Published in final edited form as:

Nat Methods. 2009 September ; 6(9): 673–676. doi:10.1038/nmeth.1358.

Phymm and PhymmBL: Metagenomic Phylogenetic Classification with Interpolated Markov Models

Arthur Brady¹ and Steven L. Salzberg¹

¹ Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742

Abstract

Metagenomics projects collect DNA from uncharacterized environments that may contain thousands of species per sample. One main challenge facing metagenomic analysis is phylogenetic classification of raw sequence reads into groups representing the same or similar taxa, a prerequisite for genome assembly and for analyzing the biological diversity of a sample. New sequencing technologies have made metagenomics easier, by making sequencing faster, and more difficult, by producing shorter reads than previous technologies. Classifying sequences from reads as short as 100 base pairs has until now been relatively inaccurate, requiring researchers to use older, long-read technologies. We present Phymm, a classifier for metagenomic data, that has been trained on 539 complete, curated genomes and can accurately classify reads as short as 100 bp, representing a substantial leap forward over previous composition-based classification methods. We also describe how combining Phymm with sequence alignment algorithms, further improves accuracy.

Introduction

Dramatic improvements in the speed and efficiency of DNA sequencing have encouraged the rapid growth of metagenomics, the study of DNA collected directly from environmental samples. This new field, which promises to uncover thousands of previously unknown species, has been compared to “a reinvention of the microscope in the expanse of research questions it opens to investigation.”¹ Only a small fraction of microbial organisms can be cultured in a laboratory, a prerequisite for traditional genome sequencing and analysis. Single-organism genome sequencing projects have yielded a wealth of new scientific knowledge, which we are only beginning to exploit. Metagenomics promises to take these discoveries even further, by enabling scientists to study the full diversity of the microbial world. By sequencing collections of organisms from environments ranging from the human body to soil to the ocean floor, metagenomics projects are vastly increasing the range of organisms that can be analyzed, allowing for systems-level study of microbial environments, and revealing a heretofore hidden world of biological complexity².

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use:http://www.nature.com/authors/editorial_policies/license.html#terms

Email: abrady@umiacs.umd.edu, salzberg@umiacs.umd.edu.

Although sequencing technology allows us to collect and sequence vast samples of uncultured DNA, many technical challenges must be overcome in order to make sense of these data. In order to identify the species and genes in a sample, DNA fragments (“reads”) from common species need to be grouped together and assembled, if possible.

The newest sequencing technologies (increasingly preferred by metagenomics researchers due to reduced cost) produce relatively short reads, ranging from 25–400 bp, making the taxonomic classification problem considerably more difficult than with longer reads, which contain more identifying information. One algorithm, CARMA3, attempts to match short reads to known Pfam domains (structural components conserved across multiple proteins) and protein families. However, only ~15% of random 100-bp shotgun reads could be matched to extant Pfam groups; even among this reduced set of reads (excluding 85% of the input data due to the absence of a match), the average sensitivity of this approach at the genus level was 40%, meaning only 6% of the input reads were correctly classified. PhyloPythia4, a classification method based on support vector machines (SVMs), examines oligonucleotide frequencies to characterize taxonomic groups. This method is effective for DNA fragments of 3,000 bp and longer, but for 1,000-bp sequences, sensitivity drops drastically (to just 7.1% at the genus level). It has been observed⁵ that 1,000 bp represents a critical barrier that classification methods need to break.

Another approach is to use sequence homology, aligning reads to known sequences using BLAST⁶ and assigning taxonomy based on the best match^{7,8}. BLAST is highly accurate if the source organism; however, if the source species is missing from the database, accuracy drops dramatically, as we show below. The MEGAN software system⁹ classifies reads based on multiple high-scoring BLAST hits, assigning reads to a common ancestor of those BLAST matches that exceed a bit-score threshold. It is worth noting that in a recent coral reef study¹⁰, only 12% of reads had matches in a comprehensive microbial BLAST database.

Here we present Phymm, a new method for phylogenetically classifying short sequence fragments such as those generated by metagenomics sequencing projects. Throughout this paper, we use “classification” to mean the assignment of a specific label (in our case, a phylogenetic group) to members of a data set (in this case, DNA reads). We want to draw a clear distinction to ‘binning’ which, though often used interchangeably with classifying, refers to the grouping of a data set into subgroups which are distinct from one another; these subgroups may remain unlabeled.

Phymm uses interpolated Markov models (IMMs) to characterize variable-length oligonucleotides typical of a phylogenetic grouping. IMMs have been employed with great success for bacterial gene finding in the GLIMMER system¹¹, but this is the first use of IMMs for the general phylogenetic classification problem. Our results demonstrate that for short reads, Phymm represents a dramatic improvement over previous methods such as PhyloPythia4, accurately classifying unknown fragments as short as 100 bp. We also present a hybrid method that incorporates information from both Phymm and BLAST, and show that this hybrid method outperforms either of the two single methods.

Results

In an earlier study describing the program GLIMMER12, IMMs were shown to be highly accurate at distinguishing reads derived from a bacterial symbiont *P. didemni* from those of its eukaryotic sea squirt host *L. patella*. GLIMMER discriminated with 99% accuracy between reads from two evolutionarily distant species. Our goal in the present study was to see how well this discrimination generalizes to the problem of fully classifying much larger metagenomics samples that included many closely-related organisms. We tested our method on both synthetic and real metagenome data.

Synthetic metagenome data

We conducted three groups of classification experiments, in which test sets of synthetic metagenomic reads were assigned taxonomic labels using Phymm, BLAST and PhymmBL. Phymm contains a large suite of IMMs trained on chromosomes and plasmids from organisms collected from the NCBI RefSeq database¹³ (Phymm's "reference library"). When used to score a DNA sequence, an IMM computes a score corresponding to the probability that the IMM generated that sequence, which can be used to estimate the probability that the sequence belongs to the class of sequences on which the IMM was trained. In the Phymm experiment, each read was scored with each IMM in the reference library, and the read was then classified using the clade labels belonging to the organism whose IMM generated the best score for that read. In the BLAST experiment, each read was submitted as a BLASTN query, searching against a background database built from the same genomes used to generate the IMMs, and clade labels were assigned using the known labels of the best BLAST hit. Finally, we used PhymmBL to score reads using both stand-alone methods in parallel, assigning a "best hit" using a weighted combination of scores from both methods. (See Online Methods for reference sequence data, test set construction, experimental configuration, and the PhymmBL scoring algorithm.)

Clade-level exclusions

Our central goal was to model the problem of classifying a sequence from a species that has never before been observed; we expect that metagenomic data will lead to the discovery of thousands of new species, and that this problem will be a common one. By definition, given a read from an organism which has not been sequenced, no classification method can predict the correct species label (because that label does not exist). For higher-level phylogenetic classifications, on the other hand, we may have previously seen the genus, family, or other higher-level clade. We therefore repeated all three groups of experiments multiple times, with each iteration configured to explicitly exclude comparisons of each query read to related species at increasingly general clade levels.

Classification accuracy

We focus first on the experiment in which Phymm was used to classify reads for which species-level matches were masked. This experiment was repeated for reads of lengths 100, 200, 400, 800 and 1,000 bp (**Online Methods**). Accuracy results for this experiment on all five read lengths are given in Fig. 1 and in Supplementary Table 1. As expected, classification at the genus level was the most difficult task, with 53 possible genera available

as labels. Moving up the phylogenetic tree, the data sets contained 48 distinct bacterial and archaeal families, 34 orders, 21 classes, and 14 phyla. Phymm's accuracy improves with greater read length, as expected, ranging from 32.8% for the problem of assigning the correct genus to 100-bp reads, up to 89.8% for assigning the correct phylum to 1,000-bp reads. At 400 bp the read length provided by current-generation Roche 454 pyrosequencers accuracy at the genus level is 60.3%. These results represent very substantial improvements over previous methods such as the SVM-based PhyloPythia, which reported 7.1% accuracy at the genus level for 1,000-bp sequences⁴ and C ARMA, which labeled 6% of 100-bp reads correctly at the genus level.³

In most cases, BLAST by itself is superior to Phymm by itself, but for the 800-bp and 1,000-bp reads, Phymm outperforms BLAST at the class and phylum levels (Supplementary Table 2).

The hybrid PhymmBL classifier produces further improvements (Table 1, Supplementary Table 3). For all read lengths and clade levels, PhymmBL outperforms both Phymm and BLAST, showing approximately 6% improvement over BLAST by itself at all taxonomic levels for the 1,000-bp query set. These results indicate that the two approaches are somewhat complementary and that PhymmBL is able to utilize information from both. Table 1 shows a comparison of all three methods, along with results for PhyloPythia, at the 1,000-bp read length.

Both Phymm and PhymmBL give highly robust, reproducible results: in all cases, the observed standard deviation in accuracy was less than 1%.

Finally, because more than 1,100 IMM scores are assigned to each read, we considered whether looking beyond the single top-scoring IMM might improve accuracy. We therefore examined the top five scores assigned by PhymmBL to each read in the 1,000-bp test set, and counted how often the correct clade appeared in at least one of the top five predictions. This produced an increase of 5–9% at all levels, as follows (with the accuracy of the top-scoring prediction alone shown in parentheses): phylum, 97.9% (93.8%); class, 96.6% (90.6%); order, 94.5% (86.9%); family, 92.6% (84.8%); genus, 89.3% (78.4%). This suggests that Phymm and PhymmBL might be improved even further if they can make use of additional complementary signals.

Metagenome data from an acid mine

Evaluating classification methods on real metagenomes can be problematic: ordinarily, the true taxonomic composition of the data cannot be established with certainty. The composition of an acid mine drainage (AMD) metagenome first introduced in a 2004 study¹⁴, however, has been substantially characterized. It contains three dominant populations, namely the archaeon *Ferroplasma acidarmanus* and two groups of bacteria, *Leptospirillum* groups II and III.

We ran two experiments classifying the AMD data: one in which we used the unaltered RefSeq library as a reference, and one in which we added the draft genomes for all three groups to the RefSeq data to create an augmented reference library. This allowed us to

compare – as we did for the synthetic data set – performance in the presence of more and less specific reference data.

We present three different characterizations of the AMD data in terms of population distribution: Fig. 2 compares population breakdown at the phylum level as predicted by PhymmBL across two runs: one using the augmented reference library (including the three new draft genomes), the other using only the RefSeq data. Fig. 3 gives the population breakdown by species given by PhymmBL using the augmented reference library.

In addition to these high-level characterizations, we examined how well Phymm, BLAST and PhymmBL – using only the RefSeq library *without* the added draft genomes – could characterize the three dominant species groups themselves. We aligned all the AMD reads to the three draft genomes, in order to establish a “correct” read set for each species group, and then examined the classification performance for each of the three methods on each of the three positively-identified groups of reads. Phylum-level predictions for these groups are given in Supplementary Figs. 1–9. PhymmBL correctly predicted the phylum (*Euryarchaeota*) for *F. acidarmanus* 61.0% of the time; it assigned approximately 80% of *Leptospirillum* reads to *Proteobacteria*. (Note that members of the phylum (*Nitrospira*) assigned to the *Leptospirillum* genus were originally provisionally assigned to *Deltaproteobacteria*15, and indeed a majority of reads from both groups were labeled with this class: 59.9% for group II, and 51.7% for group III.)

When the three positively-identified sets of reads were analyzed using the reference library augmented *with* the draft genomes for the three groups, PhymmBL’s species-level prediction accuracy was 98% or higher for each group (data not shown). (We believe these accuracy measurements to be more anecdotal than evidentiary, though, due to the fact that the draft genomes, which were used to train IMMs, were presumably assembled from the same raw reads which we used as test data for this particular run, although these results do give an indication of PhymmBL’s robustness in the presence of sequencing errors and natural mutations, which were certainly present in the raw read data.)

Discussion

The new *ab initio* classification method presented here, Phymm, is substantially more accurate than previous approaches, particularly for short read lengths. One advantage of interpolated Markov models over other methods, particularly those that use oligonucleotide counts16, is that IMMs utilize information from multiple oligonucleotides of different lengths and integrate the results. Thus instead of having to choose between 5-mers and 6-mers for classification at the class or phylum levels (as is done in PhyloPythia4), Phymm can use both. For our experiments we considered oligomers ranging from 1–12 in length, and Phymm automatically selected those oligomers that best characterized each species. Note also that these are comparable to “spaced seeds” in that the positions from which information is extracted are not necessarily adjacent. The program GLIMMER12 showed the effectiveness of IMMs for a binary classification problem; here, we classified reads from hundreds of species, all of which are more closely related to one another than the bacterium-

sea squirt pair analyzed by GLIMMER, and found that IMM is also effective at this much more difficult task.

In contrast to other approaches, Phymm classifies all reads, and its accuracy at the genus level is 32.8% for 100-bp reads (compared to 6% for CARMA); for 1,000-bp reads, genus-level accuracy rises to 71.1% (compared to 7.1% for PhyloPythia).

However, as our experiments demonstrate, the best standalone method for classifying reads from metagenomics projects – at least when other species from the same genus are known – is BLAST. Although BLAST has been the standard classification method for long reads^{7,14}, some previous studies excluded it, which may have made their results appear more positive. BLAST has shortcomings when a sequence is truly different from anything in the database – as is often the case with actual environmental samples – but this is a universal problem. As has been previously observed¹⁷, marker gene approaches based on 16S rDNA do not appear to improve classification accuracy over that of BLAST alone. In contrast, our IMM-based method provides a clear boost to BLAST: our hybrid method, PhymmBL, outperformed each of its two component methods alone, and we note that both Phymm and BLAST contributed substantial numbers of correct assignments that were missed by the other method.

One of the main advantages of our phylogenetic classification method is that preprocessing reads is unnecessary: no gene finding, protein-domain matching, or conserved-sequence identification steps were needed, allowing predictions to be made for all reads in a query set without sacrificing accuracy as compared to existing methods. The 1Kb “barrier”⁵ does not seem to be a problem for Phymm, although as with all methods, accuracy improves with longer reads. As read lengths for new sequencing technologies increase, the ability to accurately classify short reads, directly as they emerge from the sequencers, will continue to improve. Current metagenomics analysis pipelines⁵ postpone classification until after assembly has been attempted, due to the unreliability of existing composition-based methods at accurately generating phylogenetic classifications for sequences less than 1,000 bp. As shown in the sea squirt/symbiont study¹², accurate binning can improve assembly; methods like Phymm and PhymmBL (when used as binning tools) should thus be able to improve all downstream analysis of metagenomes, including assembly and gene finding.

Online Methods

Synthetic metagenome data

At present there are no benchmark metagenomics datasets based on real environmental sequences for which the correct taxonomy has been fully and confidently characterized. Indeed, all existing metagenomic samples from natural environments contain multiple unknown species. Following the work of others¹⁸, therefore, we worked with simulated metagenomic samples drawn from existing sequences. For our main data source, we built a core library of all complete bacterial and archaeal genomes available in RefSeq as of October, 2008¹³, comprising 539 distinct species containing 1,146 chromosomes and plasmids. (The full taxonomic composition of this data set can be found in Supplementary Tables 7–11, along with the amount of sequence data present for each clade.) All reads used

in our experiments with synthetic data were drawn randomly from their source genomes, with no alteration. For the first set of experiments in – which no matches were masked, and comparisons were allowed between reads and their source species – query reads were extracted from each genome, then masked out *before* training IMMs or building our BLAST database, so that our training set (genomes with extracted reads masked) and our test set (extracted reads) did not overlap. For all other experiments, IMMs and BLAST databases were constructed using whole genomes, and separation of training and test data was automatic: the entire genome of the species from which each query read was drawn was masked, by design, during the classification process.

Synthetic test set construction and filtering

In order to control against under-representation of some clades in the available data, query sets were filtered so that all species under consideration had at least two sister species within the clade under consideration. For example, in the experiment which masked exact species matches but allowed intra-genus comparisons, without this filtering step, if a given species was the only sequenced representative of its genus, then it would have been impossible to assign a correct genus label to reads from that species (because the species itself was excluded from the scoring process by design), which in turn would artificially depress the results. Analogous filters were used at higher levels; e.g., phylum-level predictions were made only for reads which had at least two other species in their home phylum.

Each synthetic test set initially contained 5 randomly-selected “reads” from each of the 1,146 chromosomes and plasmids in the RefSeq reference data, totaling 5,730 reads representing 539 bacterial and archaeal species. After filtering out species that did not meet the criteria above, test set sizes for the species-masked, genus-masked, family-masked, order-masked and class-masked experiments contained 2,870, 3,255, 3,335, 4,575 and 4,390 reads, respectively.

The most sparse of these the – experiment in which exact species matches were masked and the lowest-level predictions were at the genus level – represented 573 chromosomes and plasmids from 254 species across 48 genera. This is still a very broad group of genera compared to previous studies; e.g., McHardy *et al.* used 31 genera⁴.

Acid mine drainage test set

We downloaded the entire set of raw sequence reads for the AMD metagenome data set presented in Tyson *et al.*¹⁴ from the NCBI trace archive. Vector sequences were removed from reads using Figaro¹⁹. Each read was then truncated to include only the longest contiguous stretch of bases for which base-calling quality scores were at least 17. Finally, reads less than 100 bp were discarded. Of the original 180,713 raw reads in the data set, 166,345 remained after all quality filtering was complete.

We established “true positives” for reads belonging to *F. acidarmanus* and the two *Leptospirillum* groups by using MUMmer²⁰ to align all reads in the data set to the draft genomes for the three species groups. Three sets of positive matches, one for each species group, were identified by selecting reads which aligned to exactly one of the three draft

genomes. When complete, the *F. acidarmanus* set contained 15,628 reads; the *Leptospirillum* group II set contained 48,589 reads; and the *Leptospirillum* group III set contained 10,104 reads. These sets were used to generate the per-group predictive accuracy results presented in Supplementary Figures 1–9.

Classification infrastructure

We built one IMM per molecule (chromosome or plasmid), with each IMM trained on the entire molecular input sequence, yielding a total of 1,146 IMMs. To compare our method to BLAST on the same data, we also constructed a BLAST database containing all 1,146 molecular sequences.

In designing PhymmBL, we attempted to boost accuracy by adding several common composition-based measures including GC content and dinucleotide frequencies²¹, but none was found to improve accuracy, likely because the statistics captured by these measures closely overlap those already captured by the Phymm IMMs.

For more details on experimental methods, including a study examining the relationship between the amount of training data available per clade and predictive accuracy; construction and configuration of the synthetic data sets; details on our use of IMMs; a description of all preprocessing performed on the raw AMD read set before classification; parameter estimation for PhymmBL's weighted voting scheme; and the direct estimation of confidence levels for Phymm predictions, please refer to the supplementary material accompanying this article.

Impact of training data size on classification accuracy

Taxonomic groups are represented in sequence databases to vastly different degrees; among the RefSeq genera we used for this study, for example, the amount of sequence data differed by up to two orders of magnitude. To explore potential biases in classification accuracy resulting from such a wide range in the volume of available training data for each clade, we plotted clade-specific classification accuracy as a function of training set size. Results at the phylum level for 100-bp queries, classified by Phymm (with exact species matches excluded) are given in Supplementary Figure 10. This analysis was conducted 10 times to establish variance in accuracy; a tabular version of Supplementary Figure 10, including standard deviations in accuracy for each phylum, along with analogous results at the class, order, family and genus levels are given in Supplementary Tables 7–11.

Somewhat unexpectedly, little if any correlation was observed between the amount of sequence data available for IMM training and predictive power. Also unexpectedly, the standard deviations for predictive accuracy – which one might expect to vary widely and exhibit correlations with amount of training data – exhibited neither of these trends, with deviations ranging from less than 1% to approximately 17%, with no apparent correspondence between variance and amount of training data. We hypothesize that accuracy is far more dependent on evolutionary diversity (i.e., mutational diversity) within each clade: a phylogenetic group containing species that are more evolutionarily distant

from one another will be inherently harder to classify, using methods based on sequence composition, than one containing more closely-related species.

IMMs and phylogenetic classification

Interpolated Markov models are a form of Markov chain that uses a variable number of states to compute the probability of the next state. IMMs, which are also called variable-order Markov models, were described in detail in the original GLIMMER papers^{11,22}. For our purposes, the main idea is that IMMs can be used to classify sequences based on patterns of DNA distinct to a clade, whether the clade is a species, genus, or higher-level phylogenetic group. During training, the IMM algorithm constructs probability distributions representing observed patterns of nucleotides that characterize each species. The model used by GLIMMER and Phymm captures non-adjacent patterns when necessary; e.g., it can use 8 positions spaced across a window of 12 bases. During classification, we use each IMM as a scoring method: it examines the nucleotides in a given query sequence and outputs a score corresponding to the probability that the query was generated from the same distribution as that used to train the IMM.

Parameter settings for weighted voting in PhymmBL

For PhymmBL, we empirically we determined the combined score using the function

$$Score = IMM + 1.2(4 - \log(E))$$

where *IMM* is the score from the best-matching IMM and *E* is the smallest (best) E-value returned by BLAST. Our IMMs return log-likelihood scores as integers, generally in the range between -500 and -100, with higher scores representing better matches; the log-transformation brought the BLAST scores into this scale. The constant 4 was determined experimentally to be optimal via binary search on small positive integers, and the weight of 1.2 was subsequently determined to be optimal via binary search on values between 1 and 3. The ranges for both searches (integers between 0 and 5 to find the additive constant 4, and values between 1 and 3, in increments of 0.1, to determine the multiplicative weight of 1.2) were established by identifying values at which the predictions of one method completely dominated those of the other. For example: multiplicative weights less than 1 generated combined scores essentially identical to those produced by the IMMs alone, while those greater than 3 generated combined scores which were the same as those produced by BLAST by itself. These settings may only represent a local optimum, but different values of the weights had only a marginal effect on overall accuracy.

Confidence scores

We conducted some preliminary experiments with the goal of correlating raw Phymm and PhymmBL scores with predictive accuracy, in order to establish a function mapping these scores to the probability of generating a correct taxonomic prediction. A smooth, monotonic function mapping score to accuracy was indeed observed for reads 100 bp in length, but reads of different lengths exhibited different score ranges, and maps between score and predictive power across various lengths were not scaled, one to another, in any obvious

(constant or linear) way. While we believe such a relationship can be established, more complex investigation will be required in order to properly determine a closed-form relationship between raw scores and predictive accuracy.

Availability

A one-stop installer for the Phymm/PhymmBL system is available for download at <http://www.cbcb.umd.edu/software/phymm/>, along with a README describing system requirements and configuration.

Copies of all synthetic metagenomic data described in this paper are available for download and are linked off the Phymm/PhymmBL download page.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Thanks to Art Delcher for helpful discussions regarding IMM configuration. Thanks also to our anonymous reviewers, whose comments without doubt helped make this manuscript substantially stronger. This work was supported in part by NIH grants R01-LM006845 and R01-GM083873.

References

1. Handelsman, J.; Tiedje, J.; Alvarez-Cohen, L., et al. *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. The National Academies Press; Washington, DC: 2007.
2. Rondon MR, August PR, Bettermann AD, et al. *Appl Environ Microbiol*. 2000; 66(6):2541. [PubMed: 10831436]
3. Krause L, Diaz NN, Goesmann A, et al. *Nucleic Acids Res*. 2008; 36(7):2230. [PubMed: 18285365]
4. McHardy AC, Martin HG, Tsirigos A, et al. *Nat Methods*. 2007; 4(1):63. [PubMed: 17179938]
5. Kunin V, Copeland A, Lapidus A, et al. *Microbiol Mol Biol Rev*. 2008; 72(4):557. [PubMed: 19052320]
6. Altschul SF, Madden TL, Schaffer AA, et al. *Nucleic Acids Res*. 1997; 25(17):3389. [PubMed: 9254694]
7. Tringe SG, von Mering C, Kobayashi A, et al. *Science*. 2005; 308(5721):554. [PubMed: 15845853]
8. Tito RY, Macmil S, Wiley G, et al. *PLoS ONE*. 2008; 3(11):e3703. [PubMed: 19002248]
9. Huson DH, Auch AF, Qi J, et al. *Genome Res*. 2007; 17(3):377. [PubMed: 17255551]
10. Dinsdale EA, Pantos O, Smriga S, et al. *PLoS ONE*. 2008; 3(2):e1584. [PubMed: 18301735]
11. Salzberg SL, Delcher AL, Kasif S, et al. *Nucleic Acids Res*. 1998; 26(2):544. [PubMed: 9421513]
12. Delcher AL, Bratke KA, Powers EC, et al. *Bioinformatics*. 2007; 23(6):673. [PubMed: 17237039]
13. Pruitt KD, Tatusova T, Maglott DR. *Nucleic Acids Res*. 2007; 35(Database issue):D61. [PubMed: 17130148]
14. Tyson GW, Chapman J, Hugenholtz P, et al. *Nature*. 2004; 428(6978):37. [PubMed: 14961025]
15. Bock, E.; Wagner, M. *The Prokaryotes*. Dworkin, M.; Falkow, S., editors. Vol. 3. 2006.
16. Chapus C, Dufraigne C, Edwards S, et al. *BMC Evol Biol*. 2005; 5:63. [PubMed: 16280081]
17. Manichanh C, Chapple CE, Frangeul L, et al. *Nucleic Acids Res*. 2008; 36(16):5180. [PubMed: 18682527]
18. Mavromatis K, Ivanova N, Barry K, et al. *Nat Methods*. 2007; 4(6):495. [PubMed: 17468765]
19. White JR, Roberts M, Yorke JA, et al. *Bioinformatics*. 2008; 24(4):462. [PubMed: 18202027]

20. Delcher AL, Salzberg SL, Phillippy AM. Curr Protoc Bioinformatics. 2003; 10(10):3. [PubMed: 18428693]
21. Karlin S, Burge C. Trends Genet. 1995; 11(7):283. [PubMed: 7482779]
22. Delcher AL, Harmon D, Kasif S, et al. Nucleic Acids Res. 1999; 27(23):4636. [PubMed: 10556321]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1. Percent accuracy of Phymm, with species-level matches masked, for read lengths from 100–1000 bp

Colored dots show classification accuracy reported for PhyloPythia at 1000 bp for genus-through phylum-level predictions, and for CARMA at 100 bp (as a percentage of the entire input data set) for genus- and phylum-level predictions.

Figure 2. PhymmBL's phylum-level population characterization of the AMD data using (A) the RefSeq-generated IMM plus IMM generated from the draft genomes of the three dominant species in the AMD set, and (B) the RefSeq-generated IMM on their own.

Figure 3. PhymmBL's species-level population characterization of the AMD data using the RefSeq-generated IMMs plus IMMs generated from the draft genomes of the three dominant species in the AMD set.

Table 1

Comparison of performance accuracy, with same-species matches masked, for 1,000-bp reads, showing all three methods described in this study alongside PhyloPythia.

Query Length	Phymm	BLAST	PhymmBL	PhyloPythia
Genus	71.1	73.8	78.4	7.1
Family	77.5	79.2	84.8	–
Order	80.6	80.8	86.9	25.1
Class	85.4	84.1	90.6	30.8
Phylum	89.8	88.0	93.8	50.3

Accuracy is measured as the percentage of all reads for which each method produced the correct phylogenetic label. We performed this set of experiments once for each read length. All other experiments with our synthetic metagenome data set were performed only for 100-bp reads; each of these experiments was repeated 10 times. Results for experiments at all levels of the phylogeny, along with standard deviations for each result, are provided in Supplementary Tables 4–6.