principal coordinates (combined variance of 32.5%) to visualize the dissimilarity distances and variation between samples. The UniFrac distance between samples represents the dissimilarity between microbial communities related to the fraction of evolutionary history in a phylogenetic tree that is unique to one of the communities rather than shared by both (Ramette, 2007). The PCoA showed a clear distinction between AC and WW/CARB, which indicated the distinct communities enriched from acetate compared to other feed substrates. The second coordinate (PC2) indicates a wide spread between wastewater samples, whereas carbohydrate samples cluster tightly, suggesting that the carbohydrates fed microbial community are homogeneous and wastewater fed communities are more diverse, as expected given the complex nature of the feed. This result is contrary to the overall OTU-based alpha diversity analysis (Fig. S2), possibly due to the fact that the microbial community diversity was phylogenetically similar but highly variable at different taxonomic ranks for divergent carbohydrates degradation (Fig. S3). Similar results were also found with the PCoA and NMDS analysis using unweighted Unifrac and weighted Unifrac metrics, respectively (Fig. S4). The PCoA and NMDS statistical analysis results further indicated the ability to predict feed substrates in MFC incorporating microbial community analysis.

To facilitate data visualization and improve prediction performance, random forest variable importance measures were used to determine the impact of each predictor variable individually as well as in multivariate interactions with other predictor variables (Strobl et al., 2008). The variables evaluated were the relative abundances of dominant phyla across all samples. The phyla determined to be the best predictors of substrate type according to correlation and impact on the response prediction were *Spirochaetae*, *Synergistetes*, *Actinobacteria*, *Bacteroidetes*, *Tenericutes*, *Deferribacteres*, *Proteobacteria*, and *Firmicutes* (Fig. 3a). MFC samples fed acetate had greater relative abundance of *Proteobacteria* (46.38 ± 13.42%), *Deferribacteres* (1.88 ± 3.23%), *Actinoacteria* (5.70 ± 4.89%), and *Bacteroidetes* (24.02 ± 8.51%) than that fed carbohydrates and wastewater. Prominent phyla in the MFC samples fed with carbohydrates included *Synergistetes* (32.27 ± 9.39%), *Spirochaetae* (11.18 ± 6.15%) and *Firmicutes* (23.24 ± 16.76%). Dominant phyla with wastewater included *Proteobacteria* (37.99 ± 26.56%), *Bacteroidetes* (19.96 ± 10.74%), *Firmicutes* (18.64 ± 19.57%) and *Euryarchaeota* (8.86 ± 17.19%) (Fig. 4).

The most influential predictor variables at the family level included *Phyllobacteriaceae*, *Oceanospirillaceae*, *Spirochaetaceae*, *Synergistaceae*, *Clostridiacear_1*, *Campylobacteraceae*, *Porphyromonadaceae*, and *Geobacteraceae* (Fig. 3b). Samples fed with acetate showed increased abundance of *Phyllobacteriaceae* (2.93 ± 1.79%), *Oceanospirillaceae* (2.65 ± 2.93%), *Synergistaceae* (9.89 ± 7.32%), *Porphyromonadaceae (7.73 ± 2.90%)*, *Rikenellaceae* (6.44 ± 3.81%), *Geobacteraceae* (18.10 ± 14.85%), *Lentimicrobiaceae* (2.58 ± 2.87%) and *Deferribacteraceae* (1.99 ± 3.14%) (Fig. 4a). Families in the samples fed with carbohydrates had a high abundance of *Synergistaceae* (33.47 ± 9.86%), *Spirochaetaceae* (11.60 ± 6.43%*)*, *Eubacteriaceae* (14.51 ± 18.52%) and *Actinomycetaceae* (3.36 ± 3.52%), and *Geobacteraceae* (6.42 ± 5.56%), whereas dominant families with wastewater fed samples included higher abundances of *Porphyromonadaceae* (7.73 ± 5.96%), *Rikenellaceae* (7.40 ± 5.48%), *Geobacteraceae* (14.35 ± 13.50%), *Erysipelotrichaceae* (10.67 ± 3.32%), and *Clostridiacear_1* (1.44 ± 2.60%) (Fig. 4b and c). *Proteobacteria* families of *Phyllobacteriaceae*, *Geobacteraceae*, *Oceanoscpirillaceae*, *Campylobacteraceae* and *Brucellaceae* were increased in the acetate fed MFCs compared to those fed carbohydrates and wastewater. However, overall the *Proteobacteria* phylum was a relatively minor important feature in these datasets (Fig. 3a). Even so, *Phyllobacteriaceae* and *Oceanoscpirillaceae* were the strongest predictors among all families, although they were the low abundant constituents (Fig. 3b). *Porphyromonadaceae* and *Rikenellaceae* family belong to the *Bacteroidetes* phyla, and their abundances with acetate and wastewater were homogeneous, while their influences on model prediction were significantly different (Figs. 3b and 4). The difference of microbial community abundance at phylum and family taxonomy led to a heterogeneous data input to the machine learning algorithms, which exert different model training results capable of affecting model prediction accuracy.

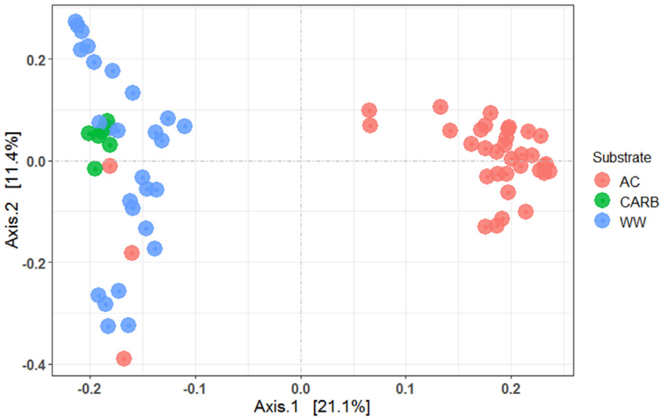

Fig. 2. Phylogenetic distances between samples determined via weighted UniFrac PCoA of the overall ASVs.

3.3. Prediction of substrate classification by machine learning

To identify and detect the correlation between microbial population abundance and feed substrate, six machine learning algorithms (GLMNET, RF, XGBOOST, NNET, KNN and SVM with radial kernel) were trained and evaluated in terms of their ability to precisely predict substrate composition from the different taxonomic ranks of genomic data. The initial model was trained and validated using 15 phyla. The model developed using the NNET algorithm had the highest accuracy (93 ± 6%), corresponding to *kappa* values (inter-rater agreement) of 0.87 ± 0.10, followed by the RF, GLMNET, and XGBOOST algorithms with accuracies at around 80% and *kappa* values of 0.76 ± 0.10, 0.75 ± 0.09, and 0.66 ± 0.12, respectively (Fig. 5a). The KNN algorithmic accuracy was ~65% identification and *kappa* value of 0.28 ± 0.12, which was the lowest prediction accuracy and is possibly due to its inability to handle imbalanced datasets (Zhang and Zhou, 2007).

The success of data mining techniques is dependent on the quality of the dataset. To verify the interaction between divergent datasets from

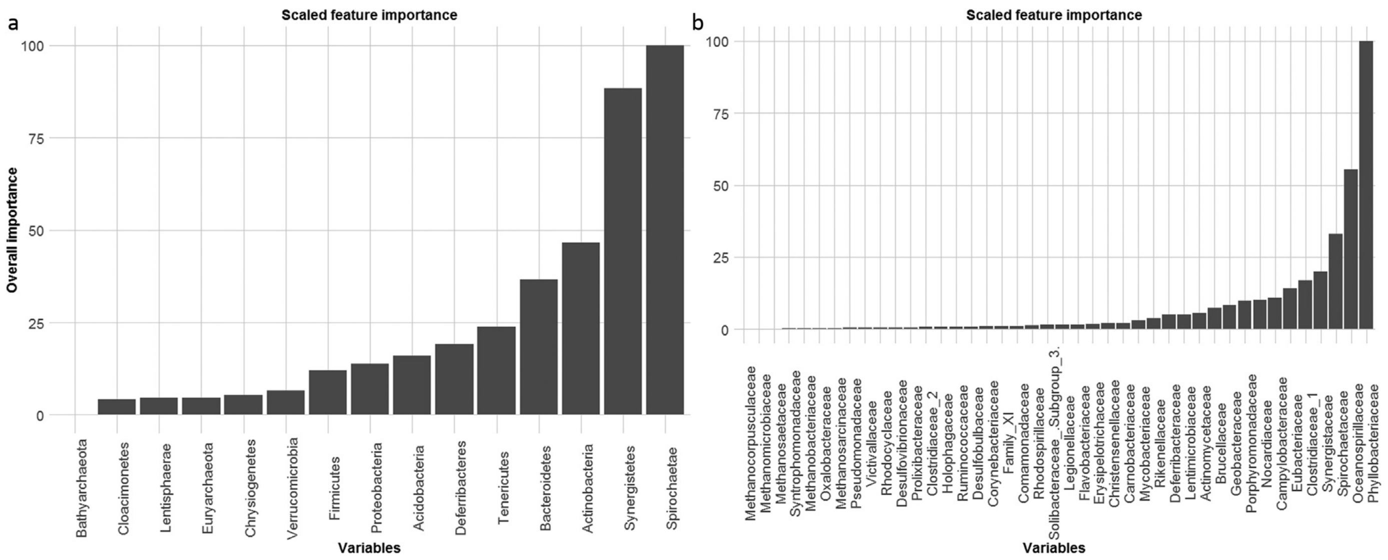

Fig. 3. The feature importance of (a) phyla and (b) family selected by random forest with 10-fold cross-validation.

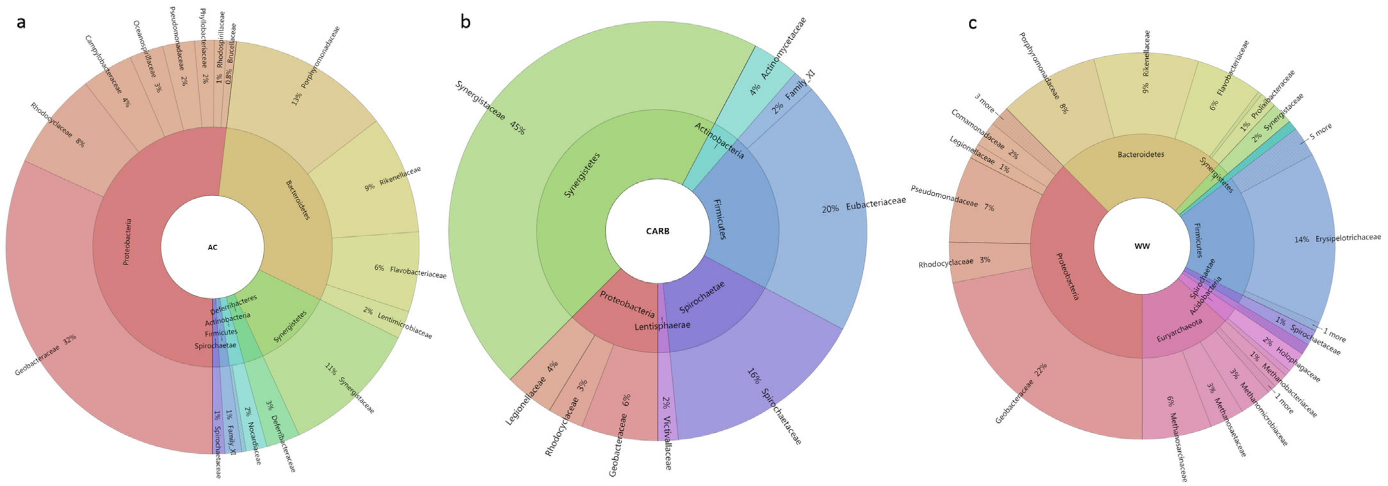

Fig. 4. The metagenome of (a) AC, (b) CARB and (c) WW feeds displayed using Krona. Taxonomy nodes are shown as nested sectors arranged from the phyla and family level.

varying sequencing methods (Roche 454 and Illumina sequencing) and prediction performance, the sample sequencing method was employed as an additional input variable. The accuracy values of all algorithms remained higher than 80% and the *kappa* values remained above 0.7 except for the SVM algorithm, indicating that, surprisingly, both sample sequencing methods provide similar taxonomic results and consistent prediction (Fig. S5). It has been reported that the biases in genomic data when using different sequencing method can be negligible and beta diversity can remain robust, although varying amplification primers, sequencing primers, sequencing methods, as well as differences in quality filtering and clustering can affect data quality and quantitative abundance (Nelson et al., 2014; Tremblay et al., 2015). Therefore, we conclude that the difference in sequencing methods was irrelevant for prediction accuracy and can be neglected in future models.

The effect of taxonomic level classification on prediction accuracy was also tested. 49 family abundances were selected to include as a separate family level training dataset. For all algorithms, except SVM and KNN, accuracy values were higher than 80%, in keeping with *kappa* values higher than 0.7 (Fig. 5b). The GLMNET algorithm incorporating family data produced an insignificant increase of ~4% prediction accuracy and ~0.07 *kappa* compared with phyla-dependent models. The *kappa* values of the SVM algorithm was 0.40 ± 0.11, indicating a good prediction accuracy. Only the *kappa* values for the KNN algorithm were poor (0.04 ± 0.07), due to the challenges discussed previously. This slight increase of accuracy and *kappa* values indicated a sustained predictability for larger but more narrowly classified datasets.

The full genomic dataset was compressed using NMDS and PCoA as model inputs to evaluate model predictability. Heterogeneous microbial community abundances with different feed substrates were successfully characterized using PCoA and NMDS with Unifrac dissimilarity distance metrics and provides a reliable method for classification prediction (Fig. 2 and Fig. S4). The most accurate prediction model was the XGBOOST algorithm for both PCoA and NMDS compressed datasets with an accuracy of 88 ± 4% corresponding to a *kappa* value of 0.78 ± 0.07, which were similar to that for the phyla and family datasets (Fig. 6). This is likely due to the scalability of the XGBOOST model and ability to improve predictability through stacking numerous tree models (Chen and Guestrin, 2016). Accuracies of the NNET and SVM algorithms were reduced to around 50% with the PCoA compressed data, corresponding to *kappa* values of zero. The accuracy and *kappa* of GLMNET and RF were slightly lower than both phyla and family ranked models from the PCoA compressed data (Fig. 6a). Conversely, substrate classification of different algorithms from NMDS compressed dataset inputs were significantly more accurate than that

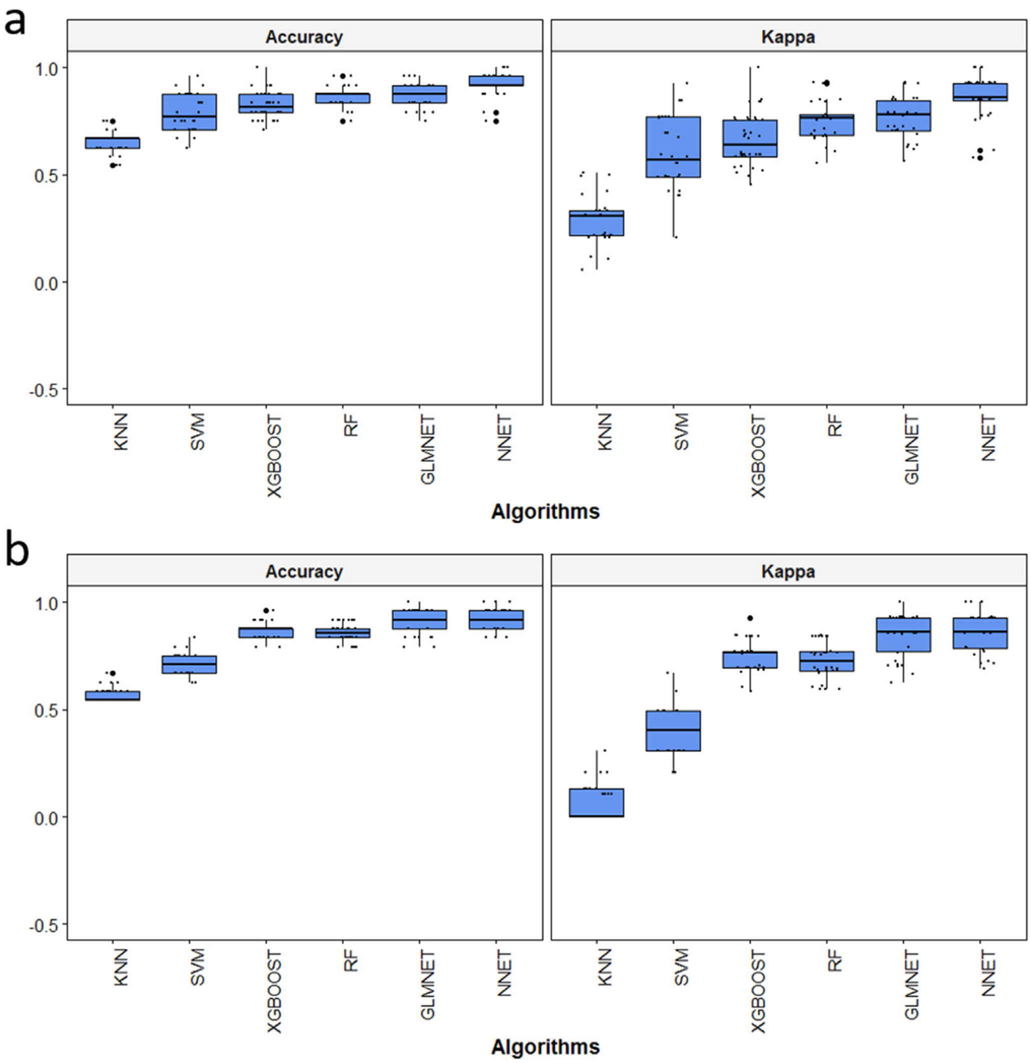

Fig. 5. Accuracy and *kappa* metrics of different algorithms incorporated with (a) phyla datasets and (b) family datasets.

for the PCoA compressed dataset. Models using NMDS compressed inputs (apart from those using KNN algorithms) had similar prediction accuracy of around 80% and *kappa* values close to 0.5 (Fig. 6b).

The trained and validated algorithms successfully discriminated between the three feed substrates, as expected. The higher accuracy of substrate classification models based on the family ranked datasets suggested that the best prediction variables for identifying feed substrate can be obtained by having more specific microbial communities. Furthermore, the NMDS compressed ASV dataset yielded a substantially higher accuracy of substrate classification compared to the PCoA compressed ASV datasets indicating that NMDS is a better predictor of large ASV datasets for prediction interaction between microbial communities and feed substrate.

4. Conclusions

This study provides a new implementation of machine learning algorithms that has potential for substantial practical applications in biotechnology by demonstrating that incorporation of genomic data can be used for predicting the emergence of substrate-specific microbial communities in biosensors/bioreactors. This study is the first to link qualitative identification of chemicals presented in water with genomic data, and the first to evaluate the predictability of six different machine-learning algorithms comprehensively using four different input features. Heterogeneous microbial community diversity with different feed substrate led to significantly higher accuracy (around 80%) and kappa values (around 0.6) for algorithms using NMDS compressed, and family ranked genomic datasets. Good prediction of feed substrate identification was also achieved for the phyla, family and NMDS compressed ASV datasets. Surprisingly, incorporation of a sequencing method parameter did not affect the prediction accuracies. The success of this approach is due to the interaction between microbial community composition and feed substrates. This provides an opportunity for improving the specificity of chemical detection with biosensors compared to those using both electrical signals and fundamental chemical properties as inputs, where accuracy was limited to 26% (Feng et al., 2013). The predictability of substrate classification could be improved by suitable algorithm selection and optimization of model input variables. Further development of this approach not only could provide a means to accurately classify a broad spectrum of unknown chemicals for water quality monitoring, but also for toxin detection connected to the diverse and dynamic microbial communities considering the discrepancies in microbial resistance to different types of toxic pollutants, and divergent toxic shocks, which may also lead to community structure shifts. However, to ensure its use for practical biosensing applications, significantly more samples and input features need to be considered in model training and performance evaluation. Conversely, using this approach we can also better identify the significant members and proportions of microbial community populations, even in a known feed substrate system, due to the interaction between microbial community

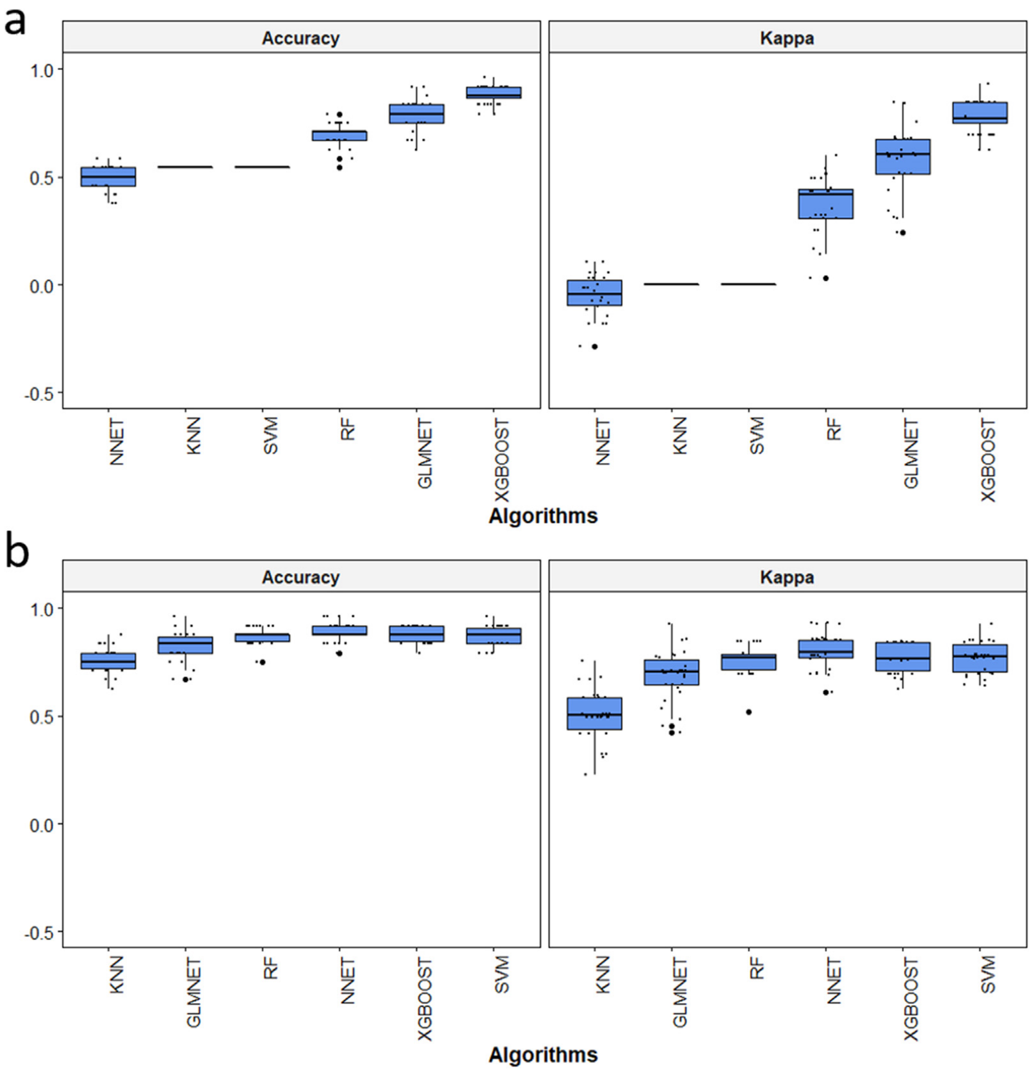

Fig. 6. Accuracy and *kappa* metrics of different algorithms incorporated with (a) PCoA compressed overall ASVs and (b) different distance metrics of NMDS compressed overall ASVs.

and organic matter. The metabolism of complex chemical compounds, such as polysaccharides, proteins and lipids, is usually a multi-step process (Morris et al., 2013). Relative to our current knowledge of biochemical metabolic pathways used by both aerobic and anaerobic microorganisms for some carbonaceous matters (McInerney et al., 2009), more knowledge is required regarding the complete metabolic pathways of specific food chains. This approach can be further extended to linking the flux of organic matter with the greenhouse gas emission process by exploring the interaction between microbial populations and organic matters. This information can be used to construct realistic food webs and metabolic pathways that can then predict the consequences of global climate change. Overall, the method detailed in this study can be used for understanding the interactions between substrate and microbial community population for economic, environmental and societal benefits.

Acknowledgments

This publication was made possible through funding support provided by the U.S. National Science Foundation (CBET 0955124), the Natural Science Foundation of China (No. 21828802, 21878242), and Newcastle University and University of Glasgow (EPSRC SUPERGEN grant: EP/H019480/1). W.F. Cai acknowledges the support from the China Scholarship Council (201706280278). M.J. Wade acknowledges the support from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 702408 (DRAMATIC).

Declaration of interests

None.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.bios.2019.03.021.

References

Beals, E.W., 1984. Bray-Curtis ordination: an effective strategy for analysis of multivariate ecological data. Adv. Ecol. Res. Academic Press, pp. 1–55.
Billard, A., Laval, V., Fillinger, S., Leroux, P., Lachaise, H., Beffa, R., Debieu, D., 2012. Appl. Environ. Microbiol. 78, 1063–1068.
Biswas, P., Karn, A.K., Balasubramanian, P., Kale, P.G., 2017. Biosens. Bioelectron. 94, 589–604.
Bongards, M., Gaida, D., Trauer, O., Wolf, C., 2014. Energy Sustain. Soc. 4, 19.
Cai, W.F., Geng, J.F., Pu, K.B., Ma, Q., Jing, D.W., Wang, Y.H., Chen, Q.Y., Liu, H., 2018. Chem. Eng. J. 333, 572–582.
Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J., Holmes, S.P., 2016. Nat. Methods 13, 581–583.

Catal, T., Li, K., Bermek, H., Liu, H., 2008. J. Power Sources 175, 196–200.
Chang, I.S., Jang, J.K., Gil, G.C., Kim, M., Kim, H.J., Cho, B.W., Kim, B.H., 2004. Biosens. Bioelectron. 19, 607–613.
Chang, J., Chapman, B., Friedberg, I., Hamelryck, T., De Hoon, M., Cock, P., Antao, T., Talevich, E., 2010. Biopython Tutorial and Cookbook. pp. 51–53.
Chen, T., Guestrin, C., 2016. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.
Chen, X., Ishwaran, H., 2012. Genomics 99, 323–329.
Chouler, J., Cruz-Izquierdo, A., Rengaraj, S., Scott, J.L., Di Lorenzo, M., 2018. Biosens. Bioelectron. 102, 49–56.
Di Lorenzo, M., Thomson, A.R., Schneider, K., Cameron, P.J., Ieropoulos, I., 2014. Biosens. Bioelectron. 62, 182–188.
Dixon, P., 2003. J. Veg. Sci. 14, 927–930.
Dubinsky, E.A., Butkus, S.R., Andersen, G.L., 2016. Water Res. 105, 56–64.
Feng, Y., Barr, W., Harper Jr., W.F., 2013. J. Environ. Manag. 120, 84–92.
Feng, Y., Harper Jr., W.F., 2013. J. Environ. Manag. 130, 369–374.
Feng, Y., Wang, X., Logan, B.E., Lee, H., 2008. Appl. Microbiol. Biotechnol. 78, 873–880.
Friedman, J., Hastie, T., Tibshirani, R., 2009. R package version 1.
Gieg, L.M., Fowler, S.J., Berdugo-Clavijo, C., 2014. Curr. Opin. Biotechnol. 27, 21–29.
Goldberger, J., Hinton, G.E., Roweis, S.T., Salakhutdinov, R.R., 2005. Adv. Neural Inf. Process. Syst. 513–520.
Han, H., Qiao, J., 2013. IEEE Trans. Control Syst. Technol. 21, 2423–2431.
Heidrich, E.S., Dolfing, J., Wade, M.J., Sloan, W.T., Quince, C., Curtis, T.P., 2018. Bioelectrochemistry 119, 43–50.
Hill, T.C.J., Walsh, K.A., Harris, J.A., Moffett, B.F., 2003. FEMS Microbiol. Ecol. 43, 1–11.
Ishii, S., Suzuki, S., Norden-Krichmar, T.M., Phan, T., Wanger, G., Nealson, K.H., Sekiguchi, Y., Gorby, Y.A., Bretschger, O., 2014. ISME J. 8, 963–978.
Jiang, Y., Liang, P., Liu, P., Wang, D., Miao, B., Huang, X., 2017a. Biosens. Bioelectron. 94, 344–350.
Jiang, Y., Liang, P., Liu, P., Yan, X., Bian, Y., Huang, X., 2017b. Int. J. Hydrog. Energy 42, 4342–4348.
Jiang, Y., Yang, X., Liang, P., Liu, P., Huang, X., 2018. Renew. Sustain. Energy Rev. 81, 292–305.
Kiely, P.D., Regan, J.M., Logan, B.E., 2011. Curr. Opin. Biotechnol. 22, 378–385.
Kim, J.R., Jung, S.H., Regan, J.M., Logan, B.E., 2007. Bioresour. Technol. 98, 2568–2577.
King, S.T., Sylvander, M., Kheperu, M., Racz, L., Harper Jr., W.F., 2014. Sci. Total Environ. 497–498, 527–533.
Korotcov, A., Tkachenko, V., Russo, D.P., Ekins, S., 2017. Mol. Pharm. 14, 4462–4475.
Krogh, A., 2008. Nat. Biotechnol. 26, 195–197.
Kuhn, M., 2008. J. Stat. Softw. 28, 1–26.
Lande, R., 1996. Oikos 5–13.
Larsen, P., Dai, Y., Collart, F.R., 2015. Methods Mol. Biol. 1260, 33–43.
Lesnik, K.L., Liu, H., 2017. Environ. Sci. Technol. 51, 10881–10892.
Li, C., Wang, L.G., Liu, H., 2018. Appl. Microbiol. Biotechnol. 102, 7611–7621.
Libbrecht, M.W., Noble, W.S., 2015. Nat. Rev. Genet. 16, 321–332.
Liu, B., Lei, Y., Li, B., 2014. Biosens. Bioelectron. 62, 308–314.
Liu, C.M., Aziz, M., Kachur, S., Hsueh, P.R., Huang, Y.T., Keim, P., Price, L.B., 2012. BMC Microbiol. 12, 56.
Liu, H., Cheng, S.A., Logan, B.E., 2005. Environ. Sci. Technol. 39, 658–662.
Liu, H., Logan, B.E., 2004. Environ. Sci. Technol. 38, 4040–4046.
Lozupone, C., Knight, R., 2005. Appl. Environ. Microbiol. 71, 8228–8235.
McInerney, M.J., Sieber, J.R., Gunsalus, R.P., 2009. Curr. Opin. Biotechnol. 20, 623–632.
Mjolsness, E., DeCoste, D., 2001. Science 293, 2051–2055.
Morris, B.E., Henneberger, R., Huber, H., Moissl-Eichinger, C., 2013. FEMS Microbiol. Rev. 37, 384–406.
Nelson, M.C., Morrison, H.G., Benjamino, J., Grim, S.L., Graf, J., 2014. PLoS One 9, e94249.
Ondov, B.D., Bergman, N.H., Phillippy, A.M., 2011. BMC Bioinform. 12, 385.
Ou, S., Zhao, Y., Aaron, D.S., Regan, J.M., Mench, M.M., 2016. J. Power Sources 328, 385–396.
Pal, M., 2007. Int. J. Remote Sens. 26, 217–222.
Quek, S.B., Cheng, L., Cord-Ruwisch, R., 2015. Water Res. 77, 64–71.
Ramette, A., 2007. FEMS Microbiol. Ecol. 62, 142–160.
Ripley, B., Venables, W., Ripley, M.B., 2016. R package version, pp. 7–3.
Scholkopf, B., Sung, K.K., Burges, C.J.C., Girosi, F., Niyogi, P., Poggio, T., Vapnik, V., 1997. ITSP 45, 2758–2765.
Sharma, Y., Li, B., 2010. Bioresour. Technol. 101, 1844–1850.
Shi, S., Xu, G., 2018. Chem. Eng. J. 347, 280–290.
Stein, N.E., Keesman, K.J., Hamelers, H.V., van Straten, G., 2011. Biosens. Bioelectron. 26, 3115–3120.
Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T., Zeileis, A., 2008. BMC Bioinform. 9, 307.
Tremblay, J., Singh, K., Fern, A., Kirton, E.S., He, S., Woyke, T., Lee, J., Chen, F., Dangl, J.L., Tringe, S.G., 2015. Front. Microbiol. 6, 771.
Vilajeliu-Pons, A., Baneras, L., Puig, S., Molognoni, D., Vila-Rovira, A., Hernandez-Del Amo, E., Balaguer, M.D., Colprim, J., 2016. PLoS One 11, e0164044.
Wang, D., Liang, P., Jiang, Y., Liu, P., Miao, B., Hao, W., Huang, X., 2018. Biosens. Bioelectron. 111, 97–101.
Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R., 2007. Appl. Environ. Microbiol. 73, 5261–5267.
Wang, X., Gao, N., Zhou, Q., 2013. Biosens. Bioelectron. 43, 264–267.
Wang, X., Williams, C., Liu, Z.H., Croghan, J., 2019. Brief. Bioinform. 20, 156–167.
Wu, Y.C., Shiledar, A., Li, Y.C., Wong, J., Feng, S., Chen, X., Chen, C., Jin, K., Janamian, S., Yang, Z., Ballard, Z.S., Gorocs, Z., Feizi, A., Ozcan, A., 2017. Light Sci. Appl. 6, e17046.
Yang, Y., Yu, Y.Y., Wang, Y.Z., Zhang, C.L., Wang, J.X., Fang, Z., Lv, H., Zhong, J.J., Yong, Y.C., 2017. Biosens. Bioelectron. 98, 338–344.
Zhang, M.L., Zhou, Z.H., 2007. Pattern Recognit. 40, 2038–2048.