Taxonomic Classification of Yeast Using Amplicon Sequencing by Combining UNITE and DITSY Databases

Samir Akre[1,2], Diana H. Taft[2,3], Julian Lopez[2], Kyria Boundy-Mills[2], and David A. Mills[2,3,4].

[1]Department of Biomedical Engineering, [2]Department of Food Science and Technology, [3]Foods for Health Institute, [4]Department of Viticulture and Enology, University of California, Davis, CA 95616.

Abstract

Internal Transcribed Spacer (ITS) sequencing is a current marker gene sequencing target for describing fungi in complex communities. A challenge in classifying yeasts from ITS sequencing is the database used for fungal classification, UNITE, does not focus on yeast taxonomy. A second database, DITSY (Davis ITS Yeast version 2), focuses on yeasts, but does not include enough information on non-yeasts to ensure accurate classification. This project compares the accuracy of UNITE, DITSY, and a hybrid UNITE-DITSY database at classifying ITS sequences. A mock community of ITS sequences dominated by known yeasts was processed using Deblur in QIIME2 and taxonomy assigned with a Bayesian classifier trained with either UNITE, DITSY, or the hybrid database. The classifications were compared to expected frequencies in mock communities at each taxonomic level using the Bray-Curtis distance (0=perfect; 1=poor performance). The Bray-Curtis distribution revealed that at the family level UNITE, DITSY and the hybrid database trained classifiers performed with a mean Bray-Curtis distance of 0.861, 0.878, and 0.734 respectively. The hybrid UNITE-DITSY database produced more accurate classifications of fungi in these tests, suggesting that it will be more effective than the current UNITE database when studying fungal communities containing yeasts.

Introduction

- Yeasts are important members of microbial communities in many food products, and accurate classification of high throughput ITS sequencing is important to studying these communities
- Current databases used for classifying fungi by ITS sequence (UNITE) need updates on yeast taxonomy
- Davis ITS Yeast (DITSY) v2 database contains updated information on yeast ITS
- DITSY contains only limited information on non-yeast fungi, limiting utility when communities contain both yeasts and non-yeasts
- Methods to evaluate classifier performance evaluate each level of taxonomy separately [1]
  - Current methods may over estimate performance at lower taxonomic levels when a misclassification occurs at a higher level

Hypothesis

A UNITE-DITSY hybrid database will provide more accurate classification of short ITS sequences than either UNITE or DITSY databases alone.

Acknowledgements

This work was supported by National Institutes of Health awards R01AT007079, R01AT008759, F32HD093185, the Peter J. Shields Endowed Chair in Dairy Food Science and the Sloan Foundation.

Materials & Methods

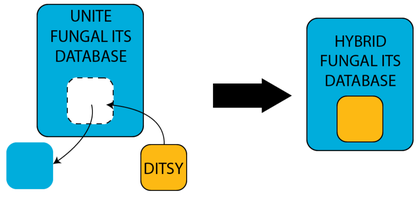

Figure 1: Generating A Hybrid Database
A Hybrid UNITE-DITSY database was developed by removing specific yeast ITS Sequencing data from UNITE and replacing it with cognate DITSY data.

- Mock communities focused on yeast were downloaded from mockrobiota [2]
- ITS Sequence data processed using Deblur[3] in QIIME2[4]
- Taxonomy was used using naïve Bayesian classifiers in QIIME2[5] trained on either UNITE, DITSY, or the UNITE-DITSY hybrid (see Figure 1)

Evaluation Not Accounting for Higher Level Errors

	Expected Results	Classifier Results	Bray-Curtis Distance
Kingdom	a	a	0
Phylum	b	b	0
Class	c	c	0
Order	d	d	0
Family	e	e	0
Genus	f	q	1
Species	Unknown	Unknown	0

Evaluation Accounting for Higher Level Errors

	Expected Result	Classifier Results	Bray-Curtis Distance
Kingdom	a	a	0
Phylum	b	b	0
Class	c	c	0
Order	d	d	0
Family	e	e	0
Genus	f	q	1
Species	Unknown Genus f	Unknown Genus q	1

Figure 2: New Method for Evaluating Classification
The new method of evaluation retains information on the highest taxonomic level a classifier reaches to reduce calculating correct assignment when an error occurred at a higher taxonomic level.

- For each database, a table of expected results was generated based on the known composition of the mock community and the species included in the database
- Actual results were compared to expected results using Bray-Curtis distance, where a value of 0 means the two were identical and a value of 1 means the expected and actual values were completely different

Results

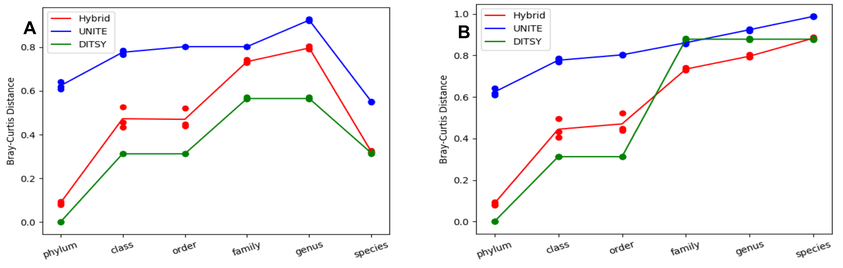

Figure 3: Old and New Evaluation of Performance Trained by UNITE, DITSY, and a Hybrid Database at Classification of Yeast Mock Community
The traditional classification methods (A) resulted in a lower Bray-Curtis value at the species level, the new method (B) more accurately represents performance at the species level. The Hybrid database resulted in a lower Bray-Curtis distance than UNITE alone at almost all taxonomic levels, and a lower distance than DITSY alone at family, genus, and species.

Taxonomic Level	ANOVA P-Value	Group 1	Group 2	p-value
Phylum	9.53E-10	Ditsy	Hybrid	< 0.001
		Ditsy	Unite	< 0.001
		Hybrid	Unite	< 0.001
Class	2.28E-06	Ditsy	Hybrid	0.0263
		Ditsy	Unite	< 0.001
		Hybrid	Unite	< 0.001
Order	1.19E-06	Ditsy	Hybrid	< 0.001
		Ditsy	Unite	< 0.001
		Hybrid	Unite	< 0.001
Family	1.47E-07	Ditsy	Hybrid	0.0268
		Ditsy	Unite	< 0.001
		Hybrid	Unite	< 0.001
Genus	4.77E-07	Ditsy	Hybrid	< 0.001
		Ditsy	Unite	< 0.001
		Hybrid	Unite	< 0.001
Species	6.48E-08	Ditsy	Hybrid	0.293
		Ditsy	Unite	< 0.001
		Hybrid	Unite	< 0.001

Table 1: Tukey's Post-hoc Tests
After ANOVA indicated significant differences in Bray-Curtis values between the expected and actual results obtained using UNITE, DITSY, and the Hybrid database, Tukey's post-hoc test was used to check which of the databases differed significantly. Except for DITSY and the Hybrid database, at the species level, all comparisons were significantly different.

Conclusions

- Hybrid Database resulted in better classification of results than UNITE alone and superior classification of family, genus, and species than DITSY
- New method of evaluating classifiers is a better indicator of performance at species and genus level
- Future work will include testing classifiers on additional mock communities focused on yeasts; communities are currently being sequenced

References

[1] "TAX CREdiT: TAXonomic ClassifieR Evaluation Tool", *GitHub*, 2018. [Online]. Available: https://github.com/caporaso-lab/tax-credit.
[2] N. A. Bokulich *et al.*, "mockrobiota: a Public Resource for Microbiome Bioinformatics Benchmarking," *mSystems*, vol. 1, no. 5, pp. e00062-16, Sep-Oct
[3] A. Amir *et al.*, "Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns," *mSystems*, vol. 2, no. 2, pp. e00191-16, Mar-Apr
[4] "QIIME 2", *Qiime2.org*, 2018. [Online]. Available: https://qiime2.org/.
[5] "qiime2/q2-feature-classifier", *GitHub*, 2018. [Online]. Available: https://github.com/qiime2/q2-feature-classifier.