UNIVERSITY OF CALIFORNIA

SANTA CRUZ

Rapid Classification of NifH Protein Sequences using Classification and Regression Trees

A thesis submitted in partial satisfaction

of the requirements for the degree of

MASTER OF SCIENCE

in

OCEAN SCIENCES

by

Ildiko Frank

June 2014

The Thesis of Ildiko Frank
is approved:

____

Professor Jonathan Zehr, Chair

____

Professor Chad Saltikov

____

Professor Marilou Sison-Mangus

____

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

Abstract

Rapid Classification of NifH Protein Sequences using

Classification and Regression Trees

by

Ildiko Frank

Grouping and classifying *nifH* gene sequences, molecular proxies for studying nitrogen fixation, are essential steps in diazotroph community analysis, and the increasing size of environmental sequence libraries necessitates a fast and automated solution. We present a novel approach to classify NifH protein sequences into well-defined phylogenetic clusters that provide a common platform for cross-ecosystem comparative analysis. Cluster membership can be accurately predicted with Classification and Regression Trees (CART) statistical models that identify and utilize signature residues in the protein sequences. The decision tree-based classification models were trained and evaluated with the publicly available cluster-annotated *nifH* gene database and further assessed with model-independent sequence sets from diverse ecosystems. Network graph-based exploration of cluster structures led to models for sequence classification even at finer taxonomic levels. We demonstrate the utility of this novel sequence binning approach in a comparative study where joint treatment of diazotroph assemblages from a wide range of habitats identified specialists and generalists and revealed a marine – terrestrial distinction in

the community composition. Our rapid and automated cluster assignment circumvents extensive analysis of the *nifH* database and calculating phylogenies; hence, saves time and resources in studying nitrogen fixation.

Acknowledgement

I would like to thank my advisor, Jonathan Zehr, for all the support and encouragement I received starting from the day I walked into his office as an undergraduate student.

My fellow lab members have been a constant source of valuable information and feedback. I am especially grateful to Kendra Turk-Kubo, Irina Shilova, and Jim Tripp.

Chapter I: Introduction

Microorganisms drive Earth's biogeochemical cycles and shape the ecology of marine as well as terrestrial provinces (Falkowski et al., 2008). Microbial communities are characterized by a few dominant species and a long tail of low-abundance taxa (Sogin et al., 2006). The majority of microbial diversity is found in a mere fraction of soil (Elshahed et al., 2008) and marine (Pedros-Alio, 2006) biomass. The number of microbial species in the ocean is estimated to be a few million; it could be a magnitude larger in soil (Curtis et al., 2002). Some studies suggest even greater diversity, but its significance in microbial community function and evolution remains an open question (Zehr, 2010).

Since less than 1% of microorganisms can be grown in the laboratory, and half of the prokaryotic phyla lack cultured representatives (Rappe and Giovannoni, 2003), molecular techniques are essential proxies for studying environmental microbial assemblages (Zehr et al., 2009). For most ecosystems, cultivation-independent molecular surveys are indispensable, because biogeochemical and ecological interactions cannot be investigated in a laboratory setting (DeLong, 2009). Instead of directly observing phenotypes, molecular proxies are used to decipher "who is there" and "what they are capable of doing". In the last few decades, molecular tools have greatly expanded our knowledge of phylogenetic as well as functional diversity of microbial communities; yet, microbial ecology is still far behind plant and animal ecology in understanding spatial and temporal patterns and large-scale variability (Fierer and Ladau, 2012).

Marker gene amplification is one of the molecular techniques first used in microbial ecology. In contrast to metagenomics, which is the genomic analysis of microbial assemblages (Handelsman, 2004), the marker gene approach targets a single conserved gene, usually with DNA amplification and sequencing, to quantify microbial diversity or assess specific biogeochemical functions. Pace and Woese pioneered the universally conserved small subunit ribosomal RNA, 16S rRNA, as a primary phylogenetic marker for describing prokaryotic phylogeny (Fox et al., 1980; Lane et al., 1985). Later, protein-coding markers emerged and provided a finer resolution and better distinction of closely related strains in some taxa than the 16S rRNA gene, which often falls short in informative characters (Santos and Ochman, 2004). "Functional" gene targets, that represent biogeochemically-important processes, include *nasA* in nitrate assimilation (Allen et al., 2001), *nirS* in denitrification (Ward et al., 2009), *dsr* in dissimilatory sulfate reduction (Wagner et al., 1998), *nosZ* in nitrous oxide reduction (Jones et al., 2013), and *nifH* in biological nitrogen fixation (Zehr and McReynolds, 1989).

Biological nitrogen fixation, the reduction of atmospheric N_2 gas to ammonium, is an important biogeochemical process that plays an indispensable role in the global nitrogen cycle (Ward et al., 2007). The amount of fixed nitrogen is estimated to be 128 Tg per year (Galloway et al., 2004). This new source of nitrogen bolsters the trophic web in nutrient-limited provinces including vast areas of the ocean (Vitousek and Howarth, 1991) and is linked to atmospheric carbon dioxide fixation and carbon export from surface waters (Falkowski, 1997). The rest of the N cycle is composed of unrelated reversible metabolic pathways and is catalyzed by different multispecies microbial communities that are often spatially and temporally separated (Falkowski et al., 2008). The oxidative pathway, called nitrification, involves two groups of prokaryotes; the first group converts ammonium to nitrite, which is further oxidized to nitrate by a second group (Zehr and Kudela, 2011). These low energy compounds are then utilized as electron acceptors in anaerobic respiration, called denitrification. The anammox process, a redox reaction that involves ammonium and nitrite, is a recently discovered third component in the cycle of opposing processes (Mulder et al., 1995).

Nitrogen-fixing microbes, called diazotrophs, comprise just a small subset of prokaryotes from diverse taxonomic groups with physiologies that range from aerobic to anaerobic and includes oxygen-evolving as well as non-oxygen-evolving photoautotrophy (Triplett, 2000). Nitrogen-fixers, autotrophs as well as heterotrophs, thrive in all marine environments from surface waters (Farnelid et al., 2011) to the deep sea (Mehta et al., 2003), as well as in terrestrial ecosystems in bulk soil (Hsu and Buckley, 2009), the rhizosphere (Deslippe and Egger, 2006), and phyllosphere (Furnkranz et al., 2008).

Despite the large variation in genome size and physiology, all diazotrophs rely on the same oxygen sensitive enzyme, called nitrogenase, for nitrogen reduction (Stacey et al., 1992). Nitrogenase is composed of two proteins: the heterotetramer

($\alpha_2\beta_2$) dinitrogenase performs the reduction, and the homodimer dinitrogenase reductase provides the electrons (Hamilton et al., 2011). Based on their metal cofactors, the former is called iron-molybdenum protein and the latter is referred to as iron protein, where the two subunits are covalently linked by a 4Fe:4S cluster. Although the operon(s) encoding and regulating the enzyme may contain as many as twenty genes, the main structural genes are *nifD* (α component), *nifK* (β component), and *nifH* (iron protein). In addition to these key coding regions, three more genes, *nifE*, *nifN*, and *nifB*, are assumed to be essential for a functional enzyme (Dos Santos et al., 2012). All diazotrophs contain one or more subtypes of the enzyme; the *nif* gene set that codes the molybdenum-dependent (Mo-Fe) enzyme is replaced by the *vnf* set in the vanadium-dependent (V-Fe) alternative enzyme or by the *anf* set in the iron-only (Fe-Fe) alternative enzyme. The first alternative nitrogenase was identified in *Azotobacter vinelandii* (Bishop and Joerger, 1990), and the number of diazotrophs identified with *anf* and *vnf* operons has been growing since (Betancourt et al., 2008).

Sequence and structure of the *nifH* encoded iron protein have been well studied. Crystallography provided the first 3D details of the protein isolated from *Azotobacter vinelandii* more than 20 years ago (Georgiadis et al., 1992). Several highly conserved regions sandwiched between variable ones were identified based on 59 full genome sequences (Schlessman et al., 1998). Conserved elements include the P loop and two switch regions that are essential in nucleotide binding. These positions can be used to screen sequences, but contain no phylogenetic information. The potential of the so called 60's loop in distinguishing sequences from different taxa was recognized early on (Schlessman et al., 1998), but the relationship between diverged positions and the established *nifH* clusters has yet to be examined. The extant large collection of *nifH* sequences does contain such relational information that could be explored with statistical models including feature selection and utilized for rapid sequence classification.

The highly conserved *nifH* gene is well-suited for phylogenetic and ecological analyses of nitrogen-fixing organisms. A phylogenetic tree published two decades ago groups the *nifH* sequences into four main clusters labeled by Roman numerals (Chien and Zinder, 1994). Despite some disagreement (Raymond et al., 2004; Gaby and Buckley, 2011), this division is widely accepted. The bulk of the sequences that form cluster I originate mainly from Cyanobacteria, Proteobacteria, Firmicutes, or Actinobacteria (Zehr et al., 2003). Most organisms in this cluster contain Fe-Mo nitrogenase. Cluster II, a considerably smaller group, contains sequences from Bacteria with alternative V-Fe or Fe-Fe enzymes and from some methanogenic Archaea. The distantly related sequences in Cluster III come predominantly from anaerobic Bacteria or Archaea. Paralogs of *nifH* that are not involved in nitrogen fixation group in cluster IV. With the steady accumulation of environmental sequences came the need for a finer level characterization. Subclusters were defined based on the phylogenetic information amassed in the large publicly available *nifH* database (Zehr et al., 2003) http://www.pmc.ucsc.edu/~wwwzehr/research/database/.

This sequence database is a valuable resource that has bolstered numerous studies of nitrogen fixation in a wide range of habitats including open ocean (Farnelid et al., 2011; Turk et al., 2011; Halm et al., 2012; Bonnet et al., 2013), coastal waters (Moisander et al., 2007; Bombar et al., 2011; Hamersley et al., 2011), lakes (Steward et al., 2004), caves (Desai et al., 2013), phyllosphere (Furnkranz et al., 2008), rhizosphere (Duc et al., 2009), and symbiont hosts (Yamada et al., 2007; Mohamed et al., 2008; Desai and Brune, 2012; Lema et al., 2012). Although not as extensive as the rRNA sequence collections in the Ribosomal Database Project (Cole et al., 2014), GreenGenes (DeSantis et al., 2006), or SILVA (Pruesse et al., 2007), the *nifH* database similarly poses great challenges to its curators and users in collecting, organizing and analyzing large amounts of molecular data. Clustering and classification of sequences are two key components of the analysis pipeline typically followed in ecological studies including those focused on diazotrophs.

Clustering, often the first processing step, is used to group sequences into so called operational taxonomic units (OTU) (Wooley et al., 2010). Number and taxonomic annotations of the emerging OTUs are a priori unknown. Sequences are merged based on pairwise similarity or distance measures. The resulting OTUs depend on the sequence region considered, the measure applied, and the linkage algorithm selected. The mothur software, which is popular in analyzing environmental sequence data, clusters sequences using nearest, furthest, or average neighbor linkage (Schloss et al., 2009). Non-distance-based clustering algorithms, like cd-hit (Li and Godzik, 2006) and UCLUST (Edgar, 2010) are the method of choice for large data sets where speed and size are critical issues. Similar to distance-based methods, these fast algorithms do not guarantee intra- and inter-OTU similarity properties.

In contrast to clustering, classification assigns sequences to categories defined in the reference database (Bazinet and Cummings, 2012). Three different methodologies are widely used to classify sequences into predefined taxonomic groups: sequence similarity search, sequence composition model, and phylogenetic method. BLAST, the almost exclusively used sequence similarity search, matches sequences against annotated databases and identifies closest relatives, often with known taxa labels (Altschul et al., 1990). This procedure is implemented in CAMERA (Seshadri et al., 2007), and MG-RAST (Meyer et al., 2008) environmental sequence analysis platforms. Naïve Bayesian Classifier is one of the most popular sequence composition models. It is based on oligonucleotide (8-mer) frequencies and is implemented in Ribosomal Database Project (Wang et al., 2007). The algorithm is fast, does not require sequence alignment, and works well with sequence fragments. The phylogenetic method works only with marker genes, for example, FastTree (Price et al., 2010); it attempts to "place" query sequences on a phylogenetic tree where branch clusters are defined and labeled. This approach is implemented in the ARB software environment (Ludwig et al., 2004), and used in 16S rRNA as well as in *nifH* sequence-based analyses. None of these solutions are simple and cannot be represented visually. The decision algorithms behind the classification are implemented as black-boxes without a user-friendly interface to quickly identify mislabeled sequences. Consequently, there is a need for a simple, graphically enhanced methodology to improve the ease of application and transparency of the results.

Classification And Regression Trees (CART) methodology might fit this need. CART, popular in data mining and machine learning, is a classification method specifically designed to model large complex data sets characterized by non-linear relationships among the variables (Breiman et al., 1983). Category prediction is based on a hierarchy of simple decision rules that are organized and displayed as a binary decision tree. It has been applied in a wide range of studies including ecological analyses (De'ath and Fabricius, 2000; Usio et al., 2006; Clarke et al., 2008; Pesch et al., 2011). Although, CART has not been utilized for sequence classification, it may be a successful competitor to the above discussed methodologies.

This thesis presents a novel NifH protein sequence classification based on CART models, tests its accuracy, and demonstrates the utility of the newly derived sequence characterization in a comparative study of ecosystems. The sequence labeling procedure is automated in a Python script and is available for general use at the website: http://pmc.ucsc.edu/~wwwzehr/research/. In-depth discussion of the classification models includes assessment of their accuracy using statistical estimation and network analysis. The latter is based on sequence groups resulting from a novel algorithm that ensures specified intra- and inter-group similarity. A cross-ecosystem analysis of published sequence sets originating from a wide range of marine and terrestrial habitats illustrates the power of uniformly applied cluster labels that create a common platform for comparative analysis.

Chapter II: Rapid Classification of NifH Protein Sequences using Classification and Regression Trees

Abstract

Grouping and classifying *nifH* gene sequences, molecular proxies for studying nitrogen fixation, are essential steps in diazotroph community analysis, and the increasing size of environmental sequence libraries necessitates a fast and automated solution. We present a novel approach to classify NifH protein sequences into well-defined phylogenetic clusters that provide a common platform for cross-ecosystem comparative analysis. Cluster membership can be accurately predicted with Classification and Regression Trees (CART) statistical models that identify and utilize signature residues in the protein sequences. The decision tree-based classification models were trained and evaluated with the publicly available cluster-annotated *nifH* gene database and further assessed with model-independent sequence sets from diverse ecosystems. Network graph-based exploration of cluster structures led to models for sequence classification even at finer taxonomic levels. We demonstrate the utility of this novel sequence binning approach in a comparative study where joint treatment of diazotroph assemblages from a wide range of habitats identified specialists and generalists and revealed a marine – terrestrial distinction in the community composition. Our rapid and automated cluster assignment circumvents

extensive analysis of the *nifH* database and calculating phylogenies; hence, saves time and resources in studying nitrogen fixation.

Introduction

Biological nitrogen fixation is a prokaryote-driven biogeochemical process that sustains the trophic web in nitrogen limited habitats including vast areas of the ocean (Vitousek and Howarth, 1991) and is linked to atmospheric carbon dioxide fixation and carbon export from surface waters (Falkowski, 1997). Since the microbial majority is recalcitrant to cultivation (Rappe and Giovannoni, 2003), and biogeochemical interactions cannot be investigated in a laboratory setting (DeLong, 2009), cultivation-independent molecular surveys are indispensable in assessing microbial diversity and metabolic complexity (Zehr et al., 2009). While the small subunit ribosomal RNA became the primary phylogenetic marker (Lane et al., 1985), *nifH* gene was proposed as molecular proxy in nitrogen fixation studies (Zehr and McReynolds, 1989) that led to the recognition of unexpected diazotroph diversity (Ueda et al., 1995; Zehr et al., 1995) and the discovery of widely distributed nitrogen-fixers with unusual physiology (Zehr, 2011). The publicly available *nifH* sequence database available at http://www.pmc.ucsc.edu/~wwwzehr/research/database/ (Zehr et al., 2003) is a valuable resource that has bolstered numerous investigations of nitrogen-fixing assemblages in marine (Moisander et al., 2007; Zehr et al., 2007; Fong et al., 2008; Moisander et al., 2008; Bombar et al., 2011; Farnelid et al., 2011; Hamersley et al., 2011; Turk et al., 2011; Halm et al., 2012; Bonnet et al., 2013;

Rahav et al., 2013) and terrestrial environments (Steward et al., 2004; Furnkranz et al., 2008; Duc et al., 2009; Desai et al., 2013) as well as in host symbionts (Yamada et al., 2007; Mohamed et al., 2008; Desai and Brune, 2012; Lema et al., 2012).

Diazotroph diversity assessment necessitates classifying *nifH* sequences into annotated taxonomic groups. Despite some disagreement (Raymond et al., 2004; Gaby and Buckley, 2011), a phylogenetic division of four main clusters (Chien and Zinder, 1994) is largely accepted. Cluster I is composed mainly of Cyanobacteria, alpha-, beta-, and gamma-Proteobacteria, Firmicutes, or Actinobacteria (Zehr et al., 2003). Sequences from prokaryotes with alternative nitrogenase enzymes (Betancourt et al., 2008) and from methanogenic Archaea (Chien et al., 2000) form cluster II. The distantly related sequences in cluster III come predominantly from anaerobic organisms, whereas the cluster IV sequences are *nifH* paralogs that are not involved in nitrogen fixation.

Exponential growth of the *nifH* database and increasing size of environmental *nifH* libraries necessitated finer-level sequence grouping in diazotroph community analyses. Such groups are derived either by merging sequences into more manageable but a priori unknown number of operational taxonomic units, OTUs, or by classifying sequences into subclusters, intra-cluster branches of the *nifH* phylogenetic tree (Zehr et al., 2003). OTUs can be calculated by distance-based hierarchical clustering, implemented for example in the mothur package (Schloss et al., 2009), or by fast clustering algorithms suited for large data sets, for example, cd-hit (Li and Godzik, 2006) or UCLUST (Edgar, 2010). The resulting groups are study specific, not comparable across ecosystems and, in contrast to subclusters, do not provide a common platform for abundance and diversity analyses. Subcluster assignment of newly acquired sequences currently is performed by a time-consuming and computationally demanding "placing on the tree" approach (Price et al., 2010). A rapid solution would greatly facilitate this essential step of ecological analyses.

In addition to the above phylogeny-based sequence characterization, implemented for example in the ARB software environment (Ludwig et al., 2004), sequence similarity and sequence composition are also utilized to classify sequences into established taxonomic groups (Bazinet and Cummings, 2012). BLAST (Altschul et al., 1990), available in metagenomic platforms CAMERA (Seshadri et al., 2007) and MG-RAST (Meyer et al., 2008), matches sequences against an annotated database. However, sequences without close relatives are likely to be misclassified in this sequence similarity-based approach. Naïve Bayesian Classifier is a fast sequence composition-based technic that calculates oligonucleotide (8-mer) frequencies. It is implemented and widely used for rRNA sequence classification in the Ribosomal Database Project (Wang et al., 2007), but was found to be inferior to BLAST in classifying sequences of *pmoA*, a functional marker gene of methanotrophs (Dumont et al., 2014).

Classification And Regression Trees (CART), a statistical methodology popular in data mining and machine learning, was specifically designed to handle large complex data sets (Breiman et al., 1983). The CART model consists of a hierarchy of simple decision rules, each based on a single predictor, which are organized and graphically presented as a binary decision tree. Among its many applications, it has been used in various ecological studies to model abundance data and correlate environmental and biological parameters (De'ath and Fabricius, 2000; Usio et al., 2006; Clarke et al., 2008; Pesch et al., 2011), but has not been tested for environmental amplicon classification.

This study presents and evaluates a novel cluster assignment of NifH protein sequences based on CART classification models and demonstrates the utility of a uniform sequence grouping and classification in a comparative ecosystem analysis. The rapid cluster assignment proved to be a successful replacement for the currently used time-consuming phylogeny-based procedure. CART models identified signature residues that contain sufficient information to distinguish among the established phylogenetic clusters as well as to screen for key nitrogen-fixing Cyanobacteria.

Materials and Methods

Training Set for CART Modeling

The publicly available *nifH* sequence database (Zehr et al., 2003) at http://www.pmc.ucsc.edu/~wwwzehr/research/database/ was utilized to train classification models and calculate sequence network graphs. It contains 22,497 sequences that are assigned to main clusters I (17,321), II (542), III (3,876), and IV (758). A subset of sequences (16,567) is further assigned to 43 subclusters of uneven sizes. The largest groups are 1B (3,304) and 1K (3,239), whereas the smallest subclusters 1, 2, 2D, 3, 3B, 3S, 4, and 4G contain only a dozen or less sequences.

Most of the sequences originate from environmental samples and only a small portion (663) was obtained from fully sequenced genomes. CART models were trained with protein sequences, where positions were labeled according to the *Azotobacter vinelandii* residues from A1 to A290 (Schlessman et al., 1998).

Test Set for CART Evaluation

A set of 1,558 unique *nifH* sequences derived from soil samples were imported into the ARB database and assigned to four main and seventeen subclusters (Collavino et al., 2014). As in the training set, most sequences (90%) belong to cluster I. This training set-independent data is composed of protein fragments covering positions between A45 and A153 and was used to evaluate the CART models' cluster assignment accuracy.

Environmental Data Sets for Ecosystem Comparison

Thirteen published *nifH* coded protein sequence sets, were downloaded from the GenBank database (http://www.ncbi.nlm.nih.gov) for further model accuracy assessment and cross-ecosystem analysis (Table 1). Sequence coverage varies between positions A7 and A206, but most sets include the A45 – A153 range. Six terrestrial sets originate from bulk soil (G, S), rhizosphere (AR), phyllosphere (R), geothermal mats (Y), and termite gut (T). Seven marine sets encompass a wide geographical and depth range representing coastal waters (C, M, MK), open ocean surfaces (A, P, SP), and extreme water depth (D). In all cases, sequences are assigned into the four main clusters, but the finer level grouping is study specific.

Data Analysis

Aligned and cluster-annotated protein sequences were exported in fasta format from a *nifH* gene sequence database stored in ARB (Ludwig et al., 2004). Sequence logos for visual exploration of the amino acid variation along the NifH protein sequence were created by WebLogo (Crooks et al., 2004). Statistical analysis was performed in R, an open source data analysis environment (R Development Core Team, 2013). Sequences were imported into R using package “seqinr” (Charif et al., 2012). Mantel test to compare ecosystem similarities calculated from cluster versus OTU counts was calculated in R package “vegan” (Oksanen et al., 2013). Correspondence analysis to visualize ecosystem similarity was performed in R package “ca” (Nenadic and Greenacre, 2007).

CART (Classification And Regression Trees) models (Breiman et al., 1983) were calculated to predict cluster assignments of NifH protein sequences. One model classifies into main clusters, and four separate models further characterize sequences by subclusters. Positions in the protein sequence were used as categorical predictor variables that may have twenty different amino acids as levels. Cluster assignment defined in the database was the categorical response to be predicted. Each decision node of the tree was defined in terms of a primary protein sequence position and a list of amino acids that determined how sequences traversed down the tree all the way to the terminal nodes corresponding to *nifH* clusters. When a primary position was missing from a sequence, cluster prediction was based on the corresponding surrogate protein position. Such “backup” positions, identified for each decision node in the model, are highly correlated with the primary positions and their use does not diminish the classification performance. Due to the uneven sizes of the categories, main clusters and subclusters, categories were weighted in inverse proportion to their size. Ten-fold cross-validation was applied to quantify the predictive power of each model. R packages “rpart” (Therneau et al., 2014) and “rpart.plot” (Milborrow, 2014) were used to calculate, evaluate, and display the CART models.

Similarity between sequence pairs was quantified as normalized Hamming distance, i.e. number of sequence positions with different amino acids divided by the sequence length. This measure ranges from 0 to 1, where 0 indicates identical sequences, and was calculated by the command “daisy” in R package “cluster” (Maechler et al., 2014). A novel algorithm was developed to group protein sequences at a finer than subcluster level. The resulting groups are referred to as OTUs (operational taxonomic units) with similarity level indicated; for example, OTU98 for groups merged at 98% similarity. OTUs were defined by the following algorithm:

```
Calculate distances between all sequence pairs: $D_{ij}$ for i and j = 1, nseq.
Define sequence connectivity $d_{ij}$ at specified similarity level set by dmax (e.g. 98% similarity corresponds to dmax = 0.02): if $D_{ij} < \text{dmax}$ then $d_{ij} = 1$ (sequence pair connected), else $d_{ij} = 0$.
Loop through the following steps until all sequences are assigned to an OTU:
- count number of connections for each sequence: $\Sigma_j d_{ij}$ for i = 1, nseq;
```

- select the sequence with the largest number of connections as representative of the next OTU;

- representative and all its connections form the next OTU; exclude them from further grouping.

Update sequence connectivity by removing inter-OTU connections.

The resulting sequence connectivity matrices were transformed into networks where vertices represent sequences and edges indicate intra-OTU sequence connections. Networks were created and plotted with package “network” (Butts et al., 2014).

Results and Discussion

Amino Acid Variation and Sequence Coverage

We hypothesized that a small set of variable NifH protein sequence positions may contain sufficient information for accurate cluster assignment based on statistical modeling. Graphical exploration with sequence logo (Figure S1) confirms previously identified conserved residues (Schlessman et al., 1998). Strings of single letters clearly set apart four formerly named regions: P loop (A9-A19), Switch I (A38-A48), Metal Cluster Coordination (A86-A102), and Switch II (A125-A142). These nearly constant positions are useful to screen environmental sequences, but can be ignored in sequence similarity calculation and in statistical modeling. Between the extended constant regions, however, the sequence contains variable positions, including the previously identified 60’s loop (Schlessman et al., 1998). Sequence logos calculated for each main cluster separately reveal positions where the amino acid content is cluster dependent and exhibit high inter-class coupled with low intra-class variability (Figure 1.)

Sequence coverage in the training set that varies among residues (Figure S2) may also affect which positions are included in a model. Dominance of environmental sequence fragments explains the observed high coverage between positions A45 and A153 defined by the commonly used primer sets (Gaby and Buckley, 2012). Number of sequences including positions before A39 (start of the nifH3 primer) and after position A153 (end of the PolR primer) is extremely low. There are notable dips in the number of sequences at positions A67 and A68, and especially at position A119. The first two anomalies are mainly due to deletions in cluster III sequences, whereas at position A119 the gap occurs mainly in cluster I.

Classification Model for Main Clusters

A CART classification model, developed on the training set, successfully assigns sequences into the well-established four main *nifH* clusters. Instead of considering phylogeny and sequence similarity, our model labels sequences based on amino acids at select residues. The streamlined CART tree contains only three decision nodes and four terminal nodes that correspond to the main clusters (Figure 2). All three primary positions (A109, A49, and A53) are within the range of high sequence coverage. The CART cluster assignment shows very good agreement (95% to 99%) with the main cluster labels defined in the database, and high classification accuracy was obtained also by ten-fold cross-validation. Classification of a model-independent sequence set confirmed the good results observed on the training set: overall accuracy of main cluster assignment was 98%.

The CART model trained on the smaller set of full genome derived sequences has a similar structure (decision tree not shown). The number of splits and resulting terminal nodes in the two models are identical, but the selected primary positions match only in the first decision node. The second split is based on position A37 and the third split on position A144. Since these residues are outside of the typical environmental sequence range, their surrogates, A49 and A53, are used in classifying environmental sequence fragments. These surrogates are the same positions as the primary positions in the model calculated from the training set. This match between the two decision trees suggests that environmental sequences do not contain additional information for main cluster assignment.

Two-class models, separating one main cluster from the other three, identify signature residues unique to each main cluster. Sequence logos, constructed for each main cluster separately, confirm divergence at the selected primary and surrogate positions. These signature residues, many located in the so called 60’s loop, form a fingerprint of the corresponding main cluster (Figure 1). In clusters I and III the red arrow points to the primary positions highly conserved within the cluster, i.e. dominated by a single amino acid. The signature phenylalanine residue at position A109 in cluster I is replaced by a similarly hydrophobic leucine and methionine in clusters II and III. The highly conserved basic lysine at position A53 observed in cluster I and II is uniformly replaced by the hydrophobic leucine in cluster III sequences. In contrast, the primary positions in cluster II and IV show high intra-cluster amino acid variation.

At the highest taxonomic level, the *nifH* phylogeny is strikingly different from the phylogeny based on 16S rRNA; main cluster division precedes the split between Bacteria and Archaea (Zehr et al., 2003). This phylogeny mismatch is confirmed by a classification model distinguishing seven categories (only Bacteria in cluster I) where decision nodes split sequences into the four main clusters before the Archaea – Bacteria division (Figure S3). Classification results perfectly match the domain and main cluster labels for all Archaea and for Bacteria sequences in clusters I and II. In cluster III and IV this match is 98% and 97%, respectively. A closer examination of the last two decision nodes revealed that all Bacterial sequences that were separated at these nodes from Archaea belong to organisms living in extreme environments; for example, *Desulforudis audaxviator* was isolated from groundwater miles below the surface and *Dethiobacter alkaliphilus* was extracted from Mongolian soda lake sediments. Such high similarity between sequences from different domains might indicate lateral transfer of the nitrogenase genes.

Classification Model for Subclusters

CART decision trees based on a handful of signature residues were also found effective in characterizing sequences by subclusters. Cluster I, mainly Proteobacteria and Cyanobacteria, includes about 80% of the training set sequences. This group is split into twelve subclusters that exhibit approximate correspondence with the 16S rRNA phylogeny (Zehr et al., 2003). The classification tree ends in twelve terminal nodes, one for each subcluster (Figure S4). The decision nodes involve only eight residues, most in the so called 60’s loop located at the interface between the two nitrogenase components in the 3D protein structure. Classification accuracy is above 96% in most subclusters. Most test set sequences (1,423 / 1,558) group with cluster I, where the overall agreement between CART’s prediction and the published subcluster assignments is 91%.

Subcluster 1B, composed exclusively of Cyanobacteria sequences, holds special interest in ecological studies; hence merits a rapid screening. A two-class CART model can distinguish Cyanobacteria from other cluster I sequences with 97% accuracy based on a single decision node (decision tree not shown). The primary position, A103, is located in the same alpha helix as the signature residue A109 of cluster I. These two residues contain sufficient information for a simple screen: if A109=F (phenylalanine) and A103=I (isoleucine), then with high probability the sequence belongs to a Cyanobacteria. This algorithm resulted in 7% false negative (3,087/3,304 Cyanobacteria identified) and 1% false positive (138 / 13,263) in the training set.

CART subcluster labeling has high accuracy also in Cluster II, where sequences, mainly from organisms with alternative nitrogenase, are split into five subclusters**.** Classification, based on a tree with four decision nodes (A54, A67, A115, and A117) and five terminal nodes, matches 99% of the subcluster labels in the database (Figure S5). Except for the two poorly represented small subclusters (2 and 2D), the predictive power of the model is equally impressive.

Classification accuracy is disappointing within cluster III, which is composed of sequences mostly from anaerobic organisms from diverse Archaea and Bacteria taxa. Currently, the database defines eighteen subclusters, which is possibly an overfit of sequence variation. The overall match is 76%, and it drops considerably lower in some groups (Figure S6). Most primary positions in the model are between A76 and A87, a region distinct from the 60's loop featured in cluster I and II. This sequence range straddles two beta sheets towards the edge of the 3D structure. We recommend revisiting the phylogeny-based subcluster definition in this group.

Cluster IV protein sequences are the most divergent since they belong to non-nitrogen-fixing Archaea and Bacteria and consequently are under less evolutionary pressure. Despite the high amino acid variability at most positions (Figure 1), CART classification matches the database labels with 97% overall accuracy (Figure S7). The primary positions in the decision nodes are located considerably further from the N terminus than those selected in models for the other main clusters.

Classification Model for Cyanobacteria

Classification models can also be developed to identify sequences at genus, species, strain, or ecotype level when an annotated training set is available. We successfully modeled UCYN-A (*Candidatus* Atelocyanobacterium thalassa) and *Trichodesmium* spp., two Cyanobacteria that hold special importance in the marine environment (Zehr, 2011). Since sequences are labeled only at main and subcluster levels in the database, OTU groups within subcluster 1B were examined in order to identify sequence groups from single taxa and annotate training sets for classification. For statistical purposes, a single taxon OTU is useful only if its size reaches at least 8-10% of the 2,069 unique Cyanobacteria sequences. Distance metric based on the A45 - A153 range and our novel algorithm was used to group Cyanobacteria sequences. This algorithm assures that within an OTU, similarity between a member and a representative sequence is equal to or higher than the specified level. Furthermore, each sequence is connected to one and only one OTU representative, and similarity between representatives from different OTUs is always less than the specified level. At 98% similarity, 495 OTU98s were generated, whereas at 95% similarity, sequences grouped into 179 OTU95s. Protein BLAST search against the reference protein database was used to identify the representative sequences.

At 98% similarity, the largest group contained 151 sequences, and at 95% similarity, the size of this group increased to 176 sequences. In both cases, the representative sequences were identified as UCYN-A at 100% identity (second match only at 91%). A two-class CART model based on residues A78 and A85 correctly classified all 176 sequences and identified additional two sequences as UCYN-A. According to pBLAST, the closest relative of these sequences was also UCYN-A, but only at 93% identity.

The second largest OTU98 contained 140 sequences and the representative sequence was a perfect match with *Trichodesmium*. Grouping sequences at 95%, this group contained 222 sequences. A two-class CART model identified a sequence as *Trichodesmium* based on residues A61 and A83. In addition to the 222 sequences, eight other sequences were labeled as *Trichodesmium*. Protein BLAST search matched them at 92-94% identity with *Trichodesmium* (second match only at 81-82%).

These results indicate that sequences of *Trichodesmium* and UCYN-A form two groups that stand apart from other Cyanobacteria. In both cases, the intra-group distances vary between 1 – 7% in the A45 - A153 range. Identification of other Cyanobacteria taxa (e.g. *Crocosphaera*) for model training has failed. Representative of the third largest OTU98 had no match in the reference protein database above 90% identity, and the next largest three groups contained sequences from three or more genera. Consequently, other Cyanobacteria taxa could not be uniquely identified based on the A45 – A153 region, which is the typical fragment of environmental NifH protein sequences.

Model Evaluation with Network Graphs

Good performance of the CART model-based cluster assignment was further confirmed with thirteen *nifH* sequence sets selected from the literature (Table 1). All studies split the sequences into main clusters, presented on one or more phylogenetic trees. Only three sets (D, M, and R) encompass all four main clusters (Table 2), cluster I is the only group in the Pacific and Soil sets, but missing from the Termite set. CART cluster prediction matches the published groups with less than 1% error rate.

Unfortunately, group label comparison is not possible beyond the main clusters because each paper relies on a unique grouping method, group labels are not deposited as metadata, and NCBI accession numbers are rarely reported on the trees. Matching the CART derived subcluster labels with groups defined in these studies is challenging and often can be done only qualitatively by comparing group proportions. Our *Trichodesmium* and UCYN-A recognition algorithm proved reliable in five data sets (A, M, MK, P, SP) where these Cyanobacteria were reported.

Since direct comparison between the published sequence groups and the CART derived subclusters was only qualitative, our subcluster prediction was evaluated based on sequence networks that provide visual and numerical identification of potentially mislabeled sequences as well as graphical exploration of intra-cluster sequence variations. Networks from sequence sets with high variability (e.g. Termite), contain many singletons and disconnected small OTUs (Figure 3, top panel), whereas networks from low variability sequence sets (e.g. Atlantic), are dominated by few large OTUs that comprise the majority of sequences (Figure 3, bottom panel). OTUs and sequence connections for network analysis were defined using our novel sequence grouping algorithm. Networks of OTU98 and OTU90 (Table 2) were explored because 98% represents species level similarity, whereas 90% similarity is close to the average subcluster variation. Indeed, in most sequence sets, the number of OTU90s were close to the number of subclusters (Table 2); the strong positive correlation between these numbers is highly significant (Spearman correlation = 0.87, p = 0.0001). Proportions of mislabeled sequences in each data set were calculated based on OTU98 and OTU90 label purity (Table 3). An OTU is called pure if all its sequences are assigned to the same subcluster and a sequence is flagged mislabeled if its CART derived cluster assignment does not match the majority label within the OTU.

With OTU98, network-based classification error rate is very low (2 % or less), and with OTU90, it goes above 10% only in three data sets (C, M, MK). Two OTU90s in the Chesapeake set contain mixed cluster III labels. As noted previously, grouping within cluster III is questionable and ill-modeled. The high error rate in the Mediterranean set at 90% similarity is mainly due to merging 1O and 1P labeled sequences. These are neighboring subclusters on the phylogenetic tree; therefore, it is reasonable that they form a single OTU. The CART model separates 1O and 1P sequences only at the bottom of the decision tree, which is another indication of the potential overlap of these two groups. Fusion of 1O and 1P sequences at 90% similarity is also observed in the Atlantic and Mekong sets. Another overlapping subcluster pair is 1J and 1K. Again, 1K and 1J are next to each other on the phylogenetic tree and separated only at the bottom of the CART tree. We recommend to annotate these subcluster pairs with a single label.

Cross-ecosystem Analysis

Analyzing sequences sets of different origins together illustrated the power of uniformly applied cluster labels, and joint sequence binning helped to identify generalist and specialist diazotrophs. Cluster structure and OTU identities across ecosystems were compared by merging thirteen data sets and binning the sequences together. Two combined sets were analyzed: the complete set (2,433) contained all sequences and reflected OTU abundances, whereas the unique set (1,233) included only sequences unique within each set and was used to explore intra-group sequence variation. In both cases, the A45 – A153 primer-defined region was considered in calculating sequence similarity and OTU binning. These merged sets represented all four main clusters and 38 of the 43 subclusters that are defined in the database. As in the training set, 80% of sequences belong to cluster I, and the dominant subclusters are 1K+1J, 1B, 1G, and 1A.

At 98% similarity, the unique set sequences grouped into 408 OTU98s from which 276 are singletons, whereas the complete set resulted in 406 OTU98s with 247 singletons. Majority cluster label and origin of sequences were recorded for the largest groups in the unique set, and their representative sequences were identified by protein BLAST search (Table 4). Only few representatives were unidentifiable and many matched at 98% - 100% identity with known diazotrophs. Although the majority of non-singleton OTUs contained sequences from a single origin, 29 OTU98s were composed of sequences from multiple habitats (Figure 4).

The largest group was of marine origin mainly composed of Atlantic transcripts, was labeled 1G, and its representative matched closest to *Pseudomonas stutzeri*. The second and third largest groups contained sequences of marine as well as terrestrial origins, both were labeled 1K, and their representative sequences matched with *Burkholderia xenovorans* and *Azospirillum amazonense*, respectively. These sequences may be either contaminants or extracted from truly generalist diazotrophs. The same three groups led also the list obtained from the complete set. Several other OTU98s merged terrestrial and marine sequences and were identified as *Nodosilinea nodulosa, Bradyrhizobium elkanii, Nostoc punctiforme, Dechlorosoma suillum*, and *Fischerella muscicola*.

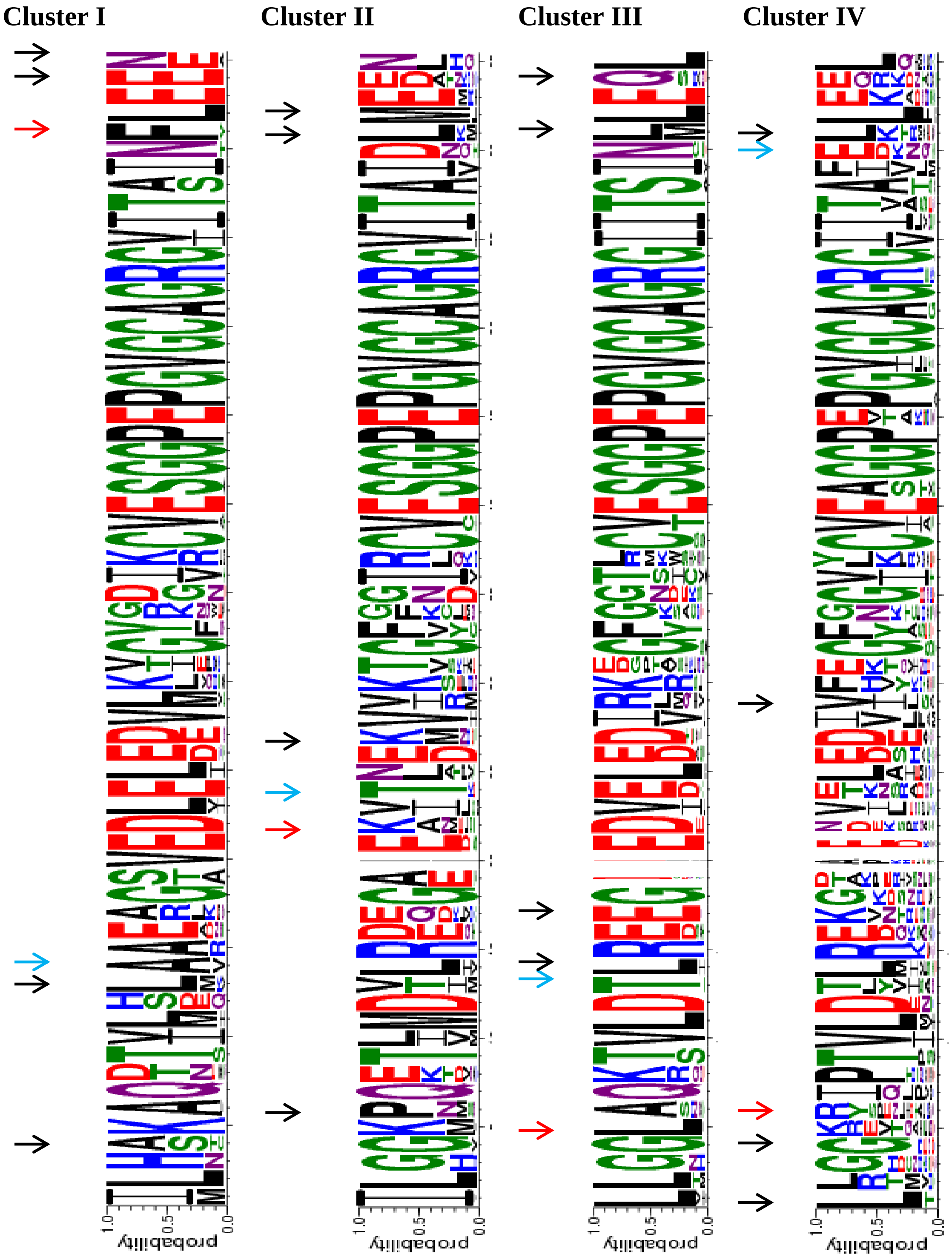
Sequences in OTU98 #5 originated from three marine locations and the group representative was identified as UCYN-A. *Trichodesmium* was identified as representative of OTU98 #6, another group of purely marine origin. Both OTUs were among the eight largest bins in the complete set indicating great abundance of these marine Cyanobacteria.

Subclusters approximately represent sequence groups at 90% similarity. Ecosystem similarities calculated from subcluster proportions and from OTU90 proportions were significantly correlated (Mantel test cor = 0.80, p = 0.001). Similar diazotroph communities have similar cluster composition and group on a correspondence analysis projection (Figure 5, top panel). Not surprisingly, communities from the deep sea (D) and termite gut (T) stand apart because they harbor the most unique diazotroph compositions. Except for the rainforest (R) community, sequence sets of terrestrial origin (Y, G, AR, S) tightly group at the lower right corner, whereas marine assemblages (A, P, SP, MK, M, C) that stretch along a diagonal vary more in composition. Projection of communities characterized by ten largest OTU90 proportions show a similar divide between marine and terrestrial communities (Figure 5, bottom panel).

Conclusion

With statistical modeling, we supported a hypothesis that the *nifH* coded protein sequence includes signature residues with sufficient information for phylogenetic cluster membership prediction. A potential of distinguishing sequences from different taxa based on certain sequence regions was suggested early on (Zehr et al., 1997; Schlessman et al., 1998), but the hypothesized relationship was not quantified with mathematical models. Similar classification models could be developed for other functional genes, making use of available annotated training sets. Although subcluster divisions have been applied to characterize sequences from a wide range of ecosystems (Mohamed et al., 2008; Moisander et al., 2008; Duc et al., 2009; Hamersley et al., 2011; Bonnet et al., 2013; Collavino et al., 2014), these phylogenetically defined groups are not prevalent in the diazotroph studies. Instead, diversity and sample similarity analyses are often based on operational taxonomic units defined at various similarity levels (Hsu and Buckley, 2009; Hamilton et al., 2011; Turk et al., 2011), or on study specific sequence groups called clades (Deslippe and Egger, 2006), operational protein units (Lema et al., 2012), or simply groups (Man-Aharonovich et al., 2007). As demonstrated in our cross-ecosystem study, uniformly applied sequence characterization reveals information not present in individual studies. The rapid and automated cluster assignment that is presented here will be available for general use in the form of a Python script at http://www.pmc.ucsc.edu/~wwwzehr/research/. This novel sequence classification does not require accessing the database or calculating phylogenies and it can be accomplished with less resources and expertise; hence, it will greatly facilitate the exploration and comparison of diazotroph communities.

Figure 1. Sequence logos of NifH in A49 – A113 range are calculated for each main cluster. Arrows point to signature residues identified by two-class CART models: red=primary position, blue=first surrogate, black=subsequent surrogate positions.



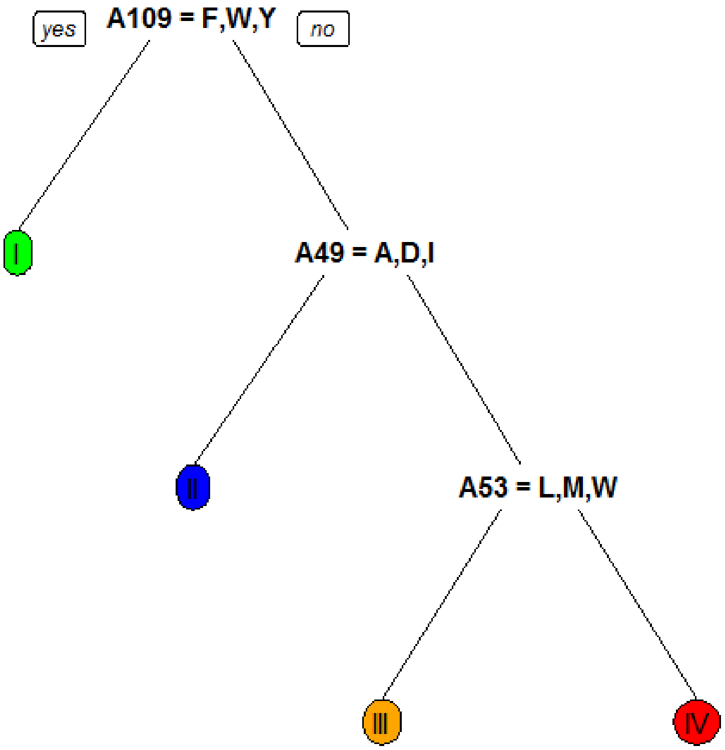
Cluster I

Cluster II

Cluster III

Cluster IV

Figure 2. Graphical representation of the CART classification model that successfully assigns sequences into main clusters based on three residues. The four terminal nodes correspond to clusters I, II, III, and IV. Each decision node lists the protein sequence position and the amino acids in the left group of sequences. For example, if a sequence has phenylalanine (F), tryptophan (W), or tyrosine (Y) at position A109, then it belongs to cluster I. Accuracy of the classification was calculated on the training set (Fit%), on the test set (Test%), and estimated by ten-fold cross-validation (Pred%).



Cluster	I	II	III	IV	TOTAL
Size	17,321	542	3,876	758	22,497
Fit%	99	98	95	96	98
Pred%	99	96	94	95	98
Test%	99	100	92	100	98

Figure 3. Network graphs of NifH sequences at 98% similarity. Top panel: Termite set of 33 unique sequences grouped into 27 OTU98s. Bottom panel: Atlantic set of 281 unique sequences grouped into 66 OTU98s. Sequences are color coded by main cluster: **I, II, III, IV** and labeled by subcluster.

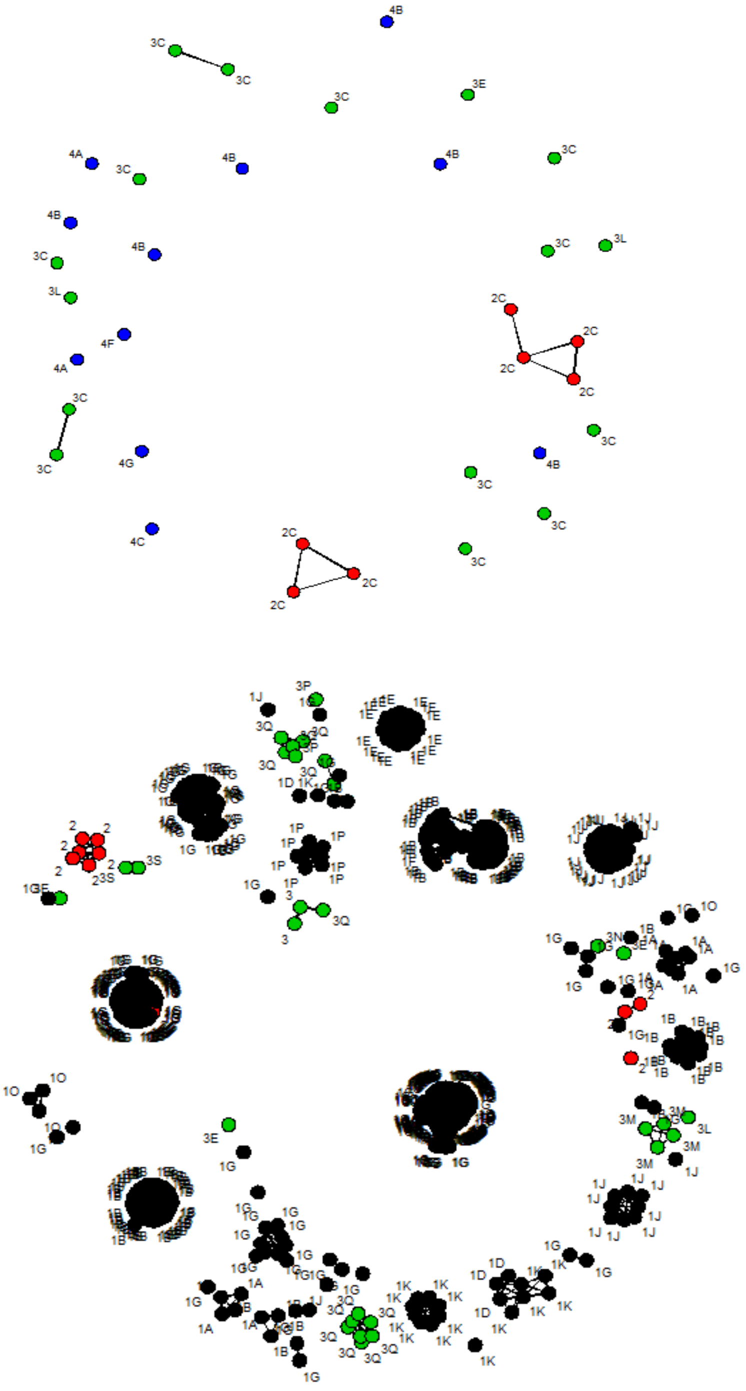


Figure 4. Network of 1,233 sequences merged from thirteen ecosystems and grouped at 98% similarity. Sequences color coded as **terrestrial** and **marine** and labeled by origin.

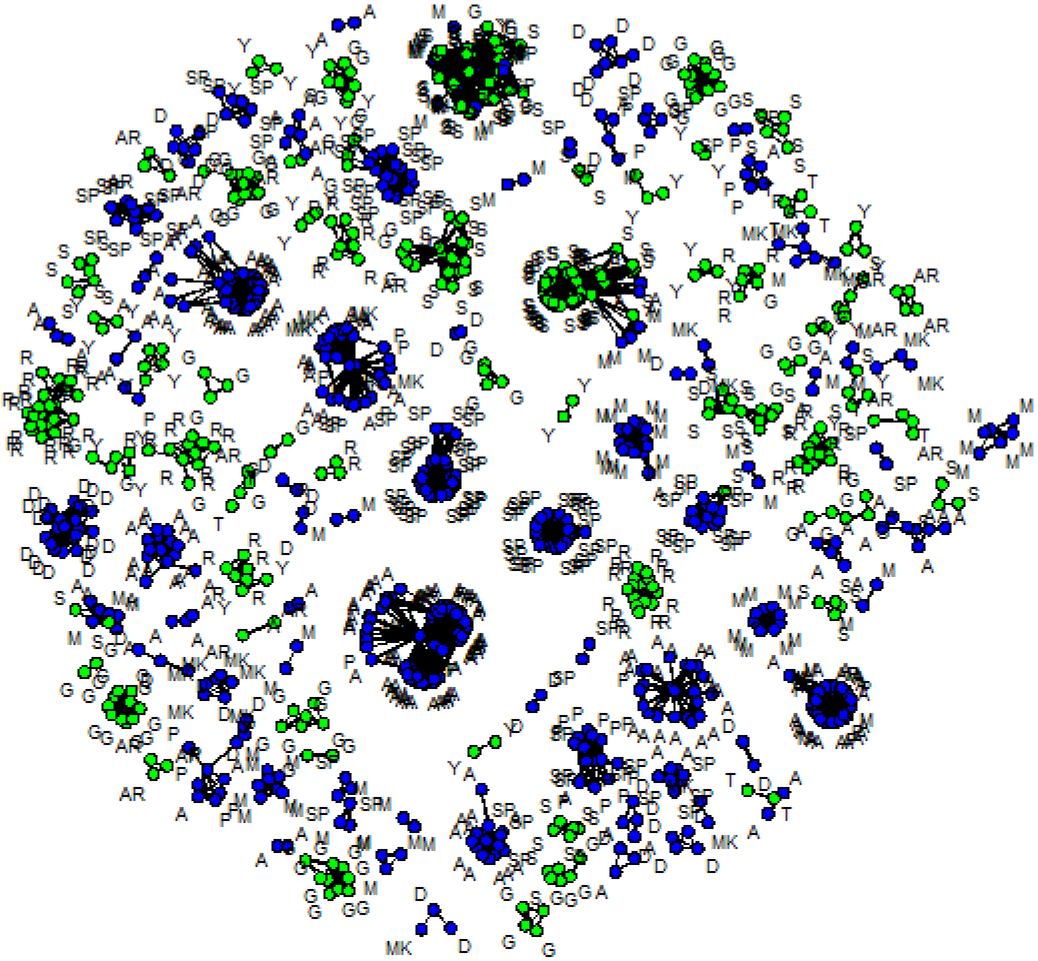


Figure 5. Similarity among diazotroph communities visualized by projections on the first two correspondence analysis components. Communities are color coded as **terrestrial** or **marine** and labeled by origin.
Top panel: projection calculated from contingency table of subcluster proportions.
Bottom panel: projection calculated from ten largest OTU90 proportions.

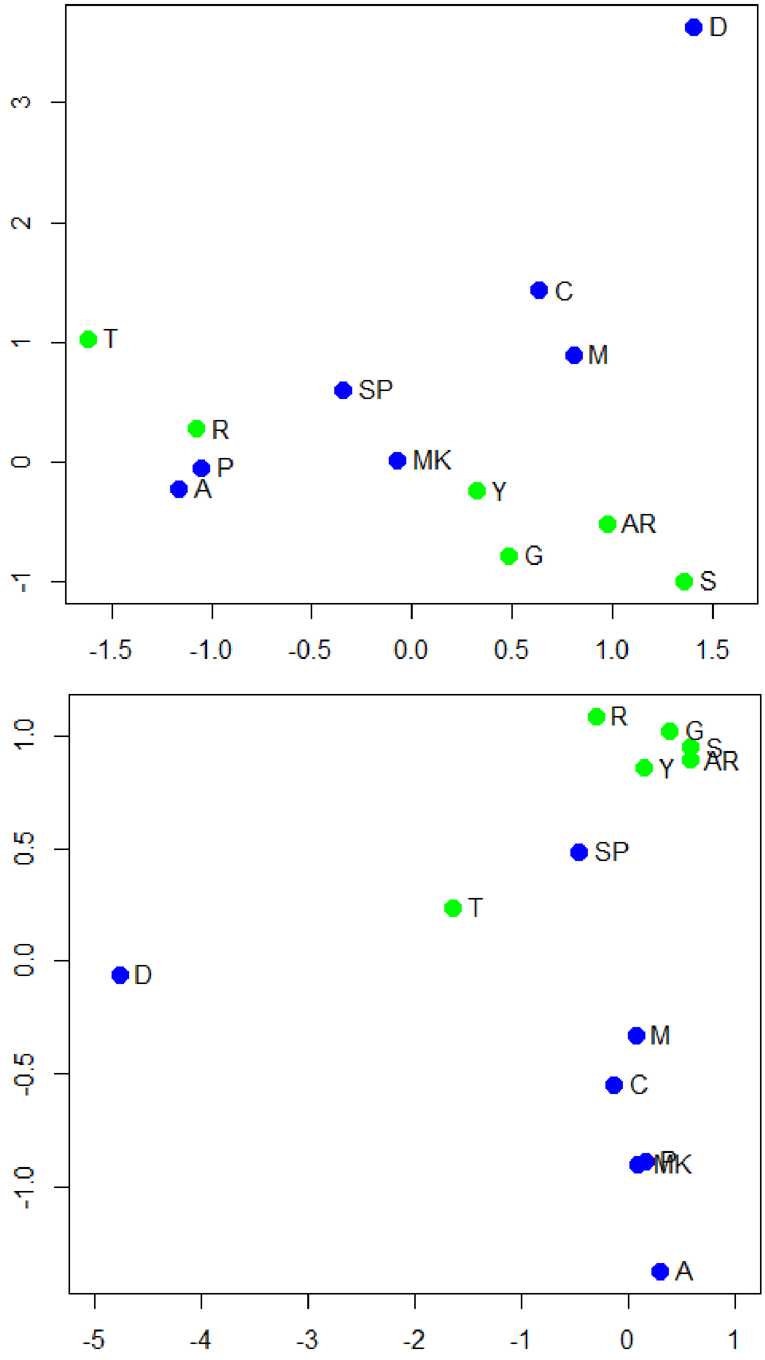


Figure S1. Sequence logo calculated from 22,497 NifH protein sequences. Amino acids are represented by single letter code. Letter size indicates amino acid proportion, and color indicates polarity (red=acidic, blue=basic, green=polar, purple=neutral, black=hydrophobic).

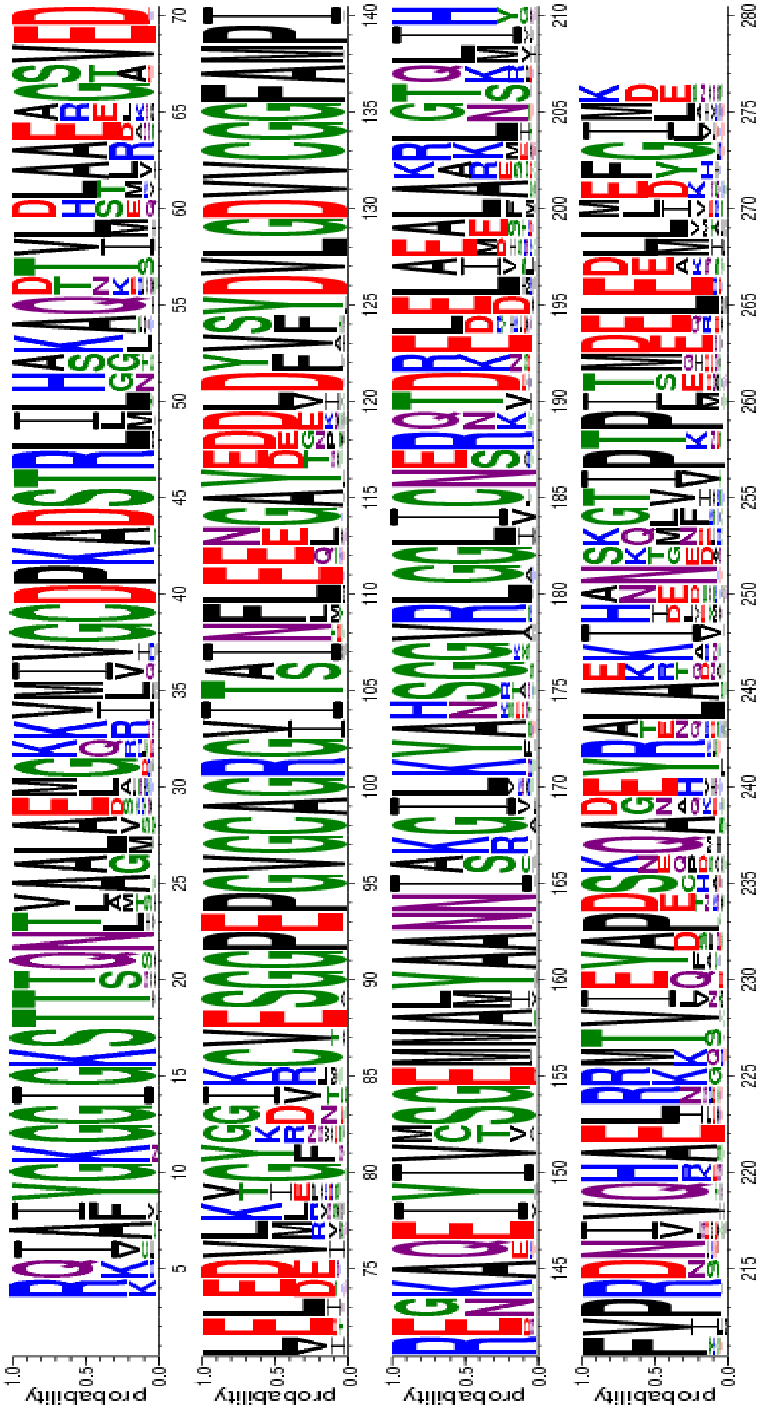


Figure S2. Sequence coverage is uneven in the database. Sequence positions are numbered according to the *Azotobacter vinelandii* NifH protein positions. Green bars indicate conserved residues where at least 98% of sequences have identical amino acids.

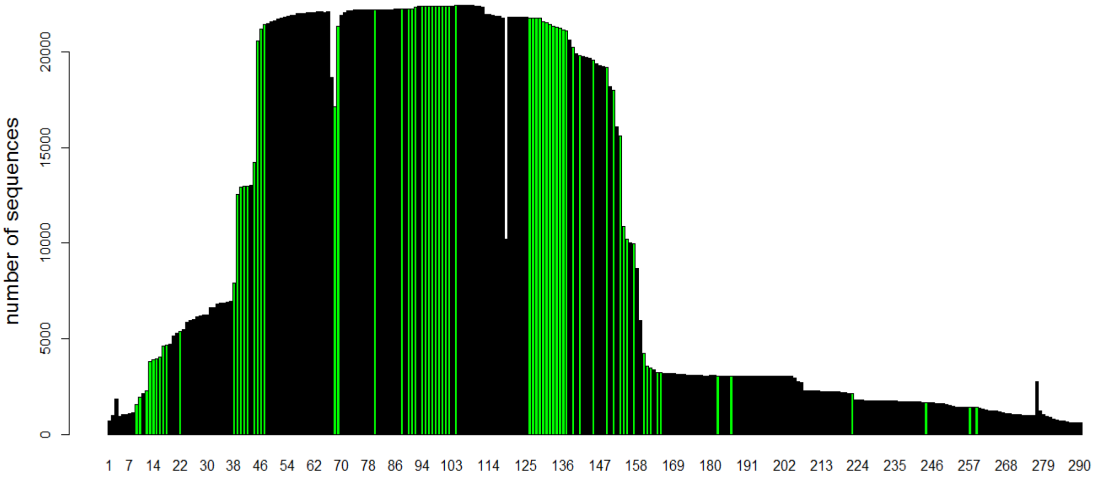


Figure S3. Graphical representation of the CART classification model that splits sequences according to the main clusters before Bacteria and Archaea domains are distinguished. Terminal node labels B1 – B4 indicate Bacteria sequence groups in main clusters I – IV, and labels A2 – A4 denote Archaea sequence groups in main clusters II – IV.

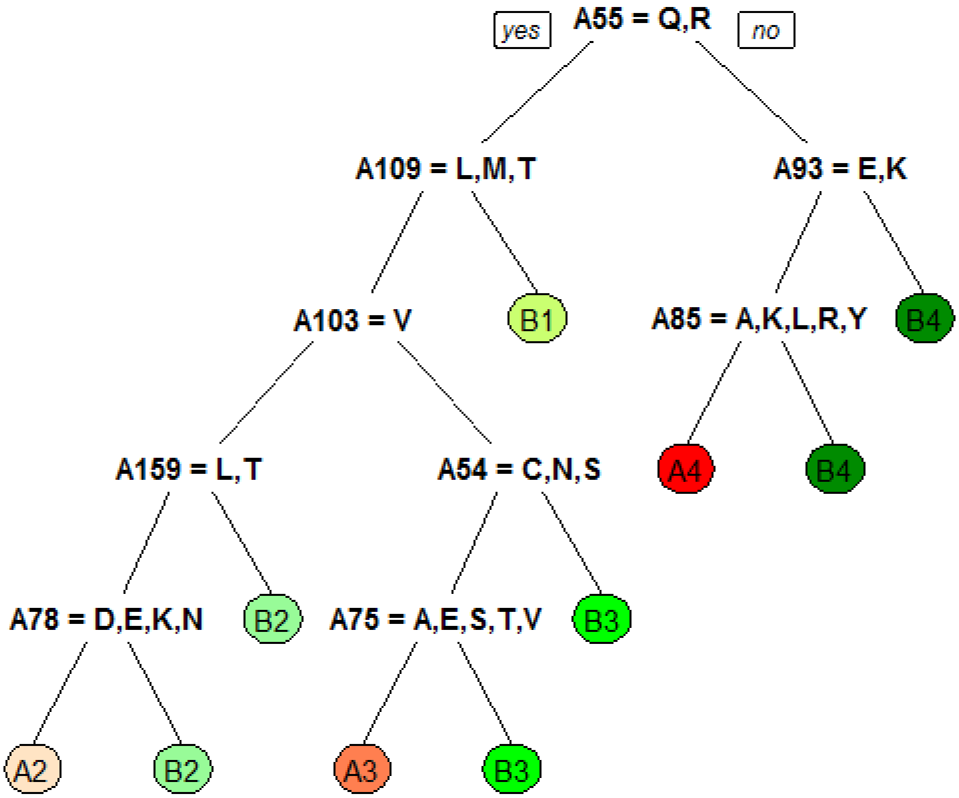
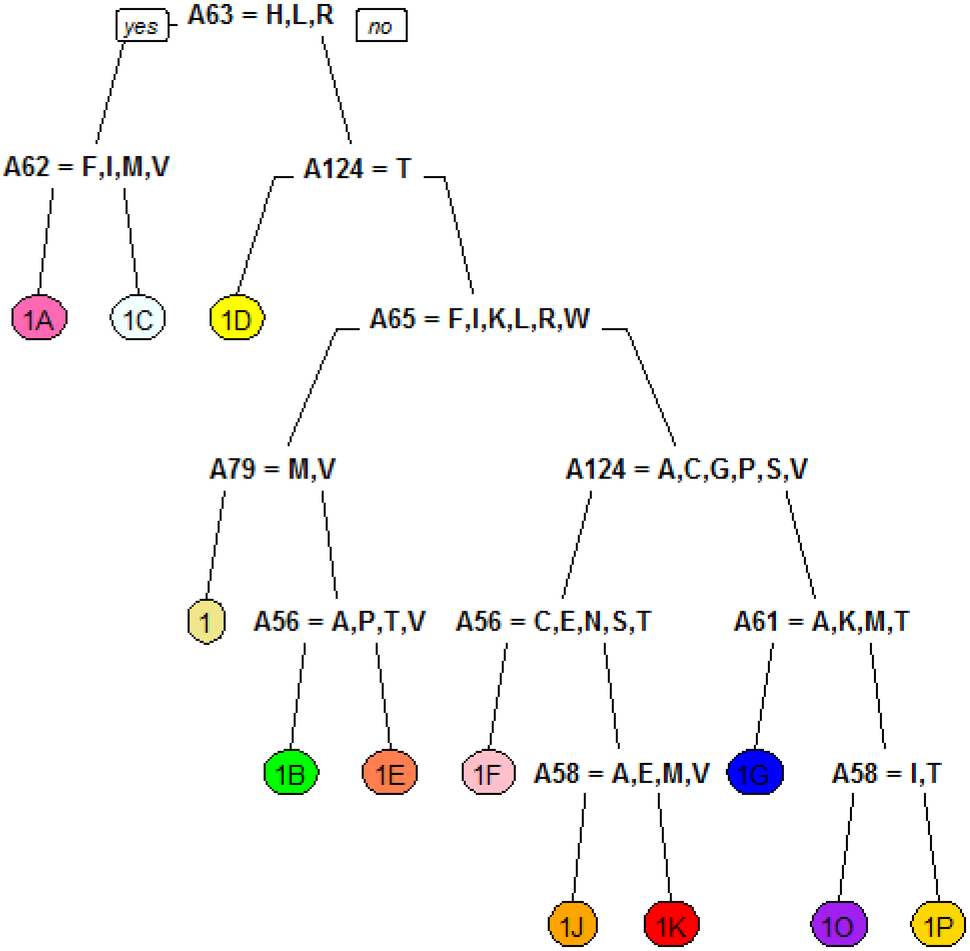


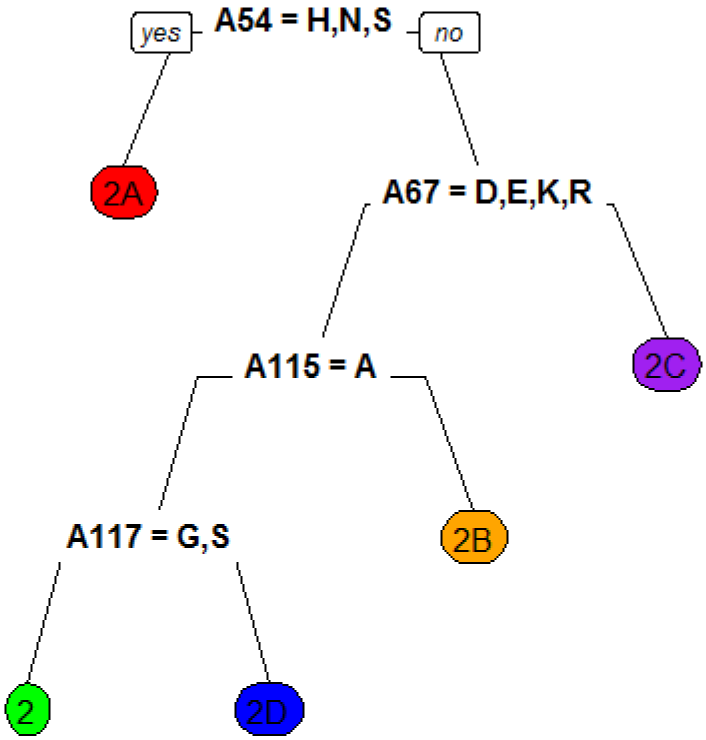
Figure S4. Graphical representation of the CART classification model that assigns sequences into subclusters within cluster I. Terminal nodes correspond to the twelve subclusters defined in the database. Accuracy of the classification was calculated on the training set (Fit%), on the test set (Test%), and estimated by ten-fold cross-validation (Pred%).



Cluster	**1**	**1A**	**1B**	**1C**	**1D**	**1E**	**1F**	**1G**	**1J**	**1K**
Size	4	1,436	3,304	260	566	131	60	1,461	1,451	3,239
Fit%	100	99	96	98	99	96	98	98	94	93
Pred%	75	99	95	97	99	97	95	98	94	93
Test%	0	99	100	100	17	94	--	100	90	82

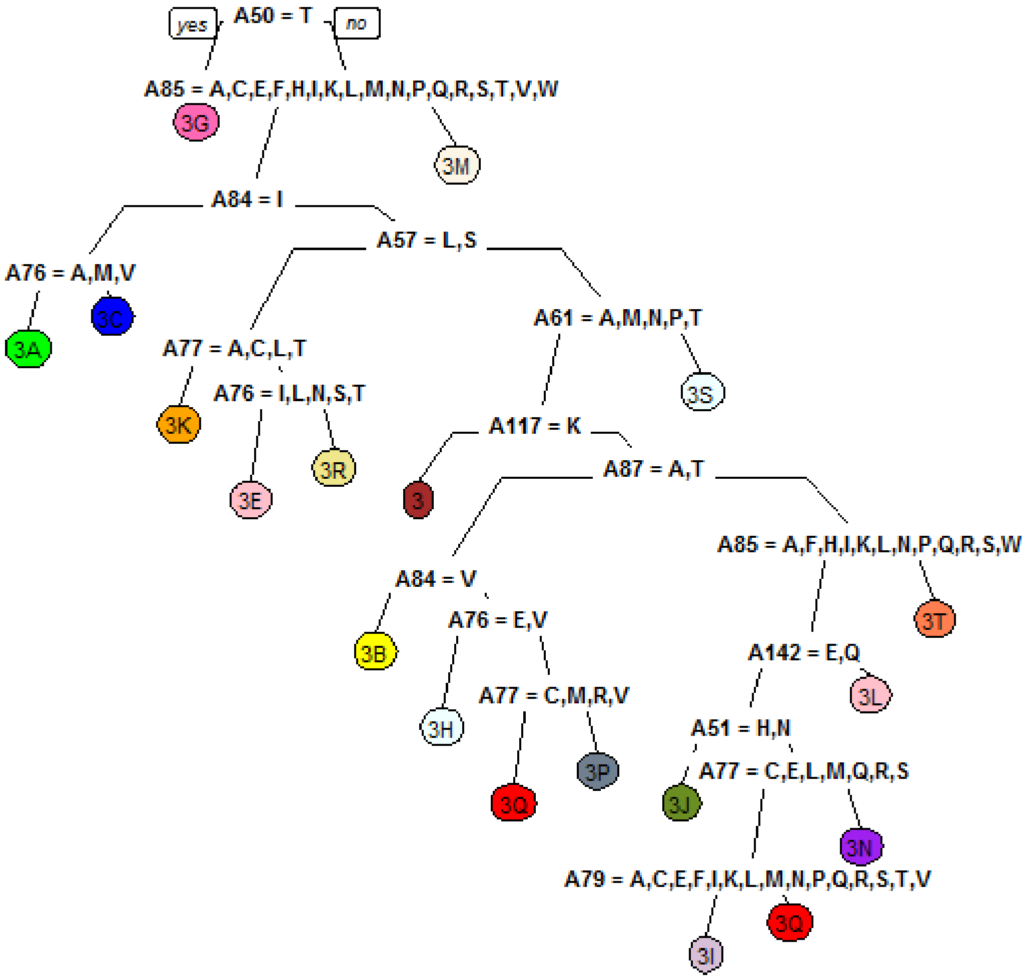
Cluster	**1O**	**1P**	**TOTAL**
Size	304	421	12,637
Fit%	91	96	96
Pred%	89	95	96
Test%	100	100	91

Figure S5. Graphical representation of the CART classification model that assigns sequences into subclusters within cluster II. Terminal nodes correspond to the five subclusters defined in the database. Accuracy of the classification was calculated on the training set (Fit%) and estimated by ten-fold cross-validation (Pred%).



Cluster	**2**	**2A**	**2B**	**2C**	**2D**	**TOTAL**
Size	6	80	78	171	11	346
Fit%	83	99	97	99	100	99
Pred%	0	99	95	99	36	95

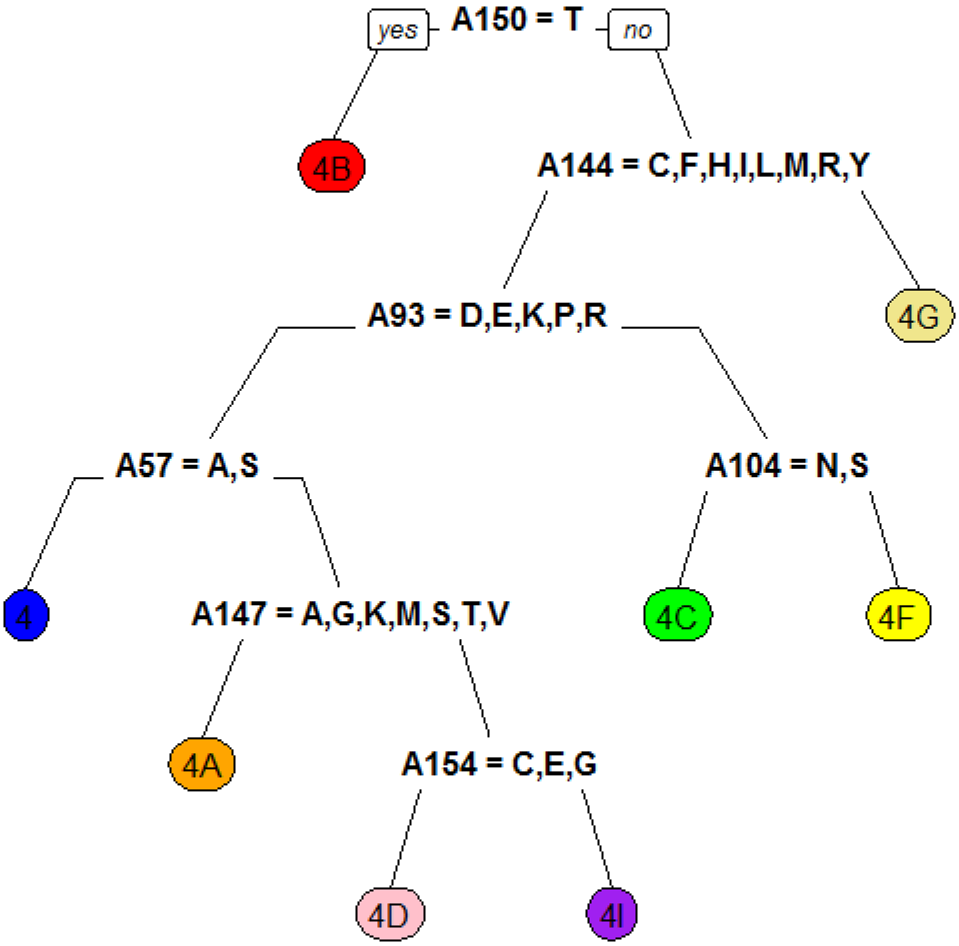
Figure S6. Graphical representation of the CART classification model that assigns sequences into subclusters within cluster III. One of the eighteen subclusters defined in the database, 3Q, is represented by two terminal nodes. Accuracy of the classification was calculated on the training set (Fit%) and estimated by ten-fold cross-validation (Pred%).



Cluster	3	3A	3B	3C	3E	3G	3H	3I	3J	3K
Size	10	167	5	228	771	255	192	341	196	50
Fit%	80	92	100	77	84	90	93	54	74	90
Pred%	70	91	80	77	82	91	92	66	76	86

Cluster	3L	3M	3N	3P	3Q	3R	3S	3T	TOTAL
Size	263	28	114	331	50	34	14	72	3,121
Fit%	73	100	74	44	90	97	86	90	76
Pred%	75	98	75	35	66	97	79	81	76

Figure S7. Graphical representation of the CART classification model that assigns sequences into subclusters within cluster IV. Terminal nodes correspond to the eight subclusters defined in the database. Accuracy of the classification was calculated on the training set (Fit%) and estimated by ten-fold cross-validation (Pred%).



Cluster	4	4A	4B	4C	4D	4F	4G	4I	TOTAL
Size	9	62	88	24	94	150	12	24	463
Fit%	100	87	100	100	95	100	100	96	97
Pred%	89	77	100	100	95	97	100	83	94

Table 1. Origin and type of the thirteen environmental sequence sets used in cross-ecosystem analysis.

Symbol **Name**	**Reference**	**Type of Environment** **Origin of Sequences**
AR Arctic	(Deslippe and Egger, 2006)	Terrestrial, Rhizosphere, Arctic Canada shrubs
A Atlantic	(Turk et al., 2011)	Marine, Surface, Cape Verde, Atlantic Ocean
C Chesapeake	(Burns et al., 2002)	Marine, Sediment, Chesapeake Bay and Neuse river
D Deep	(Mehta et al., 2003)	Marine, Deep, NE Pacific Ocean
G Glacier	(Duc et al., 2009)	Terrestrial, Soil, Swiss Alps glaciers
M Mediterranean	(Man-Aharonovich et al., 2007)	Marine, Surface, East Mediterranean Sea
MK Mekong	(Bombar et al., 2011)	Marine, Surface, Mekong River Plume, South China Sea
P Pacific	(Zehr et al., 2007)	Marine, Surface, Pacific Ocean, near Hawaii
R Rainforest	(Furnkranz et al., 2008)	Terrestrial, Phyllosphere, Costa Rica rainforest
S Soil	(Hsu and Buckley, 2009)	Terrestrial, Soil, New York State
SP Sponge	(Mohamed et al., 2008)	Marine, Symbiont, Key Largo sponges
T Termite	(Du et al., 2012)	Terrestrial, Symbiont, Chinese termite gut
Y Yellowstone	(Hamilton et al., 2011)	Terrestrial, Microbial mat, Yellowstone Park geothermal springs

Table 2. Size and structure of the thirteen environmental sequence sets used in cross-ecosystem analysis.

Set	Sequence	Unique	OTU98	OTU90	Main clusters	Subclusters
AR	42	30	20	9	I, IV	7, 1
A	603	281	66	27	I, II, III	10, 1, 8
C	17	17	17	11	I, III	4, 6
D	120	85	36	17	I, II, III, IV	4, 2, 4, 5
G	318	139	56	20	I, II, III	9, 2, 7
M	191	103	38	15	I, II, III, IV	9, 2, 4, 1
MK	57	40	24	7	I, III	6, 3
P	86	60	23	7	I	4
R	137	103	28	5	I, II, III, IV	3, 1, 1, 1
S	415	162	55	8	I	9
SP	347	123	26	7	I, II, III	3, 1, 3
T	34	33	27	18	II, III, IV	1, 3, 5
Y	66	57	27	11	I, II, III	9, 1, 1
ALL	2433	1233	443	162	I, II, III, IV	12, 4, 15, 7

Table 3. Potentially mislabeled sequence proportions estimated by network analysis of thirteen environmental sequence sets. An OTU is called pure if all its sequences are assigned to the same subcluster and a sequence is flagged mislabeled if its CART derived cluster assignment does not match the dominant label within the OTU.

	98% similarity				**90% similarity**			
Set	**OTU**	**Pure OTU**	**Mislabeled sequences**	**% miss**	**OTU**	**Pure OTU**	**Mislabeled sequences**	**% miss**
AR	20	20	0/42	0	9	8	1/42	2
A	66	61	9/603	1	27	22	48/603	8
C	17	17	0/17	0	11	9	2/17	12
D	36	34	2/120	2	17	13	6/120	5
G	56	55	1/318	<1	20	17	23/318	7
M	38	35	3/191	2	15	10	38/191	20
MK	24	24	0/57	0	7	6	8/57	14
P	23	23	0/86	0	7	7	0/86	0
R	28	26	2/137	1	5	4	3/137	2
S	55	54	1/415	<1	8	4	42/415	10
SP	26	26	0/347	0	7	6	1/347	<1
T	27	27	0/34	0	18	18	0/34	0
Y	27	27	0/66	0	11	9	5/66	8

Table 4. Identity of the largest OTU98s resulting from joint binning of thirteen environmental sequence sets. Each group is identified by the majority subcluster label, by the origin of its sequences, and by the closest relative (pBLAST) of its representative sequence. Type is called “unique” if all sequences are from a single data set, and “mixed” if sequences originate from terrestrial and marine habitats.

ID	Size	Label	Origin	Type	Representative’s match	Identity %
1	75	1G	A, M, MK, P	marine	*Pseudomonas stutzeri*	94
2	64	1K	A, G, M, P, S, SP, Y	mixed	*Burkholderia xenovorans*	100
3	51	1K	A, M, S	mixed	*Azospirillum amazonense*	100
4	38	1G	A	unique	*Azotobacter vinelandii*	97
5	34	1B	A, M, P	marine	UCYN-A	100
6	32	1B	A, MK, P, SP	marine	*Trichodesmium erythraeum*	100
7	21	1G	A, P	marine	*Pseudomonas stutzeri*	95
8	21	1B	R	unique	*Nostoc*. sp. PCC7107	100
9	21	1B	SP	unique	*Cyanothece* sp.	94
10	19	2B	D	unique	uncultivated	--
11	19	3E	SP	unique	*Desulfovibrio oxyclinae*	94
12	18	1A	AR, G, S	terrestrial	*Geobacter uraniireducens*	97
13	18	1B	G, SP, Y	mixed	*Nodosilinea nodulosa*	98
14	15	1B	A, P, SP	marine	*Fischerella muscicola*	98
15	14	1J	A	unique	*Sinorhizobium meliloti*	96
16	14	2	R	unique	*Dickeya dadantii*	98
17	14	1B	A, SP	marine	*Trichodesmium erythraeum*	97
18	13	1P	A, M	marine	*Dechloromonas aromatica*	96
19	12	1K	S, SP	mixed	*Bradyrhizobium elkanii*	97
20	11	3M	G	unique	uncultivated	--
25	10	1B	G, R	mixed	*Nostoc punctiforme*	97
29	8	1J	A, M, S	mixed	*Dechlorosoma suillum*	100
30	8	1K	G, Y	terrestrial	*Methylocystis parvus*	96
31	8	1G	A, P	marine	*Allochromatium vinosum*	96
33	7	2	A, SP	marine	*Desulfovibrio desulfuricans*	100
36	7	1B	P, Y	mixed	*Fischerella muscicola*	100
40	6	1A	G, S	terrestrial	*Geobacter uraniireducens*	96
42	6	1A	A, M, MK	marine	*Desulfuromonas acetoxidans*	99

References

Allen, A.E., Booth, M.G., Frischer, M.E., Verity, P.G., Zehr, J.P., and Zani, S. (2001) Diversity and detection of nitrate assimilation genes in marine bacteria. *Applied and Environmental Microbiology* **67**: 5343-5348.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology* **215**: 403-410.

Bazinet, A.L., and Cummings, M.P. (2012) A comparative evaluation of sequence classification programs. *Bmc Bioinformatics* **13**: 92

Betancourt, D.A., Loveless, T.M., Brown, J.W., and Bishop, P.E. (2008) Characterization of diazotrophs containing Mo-independent nitrogenases, isolated from diverse natural environments. *Applied and Environmental Microbiology* **74**: 3471-3480.

Bishop, P.E., and Joerger, R.D. (1990) Genetics and molecular-biology of alternative nitrogen-fixation systems. *Annual Review of Plant Physiology and Plant Molecular Biology* **41**: 109-125.

Bombar, D., Moisander, P.H., Dippner, J.W., Foster, R.A., Voss, M., Karfeld, B., and Zehr, J.P. (2011) Distribution of diazotrophic microorganisms and nifH gene expression in the Mekong River plume during intermonsoon. *Marine Ecology Progress Series* **424**: 39-U55.

Bonnet, S., Dekaezemacker, J., Turk-Kubo, K.A., Moutin, T., Hamersley, R.M., Grosso, O. et al. (2013) Aphotic N-2 Fixation in the Eastern Tropical South Pacific Ocean. *Plos One* **8**.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1983) *Classification and Regression Trees*: Wadsworth.

Burns, J.A., Zehr, J.P., and Capone, D.G. (2002) Nitrogen-fixing phylotypes of Chesapeake Bay and Neuse River estuary sediments. *Microbial Ecology* **44**: 336-343.

Butts, C., Handcock, M., and Hunter, D. (2014) network: Classes for Relational Data. R package version 1.9.0.

Charif, D., Lobry, J., Necsulea, A., Palmeira, L., Penel, S., and Perriere, G. (2012) seqinr: Biological Sequences Retrieval and Analysis. R package version 3.0.6.

Chien, Y.T., and Zinder, S.H. (1994) Cloning, DNA sequencing, and characterization of *nifD*-homologous gene from the archaeon *Methanosarcina barkeri* 227 which resembles *nifD*1 from the eubacterium *Clostridium pasteurianum*. *Journal of Bacteriology* **176**: 6590-6598.

Chien, Y.T., Auerbuch, V., Brabban, A.D., and Zinder, S.H. (2000) Analysis of genes encoding an alternative nitrogenase in the archaeon Methanosarcina barkeri 227. *Journal of Bacteriology* **182**: 3247-3253.

Clarke, K.R., Somerfield, P.J., and Gorley, R.N. (2008) Testing of null hypotheses in exploratory community analyses: similarity profiles and biota-environment linkage. *Journal of Experimental Marine Biology and Ecology* **366**: 56-69.

Cole, J.R., Wang, Q., Fish, J.A., Chai, B.L., McGarrell, D.M., Sun, Y.N. et al. (2014) Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research* **42**: D633-D642.

Collavino, M., Tripp, J., Frank, I., Vidoz, M., Calderoli, P., Donato, M. et al. (2014) nifH pyrosequencing reveals the potential for location-specific soil chemistry to influence N2-fixing community dynamics. *Environmental Microbiology* doi: 10.1111/1462-2920.12423.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004) WebLogo: A sequence logo generator. *Genome Research* **14**: 1188-1190.

Curtis, T.P., Sloan, W.T., and Scannell, J.W. (2002) Estimating prokaryotic diversity and its limits. *Proceedings of the National Academy of Sciences of the United States of America* **99**: 10494-10499.

De'ath, G., and Fabricius, K.E. (2000) Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology* **81**: 3178-3192.

DeLong, E.F. (2009) The microbial ocean from genomes to biomes. *Nature* **459**: 200-206.

Desai, M.S., and Brune, A. (2012) Bacteroidales ectosymbionts of gut flagellates shape the nitrogen-fixing community in dry-wood termites. *Isme Journal* **6**: 1302-1313.

Desai, M.S., Assig, K., and Dattagupta, S. (2013) Nitrogen fixation in distinct microbial niches within a chemoautotrophy-driven cave ecosystem. *Isme Journal* **7**: 2411-2423.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K. et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology* **72**: 5069-5072.

Deslippe, J.R., and Egger, K.N. (2006) Molecular diversity of nifH genes from bacteria associated with high arctic dwarf shrubs. *Microbial Ecology* **51**: 516-525.

Dos Santos, P., Fang, Z., Mason, S., Setubal, J., and Dixon, R. (2012) Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. *BMC Genomics* **13**: 1-12.

Du, X., Li, X.J., Wang, Y., Peng, J.X., Hong, H.Z., and Yang, H. (2012) Phylogenetic Diversity of Nitrogen Fixation Genes in the Intestinal Tract of Reticulitermes chinensis Snyder. *Current Microbiology* **65**: 547-551.

Duc, L., Noll, M., Meier, B.E., Burgmann, H., and Zeyer, J. (2009) High Diversity of Diazotrophs in the Forefield of a Receding Alpine Glacier. *Microbial Ecology* **57**: 179-190.

Dumont, M.G., Luke, C., Deng, Y.C., and Frenzel, P. (2014) Classification of pmoA amplicon pyrosequences using BLAST and the lowest common ancestor method in MEGAN. *Frontiers in Microbiology* **5**: 34

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460-2461.

Elshahed, M.S., Youssef, N.H., Spain, A.M., Sheik, C., Najar, F.Z., Sukharnikov, L.O. et al. (2008) Novelty and uniqueness patterns of rare members of the soil biosphere. *Applied and Environmental Microbiology* **74**: 5422-5428.

Falkowski, P.G. (1997) Evolution of the nitrogen cycle and its influence on the biological sequestration of CO2 in the ocean. *Nature* **387**: 272-275.

Falkowski, P.G., Fenchel, T., and Delong, E.F. (2008) The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**: 1034-1039.

Farnelid, H., Andersson, A.F., Bertilsson, S., Abu Al-Soud, W., Hansen, L.H., Sorensen, S. et al. (2011) Nitrogenase Gene Amplicons from Global Marine Surface Waters Are Dominated by Genes of Non-Cyanobacteria. *Plos One* **6** (4): e19223.

Fierer, N., and Ladau, J. (2012) Predicting microbial distributions in space and time. *Nature Methods* **9**: 549-551.

Fong, A.A., Karl, D.M., Lukas, R., Letelier, R.M., Zehr, J.P., and Church, M.J. (2008) Nitrogen fixation in an anticyclonic eddy in the oligotrophic North Pacific Ocean. *Isme Journal* **2**: 663-676.

Fox, G.E., Stackebrandt, E., Hespell, R.B., Gibson, J., Maniloff, J., Dyer, T.A. et al. (1980) The phylogeny of prokaryotes. *Science* **209**: 457-463.

Furnkranz, M., Wanek, W., Richter, A., Abell, G., Rasche, F., and Sessitsch, A. (2008) Nitrogen fixation by phyllosphere bacteria associated with higher plants and their colonizing epiphytes of a tropical lowland rainforest of Costa Rica. *Isme Journal* **2**: 561-570.

Gaby, J.C., and Buckley, D.H. (2011) A global census of nitrogenase diversity. *Environmental Microbiology* **13**: 1790-1799.

Gaby, J.C., and Buckley, D.H. (2012) A Comprehensive Evaluation of PCR Primers to Amplify the nifH Gene of Nitrogenase. *Plos One* **7** (7): e42149.

Galloway, J.N., Dentener, F.J., Capone, D.G., Boyer, E.W., Howarth, R.W., Seitzinger, S.P. et al. (2004) Nitrogen cycles: past, present, and future. *Biogeochemistry* **70**: 153-226.

Georgiadis, M.M., Komiya, H., Chakrabarti, P., Woo, D., Kornuc, J.J., and Rees, D.C. (1992) Crystallographic structure of the nitrogenase iron protein from *Azotobacter vinelandii*. *Science* **257**: 1653-1659.

Halm, H., Lam, P., Ferdelman, T.G., Lavik, G., Dittmar, T., LaRoche, J. et al. (2012) Heterotrophic organisms dominate nitrogen fixation in the South Pacific Gyre. *Isme Journal* **6**: 1238-1249.

Hamersley, M.R., Turk, K.A., Leinweber, A., Gruber, N., Zehr, J.P., Gunderson, T., and Capone, D.G. (2011) Nitrogen fixation within the water column associated with two hypoxic basins in the Southern California Bight. *Aquatic Microbial Ecology* **63**: 193-205.

Hamilton, T.L., Boyd, E.S., and Peters, J.W. (2011) Environmental Constraints Underpin the Distribution and Phylogenetic Diversity of nifH in the Yellowstone Geothermal Complex. *Microbial Ecology* **61**: 860-870.

Handelsman, J. (2004) Metagenomics: Application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews* **68**: 669-685.

Hsu, S.F., and Buckley, D.H. (2009) Evidence for the functional significance of diazotroph community structure in soil. *Isme Journal* **3**: 124-136.

Jones, C.M., Graf, D.R.H., Bru, D., Philippot, L., and Hallin, S. (2013) The unaccounted yet abundant nitrous oxide-reducing microbial community: a potential nitrous oxide sink. *Isme Journal* **7**: 417-426.

Lane, D.J., Pace, B., Olsen, G.J., Stahl, D.A., Sogin, M.L., and Pace, N.R. (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proceedings of the National Academy of Sciences of the United States of America* **82**: 6955-6959.

Lema, K.A., Willis, B.L., and Bourne, D.G. (2012) Corals Form Characteristic Associations with Symbiotic Nitrogen-Fixing Bacteria. *Applied and Environmental Microbiology* **78**: 3136-3144.

Li, W.Z., and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658-1659.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar et al. (2004) ARB: a software environment for sequence data. *Nucleic Acids Research* **32**: 1363-1371.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2014) cluster: Cluster Analysis Basics and Extensions. R package version 1.15.2.

Man-Aharonovich, D., Kress, N., Bar Zeev, E., Berman-Frank, I., and Beja, O. (2007) Molecular ecology of nifH genes and transcripts in the eastern Mediterranean Sea. *Environmental Microbiology* **9**: 2354-2363.

Mehta, M.P., Butterfield, D.A., and Baross, J.A. (2003) Phylogenetic diversity of nitrogenase (nifH) genes in deep-sea and hydrothermal vent environments of the Juan de Fuca ridge. *Applied and Environmental Microbiology* **69**: 960-970.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M. et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *Bmc Bioinformatics* **9**: 386

Milborrow, S. (2014) Plot rpart models. An enhanced version of plot.rpart. R package version 1.4.4.

Mohamed, N.M., Colman, A.S., Tal, Y., and Hill, R.T. (2008) Diversity and expression of nitrogen fixation genes in bacterial symbionts of marine sponges. *Environmental Microbiology* **10**: 2910-2921.

Moisander, P.H., Beinart, R.A., Voss, M., and Zehr, J.P. (2008) Diversity and abundance of diazotrophic microorganisms in the South China Sea during intermonsoon. *Isme Journal* **2**: 954-967.

Moisander, P.H., Morrison, A.E., Ward, B.B., Jenkins, B.D., and Zehr, J.P. (2007) Spatial-temporal variability in diazotroph assemblages in Chesapeake Bay using an oligonucleotide nifH microarray. *Environmental Microbiology* **9**: 1823-1835.

Mulder, A., Vandegraaf, A.A., Robertson, L.A., and Kuenen, J.G. (1995) Anaerobic ammonium oxidation discovered in a denitrifying fluidized-bed reactor. *Fems Microbiology Ecology* **16**: 177-183.

Nenadic, O., and Greenacre, M. (2007) Correspondence analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software* **20**.

Oksanen, J., Blanchet, F., Kindt, R., Legendre, P., Minchin, P., O'Hara, R. et al. (2013) vegan: Community Ecology Package. R package version 2.0.10.

Pedros-Alio, C. (2006) Marine microbial diversity: can it be determined? *Trends in Microbiology* **14**: 257-263.

Pesch, R., Schmidt, G., Schroeder, W., and Weustermann, I. (2011) Application of CART in ecological landscape mapping: Two case studies. *Ecological Indicators* **11**: 115-122.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010) FastTree 2-Approximately Maximum-Likelihood Trees for Large Alignments. *Plos One* **5**.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W.G., Peplies, J., and Glockner, F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* **35**: 7188-7196.

R Development Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rahav, E., Bar-Zeev, E., Ohayon, S., Elifantz, H., Belkin, N., Herut, B. et al. (2013) Dinitrogen fixation in aphotic oxygenated marine environments. *Frontiers in Microbiology* **4**: 227.

Rappe, M.S., and Giovannoni, S.J. (2003) The uncultured microbial majority. *Annual Review of Microbiology* **57**: 369-394.

Raymond, J., Siefert, J.L., Staples, C.R., and Blankenship, R.E. (2004) The natural history of nitrogen fixation. *Molecular Biology and Evolution* **21**: 541-554.

Santos, S.R., and Ochman, H. (2004) Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins. *Environmental Microbiology* **6**: 754-759.

Schlessman, J.L., Woo, D., Joshua-Tor, L., Howard, J.B., and Rees, D.C. (1998) Conformational variability in structures of the nitrogenase iron proteins from Azotobacter vinelandii and Clostridium pasteurianum. *Journal of Molecular Biology* **280**: 669-685.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B. et al. (2009) Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology* **75**: 7537-7541.

Seshadri, R., Kravitz, S.A., Smarr, L., Gilna, P., and Frazier, M. (2007) CAMERA: A community resource for metagenomics. *Plos Biology* **5**: 394-397.

Sogin, M.L., Morrison, H.G., Huber, J.A., Mark Welch, D., Huse, S.M., Neal, P.R. et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America* **103**: 12115-12120.

Stacey, G., Burris, R.H., and Evans, H.J. (eds) (1992) *Biological nitrogen fixation*: Springer.

Steward, G.F., Zehr, J.P., Jellison, R., Montoya, J.P., and Hollibaugh, J.T. (2004) Vertical distribution of nitrogen-fixing phylotypes in a meromictic, hypersaline lake. *Microbial Ecology* **47**: 30-40.

Therneau, T., Atkinson, B., and Ripley, B. (2014) rpart: Recursive Partitioning and Regression Trees. R package version 4.1.8.

Triplett, E.W. (ed) (2000) *Prokaryotic nitrogen fixation: a model system for the analysis of a biological process*. Wymondham, Great Britain: Horizon Scientific Press.

Turk, K.A., Rees, A.P., Zehr, J.P., Pereira, N., Swift, P., Shelley, R. et al. (2011) Nitrogen fixation and nitrogenase (nifH) expression in tropical waters of the eastern North Atlantic. *Isme Journal* **5**: 1201-1212.

Ueda, T., Suga, Y., Yahiro, N., and Matsuguchi, T. (1995) Remarkable N_2 fixing bacterial diversity detected in rice roots by molecular evolutionary analysis of *nifH* gene sequences. *Journal of Bacteriology* **177**: 1414-1417.

Usio, N., Nakajima, H., Kamiyama, R., Wakana, I., Hiruta, S., and Takamura, N. (2006) Predicting the distribution of invasive crayfish (Pacifastacus leniusculus) in a Kusiro Moor marsh (Japan) using classification and regression trees. *Ecological Research* **21**: 271-277.

Vitousek, P.M., and Howarth, R.W. (1991) Nitrogen limitation on land and in the sea: how can it occur? *Biogeochemistry* **13**: 87-115.

Wagner, M., Roger, A.J., Flax, J.L., Brusseau, G.A., and Stahl, D.A. (1998) Phylogeny of dissimilatory sulfite reductases supports an early origin of sulfate respiration. *Journal of Bacteriology* **180**: 2975-2982.

Wang, Q., Garrity, G.M., Tiedje, J.M., and Cole, J.R. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* **73**: 5261-5267.

Ward, B.B., Capone, D.G., and Zehr, J.P. (2007) What's New in the Nitrogen Cycle? *Oceanography* **20**: 101-109.

Ward, B.B., Devol, A.H., Rich, J.J., Chang, B.X., Bulow, S.E., Naik, H. et al. (2009) Denitrification as the dominant nitrogen loss process in the Arabian Sea. *Nature* **461**: 78-81.

Wooley, J.C., Godzik, A., and Friedberg, I. (2010) A Primer on Metagenomics. *Plos Computational Biology* **6** (2): e1000667

Yamada, A., Inoue, T., Noda, S., Hongoh, Y., and Ohkuma, M. (2007) Evolutionary trend of phylogenetic diversity of nitrogen fixation genes in the gut community of wood-feeding termites. *Molecular Ecology* **16**: 3768-3777.

Zehr, J.P. (2010) Microbes in Earth's aqueous environments. *Frontiers in Microbiology* **1**: 4.

Zehr, J.P. (2011) Nitrogen fixation by marine cyanobacteria. *Trends in Microbiology* **19**: 162-173.

Zehr, J.P., and McReynolds, L.A. (1989) Use of Degenerate Oligonucleotides for Amplification of the *nifH* Gene from the Marine Cyanobacterium *Trichodesmium thiebautii*. *Applied and Environmental Microbiology* **55**: 2522-2526.

Zehr, J.P., and Kudela, R.M. (2011) Nitrogen Cycle of the Open Ocean: From Genes to Ecosystems. *Annual Review of Marine Science, Vol 3* **3**: 197-225.

Zehr, J.P., Hewson, I., and Moisander, P. (2009) Molecular biology techniques and applications for ocean sensing. *Ocean Science* **5**: 101-113.

Zehr, J.P., Harris, D., Dominic, B., and Salerno, J. (1997) Structural analysis of the Trichodesmium nitrogenase iron protein: implications for aerobic nitrogen fixation activity. *Fems Microbiology Letters* **153**: 303-309.

Zehr, J.P., Jenkins, B.D., Short, S.M., and Steward, G.F. (2003) Nitrogenase gene diversity and microbial community structure: a cross-system comparison. *Environmental Microbiology* **5**: 539-554.

Zehr, J.P., Mellon, M., Braun, S., Litaker, W., Steppe, T., and Paerl, H.W. (1995) Diversity of Heterotrophic Nitrogen-fixation Genes in a Marine Cyanobacterial Mat. *Applied and Environmental Microbiology* **61**: 2527-2532.

Zehr, J.P., Montoya, J.P., Jenkins, B.D., Hewson, I., Mondragon, E., Short, C.M. et al. (2007) Experiments linking nitrogenase gene expression to nitrogen fixation in the North Pacific subtropical gyre. *Limnology and Oceanography* **52**: 169-183.