SOFTWARE ARTICLE

HiTaC: Hierarchical Taxonomic Classification of Fungal ITS Sequences

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

[*]Correspondence: vitor.cedranpiro@hpi.de
[1]Data Analytics and Computational Statistics, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam, Prof.-Dr.-Helmert-Straße 2-3, 14482 Potsdam, Germany
Full list of author information is available at the end of the article
[†]These authors share senior authorship

Abstract

Background: 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.

Conclusions: HiTaC outperformed state-of-the-art taxonomic classifiers in most metrics. A user-friendly QIIME2 plugin allows users to easily include HiTaC in existing pipelines. HiTaC is publicly available and freely accessible at `https://gitlab.com/dacs-hpi/hitac`.

Keywords: Taxonomic classification; Fungal ITS sequences; Amplicon sequencing; Mycobiome; Hierarchical classification; Machine learning; Logistic regression

Background

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 [1]. Moreover, fungi naturally inhabit the human microbiome and can be causative agents of infections that may lead to death [2, 3], yet very little is known about the fungal microbiota of healthy

individuals when compared to the bacterial microbiome [4]. 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 [5]. Hence, it became common practice to use DNA sequencing in addition to morphological studies in contemporary mycology [6].

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 [7]. 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 [8] and also widely used to classify and identify archaea [9]. In mycology, the internal transcribed spacer (ITS) region was adopted as the universal marker for classifying fungi [10]. Comparatively, ITS has far more sequences available than fungal WGS, making it a viable cheaper alternative in mycology.

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 [11]. 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 [10]. 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 [5].

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 [12] such as Silva [13], Ribosomal Database Project (RDP) [14] and UNITE [15]. There are several homology-based software currently available for the analysis of environmental samples, such as EzTaxon [16], MG-RAST [17], PyNAST [18], UCLUST [19], QIIME 2 [20], MEGAN [21] and FindFungi [22]. However, a major limitation of homology-based approaches is the reliance on the availability of homologous sequences in reference databases [23]. Alternatively, prediction-based approaches might be useful to overcome this limitation.

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 [24]. Improving upon the RDP-Classifier, a novel Naive Bayes classifier with multinomial models was developed to increase the accuracy [25]. This multinomial classifier was recently reimplemented and optimized on Microclass [26].

Other prediction-based approaches have also been proposed in the literature, such as the k-Nearest Neighbor (KNN) algorithm [27]. 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. Q2_SK [28] implements several supervised learning classifiers from scikit-learn [29], 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 [30]. 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.

Implementation

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.

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 [31]. For example, 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 [30].

Each parent node in our hierarchical model consists of a logistic regression classifier from the package scikit-learn v0.24.1 [29]. Logistic regression is a statistical model that is frequently used for classification and predictive analysis. In summary, logistic regression estimates possible outcomes using a logistic function (for a mathematical definition, we redirect the reader to scikit-learn's documentation [32]), and it was chosen because it is a multi-class classifier known for requiring a small amount of data for training, and it copes well with imbalanced data when corrections are applied [33].

Furthermore, a filter was implemented in HiTaC to improve the quality of the predictions (Fig. 2), reducing classification errors when dealing with novel sequences or when the algorithm has a low confidence the prediction is correct. In this filter, another logistic regression classifier is trained for each taxonomic level available in the training data, employing the local hierarchical classifier per level implementation from HiClass [31]. Afterwards, correction is performed in a bottom-up approach, i.e., correcting labels from the leaf nodes towards the root. The confidence threshold of this filter has a default value of 0.7, however, it can also be defined by the user via a command line argument.

HiTaC was implemented in Python as a QIIME2 plugin in order to facilitate adoption in existing ITS sequence analysis pipelines, and it can be easily installed through conda, pip or docker, 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 available in the literature.

Evaluation

The TAXXI benchmark [34] 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 95% 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.

Datasets

We adopted five datasets containing both training and test data, which were previously available in the TAXXI benchmark [34]. 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 [35], RDP [36], the Warcup fungal ITS training set v2 (WITS) [37], UNITE [38] 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% [34] (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%.

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.

Evaluation Metrics

To measure the performance of our model, we adopted all metrics available in the TAXXI benchmark [34], 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, while *over-classification errors* happen when too many ranks are predicted. Finally, *misclassification errors* occur when a known name is incorrectly predicted.

Resources benchmark

We enhanced the TAXXI benchmark by porting it to Snakemake [39] in order to make it more reproducible, and we introduced a resources benchmark. The benchmark scripts are available in our GitLab repository[1], 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.

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 GitLab repository[2].

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

[1] `https://gitlab.com/dacs-hpi/hitac/-/tree/master/benchmark`

[2] `https://gitlab.com/dacs-hpi/hitac/-/tree/master/benchmark/results/metrics`

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 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.

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 [40]. Nevertheless, ITS has been successfully employed to accurately catalogue bacterial communities at (sub)species resolution [41]. Moreover, 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 [42].

Conclusions

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://gitlab.com/dacs-hpi/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 Yuu (HPI) for proofreading.

Author's contributions
FMM designed and implemented HiTaC's algorithm, performed analysis, benchmarked the experiments and wrote the manuscript. VAA assisted with analysis and wrote the manuscript. BYR assisted with analysis, visualizations and wrote the manuscript. VCP oversaw experimental and analysis aspects of the project and wrote the manuscript. RTR oversaw the computational and analysis aspects of the project and edited the manuscript. All authors edited and approved the final manuscript.

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.

Availability of data and materials
A reproducible Snakemake pipeline is available at
`https://gitlab.com/dacs-hpi/hitac/-/tree/master/benchmark`.
Project name: HiTaC
Project home page: `https://gitlab.com/dacs-hpi/hitac`
Operating system: Platform independent (tested on GNU/Linux)
Programming language: Python3
License: BSD 3-Clause

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Data Analytics and Computational Statistics, Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam, Prof.-Dr.-Helmert-Straße 2-3, 14482 Potsdam, Germany. [2]Department of Mathematics and Computer Science, Freie Universität Berlin, Takustr. 9, 14195 Berlin, Germany. [3]Institute of Biological Sciences, Universidade Federal de Minas Gerais, Av. Pres. Antônio Carlos, 6627, 31270–901 Belo Horizonte, Brazil. [4]Institute of Biological Sciences, Universidade Federal do Pará, R. Augusto Corrêa, 01, 66075–110 Belém, Brazil.

References

1. Hawksworth, D.L.: The magnitude of fungal diversity: the 1· 5 million species estimate revisited. Mycological research **105**(12), 1422–1432 (2001)
2. Brown, G.D., Denning, D.W., Gow, N.A., Levitz, S.M., Netea, M.G., White, T.C.: Hidden killers: human fungal infections. Science translational medicine **4**(165), 165–1316513 (2012)
3. Hoggard, M., Vesty, A., Wong, G., Montgomery, J.M., Fourie, C., Douglas, R.G., Biswas, K., Taylor, M.W.: Characterizing the human mycobiota: a comparison of small subunit rrna, its1, its2, and large subunit rrna genomic targets. Frontiers in microbiology **9**, 2208 (2018)
4. Huffnagle, G.B., Noverr, M.C.: The emerging world of the fungal microbiome. Trends in microbiology **21**(7), 334–341 (2013)
5. Nilsson, R.H., Ryberg, M., Abarenkov, K., Sjökvist, E., Kristiansson, E.: The its region as a target for characterization of fungal communities using emerging sequencing technologies. FEMS Microbiology Letters **296**(1), 97–101 (2009)
6. Hibbett, D.S.: 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 (2007)
7. Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., *et al.*: Environmental genome shotgun sequencing of the sargasso sea. science **304**(5667), 66–74 (2004)
8. Fuhrman, J.A.: Metagenomics and its connection to microbial community organization. F1000 Biology Reports **4** (2012)
9. Kim, M., Chun, J.: 16s rrna gene–based identification of bacteria and archaea using the eztaxon server. In: Methods in Microbiology vol. 41, pp. 61–74. Elsevier, ??? (2014)
10. Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., Chen, W., Consortium, F.B., *et al.*: 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 (2012)
11. Chakravorty, S., Helb, D., Burday, M., Connell, N., Alland, D.: A detailed analysis of 16s ribosomal rna gene segments for the diagnosis of pathogenic bacteria. Journal of microbiological methods **69**(2), 330–339 (2007)
12. Jonasson, J., Olofsson, M., Monstein, H.-J.: Classification, identification and subtyping of bacteria based on pyrosequencing and signature matching of 16s rdna fragments. APMIS **110**(3), 263–272 (2002)
13. Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., Glöckner, F.O.: Silva: a comprehensive online resource for quality checked and aligned ribosomal rna sequence data compatible with arb. Nucleic Acids Research **35**(21), 7188–7196 (2007)
14. Cole, J.R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R.J., Kulam-Syed-Mohideen, A., McGarrell, D.M., Marsh, T., Garrity, G.M., *et al.*: The ribosomal database project: improved alignments and new tools for rrna analysis. Nucleic Acids Research **37**(suppl_1), 141–145 (2008)
15. Nilsson, R.H., Larsson, K.-H., Taylor, A.F.S., Bengtsson-Palme, J., Jeppesen, T.S., Schigel, D., Kennedy, P., Picard, K., Glöckner, F.O., Tedersoo, L., *et al.*: The unite database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. Nucleic Acids Research **47**(D1), 259–264 (2018)
16. Chun, J., Lee, J.-H., Jung, Y., Kim, M., Kim, S., Kim, B.K., Lim, Y.-W.: 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 (2007)
17. Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., *et al.*: The metagenomics rast server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics **9**(1), 386 (2008)
18. Caporaso, J.G., Bittinger, K., Bushman, F.D., DeSantis, T.Z., Andersen, G.L., Knight, R.: Pynast: a flexible tool for aligning sequences to a template alignment. Bioinformatics **26**(2), 266–267 (2009)
19. Edgar, R.C.: Search and clustering orders of magnitude faster than blast. Bioinformatics **26**(19), 2460–2461 (2010)
20. Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., *et al.*: Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. Nature biotechnology **37**(8), 852–857 (2019)
21. Mitra, S., Stärk, M., Huson, D.H.: Analysis of 16s rrna environmental sequences using megan. BMC Genomics **12**(3), 17 (2011). BioMed Central
22. Donovan, P.D., Gonzalez, G., Higgins, D.G., Butler, G., Ito, K.: Identification of fungi in shotgun metagenomics datasets. PLOS ONE **13**(2), 0192898 (2018)
23. Gupta, A., Kapil, R., Dhakan, D.B., Sharma, V.K.: Mp3: a software tool for the prediction of pathogenic proteins in genomic and metagenomic data. PLOS ONE **9**(4), 93907 (2014)
24. Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R.: Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. Applied and Environmental Microbiology **73**(16), 5261–5267 (2007)
25. Liu, K.-L., Wong, T.-T.: Naïve bayesian classifiers with multinomial models for rrna taxonomic assignment. IEEE/ACM Transactions on computational biology and bioinformatics **10**(5), 1–1 (2013)
26. Liland, K.H., Vinje, H., Snipen, L.: microclass: an r-package for 16s taxonomy classification. BMC bioinformatics **18**(1), 172 (2017)
27. Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., *et al.*: Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. **75**(23), 7537–7541 (2009)
28. Bokulich, N.A., Dillon, M.R., Bolyen, E., Kaehler, B.D., Huttley, G.A., Caporaso, J.G.: q2-sample-classifier: machine-learning tools for microbiome classification and regression. Journal of open research software **3**(30) (2018)

29. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., *et al.*: Scikit-learn: Machine learning in python. Journal of machine learning research **12**(Oct), 2825–2830 (2011)
30. Silla, C.N., Freitas, A.A.: A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery **22**(1-2), 31–72 (2011)
31. Miranda, F.M., Köhnecke, N., Renard, B.Y.: Hiclass: a python library for local hierarchical classification compatible with scikit-learn. arXiv preprint arXiv:2112.06560 (2021)
32. Scikit-learn: Logistic Regression. `https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression`. [Online; accessed 19-October-2022] (2022)
33. Maalouf, M.: Logistic regression in data analysis: an overview. International Journal of Data Analysis Techniques and Strategies **3**(3), 281–299 (2011)
34. Edgar, R.C.: Accuracy of taxonomy prediction for 16s rrna and fungal its sequences. PeerJ **6**, 4652 (2018)
35. Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Federhen, S., *et al.*: Database resources of the national center for biotechnology information. Nucleic acids research **40**(D1), 13–25 (2012)
36. Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., Tiedje, J.M.: Ribosomal database project: data and tools for high throughput rrna analysis. Nucleic acids research **42**(D1), 633–642 (2014)
37. Deshpande, V., Wang, Q., Greenfield, P., Charleston, M., Porras-Alfaro, A., Kuske, C.R., Cole, J.R., Midgley, D.J., Tran-Dinh, N.: Fungal identification using a bayesian classifier and the warcup training set of internal transcribed spacer sequences. Mycologia **108**(1), 1–5 (2016)
38. Kõljalg, U., Larsson, K.-H., Abarenkov, K., Nilsson, R.H., Alexander, I.J., Eberhardt, U., Erland, S., Høiland, K., Kjøller, R., Larsson, E., *et al.*: Unite: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. New Phytologist **166**(3), 1063–1068 (2005)
39. Köster, J., Rahmann, S.: Snakemake—a scalable bioinformatics workflow engine. Bioinformatics **28**(19), 2520–2522 (2012)
40. Badotti, F., de Oliveira, F.S., Garcia, C.F., Vaz, A.B.M., Fonseca, P.L.C., Nahum, L.A., Oliveira, G., Góes-Neto, A.: Effectiveness of its and sub-regions as dna barcode markers for the identification of basidiomycota (fungi). BMC microbiology **17**(1), 42 (2017)
41. Milani, C., Alessandri, G., Mangifesta, M., Mancabelli, L., Lugli, G.A., Fontana, F., Longhi, G., Anzalone, R., Viappiani, A., Duranti, S., et al.: Untangling species-level composition of complex bacterial communities through a novel metagenomic approach. Msystems **5**(4) (2020)
42. Tessler, M., Neumann, J.S., Afshinnekoo, E., Pineda, M., Hersch, R., Velho, L.F.M., Segovia, B.T., Lansac-Toha, F.A., Lemke, M., DeSalle, R., *et al.*: Large-scale differences in microbial biodiversity discovery between 16s amplicon and shotgun sequencing. Scientific reports **7**(1), 1–14 (2017)
43. Gao, X., Lin, H., Revanna, K., Dong, Q.: A bayesian taxonomic classification method for 16s rrna gene sequences with improved species-level accuracy. BMC bioinformatics **18**(1), 247 (2017)
44. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of molecular biology **215**(3), 403–410 (1990)
45. Bengtsson-Palme, J., Hartmann, M., Eriksson, K.M., Pal, C., Thorell, K., Larsson, D.G.J., Nilsson, R.H.: Metaxa2: improved identification and taxonomic classification of small and large subunit rrna in metagenomic data. Molecular Ecology Resources **15**(6), 1403–1414 (2015)
46. Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I., *et al.*: Qiime allows analysis of high–throughput community sequencing data. Nature Methods **7**(5), 335 (2010)
47. Bokulich, N.A., Kaehler, B.D., Rideout, J.R., Dillon, M., Bolyen, E., Knight, R., Huttley, G.A., Caporaso, J.G.: Optimizing taxonomic classification of marker gene sequences. PeerJ Preprints (2017)
48. Edgar, R.: Sintax: a simple non-bayesian taxonomy classifier for 16s and its sequences. BioRxiv, 074161 (2016)
49. Allard, G., Ryan, F.J., Jeffery, I.B., Claesson, M.J.: Spingo: a rapid species-classifier for microbial amplicon sequences. BMC bioinformatics **16**(1), 324 (2015)

Figures

Tables

Table 1: List of software available in the TAXXI benchmark [34], which were used for comparison with HiTaC.

Method	Software	Algorithm reference
BLCA	BLCA v2.3-alpha	[43]
BTOP	blastn v2.10.1+	[44]
CT1	usearch v11.0.667	Re-implements Q1 [34]
CT2	usearch v11.0.667	Re-implements Q2_VS [34]
HiTaC	HiTaC v2.0.10	(This paper)
KNN	mothur v1.48.0	[27]
KTOP	usearch v11.0.667	[34]
Metaxa2	Metaxa 2.2.3	[45]
Microclass	microclass v1.2	[26]
NBC	usearch v11.0.667	Re-implements RDP [34]
Q1	QIIME v1.9	[46]
Q2_BLAST	QIIME v2.2022.2	[47]
Q2_SK	QIIME v2.2022.2	[47]
Q2_VS	QIIME v2.2022.2	[47]
RDP	Classifier v2.13	[24]
SINTAX	usearch v11.0.667	[48]
SPINGO	SPINGO v1.3	[49]
TOP	usearch v11.0.667	[34]

Table 2: List of datasets available in the TAXXI benchmark [34], 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

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	67.66
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	14817.23
NBC	-	Memory Limit Exceeded

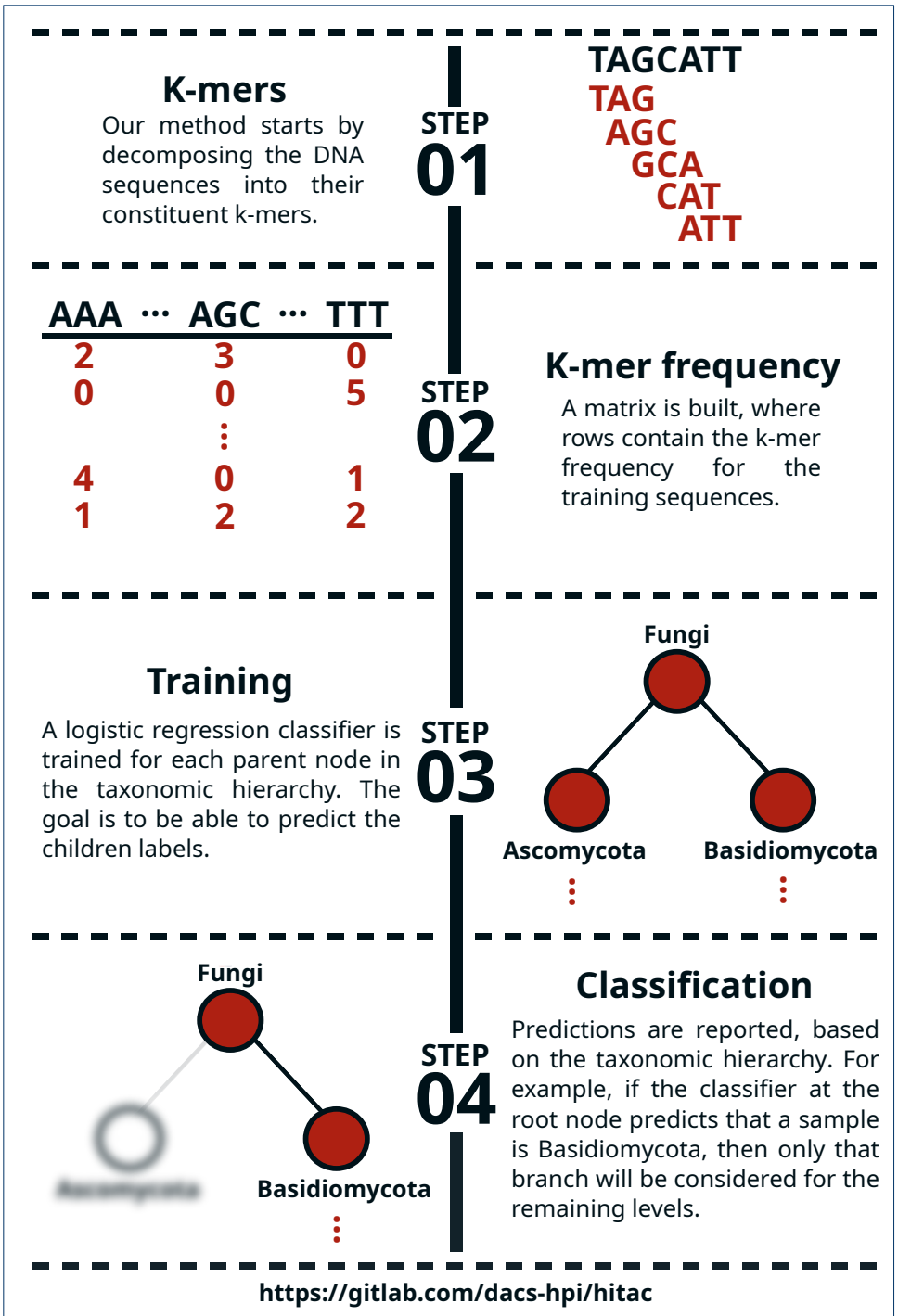

Figure 1: HiTaC algorithm overview. To compute the k-mer frequency of DNA sequences, we first decompose them into their constituent k-mers. In the example shown in step 01, $k = 3$, resulting in 5 k-mers. Afterwards, HiTaC computes a matrix containing the k-mer frequency for all training sequences, where 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.

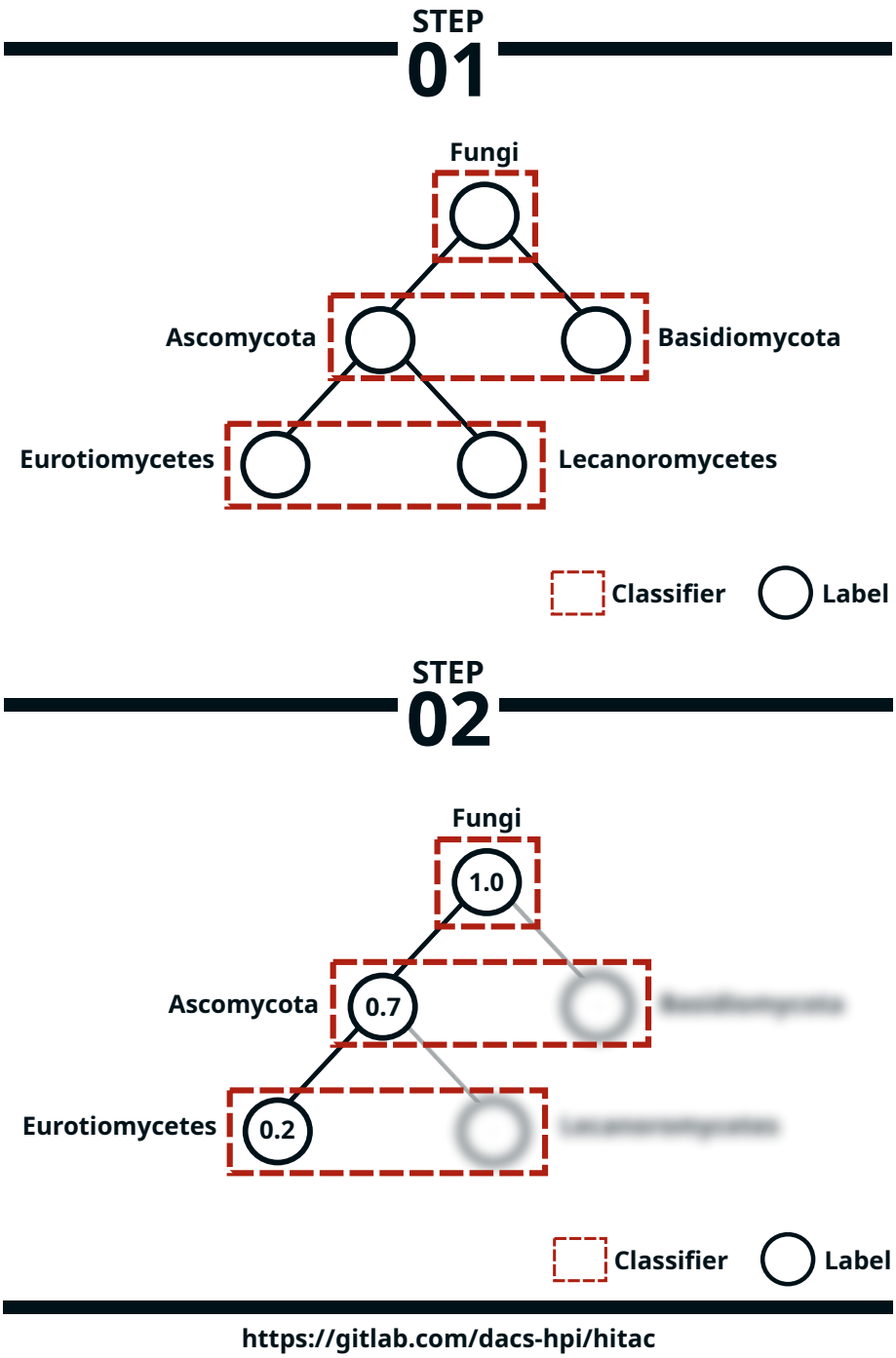

Figure 2: HiTaC filter overview. To build the filter, we first train a logistic regression classifier for each level of the hierarchy, using the same k-mer frequency algorithm introduced previously on Fig. 1. In the example shown in step 01, there are only 3 levels in the hierarchy for simplicity. Afterwards, 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.

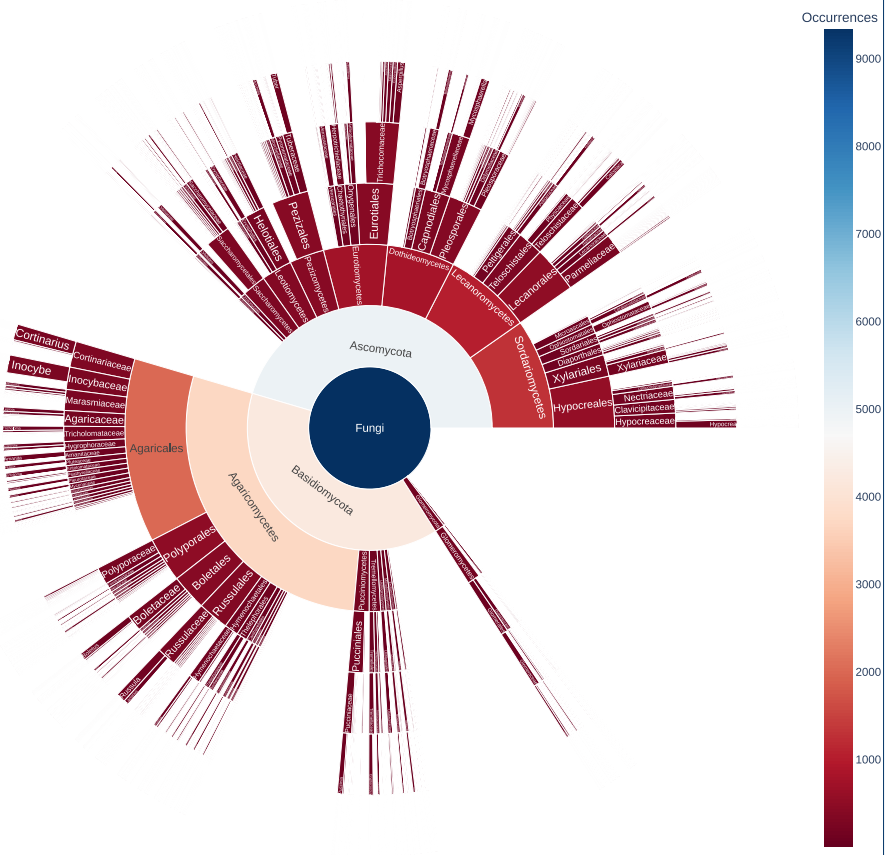

Figure 3: Depiction of imbalanced training set. For instance, the phyla Ascomycota and Basidiomycota together contain approximately 96% of total instances between the seven existing phyla in dataset SP RDP ITS 90. In the remaining taxonomic ranks and datasets we can also observe similar patterns.

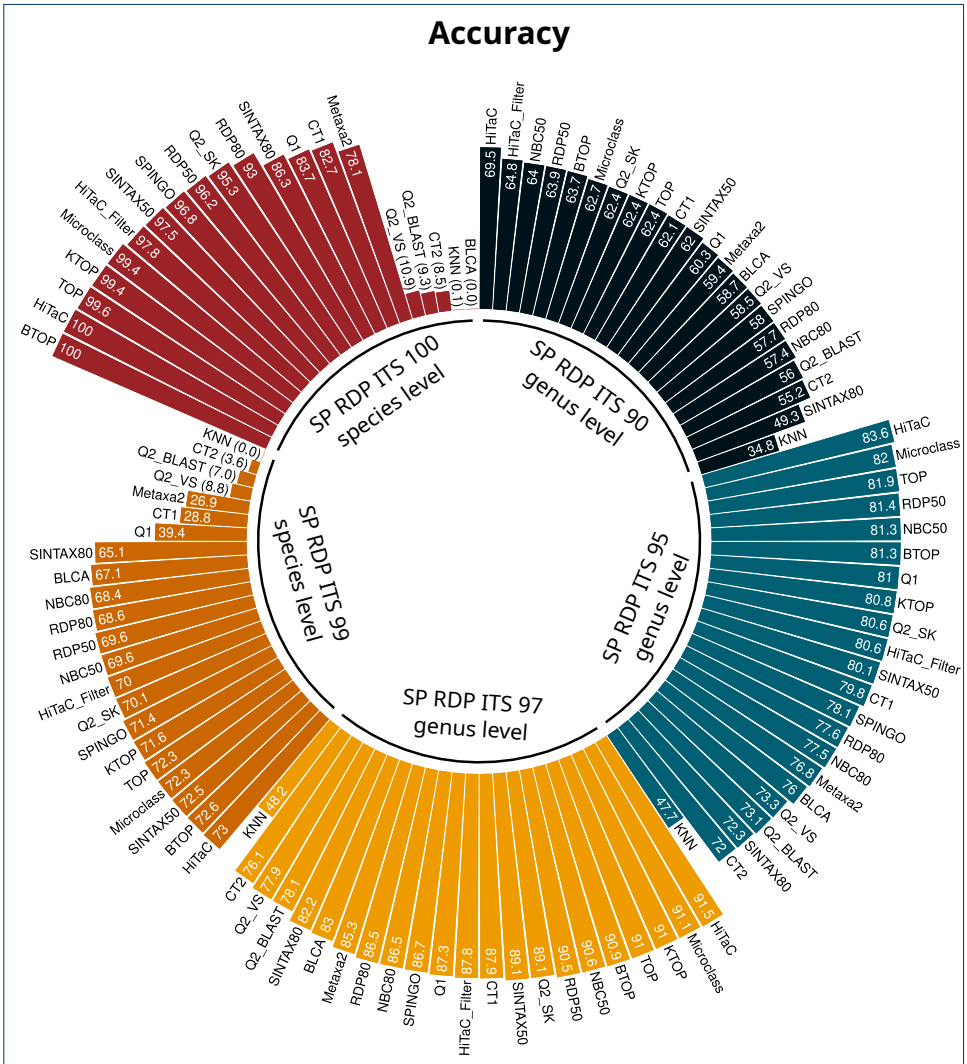

Figure 4: Accuracy results for datasets SP RDP ITS 90 at genus level, SP RDP ITS 95 at genus level, SP RDP ITS 97 at genus level, SP RDP ITS 99 at species level and **(e)** SP RDP ITS 100 at species level. HiTaC and HiTaC_Filtered achieved the highest accuracy for dataset SP RDP ITS 90, and HiTaC outperformed the third best method (NBC50) by 5.5 percentage points. HiTaC surpassed top methods in the remaining datasets SP RDP ITS 95-99, while for dataset SP RDP ITS 100 HiTaC tied with BTOP as the best method. Missing results for NBC50 and NBC80 for dataset SP RDP ITS 100 are due to a memory limit exceeded error.

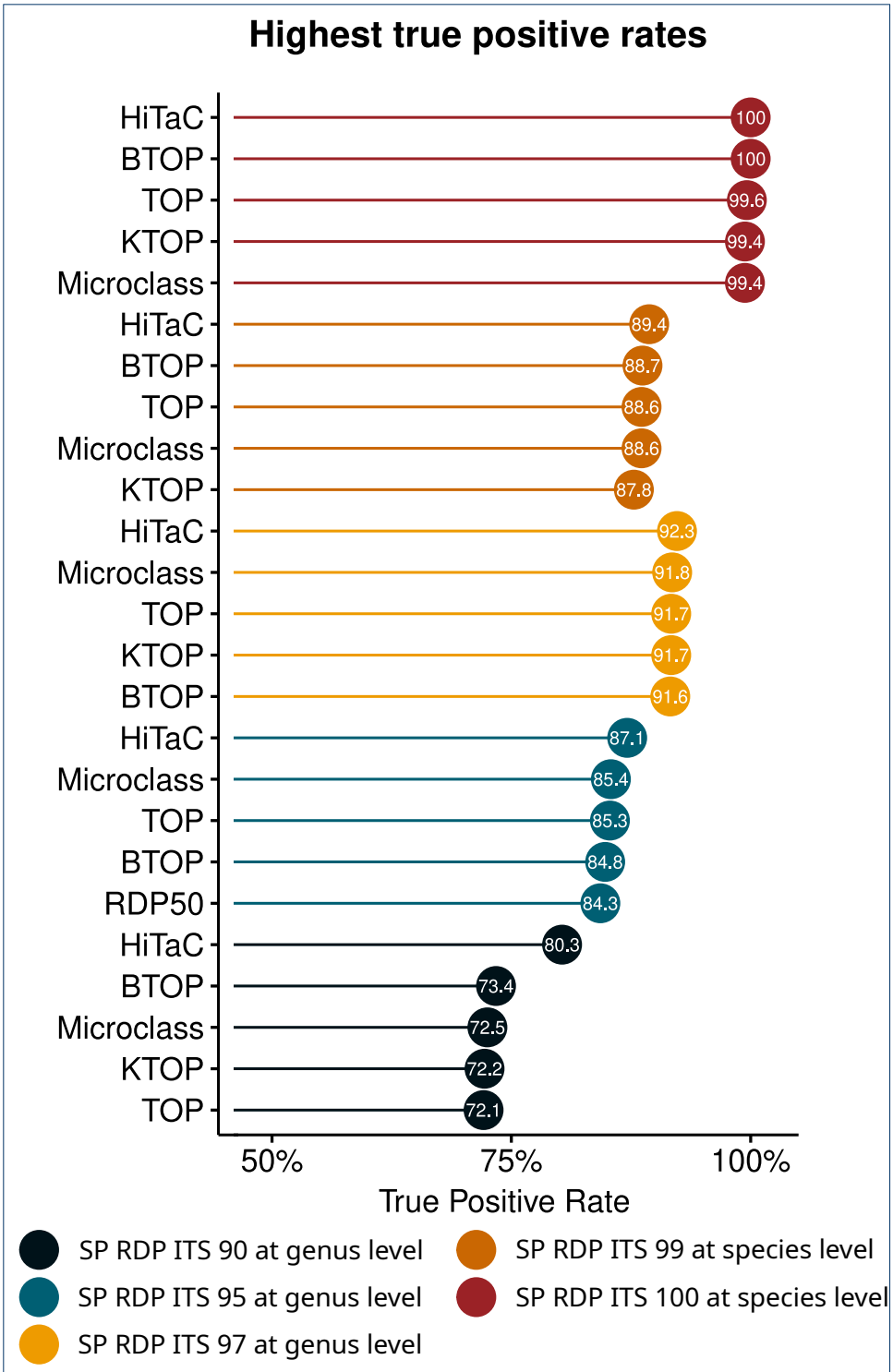

Figure 5: True positive rates for the top 5 methods for datasets SP RDP ITS 90 at genus level, SP RDP ITS 95 at genus level, SP RDP ITS 97 at genus level, SP RDP ITS 99 at species level and SP RDP ITS 100 at species level. The trend is that HiTaC achieved a sensitivity higher or equal to top methods. For instance, HiTaC tied with the top method for the dataset with 100% identity, obtaining a perfect score. Moreover, for the datasets with 90-99% identity HiTaC improved the true positive rates upon top methods.

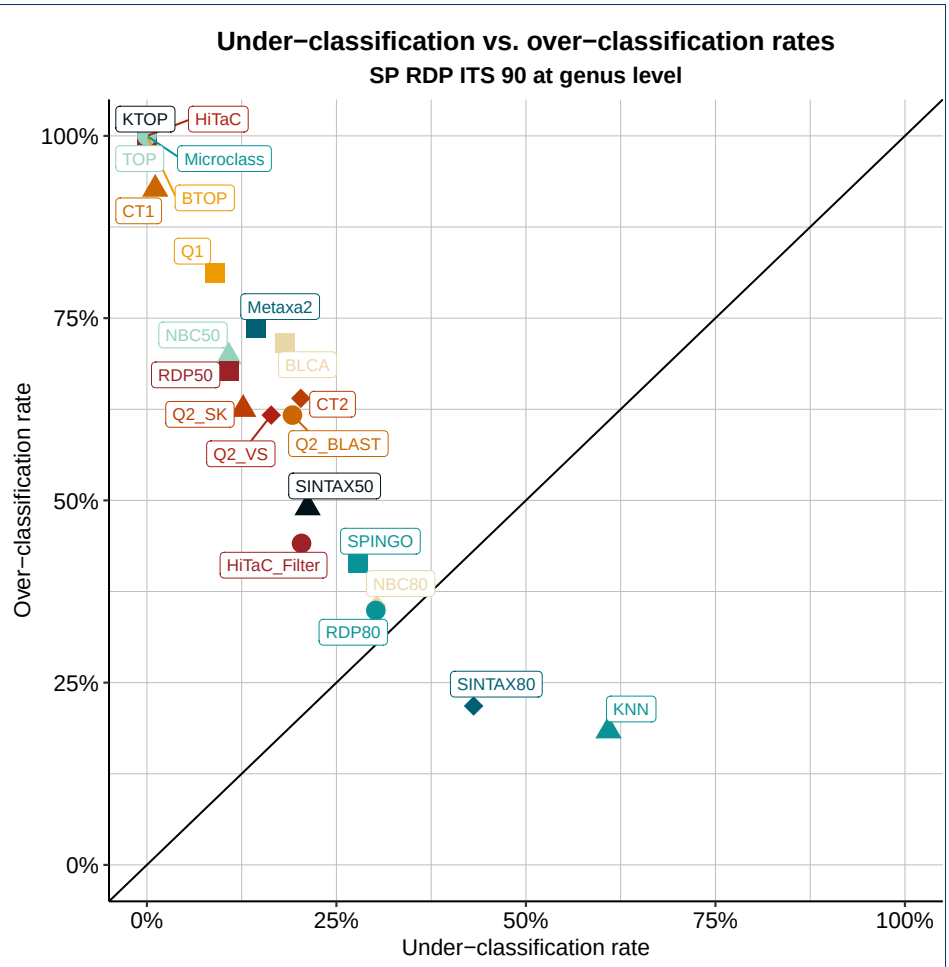

Figure 6: Under-classification versus over-classification rates for dataset SP RDP ITS 90 at genus level. The introduction of the filter in HiTaC proved to be effective by sharply decreasing the over-classification rate from 100% to 44.1%. Contrarily, the filter slightly increased the under-classification rate from 0% to 20.4%, but it still ranked among the methods with the best trade-off between over-classification and under-classification rates, which are the methods closer to the diagonal on the bottom left. An ideal method should be able to classify all known sequences and leave novel sequences unclassified, thus avoiding both over-classification and under-classification errors.

Method	Misclassification rate				
	SP RDP ITS 90 genus level	SP RDP ITS 95 genus level	SP RDP ITS 97 genus level	SP RDP ITS 99 species level	SP RDP ITS 100 species level
BLCA	16.5	10	6.9	7.9	0
BTOP	26.6	15.2	8.4	11.3	0
CT1	27.9	16.1	10.6	28.5	8.7
CT2	19	10.6	11.3	0.9	1.1
HiTaC	19.7	12.9	7.7	10.6	0
HiTaC_Filter	10.3	8.3	5	2.7	0
KNN	3.3	1.9	1.9	0	0
KTOP	27.8	15.8	8.3	12.2	0.6
Metaxa2	19.4	10.5	7.2	2.3	0
Microclass	27.5	14.6	8.2	11.4	0.6
NBC50	18.2	13.3	7.5	13.5	Memory exceeded
NBC80	9.1	8	4.9	7	Memory exceeded
Q1	23	14.7	11.4	21.7	7.8
Q2_BLAST	19.4	12.4	11.1	1.7	1.3
Q2_SK	18.8	11.8	6.7	8.4	0.8
Q2_VS	19.5	11	10.4	2.1	1.3
RDP50	18.6	12.9	7.1	13.2	2.9
RDP80	9	8.3	4.7	6.4	1.5
SINTAX50	12.1	11	7	6.7	0.4
SINTAX80	6	5.4	3.8	1.4	0
SPINGO	10.4	9.1	5	3.5	0
TOP	27.9	14.7	8.3	11.4	0.4

Figure 7: Misclassification rate results for the SP RDP ITS datasets at genus and species level. The misclassification rates for HiTaC, HiTaC_Filter, BTOP, BLCA, SPINGO, SINTAX80, Metaxa2 and KNN were 0% for the dataset SP RDP ITS 100. HiTaC's misclassification rate can be deemed adequate for the datasets with 90–99% identity, while misclassification rate reduced even further when the filter was applied to HiTaC. NBC50 and NBC80 crashed with memory limit exceeded for dataset SP RDP ITS 100, so their results are missing.