An Accurate Fungal Classification Tool and Evolutionary Dynamics of Aflatoxin Genes

FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

Vinita Deshpande | Honours Student

A/Prof Michael Charleston | School of Information Technologies

Paul Greenfield | CSIRO Computational Informatics

THE UNIVERSITY OF SYDNEY

Outline

- **Introduction and Objectives**
- **Part I: Fungal Classification**
  - Background
  - Methods
  - Results
- **Part II: Evolutionary Dynamics of Aflatoxin Genes**
  - Background
  - Methods
  - Results
- **Future Work and Conclusions**

Importance of Fungi

Beneficial Applications

- Ecology
  - Decomposition, recycling of nutrients
- Industry
  - Food fermentation processes
- Medicine
  - Penicillin from *Penicillium notatum*
  - Pharmaceutical drugs

Harmful Effects

- Aflatoxins
  - Carcinogenic toxin naturally produced by *Aspergillus flavus* and *Aspergillus parasiticus*
  - Contaminate food crops consumed by humans
  - 5 billion people worldwide exposed to aflatoxin contamination from their daily diet

Objectives

- A tool that achieves both rapid and accurate classification of novel and unknown fungal organisms

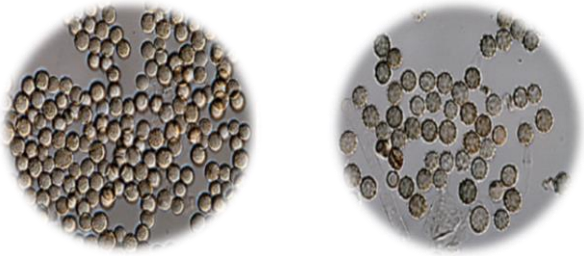
- A greater understanding of the genes responsible for aflatoxin production

PART I

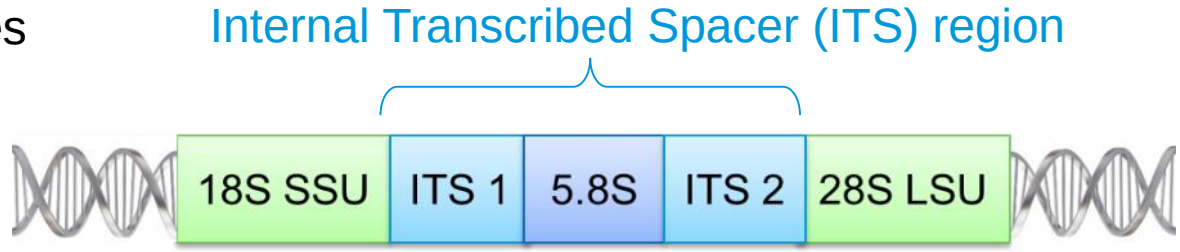
FUNGAL CLASSIFICATION

Gene Sequence-based Classification

- Traditional methods use morphological characteristics
  - Shape
  - Size



- Criteria for a gene target:
  - Universally present in all organisms
  - High variation between species
  - Low variation within species

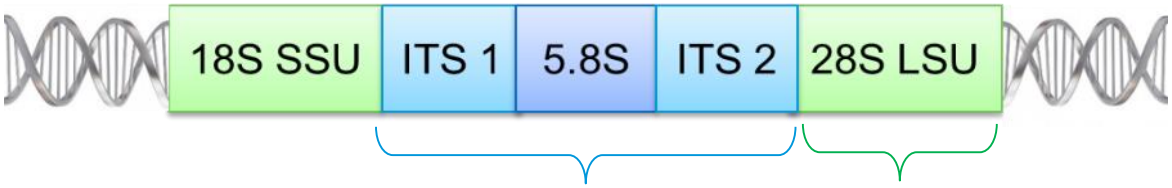


Fungal Classification Tools

- Similarity-based
  - Perform alignment with sequences in reference database
  - Computationally inefficient
  - Statistically unreliable
- Phylogeny-based
  - Use evolutionary relationships
  - Not highly accurate
- Composition-based
  - Machine learning techniques using information about the sequence itself
  - Fast, accurate, reliable

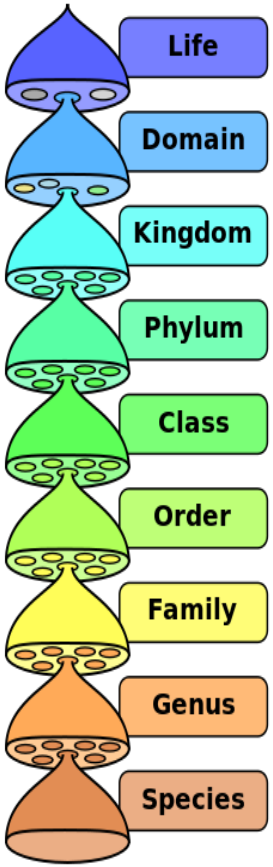
Composition-based: RDP LSU Classifier

- Offered by the Ribosomal Database Project (RDP)



- Uses 28S LSU gene sequences

- Classification to **genus** rank only



http://en.wikipedia.org/wiki/Biological_classification

Naïve Bayes Classifier

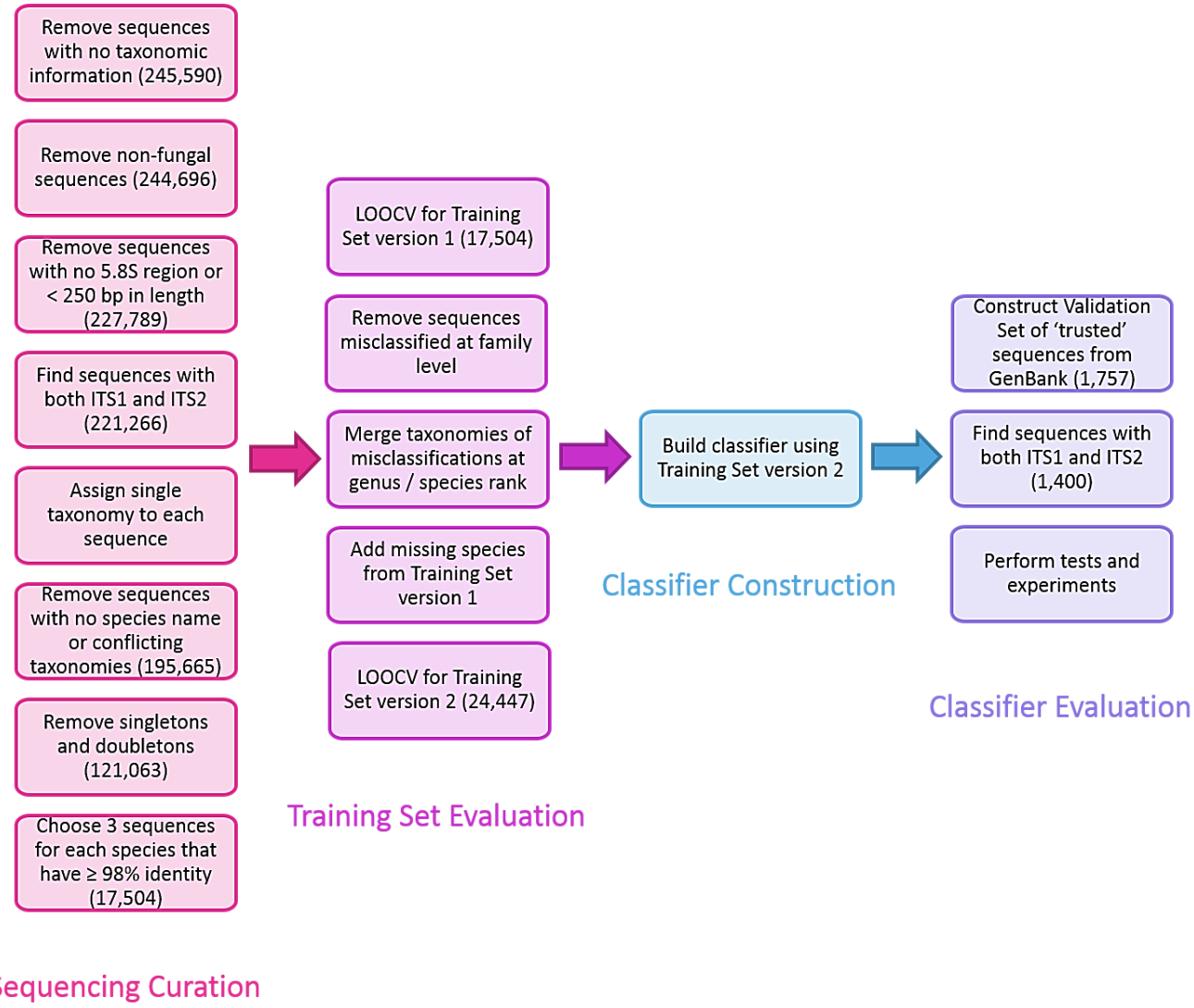
- Statistical-based, supervised classification
- Feature space: 8-base “words” or subsequences = 4^8 = 65,536 features
- Probability that unknown query sequence Q is species S (Bayes’ Theorem):

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

- Bootstrapping with 100 replicates to provide confidence values for each assignment

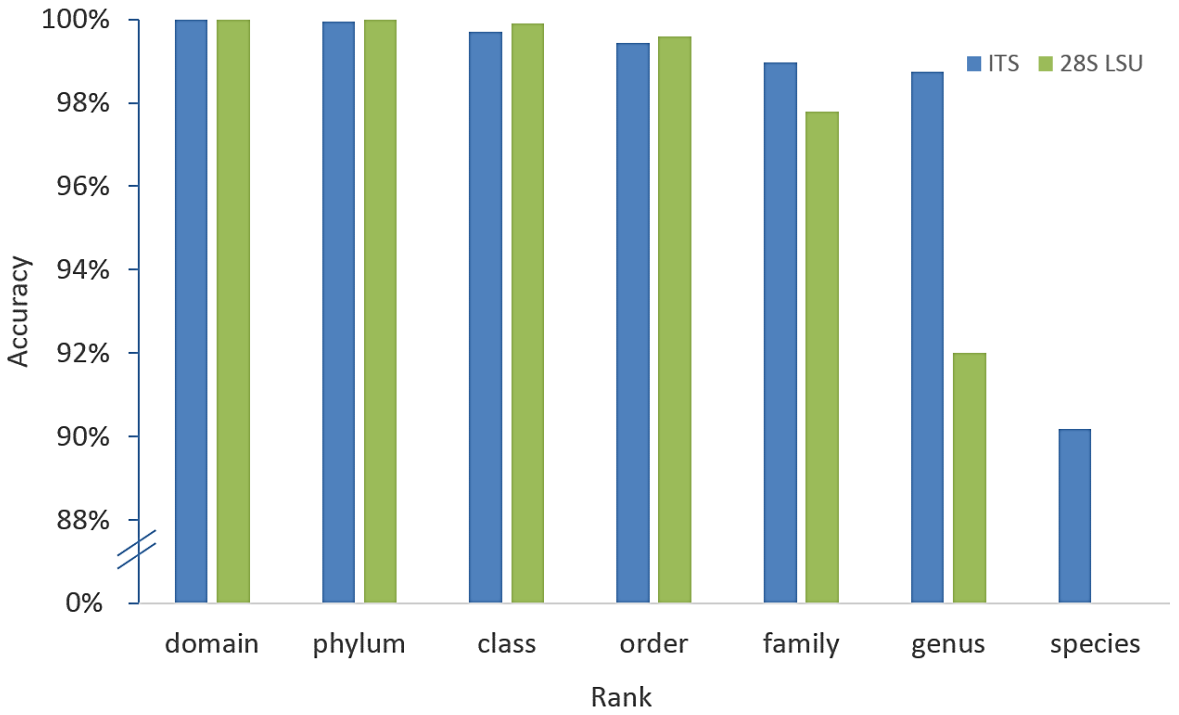
Methods

- Original dataset contained 343,809 sequences
- Final training set contained **24,447 sequences** spanning **9,073 species**



Results: Training Set Accuracy

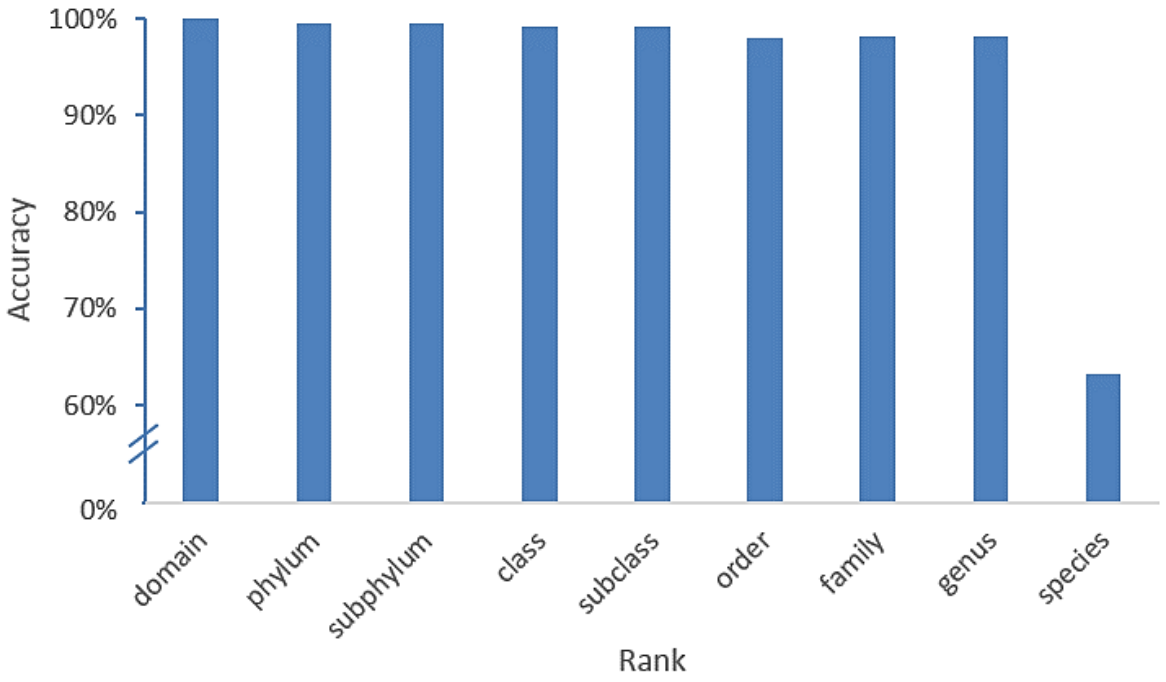
- +1.2% increase from 97.8% to 99.0% at family rank
- +6.8% increase from 92.0% to 98.8% at genus rank
- 90.2% accuracy at species rank



Comparison of LOOCV accuracy of our ITS classifier (blue) with the RDP LSU classifier (green).

Results: Validation Set Accuracy

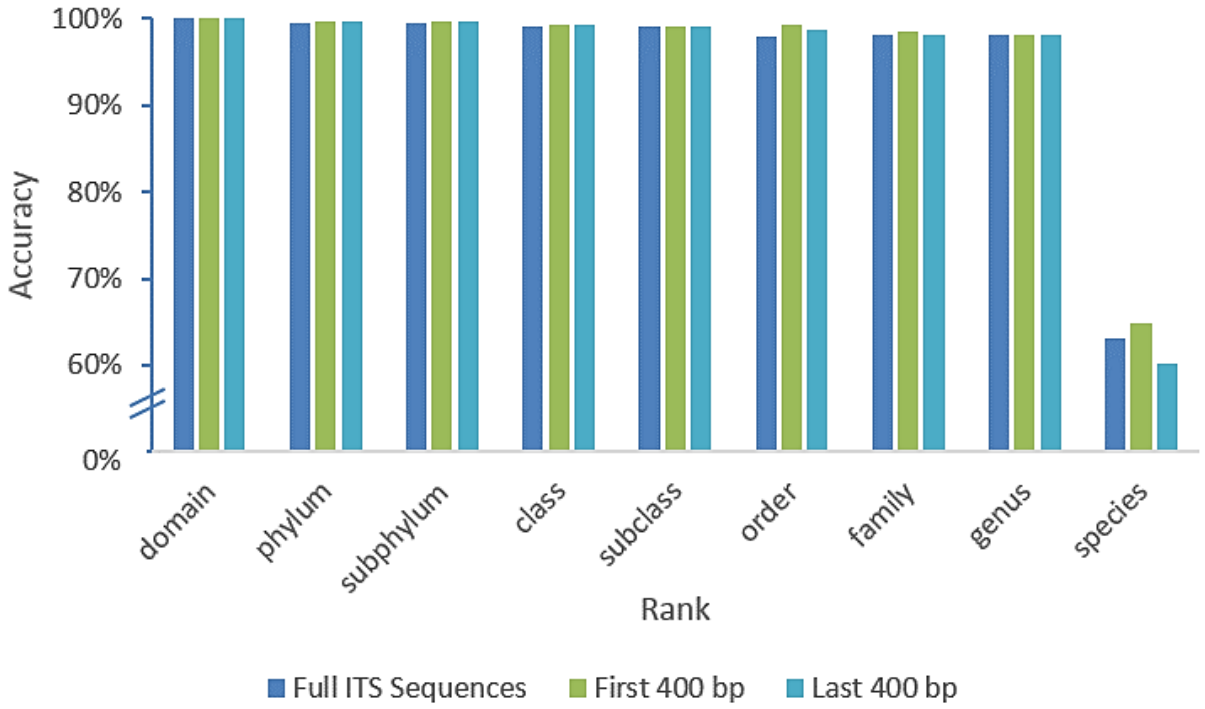
- 1400 test sequences
- > 98% accuracy for genus ranks and above
- 63% accuracy at species level



Validation set accuracy of the ITS classifier.

Results: Amplicon Sequencing Simulation

- > 98% accuracy for genus ranks and above
- 63% accuracy at species level
- Results for short 400 bp sequences comparable to full length sequences up to 1500 bp



Comparison of validation set accuracy of the ITS classifier using full length ITS sequences, the first 400 bp and the last 400 bp.

PART II

EVOLUTIONARY DYNAMICS OF AFLATOXIN GENES

The *Aspergillus* genus

- Food fermentation processes
  - *Aspergillus oryzae*
  - *Aspergillus sojae*

- Aflatoxin-producers
  - *Aspergillus flavus* ← B-type
  - *Aspergillus minisclerotigenes*
  - *Aspergillus parasiticus*
  - *Aspergillus nomius*

  (*Aspergillus minisclerotigenes*, *Aspergillus parasiticus*, *Aspergillus nomius*) → B-type, G-type

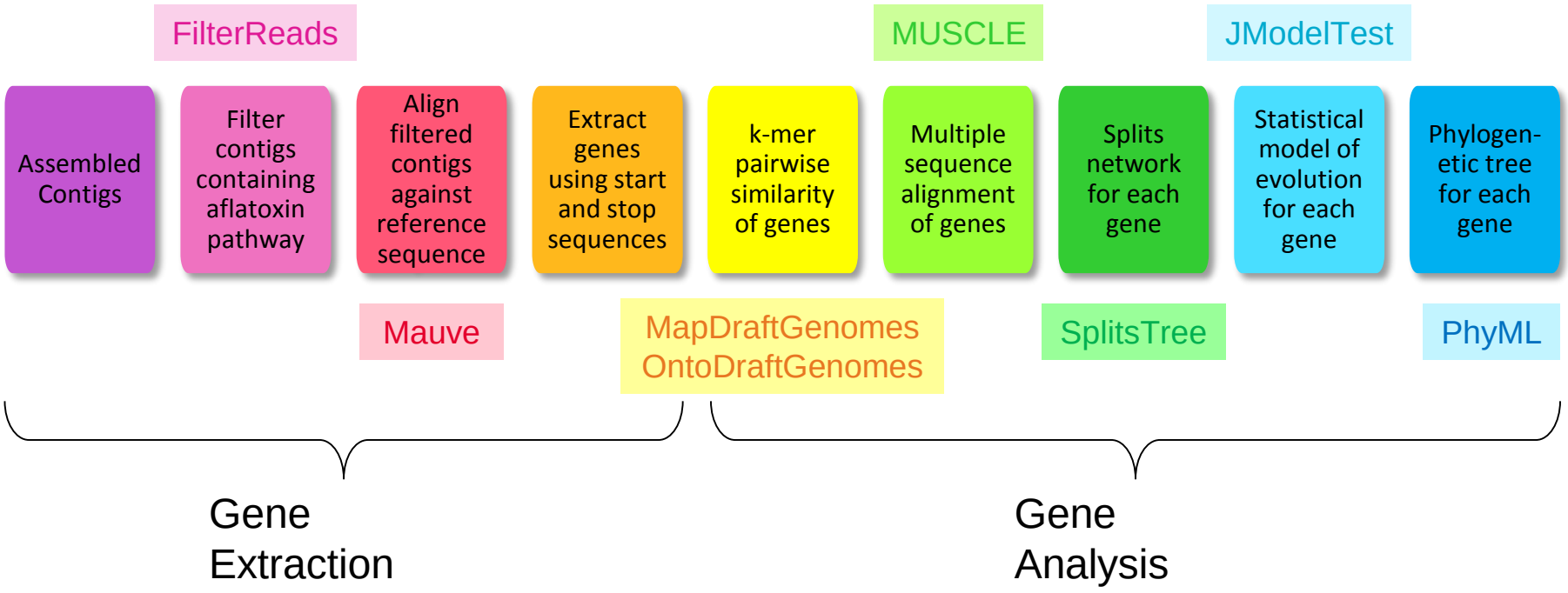
Research Questions

- Why do some strains produce aflatoxins and others do not?
- Why have the aflatoxin genes persisted in genomes of non-aflatoxin producing species?
- What is the exact relationship between *A flavus* and *A oryzae?*

⇨ **Evolutionary dynamics** to understand behaviour at evolutionary level

Methods

- 17 novel fungal genomes
  - 5 *A flavus*, 3 *A flavus-oryzae*, 3 *A minisclerotigenes*, 3 *A parasiticus*, 3 *A nomius*

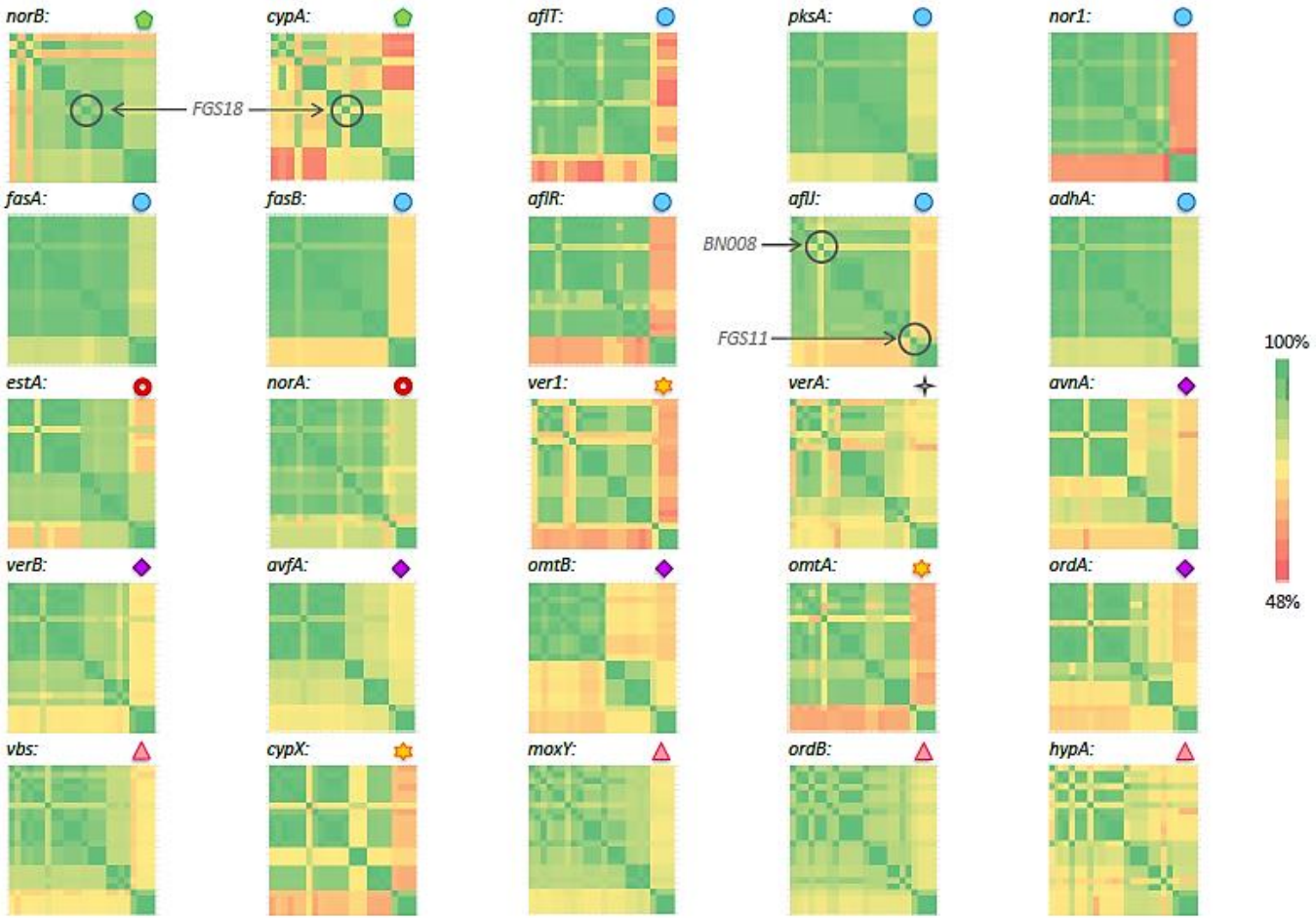


Aflatoxin Pathway Structure

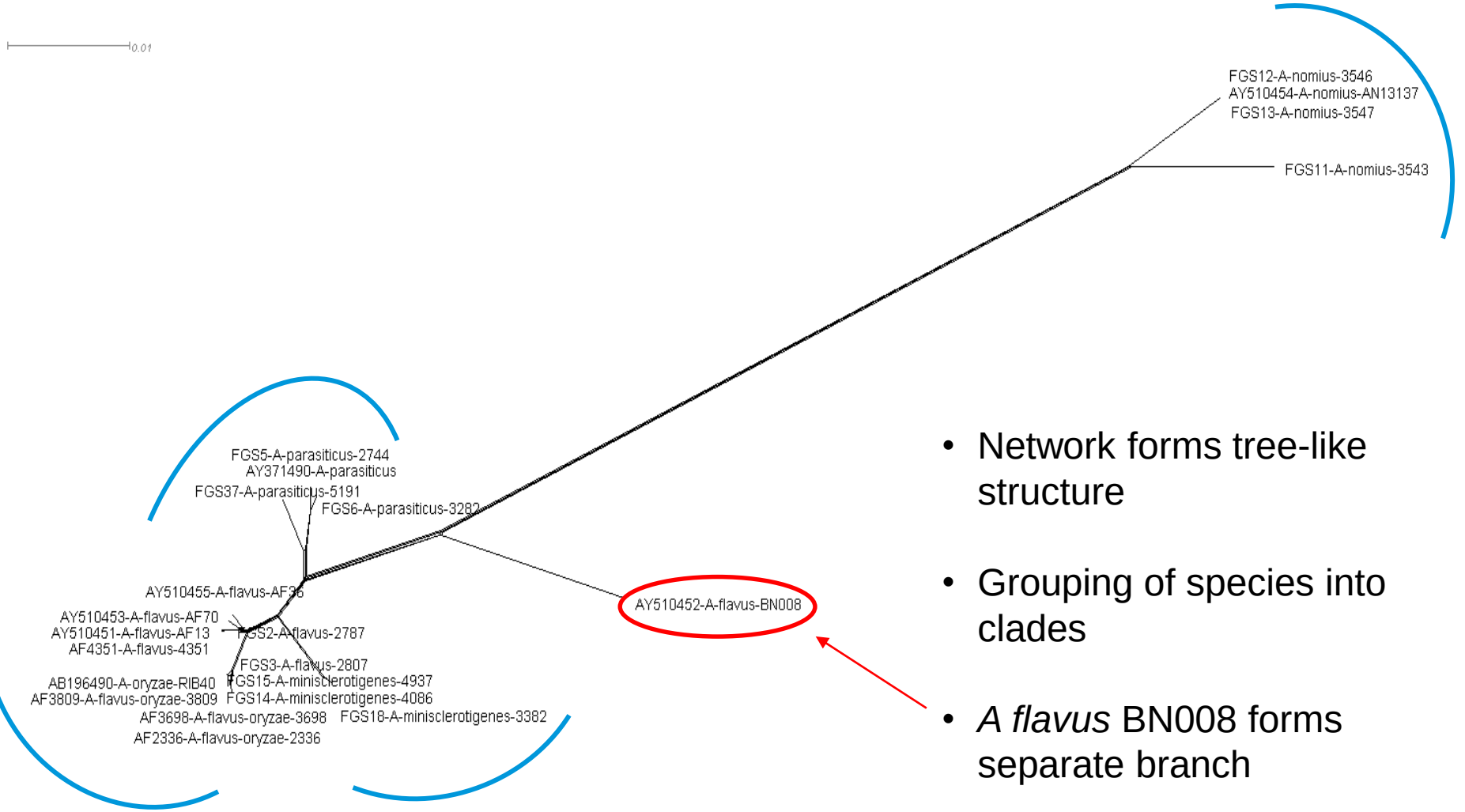
- Conservation of order
- *A flavus-oryzae* also contain full pathway
- Large deletions in FGS1 and FGS20 (both *A flavus*)
- All *A flavus, A flavus-oryzae and FGS18 (A minisclerotigenes)* contain deletions in *norB* and *cypA*

Pairwise Gene Similarities

- Highly diverse patterns
- *A flavus-oryzae* are nearly identical to *A flavus*
- BN008 is highly divergent from all *A flavus* and *A flavus-oryzae*
- FGS18 different to other *minisclerotigenes* for *norB* and *cypA* genes



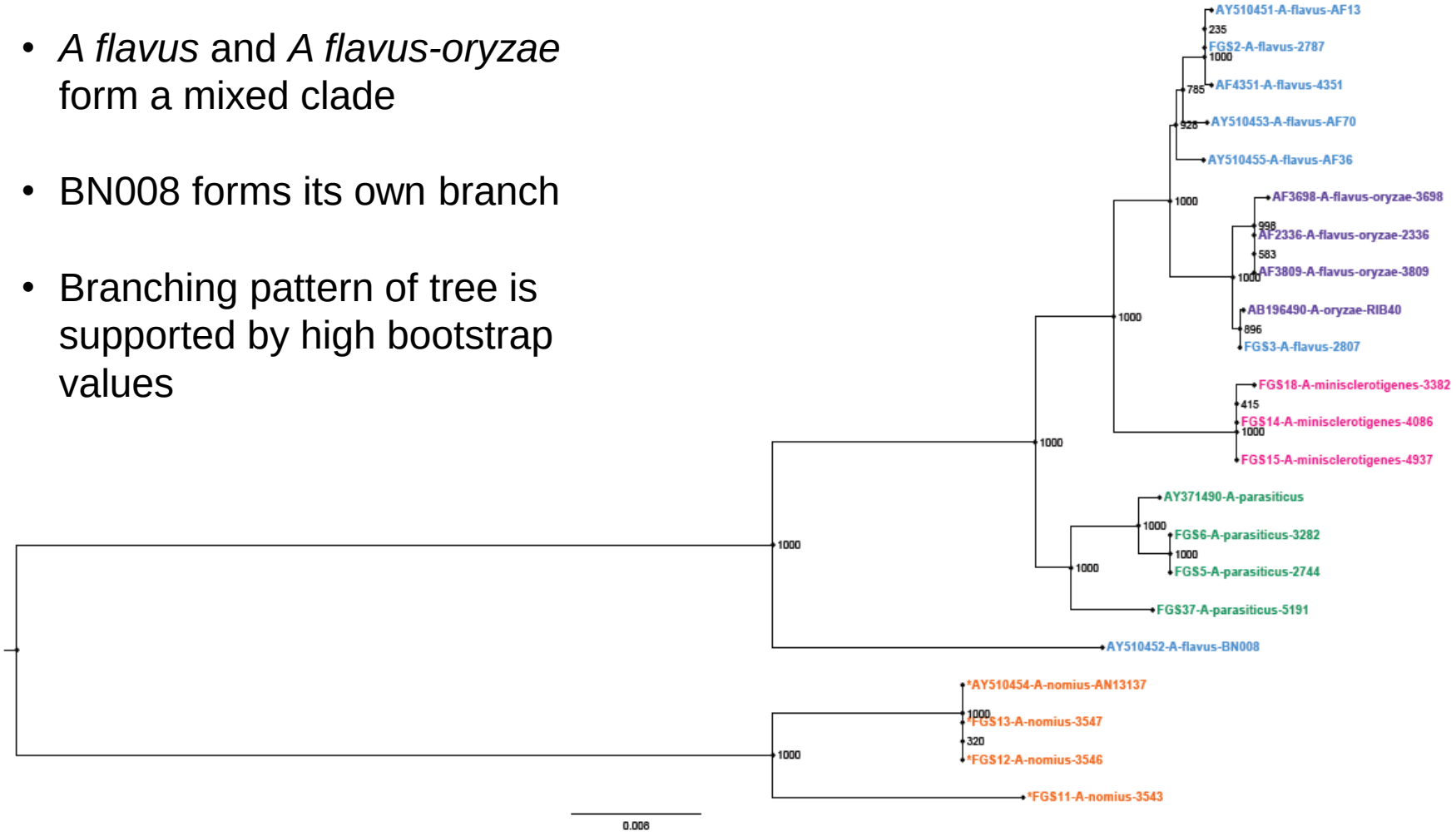
Splits Network: *fasB* gene



- Network forms tree-like structure
- Grouping of species into clades
- *A flavus* BN008 forms separate branch

Phylogenetic Tree: *fasB* gene

- *A flavus* and *A flavus-oryzae* form a mixed clade
- BN008 forms its own branch
- Branching pattern of tree is supported by high bootstrap values



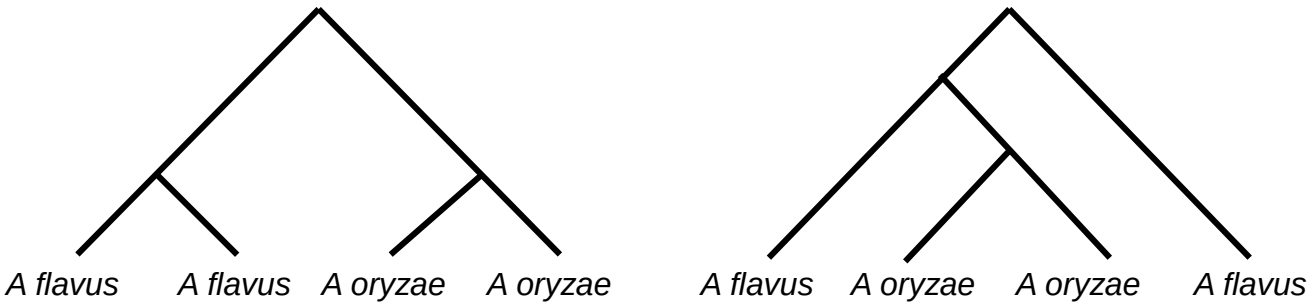
Conclusions and Discoveries

Part I:

- A novel ITS classifier that is rapid and accurate down to the **species** rank
- High quality reference set of ITS sequences

Part II:

- Aflatoxin genes undergoing similar evolutionary processes
- Strain BN008 is not *A flavus,* as labelled
- G aflatoxin producing ability of FGS18 (*A minisclerotigenes*) doubtful
- *A flavus* and *A oryzae* are different strains of the same species



Future Work

Part I:

- Another iteration of training set
- Test with different values for word size
- Test with different validation methods, e.g. 10-fold Cross Validation
- Classifier using combined ITS + LSU genes

Part II:

- Extend to all 25 genes
- Use a distance metric to cluster matrices
- Produce concatenated gene trees

What next?

- Journal publications of our findings

- Collaborate with RDP LSU Classifier developers to make our ITS classifier publicly available

THE UNIVERSITY OF SYDNEY

THANK YOU