bioRxiv, 2023, pp. 1–2

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Application Note

APPLICATION NOTE

HiTaC: a Hierarchical Taxonomic Classifier for Fungal ITS Sequences

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

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

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

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Summary: Fungi naturally inhabit the human microbiome and can be causative agents of infections. Hence, an accurate and robust method for fungal ITS classification can 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. Therefore, in this work we introduce HiTaC, a robust hierarchical machine learning model for accurate ITS classification. HiTaC is implemented as a QIIME2 plugin in order to facilitate its adoption in existing mycobiome analysis pipelines. We benchmarked it with seventeen popular software for taxonomic classification on real environmental sequencing data. HiTaC shows superior accuracy and sensitivity in classifying ITS sequences across different taxonomic ranks.
Availability and implementation: HiTaC is freely available on GitLab `https://gitlab.com/dacs-hpi/hitac`
Contact: vitor.piro@fu-berlin.de
Supplementary information: Supplementary data are available at *bioRxiv* online.

Introduction

Fungi naturally inhabit the human microbiome and can be causative agents of infections that may be fatal (Brown et al., 2012; Hoggard et al., 2018), yet little is known about the fungal microbiota of healthy individuals when compared to the bacterial microbiome (Huffnagle and Noverr, 2013). One of the approaches that can be employed in biodiversity studies of mycobiomes is to sequence only the internal transcribed spacer (ITS) region, which is established as the universal marker for classifying fungi (Schoch et al., 2012).

To identify fungal organisms in amplicon sequencing data, the taxonomic classification of ITS sequences can be performed by using either homology-based or prediction-based techniques. Homology-based approaches, such as BLAST (Altschul et al., 1990) and UCLUST (Edgar, 2010), require the alignment of a query sequence against all sequences available in a reference database. However, a major limitation of homology-based approaches is the reliance on the availability of similar sequences in reference databases (Gupta et al., 2014). Alternatively, prediction-based approaches might be useful to overcome this limitation.

Here, we introduce HiTaC, an easy-to-use yet effective prediction-based approach that applies local hierarchical classification for fungi detection in amplicon sequencing data. It takes rRNA sequences as input and yields the taxonomic labels for the detected fungal organisms. We benchmark HiTaC in various scenarios and compare it with other homology and prediction-based methods.

Materials and methods

HiTaC performs feature extraction, building of the hierarchical model, taxonomic classification, and reporting of the predictions. Figure 1a illustrates the algorithmic steps of HiTaC, where it starts by decomposing the DNA sequences from both reference and query files into their constituents k-mers. Next, two k-mer frequency matrices are built for the reference and query sequences, respectively, where each row contains the k-mer frequency for a sequence.

Logistic regression classifiers are trained for each parent node in the taxonomic hierarchy (Miranda et al., 2023), using the k-mer frequency matrix and taxonomic labels of the reference sequences as training data, with the goal of predicting children labels (Fig. S1). Predictions are performed using the k-mer frequency matrix computed for the query sequences and are reported in a top-down approach, based on the taxonomic hierarchy. For example, if the classifier at the root node predicts that a query sequence belongs to the phylum Ascomycota, then only that branch of the taxonomic tree will be considered for the remaining taxonomic levels.

© The Authors 2023.

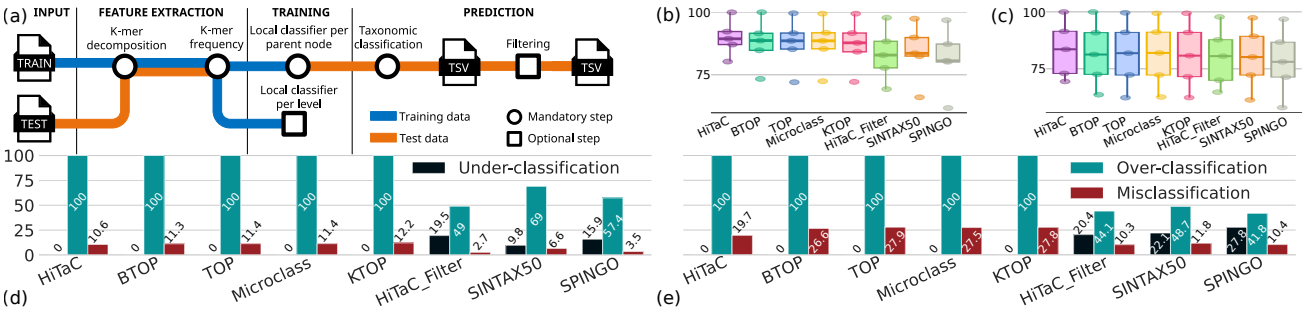

Fig. 1. (a) Sketch of the key algorithmic components of HiTaC. (b) Sensitivity performance at genus level on the datasets with 90, 95, and 97% identity and at species level on the datasets with 99 and 100% identity. (c) Accuracy performance at genus level on the datasets with 90-97% identity and 99-100% identity at species level. (d) Under-classification, over-classification, and misclassification rates at species level on the dataset with 99% identity and (e) at genus level on the dataset with 90% identity.

In order to reduce over-classification and misclassification errors, we also construct a filter by training a logistic regression classifier for each level of the hierarchy. This filter computes confidence scores for all taxonomic ranks assigned by the local classifier per parent node and excludes labels that are below a user-defined threshold, in a bottom-up approach (Fig. S2).

Results

We evaluated HiTaC in five datasets and compared its performance against seventeen taxonomic classifiers using the TAXXI benchmark (Edgar, 2018). The commands and parameters executed are summarized in Tables S4-S22.

The datasets in the TAXXI benchmark are created using real data, with a strategy known as cross-validation by identity, where varying distances between query sequences and the reference database are accounted for. In practice, this means that in the dataset with 90% identity all species are novel and there is a mix of novel and known taxa at the genus level, making it the most difficult dataset in the benchmark. As the identity increases, so does the amount of known labels at all taxonomic ranks; hence, the dataset with 100% identity can be considered the easiest in the benchmark.

We first evaluated HiTaC on the datasets with 90, 95 and 97% identity at genus level and the datasets with 99 and 100% identity at species level. As shown in Figure 1b, HiTaC achieved a sensitivity higher or equal to top methods. For instance, HiTaC tied with BTOP on the dataset with 100% identity, obtaining a perfect score. Moreover, HiTaC improved the sensitivity upon the best methods on the datasets with 90-99% identity, gaining a maximum of 6.9 percentage points.

To assess the accuracy, we also evaluated the datasets with 90-97% identity at genus level and the datasets with 99-100% identity at species level. As shown in Figure 1c, HiTaC outperformed the second best method by 4.6 percentage points on the dataset with 90% identity, while on the dataset with 100% identity it tied with BTOP, achieving 100% accuracy.

In Figures 1d-e, we assessed the over-classification, under-classification, and misclassification rates of the best-performing methods. We noticed that the best methods tend to have a high over-classification rate, which means that they are wrongly labeling novel sequences that do not exist in the training data. The introduction of the filter proved effective to reduce HiTaC's over-classification and misclassification rates, 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 0% for both over-classification and under-classification rates.

Conclusion

In summary, here we developed a novel taxonomic classifier, called HiTaC, which employs a local hierarchical machine learning model for accurate fungal ITS classification and is implemented as a QIIME2 plugin for accessibility. Our algorithm is able to handle imbalanced data and outperforms state-of-the-art classification methods on ITS datasets, consistently achieving higher accuracy and sensitivity across different taxonomic ranks, especially on lower taxonomic levels.

Funding

VCP was financed by the Deutsche Forschungsgemeinschaft (DFG) under grant number 458163427.
Conflict of Interest: none declared.

References

Altschul, S. et al. (1990) Basic local alignment search tool. *J. Mol. Biol*, **215**, 403–410.

Brown, G. et al. (2012) Hidden killers: human fungal infections. *Sci. Transl. Med.*, **4**, 165rv13–165rv13.

Edgar, R. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinform.*, **26**, 2460–2461.

Edgar, R. (2018) Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. *PeerJ*, **6**, e4652.

Gupta, A. et al. (2014) MP3: A Software Tool for the Prediction of Pathogenic Proteins in Genomic and Metagenomic Data. *PLOS ONE*, **9**, e93907.

Hoggard, M. et al. (2018) Characterizing the Human Mycobiota: A Comparison of Small Subunit rRNA, ITS1, ITS2, and Large Subunit rRNA Genomic Targets. *Front. Microbiol.*, **9**, 2208.

Huffnagle, G. and Noverr, M. (2013) The emerging world of the fungal microbiome. *Trends Microbiol.*, **21**, 334–341.

Miranda, F. et al. (2023) HiClass: a python library for local hierarchical classification compatible with scikit-learn. *JMLR*, **24**, 1–17.

Schoch, C. et al. (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 6241–6246.