Improving the accuracy of taxonomic classification for identifying taxa in microbiome samples

Adithya Murali[1], Aniruddha Bhargava[2], & **Erik S. Wright**[3]
[1]Department of Computer Sciences, University of Wisconsin–Madison
[2]Robotics, Amazon, Inc.
[3]Department of Biomedical Informatics, University of Pittsburgh

Contact info: eswright@pitt.edu
WrightLabScience.com

Introduction

It has become increasingly clear that the microbiome is an essential component of human and ecosystem health. Microbiome studies frequently involve sequencing a taxonomic marker, such as the 16S rRNA or ITS, to identify the microorganisms that are present in a sample of interest. We have developed a new method, named IDTAXA, for taxonomic classification of marker gene sequences that exhibits a substantially lower error rate than previous approaches.

Results

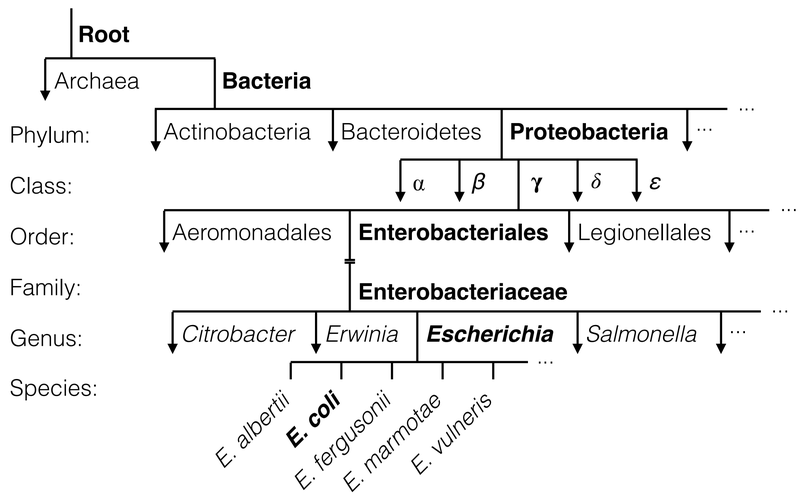
IDTAXA avoids misclassifying sequences belonging to novel taxonomic groups that are not represented in existing taxonomic databases, which is the predominant type of error made by current classifiers. For example, the popular RDP Classifier incorrectly assigns 26.0% of novel 16S rRNA sequences to an existing taxonomic group when the organism actually belongs to a novel taxonomic group. In contrast, IDTAXA only incorrectly classifies 13.6% of such sequences, while correspondingly improving on the fraction of sequences correctly classified to known taxonomic groups.

Conclusions

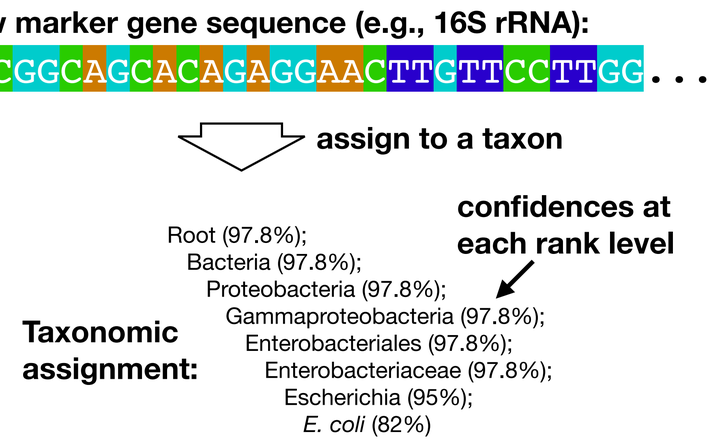
IDTAXA exhibits substantially lower error rates when classifying novel sequences that are unrepresented in the reference taxonomy. This considerably alters the interpretation of microbiome data because many microbial communities contain a large fraction of previously unknown microorganisms that are not yet represented in taxonomic databases. Collectively, these improvements often lead to substantially different classifications on real microbiome data, which may considerably alter its interpretation.

1. Classifying marker gene (e.g., 16S) sequences into a taxonomy

a. Given a reference taxonomy with sequence representatives

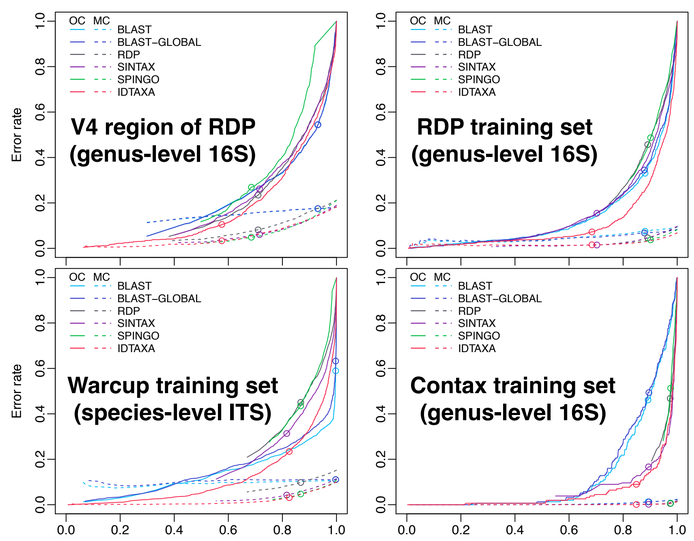


b. The goal is to predict the taxon of a microbiome sequence

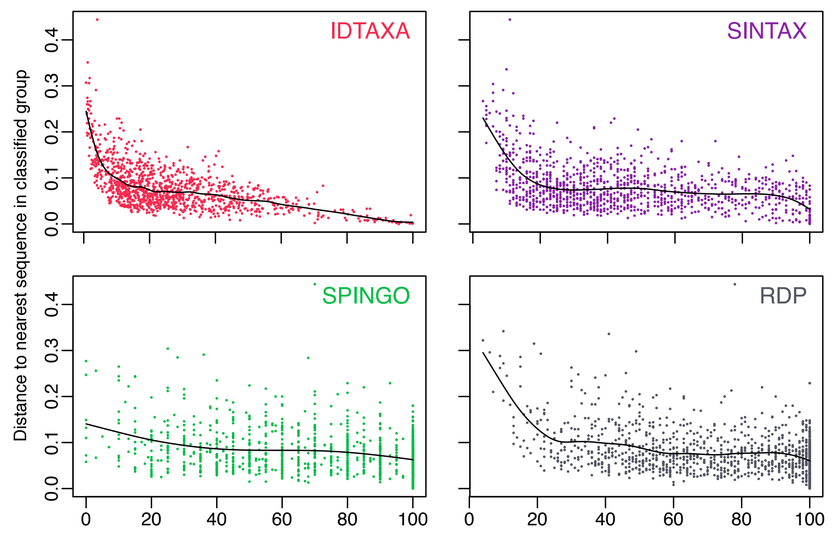


2. How well does the IDTAXA algorithm work for taxonomic assignment?

a. Comparing performance with different reference taxonomies

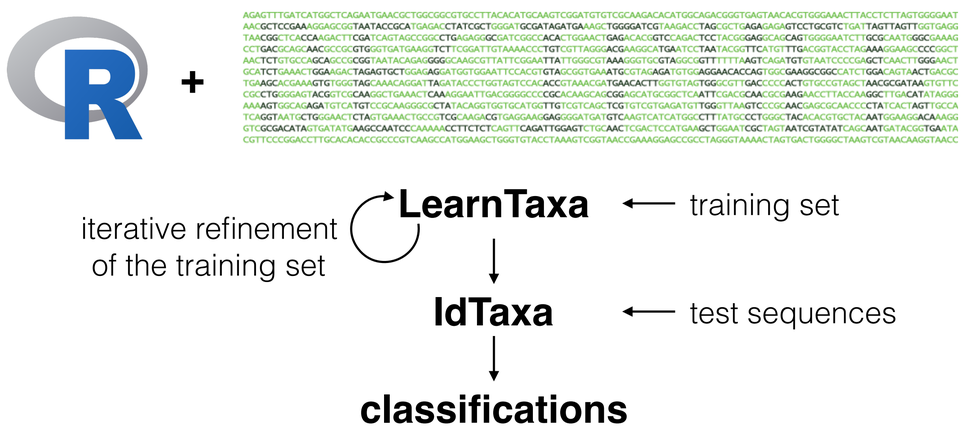


b. IDTAXA's confidence is more correlated with percent identity



3. IDTAXA is available as a part of the R package DECIPHER or online

a. How to classify new sequences with the IDTAXA algorithm:



b. Visit http://DECIPHER.codes