HiTaC: Hierarchical Taxonomic Classification of Fungal ITS Sequences

Fábio M Miranda1,2,3, Vasco AC Azevedo3, Bernhard Y Renard,1 Vitor C Piro,1,2,*1 and Rommel TJ Ramos3,4,†

1Data Analytics and Computational Statistics, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam, 14482 Potsdam, Germany,
2Department of Mathematics and Computer Science, Freie Universität Berlin, 14195 Berlin, Germany,
3Institute of Biological Sciences, Universidade Federal de Minas Gerais, 31270–901 Belo Horizonte, Brazil and
4Institute of Biological Sciences, Universidade Federal do Pará, 66075–110 Belém, Brazil

* To whom correspondence should be addressed.
† These authors share senior co-authorship.

Abstract

Motivation: Fungi play a key role in several important ecological functions, ranging from organic matter decomposition to symbiotic associations with plants. Moreover, fungi naturally inhabit the human microbiome and can be causative agents of human infections. An accurate and robust method for fungal ITS classification is not only desired for the purpose of better diversity estimation, but it can also help us gain a deeper insight into the dynamics of environmental communities and ultimately comprehend whether the abundance of certain species correlate with health and disease. Although many methods have been proposed for taxonomic classification, to the best of our knowledge, none of them consider the taxonomic tree hierarchy when building their models. This in turn, leads to lower generalization power and higher risk of committing classification errors.

Results: Here we introduce HiTaC, a robust hierarchical machine learning model for accurate ITS classification, which requires a small amount of data for training and is able to handle imbalanced datasets. HiTaC was thoroughly evaluated with the established TAXXI benchmark, characterizing its ability to correctly classify fungal ITS sequences of varying lengths and a range of identity differences between the training and test data. HiTaC outperforms state-of-the-art methods when trained over noisy data, consistently achieving higher accuracy and sensitivity across different taxonomic ranks, improving accuracy by 5.5 and sensitivity by 6.9 percentage points over top methods in the most noisy dataset available on TAXXI.

Availability: HiTaC is publicly available and freely accessible at https://github.com/mirand863/hitac.

Contact: vitor.piro@fu-berlin.de

1 Introduction

In the last decades, there has been a surge of interest in characterizing the mycoflora of communities. This is due to the fact that fungi are key organisms in several important ecological functions, ranging from nutrient recycling to mycorrhizal associations and decomposition of wood and litter (Hawksworth, 2001). Moreover, fungi naturally inhabit the human microbiome and can be causative agents of infections that may lead to death (Brown et al., 2012; Hoggard et al., 2018), yet very little is known about the fungal microbiota of healthy individuals when compared to the bacterial microbiome (Huffnagle and Noverr, 2013). Furthermore, identifying the presence of fungi can be challenging, as they rarely form structures visible to the naked eye and similar structures are frequently composed of several distinct species (Nilsson et al., 2009). Hence, it became common practice to use DNA sequencing in addition to morphological studies in contemporary mycology (Hibbett, 2007).

One of the approaches that can be employed is whole-genome shotgun sequencing (WGS), where the purpose is to obtain the genetic material of all organisms present in a given microbiome simultaneously (Venter et al., 2004). Nonetheless, in studies where the end goal is only to estimate the diversity or discover the evolutionary distance of organisms in a microbiome, it is cheaper and more convenient to sequence only genetic markers such as 16S rRNA, which is well conserved in all bacteria (Fuhrman, 2012) and also widely used to classify and identify archaea (Kim and Chun, 2014). In mycology, the internal transcribed spacer (ITS) region was adopted as the universal marker for classifying fungi (Schoch et al., 2012). Comparatively, ITS has far more sequences available than fungal WGS, making it a viable cheaper alternative in mycology.

© The Authors 2023.
The main characteristic of 16S rRNA that makes it a good genetic marker is the presence of nine hypervariable regions V1–V9, which are flanked by conserved regions that can be used to amplify target sequences using universal primers (Chakravorty et al., 2007). However, 16S rRNA has fewer hypervariable domains in fungi, and consequently it is more appropriate to use the ITS regions which are composed by 18S–26S rRNAs (Schoch et al., 2012). Additionally, the ease with which ITS is amplified makes it an appealing target for sequencing samples from environments as challenging as soil and wood, where the initial quantity and quality of DNA is low (Nilsson et al., 2009).

From a computational perspective, the taxonomic classification of ITS fungal sequences can be performed by using either homology-based or prediction-based techniques. Homology-based approaches require the alignment of a query sequence against all sequences available in a reference database (Jonassen et al., 2002) such as Silva (Pruesse et al., 2007), Ribosomal Database Project (RDP) (Cole et al., 2008) and UNITE (Nilsson et al., 2018). There are several homology-based software currently available for the analysis of environmental samples, such as EzTaxon (Chun et al., 2007), MG-RAST (Meyer et al., 2008), PyNAST (Caporaso et al., 2009), UCLUST (Edgar, 2010), QIIME 2 (Bolyen et al., 2019), MEGAN (Mitra et al., 2011) and FindFungi (Donovan et al., 2018).

However, a major limitation of homology-based approaches is the reliance on the availability of homologous sequences in reference databases (Gupta et al., 2009). There are several homology-based supervised learning classifiers from scikit-learn (Pedregosa et al., 2018b), e.g., Random Forest, Support Vector Machines (SVM) and Gradient Boosting. Q2_SK is flexible and allows the selection of features and classifiers. Despite all these advances made in prediction-based taxonomic classification, to the best of our knowledge, all the tools available perform flat classification, which means that they predict only leaf nodes and do not fully explore the hierarchical structure of the taxonomic tree when building their models (see Fig. 3 for a depiction of the hierarchical structure of the taxonomic tree). The main disadvantage of this approach is that the information about parent-child class relationships in the hierarchical structure is not explored, which may cause flat classifiers to commit more errors than hierarchical approaches (Silla and Freitas, 2011). Therefore, we developed a new hierarchical taxonomic classifier for fungal ITS sequences, denominated HiTaC, which explores the hierarchical structure of the taxonomic tree and improves the accuracy and sensitivity of fungal ITS predictions. Furthermore, HiTaC can be easily installed and is compatible with QIIME2, facilitating its adoption in ITS sequence analysis pipelines.

2 Materials and methods

2.1 Algorithm

HiTaC automatically performs feature extraction, building of the hierarchical model, taxonomic classification and reporting of the predictions. Fig. 1 illustrates the algorithmic steps of HiTaC, where it starts by extracting the 6-mer frequencies from both training and test files. Although 6-mer was chosen as the standard value, it can be changed via a command line parameter, however, we did not see much improvement for larger k-mer values that justify the increase in computational complexity. Afterwards, a logistic regression classifier is trained for each parent node in the taxonomic tree. Once the hierarchical model is trained, predictions are made based on the taxonomic tree structure and HiTaC finishes execution reporting the predictions made.

Fig. 1. HiTaC algorithm overview. To compute the k-mer frequency of DNA sequences, we first decompose them into their constituent k-mers. In the example shown in step 01, k = 3, resulting in 5 k-mers. Afterwards, HiTaC computes a matrix containing the k-mer frequency for all training sequences, where each row contains the k-mer frequency for a given sequence and the columns are all possible k-mers for a given k, i.e., all k nucleotide permutations of A, C, G and T. Next, HiTaC trains a logistic regression classifier for each parent node in the taxonomic tree, where the features are the k-mer frequencies computed in step 2 and the labels are the taxonomic annotations of the training sequences. In step 03, only the first 2 levels are shown for brevity. Lastly, HiTaC makes predictions in a top-down approach, based on the taxonomic hierarchy.

HiTaC trains a local classifier for each parent node of the taxonomic tree, using the local hierarchical classifier per parent node implementation provided in the library HiClass (Miranda et al., 2021). For example,
HiTaC: Hierarchical Taxonomic Classification of Fungal ITS Sequences

a classifier in the root node of the taxonomic tree decides whether a given sequence belongs to the phylum Ascomycota or Basidiomycota. If the sequence is classified as Ascomycota, then a child classifier decides whether the sequence belongs to the class Sordariomycetes, Dothideomycetes, or Leotiomycetes, and similarly for the remaining standard taxonomic ranks, namely order, family, genus and species (see Fig. 3 for an example). This hierarchical model was chosen instead of the local hierarchical classifiers per level or per node because it respects the natural relationships between the hierarchical categories, thus avoiding inconsistent predictions across different taxonomic ranks. A detailed review of different hierarchical models and applications is provided by (Silla and Freitas, 2011).

Each parent node in our hierarchical model consists of a logistic regression classifier for each level of the hierarchy, using the same k-mer frequency algorithm introduced previously on Fig. 1. In the example shown in step 01, there are only 7 levels in the hierarchy for simplicity. Afterwards, HiTaC computes the confidence score for all taxonomic ranks assigned to a sample by the local classifier per parent node, then on a bottom-up approach it removes labels that are below a certain threshold. In step 02, supposing that the threshold is 0.7 and that the given sample has a confidence score of 0.2 assigned for the lowest rank, then the final prediction for this sample will only contain the labels Fungi and Ascomycota.

Table 1. List of software available in the TAXXI benchmark (Edgar, 2018), which were used for comparison with HiTaC.

| Method   | Software                | Algorithm reference |
|----------|-------------------------|---------------------|
| BLCA     | BLCA v2.3-alpha         | (Gao et al., 2017)  |
| BTOp     | blastn v2.10.1+         | (Altschul et al., 1990) |
| CT1      | usearch v11.0.667       | Re-implements Q1 (Edgar, 2018) |
| CT2      | usearch v11.0.667       | Re-implements Q2_VS (Edgar, 2018) |
| HiTaC    | HiTaC v2.0.10          | (This paper)         |
| KNN      | mothur v1.48.0         | (Schloss et al., 2009) |
| KTOP     | usearch v11.0.667       | (Edgar, 2018)        |
| Meta2    | MetaXas 2.2.3           | (Bengtsson-Palme et al., 2015) |
| Microclass | microclass v1.2        | (Liland et al., 2017) |
| NBC      | usearch v11.0.667       | Re-implements RDP (Edgar, 2018) |
| Q1       | QIIME v1.9             | (Caporaso et al., 2010) |
| Q2_BLAST | QIIME v2.2022.2        | (Bokulich et al., 2018a) |
| Q2_SK    | QIIME v2.2022.2        | (Bokulich et al., 2018a) |
| Q2_VS    | QIIME v2.2022.2        | (Bokulich et al., 2018a) |
| RDP      | Classifier v2.13       | (Wang et al., 2007)  |
| SINTAX   | usearch v11.0.667       | (Edgar, 2016)        |
| SPINGO   | SPINGO v1.3            | (Allard et al., 2015) |
| TOP      | usearch v11.0.667       | (Edgar, 2018)        |

The TAXXI benchmark (Edgar, 2018) was employed to facilitate the comparison between HiTaC and seventeen other similar software previously available in the benchmark (Table 1), some of which are homology based approaches, while others are alignment free. We relied on this benchmark because it mimics more accurately real world scenarios including novel sequences, as well as its diversity in software, reference databases and use of real environmental data. Furthermore, the TAXXI benchmark assesses the accuracy using cross-validation by identity, a strategy which models the variation in distances between query sequences and the closest entry in reference databases. In this approach, a 90% identity enforces sequences to only belong to species that are absent from the training set, while a 90% identity allows for a mix of present and absent species. In other words, the cross-validation by identity enables the creation of datasets whose lowest taxonomic ranks were nonexistent in the training data while still knowing the ground truth, which allows for a better evaluation on how a classifier behaves when trying to predict novel sequences.
2.2.1 Datasets
We adopted five datasets containing both training and test data, which were previously available in the TAXXI benchmark (Edgar, 2018). These datasets were selected because they were the only ones with taxonomic annotations of fungal ITS sequences up to species level. Although there were other five datasets available in this benchmark possessing fungal ITS sequences, we noticed that they were copies of the same datasets we selected, except that they were missing the species-level annotation in the training datasets and ground truth. Thus, there was no added value in using these other datasets in our evaluation.

The fungal ITS datasets selected from the TAXXI benchmark were based on real environmental data, publicly available on NCBI (Sayers et al., 2012), RDP (Cole et al., 2014), the Warcup fungal ITS training set v2 (WITS) (Deshpande et al., 2016), UNITE (Kõljalg et al., 2005) and other in vivo samples. They were created through cross-validation by identity, hence the identity between the query sequences and the closest entry in a reference database varied on them, ranging from 90% to 100% (Edgar, 2018) (Table 2). In practice, an identity of 90% means that the training data does not contain any of the labels that are in the test dataset at the species level, but most labels are known at the genus level. The known taxa at the species level increase with identity, i.e., all species are present in the training data when the identity is 100%.

Table 2. List of datasets available in the TAXXI benchmark (Edgar, 2018), which were selected for possessing taxonomic annotations of fungal ITS sequences up to species level and were used to compare HiTaC with similar software.

| Name          | Lowest Rank | Identity (%) | # Training Sequences | # Test Sequences |
|---------------|-------------|--------------|----------------------|------------------|
| SP RDP ITS 90 | Species     | 90           | 9336                 | 1935             |
| SP RDP ITS 95 | Species     | 95           | 11439                | 1265             |
| SP RDP ITS 97 | Species     | 97           | 12545                | 1420             |
| SP RDP ITS 99 | Species     | 99           | 10901                | 3945             |
| SP RDP ITS 100| Species     | 100          | 16055                | 16055            |

It is also important to mention that the training datasets are extremely imbalanced. For example, among the seven existing phyla on dataset SP RDP ITS 90, Ascomycota and Basidiomycota contain approximately 52% and 44% of total instances, respectively (Fig. 3). A similar pattern can be observed in the other datasets with higher identities.

2.2.2 Evaluation Metrics
To measure the performance of our model, we adopted all metrics available in the TAXXI benchmark (Edgar, 2018), namely:

- **Accuracy (ACC)** - Accounts for all test sequences predicted by the algorithm with names that are known and/or over-predicted, ranging from 0% (no correct predictions) to 100% (no errors);
- **Misclassification Rate (MCR)** - The percentage of known sequences incorrectly predicted;
- **Over-classification Rate (OCR)** - The percentage of unknown sequences assigned a label;
- **True Positive Rate (TPR)** - The percentage of known sequences correctly predicted;
- **Under-classification Rate (UCR)** - The percentage of known sequences that were not assigned a label.

Defining these metrics in mathematical terms, we have:

\[
\text{ACC} = \frac{\text{True Positive} \times \text{Known Sequences}}{\text{Known Sequences} + \text{Over-classification Errors}}
\]

\[
\text{MCR} = \frac{\text{Misclassification Errors}}{\text{Known Sequences}}
\]

\[
\text{OCR} = \frac{\text{Over-classification Errors}}{\text{Novel Sequences}}
\]

\[
\text{TPR} = \frac{\text{True Positive}}{\text{Known Sequences}}
\]

\[
\text{UCR} = \frac{\text{Under-classification Errors}}{\text{Known Sequences}}
\]

where known sequences have labels in the training dataset and novel sequences do not have labels in the training dataset. A true positive is the number of correctly predicted sequences. Under-classification errors occur when too few ranks are predicted, while over-classification errors happen when too many ranks are predicted. Finally, misclassification errors occur when a known name is incorrectly predicted.

2.2.3 Resources benchmark
We enhanced the TAXXI benchmark by porting it to Snakemake (Köster and Rahmann, 2012) in order to make it more reproducible, and we introduced a resources benchmark. The benchmark scripts are available in our GitHub repository1, and it was computed on a cluster node running GNU/Linux with 512 GB physical memory and 128 cores provided by two AMD EPYC™ 7742 processors, where 8 cores were allocated for each method. The only exception was Q1 because one of its dependencies, uclust, is only provided inside a virtual disk image. Hence, Q1 was executed inside a GNU/Linux virtual machine on a laptop running Windows 10 with 16 GB physical memory and 12 cores provided by an AMD Ryzen™ 5 4600H processor, where 10 GB and 6 cores were allocated.

1 https://github.com/mirand863/hitac/tree/main/benchmark
HiTaC: Hierarchical Taxonomic Classification of Fungal ITS Sequences

3 Results

We comprehensively evaluated the performance of HiTaC at different taxonomic ranks on five datasets containing fungal ITS sequences with identities varying between 90% and 100%. We showed that HiTaC outperformed state-of-the-art taxonomic classifiers at the genus level, challenging it with novel sequences and noisy regions in sequencing data in the datasets SP RDP ITS 90–97. We also evaluated the robustness of HiTaC by testing it at the species level, the most difficult rank that is possible to predict with fungal ITS sequences, on datasets SP RDP ITS 99–100 from the TAXXI benchmark. The results presented here and further taxonomic ranks can be found in our GitHub repository.

We first evaluated the accuracy at the genus and species level on the SP RDP ITS datasets, reporting the accuracy for all methods available in the TAXXI benchmark. As shown in Fig. 4, at the species level HiTaC and BTOP achieved 100% accuracy on the dataset with 100% identity. Furthermore, HiTaC obtained a significantly higher accuracy for the dataset with 90% identity at the genus level, with a 5.5 percentage points improvement over NBC50, the second best third party method for this dataset, while HiTaC’s filter also promoted a slight improvement over NBC50 for the same dataset. HiTaC also achieved top accuracy for the datasets with identities varying between 95–99%. On the other hand, while HiTaC’s filter did not achieve top accuracy for these datasets, it was still quite competitive, ranking just a few positions below the top performing methods and it was planned to reduce the over-classification rate as we will show later.

Similar conclusions can be drawn from Fig. 5, in which we evaluated the true positive rate (TPR) for the top 5 methods for each dataset in the TAXXI benchmark. The trend is that HiTaC achieved a TPR higher or equal to top methods. To illustrate this point, the best true positive rate at genus level for dataset SP RDP ITS 90 was achieved by HiTaC, which sharply increased the TPR by 6.9 percentage points when compared with BTOP. Regarding dataset SP RDP ITS 95 at genus level, HiTaC obtained 1.7 percentage points improvement over Microclass, while for dataset SP RDP ITS 97 at genus level HiTaC surpassed Microclass with a 0.5 increase in percentage points. At species level, HiTaC achieved a 0.7 percentage points gain over BTOP for dataset SP RDP ITS 99, while for dataset SP RDP ITS 100 HiTaC tied with BTOP, obtaining a perfect score of 100% TPR.

On Fig 6, we show the results for under-classification and over-classification rates for dataset SP RDP ITS 90 at genus level. HiTaC, BTOP, KTOP, Microclass and TOP did not leave any known sequences unclassified. KNN, on the other hand, left 60.9% known sequences unclassified. Regarding over-classification, KNN achieved the lowest rate overall at 18.4%. Conversely, HiTaC, BTOP, KTOP, Microclass and TOP

---

2 https://github.com/mirand863/hitac/tree/main/benchmark/results/metrics
scored 100% over-classification rates for most datasets, which means that they are wrongly labeling novel sequences that do not exist in the training datasets. The introduction of the filter proved effective to reduce HiTaC’s over-classification rate, but slightly increased the under-classification rate. An ideal algorithm would be able to classify all known sequences and leave novel sequences unclassified, thus achieving 100% accuracy.

The misclassification rate results are shown in Fig. 7. The methods that predominantly attained the lowest misclassification rates were KNN, NBC80, RDP80 and SINTAX80. In general, HiTaC achieved a median misclassification rate when compared to the other methods, reaching 19.7% for the dataset SP RDP ITS 90 and lower values for the remaining datasets with higher identities, eventually reaching a misclassification rate of 0% for the dataset with 100% identity. The filter implemented in HiTaC successfully reduced the misclassification rate to 10.3% for the dataset SP RDP ITS 90 and even lower values were achieved for the remaining datasets, reaching a misclassification rate as low as 0% for the dataset SP RDP ITS 100.

Lastly, Table 3 displays the results for the resources benchmark for dataset SP RDP ITS 100, at species level. In terms of wall clock time, HiTaC ranked in the 10th position, being slower only when compared with methods implemented in compiled languages such as C and C++. NBC presented a memory limit exceeded error, since its 32 bit version can only use at most 4 GB of RAM.

4 Discussion
Taxonomic classification is the first step in studies targeting the taxonomic analysis of environmental samples and its results are crucial for downstream analysis. In this work we developed a robust hierarchical taxonomic classifier for accurate fungal ITS classification. Our results showed that with the hierarchical machine learning model, HiTaC could achieve significantly better overall performance than state-of-the-art taxonomic classifiers even with imbalanced training data.

We noticed that HiTaC, BTOP, KTOP, Microclass and TOP erroneously classify novel sequences that do not have labels in the training datasets, when query identities are below 100%. For this reason, the predictions made by these methods should be taken with caution. On the other hand, conservative methods such as CT2, Q2_VS, Q2_BLAST, Metaxa2 and KNN have an exceptionally high under-classification rate at the species level, when query identity is quite high in our tests, thus possibly missing the opportunity to classify many sequences already deposited in databases. Both extremes can negatively impact the taxonomic analysis of environmental samples, hence a middle ground would be preferred. Therefore, a filter was introduced in HiTaC in order to reduce the over-classification rate by reporting confidence scores for the predictions made and it can also be set to detect and remove novel or difficult lower ranks that are unlikely to be correct.

In our view, it is challenging to rely on ITS markers alone to classify fungi at the species level, given the current limitations in ITS databases such as low quality sequencing and misidentifications (Badotti et al., 2017). Nevertheless, ITS has been successfully employed to accurately catalogue bacterial communities at (sub)species resolution (Milani et al., 2020). Moreover, from an ecological perspective, it is interesting to use
Table 3. Results obtained by the resources benchmark for the dataset SP RDP ITS 100. HiTaC ranked 10th in terms of processing speed, surpassed mostly by methods based on compiled languages. The filter proved to be slow during the training stage, which placed it among the slowest methods.

| Method     | Time (hh:mm:ss) | Memory (MB) |
|------------|-----------------|-------------|
| CT1        | 00:00:04        | 68.24       |
| KTOP       | 00:00:04        | 60.24       |
| TOP        | 00:00:04        | 66.64       |
| CT2        | 00:00:05        | 71.36       |
| Q1         | 00:00:12        | 336.92      |
| KNN        | 00:00:18        | 84.96       |
| SINTAX     | 00:00:26        | 63.69       |
| SPINGO     | 00:00:31        | 55.35       |
| Microclass | 00:01:00        | 4663.38     |
| HiTaC      | 00:04:58        | 10112.29    |
| BTOp       | 00:10:10        | 51.9        |
| BLCA       | 00:10:54        | 136.14      |
| Q2_SK      | 00:11:44        | 7491.52     |
| RDP        | 00:14:02        | 2887.86     |
| Q2_VS      | 00:20:04        | 693.37      |
| Q2_BLAST   | 00:38:27        | 603.39      |
| Metaxa2    | 04:19:04        | 8983.89     |
| HiTaC_Filter| 07:26:28        | 14817.29    |
| NBC        | -               | Memory Limit Exceeded |

5 Conclusion
In summary, in this study we developed a novel taxonomic classifier, HiTaC, which employs a hierarchical machine learning model for accurate fungal ITS classification. Our algorithm is able to handle imbalanced data and outperforms state-of-the-art classification methods on ITS datasets available in the TAXXI benchmark, consistently achieving higher accuracy and sensitivity across different taxonomic ranks, especially on lower taxonomic levels (genus and species). HiTaC is an open-source software, available at https://github.com/mirand863/hitac.

Acknowledgements
FMM gratefully acknowledges Marcos Augusto dos Santos (UFMG) for initial inspiration, Robert C. Edgar (independent researcher) for providing the scripts and datasets in the TAXXI benchmark, Sven Giese (HPI) for code improvement suggestions and Elizabeth Yiu (HPI) for proofreading.

Funding
This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) under grant number 307584/2018-6.

References
Allard, G. et al. (2015). Spingo: a rapid species-classifier for microbial amplicon sequences. BMC bioinformatics, 16(1), 324.

Altschul, S. F. et al. (1990). Basic local alignment search tool. Journal of molecular biology, 215(3), 403–410.

Badotti, F. et al. (2017). Effectiveness of its and sub-regions as dna barcode markers for the identification of basidiomycota (fungi). BMC microbiology, 17(1), 42.

Bengtsson-Palme, J. et al. (2015). Metaxa2: improved identification and taxonomic classification of small and large subunit rna in metagenomic data. Molecular Ecology Resources, 15(6), 1403–1414.

Bokulich, N. A. et al. (2018a). Optimizing taxonomic classification of marker-gene amplicon sequences with qiime 2’s q2-feature-classifier plugin. Microbiome, 6(1), 1–17.

Bokulich, N. A. et al. (2018b). q2-sample-classifier: machine-learning tools for microbiome classification and regression. Journal of open research software, 3(30).

Bolyen, E. et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. Nature biotechnology, 37(8), 852–857.

Brown, G. D. et al. (2012). Hidden killers: human fungal infections. Science translational medicine, 4(165), 165rv13–165rv13.

Caporaso, J. G. et al. (2009). Nytax: a flexible tool for aligning sequences to a template alignment. bioinformatics, 26(2), 266–267.

Caporaso, J. G. et al. (2010). Qlime allows analysis of high-throughput community sequencing data. Nature Methods, 7(5), 335.

Chakravorty, S. et al. (2007). A detailed analysis of 16s ribosomal rna gene segments for the diagnosis of pathogenic bacteria. Journal of microbiological methods, 69(2), 330–339.

Chun, J. et al. (2007). Extaxon: a web-based tool for the identification of prokaryotes based on 16s ribosomal rna gene sequences. International Journal of Systematic and Evolutionary Microbiology, 57(10), 2259–2261.

Cole, J. R. et al. (2008). The ribosomal database project: improved alignments and new tools for rna analysis. Nucleic Acids Research, 37(suppl_1), D141–D145.

Cole, J. R. et al. (2014). Ribosomal database project: data and tools for high throughput rna analysis. Nucleic acids research, 42(D1), D633–D642.

Deshpande, V. et al. (2016). Fungal identification using a bayesian classifier and the warcup training set of internal transcribed spacer sequences. Mycologia, 108(1), 1–5.

Donovan, P. D. et al. (2018). Identification of fungi in shotgun metagenomics datasets. PLOS ONE, 13(2), e0192898.

Edgar, R. (2016). Sintax: a simple non-bayesian taxonomy classifier for 16s and its sequences. BioRxiv, page 074161.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than blast. Bioinformatics, 26(19), 2460–2461.

Edgar, R. C. (2018). Accuracy of taxonomy prediction for 16s rna and fungal its sequences. PeerJ, 6, e4652.

Fuhrman, J. A. (2012). Metagenomics and its connection to microbial community organization. F1000 Biology Reports, 4.

Gao, X. et al. (2017). A bayesian taxonomic classification method for 16s rna gene sequences with improved species-level accuracy. BMC bioinformatics, 18(1), 247.

Gupta, A. et al. (2014). M3p: a software tool for the prediction of pathogenic proteins in genomic and metagenomic data. PLOS ONE, 9(4), e93907.

Hawksworth, D. L. (2001). The magnitude of fungal diversity: the 1· 5 million species estimate revisited. Mycological research, 105(12), 1422–1432.
Liu, K.-L. and Wong, T.-T. (2013). Naïve bayesian classifiers with
Huffnagle, G. B. and Noverr, M. C. (2013). The emerging world of the
et al.
Hibbett, D. S. (2007). After the gold rush, or before the flood? evolutionary
8
Miranda et al.
Köster, J. and Rahmann, S. (2012). Snakemake—a scalable bioinformatics
et al.
Kõljalg, U.
Meyer, F.
et al.
Mitra, S.
et al.
Kőlajl, U. et al. (2005). Unite: a database providing web-based
Methods in Microbiology, volume 41, pages 61–74. Elsevier.
Köster, J. and Rahmann, S. (2012). Snakemake—a scalable bioinformatics
workflow engine. Bioinformatics, 28(19), 2520–2522.
Liland, K. H. et al. (2017). microclass: an r-package for 16s taxonomy
classification. BMC bioinformatics, 18(1), 172.
Liu, K.-L. and Wong, T.-T. (2013). Naïve bayesian classifiers with
multinomial models for rrna taxonomic assignment. IEEE/ACM
Transactions on computational biology and bioinformatics, 10(5), 1–1.
Maalouf, M. (2011). Logistic regression in data analysis: an overview.
International Journal of Data Analysis Techniques and Strategies, 3(3),
281–299.
Meyer, F. et al. (2008). The metagenomics rast server -- a public resource
for the automatic phylogenetic and functional analysis of metagenomes.
BMC bioinformatics, 9(1), 386.
Milani, C. et al. (2020). Untangling species-level composition of
complex bacterial communities through a novel metagenomic approach.
Mysteme, 5(4).
Miranda, F. M. et al. (2021). Hiclass: a python library for local
hierarchical classification compatible with scikit-learn. arXiv preprint
arXiv:2112.06560.
Mitra, S. et al. (2011). Analysis of 16s rrna environmental sequences using
megan. BMC Genomics, 12(3), S17.
Nilsson, R. H. et al. (2009). The its region as a target for characterization
of fungal communities using emerging sequencing technologies. FEMS
Microbiology Letters, 296(1), 97–101.
Nilsson, R. H. et al. (2018). The unite database for molecular identification
of fungi: handling dark taxa and parallel taxonomic classifications.
Nucleic Acids Research, 47(D1), D259–D264.
Pedregosa, F. et al. (2011). Scikit-learn: Machine learning in python.
Journal of machine learning research, 12(Oct), 2825–2830.
Pruesse, E. et al. (2007). Silva: a comprehensive online resource for quality
checked and aligned ribosomal rrna sequence data compatible with arb.
Nucleic Acids Research, 35(21), 7188–7196.
Sayers, E. W. et al. (2012). Database resources of the national center for
biotechnology information. Nucleic acids research, 40(D1), D13–D25.
Schloss, P. D. et al. (2009). Introducing mothur: open-source,
platform-independent, community-supported software for describing
and comparing microbial communities. Appl Environ. Microbiol.,
75(23), 7537–7541.
Schoch, C. L. et al. (2012). Nuclear ribosomal internal transcribed spacer
(its) region as a universal dna barcode marker for fungi. Proceedings of
the National Academy of Sciences, 109(16), 6241–6246.
Scikit-learn (2022). Logistic Regression. https://
sklearn-learn.org/stable/modules/linear_model.
hn#logistic-regression. [Online; accessed 19-October-2022].
Silla, C. N. and Freitas, A. A. (2011). A survey of hierarchical classification
across different application domains. Data Mining and Knowledge
Discovery, 22(1-2), 31–72.
Tessler, M. et al. (2017). Large-scale differences in microbial biodiversity
discovery between 16s amplicon and shotgun sequencing. Scientific
reports, 7(1), 1–14.
Venter, J. C. et al. (2004). Environmental genome shotgun sequencing of
the sargasso sea. science, 304(5667), 66–74.
Wang, Q. et al. (2007). Naive bayesian classifier for rapid assignment
of rrna sequences into the new bacterial taxonomy. Applied and
Environmental Microbiology, 73(16), 5261–5267.