Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2011, Article ID 495849, 11 pages
doi:10.1155/2011/495849

Research Article

Discovering the Unknown: Improving Detection of Novel Species and Genera from Short Reads

**Gail L. Rosen,[1] Robi Polikar,[2] Diamantino A. Caseiro,[3]
Steven D. Essinger,[1] and Bahrad A. Sokhansanj[4]**

[1] *Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA 19104, USA*
[2] *Department of Electrical and Computer Engineering, Rowan University, Glassboro, NJ 08028, USA*
[3] *Spoken Language Systems Laboratory, Instituto Superior Técnico, 1049-001 Lisbon, Portugal*
[4] *School of Biomedical Engineering, Science, and Health Systems, Drexel University, Philadelphia, PA 19104, USA*

Correspondence should be addressed to Gail L. Rosen, gailr@ece.drexel.edu

Received 20 September 2010; Revised 23 December 2010; Accepted 25 January 2011

Academic Editor: Chiu-Chung Young

High-throughput sequencing technologies enable metagenome profiling, simultaneous sequencing of multiple microbial species present within an environmental sample. Since metagenomic data includes sequence fragments ("reads") from organisms that are absent from any database, new algorithms must be developed for the identification and annotation of novel sequence fragments. Homology-based techniques have been modified to detect novel species and genera, but, composition-based methods, have not been adapted. We develop a detection technique that can discriminate between "known" and "unknown" taxa, which can be used with composition-based methods, as well as a hybrid method. Unlike previous studies, we rigorously evaluate all algorithms for their ability to detect novel taxa. First, we show that the integration of a detector with a composition-based method performs significantly better than homology-based methods for the detection of novel species and genera, with best performance at finer taxonomic resolutions. Most importantly, we evaluate all the algorithms by introducing an "unknown" class and show that the modified version of PhymmBL has similar or better overall classification performance than the other modified algorithms, especially for the species-level and ultrashort reads. Finally, we evaluate the performance of several algorithms on a real acid mine drainage dataset.

1. Introduction

Mass amounts of high-throughput sequenced DNA are being produced as a result of metagenomics projects, and new tools are needed to identify the taxonomic content of these environmental samples. Currently, biologists have two main goals: (1) classify as many organisms as possible, and (2) assess the genes and functions within the sample [1]. This is especially difficult when the sample contains many uncultivated organisms that have no known reference genome. In addition, the reads obtained from these samples can have short lengths from next-generation technologies, complicating the identification process. Such technologies are pivotal in order to sequence as much DNA as possible within such a sample in a timely fashion but have a "size" (associated with accuracy) tradeoff.

Currently, taxonomic identification is plagued by several obstacles. Typically, researchers classify metagenomic reads, or sequenced DNA reads, by scoring their alignment to previously sequenced organisms using BLAST [2–4]. Unfortunately, only a thousand out of millions of possible species have their genomes fully sequenced. This under-representation has severely restricted the development of an automated system that recognizes sequences from taxa without completely sequenced genomes [5, 6]. In fact, several researchers believe that sequencing error skews our estimates of the abundance of taxa since errors can cause artificial "divergence" in reads, enough to falsely predict new operational taxonomic units (OTUs) [5, 7]. Recent papers call for ways to infer which species are truly known or unknown from metagenomics samples due to this unwanted variation [5, 7]. Huson et al. show that anywhere between

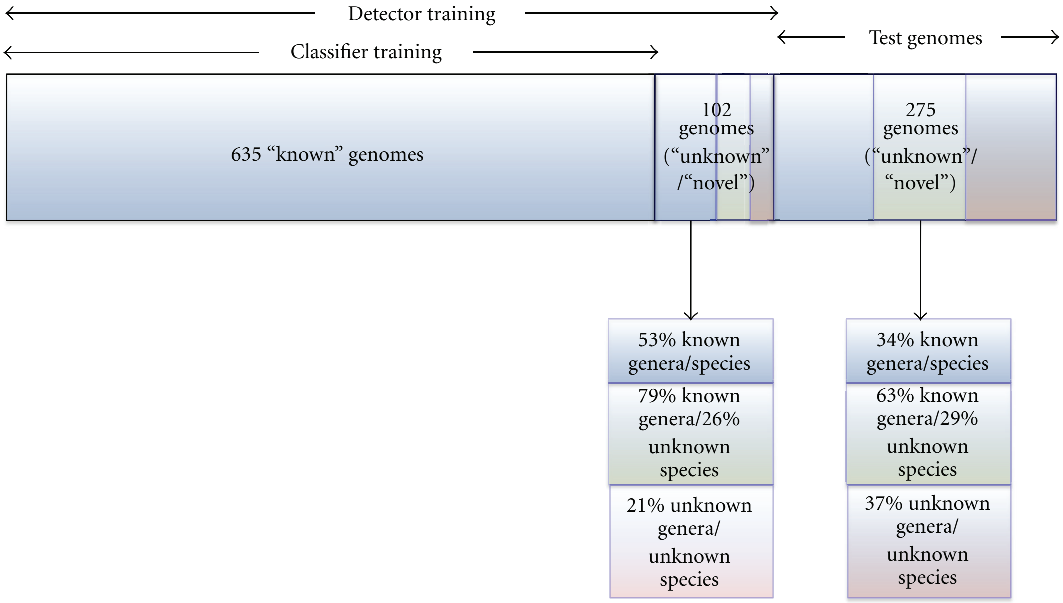
Detector training
Classifier training
Test genomes
635 "known" genomes
102 genomes ("unknown" /"novel")
275 genomes ("unknown"/ "novel")
53% known genera/species
79% known genera/26% unknown species
21% unknown genera/ unknown species
34% known genera/species
63% known genera/29% unknown species
37% unknown genera/ unknown species

Figure 1: The datasets used in this paper are composed of a classifier training set/database, novel genomes used to train the detector, and a separate novel-genome test set. The *blue* areas represent the percentage of genomes that have "known" genera/species; the *green* areas represent the percentage of genomes that are "known" at the genus level but "unknown" on the species level; the *red* areas represent the percentage of genomes that are "unknown" at both the species and genus levels.

10% and 90% of all reads may fail to produce any hits to known genomes when analyzed with BLAST, and they develop a last common ancestor (LCA) algorithm to assign reads to the most confident taxon (class) [3]. In [8], Brady and Salzberg also state that they "investigate" confidence scores for predicting the correctness of the classifier, but they are unable to solve this problem. Therefore, a detector is clearly needed for composition-based methods to accept/reject reads based on their known/unknown status. In this paper, we address this detection problem and show that we can use the likelihood scores as confidence scores and can interpolate the scores between different read lengths to obtain a consistent detector of "known" reads.

In this paper, we develop a detector for "unknown" novel genome reads for use with composition-based methods that can accept/reject reads from novel taxa anywhere in the taxonomic hierarchy. For example, a species-level confidence detector may reject a read but the genus-level detector may accept it. This would indicate that this is a new species within a known genus. To detect these novel genomes, we show that composition-based methods perform better than homology-based methods on the finer resolution of taxonomic levels. The nave Bayes classifier (NBC) and PhymmBL offer fast and attractive solutions, because they are based on the DNA's compositional word-frequency occurrence and are able to give a log-likelihood score or similar that can be used to develop such a detector [8, 9]. NBC and Phymm's ease of use, speed of training and testing, combined with its log-likelihood output (or in the case of PhymmBL, hybrid scores) make them attractive, simple, yet elegant designs for large-scale metagenomic classification and comparison.

2. Materials and Methods

2.1. Dataset for Detector Design. In order to design a detector, we partition the available data into three datasets as shown in Figure 1. The first is the "known" dataset, composed of 635 known completed microbial genomes in Genbank [10] that were available as of February 2008. This known dataset is used to train the supervised *classifier*. Then, we use a second dataset consisting of 102 "novel" genomes, those strains added to Genbank from February 2008 to August 2008. This second dataset is used to train the *detector* to determine whether a read belongs to a known or unknown species/genus. Finally, we have a third "test unknowns" dataset composed of 275 genomes, added to Genbank from August 2008 to November 2009. This dataset is then used to evaluate the detector/classifier combination. We specifically define the term "unknown" as a read's taxonomic class does not exist in the training database for the particular taxonomic level/rank being classified. For example, some reads may originate from "unknown" species/genera but still be known at the phyla level. We illustrate the "known," "unknown"/"novel" concept for each dataset in Figure 1, and describe their composition in detail in the next section. Ultimately, we show that such a detector, in conjunction with composition-based likelihood scores, is able to determine if a read originates from unknown species/genera.

In order to accomplish this goal, we start by evaluating the accuracy of NBC, BLAST, and PhymmBL on 102 novel genomes for the species and genus levels (after being trained on the 635 genome training dataset). Based on this evaluation, we then design a detector: first by using BLAST scores followed by using the NBC likelihood scores. We also design a detector for PhymmBL using the combination of its Phymm-likelihood and BLAST scores. Finally, we compare the performance of the classifier+detectors for the NBC and PhymmBL approaches on the test dataset of 275 novel genomes to the BLAST-based methods: MEGAN, SOrt-ITEMS [11], and CARMA [12]. Because homology-based methods have never been benchmarked for known/novel detection, we will benchmark these for the first time while also developing and benchmarking composition-based detectors. Finally, we demonstrate NBC's, PhymmBL's, and SOrt-ITEM's performance on an experimentally acquired acid mine drainage dataset.

The 635 microbes of the "known" dataset, from [13], belong to 470 distinct species and 260 distinct genera. 404 strains are the sole member of their species-class while 171 strains are the sole member of their genus in the dataset. This shows that some knowledge will be lacking when it comes to species- and genus-class diversity. While 66 species contain more than one strain, 89 genera contain more than one strain. The microbial strains genome lengths range from 160 K(bp) for *Candidatus Carsonella* to 13 Mil(bp) for *Sorangium cellulosum*.

In order to design the detector, another 102 strains were acquired from Genbank and labeled "novel" and not represented in the "known" training database. 54 of these 102 novel strains belong to 36 known species while the remaining 48 comprise 46 "novel" species (with respect to the "known" database). At the next taxonomic level, 81 of the strains belong to 55 "known" genera while the remaining 21 comprise 21 "novel" genera. This is a good known/unknown representation for different levels of the detector because all strains are novel (with respect to the training dataset), approximately 1/2 of the strains' species classification is unknown, and approximately 1/5 of the strains' genera classification is unknown.

2.2. Test Dataset. 275 strains were acquired in November 2009 from NCBI, which were all new, completely sequenced microbial genomes, since August 2008. The 275 genomes comprise 156 unique genera, of which 64 are in the "known" database and 92 are not, and 216 unique species, of which 48 are "known" and 168 are not. 172 (63%) of the genomes belong to the 64 known genera, and 96 (34%) of the them belong to the 48 known species. The "unknown" strains belong to a diverse set of genera and species; the $275 - 172 = 103$ (37%) "unknown" strains belong to 92 novel genera, and $175 - 96 = 179$ (66%) of the strains belong to 168 novel species. 5 strains from the test set's genera overlap with the "unknown genera" in 102 genomes used to train the detector. This means that the detector trained on "unknown" genera that is also represented in 2% of the test set. There is a concern that this overlap may have artificially raised accuracy, but the overlap affects only 2% of the sequences, so we conclude that the artificial increase in performance, if any, is negligible. Also, all classifiers have the same training advantage, so it is still a fair comparison. Finally, we note that there is no overlap at the species level. Therefore, we do not "overtrain" our detector on many examples of "unknowns" that also occur in the test set. It is possible that when designing a detector with a different dataset, some of the unknowns may exist in the "novel" training data, so this is a realistic dataset. Also, there is a good distribution of novel to known strains based on which we would expect 37% of the genomes to be rejected (i.e., declared by the detector to be unknown) at the genus level while 66% to be rejected at the species level.

2.3. Detector Development. To develop a detector, we compose a ROC (receiver operating characteristic) curve using the likelihood scores of the composition-based methods on the training dataset. Each score is associated with the binary decision of whether the genome exists in the database or not. The best operating point on the training dataset is determined as the threshold that obtains the best combined sensitivity and specificity, defined by the the maximum point of the summed sensitivity and specificity metrics. The development of the detector is summarized as follows.

(1) Acquire 635 known genomes.

(2) Train NBC/PhymmBL on the 635 genomes.

(3) Acquire 102 unknown genomes.

(4) Draw 100 L-length reads from each of the 737 full genomes (coding and noncoding), where $L = 500, 100, 25$ bp.

(5) Score the L-length reads (using NBC, PhymmBL), where the scores can be interpreted as posterior probabilities of the genomes predicted by the classifiers.

(6) Construct an ROC curve using the algorithm's scores and known/unknown labels.

(7) Determine best operating point by maximizing the sensitivity+specificity.

(8) Select score threshold corresponding to best operating point for the training data (to be subsequently used on test data).

2.4. Measures for Comparison. We define the following measures which will be used to compare the methods.

(i) Detector sensitivity = TP/(TP + FN), where TP is the number of true positives (reads from "known" taxa correctly identified) and FN is the number of false negatives (reads from "known" taxa incorrectly identified as unknown).

(ii) Detector specificity = TN/(TN + FP), where TN is the number of true negatives ("unknown" reads correctly identified as unknown), and FP is the number of false positives (unknown reads labeled as "known") number.

(iii) Detector accuracy = total correct decision/total number of reads = (TP + TN)/(TP + FN + FP + TN).

(iv) Classifier accuracy = total correctly classified/total number of reads, where correctly classified means classified correctly into its taxonomic rank.

(v) Overall classification accuracy = (total detected as known that are also correctly classified)+TN/ total number of reads, where the first term can be approximated by (TP + FP) ∗ (classifier accuracy).

2.5. Methods for Comparison. The methods, in addition to NBC, can be accessed via the web. NBC [9] is available for download and online at http://nbc.ece.drexel.edu/. PhymmBL [8] is available for download from http://www.cbcb.umd.edu/software/phymm/. SOrt-ITEMS [11] is available for download from http://metagenomics.atc.tcs.com/binning/SOrt-ITEMS/. MEGAN [3] is available for download from http://www-ab.informatik.uni-tuebingen.de/software/megan. WebCarma [12] is available online at http://webcarma.cebitec.uni-bielefeld.de/cgi-bin/webcarma.cgi.

3. Results

To compare several methods for the task of detecting novel genomes, we simulate several scenarios. In the first section, we develop a detector (BLAST, NBC, and PhymmBL) that accepts/rejects reads of unknown species and genera based on each method's score; species/genus levels are informative levels where we would expect to see larger differences between the various methods. Also, we narrow our performance comparisons to these levels since many taxa do not have all levels defined (e.g., some taxa have species-, genus-, family- and phyla-level labels but are missing family-, and class-level labels in the Genbank taxonomy database).

We only develop the detector for PhymmBL and NBC, since there are other BLAST-based detectors, such as MEGAN, CARMA, and SOrt-ITEMS that perform such a detection task. As a measure of comparison between ROC curves in all sections, we assess the area-under-the-curve (AUC) metric, a standard measure for detector performance. We then compare the performance of the NBC and PhymmBL detectors on a test set and show that they can improve the detector accuracy of the raw method and outperform BLAST-based methods. We also show the overall classification performance (binning each read into their associated bins) in addition to the unknown class.

3.1. Detector Development for BLAST, NBC, and PhymmBL (for the 635 Training Genomes Plus 102-Test Genomes). The poor classifier accuracies on novel genomes leads us to ask how well can classifiers predict "unknown" taxa. In other words, can BLAST's bit score or NBC's/PhymmBL's score be used to indicate whether a fragment is truly from a new species, genus, and so forth?

3.1.1. BLAST Bitscore for Detection of 635 Known Plus 102 Unknown Genomes. Here, we show the utility of BLAST's bitscore when accepting/rejecting known/unknown reads. In creating the receiver operating characteristic (ROC) curve for BLAST's bitscore, reads are marked as correct if they are correctly classified. In Figure 2, we show BLAST's ability to accept/reject reads from known/unknown strains for strain, species, and genus levels. We see that each of the optimal operating points is near 100% sensitivity for the three taxonomic levels while the specificities are around 20%, 35%, and 60% for strains, species, and genera, respectively.

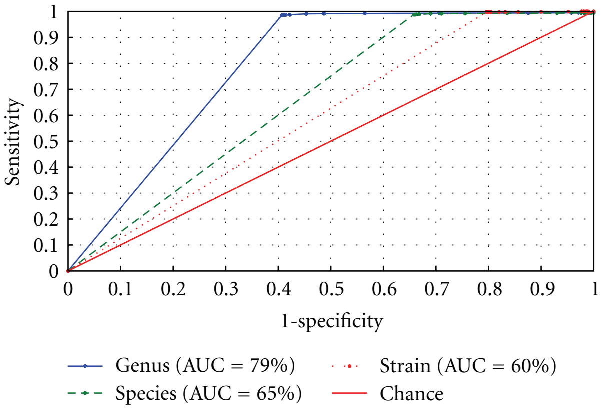

Figure 2: The ability of BLAST to discern the correct strain genome (in red-dashed), species genome (in green-dash) and the correct genus label (in blue) for the known 63, 500 (100 randomly selected reads from each of 635 known genomes) plus unknown 10, 200 (100 randomly selected reads from 102 novel genomes) 25 bp reads. The ROC curves compare BLAST's bit scores against a varying threshold. The plot demonstrates that BLAST predicts most "known" genomes correctly at the optimal operating point, but incorrectly detects "unknown" genomes. For the strain detection, the area-under-the-curve is 60.1% with the best threshold yielding a sensitivity of 99.8% and specificity of 20.4%. For the species-level detection, the AUC is 65% with 99.1% sensitivity and a specificity of 34.7%. For the genus detection, the area-under-the-curve is 78.9% with the best threshold yielding a sensitivity of 98.6% and specificity of 59.3%. The red line represents the 50% chance line.

BLAST's ability to predict taxonomically known and unknown reads using the bit score/e-value has mixed results. While BLAST is clearly a good classifier for known organisms within its database, it lacks the ability to reject (declare "unknown") novel genomes with both high sensitivity and specificity. We now investigate the feasibility of PhymmBL/NBC's score to develop a better detector for known and unknown taxa from using very short-read reads.

3.2. NBC Scores for Detection of 635 Known Plus 102 Novel Genomes. To develop a detector using the NBC likelihood scores, we vary a threshold and mark reads correct if they are in the database (a looser constraint than being correctly classified). The NBC detector's ROC curves are shown in Figure 3. The AUC for the 500 bp reads are marginally better than the 25 bp reads, and, interestingly, the genus and species levels perform the same.

3.3. PhymmBL Scores or Detection of 635 Known Plus 102 Unknown Genomes. The ROC curves are also constructed for PhymmBL for species- and genus-level performance and 500 bp and 25 bp reads. As seen in Figure 4, the ROC shapes are similar to those of BLAST's, but with better specificity. The 25 bp species-level specificity for PhymmBL is 46% compared to BLAST's 35%, and the genus-level specificity

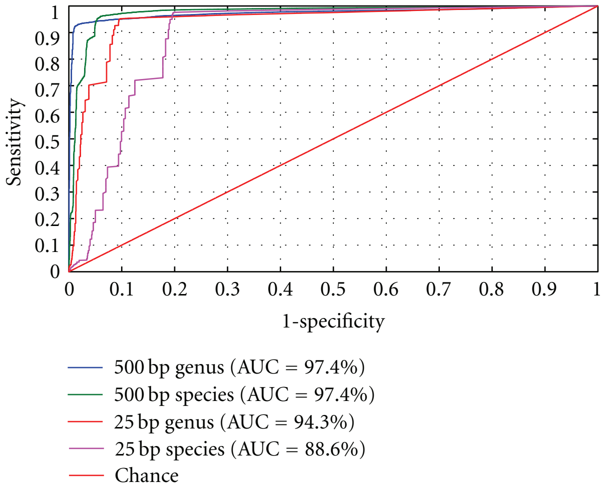

Figure 3: Comparison of ROC curves using the likelihood-based NBC scores. The AUC metric shows that the detector performs best on 500 bp reads (and not that much lower for 25 bp) for both the species and genus levels for the 100 reads each from the 635 known and 102 unknown training genomes.

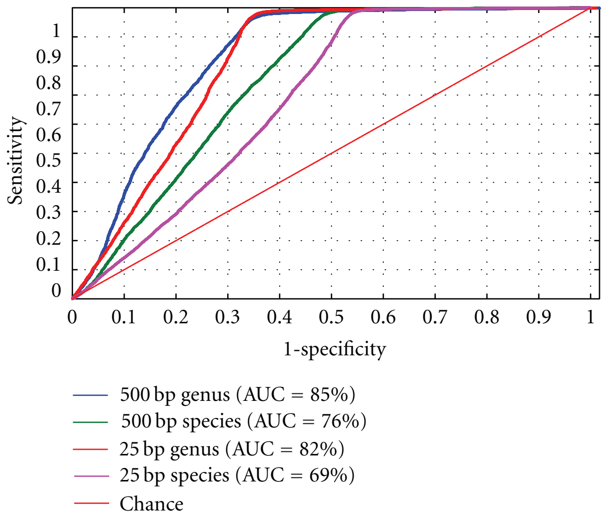

Figure 4: The PhymmBL ROCs follow a similar shape to the BLAST ROCs but have significantly better optimal operating points. Yet, the overall AUC of each of the curves falls below that of NBC curves.

is 65% compared to BLAST's 60%. The sensitivity that is sacrificed is only a 1-2% decrease from BLAST's 99%. While the 500 bp and 25 bp genus-level operating points are the same, the 500 bp read's AUC is better than the 25 bp read's AUC. For the species level, the 500 bp operating point and AUC are better than the 25 bp operating point.

3.4. Implementation of the Detector: Extrapolating the Operating-Point Thresholds to All Fragment Sizes. One of the obstacles to using the best derived ROC operating point is that the threshold changes according to the fragment length. This is an important aspect to address since in any given dataset from next-generation sequencing technology, the read lengths are variable (usually with an average length). To overcome this obstacle and adjust the threshold for each read length, we interpolate between different operating thresholds for the three read lengths. This gives a heuristic equation that can then be used for any read length. Previously, the best operating threshold point was chosen for the 500 bp, 100 bp, and 25 bp reads for each of the strains, species, and genus classifications. The best operating point is determined as the point that obtained the combined best sensitivity and specificity, which sum closest to 200% (100% for specificity/sensitivity, resp.).

For NBC thresholds, we use a linear interpolation. For example, using this method on the species level, the NBC log-likelihood best operating thresholds were determined to be −8079, −1445, and −185 for 500 bp, 100 bp, and 25 bp, respectively. A linear interpolation between these points yields a good fit (where the R^2 fit value [14] is $1 - 3e^{-}6$ or 1 when rounded). On average for the strain, species, and genus classification, the linear log-likelihood fit is $y = -16.6x+210$, where x is the length of the read and y is the likelihood detector threshold.

For PhymmBL, we found that the best thresholds for species level were −270.6, −27.4, and −18.5 for 500 bp, 100 bp, and 25 bp reads. Even though the PhymmBL scores are not truly likelihood scores (since they combines Phymm's likelihood score and BLAST's e-value), the thresholds follow a trend, and we can develop a heuristic interpolation for them. A parabolic curve approximated the interpolation better in this case. For example, the species-level fit is modeled by $y = -0.001x^2 - 0.0074x - 17.75$ (where the R^2 fit is again nearly 1).

3.5. Testing the Detector on 275 Novel Genomes. The detectors, developed using the 635 known genomes and the 102 unknown genomes, are used to accept or reject 100 reads from each of the 275 new genomes as "known" or "unknown", respectively. The new 275 genomes are all "unknown" on the strain level but some have "known" status on higher levels (as described in the Materials and Methods section). In addition to evaluating the detectors, we also assess the ability of MEGAN and SOrt-ITEMS to accept/reject taxa via their capability to classify a read at the species/genus level. The sensitivity, specificity, and detector accuracy are calculated for all methods and can be seen in Table 1. The detectors' ability to accept/reject reads from novel species was better than accepting/rejecting reads from novel genera, unlike homology-based LCA algorithms that perform better at higher-level taxa. This can be due to the fact that there are more unknown species in the set, and, therefore, if the detector's specificity is high, it will do better on classes with more "unknowns".

To compare the implementations of NBC+detector and PhymmBL+detector against current methods, we downloaded MEGAN version 3.7.2. Also, we downloaded the SOrt-ITEMS that was last updated January 7th, 2010 and used TBLASTX since it requires a protein-BLAST search. We also benchmarked against WebCarma 1.0, run on March 8, 2010. For WebCARMA, we infered whether a taxon was in

Table 1: Sensitivity, specificity, and (detector) accuracy rates of *detectors* for accepting/rejecting reads as "known" from a 275-strain test-set. Using 5-fold cross-validation, the maximum standard deviation is 1%. If all fragments were rejected, the species level would obtain 66% accuracy and the genus level 37% accuracy, and PhymmBL+Detector achieves 15–30% above this threshold. SOrt-ITEMS did not classify any fragment below the genus level, so N/A is designated for the species level. WebCarma's performance using 500 bp fragments resulted in a 20.1% sensitivity, 86.9% specificity, and 54% detector accuracy for the species level, and 23% sensitivity, 85% specificity, and 40.3% detector accuracy for the genus level. WebCarma only classified about 10 K of the 27.5 K reads. Due to its poor performance, we did not include it in the table.

NBC detector						
	Species			Genus		
Fragment length	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
500 bp	53.7%	96.3%	81.9%	32.9%	99.9%	58.0%
100 bp	62.2%	95.5%	84.3%	39.3%	99.5%	61.8%
25 bp	77.4%	89.6%	85.5%	61.7%	76.6%	67.3%
PhymmBL detector						
	Species			Genus		
Fragment length	Sensivitiy	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
500 bp	84.0%	88.3%	86.8%	58.5%	97.4%	73.0%
100 bp	79.9%	92.0%	87.9%	52.5%	98.3%	69.6%
25 bp	77.2%	86.8%	83.5%	51.2%	92.6%	66.7%
MEGAN as a detector						
	Species			Genus		
Fragment length	Sensivitiy	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
500 bp	83.3%	60.0%	68.1%	76.6%	66.5%	72.8%
100 bp	79.5%	71.4%	74.2%	66.9%	76.8%	70.6%
25 bp	71.0%	74.5%	73.2%	55.3%	73.4%	62.1%
SOrt-ITEMS as a detector						
	Species			Genus		
Fragment length	Sensivitiy	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
500 bp	N/A	N/A	N/A	57.1%	96.5%	71.2%
100 bp	N/A	N/A	N/A	44.8%	97.9%	64.5%
25 bp	N/A	N/A	N/A	6.1%	98.7%	40.5%

the database or not by checking if that taxon showed up in the results file. If it did not, we declared it as unknown. WebCARMA only annotated approximately 10,000 of the 27,500 reads resulting in poor genera/species detection and detection accuracy.

3.5.1. Comparison of Detector Performances. The results in Table 1, which summarize the sensitivity, specificity, and detection-accuracy of NBC and PhymmBL detectors, indicate that the performance of the PhymmBL detector is better than the NBC detector for 500 bp and 100 bp reads whereas NBC detector is better for the 25 bp reads. We conjecture that NBC is overfitting the operating-point thresholds due to the linear-interpolation heuristic, which interpolates the operating threshold between different-sized read lengths, and a more intelligent interpolation may be needed.

On the test data, PhymmBL's sensitivities were better than NBC's, but the specificity rates were not as good as NBC's. Nonetheless, PhymmBL+detector worked better for most of the reads and species/genus levels. It can achieve around 80% sensitivity and 90% specificity for the species level and 50+% sensitivity and 90+% specificity for the genus level. We hypothesize that NBC, because of its dependence on fixed Nmer size overfits the data compared to PhymmBL, and therefore the thresholds derived on the training dataset do not extend to the test set as well.

MEGAN and SOrt-ITEMS can also be used as detectors. MEGAN uses an LCA algorithm to determine if a read should be assigned to a particular taxonomic level, and SOrt-ITEMS uses additional alignment information. We used the default parameters for the MEGAN and SOrt-ITEMS. In order to do a fair comparison, we used the same BLAST reports that were generated for the 500 bp PhymmBL analysis for MEGAN, and we obtained a TBLASTX report for SOrt-ITEMS. For the 27,500 reads, 4 reads did not get scored by BLAST and, therefore, were not even assessed by the methods. If the method assigned a read at the genus/species level, we determined that this read "passed" its built in detector, regardless of the accuracy of the assignment. In other words, if a read is assigned to the family level, it is considered "unknown" at the species/genus levels. If a read is assigned to the species level (and, therefore, consequently has most upper-level assignments as well) but is the wrong species and genus, we declare that MEGAN/SOrt-ITEMS detector

Table 2: Comparison of overall classification accuracies (the number of reads that are identified as "known" that are classified into their correct class plus the no. of unknowns that are correctly rejected divided by all reads) on the 275-strain test set. Using 5-fold cross-validation, the maximum standard deviation is 1%. NBC, BLAST, and PhymmBL, in their native form, cannot detect "unknown" classes while the methods combined with a detector can. Performance is also compared to MEGAN and SOrt-ITEMS accuracy. N/A is designated for the species-level for SOrt-ITEMS since it did not classify anything below the genus level. SOrt-ITEMS obtains the best performance for 500 bp reads for the genus level but is under the 1% standard deviation threshold to be statistically significant. WebCarma was not included because its overall performance for 500 bp reads was 50% for the species level and 37% for the genus level. Note that the overall classification performance increases dramatically when a detector is added to NBC and PhymmBL.

				Species			
Fragment length	NBC	BLAST	PhymmBL	MEGAN	SOrt-ITEMS	NBC + detector	PhymmBL + detector
500 bp	27.5%	28.1%	28.0%	63.2%	N/A	78.0%	78.6%
100 bp	25.3%	26.1%	26.9%	69.4%	N/A	78.3%	81.1%
25 bp	20.9%	22.8%	23.5%	68.1%	N/A	74.7%	73.6%
				Genus			
Fragment length	NBC	BLAST	PhymmBL	MEGAN	SOrt-ITEMS	NBC + detector	PhymmBL + detector
500 bp	43.4%	49.2%	51.4%	68.8%	71.0%	53.6%	70.8%
100 bp	37.6%	42.8%	44.4%	66.5%	64.0%	54.9%	67.4%
25 bp	30.0%	32.7%	33.5%	54.8%	40.1%	45.3%	60.3%

labels it as "known" (passed it) at the species-level and genus-level but misclassified it. If method did not assign a read or assigned it above the species/genus levels (e.g., as its family-label), we mark it as "rejected" for the species/genus level detection. We observe from Table 1 that MEGAN has worse specificity for both species and genus levels, than either NBC, or PhymmBL-based detectors. For the genus level, MEGAN has the best sensitivity out of all the methods, although its specificity remains low.

In order to be able to take the classifier accuracy into consideration in addition to the detector, we define an *"overall" classification accuracy* measure as (number of reads correctly classified) + (number of reads correctly rejected)/(total number of reads). The (number of reads correctly classified) can be calculated as (# of reads that pass the detector that are correctly classified).

For the species level, MEGAN's detection accuracy, shown in Table 2 is worse than the other detection methods. Consequently, its accuracy as a detector is about the same or lower than PhymmBL's detector. SOrt-ITEMS does not classify below the genus level. As a detector, it has low sensitivity and high specificity, but the sensitivity drastically decreases as the reads get shorter. On the other hand, SOrt-ITEMS has a high positive predictive value for the reads it does pass, and, therefore, it has a high detection accuracy for 500 bp on the genus level, which decreases for shorter reads.

Note that only 102 "unknown" genomes were used in the development of the detector as opposed to the 635 known genomes. For NBC+detector and PhymmBL+detector, we hypothesize that their performance lowers for genus-level classification compared to the species level because fewer "unknown" genus-level examples were available in the design of the detector. We conjecture that as the database grows and more "unknown" examples are available, a detector will be more accurate.

3.5.2. Comparison of Classifier Accuracy Rates with and without the Detectors. Finally, in Table 2, we show how the methods' classification accuracy (without an unknown class) can be improved with the detector. In this experiment, the classifier method first scores the reads; then, the reads are accepted or rejected using the detector. The classification accuracy of the reads that pass the detector is then computed. NBC, BLAST, and PhymmBL all perform 20–35% accuracy for the species and genus level, when used in their native form without a detector (we do not benchmark BLAST+detector, since PhymmBL uses BLAST and performs better than BLAST). However, when the detector is added, the classifiers' overall accuracy, defined as the number of reads correctly rejected plus the number of reads identified as "known" that were correctly classified divided by the total number of reads, significantly improves for both species level and genus-level classification.

To calculate MEGAN and SOrt-ITEMS overall classification accuracy, we scored both algorithms' output as a true positive if it correctly identified the correct genus or species, and we score the output as a true negative if the method only assigned unknown reads to higher than the genus level (in other words, it could not resolve the correct species/genus). For the species level, MEGAN performed worse than classifiers with detectors, as seen in Table 2. For the genus level, both methods are better than NBC+detector but is still worse than PhymmBL+detector. We conclude that the purely BLAST-based or composition based algorithms are not as good as the hybrid PhymmBL + detector for determining novel species/genera.

3.6. Coding versus Noncoding Detector Accuracy for the Composition and Hybrid Methods. A question arises; how does the structure of the genomes relate to the novel/known detection. Since noncoding regions are much more variable

Table 3: PhymmBL and NBC detector accuracy rates versus coding/noncoding reads (coding includes full and partial coding regions).

	PhymmBL			NBC		
Method	All	Coding	Noncoding	All	Coding	Noncoding
Species	87.5%	**87.8**%	79.4%	82.8%	82.7%	**83.2**%
Genus	72.6%	**73.7**%	68.2%	57.1%	56.9%	**62.4**%

than coding regions, are they more susceptible to errors in the detector methods?

We analyzed the 275-genome test set for the composition-based approaches to get an idea of how these different regions perform. We first ran the 27,500 500 bp reads through MetaGeneMark [15] to annotate the coding regions. We have shown that MetaGeneMark has almost 90% accuracy for predicting coding regions [16] for 500 bp fragments. MetaGeneMark did not annotate 766 of the reads, resulting in 2.8% of the 500 bp reads as noncoding. Since this is a significant stretch of noncoding DNA, we expect to have a small proportion.

We then examined the accuracy rates of the detector on the coding/noncoding regions in Table 3. We can see that PhymmBL has higher accuracy for coding regions than noncoding regions, and NBC is opposite. We hypothesize that since PhymmBL partially uses BLAST, that it is more likely to predict homologous gene regions correctly than noncoding regions. In fact, 32% of noncoding regions failed in the genus-level detection in PhymmBL. NBC, on the other hand, which is fully composition based, is more likely to recognize the unique signatures of the noncoding regions—and its noncoding genus-level detection accuracy is almost as high as PhymmBL's, but its discrimination between known and novel gene regions is not sufficient. This provides insight that the two methods may be complimentary.

4. Demonstration of Detector on a Difficult (Real) Dataset

While it is important to benchmark methods rigorously on a test set, we can only get a true insight into their usage by examining a real dataset. We, therefore, analyze the Soudan acid mine red drainage dataset, which is a sample taken near a borehole of the mine [17]. This is a challenging metagenomics sample because there are 317 K reads, where the average read length in this sample is 100 bp, which some researchers claim are very short reads [18]. In the red Soudan Mine set, the number of organisms found without the detector was 628/631, using NBC/PhymmBL, respectively, out of the 635. This is most likely false, since acid mine drainage is known to be of low complexity [19]. We also found the median number of reads per organism is 214/340 while the mean is 506/503 with standard deviation of +/− 954/545. So, while there is high variance for high-abundance organisms, there is an "even spread" of hits across the genome training set, highlighting a higher than expected diversity. While the Soudan sludge's diversity may be higher than that studied by Tyson et al. [19], it is doubtful that it is this high, and this issue highlights the difficulty of analysis on soil and water samples complicated by short reads. Therefore, the results in this section should be examined in a critical light, as all of these classifiers performed better for longer reads. The idea is to highlight the advantage of using methods that will "filter out" unknown taxa accurately, so that we can gain an accurate assessment of "known" taxa at particular levels.

The raw NBC scores found the four most abundant genomes to be (1) *Flavobacterium johnsoniae* with 12,816 reads, (2) *Trichodesmium erythraeum IMS101* with 9641 reads, (3) *Sorangium cellulosum (So ce56)* with 8747 reads, and (4) *Clostridium beijerinckii* with 7300 reads, *Johnsoniae* et al. are not usually found in marine environments while *Trichodesmium erythraeum IMS101* is. While these organisms could come from the soil part of the sludge, it is unlikely that they would survive in such a salty environment.

For PhymmBL (shown in Table 4), the most abundant four genomes are (1) *Gramella forsetii* with 4102 reads, (2) *Marinobacter aquaeolei* with 3885 reads, (3) *Flavobacterium johnsoniae* with 3480 reads, and (4) *Dinoroseobacter shibae* with 3402 reads. PhymmBL has identified marine microbes in 1, 2, and 4 that are indeed more likely to be present in the marine sludge.

Next, the species/genus detectors are evaluated on the dataset. For NBC, only 141 reads out of 317 K passed the species-level detector, and 179 reads passed the genus-level detector. For PhymmBL, 794 reads out of the 317 K reads passed the species-level detector while 1053 reads passed the genus-level detector. The detectors may seem too selective because sensitivities of the detectors are suboptimal, as shown in the Results section. Although as expected, the sensitivity is much higher for the PhymmBL detector because it passes more reads. Tables 4 and 5 illustrate the distribution of reads that passed the NBC and PhymmBL detectors.

For NBC, the top hit is *Marinobacter hydrocarbonoclasticus*, which can degrade hydrocarbons and is found in pollution—therefore, likely to be present in the sample [20]. The second hit, *Ruegeria* is known to metabolize sulfur and could play an important part in the acid drainage of the mine, therefore, it is also quite likely to be present in the sample [21]. *Rhodobacter sphaeroides* can metabolize sulfur compounds and is highly likely in the sample [22]. *Dinoroseobacter shibae* is known for its ability to perform aerobic anoxygenic photosynthesis, and since the red sample of the acid mine drainage is near the surface, it is also likely to be present [23]. In [17], authors found a wide range of metabolisms in the sample, and NBC passed organisms that had a diversity of metabolisms. Therefore, we show the power of the detector to discriminate reads that are most likely to be in our database, as opposed to not applying the detector, in which case "unknown" strains are simply misclassified. PhymmBL finds similar organisms but in a different order and since it has higher sensitivity, it is able to find more species/genus reads that pass the detector.

The abundant species that are in the top 10 raw reads but did not pass the detector are *Gramella forsetii*, *Flavobacterium johnsoniae*, *Polaromonas naphthalenivorans*, *Aeromonas salmonicida*, and *Rhizobium leguminosarum*.

Table 4: The table shows the distribution of top 10 most abundant species reads of PhymmBL and the top-8 species-reads passed the species resolution detectors for the red soudan acid mine drainage dataset, using the 635-genome training database.

(a)

PhymmBL	
Organism	Matched reads
Gramella forsetii	4102
Marinobacter hydrocarbonoclasticus	3885
Flavobacterium johnsoniae	3480
Dinoroseobacter shibae	3402
Ruegeria pomeroyi	3119
Polaromonas naphthalenivorans	3116
Aeromonas salmonicida	2899
Rhodobacter sphaeroides	2616
Rhizobium leguminosarum	2541
Paracoccus denitrificans	2533

(b)

NBC detector		PhymmBL detector	
Organism	Matched reads	Organism	Matched reads
Marinobacter hydrocarbonoclasticus	31	Dinoroseobacter shibae	85
Dinoroseobacter shibae	18	Marinobacter hydrocarbonoclasticus	62
Ruegeria sp. TM1040	17	Rhodobacter sphaeroides	24
Rhodobacter sphaeroides	15	Ruegeria pomeroyi	24
Shewanella sp. ANA-3	11	Ruegeria sp. TM1040	22
Shewanella baltica	5	Paracoccus denitrificans	20
Desulfotalea psychrophila	4	Shewanella baltica	17
Paracoccus denitrificans	4	Shewanella sp. ANA-3	14

Table 5: The table shows the distribution of top 8 most abundant genus reads that passed the genus-resolution detectors for the red soudan acid mine drainage dataset, using the 635-genome training database.

NBC detector		PhymmBL detector		SOrt-ITEMS	
Organism	Matched reads	Organism	Matched reads	Organism	Matched reads
Marinobacter	40	Dinoroseobacter	101	Marinobacter	476
Dinoroseobacter	24	Marinobacter	73	Gramella	388
Rhodobacter	23	Ruegeria	59	Dinoroseobacter	297
Shewanella	20	Rhodobacter	41	Rhodobacter	264
Ruegeria	19	Shewanella	41	Flavobacterium	161
Paracoccus	9	Pseudomonas	26	Pseudomonas	131
Desulfotalea	4	Bacillus	21	Alkalilimnicola	111
Bartonella	4	Clostridium	21	Roseobacter	101

Gramella forsetii may be truly present as it is reasonable to find in the sample since it degrades polymeric organic matter [24], but it is usually found in marine environments, so it is difficult to conclude. *Flavobacterium johnsoniae* and *Aeromonas salmonicida* are fish-born pathogens and could actually be present [25, 26], but this again is hard to conclude. *Polaromonas naphthalenivorans* is likely since it grows on hydrocarbons found in contaminated sediment [27], so that is a misrejection of the classifier. On the other hand, rejecting *Rhizobium leguminosarum* is quite reasonable since it is found in plants. So, these bacteria are still found after the detector but not as abundant. This can be for several reasons. One hypothesis is that reads that correspond to these bacteria may actually be from closely related but unknown species. Another hypothesis is that these reads are from horizontally transferred elements which are responsible for particular metabolisms but actually belong to an unknown species.

We also compared against SOrt-ITEMS, since this homology-based method had the best 500 bp classification accuracy for the genus level and was comparable to MEGAN for 100 bp reads. Compared to NBC/Phymm+detector, the main differences are that it accepts *Gramella* and *Flavobacterium* as a likely genera, although the other detectors reject these. It also finds a likely presence of *Alkalilimnicola* [28], which is an arsenite oxidizing bacterium and is found commonly in contaminated waters. This intuitively seems likely in the acid mine setting and seems to have an advantage over the other methods. SOrt-ITEMS passes *Ruegeria*, which is known to be in marine-only settings but could be present due to the saline nature.

We can see that the PhymmBL/NBC+detector methods are more selective in confidently "passing" reads through the detector than SOrt-ITEMS, and we hypothesize that the composition-based methods, such as NBC and PhymmBL, better reject some organisms that are unknown at the genus level. However, it may also be true that the sensitivity of these methods may not be as accurate on a real dataset, as seen with the SOrt-ITEMS discovery of the abundance of *Alkalilimnicola*.

All the methods agree that under 1% of the data originates from previously sequenced genera! This agrees with the hypothesis that 90+% of species cannot be cultured and, therefore, have not been previously sequenced [1]. This is an amazing result and shows us the difficult problems that metagenomics samples pose. We only examine species/genus classifications in this paper, but we hypothesize that more than 1% can be successfully classified into higher taxonomic levels. We emphasize that this discussion is an exercise in analyzing an extremely challenging dataset that has a diversity of organisms and short ~100 bp reads.

5. Discussion

In this paper, we introduce a novel/known organism detector, an automated approach to determine whether a given organism is previously known or novel. The approach can be used with composition- or hybrid-based taxonomic classifiers. We also rigorously benchmarked all relevant algorithms to assess their performances and to distinguish novel from known reads. Being able to discriminate between next-generation sequencing reads that originate from known novel organisms will allow biologists to make new organism discoveries, have higher confidence in those reads that come from previously sequenced species, and discern other domain-level contaminants (such as viral or eukaryotic DNA) in the sample. Previously developed homology-based methods can be used for such detection, but, as we have shown, those methods are suboptimal, especially at the species level. Overall, the PhymmBL detector is the best and obtains ~85+% accuracy for accepting/rejecting species reads (in novel detection) and ~70+% for genus reads. In PhymmBL, misclassification of reads that pass the detector causes a 10%/3% drop in overall accuracy for 500 bp reads. For example, for the species-level, the detector accuracy is 86.8%, but it is overall classification accuracy (correctly classifying the known reads plus labeling the unknowns) is 78.6%. This detection → classification drop is about 4-5% for MEGAN, but MEGAN has worse detector/classification accuracy to begin with. An impressive factor in SOrt-ITEMS, is that it only drops 0.2% when classifying reads that pass the detector, meaning this method is very confident in correctly classifying any read that passes its detector. This makes SOrt-ITEMS overall accuracy one of the best for the genus level at 500 bp.

The next step will be to implement such a detector for upper levels on the tree of life. The end product will then be to supply a probabilistic threshold to users, so that reads that have a likelihood score above this threshold can be confidently labeled as "known" whereas those whose score falls under this threshold can be confidently labeled as "novel". Such a detector can give a first pass of all the reads that may belong to the database, and would be useful in determining new species. On the other hand, we can also use this system to determine whether a given taxon is novel at deep branching within the tree of life. Once reads are identified to come from novel species, they can then be placed in the phylogenetic tree to determine their position in the tree of life.

The limitation of this approach is that it is only a "first pass" at labeling the reads that originate from known/novel organisms. Further interpretation is then required to determine novelty. For example, a read that is on the "threshold of detection" may just be another allele of a gene within the same species. Those with very low scores are more likely to be novel, and then they will have to be aligned and placed in a tree with other sequences to determine lineages. But with the vast amount of information coming from metagenomics datasets (with millions of reads), obtaining a "first-pass" set of sequences that is a fraction of the original number can significantly reduce computational time of subsequent phylogenetic analysis.

6. Conclusions

This work develops a detector and demonstrates its application to identify known and unknown genomes for composition-based classification methods, and we demonstrate that our detector with a hybrid method outperforms current homology-based methods. Effectively, the detector introduces an "unknown" class and enables classification methods to filter out reads that will not be classified correctly, resulting in improved classification accuracy. In addition to detecting novel genomes, we also propose that the detector can be used to filter out noisy reads that have lowconfidence when scored.

We use the previously implemented naÏve Bayes classifier and PhymmBL (interpolated Markov model plus BLAST), which assigns a read to the closest match in the database, to design a detector that can detect reads from previously sequenced organisms. We show that NBC and PhymmBL scores can be used to determine if a read is from a novel organism in respect to the training database. The overall classification accuracies of composition-based methods are greatly improved when detectors are added to filter out "unknown" organisms. Also, there is only a mild decrease in performance when classifying ultrashort reads as opposed to Roche 454 length. We determine that the PhymmBL + detector classification performs similarly or better than all the methods. Overall, the PhymmBL detector obtains ∼85+% accuracy for accepting/rejecting species reads and ∼70+% for genus reads and only slightly lower in the overall classification accuracy.

Funding

This paper was supported by the National Science Foundation Career award no. 0845827 and DOE award DE-SC0004335.

Conflict of Interests

The authors declare no conflict of interests.

References

[1] J. Handelsman, *Committee on Metagenomics: Challenges and Functional Applications*, The National Academies Press, Washington, DC, USA, 2007.

[2] J. C. Venter, K. Remington, J. F. Heidelberg et al., "Environmental genome shotgun sequencing of the Sargasso Sea," *Science*, vol. 304, no. 5667, pp. 66–74, 2004.

[3] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster, "MEGAN analysis of metagenomic data," *Genome Research*, vol. 17, no. 3, pp. 377–386, 2007.

[4] E. A. Dinsdale, O. Pantos, S. Smriga et al., "Microbial ecology of four coral atolls in the Northern Line Islands," *PLoS One*, vol. 3, no. 2, Article ID e1584, 2008.

[5] J. Reeder and R. Knight, "The 'rare biosphere': a reality check," *Nature Methods*, vol. 6, no. 9, pp. 636–637, 2009.

[6] S. Essinger and G. L. Rosen, "Benchmarking blast accuracy of genus/phyla classification of metagenomic reads," in *Pacific Symposium on Biocomputing*, 2010.

[7] C. Quince, A. Lanzén, T. P. Curtis et al., "Accurate determination of microbial diversity from 454 pyrosequencing data," *Nature Methods*, vol. 6, no. 9, pp. 639–641, 2009.

[8] A. Brady and S. L. Salzberg, "Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models," *Nature Methods*, vol. 6, no. 9, pp. 673–676, 2009.

[9] G. L. Rosen, E. R. Reichenberger, and A. M. Rosenfeld, "NBC: the Naïve Bayes classification tool webserver for taxonomic classification of metagenomic reads," *Bioinformatics*, vol. 27, no. 1, pp. 127–129, 2011.

[10] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler, "GenBank," *Nucleic Acids Research*, vol. 36, no. 1, pp. D25–D30, 2008.

[11] M. M. Haque, T. S. Ghosh, D. Komanduri, and S. S. Mande, "SOrt-ITEMS: sequence orthology based approach for improved taxonomic estimation of metagenomic sequences," *Bioinformatics*, vol. 25, no. 14, pp. 1722–1730, 2009.

[12] W. Gerlach, S. Jünemann, F. Tille, A. Goesmann, and J. Stoye, "WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads," *BMC Bioinformatics*, vol. 10, article 430, 2009.

[13] G. Rosen, E. Garbarine, D. Caseiro, R. Polikar, and B. Sokhansanj, "Metagenome fragment classification using n-mer frequency profiles," *Advances in Bioinformatics*, vol. 2008, Article ID 205969, 12 pages, 2008.

[14] N. R. Draper and H. Smith, *Applied Regression Analysis*, Wiley-Interscience, New York, NY, USA, 1998.

[15] W. Zhu, A. Lomsadze, and M. Borodovsky, "Ab initio gene identification in metagenomic sequences," *Nucleic Acids Research*, vol. 38, no. 12, p. e132, 2010.

[16] N. G. Yok and G. L. Rosen, "Combining gene prediction methods to improve metagenomic gene annotation," *BMC Bioinformatics*, vol. 12, no. 1, 2011.

[17] R. A. Edwards, B. Rodriguez-Brito, L. Wegley et al., "Using pyrosequencing to shed light on deep mine microbial ecology," *BMC Genomics*, vol. 7, article 57, 2006.

[18] K. E. Wommack, J. Bhavsar, and J. Ravel, "Metagenomics: read length matters," *Applied and Environmental Microbiology*, vol. 74, no. 5, pp. 1453–1463, 2008.

[19] G. W. Tyson, J. Chapman, P. Hugenholtz et al., "Community structure and metabolism through reconstruction of microbial genomes from the environment," *Nature*, vol. 428, no. 6978, pp. 37–43, 2004.

[20] A. Lattuati, P. Metzger, M. Acquaviva, J. C. Bertrand, and C. Largeau, "n-Alkane degradation by Marinobacter hydrocarbonoclasticus strain SP 17: long chain β-hydroxy acids as indicators of bacterial activity," *Organic Geochemistry*, vol. 33, no. 1, pp. 37–45, 2002.

[21] J. M. González, J. S. Covert, W. B. Whitman et al., "Silicibacter pomeroyi sp. nov. and Roseovarius nubinhibens sp. nov., dimethylsulfoniopropionate-demethylating bacteria from marine environments," *International Journal of Systematic and Evolutionary Microbiology*, vol. 53, no. 5, pp. 1261–1269, 2003.

[22] V. van Fleet-Stalder, T. G. Chasteen, I. J. Pickering, G. N. George, and R. C. Prince, "Fate of selenate and selenite metabolized by Rhodobacter sphaeroides," *Applied and Environmental Microbiology*, vol. 66, no. 11, pp. 4849–4853, 2000.

[23] H. Biebl, M. Allgaier, B. J. Tindall et al., "Dinoroseobacter shibae gen. nov., sp. nov., a new aerobic phototrophic bacterium isolated from dinoflagellates," *International Journal of Systematic and Evolutionary Microbiology*, vol. 55, no. 3, pp. 1089–1096, 2005.

[24] M. Bauer, M. Kube, H. Teeling et al., "Whole genome analysis of the marine Bacteroidetes 'Gramella forsetii' reveals adaptations to degradation of polymeric organic matter," *Environmental Microbiology*, vol. 8, no. 12, pp. 2201–2213, 2006.

[25] L. Flemming, D. Rawlings, and H. Chenia, "Phenotypic and molecular characterisation of fish-borne Flavobacterium johnsoniae-like isolates from aquaculture systems in South Africa," *Research in Microbiology*, vol. 158, no. 1, pp. 18–30, 2007.

[26] M. E. Reith, R. K. Singh, B. Curtis et al., "The genome of Aeromonas salmonicida subsp. salmonicida A449: insights into the evolution of a fish pathogen," *BMC Genomics*, vol. 9, article 427, 2008.

[27] C. O. Jeon, W. Park, W. C. Ghiorse, and E. L. Madsen, "Polaromonas naphthalenivorans sp. nov., a naphthalene-degrading bacterium from naphthalene-contaminated sediment," *International Journal of Systematic and Evolutionary Microbiology*, vol. 54, no. 1, pp. 93–97, 2004.

[28] S. E. Hoeft, J. S. Blum, J. F. Stolz et al., "Alkalilimnicola ehrlichii sp. nov., a novel, arsenite-oxidizing haloalkaliphilic gammaproteobacterium capable of chemoautotrophic or heterotrophic growth with nitrate or oxygen as the electron acceptor," *International Journal of Systematic and Evolutionary Microbiology*, vol. 57, no. 3, pp. 504–512, 2007.