BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 20 2014, pages 2883–2890
doi:10.1093/bioinformatics/btu417

Sequence analysis

Advance Access publication July 2, 2014

ARBitrator: a software pipeline for on-demand retrieval of auto-curated *nifH* sequences from GenBank

Philip Heller[1], H. James Tripp[2], Kendra Turk-Kubo[3] and Jonathan P. Zehr[3,*]

[1]Department of Biomolecular Engineering, University of California, Santa Cruz, CA 95064, USA, [2]Department of Energy (DOE) Joint Genome Institute, Walnut Creek, CA 94598, USA and [3]Department of Ocean Sciences, University of California, Santa Cruz, CA 95064, USA

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Studies of the biochemical functions and activities of uncultivated microorganisms in the environment require analysis of DNA sequences for phylogenetic characterization and for the development of sequence-based assays for the detection of microorganisms. The numbers of sequences for genes that are indicators of environmentally important functions such as nitrogen (N_2) fixation have been rapidly growing over the past few decades. Obtaining these sequences from the National Center for Biotechnology Information's GenBank database is problematic because of annotation errors, nomenclature variation and paralogues; moreover, GenBank's structure and tools are not conducive to searching solely by function. For some genes, such as the nifH gene commonly used to assess community potential for N_2 fixation, manual collection and curation are becoming intractable because of the large number of sequences in GenBank and the large number of highly similar paralogues. If analysis is to keep pace with sequence discovery, an automated retrieval and curation system is necessary.

Results: ARBitrator uses a two-step process composed of a broad collection of potential homologues followed by screening with a best hit strategy to conserved domains. 34 420 nifH sequences were identified in GenBank as of November 20, 2012. The false-positive rate is ~0.033%. ARBitrator rapidly updates a public nifH sequence database, and we show that it can be adapted for other genes.

Availability and implementation: Java source and executable code are freely available to non-commercial users at http://pmc.ucsc.edu/~wwwzehr/research/database/.

Contact: zehrj@ucsc.edu

Supplementary information: Supplementary information is available at *Bioinformatics* online.

Received on January 15, 2014; revised on June 10, 2014; accepted on June 25, 2014

1 INTRODUCTION

Microorganisms catalyze a variety of biogeochemical transformations, such as nitrogen (N_2) fixation, that are critical for ecosystem function. Because many microorganisms in the environment have not been cultivated, studies of such functions depend on amplification and sequencing of genes that encode proteins involved in the functions of interest (Ueda *et al.*, 1995; Zehr *et al.*, 1995, 2003). This approach facilitates detection and phylogenetic characterization of uncultivated microorganisms.

Dinitrogen (N_2) is the most abundant gas in the atmosphere, but is not bioavailable unless it is reduced to ammonia by N_2 fixation. Biological N_2 fixation can require the protein products of up to 20 *nif* genes. The *nifH*, *nifD* and *nifK* gene products are the structural components of the conventional molybdenum (Mo)-containing nitrogenase enzyme (Igarashi, 2003). The products of other *nif* genes are involved in roles such as regulation and biosynthesis (Rubio and Ludden, 2008). In environmental studies, *nifH*, which encodes the iron (Fe) protein of molybdenum (Mo), vanadium (V) and iron (Fe) nitrogenases, is the most commonly used nitrogenase gene for the investigation of microorganisms with the potential to fix N_2, as it is the most highly conserved in sequence (Young, 1992). This gene can be used to examine the diversity of N_2-fixing microorganisms in the environment, provides insight into the evolution and ecology of N_2 fixation and can indicate the potential for N_2 fixation in microbial communities (Lovell *et al.*, 2001; Zehr and Capone, 1996).

The application of degenerate *nifH* PCR primers (Zehr and McReynolds, 1989) enabled the discovery of novel *nifH* sequences in the environment. This approach fueled studies of N_2 fixation across a broad range of habitats, including woody dicotyledenous plants (Simonet *et al.*, 1991), rice roots (Ueda *et al.*, 1995), termite guts (Ohkuma *et al.*, 1996), stromatolites (Steppe *et al.*, 1996), central ocean gyres (Zehr *et al.*, 1998) and salt marshes (Lovell *et al.*, 2001). The size of phylogenetic trees based on all available nucleotide and amino acid sequences expanded from 19 sequences in 1994 (Chien and Zinder, 1994) to ~100 sequences in 1997 (Zehr *et al.*, 1997) and to ~1500 sequences in 2003 (Zehr *et al.*, 2003). Four (or five, depending on author) major phylogenetic clusters have been described (Chien and Zinder, 1994; Gaby and Buckley, 2011; Raymond *et al.*, 2004; Zehr *et al.*, 2003). Only three of these phylogenetically related clusters contain true nitrogenase-encoding *nif* genes. Cluster I *nifH* primarily comprises 'conventional' *nifH*, which encodes the Fe protein of Mo nitrogenase (Igarashi, 2003), as well as *vnfH* genes that encode the Fe protein of V nitrogenase (note that vanadium nitrogenase genes cluster differently based on *nifD* or *nifK* genes; Raymond *et al.*, 2004). Organisms that contain Cluster 1 *nifH* genes include cyanobacteria and alpha-, beta- and gamma-proteobacteria. Cluster II *nifH* genes encode the Fe protein of 'alternative' nitrogenases that contain iron but do not contain Mo or V (Joerger *et al.*, 1989; Lehman and

*To whom correspondence should be addressed.

Roberts, 1991). Cluster III is dominated by genes encoding Fe proteins of nitrogenases primarily of anaerobes, including methanogens and sulfate reducers; these nitrogenases likely contain Mo. Clusters IV and V (sometimes grouped as Cluster IV) contain *nifH* paralogues whose functions include photopigment biosynthesis (Young, 2005) and non-N_2-fixation electron transport (Raymond *et al*., 2004). It has also been suggested that the function of Cluster IV *nifH* paralogues found in non-N_2-fixing Archaea is the biosynthesis of cofactor F430, essential to the production of methane (Boyd *et al*., 2011; Staples *et al*., 2007). The gene name *nflH*, for '*nifH*-like', has been proposed for these *nifH* paralogues (Staples *et al*., 2007).

As sequences are accumulating rapidly in gene sequence databases, ongoing efforts to retrieve and analyze new *nifH* records are necessary to elucidate relationships between phylogenetic categories and identify phylotypes in different ecosystems. These efforts can be frustrated by the sheer number and growth rate of *nifH* gene sequences deposited into the National Center for Biotechnology Information (NCBI) GenBank database (Benson, 2004). For example, in February 2009, Gaby and Buckley identified *nifH* sequences from the non-redundant nucleotide collection database (nr/nt) at GenBank for a global census of nitrogenase diversity (Gaby and Buckley, 2011) and records were manually curated to form a database of ~17 000 sequences. We estimate that in the intervening time between the download and publication of the related article in 2011, at least 10 000 *nifH* sequences were added to the database. We estimate that as of January 2012 there were >32 000 *nifH* sequences in that database, with a growth rate of >300 sequences per month. Retrieval involving human intervention now requires a significant investment of manpower that will increase over time; an automated retrieval pipeline is needed to allow analysis to keep pace with data collection. However, automating retrieval of all sequences of any specific gene from GenBank is difficult. Moreover, the most common query idioms for searching GenBank are variants of BLAST (Altschul *et al*., 1990), which search for sequence similarity rather than function. Approaches based on BLAST alone are likely to be overly sensitive, as hits to *nifH* homologues with functions other than N_2 fixation are not easily distinguished from genuine *nifH* hits. Text searches that analyze sequence annotation can be misled by misannotation (Tripp *et al*., 2011) or misspellings in the annotation fields; a text search for 'nifh' in the nr protein database found only 6173 sequences, of which we believe 527 are not actually *nifH*. The Fungene database (http://fungene.cme.msu.edu) provides a repository of collections of functional genes from GenBank, classified by hidden Markov models (Eddy, 1998; Krogh *et al*., 1994); however, updates from GenBank are infrequent and the hidden Markov model approach is prone to false calls. An alternative *nifH* database available to the public, maintained at Cornell (Gaby and Buckley, 2011, 2014), requires the manual retrieval and curation of GenBank *nifH* sequences.

To resolve these issues, a software pipeline was developed, called ARBitrator, that requires little human intervention and retrieves up-to-date *nifH* sequence collections within a few hours. The software is adaptable to collecting sequences for genes other than *nifH*, and is especially helpful for discriminating genes of interest from their paralogues, as it incorporates an auto-curation feature based on best Reversed PSI-BLAST hits to GenBank's Conserved Domain Database. ARBitrator's output is formatted for input into other programs, such as the ARB phylogenetic software environment (Ludwig *et al*., 2004). Supplementary Appendix A provides a procedure for incrementally updating an existing *nifH* ARB database using the output of ARBitrator, thus facilitating maintenance of a comprehensive updated gene database.

2 SYSTEM AND METHODS

2.1 Design criteria

To retrieve sequences for a single gene (in this case the *nifH* gene) from a large public database, and to facilitate maintenance of a sequence database for that gene, a pipeline needs to meet specific design criteria. Based on experience with maintaining a public database of *nifH* sequences (http://pmc.ucsc.edu/~wwwzehr/research/database/), the following requirements were identified:

(1) The pipeline should be easy to invoke and should require no manual setup or runtime intervention. In particular, the pipeline should use the public online GenBank database and services provided by NCBI, rather than using a local copy of the database that would need to be downloaded and updated before pipeline execution.

(2) The data should not require manual curation. Given the current size of the GenBank database and the rate at which new *nifH* sequences are submitted, any step involving even trivial manual inspection of individual records would introduce excessive delays and possible errors.

(3) The pipeline should have acceptable error rates. Automated classification systems have inherent error rates that must be controlled to within acceptable tolerances. In the case of *nifH* classification, sequences with strong similarity to *nifH* but different function might be accepted on the basis of homology (false positives). Similarly, sequences with *nifH* function that have strong similarity to non-*nifH* genes might be rejected (false negatives). Both kinds of error must be minimized.

(4) Sequence identification should not rely on annotations in the database. Although annotations are useful as quality-control checks (e.g. for determining false-positive/-negative rates), classification should be based only on sequence content. A solution based on annotations could be no better than the false-annotation rate, which is unknown but may be too high for reliable classification. Moreover, any such solution would complicate the software by requiring natural language processing.

(5) The output of the pipeline should contain all metadata associated with the identified sequences, presented in a standard format. The pipeline's results should be easy to analyze. We have traditionally used ARB (Ludwig *et al*., 2004) to build phylogenetic trees of *nifH* sequences; therefore, the pipeline must produce output in EMBL format (Kanz, 2004), which ARB is able to read. All annotation information and metadata should be included in the EMBL records.

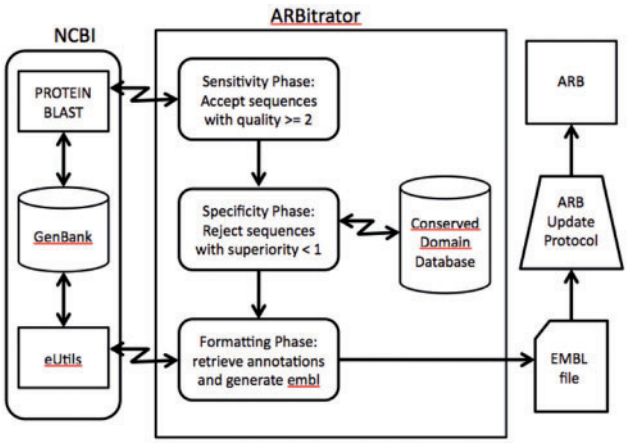

Fig. 1. ARBitrator flowchart. In the sensitivity phase, representative *nifH* sequences are BLASTed against the GenBank database at NCBI. In the specificity phase, candidate sequences retrieved by the sensitivity phase are BLASTed against a database of conserved domains. In the formatting phase, accepted sequences are output as EMBL records. After ARBitrator executes, the protocol described in Supplementary Appendix A is used to add new records to an existing ARB database

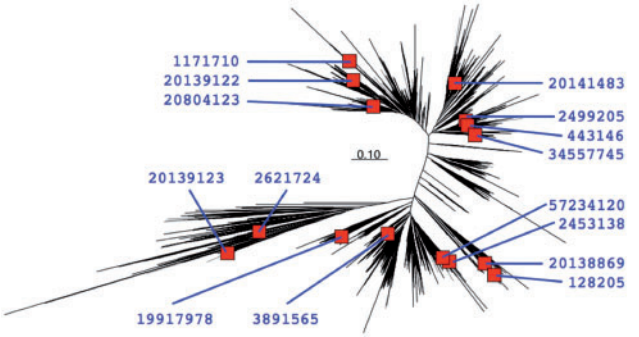

Fig. 2. Neighbor-joining tree of partial *nifH* amino acid sequences constructed using the 15 representative sequences (red boxes) and positive training set sequences. All major clusters are represented by at least one sequence. Labels are GenBank GIs

(6) The pipeline should be adaptable to the retrieval of genes other than *nifH*. To this end, code should be open source under standard licenses, and adaptation should not require a high level of programming expertise or technical support.

2.2 Algorithm

ARBitrator evaluates candidate sequences on two criteria, which we designate 'quality' and 'superiority'. Quality measures a candidate sequence's similarity to the most similar member of a repre-sentative set of 15 *nifH* sequences; superiority measures the degree to which a candidate is more similar to *nifH* than to genes coding for other known proteins. ARBitrator first executes a sensitivity phase in which it collects candidates based on quality (Fig. 1). Initial investigations showed that this phase is effective at finding *nifH* sequences; however, many non-*nifH* (false positive) sequences are also collected. In the subsequent specificity phase, false-positive candidates are eliminated based on superiority. In a final formatting phase, nucleotide sequences and annotations are retrieved for each accepted sequence, and an output file is generated in EMBL format.

2.3 Implementation

In the sensitivity phase, a set of *nifH* protein sequences is BLASTed against the nr database at GenBank using the blastp (protein query, protein subject) program. Given the large number of known *nifH* sequences (>15 000 at the inception of this project), a 'representative set' was selected consisting of 15 sequences that are evenly distributed throughout the *nifH* phylogenetic tree (Fig. 2). By only BLASTing these representative sequences, rather than all known sequences, time spent in this phase of the algorithm is reduced by three orders of magnitude. For all candidate sequences (hits) retrieved, we define 'quality' as the negative $\log_{10}$ of the *E*-value of the hit. If a subject is hit by multiple queries from the representative set, then quality is defined as the negative $\log_{10}$ of the smallest *E*-value across all hits. The sensitivity phase accepts sequences with quality ≥ 2 (i.e. all *E*-values are ≤ 0.01). The output of this phase is a collection of GIs (GenInfo Identifiers).

To support the specificity phase, a database of conserved domains was constructed based on the total set of GenBank Conserved Domains in the Subfamily Hierarchy for cd01983, Fer4_Nifh. Candidate GIs that meet the sensitivity criteria are Reversed PSI-BLASTed against this database. The Reversed PSI-BLAST (Reversed Position-Specific Iterated BLAST) algorithm uses position-specific scoring (Marchler-Bauer *et al.*, 2001, 2011), and is thus appropriate for conserved domain analysis, as differences within a conserved domain are penalized more heavily than differences in non-conserved regions. The three best-scoring hits for each candidate are analyzed. Any hits to the cd02117 (NifH_like) conserved domain are discarded as uninformative; candidate sequences are accepted if the best remaining hit is to the cd02040 (NifH) conserved domain, and the *E*-value of this hit is at least 10× smaller than the *E*-value of the next best hit. We define a sequence's *superiority* as $\log_{10}$ (*E*-value of best non-cd02040 hit) $-\log_{10}$ (*E*-value of cd02040 hit). Thus, candidates are accepted if their superiority is ≥ 1. The output of the sensitivity phase is a subset of the protein GIs generated by the sensitivity phase, representing sequences that ARBitrator classifies as *nifH*. For analysis requiring only protein GIs, the pipeline may optionally be terminated at this point.

The formatting phase retrieves the nucleotide coding sequence and annotations associated with each protein GI, and builds a record in EMBL format (Kulikova *et al.*, 2007). The GenBank protein record for each GI is retrieved from NCBI via the eUtils utility (http://www.ncbi.nlm.nih.gov/books/NBK25497/pdf/chapter2.pdf). This record is scanned for the 'coded_by' subfield of the 'CDS' (coding sequence) field to extract the identifier of a nucleotide sequence—typically a complete genome—along with coordinates for the start and end of the protein sequence. A second eUtils query retrieves the nucleotide record, from which the *nifH* nucleotide sequence is extracted. The public-domain ReadSeq utility (http://iubio.bio.indiana.edu/soft/molbio/readseq/java/) converts this record to EMBL format. Lastly, for compatibility with ARB's input filtering, multi-line literal quote field

values in the EMBL record are concatenated into single monolithic lines.

After execution of the pipeline, new records can be incorporated into an existing ARB database using a protocol described in Supplementary Appendix A. Briefly, new records are imported into ARB using a custom import filter. All protein sequences are then exported and aligned to the NifH/FrxC protein family using a hidden Markov model. The aligned sequences are then imported into ARB. Nucleotide sequences are backaligned to the protein alignment using the ARB backalign tool. The nucleotide and amino acid sequences can then be used for phylogenetic analysis and probe design.

2.4 Tuning

The software's classification algorithm is fine-tuned by three parameters: the representative set of verified *nifH* sequences used as the BLAST queries of the sensitivity phase, and the threshold values for quality and superiority. The representative set was chosen to broadly represent the known clades in the *nifH* phylogenetic tree. Members of the set are shown in Figure 2.

To determine threshold values for quality and superiority, positive and negative training sets were created. The positive training set contains the 15 513 *nifH* sequences in a manually curated *nifH* database (http://pmc.ucsc.edu/~wwwzehr/research/database/) as of September 1, 2010. The negative training set contains 766 sequences that are believed not to be *nifH* but that are moderately similar to *nifH*. Optimal quality and superiority thresholds were computed by exhaustive search of quality–superiority space. When quality threshold = 2 and superiority threshold = 1, miscalls are minimized in both training sets. These thresholds were verified by N-fold cross-validation.

2.5 Error rates

To determine false-positive and false-negative rates, ARBitrator was executed immediately after the November 2012 run reported herein, with the quality threshold relaxed to zero and the superiority threshold relaxed to –10. This generated a sample set consisting of sequences that were accepted by ARBitrator as *nifH*, as well as sequences that were rejected by a small margin. Sequences were aligned to a Pfam-curated multiple alignment of the NifH/frxC family (Fer4_NifH; PF00142) using the HMMalign module of the HMMer software package (Finn *et al*., 2011). Sequences that were too short for reliable alignment were omitted from the error-rate computation. To classify 'accept' sequences, a neighbor-joining phylogenetic tree (Saitou and Nei, 1987) was constructed using sequences that had amino acid residues in the region most widely PCR amplified by *nifH* primers (Zehr *et.al.* 2003), as the majority of the sequences submitted to GenBank are from PCR amplification. Sequences with short branches within Clusters I through III were classified as *nifH*; sequences with short branches within Cluster IV (which contains *nifH* homologues and non-functional genes) could not be classified with confidence, and were omitted from the error-rate computation; sequences with long branches within Cluster IV were classified as not *nifH*. ARBitrator's false-positive rate was computed as the number of 'accept' sequences classified phylogenetically as not *nifH*, divided by the total number of 'accept' sequences that could be aligned. A false-positive rate for sequences reported as *nifH* by Fungene was computed by the same method. To classify 'reject' sequences, any sequence that aligned poorly with the NifH/frxC family model was classified as not *nifH*. Those that aligned well but did not have amino acid residues in the region most widely amplified by *nifH* primers were omitted from the error-rate computation, as their phylogenetic analysis would not be reliable. A neighbor-joining tree was constructed using the remaining 'rejects' (i.e. those whose alignments permitted phylogenetic analysis), and the *nifH* representative set sequences. Classification was performed as described above, except that sequences with short branches to Cluster IV were conservatively classified as *nifH* to determine an upper bound for the false-positive rate.

To assess the reliability of annotations among *nifH* and similar sequences, annotated gene function was retrieved for all 'accept' and 'reject' sequences. Records whose annotated function was *nifH* or synonymous to *nifH* ('dinitrogenase reductase', 'nitrogenase Fe protein' and 138 others) were classified as annotated as *nifH*. Phylogeny-based misannotation rates were computed using the approach described to determine error rates. Sequences whose annotation contradicted phylogenetic classification were blasted against the GenBank nr database to assess the likelihood of misannotation.

2.6 Extension beyond *nifH*

To test ARBitrator's effectiveness on genes other than *nifH*, the pipeline was configured to retrieve sequences of the *nifD* gene (that encodes the alpha subunit of the Mo, V and Fe nitrogenase proteins). The representative sequence set consisted of the *nifD* sequences of the organisms that contributed to the *nifH* representative set, as well as three sequences of *vnfD* (the vanadium-using form of *nifD*). A positive training set was built by running the pipeline with the *nifD* representative and the quality/superiority settings used for *nifH*, and hand-selecting 73 sequences from the results; 43 additional sequences were added from six published studies (Dedysh, 2004; Fani *et al*., 2000; Henson *et al*., 2004; Holmes, 2004; Parker *et al*., 2002; Rodríguez-Echeverría, 2010). A negative training set was built by selecting 82 sequences of five genes that are known to have high similarity to *nifD*: *nifE*, *nifH*, *nifK*, *nifN* and protochlorophyllide reductase. Optimal quality and superiority thresholds were computed by exhaustive search of quality–superiority space, and the pipeline was executed with quality = 8.1 and superiority = 0.1. As with the *nifH* configuration, the thresholds were then relaxed to generate a sample set for computation of true- and false-negative rates.

3 RESULTS

3.1 Sequences Retrieved on November 20, 2012

On November 20, 2012, ARBitrator returned 34 420 *nifH* sequences from GenBank, of which 1757 were new to GenBank since the previous ARBitrator run on July 11, 2012. The list of protein sequence GIs may be downloaded from http://pmc.ucsc.edu/~wwwzehr/research/database/. Figure 3 shows a phylogenetic tree of representatives of the 34 420 sequences.

Figure 4 shows the distribution by quality and superiority of all sequences in the sample set. Linear regression analysis of the accepted sequences (upper-right quadrant) shows a linear

relationship with superiority = 1.3 + .45 x quality, and a coefficient of determination of 0.91. Most sequences accepted by ARBitrator cluster around the point (quality = 63, superiority = 28). Intriguingly, there is a cluster of rejected sequences around the point (quality = 2, superiority = –5); these are predominantly annotated as septum-site determining proteins or cobyrinic acid a,c-diamide synthase (*cobB*).

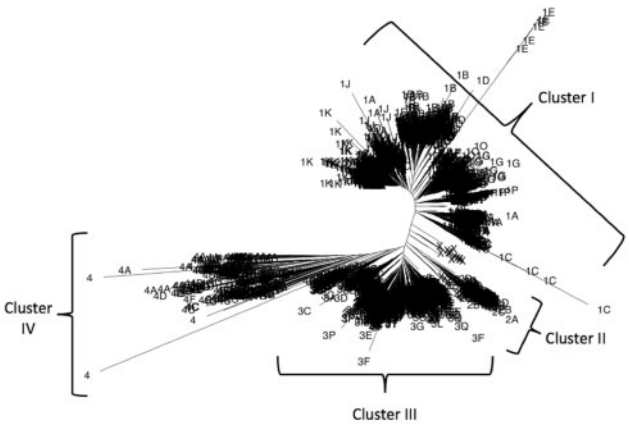

Fig. 3. Neighbor-joining tree of partial *nifH* amino acid sequences representing the 34 420 *nifH* sequences acquired in the November 20, 2012, pipeline run. Sequences were clustered at 98% amino acid identity using CD-HIT (Huang *et al*., 2010). Clusters and subclusters are marked

3.2 Error rates

Of 34 420 sequences in the sample set that are called *nifH* by ARBitrator (upper-right quadrant in Fig. 4), 2208 are too short for phylogenetic analysis and are omitted from the error-rate computation. 30 051 cluster with short branches with Clusters I, II and III, and are confirmed as *nifH*. 2151 sequences that cluster with short branches within Cluster IV cannot be confidently classified as *nifH* or not-*nifH*, and are omitted from the computation. Ten sequences associate via long branches with Cluster IV and are classified as not *nifH*. Thus, the phylogenetically derived false-positive rate is 10/(10 + 30 051) = 0.033%. Of the 32 227 sequences in the sample that are rejected by ARBitrator (upper-left, lower-left and lower-right quadrants in Fig. 4), 2067 are too short for phylogenetic analysis and are omitted from the error-rate computation. 28 846 align poorly with the NifH/frxC family and are confirmed as not *nifH*. 1134 cluster with long branches within Cluster IV, and are also confirmed as not *nifH*. Eight associate via short branches with Cluster IV and are omitted from the error-rate computation. Thus, the phylogenetic analysis detected no false-negative errors. See Supplementary Figure S1 for phylogenetic trees of accepted and rejected sequences.

104 sequences that ARBitrator accepts are annotated as not *nifH*; 87 of these are confirmed by phylogeny as *nifH* (i.e. phylogeny confirms ARBitrator's call), 10 are classified by phylogeny as not *nifH* and 10 cannot be phylogenetically classified. A total of 88 sequences in the sample set that ARBitrator rejects are annotated as *nifH*; 82 of these are classified by phylogeny as not *nifH* (i.e. phylogeny confirms ARBitrator's call), 8 are

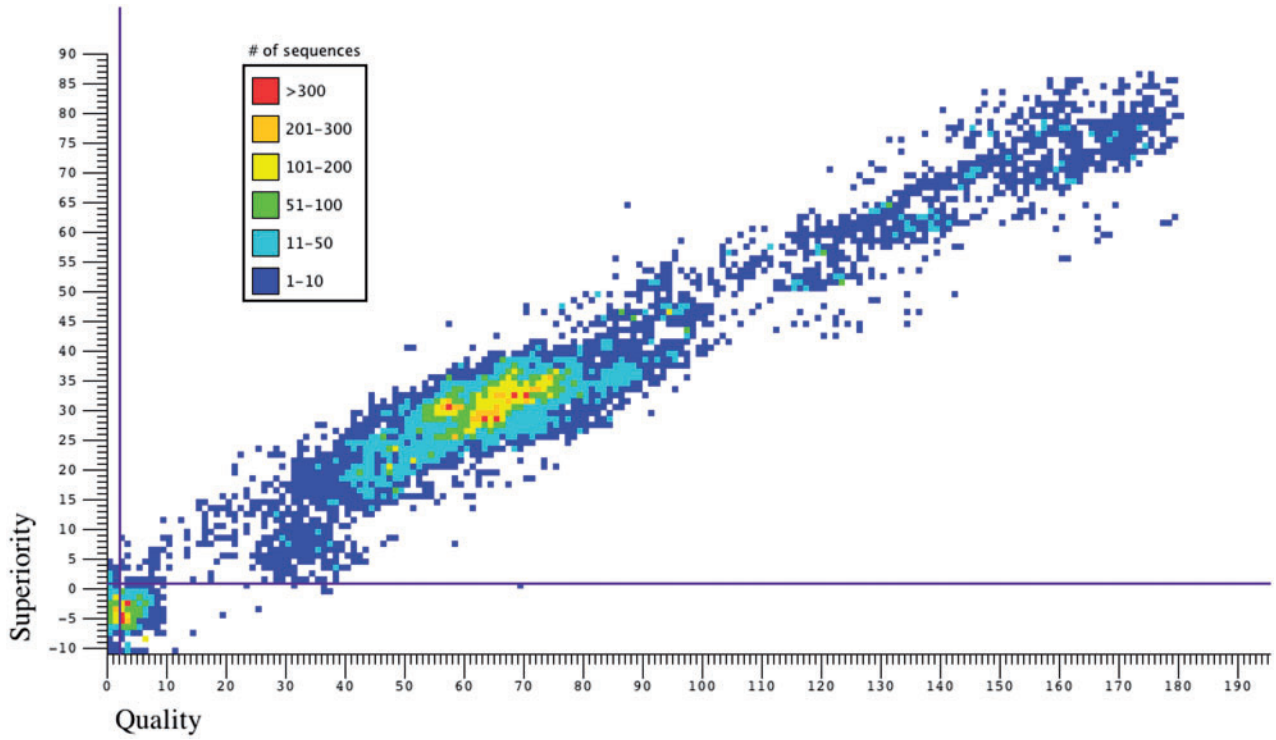

Fig. 4. Distribution of sequences by quality and superiority. Thresholds are represented by purple crosshairs; sequences classified as *nifH* by ARBitrator are above and to the right of the crosshairs

classified by phylogeny as *nifH* and 8 cannot be phylogenetically classified.

3.3 Comparison to other *nifH* databases

The sequences were compared with the 31 970 sequences in the Fungene database as of November 20, 2012. ARBitrator and Fungene had 29 836 sequences in common. ARBitrator collected 4584 sequences that Fungene rejected, and Fungene collected 2134 sequences that ARBitrator rejected; of these, ARBitrator rejected 1744 because of quality and 390 because of superiority. For the sequences reported by both ARBitrator and Fungene, the phylogenetically derived false-positive rate is 0.033%, the same as the overall ARBitrator false-positive rate. For the 2134 sequences that only Fungene reports as *nifH*, 1964 did not align with the NifH/frxC family and are classified as not *nifH*. Thus, the false-positive rate for these sequences is $\geq$1964/2134 = 92%. The overall Fungene false-positive rate is 6.2%.

The second update of the Cornell nifH database reported in Gaby and Buckley (2014)—based on a May 2012 snapshot of GenBank—contained 32 854 nt records, 28 742 of which have corresponding protein records specified by a 'db_xref' field, thus permitting comparison with ARBitrator. ARBitrator rejected 479 of these records (1.7%): 440 because of low quality, and 39 because of low superiority.

3.4 *nifD* results

The *nifD* configuration of the pipeline returned 2747 sequences, of which 2726 are annotated as *nifD*, for a false-positive rate of 0.76% conditioned on annotation accuracy. The sample set contained 8715 rejected sequences, of which 76 are annotated as *nifD*, for an annotation-conditioned false-negative rate of 0.87%. For the 2972 *nifD* sequences reported by Fungene for which unambiguous annotations could be retrieved, the annotation-conditioned false-positive rate is 48%.

4 DISCUSSION

nifH gene diversity studies, and all single-gene diversity studies, are hampered by the difficulty of collecting all sequences associated with the gene of interest. NCBI's GenBank, the database where newly discovered sequences are deposited, provides no service for selecting all records that are annotated as representing a specified gene. A direct text search of the database would be of dubious value because of misannotations. GenBank's main query idiom is the BLAST search, which selects based on sequence similarity regardless of function. Consequently, a sequence collection pipeline that simply BLASTs a representative set of query genes will accept paralogues with the wrong function. The challenge of supporting diversity studies is exacerbated by the rapid growth of GenBank: the collection of *nifH*, and presumably other genes, appears to have been growing exponentially for the past several years. Thus, there exists an opportunity to facilitate diversity studies by increasing the efficiency of sequence retrieval.

4.1 Necessity for both quality and superiority criteria

As Figure 4 shows, sequences obtained and accepted as *nifH* by ARBitrator (above and to the right of the purple crosshairs) cluster around the point (quality = 63, superiority = 28). When the quality and superiority criteria are relaxed (to the left of and/or below the crosshairs), a second cluster appears around quality = 0.1, superiority = –5; sequences in this second cluster are predominantly annotated as *MinD* (membrane ATPase of the MinC-MinD-MinE system) or *CobB* (cobyrinic acid a,c-diamide synthase, involved in the biosynthesis of vitamin B12). There is an approximately linear relationship between quality and superiority (superiority = 1.3 + .45 quality), with $r^2 = 0.91$. This relationship is not strong enough to allow either quality or superiority alone to be used as a selection criterion. For example, without the quality criterion, all sequences to the right of the vertical crosshair would be accepted, including many sequences from the *MinD*/cobyrinic acid peak. Similarly, without the superiority criterion, all sequences below the horizontal crosshair would be accepted.

4.2 Error rates

ARBitrator's low error rates (0.033% false positives, no detectable false negatives) can be understood in light of the underlying similarities between the ARBitrator algorithm and the error-rate analysis. In both approaches, candidate sequences are aligned against known *nifH* sequences. With ARBitrator, the candidate sequences are the contents of the GenBank protein database, and the known *nifH* sequences are the 15 representative sequences; during the sensitivity phase, the BLAST step aligns each query (known *nifH* representative) against each database member (candidate). Candidates are provisionally accepted if they align well enough with the representatives, with the specificity phase providing necessary additional accuracy to the measurement of alignment quality. In the error-rate analysis, the candidate sequences are the members of the ARBitrator sample set, which are aligned against the known *nifH* members of the Fer4_NifH Pfam family. The alignment algorithms in the two approaches are not identical in all implementation details, but in both cases sequences are accepted if and only if they are similar to known *nifH* sequences, with similarity measured by alignment score.

When a sequence's annotation contradicts gene its ARBitrator classification (i.e. ARBitrator accepts a sequence that is not annotated as *nifH*, or rejects a sequence that is annotated as *nifH*), phylogenetic analysis supports the ARBitrator call in 90% of cases. In all, 104 sequences are classified by ARBitrator as *nifH* but annotated as not *nifH*. When these were BLASTed against the nr database, and the top 20 non-self hits for each query were inspected, 48 queries hit exclusively to *nifH* subjects or to non-*nifH* subjects from the same study as the query. We propose that these subject sequences, which come from two studies that submitted multiple sequences to GenBank, were systematically misannotated by the researchers and should have been annotated as *nifH*. Similarly, ARBitrator rejects 88 sequences that are annotated as *nifH*. When these were blasted against the nr database and the top 20 hits of each query were inspected, it was found that 54 of the queries hit exclusively to non-*nifH* subjects, or to *nifH* subjects from the same study as the query. These subject sequences come from five studies that submitted multiple sequences to GenBank,

and we propose that these sequences were systematically misannotated as *nifH*.

4.3 Comparison with other *nifH* databases

The ARBitrator and Fungene databases as of November 20, 2012, had 29 836 *nifH* sequences in common. ARBitrator accepted 4584 sequences that Fungene rejected. The false-positive rate for these sequences is approximately the same as for the overall sample set (0.03%). Fungene accepted 2134 sequences that ARBitrator rejected; the false-positive rate for these sequences is 90%. This discrepancy can perhaps be attributed to the fundamental difference between ARBitrator's BLAST-based algorithm and the hidden Markov model that underlies the Fungene pipeline. ARBitrator classifies according to similarity to representative *nifH* sequences and to the NifH conserved domain; Fungene classifies according to similarity to a composite profile model.

ARBitrator's results are substantially in agreement with the latest update of the Buckley Laboratory database, which is not generated by auto-curating software and requires manual processing (Gaby and Buckley, 2014).

4.4 Extension beyond *nifH*

The *nifD* pipeline configuration, despite being based on cruder positive and negative training sets than the *nifH* configuration, nevertheless produced annotation-based error rates that were <1%. This result supports the applicability of the ARBitrator algorithm to other genes besides *nifH*.

The discrepancy between the superiority thresholds for *nifH* (1.0) and *nifD* (0.1) can be explained by the evolutionary history of *nifD*, which apparently originated as an ancestral gene that underwent a duplication event, giving rise to an ancestral bicistronic operon that later duplicated again to produce the present-day *nifD*, *nifE*, *nifK* and *nifN* genes (Fani *et al.*, 2000). Thus, many *nifD* genes have low superiority because their similarity to the *nifE*, *nifK* and *nifN* conserved domains is only slightly worse than their similarity to the *nifD* conserved domain. A higher superiority threshold would reject such sequences and increase the false-negative rate. The similarity of *nifD* to other genes makes it a particularly rigorous test of the extension of the ARBitrator approach.

In conclusion, by combining a quality-based sensitivity phase with a superiority-based specificity phase, we have been able to implement a pipeline that meets all design criteria. Records provide nucleotide and amino acid sequences, as well as complete annotations. Computed false-positive and false-negative rates are acceptably low, and actual rates may be even lower. Results from the November 20, 2012, run are generally in good agreement with the *nifH* sequence collection at Fungene; however, the ARBitrator results are more extensive, have a lower error rate, and can be updated whenever a user wishes to rerun the pipeline. ARBitrator is designed to support ongoing *nifH* phylogeny research into the future as the GenBank collection continues to grow exponentially. If error rates increase, adjustments can be made to the quality and superiority thresholds. If NifH/FrxC family and CobB sequences continue to be major contributors to the false-positive rate, the software may need to be adapted to detect and reject these special cases.

In addition to its original application to *nifH* phylogeny, ARBitrator can be adapted to other genes, as evidenced by the *nifD* results. An immediate use for adapted versions of the pipeline would be the collection *nif* genes other than *nifH* and *nifD*. Comparison of phylogenies of multiple *nif* genes could provide new insight into N_2 fixation diversity and ecology.

ACKNOWLEDGEMENTS

The authors are grateful to Deniz Bombar for his help in organizing this article.

Funding: This work was supported by NSF grant EF0424599 for the Center for Microbial Oceanography: Research and Education (CMORE), a Gordon and Betty Moore Marine Investigator grant (J.Z.) and the Microbial Environmental Genomics Applications: Modeling, Experimentation, and Remote Sensing (MEGAMER) facility (J.Z.) funded by the Gordon and Betty Moore Foundation.

Conflict of Interest: None declared.

REFERENCES

Altschul,S. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Benson,D.A. (2004) GenBank: update. *Nucleic Acids Res.*, **32**, 23D–26D.

Boyd,E.S. *et al.* (2011) An alternative path for the evolution of biological nitrogen fixation. *Front. Microbiol.*, **2**, 205.

Chien,Y. and Zinder,S. (1994) Cloning, DNA sequencing, and characterization of a nifD-homologous gene from the archaeon Methanosarcina barkeri 227 which resembles nifD1 from the eubacterium Clostridium pasteurianum. *J. Bacteriol.*, **176**, 6590–6598.

Dedysh,S.N. (2004) NifH and NifD phylogenies: an evolutionary basis for understanding nitrogen fixation capabilities of methanotrophic bacteria. *Microbiology*, **150**, 1301–1313.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinform. Rev.*, **9**, 755–763.

Fani,R. *et al.* (2000) Molecular evolution of nitrogen fixation: the evolutionary history of the nifD, nifK, nifE, and nifN genes. *J. Mol. Evol.*, **51**, 1–11.

Finn,R.D. *et al.* (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*, **39**, W29–W37. doi:10.1093/nar/gkr367.

Gaby,J.C. and Buckley,D.H. (2011) A global census of nitrogenase diversity. *Environ. Microbiol.*, **13**, 1790–1799.

Gaby,J.C. and Buckley,D.H. (2014) A comprehensive aligned nifH gene database: a multipurpose tool for studies of nitrogen-fixing bacteria. *Database*, **2014**, bau001.

Henson,B.J. *et al.* (2004) Molecular phylogeny of the heterocystous cyanobacteria (subsections IV and V) based on nifD. *Int. J. Syst. Evol. Microbiol.*, **54**, 493–497.

Holmes,D.E. (2004) Comparison of 16S rRNA, *nifD*, *recA*, *gyrB*, *rpoB* and *fusA* genes within the family Geobacteraceae fam. nov. *Int. J. Syst. Evol. Microbiol.*, **54**, 1591–1599.

Huang,Y. *et al.* (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.

Igarashi,R.Y. (2003) Nitrogen fixation: the mechanism of the Mo-dependent nitrogenase. *Critical Rev. Biochem. Mol. Biol.*, **38**, 351–384.

Joerger,R.D. *et al.* (1989) Two nifA-like genes required for expression of alternative nitrogenases by Azotobacter vinelandii. *J. Bacteriol.*, **171**, 3258–3267.

Kanz,C. (2004) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **33**, D29–D33.

Krogh,A. *et al.* (1994) Protein modeling using hidden Markov models. *J. Mol. Biol.*, **235**, 1501–1531.

Kulikova,T. *et al.* (2007) EMBL nucleotide sequence database in 2006. *Nucleic Acids Res.*, **35**, D16–D20.

Lehman,L.J. and Roberts,G.P. (1991) Identification of an alternative nitrogenase system in Rhodospirillum rubrum. *J. Bacteriol.*, **173**, 5705–5711.

Lovell,C.R. *et al.* (2001) Recovery and phylogenetic analysis of nifh sequences from diazotrophic bacteria associated with dead aboveground biomass of spartina alterniflora. *Appl. Environ. Microbiol.*, **67**, 5308–5314.

Ludwig,W. *et al.* (2004) ARB: a software environment for sequence data. *Nucleic Acids Res.*, **32**, 1363–1371.

Marchler-Bauer,A. *et al.* (2001) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.*, **30**, 1–3.

Marchler-Bauer,A. *et al.* (2011) CDD: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Res.*, **39**, D225–D229.

Ohkuma,M. *et al.* (1996) Diversity of nitrogen fixation genes in the symbiotic intestinal microflora of the termite Reticulitermes speratus. *Appl. Environ. Microbiol.*, **62**, 2747–2752.

Parker,M.A. *et al.* (2002) Conflicting phylogeographic patterns in rRNA and nifD indicate regionally restricted gene transfer in Bradyrhizobium. *Microbiology*, **148** (Pt. 8), 2557–2565.

Raymond,J. *et al.* (2004) The natural history of nitrogen fixation. *Mol. Biol. Evol.*, **21**, 541–554.

Rodríguez-Echeverría,S. (2010) Rhizobial hitchhikers from Down Under: invasional meltdown in a plant-bacteria mutualism? *J. Biogeogr.*, **37**, 1611–1622.

Rubio,L.M. and Ludden,P.W. (2008) Biosynthesis of the iron-molybdenum cofactor of nitrogenase. *Annu. Rev. Microbiol.*, **62**, 93–111.

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.

Simonet,P. *et al.* (1991) Frankia genus-specific characterization by polymerase chain reaction. *Appl. Environ. Microbiol.*, **57**, 3278–3286.

Staples,C.R. *et al.* (2007) Expression and association of group IV nitrogenase NifD and NifH homologs in the non-nitrogen-fixing archaeon *Methanocaldococcus jannaschii*. *J. Bacteriol.*, **189**, 7392.

Steppe,T. *et al.* (1996) Consortial N_2 fixation: a strategy for meeting nitrogen requirements of marine and terrestrial cyanobacterial mats. *FEMS Microbiol. Ecol.*, **21**, 149–156.

Tripp,H.J. *et al.* (2011) Misannotations of rRNA can now generate 90% false positive protein matches in metatranscriptomic studies. *Nucleic Acids Res.*, **39**, 8792–8802.

Ueda,T. *et al.* (1995) Remarkable N2-fixing bacterial diversity detected in rice roots by molecular evolutionary analysis of nifH gene sequences. *J. Bacteriol.*, **177**, 1414–1417.

Young,J.P.W. (1992) Phylogenetic classification of nitrogen-fixing organisms. In: Stacey,G. *et al.* (eds) *Biological nitrogen fixation.* Chapman and Hall, New York, NY, pp. 43–86.

Young,J. (2005) The phylogeny and evolution of nitrogenases. In: Palacios,R. and Newton,W.E. (eds) *Genomes and Genomics of Nitrogen-Fixing Organisms.* Netherlands, Springer, pp. 221–241.

Zehr,J.P. and McReynolds,L.A. (1989) Use of degenerate oligonucleotides for amplification of the nifH gene from the marine cyanobacterium Trichodesmium thiebautii. *Appl. Environ. Microbiol.*, **55**, 2522–2526.

Zehr,J. and Capone,D. (1996) Problems and Promises of Assaying the Genetic Potential for Nitrogen Fixation in the Marine Environment. *Microb. Ecol.*, **32**, 263–281.

Zehr,J. *et al.* (1995) Diversity of heterotrophic nitrogen fixation genes in a marine cyanobacterial mat. *Appl. Environ. Microbiol.*, **61**, 2527–2532.

Zehr,J.P. *et al.* (1997) Phylogeny of cyanobacterial nifH genes: evolutionary implications and potential applications to natural assemblages. *Microbiology*, **143**, 1443–1450.

Zehr,J. *et al.* (1998) New nitrogen-fixing microorganisms detected in oligotrophic oceans by amplification of nitrogenase (nifH) genes. *Appl. Environ. Microbiol.*, **64**, 3444–3450.

Zehr,J. *et al.* (2003) Nitrogenase gene diversity and microbial community structure: a cross-system comparison. *Environ. Microbiol.*, **5**, 539–554.