Machine learning models accurately predict clades of proteocephalidean tapeworms based on host and zoogeographical data

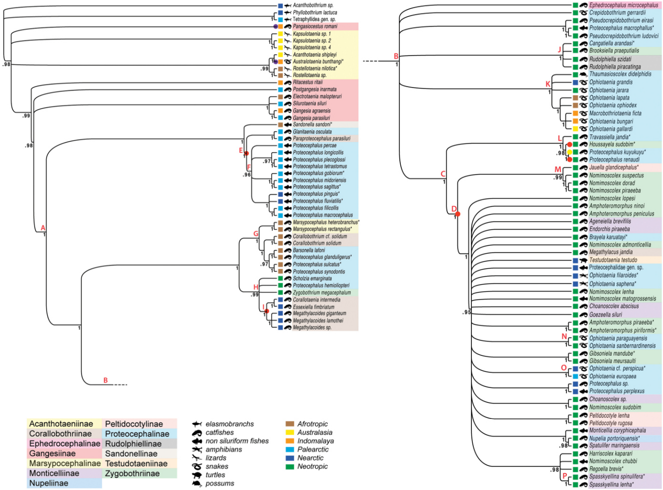
PVA, RJdS, AdC, JLL, AD, DJ, DJM

UNESP, MHNG, UFRRJ & UNC Charlotte

ASP 99th Annual Meeting – 6/16/24

Philippe Vieira Alves, Reinaldo José da Silva, Alain de Chambrier,
José Luis Luque, Anastasiia Duchenko, Daniel Janies,
& **Denis Jacob Machado**

Characters of difficult optimization

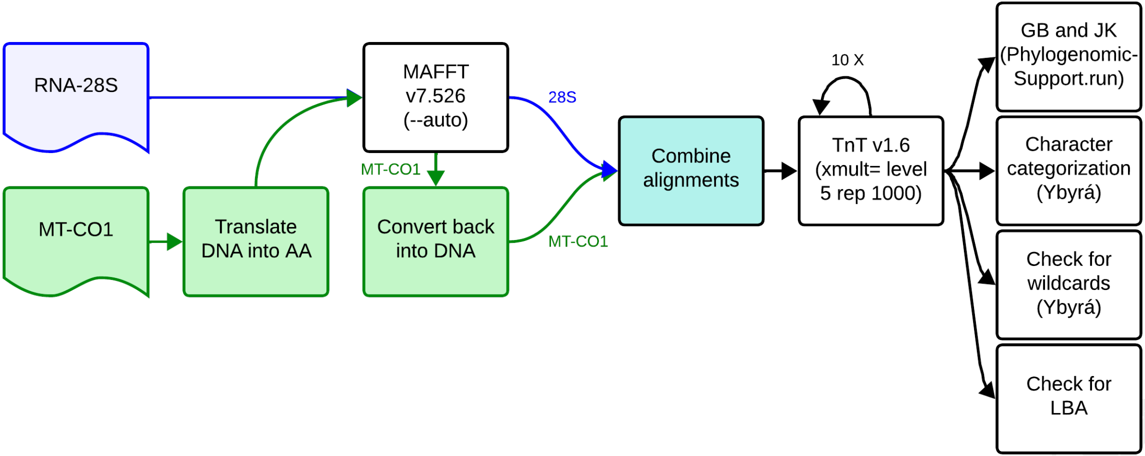


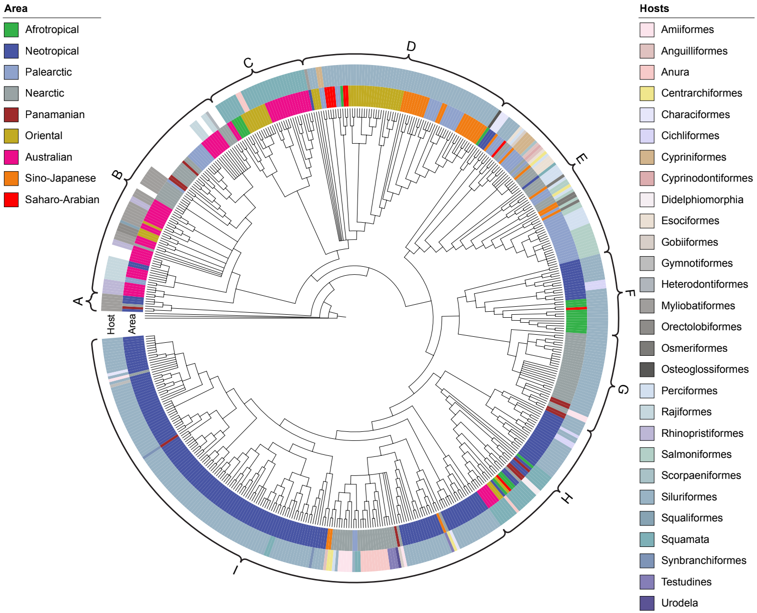
de Chambrier *et al.* (2015; doi: 10.3897/zookeys.500.9360).

Species selection and DNA data

- Our matrix of **28S rRNA** (510 terminals) and **MT-CO1** (253 terminals) contained a total of **537 terminals**.
- 58 terminals were sequenced for the first time to generate **85 new sequences** (56 for 28S and 29 for MT-CO1).
- This matrix represents **222 parasite species** from **194 host species**.
- Our **outgroup** (**87 terminals**) comprises *Acanthobothrium* (18 species), *Clistobothrium* (1; our root), *Matticestus* (2), *Pachybothrium* (1), and *Potamotrygonocestus* (2).
- Our **ingroup** (**450 terminals**) contains 63 genera of proteocephalids.

Phylogenetics workflow


RNA-28S
MT-CO1
Translate DNA into AA
MAFFT v7.526 (--auto)
MT-CO1
Convert back into DNA
28S
MT-CO1
Combine alignments
10 X
TnT v1.6 (xmult= level 5 rep 1000)
GB and JK (Phylogenomic-Support.run)
Character categorization (Ybyrá)
Check for wildcards (Ybyrá)
Check for LBA

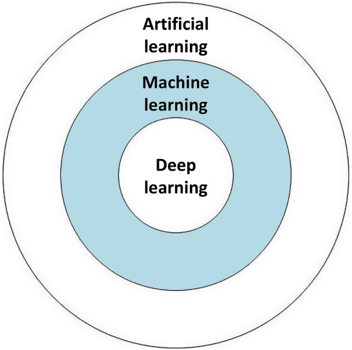

Area
Afrotropical
Neotropical
Palearctic
Nearctic
Panamanian
Oriental
Australian
Sino-Japanese
Saharo-Arabian
Hosts
Amiiformes
Anguilliformes
Anura
Centrarchiformes
Characiformes
Cichliformes
Cypriniformes
Cyprinodontiformes
Didelphiomorphia
Esociformes
Gobiiformes
Gymnotiformes
Heterodontiformes
Myliobatiformes
Orectolobiformes
Osmeriformes
Osteoglossiformes
Perciformes
Rajiformes
Rhinopristiformes
Salmoniformes
Scorpaeniformes
Siluriformes
Squaliformes
Squamata
Synbranchiformes
Testudines
Urodela
A
B
C
D
E
F
G
H
I
Host
Area

Host and biogeographical data

Ten different features (5,040 data points):

- Host taxonomy:
  - class (5)
  - order (29)
  - family (66)
  - genus (120)
  - species (176)
- Environment and habitat:
  - terrestrial or aquatic (2)
  - freshwater, brackish, or saltwater (3)
- Locality:
  - zoogeographical region (10)
  - continent (7)
  - country or river basin (42)

What are machine learning and random forests?

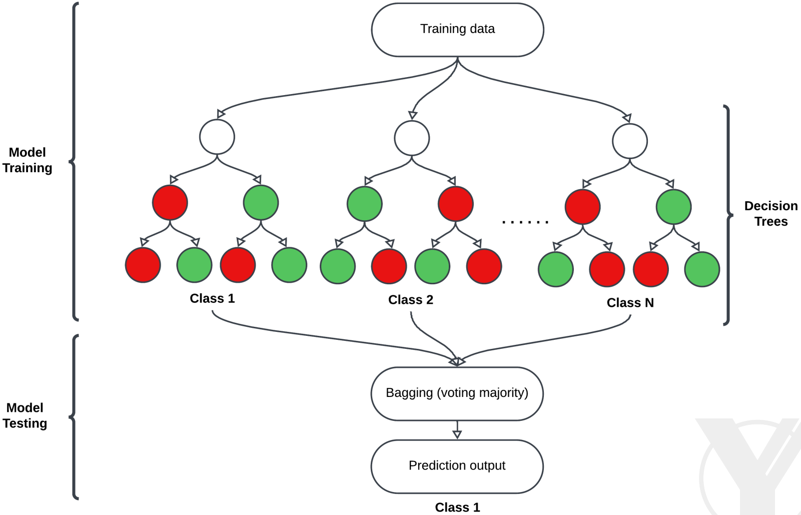


Artificial learning

Machine learning

Deep learning

Modified from Fig. 1 from Jakhar & Kaur (2020; DOI: 10.1111/ced.14029).



Training data

Model Training

Class 1

Class 2

Class N

Decision Trees

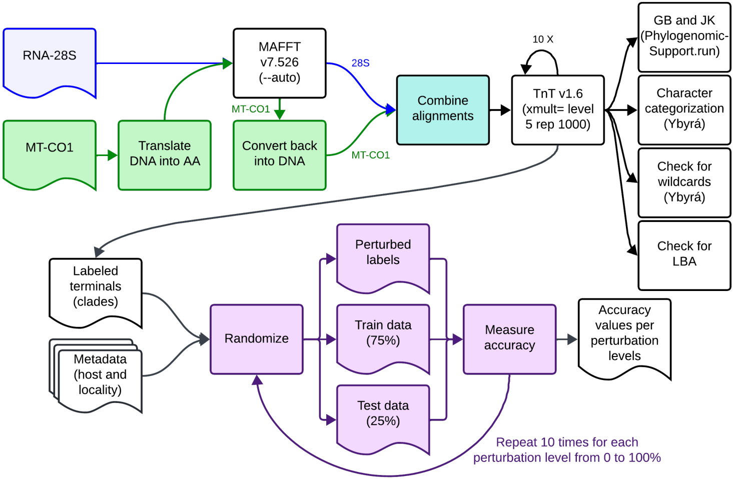
Model Testing

Bagging (voting majority)

Prediction output

Class 1

Our random forest experiment



RNA-28S

MT-CO1

Translate DNA into AA

MAFFT v7.526 (--auto)

MT-CO1

Convert back into DNA

28S

MT-CO1

Combine alignments

10 X

TnT v1.6 (xmult= level 5 rep 1000)

GB and JK (Phylogenomic-Support.run)

Character categorization (Ybyrá)

Check for wildcards (Ybyrá)

Check for LBA

Labeled terminals (clades)

Metadata (host and locality)

Randomize

Perturbed labels

Train data (75%)

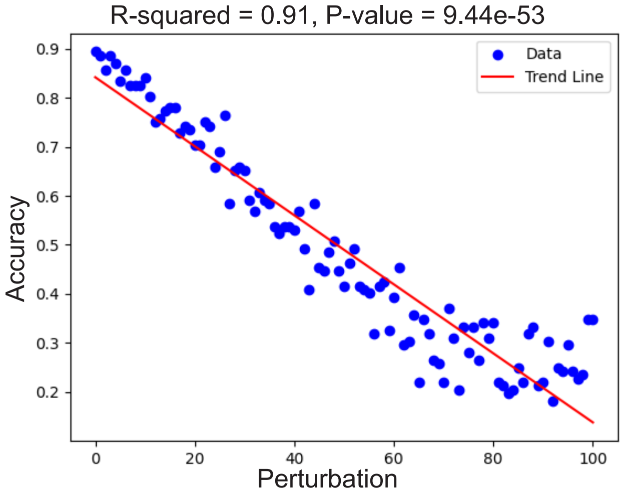
Test data (25%)

Measure accuracy

Accuracy values per perturbation levels

Repeat 10 times for each perturbation level from 0 to 100%

The effect of clade perturbation over accuracy



Acknowledgements:

FAPESP
FUNDAÇÃO DE AMPARO À PESQUISA
DO ESTADO DE SÃO PAULO

unesp

UNIVERSITY OF NORTH CAROLINA
CHARLOTTE

muséum genève

UFRRJ
UNIVERSIDADE FEDERAL RURAL
DO RIO DE JANEIRO

Funding:

FAPESP proc. no. 2023/00714-5

Contact:

Dr. Denis Jacob Machado
UNC Charlotte
Dept. of Bioinformatics and Genomics
CIPHER center

Email: dmachado@charlotte.edu
Lab page: phyloinformatics.com
Zenodo: https://doi.org/10.5281/zenodo.11307234

phyloinformatics.com

zenodo

Example application of random forests

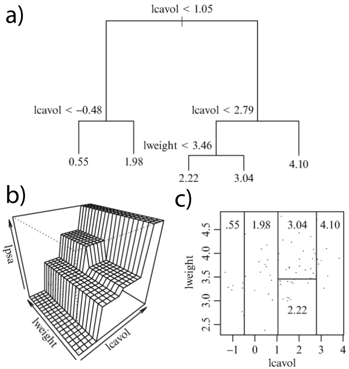


Fig. 3 from Cutler *et al*. (2012; DOI: 10.1007/978-1-4419-9326-7_5).

Prostate cancer data comes from a prostate cancer study (Stamey *et al*. 1989; Hastie *et al*. 2009).

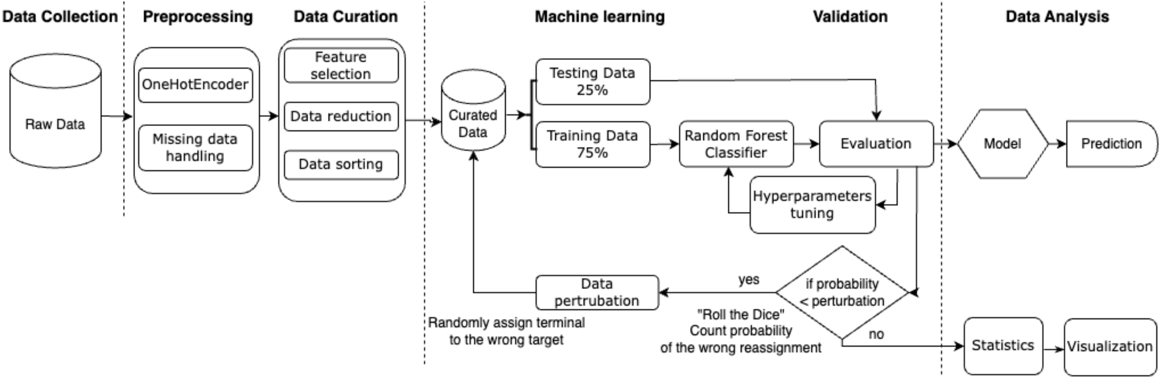
a) Tree diagram.

b) A perspective plot of the fitted regression surface.

c) Partitioning of the predictor space.

Response variable: level of prostate-specific antigen (*lpsa*). **Predictor variables:** log cancer volume (*lcavol*), log prostate weight (*lweight*), age, log of the amount of benign prostatic hyperplasia (*lbph*), seminal vesicle invasion (*svi*), log of capsular penetration (*lcp*), Gleason score (*gleason*), and percentage of Gleason scores 4 or 5 (*pgg45*).

A closer view into our random forests



Data Collection

Raw Data

Preprocessing

OneHotEncoder

Missing data handling

Data Curation

Feature selection

Data reduction

Data sorting

Curated Data

Machine learning

Testing Data 25%

Training Data 75%

Random Forest Classifier

Hyperparameters tuning

Validation

Evaluation

if probability < perturbation

yes

no

Data pertrubation

Randomly assign terminal to the wrong target

"Roll the Dice" Count probability of the wrong reassignment

Data Analysis

Model

Prediction

Statistics

Visualization