Outlier Detection in BLAST Hits*

Nidhi Shah[1], Stephen F. Altschul[2], and Mihai Pop[3]

1   Department of Computer Science and Center for Bioinformatics and
     Computational Biology, University of Maryland, College Park, MD, USA
     nidhi@cs.umd.edu
2   Computational Biology Branch, NCBI, NLM, NIH, Bethesda, MD, USA
     altschul@ncbi.nlm.nih.gov
3   Department of Computer Science and Center for Bioinformatics and
     Computational Biology, University of Maryland, College Park, MD, USA
     mpop@umiacs.umd.edu

Abstract

An important task in a metagenomic analysis is the assignment of taxonomic labels to sequences
in a sample. Most widely used methods for taxonomy assignment compare a sequence in the
sample to a database of known sequences. Many approaches use the best BLAST hit(s) to assign
the taxonomic label. However, it is known that the best BLAST hit may not always correspond
to the best taxonomic match. An alternative approach involves phylogenetic methods which take
into account alignments and a model of evolution in order to more accurately define the taxonomic
origin of sequences. The similarity-search based methods typically run faster than phylogenetic
methods and work well when the organisms in the sample are well represented in the database. On
the other hand, phylogenetic methods have the capability to identify new organisms in a sample
but are computationally quite expensive. We propose a two-step approach for metagenomic
taxon identification; i.e., use a rapid method that accurately classifies sequences using a reference
database (this is a filtering step) and then use a more complex phylogenetic method for the
sequences that were unclassified in the previous step. In this work, we explore whether and when
using top BLAST hit(s) yields a correct taxonomic label. We develop a method to detect outliers
among BLAST hits in order to separate the phylogenetically most closely related matches from
matches to sequences from more distantly related organisms. We used modified BILD (Bayesian
Integral Log Odds) scores, a multiple-alignment scoring function, to define the outliers within a
subset of top BLAST hits and assign taxonomic labels. We compared the accuracy of our method
to the RDP classifier and show that our method yields fewer misclassifications while properly
classifying organisms that are not present in the database. Finally, we evaluated the use of our
method as a pre-processing step before more expensive phylogenetic analyses (in our case TIPP)
in the context of real 16S rRNA datasets. Our experiments demonstrate the potential of our
method to be a filtering step before using phylogenetic methods.

1998 ACM Subject Classification B.2.4 Algorithms

Keywords and phrases Taxonomy classification, Metagenomics, Sequence alignment, Outlier
detection

---

1   Introduction

One of the goals of metagenomic analyses is to characterize the biological diversity of microbial communities. This is usually achieved by targeted amplicon sequencing of the 16S rRNA gene, either as a whole gene or focused on a hypervariable region within the gene [21]. The 16S rRNA gene is commonly used for this purpose because it is universally found in bacteria and contains a combination of highly conserved and highly variable regions. Advances in sequencing technology, targeted to a specific gene, have generated millions to hundreds of millions of reads per study [6]. Assigning accurate taxonomic labels to these reads is one of the critical steps for downstream analyses.

The most common approach for assigning taxonomic labels to reads involves comparing them to a database of sequences from known organisms. These similarity-based methods typically run rapidly and work well when organisms in the sample are well represented in the database. However, a majority of microorganisms cannot be easily cultured in laboratories, and even if they are culturable, a smaller number have been sequenced. Thus, not all environmental organisms may be represented in the sequence database. This prevents the similarity-based methods from accurately characterizing organisms within a sample that are only distantly related to the sequences in the reference database. Phylogenetic-tree based methods can characterize novel organisms within a sample by statistically modeling the evolutionary processes that generated these sequences [15, 13]. However, such methods incur a high computational cost, limiting their applicability in the context of the large datasets generated in current studies. Ideally, we would want to use similarity-based methods to assign labels to sequences from known organisms, and to use phylogenetic methods to assign labels to sequences from unknown organisms.

We propose a two-step method for taxonomy assignment where we use a rapid assignment method that can accurately assign labels to sequences that are well represented in the database, and then use more complex phylogenetic methods to classify only those sequences unclassified in the first step. In this work, we study whether and when a method can assign accurate taxonomic labels using a similarity search of a reference database. We employ BLAST because it is one of the most widely used similarity search methods [1]. However, it has been shown that the best BLAST hit may not always provide the correct taxonomic label [11]. Most taxonomic-assignment methods utilizing BLAST employ ad-hoc techniques such as recording the consensus label among the top five hits, or using a threshold based on E-value, percent identity, or bit-score [20, 22, 16, 8]. Here we propose an alternative approach for detecting whether and when the top BLAST hits yield correct taxonomic labels. We model the problem of separating phylogenetically correct matches from matches to sequences from similar but phylogenetically more distant organisms as a problem of outlier detection among BLAST hits. Our preliminary results involving simulated and real metagenomic datasets demonstrate the potential of employing our method as a filtering step before using phylogenetic methods.

2   Background

2.1   Taxonomy assignment using BLAST

Several metagenomic analyses use BLAST to assign taxonomic labels to uncharacterized reads in a sample [20, 22, 16]. BLAST is a sequence similarity search tool, and it calculates an E-value and a bit-score to assess the quality of each match. An E-value represents the number of hits of equal or greater score expected to arise by chance. A bit-score can be

understood as representing the size of the space one would need to search in order to find as strong a match by chance. However all 16S sequences are related, and therefore these scores, derived from a model of random sequences, do not provide simple information for separating sequences from different phylogenetic categories.

2.2 BILD scores for multiple sequence alignment

Multiple sequence alignments employ scoring functions to assess the quality of columns of aligned letters. Such functions have included Sum-of-the-Pairs (SP) scores [14], entropy scores [19], tree scores [17, 18] and the recently developed Bayesian Integral Log-Odds (BILD) score [2, 3]. For local pairwise alignment, substitution scores are implicitly of log-odds form [10]. BILD scores extend the log-odds formalism to multiple sequence alignments. They may be used in numerous contexts such as the construction of hidden Markov model profiles, the automated selection of optimal motifs, and the selection of insertion and deletion locations, and they can inform the decision of whether to include a sequence in a multiple sequence alignment. BILD scores can also be used to classify related sequences into subclasses, as we describe below.

3 Methods

Broadly, our approach constructs a multiple alignment from all the top hits obtained by comparing a query sequence to a database. We use BILD scores to determine whether the multiple alignment can be split into two groups that model the data better than does a single group. In essence, we find a subset of the sequences that are more closely related to one another and to the query than to the rest of the sequences in the multiple alignment. When there is no such subset i.e. when the single alignment models the data better, we leave the query unclassified and such a query sequence is then classified in the second step by a phylogenetic method.

3.1 Processing query sequences

Let S be the set of sequences in the reference database, each with a taxonomic label, and Q be a set of uncharacterized reads (i.e. query sequences). We first align each sequence in Q to sequences in S using BLAST (*-max_target_seqs 100 -outfmt 5 -task megablast*). For each $q \in Q$, we construct the ordered set S_q that contains the segments yielding the top 100 bit-scores, in decreasing order of their score. We discard all segments $l \in S_q$ where the BLAST alignment of q and l covers $\leq 90\%$ of q. We use the BLAST-generated local alignments involving q to impose a multiple alignment (M^q) on the sequences in $q \cup S_q$. Where several segments in S_q involves insertions at the identical location in q, we align these insertions to one another by left justification.

3.2 Scoring for Multiple Sequence Alignments and Cuts

We base our score for a multiple alignment (M^q) on the Bayesian Integral Log-Odds (BILD) scores described in [2]. For each alignment column, we take the prior for the nucleotide probabilities to be a Dirichlet distribution with parameters $\vec{\alpha}$, and define $\alpha^* = \sum_{k=1}^{4} \alpha_k$. (Here, we always use Jeffreys' prior [9], for which all $\alpha_k = 0.5$, and $\alpha^* = 2$.) For the jth column M_j^q of the alignment and ignoring null characters, the log-probability of observing
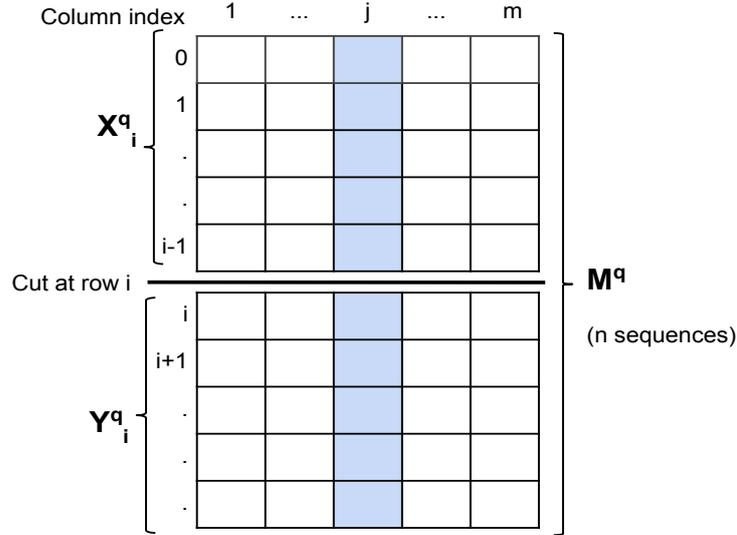
■ **Figure 1** An example of how a cut divides an MSA into two disjoint groups.

the particular collection of c_j^* nucleotides, with count vector $\vec{c_j}$, is then given by

$$L(M_j^q) = \log \left[\frac{\Gamma(\alpha^*)}{\Gamma(\alpha^* + c_j^*)} \prod_{k=1}^{4} \frac{\Gamma(\alpha_k + c_{jk})}{\Gamma(\alpha_k)} \right].$$

Here, Γ is a *gamma* function. As suggested in [2], the log-odds score for preferring a cut, at row i, of the column M_j^q into the two sub-columns X_{ji}^q and Y_{ji}^q, as illustrated in Figure 1, is given by

$$V_{ji}^q = L(X_{ji}^q) + L(Y_{ji}^q) - L(M_j^q). \tag{1}$$

Taking all columns into account, the log-odds score for preferring a cut at row i is simply formula 1 summed over all columns. However, we have found it useful to give greater weight to columns with greater diversity. Thus we adopt the score V_i^q for a cut at row i given by the formula

$$V_i^q = \sum_{j=1}^{m} e_j^a V_{ji}^q,$$

where M^q has m columns, $e_j = -\sum_{k=1}^{4} (c_{jk}/c_j^*) \log_4(c_{jk}/c_j^*)$ is the entropy (base 4) of column j, and a is an arbitrary positive parameter. Note that, using this formula, perfectly conserved columns have entropy 0 and thus weight 0, whereas columns with uniform nucleotide usage have entropy 1 and thus weight 1. We have found, by experimentation, that a useful value for the parameter a is 2.7.

3.3 Outlier detection and taxonomy assignment

We are interested in finding the phylogenetically most closely related matches in the database to the query sequence q. We proceed by computing V_i^q for cuts with increasing i, from $i = 0$, and identify first i' for which $V_{i'}^q \geq 0$, $V_{i'}^q > V_{(i'-1)}^q$, and $V_{i'}^q > V_{(i'+1)}^q$. In other words, we
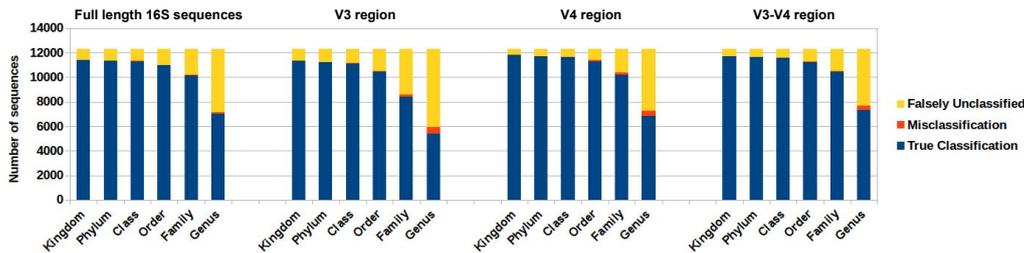
Figure 2 Leave-one-sequence-out validation of our outlier method using a simulated 16S rRNA dataset (RTS) for full-length, V3, V4, and V3-V4 regions.

find the first peak among those scores that imply the data are better explained by a split alignment. (Scores below zero favor a single alignment.) The first $i' - 1$ sequences from S_q we take as forming an outlier set $O_q = S_q[1 : i' - 1]$ for q. We extract the taxonomic labels of all sequences in O_q and assign the lowest common ancestor (LCA; [8]) of these labels to q. In the case when scores favor a single alignment, we leave the query sequence unclassified. The unclassified query sequences then should be classified, in step two of a two-step process, using a phylogenetic method.

4   Evaluation

4.1   Datasets

We used the RDP 16S rRNA gene v16 dataset (RTS), which has taxonomy annotated for each of its 13,212 sequences [5], considering only the 12,320 sequences that had taxonomic labels for all six levels - Kingdom, Phylum, Class, Order, Family, and Genus. These sequences belong to 2,320 genera with, on average, 6 sequences per genus. To evaluate our outlier detection method, we compared taxonomic labels assigned to query sequences by our method to their true labels as given in RTS. First, we used V-Xtractor with default parameters to extract the V3, V4 and V3-V4 hypervariable regions of the sequences [7]. We then used these V3 (SIM-2), V4 (SIM-3), V3-V4 (SIM-4) and full (SIM-1) sequences as query datasets and RTS sequences as a reference database. We also used a real metagenomic dataset (Dataset-1) to study the effectiveness of our method in actual practice. Dataset-1 has 58,108 sequences from the V1-V2 hypervariable region.

4.2   Leave-one-out validation

In the RTS simulated dataset, we know true taxonomic labels for all query sequences. For each taxonomic level, we compare the taxonomic labels assigned by our method to the true labels to find the number of queries that are correctly classified, misclassified or falsely unclassified. To identify correctly classified query sequences at each level, we compare, for all query sequences, the taxonomic labels assigned by our method to the true taxonomic label at that level. If the label assigned to a query by our method matches its true label, or if our method leaves the query sequence unassigned when there are no other sequences in the database with its particular label, we consider the query sequence as properly classified. For each taxonomic level, we consider misclassified those query sequences for which the assigned taxonomic label does not match the true label. We also consider falsely unclassified those sequences that were not assigned a taxonomic label at a particular level when the true label existed independently in the database.

(a)

(b)
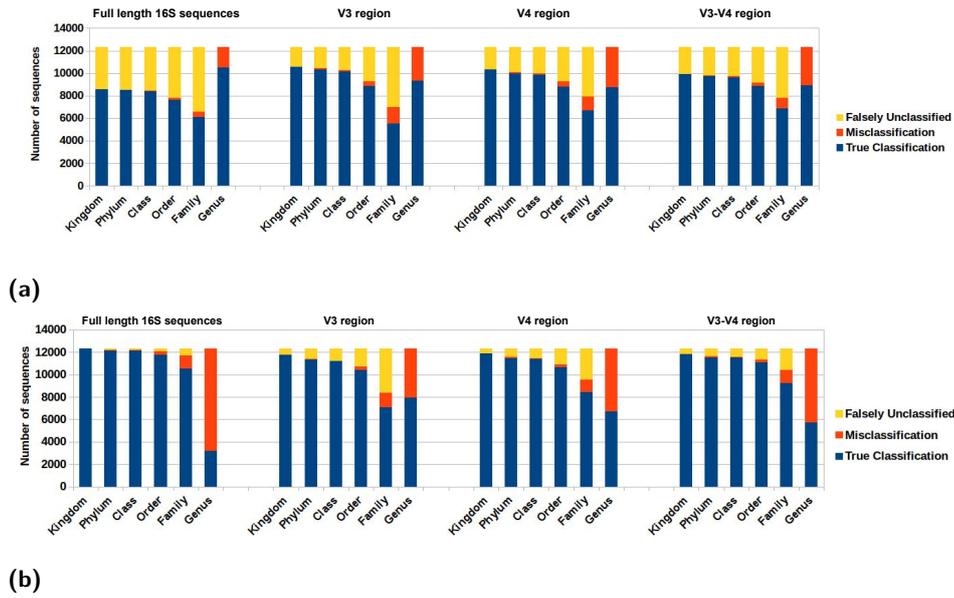
Figure 3 (a) Leave-one-genus-out validation of our outlier method using a simulated 16S rRNA dataset (RTS) for full-length, V3, V4, and V3-V4 regions. (b) Leave-one-genus-out validation of the RDP classifier on same 16S rRNA datasets.

Figure 2 shows the number of correctly classified, misclassified and falsely unclassified sequences calculated by leave-one-out cross-validation, where we assign a taxonomic label to a query sequence (full or hypervariable region) after removing its associated sequence from the database. For all query datasets, our method rarely misclassified at all taxonomic levels, generally assigned correct labels at higher levels, but tended not to assign labels at lower levels. This may be because our method uses the LCA of taxonomic labels of outlier sequences. When there are closely related sequences in the database, our method chooses to be conservative by not assigning labels at lower taxonomic levels.

To study the effectiveness of our method in classifying sequences with taxonomy unrepresented in the database, we performed genus-level leave-one-out cross-validation. Specifically, for each query, we removed all sequences from the database belonging to the same genus, and assigned taxonomic labels with our method and the RDP classifier [23]. We ran the RDP classifier using the QIIME [4] pipeline with the default confidence threshold of 80%. We calculated the number of queries that were correctly classified, misclassified and falsely unclassified as explained above. Figures 3a and 3b show results for our method and RDP respectively. Because the genus to which a query sequence belongs is never present in the database, any label assigned at genus level will result in a misclassification error, and no assignment will result in correct classification. We observed that for higher taxonomic levels (down to Order) RDP and our method have comparable misclassification rates. However, at the Family and Genus levels, our method has a lower misclassification rate. For all datasets, RDP misclassified more query sequences at the Genus level than did our method. This is primarily because RDP aggressively tries to classify as many sequences as it can, whereas our method prefers to classify only when it can do so accurately, leaving other sequences to be dealt with later by a phylogenetic method. This experiment shows that even when sequences from the same genus as the query are absent from the database, our method has high precision and makes few mistakes.
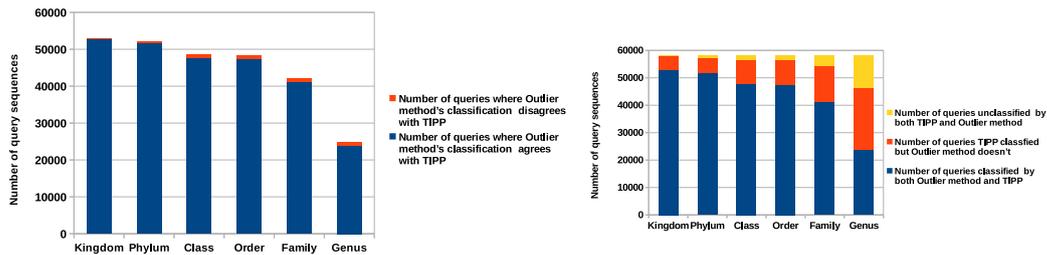
(a) Number of query sequences for which our method's classification agrees with TIPP's classification

(b) Number of query sequences classified by our method and TIPP vs. unclassfied by both

■ **Figure 4** Evaluation of our outlier method using TIPP on a real metagenomic dataset.
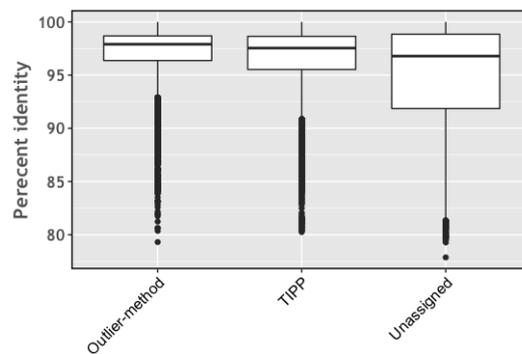
■ **Figure 5** Box plot of percent identity of the best BLAST hit for all query sequences that were assigned label at genus level by our method and TIPP vs. queries that remained unassigned by both methods.

4.3   Validation using Phylogenetic-tree based assignment

To study the effectiveness of our outlier detection method in a realistic setting, we tested it on a real metagenomic dataset. Since we do not know the true taxonomic label for all query sequences, we compared our results with those produced by TIPP [15], a phylogenetic-tree based taxonomic assignment method. We used the RDP 2014 16S reference database for both methods. In this dataset, there were 58,108 query sequences for which our method assigned 41,256 sequences at the Family level or below. Figure 4a shows that our method has a high precision for all taxonomic levels. Also, Figure 4b suggests that using our outlier method to make taxonomic assignments (at least down to the Family level) can significantly reduce the workload of a phylogenetic-tree based method like TIPP. To classify 58,108 query sequences, our method required about 29 CPU hours (including BLAST time), whereas for the same dataset TIPP needed about 300 CPU hours. This shows the potential of our method as a filtering step before using phylogenetic-tree based methods. About 11,000 sequences remained unclassified by both TIPP and our method, and we investigated whether the best BLAST hit's percent identity correlates with the ability of these programs to make classifications; see Figure 5. Unfortunately, there is no clear percent-identity cutoff one can employ to recognize sequences that will remain unassigned by both methods, although a large number of the unassigned sequences have low similarity to the nearest database sequence.
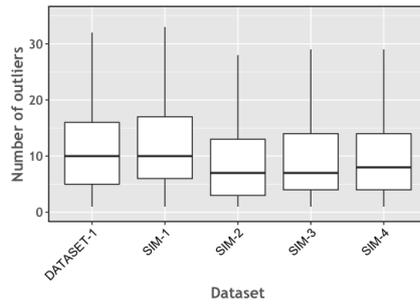
Figure 6 Box plot showing the variation in the number of outliers detected per query sequence in DATASET-1, SIM-1, SIM-2, SIM-3 and SIM-4.
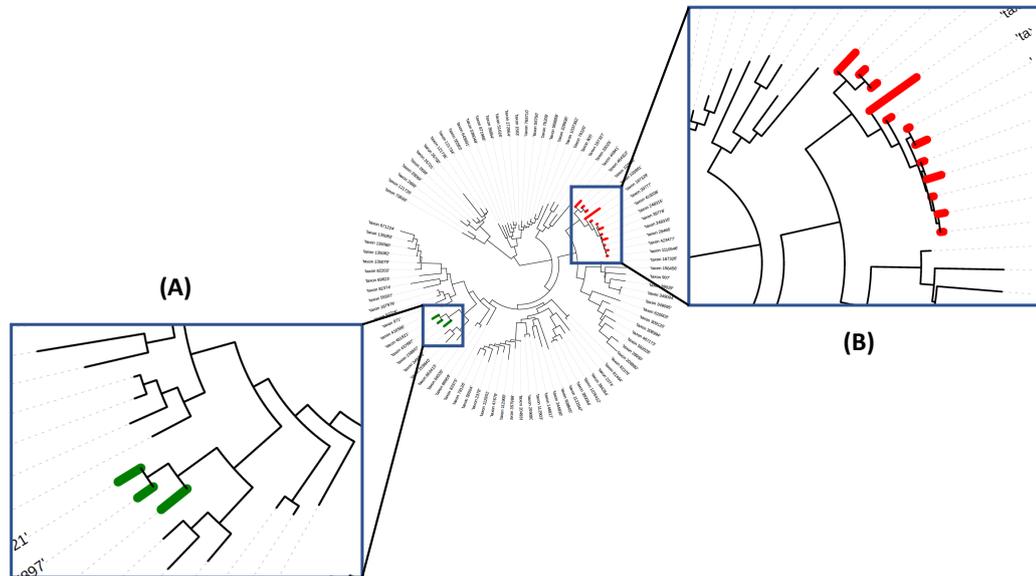
Figure 7 Phylogenetic tree showing outliers detected for two example query sequences.

4.4 Distribution of outliers

Since prior approaches restrict the analysis to just a fixed number of top hits, we evaluated the number of outliers proposed by our method. As seen in Figure 6, the number of outliers has large variance, so a single cutoff (say, the best or top five BLAST hits) will not identify all phylogenetically related matches from the database. In this case, we relied on data for which the true taxonomic label is not known. To validate whether the set of outliers detected by our method is reasonable, and to better understand the performance of our approach, we evaluated the placement of the outlier sequences within a phylogenetic tree of the database. For this, we used the phylogenetic tree for the RDP 2014 database that was bundled in the TIPP reference package, and used the Interactive Tree Of Life web tool to visualize outliers [12]. In general, we noticed that the outliers are grouped close to each other in the phylogenetic tree (see examples in Figure 7), suggesting that our method produces reasonable results. This analysis also revealed insights into the resolution level of the annotations provided by our method. When the outlier sequences cluster tightly within the phylogeny (Figure 7A), a reliable classification can be made at a low taxonomic level. When the outliers

are distributed along a broader section of the tree (Figure 7B), the classification can only be made at a higher taxonomic levels.

5 Conclusion and Discussion

We propose a two-step approach for taxonomic assignment, in which we gain as much information as we reliably can from BLAST output before using computationally expensive phylogenetic-tree based methods on sequences that are difficult to classify. In this paper, we developed an outlier detection method for taxonomy assignment using BLAST hits that separates phylogenetically correct matches from matches to sequences from similar but phylogenetically more distant organisms. This method can thus be used for step one of a two-step approach, to identify sequences that can be assigned accurate labels using just a BLAST search of a reference database.

Because all 16S rRNA sequences are related, statistics like BLAST's E-value or bit-score do not provide ready information for separating sequences from different phylogenetic categories. Our experiments show also that there isn't any single cutoff that can be used to select BLAST hits for correctly assigning taxonomic labels. We have experimented with finding outliers using bit-score distributions, but found they provided insufficient information to detect phylogenetically correct matches (data not shown). Our experiments also show that although the percent identity of its best BLAST hit is correlated to a sequence's being assigned a taxonomic label, no particular percent-identity cutoff can separate those sequences that can be classified from those that cannot. This has motivated our development of a BILD-score based method to identify when the top BLAST hits will yield accurate taxonomic labels.

Because our method is used as a filtering step, we seek to accurately classify as many query sequences as possible while making few misclassifications. The sequences that we leave unclassified are then to be handled by a phylogenetic method. Our results on simulated and real 16S rRNA metagenomic datasets show that our method has high precision at all taxonomic levels, assigning correct labels at higher levels to a majority of sequences, and that it is computationally efficient compared to phylogenetic-tree based taxonomic assignment methods. This demonstrates the promise of a two-step taxonomic assignment approach, using our method as a filtering step.

In the future, we plan to study sequences that were classified correctly by phylogenetic methods but not by ours, to gain insight for possible improvements. We also plan to study the effectiveness of restricting phylogenetic-tree based methods to the subtree spanned by our method's outliers. Finally, note that our method was developed for and tested on 16S rRNA data, and is not applicable as it stands to whole genome sequencing (WGS) datasets. However, the idea of using a two-step approach for taxonomy assignment in WGS datasets is an interesting avenue for research.

References

1 Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

2 Stephen F. Altschul, John C. Wootton, Elena Zaslavsky, and Yi-Kuo Yu. The construction and use of log-odds substitution scores for multiple sequence alignment. *PLoS Comput Biol*, 6(7):e1000852, 2010.

3   Michael Brown, Richard Hughey, Anders Krogh, I. Saira Mian, Kimmen Sjölander, and David Haussler. Using Dirichlet mixture priors to derive hidden markov models for protein families. In *Ismb*, volume 1, pages 47–55, 1993.

4   J. Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D. Bushman, Elizabeth K. Costello, Noah Fierer, Antonio Gonzalez Peña, Julia K. Goodrich, Jeffrey I. Gordon, et al. QIIME allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5):335–336, 2010.

5   James R. Cole, Qiong Wang, Jordan A. Fish, Benli Chai, Donna M. McGarrell, Yanni Sun, C. Titus Brown, Andrea Porras-Alfaro, Cheryl R. Kuske, and James M. Tiedje. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic acids research*, page gkt1244, 2013.

6   Jack A. Gilbert, Janet K. Jansson, and Rob Knight. The Earth Microbiome project: successes and aspirations. *BMC biology*, 12(1):69, 2014.

7   Martin Hartmann, Charles G. Howes, Kessy Abarenkov, William W. Mohn, and R. Henrik Nilsson. V-Xtractor: an open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16s/18s) ribosomal RNA gene sequences. *Journal of Microbiological Methods*, 83(2):250–253, 2010.

8   Daniel H. Huson, Alexander F. Auch, Ji Qi, and Stephan C. Schuster. MEGAN analysis of metagenomic data. *Genome research*, 17(3):377–386, 2007.

9   Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London a: mathematical, physical and engineering sciences*, 186(1007):453–461, 1946.

10   Samuel Karlin and Stephen F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences*, 87(6):2264–2268, 1990.

11   Liisa B. Koski and G. Brian Golding. The closest BLAST hit is often not the nearest neighbor. *Journal of molecular evolution*, 52(6):540–542, 2001.

12   Ivica Letunic and Peer Bork. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1):127–128, 2007.

13   Frederick A. Matsen, Robin B. Kodner, and E. Virginia Armbrust. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC bioinformatics*, 11(1):538, 2010.

14   Michio Murata, Jane S. Richardson, and Joel L. Sussman. Simultaneous comparison of three protein sequences. *Proceedings of the National Academy of Sciences*, 82(10):3073–3077, 1985.

15   Nam-phuong Nguyen, Siavash Mirarab, Bo Liu, Mihai Pop, and Tandy Warnow. TIPP: taxonomic identification and phylogenetic profiling. *Bioinformatics*, 30(24):3548–3555, 2014.

16   Mihai Pop, Alan W. Walker, Joseph Paulson, Brianna Lindsay, Martin Antonio, M. Anowar Hossain, Joseph Oundo, Boubou Tamboura, Volker Mai, Irina Astrovskaya, et al. Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition. *Genome biology*, 15(6):R76, 2014.

17   David Sankoff. Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, 28(1):35–42, 1975.

18   David Sankoff and Robert J. Cedergren. Simultaneous comparison of three or more sequences related by a tree. *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison/edited by David Sankoff and Joseph B. Krustal*, 1983.

19   Thomas D. Schneider, Gary D. Stormo, Larry Gold, and Andrzej Ehrenfeucht. Information content of binding sites on nucleotide sequences. *Journal of molecular biology*, 188(3):415–431, 1986.

20   Raúl Y. Tito, Simone Macmil, Graham Wiley, Fares Najar, Lauren Cleeland, Chunmei Qu, Ping Wang, Frederic Romagne, Sylvain Leonard, Agustín Jiménez Ruiz, et al. Phylotyping and functional analysis of two ancient human microbiomes. *PLoS One*, 3(11):e3703, 2008.

21   Susannah G. Tringe and Philip Hugenholtz. A renaissance for the pioneering 16S rRNA gene. *Current opinion in microbiology*, 11(5):442–446, 2008.

22   Susannah Green Tringe, Christian Von Mering, Arthur Kobayashi, Asaf A. Salamov, Kevin Chen, Hwai W. Chang, Mircea Podar, Jay M. Short, Eric J. Mathur, John C. Detter, et al. Comparative metagenomics of microbial communities. *Science*, 308(5721):554–557, 2005.

23   Qiong Wang, George M. Garrity, James M. Tiedje, and James R. Cole. Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, 73(16):5261–5267, 2007.