THE UNIVERSITY OF SYDNEY

A Fast and Accurate Classification Tool for Fungal Species Identification using Genomic Sequences

Vinita Deshpande (A/Prof Michael Charleston, Paul Greenfield)

School of Information Technologies

FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

Introduction

- Fungi are of critical importance due to their wide ranging impacts that are both beneficial and detrimental to humans and the environment.
- Of the 1.5 million species estimated to exist in the fungal kingdom, only about 70,000 have been characterised so far [1].
- DNA sequencing technologies have resulted in unprecedented volumes of fungal genomic data, of which the analysis has not been able to keep up.
- Therefore, the need for **bioinformatics tools that can perform rapid and accurate taxonomic assignment of fungi**, is ever-increasing.

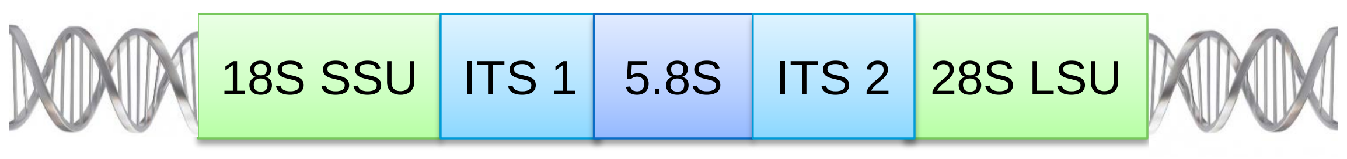
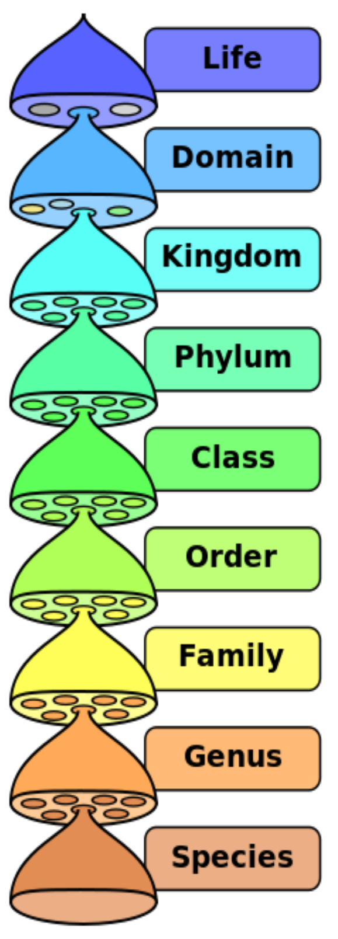


Fig 1. The 18S SSU and 28S LSU ribosomal RNA genes (green) flanking the highly variable Internal Transcribed Spacer (blue), consisting of the ITS1, 5.8S and ITS2 regions.

Contribution

- The Ribosomal Database Project (RDP) Naïve Bayes Classifier [2] uses fungal **28S LSU gene sequences** (Fig 1), and only classifies down to **genus level** (Fig 2).



- We have **built a new Naïve Bayes classifier** using fungal **Internal Transcribed Spacer (ITS) sequences** (Fig 1).
- The **higher sequence variability** and **greater discriminatory power** of the ITS region, compared to the 28S LSU, enables classification of fungi down to the **species level** (Fig 2).

Fig 2. Biological Taxonomy. Image from http://en.wikipedia.org/wiki/Biological_classification

Naïve Bayes Classifier

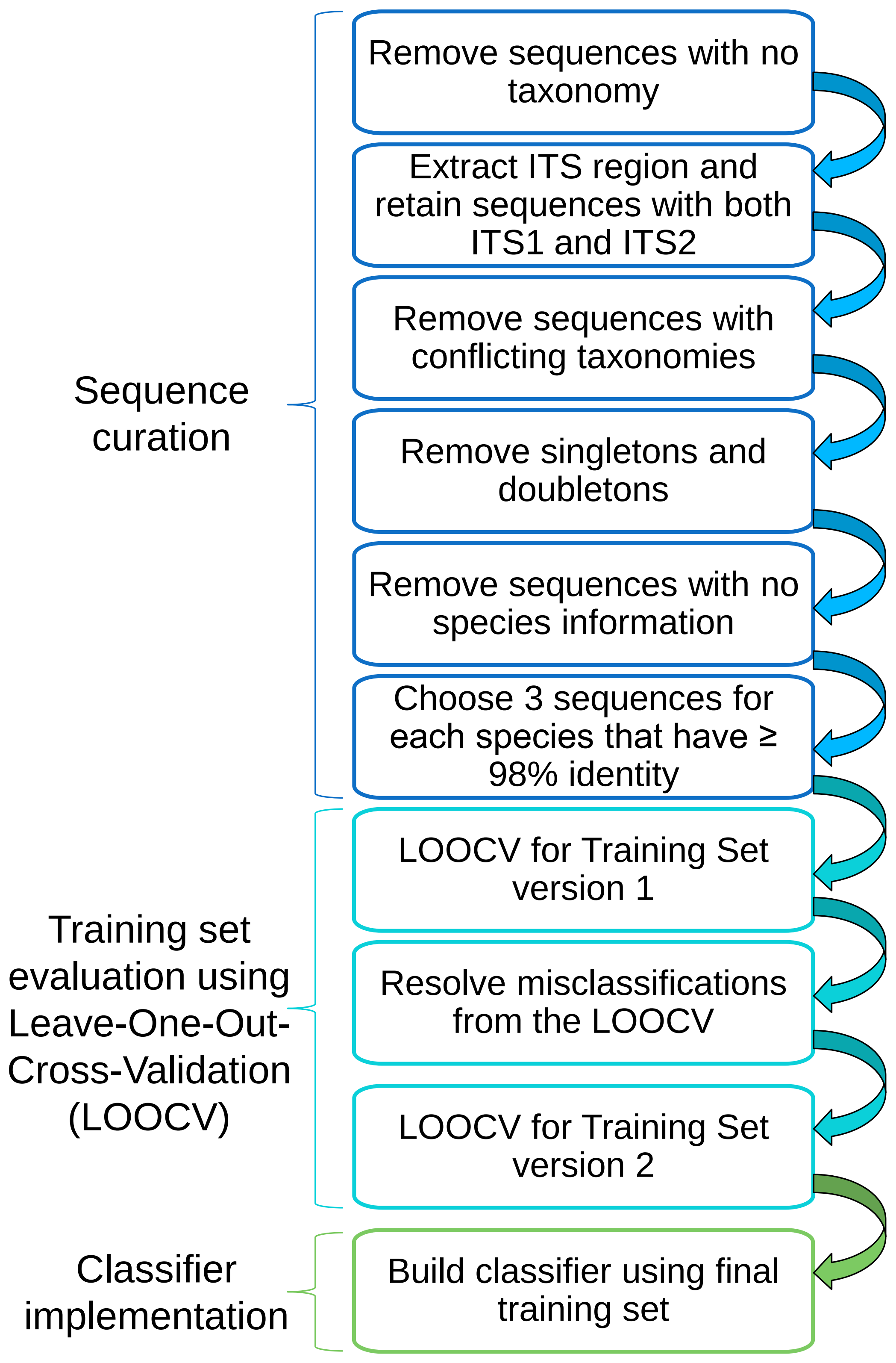
- My classifier is trained by calculating the frequencies of all possible 8-base words (subsequences) from a training set of known sequences.
- The probability that a new query sequence Q is species S is given by **Bayes' Theorem:**

$$P(S|Q) = \frac{P(Q|S) \times P(S)}{P(Q)}$$

- Assignment of Q is made to the species with the highest probability score.
- **Bootstrapping** (sampling with replacement) is performed using 100 trials. The number of times a species is chosen provides a **confidence estimate** of the assignment to that species.

Methods

- A dataset of **343,809 ITS sequences** was downloaded from UNITE (http://unite.ut.ee).
- The sequences were subject to **extensive data processing and curation** as follows:



Sequence curation

Remove sequences with no taxonomy

Extract ITS region and retain sequences with both ITS1 and ITS2

Remove sequences with conflicting taxonomies

Remove singletons and doubletons

Remove sequences with no species information

Choose 3 sequences for each species that have ≥ 98% identity

Training set evaluation using Leave-One-Out-Cross-Validation (LOOCV)

LOOCV for Training Set version 1

Resolve misclassifications from the LOOCV

LOOCV for Training Set version 2

Classifier implementation

Build classifier using final training set

- The final training set had **24,447 high-quality sequences** with **9,073 species.**

Table 1. Phylum Distribution of Training Set.

Phylum	No of Sequences	% of Total
Ascomycota	14,615	59.78%
Basidiomycota	9,195	37.61%
Zygomycota	317	1.30%
Glomeromycota	258	1.06%
Chytridiomycota	47	0.19%
Incertae sedis	8	0.03%
Neocallimastigomycota	5	0.02%
Blastocladiomycota	2	0.01%
Total	**24,447**	**100.00%**

- Ascomycota and Basidiomycota comprise 97.4% of the training set. This is desirable as *most fungal biologists will be working with species in these two phyla*.

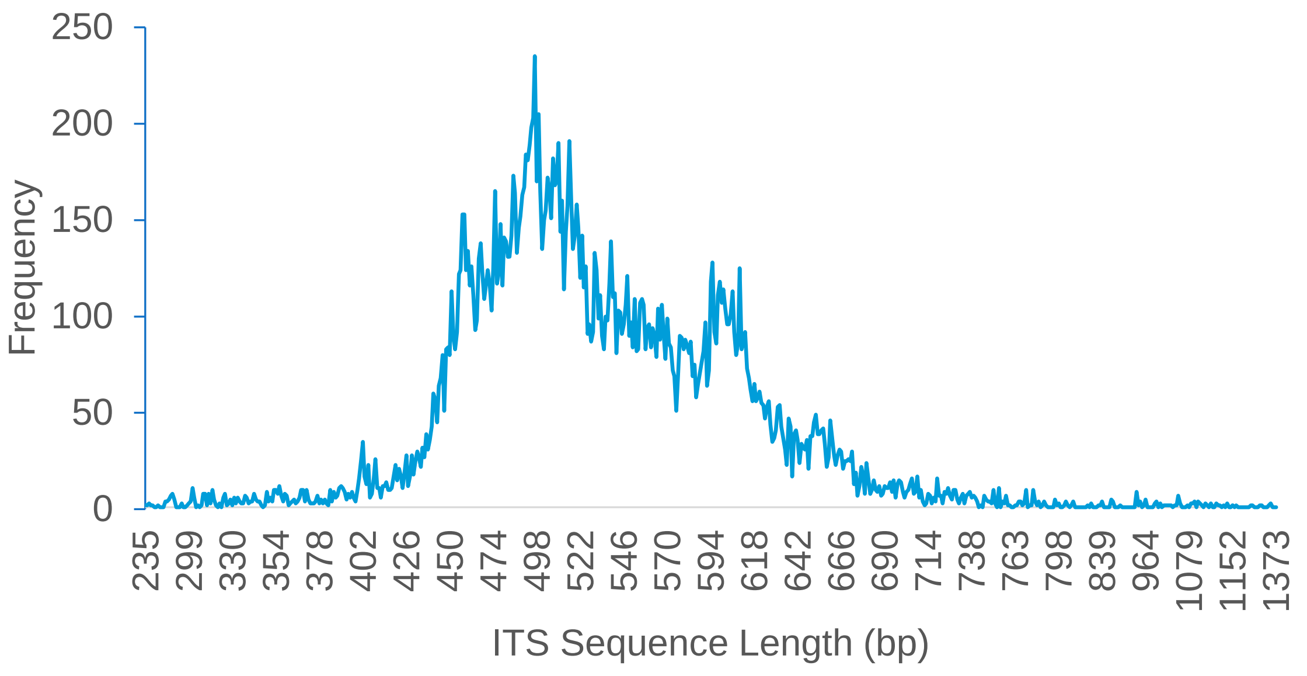


Fig 3. Length Distribution of Training Set.

- The majority of the sequences are between 400 and 700 base pairs (bp) in length, *which falls in the expected range of ITS sequences*.

Results

Training Set Accuracy using LOOCV

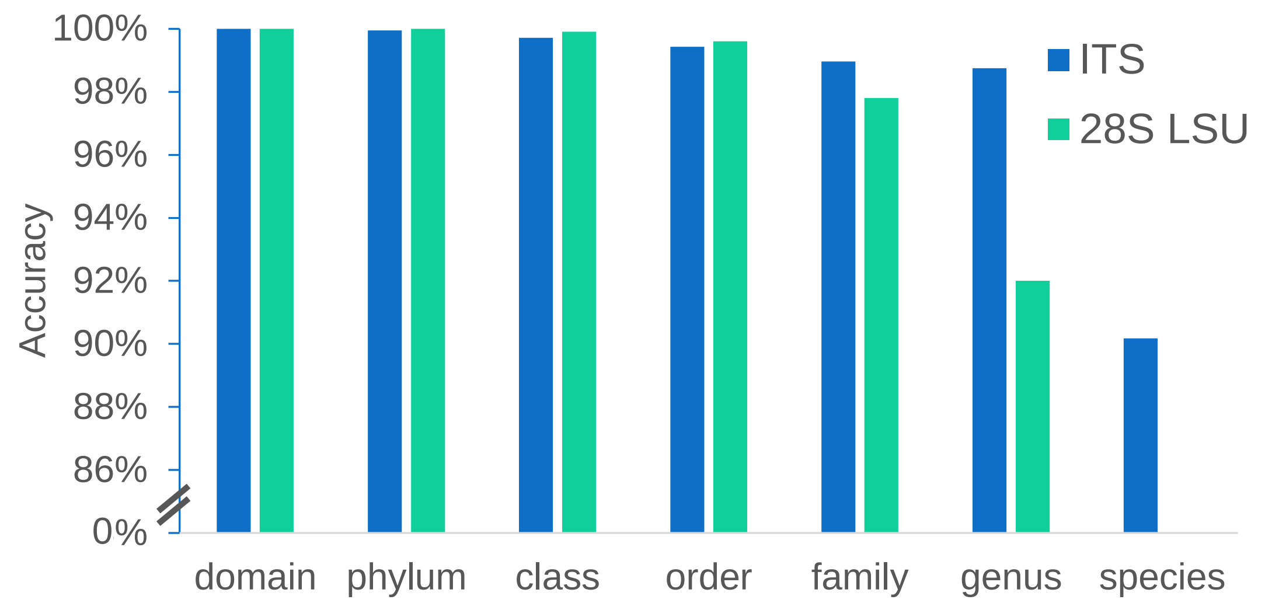


Fig 4. Comparison of LOOCV accuracy of our ITS classifier with the RDP LSU classifier [2].

- The accuracy of our classifier is similar to the LSU classifier down to order (Fig 4).
- At lower ranks, our classifier shows an increase in accuracy of **1.2% at family level to 99%** and **6.8% at genus level to 98.8%**.
- The **species level accuracy is 90.2%**.

Validation Set Accuracy

- The classifier was evaluated using a validation set of **1400 sequences** for **full length ITS**, the **first 400bp** and the **last 400bp** of the ITS region (Fig 5).

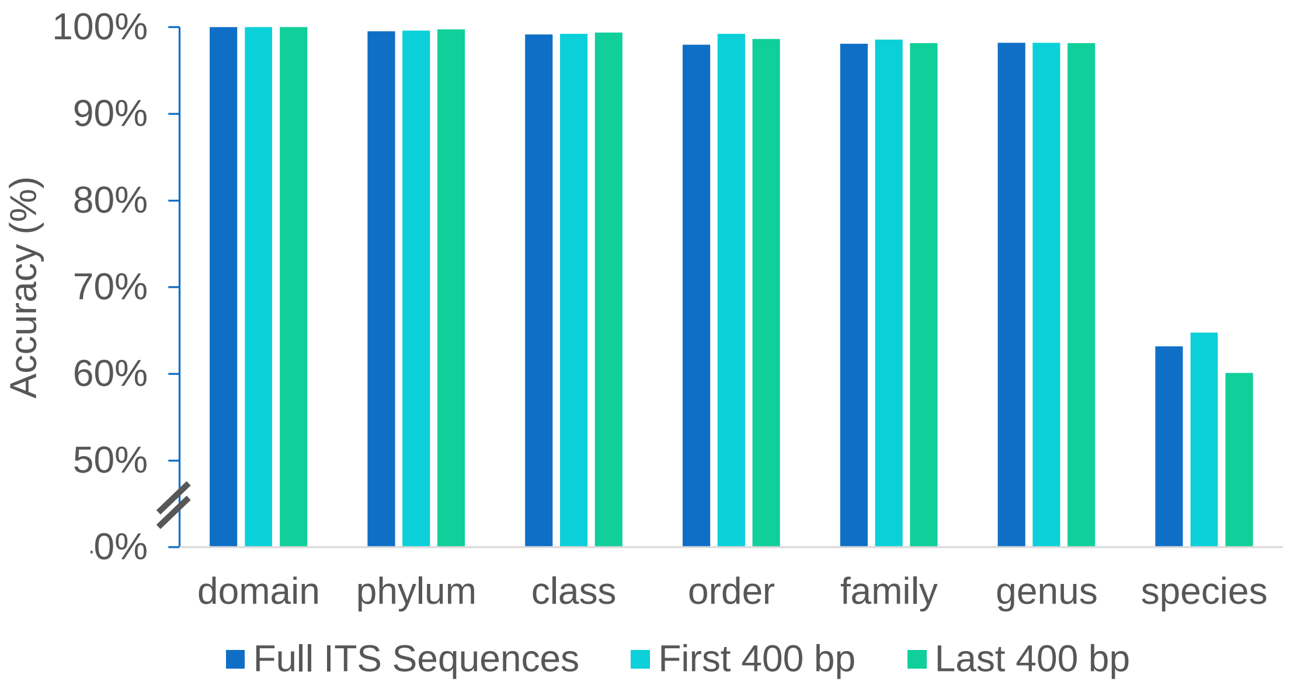


Fig 5. Classifier accuracy for the validation set.

- The results for the short 400bp sequences are comparable to the full length sequences.
- **Accuracies of 98% and 64%** were obtained at **genus and species levels** respectively.

Conclusions

- Our new ITS Classifier is **more accurate** than the current LSU classifier, with **power to resolve down to the species level**.
- The classifier, along with the curated training set, will serve as a valuable asset to fungal biologists for the **rapid and accurate taxonomic assignment of unknown or novel fungal organisms**.

Future Work

- We will include more sequences from phyla that are underrepresented in the training set.
- We will test with different validation methods, e.g., 10-fold Cross Validation.

Acknowledgements

Many thanks to Dr Nai Tran-Dinh and Dr David Midgley, from CSIRO Animal, Food and Health Sciences, for sharing their mycological expertise and for constructing the validation set used to test the classifier.

References

[1] Blackwell M *et al* (2012) Eumycota: mushrooms, sac fungi, yeast, molds, rusts, smuts, etc. http://tolweb.org/Fungi/2377/2012.01.30

[2] Liu KL *et al* (2011) Accurate, Rapid Taxonomic Classification of Fungal Large-Subunit rRNA Genes. *Appl. Environ. Microbiol.* 78(5): 1523–1533

THIS RESEARCH IS SPONSORED BY

CSIRO