HiTaC: Hierarchical Taxonomic Classification of Fungal ITS Sequences

Fábio M. Miranda [1,2,3], Vasco A. C. Azevedo [3], Bernhard Y. Renard [1,2], Vitor C. Piro [1,2,*,◕] and Rommel T. J. Ramos [3,4,◕]

[1] Data Analytics and Computational Statistics, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam, 14482 Potsdam, Germany,
[2] Bioinformatics Unit (MF1), Department for Methodology and Research Infrastructure, Robert Koch Institute, 13353 Berlin, Germany,
[3] Institute of Biological Sciences, Universidade Federal de Minas Gerais, Belo Horizonte, 31270–901, Brazil and
[4] Institute of Biological Sciences, Universidade Federal do Pará, Belém, 66075–110, Brazil.

* To whom correspondence should be addressed.
◕ These authors share senior authorship.

Abstract

Motivation: Fungi are key elements 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 of 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: In this work, we developed 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. We show that our hierarchical model, HiTaC, outperforms state-of-the-art methods when trained over noisy data, consistently achieving higher accuracy and sensitivity across different taxonomic ranks.
Availability: HiTaC is an open-source software, with documentation and source code available at https://gitlab.com/dacs-hpi/hitac.
Contact: vitor.cedranpiro@hpi.de
Supplementary information: Supplementary data are available at *bioRxiv* online.

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

© The Author 2020.

a microbiome, it is cheaper and more appropriate to sequence only genetic markers such as 16S rRNA, which is 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 not only cheaper but more reliable.

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 (Jonasson *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.*, 2014). In this scenario, prediction-based approaches might be useful.

One of the first proposed tools to perform prediction-based taxonomic classification was the RDP-Classifier, which uses 8-mer frequencies to train a Naive Bayes classification algorithm (Wang *et al.*, 2007). Improving upon the RDP-Classifier, Liu and Wong (2013) developed a Naive Bayes classifier with multinomial models to increase the accuracy. This multinomial classifier was recently reimplemented and optimized on Microclass (Liland *et al.*, 2017).

Other prediction-based approaches have also been proposed in the literature, such as the k-Nearest Neighbor (KNN) algorithm (Schloss *et al.*, 2009). Given a query sequence, the KNN algorithm identifies the k-most similar sequences in a database, and uses the taxonomic information from each of those sequences to determine the consensus taxonomy. Lastly, Q2_SK (Bokulich *et al.*, 2018) implements several supervised learning classifiers from scikit-learn (Pedregosa *et al.*, 2011), 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. According to Silla and Freitas (2011), the main disadvantages 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. 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 of fungal ITS predictions.

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. 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.

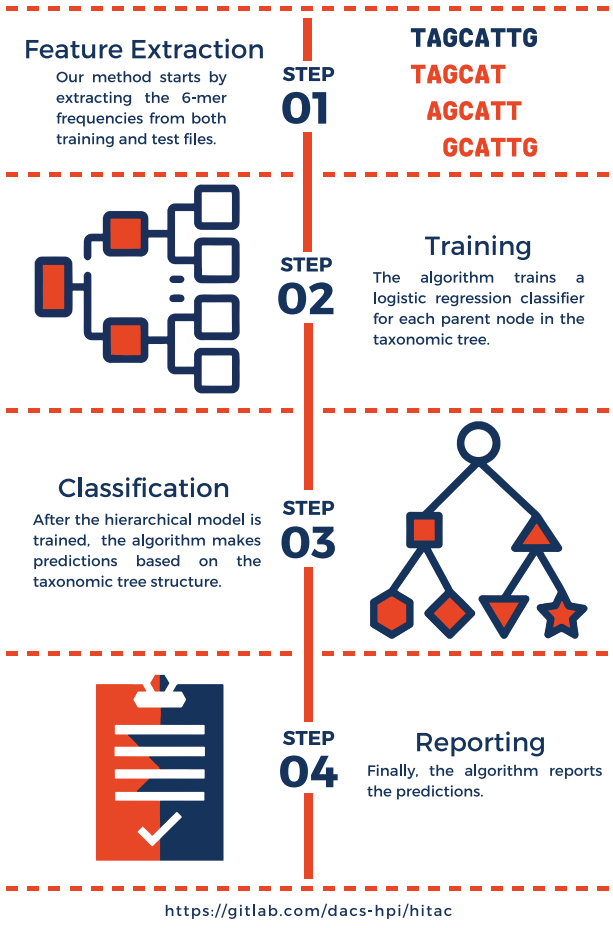


Fig. 1: HiTaC algorithm overview. Firstly, HiTaC extracts the 6-mer frequencies from both training and test files. This default 6-mer value can be easily modified by command line arguments. After feature extraction, HiTaC trains a logistic regression classifier for each parent node in the taxonomic tree, makes predictions based on the taxonomic tree structure and finally reports the predictions.

HiTaC trains a local classifier for each parent node of the taxonomic tree. For example, there is a classifier in the root node which decides whether a given sequence belongs to the phylum Ascomycota or Basidiomycota. If the sequence is classified as Ascomycota, then a new 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. This hierarchical model was chosen because it respects

the natural relationships between classes, 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 from the package scikit-learn v0.21.3 (Pedregosa *et al.*, 2011). Logistic regression was chosen because it is a multi-class classifier known for requiring a small amount of data for training, and is good at coping with imbalanced data.

2.2 Evaluation

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). We relied on this benchmark due to the fact that it more accurately mimics 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 genera that are absent from the training set, while a 95% identity allows for a mix of present and absent genera. In other words, it allows for better evaluation on how a classifier behaves when trying to predict novel sequences.

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 May 11, 2017	(Gao *et al.*, 2017)
BTOP	blastn v2.2.31+	(Altschul *et al.*, 1990)
CT1	usearch v11.0	Re-implements Q1 (Edgar, 2018)
CT2	usearch v11.0	Re-implements Q2_VS (Edgar, 2018)
HiTaC	HiTaC v1.1	(This paper)
KNN	mothur v1.39.5	(Schloss *et al.*, 2009)
KTOP	usearch v11.0	(Edgar, 2018)
Metaxa2	Metaxa 2.2-beta9	(Bengtsson-Palme *et al.*, 2015)
Microclass	microclass v1.1	(Liland *et al.*, 2017)
NBC	usearch v11.0	Re-implements RDP (Edgar, 2018)
Q1	QIIME v1.9	(Caporaso *et al.*, 2010)
Q2_BLAST	QIIME v2.2017.10	(Bokulich *et al.*, 2017)
Q2_SK	QIIME v2.2017.10	(Bokulich *et al.*, 2017)
Q2_VS	QIIME v2.2017.10	(Bokulich *et al.*, 2017)
RDP	Classifier v.2.4	(Wang *et al.*, 2007)
SINTAX	usearch v11.0	(Edgar, 2016)
SPINGO	SPINGO v1.3	(Allard *et al.*, 2015)
TOP	usearch v11.0	(Edgar, 2018)

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 datasets selected from the TAXXI benchmark were based on real environmental data, publicly available on SRA and other sources. They were created by Edgar (2018) 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% (Table 2). In practice, an identity of 90% means that the test dataset contains only novel sequences at the species level but most taxa are known at the genus level. The known taxa at the species level increase with identity.

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. 2). A similar pattern can be observed in the other datasets with higher identities.

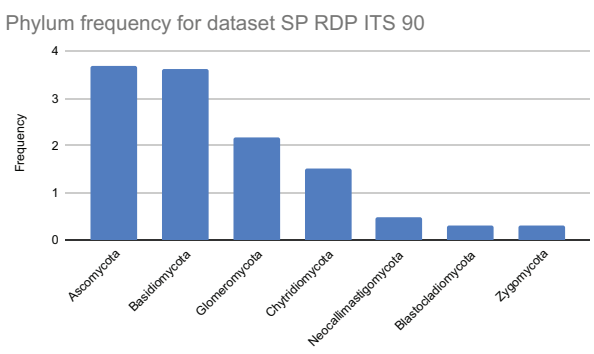


Fig. 2: Depiction of imbalanced training set on a logarithmic scale. For instance, the phyla Ascomycota and Basidiomycota together contain approximately 96% of total instances between the seven existing phyla in dataset SP RDP ITS 90. In the remaining datasets we can also observe similar patterns.

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:

$$\text{ACC} = 100 \times \frac{\text{True Positive}}{\text{Known Sequences} + \text{Over-classification Errors}}$$

$$\text{MCR} = 100 \times \frac{\text{Misclassification Errors}}{\text{Known Sequences}}$$

$$\text{OCR} = 100 \times \frac{\text{Over-classification Errors}}{\text{Novel Sequences}}$$

$$\text{TPR} = 100 \times \frac{\text{True Positive}}{\text{Known Sequences}}$$

$$\text{UCR} = 100 \times \frac{\text{Under-classification Errors}}{\text{Known Sequences}}$$

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. Over-classification errors happen when too many ranks are predicted. Finally, misclassification errors occur when a known name is incorrectly predicted.

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. A summary of the results presented here can be found on Table 3, while the results for the remaining datasets and taxonomic ranks can be found in the supplementary material.

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. 3, 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 2.8% improvement over Microclass, the second best method for this dataset. HiTaC also achieved top accuracy for the datasets with identities varying between 95–99%.

Similar conclusions can be drawn from Fig. 4, in which we evaluated the true positive rate (TPR) for all methods 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 rates at genus level for dataset SP RDP ITS 90 were achieved by HiTaC (80.5%) and BTOP (73.4%), while for dataset SP RDP ITS 95 HiTaC obtained 87.4% TPR and Microclass attained 85.4% TPR and for dataset SP RDP ITS 97 HiTaC tied with Microclass at 91.8% TPR. At species level, HiTaC achieved 89.6% TPR and BTOP obtained 88.7% TPR for dataset SP RDP ITS 99, while for dataset SP RDP ITS 100 HiTaC tied with BTOP at 100% TPR.

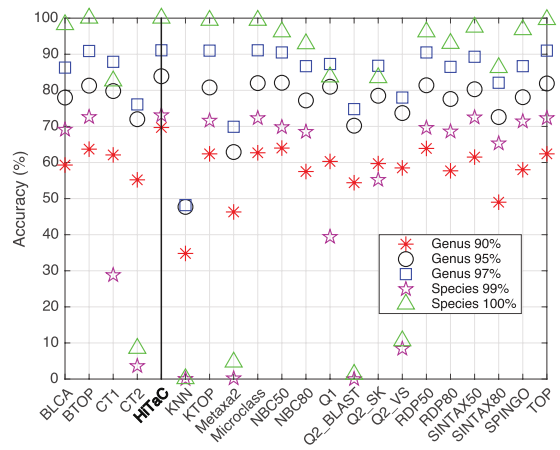


Fig. 3: Accuracy results at genus and species level for datasets SP RDP ITS, with identities varying in these datasets between 90–100%. HiTaC achieved a good accuracy overall, ranking on top for all identity values.

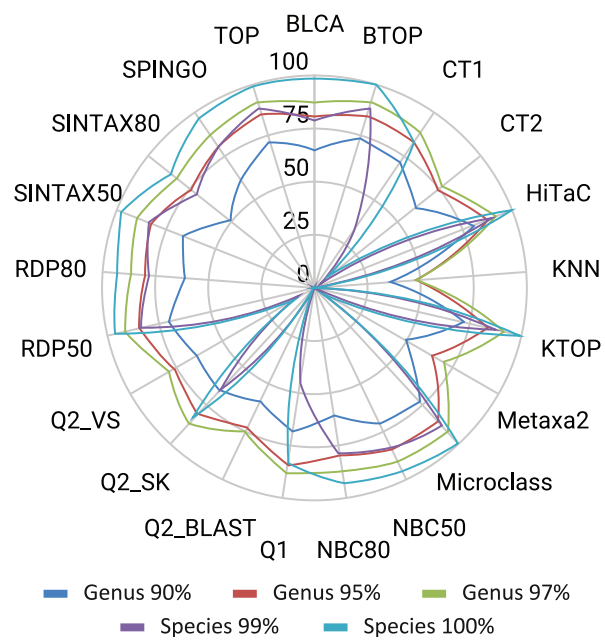
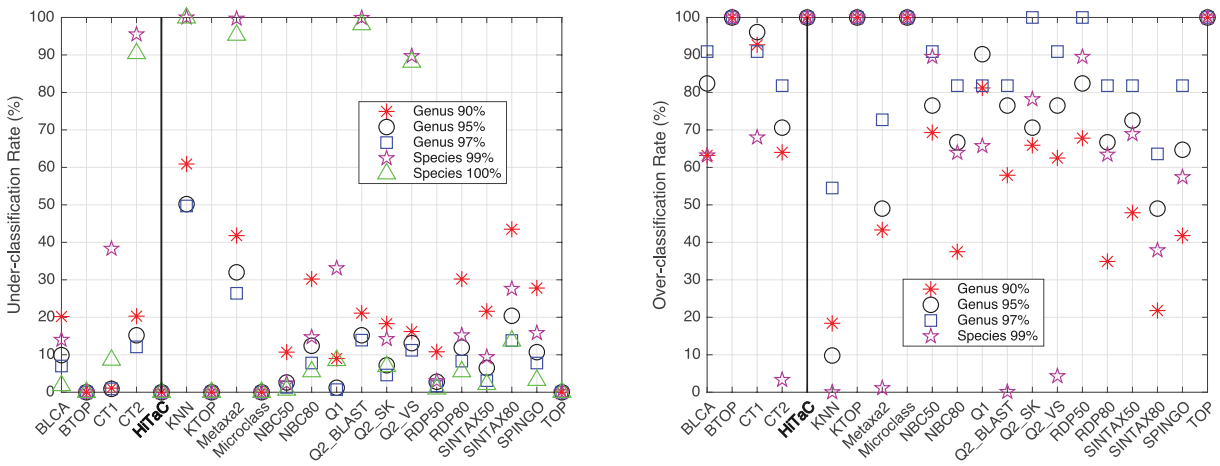


Fig. 4: True positive rates for the ITS datasets at genus and species level. The trend is that HiTaC achieved a sensitivity higher or equal to top methods. For instance, HiTaC tied with top methods for the datasets with 97 and 100% identity. However, for the datasets with 90, 95 and 99% identity HiTaC improved the true positive rates upon top methods.

On Fig 5, we show the results for under-classification and over-classification rates. HiTaC, BTOP, KTOP, Microclass and TOP did not leave any known sequences unclassified. KNN, on the other hand, left almost all known sequences unclassified for the datasets SP RDP ITS 99–100. Regarding over-classification, KNN achieved the lowest rates overall. Conversely, HiTaC, BTOP, KTOP, Microclass and TOP 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. An ideal algorithm would be able to classify all known sequences and leave novel sequences unclassified, thus achieving 100% accuracy.



(a) (b)

Fig. 5: (a) Under-classification rate results for datasets SP RDP ITS at genus and species level. HiTaC, BTOP, KTOP, Microclass and TOP achieved 0% under-classification rates for all datasets, while more conservative methods reached an under-classification rate as high as 99.9% (CT2) and 100% (KNN). (b) Over-classification rate results for datasets SP RDP ITS at genus and species level. Missing over-classification rate results for dataset SP RDP ITS 100 were not generated by the benchmark. HiTaC, BTOP, KTOP, Microclass and TOP achieved 100% over-classification rates for all datasets, while more conservative methods reached an over-classification rate as low as 0.1% (Q2_BLAST) and 0% (KNN). An ideal method would be able to classify all known sequences and leave novel sequences unclassified, thus avoiding both over-classification and under-classification errors.

Lastly, the methods that predominantly attained the lowest misclassification rates were KNN, Metaxa2 and SINTAX80. In general, HiTaC achieved a median misclassification rate, reaching 19.5% 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 misclassification rate results are shown in Fig. 6 and the results presented for all metrics so far are summarized in Table 3.

4 Discussion

Taxonomic classification is only the first step in studies targeting the taxonomic analysis of environmental samples. 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 performance than state-of-the-art taxonomic classifiers even with imbalanced training data.

Feature extraction is performed automatically by HiTaC, using 6-mer as the default value. However, this parameter is flexible and different k-mer values can be chosen. Note that while Microclass, NBC and RDP were trained with 8-mer frequencies on a single classifier, for HiTaC we used 6-mer frequencies to train multiple classifiers for each parent node in the taxonomic tree, which is much more challenging, especially for a dataset as noisy as SP RDP ITS 90. This shows that in spite of using less features, our hierarchical model still outperforms other machine learning models available in the TAXXI benchmark due to a wider exploitation of the taxonomic information.

HiTaC is an open-source software, available at https://gitlab.com/dacs-hpi/hitac. It was implemented in Python, and relies on the scikit-learn library to train local logistic regression classifiers for each parent node in the taxonomic tree. HiTaC can be easily installed through conda or pip, allowing users to install our hierarchical classifier along with its dependencies by running a single command, which is not possible for most of the methods evaluated in the TAXXI benchmark.

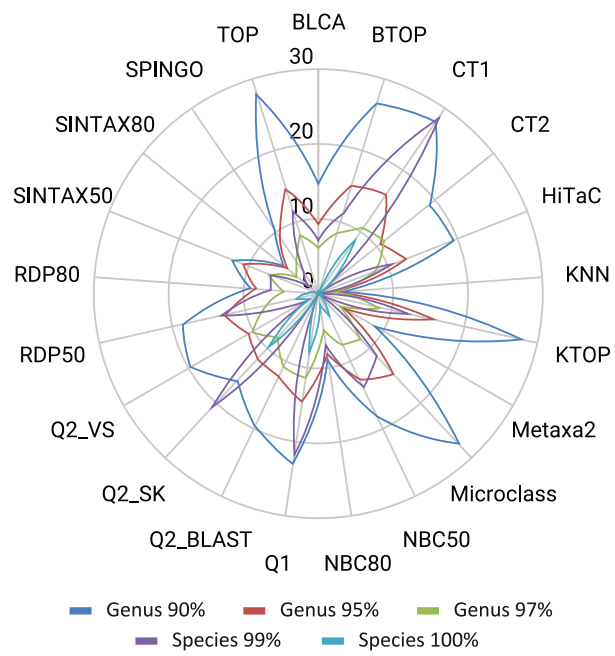


Fig. 6: Misclassification rate results for the SP RDP ITS datasets at genus and species level. The misclassification rates for HiTaC, BTOP, BLCA, SPINGO, SINTAX80, Metaxa2 and KNN were 0% for the dataset SP RDP ITS 100. HiTaC's misclassification rate is deemed adequate for the datasets with 90–99% identity.

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

Table 3. Results for genus and species levels on datasets SP RDP ITS, with best results for each metric highlighted in green and worst results colored in red. This table was sorted in descending order, based on the accuracy obtained by the methods for dataset SP RDP ITS 90. Missing over-classification rates for SP RDP ITS 100 were not generated by the benchmark. Note that higher values for accuracy (ACC) and true positive rate (TPR) are preferred, while the opposite is true for misclassification (MCR), over-classification (OCR) and under-classification (UCR) rates. Overall, HiTaC achieved the highest accuracy and sensitivity for all datasets. HiTaC, BTOP, KTOP, Microclass and TOP reached the lowest under-classification rates for all datasets, while KNN achieved the lowest misclassification and over-classification rates. HiTaC and BTOP got perfect scores for all metrics in the dataset with 100% identity.

Method	Genus at 90% Identity					Genus at 95% Identity					Genus at 97% Identity					Species at 99% Identity					Species at 100% Identity			
	ACC	MCR	OCR	TPR	UCR	ACC	MCR	OCR	TPR	UCR	ACC	MCR	OCR	TPR	UCR	ACC	MCR	OCR	TPR	UCR	ACC	MCR	TPR	UCR
HiTaC	69.7	19.5	100	80.5	0	83.9	12.6	100	87.4	0	91.1	8.2	100	91.8	0	73.1	10.4	100	89.6	0	100	0	100	0
NBC50	64	18.3	69.3	71	10.7	82.1	12.7	76.5	84.8	2.6	90.5	7.6	90.9	91.1	1.3	69.8	13.8	89.5	83.8	2.4	96.2	3.2	96.2	0.6
RDP50	63.9	18.6	67.8	70.6	10.8	81.4	12.9	82.4	84.3	2.8	90.5	7.1	100	91.2	1.7	69.6	13.3	89.5	83.7	3	96.2	2.9	96.2	0.9
BTOP	63.7	26.6	100	73.4	0	81.3	15.2	100	84.8	0	90.9	8.4	100	91.6	0	72.6	11.3	100	88.7	0	100	0	100	0
Microclass	62.7	27.5	100	72.5	0	82	14.6	100	85.4	0	91.1	8.2	100	91.8	0	72.3	11.4	100	88.6	0	99.4	0.6	99.4	0
KTOP	62.4	27.8	100	72.2	0	80.8	15.8	100	84.2	0	91	8.3	100	91.7	0	71.6	12.2	100	87.8	0	99.4	0.6	99.4	0
TOP	62.4	27.9	100	72.1	0	81.9	14.7	100	85.3	0	91	8.3	100	91.7	0	72.3	11.4	100	88.6	0	99.6	0.4	99.6	0
CT1	62.1	27.9	92.7	71	1.1	79.8	16.1	96.1	83	0.9	87.9	10.6	90.9	88.6	0.9	28.8	28.5	68	33.2	38.3	82.6	8.7	82.6	8.6
SINTAX50	61.5	12.3	47.9	66.1	21.6	80.3	10.7	72.5	82.8	6.5	89.3	7	81.8	89.9	3.1	72.5	6.8	68.9	83.8	9.4	97.5	0.4	97.5	2.1
Q1	60.3	23	81.2	68	9	81	14.7	90.2	84.1	1.2	87.3	11.4	81.8	87.9	0.7	39.4	21.7	65.7	45.2	33.1	83.7	7.8	83.7	8.5
Q2_SK	59.7	15.9	65.9	65.8	18.3	78.5	11.9	70.6	80.8	7.2	86.8	7.9	100	87.4	4.6	55.2	20.8	78.2	65	14.2	83.5	9.6	83.5	6.9
BLCA	59.3	14.7	63.2	65.1	20.2	78	9.4	82.4	80.7	9.9	86.3	6.1	90.9	86.9	7	69.1	7.1	63.2	78.9	14	98.2	0	98.2	1.8
Q2_VS	58.5	19.7	62.5	64.2	16.2	73.7	10.8	76.5	76.1	13.1	78	10.2	90.9	78.6	11.2	8.4	1.8	4.3	8.4	89.7	10.6	1.3	10.6	88.1
SPINGO	58	10.4	41.8	61.8	27.8	78.1	9.1	64.7	80.2	10.7	86.7	5	81.8	87.2	7.8	71.4	3.5	57.4	80.6	15.8	96.8	0	96.8	3.2
RDP80	57.7	9	34.9	60.8	30.2	77.6	8.3	66.7	79.8	11.9	86.5	4.7	81.8	87	8.3	68.6	6.4	63.4	78.3	15.2	93	1.5	93	5.5
NBC80	57.5	9	37.5	60.9	30.2	77.2	8.2	66.7	79.4	12.4	86.7	5	81.8	87.2	7.8	68.5	6.9	63.9	78.3	14.7	92.9	1.6	92.9	5.5
CT2	55.2	19	64	60.8	20.3	72	10.6	70.6	74.1	15.2	76.1	11.3	81.8	76.6	12.1	3.6	0.9	3.3	3.6	95.5	8.5	1.1	8.5	90.4
Q2_BLAST	54.4	19.5	57.9	59.3	21.1	70.2	12.3	76.5	72.5	15.2	74.8	10.9	81.8	75.2	13.9	0	0	0.1	0	99.9	1.3	0.5	1.3	98.1
SINTAX80	49	5.8	21.8	50.7	43.5	72.6	5.4	49	74.1	20.4	82.1	3.7	63.6	82.5	13.8	65.3	1.6	37.9	70.9	27.6	86.3	0	86.3	13.7
Metaxa2	46.3	8.8	43.3	49.4	41.8	62.9	3.9	49	64.2	32	69.9	3.3	72.7	70.3	26.4	0.2	0	1.1	0.2	99.7	4.7	0	4.7	95.3
KNN	34.8	3.3	18.4	35.8	60.9	47.7	1.9	9.8	47.9	50.2	48.2	1.9	54.5	48.4	49.7	0	0	0	0	100	0.1	0	0.1	99.9

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, in a future release, we plan to output confidence scores for the predictions made and add a threshold parameter, which can be set to detect 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, from an ecological perspective, it is interesting to use amplicon sequencing in addition to shotgun due to the fact that it can more accurately reflect the composition of microbiomes, as well as taxonomic abundance and richness, especially at higher ranks (Tessler *et al.*, 2017). The reason for that is that there are issues inherent to shotgun sequencing and a huge discrepancy between the number of genetic markers and genomes deposited in public databases.

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). For the aforementioned reasons, we believe that HiTaC is a great contribution to mycology.

Acknowledgements

We gratefully acknowledge Marcos Augusto dos Santos (UFMG) for initial inspiration, Robert C. Edgar (independent researcher) for providing the scripts and datasets in the TAXXI benchmark and Elizabeth Yuu (HPI) for proofreading.

Funding

This work was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) [001]; and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [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 rrna in metagenomic data. *Molecular Ecology Resources*, **15**(6), 1403–1414.

Bokulich, N. A. *et al.* (2017). Optimizing taxonomic classification of marker gene sequences. *PeerJ Preprints*.

Bokulich, N. A. *et al.* (2018). 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). Pynast: a flexible tool for aligning sequences to a template alignment. *Bioinformatics*, **26**(2), 266–267.

Caporaso, J. G. *et al.* (2010). Qiime 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). Eztaxon: 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 rrna analysis. *Nucleic Acids Research*, **37**(suppl_1), D141–D145.

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 rrna 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 rrna gene sequences with improved species-level accuracy. *BMC bioinformatics*, **18**(1), 247.

Gupta, A. *et al.* (2014). Mp3: 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.

Hibbett, D. S. (2007). After the gold rush, or before the flood? evolutionary morphology of mushroom-forming fungi (agaricomycetes) in the early 21st century. *Mycological research*, **111**(9), 1001–1018.

Hoggard, M. *et al.* (2018). Characterizing the human mycobiota: a comparison of small subunit rrna, its1, its2, and large subunit rrna genomic targets. *Frontiers in microbiology*, **9**, 2208.

Huffnagle, G. B. and Noverr, M. C. (2013). The emerging world of the fungal microbiome. *Trends in microbiology*, **21**(7), 334–341.

Jonasson, J. *et al.* (2002). Classification, identification and subtyping of bacteria based on pyrosequencing and signature matching of 16s rdna fragments. *APMIS*, **110**(3), 263–272.

Kim, M. and Chun, J. (2014). 16s rrna gene–based identification of bacteria and archaea using the eztaxon server. In *Methods in Microbiology*, volume 41, pages 61–74. Elsevier.

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.

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.

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 rna sequence data compatible with arb. *Nucleic Acids Research*, **35**(21), 7188–7196.

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.

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.