raxtax

A k-mer-based non-Bayesian Taxonomic Classifier

Noah A. Wahl, Georgios Koutsovoulos, Ben Bettisworth and Alexandros Stamatakis
Biodiversity Computing Group, ICS-FORTH

Taxonomic Classification

Unknown Query Sequence

AATACGTCCTAGAC

Known Reference Sequences

AAATCGTCCTGAA	Species 1
TTACGTTCCTGAA	Species 2
AATCGTTCGTGAA	Species 3
…	…
AAACGTACTTGAC	Species n

Taxonomic Classification

Useful for:

- Biodiversity Analysis
- Environmental Monitoring
- Pathogen Identification
- Many other things I don’t understand

Barcoding Genes

- Highly preserved
- Short-ish
- E.g. ITS, 16S, COX1
- Have large databases available (UNITE [1], Greengenes [2], BOLD [3])

[1] Abarenkov et. al. The UNITE database for molecular identification and taxonomic communication of fungi and other eukaryotes: sequences, taxa and classifications reconsidered. Nucleic Acids Research
[2] McDonald et. al. Greengenes2 unifies microbial data in a single reference tree. Nature Biotechnology
[3] Sujeevan Ratnasingham and Paul D N Hebert. bold: The barcode of life data system (http://www.barcodinglife.org). Mol Ecol Notes

Existing Tools and Limitations

- Tools should be **fast, accurate, easy to use and scalable**
- Existing tools don’t check all of these boxes at the same time
- Especially **scalability**
- Report only the best hit
- Compare **raxtax** against IDTAXA [1], RDP [2], BayesANT [3] and SINTAX [4]

[1] Murali, A., Bhargava, A., & Wright, E. S. (2018). IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome*
[2] Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology*
[3] Zito, A., Rigon, T., & Dunson, D. B. (2023). Inferring taxonomic placement from DNA barcoding aiding in discovery of new taxa. *Methods in Ecology and Evolution*
[4] Edgar, R. C. (2016). SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *biorxiv*

Existing Tools and Limitations

	Fast	Accurate	Easy to Use	Scalable
IDTAXA [1]				
RDP [2]				
BayesANT [3]				
SINTAX [4]				
raxtax				

[1] Murali, A., Bhargava, A., & Wright, E. S. (2018). IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome*
[2] Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology*
[3] Zito, A., Rigon, T., & Dunson, D. B. (2023). Inferring taxonomic placement from DNA barcoding aiding in discovery of new taxa. *Methods in Ecology and Evolution*
[4] Edgar, R. C. (2016). SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *biorxiv*

Features

Common features:

- Confidence scores
- Flexible lineage (raxtax, SINTAX, IDTAXA)

Unique to BayesANT: suggesting new families

Features unique to raxtax:

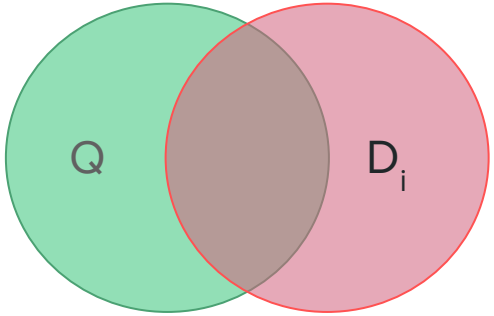
- Multiple hits
- Exact match identification, mislabeled sequence identification
- Uncertainty scores
- Checkpointing (soon)

K-mer based Similarity

4-mers

```
ATGCTAG
ATGC
 TGCT
  GCTA
   CTAG
```

Intersection of k-mer **samples**



$$\text{Similarity} = f\left(|Q|, |Q \cap D_i|\right)$$

Probabilistic Bootstrapping

Explicit bootstrapping:

- Choose a number of k-mers to sample from the query sequence
- **Sample** from the query k-mer set
- Compute matching score against **each** reference k-mer set
- **Repeat** *n* times

Slow! And scales linearly with increasing the number of samples, k-mers and repetitions!

Probabilities

What if we just **pretend** to do this sampling process?

I.e. what is the **probability** of a reference sequences having the best match score with a query if we repeat this process over and over?

How do we answer that question for one reference sequence D_i?

What is the probability of D_i having **exactly *m* matches** (PMF)

AND

all other sequences having **at most *m* matches** (CMF) ?

(for all possible match counts *m)*

Probabilities

What is the probability of D_i having **exactly *m* matches** (PMF)

AND

all other sequences having **at most *m* matches** (CMF) ?

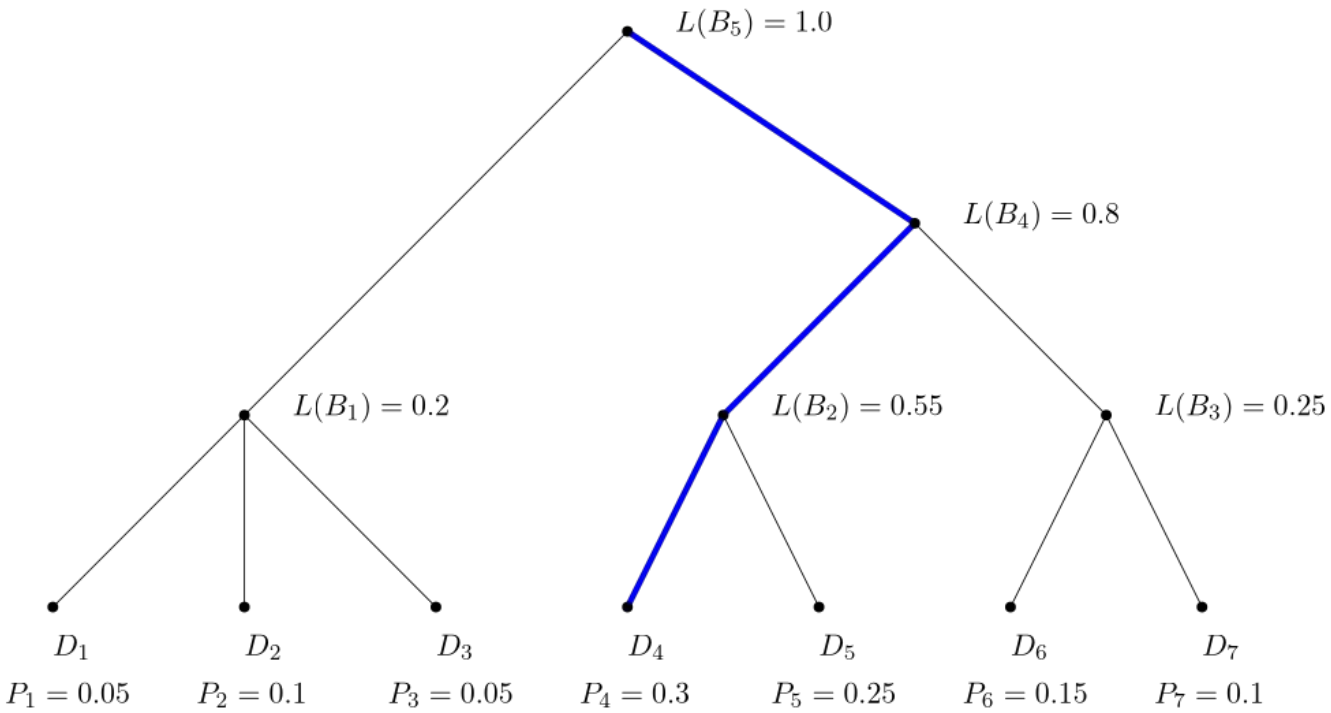
(for all possible match counts m)

$$P_i = \sum_{m=0}^{t} \left(p_i(m) \cdot \prod_{j \neq i} \left(\sum_{l=0}^{m} p_j(l) \right) \right)$$

Now what?

We have the probability of a reference sequence being the best match for every D_i and a reference taxonomic structure.

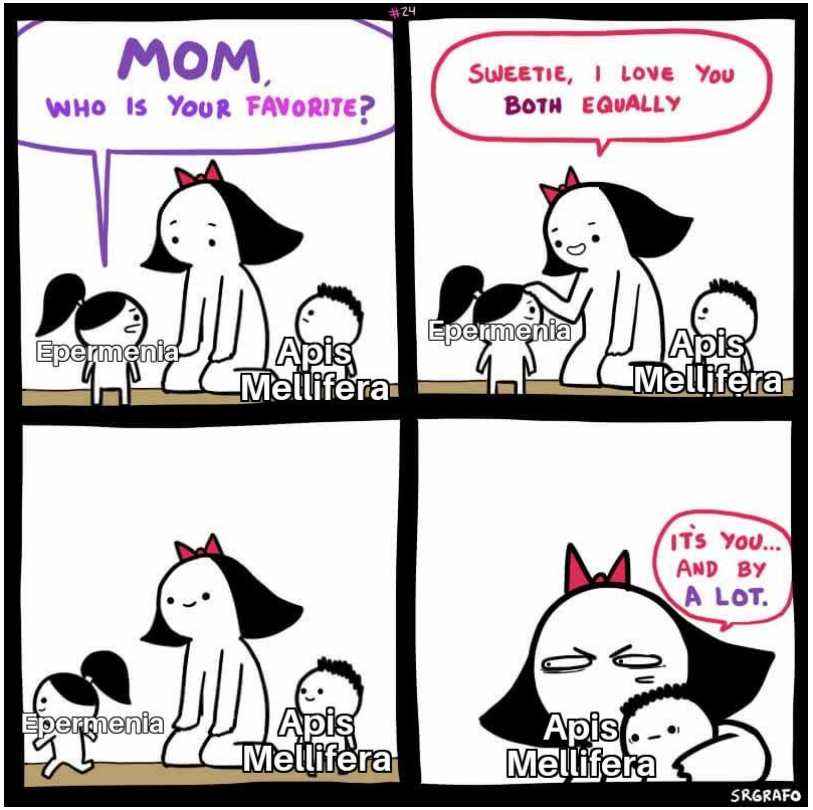
Simple Example



$$\mathcal{L}(D_4) = [1.0,\ 0.8,\ 0.55,\ 0.3]$$

Database-related Uncertainty

- Over-/Undersampled species influence the reported confidence values
- 2 additional scores based on frequencies of species/clades in the database

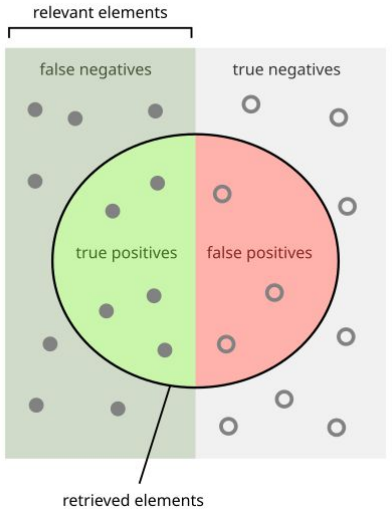


Evaluation Metric

$$\text{Recall} = \frac{TP}{TP + MC + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

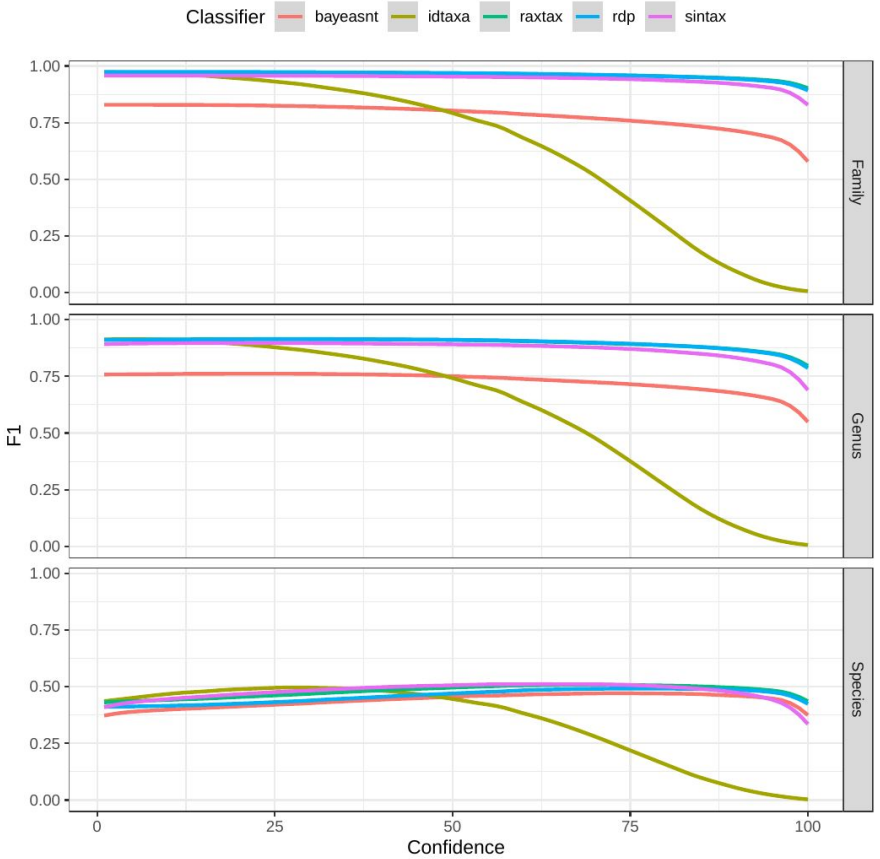
$$F_1 = 2 \cdot \frac{\text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}}$$



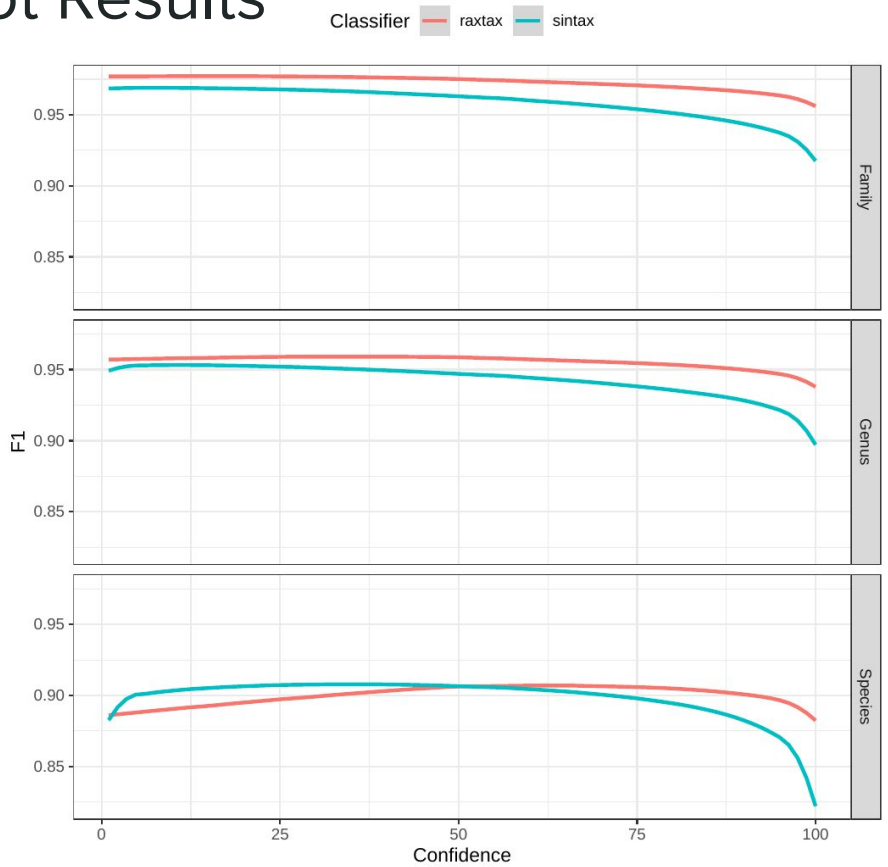
By Walber - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=36926283

UNITE Results

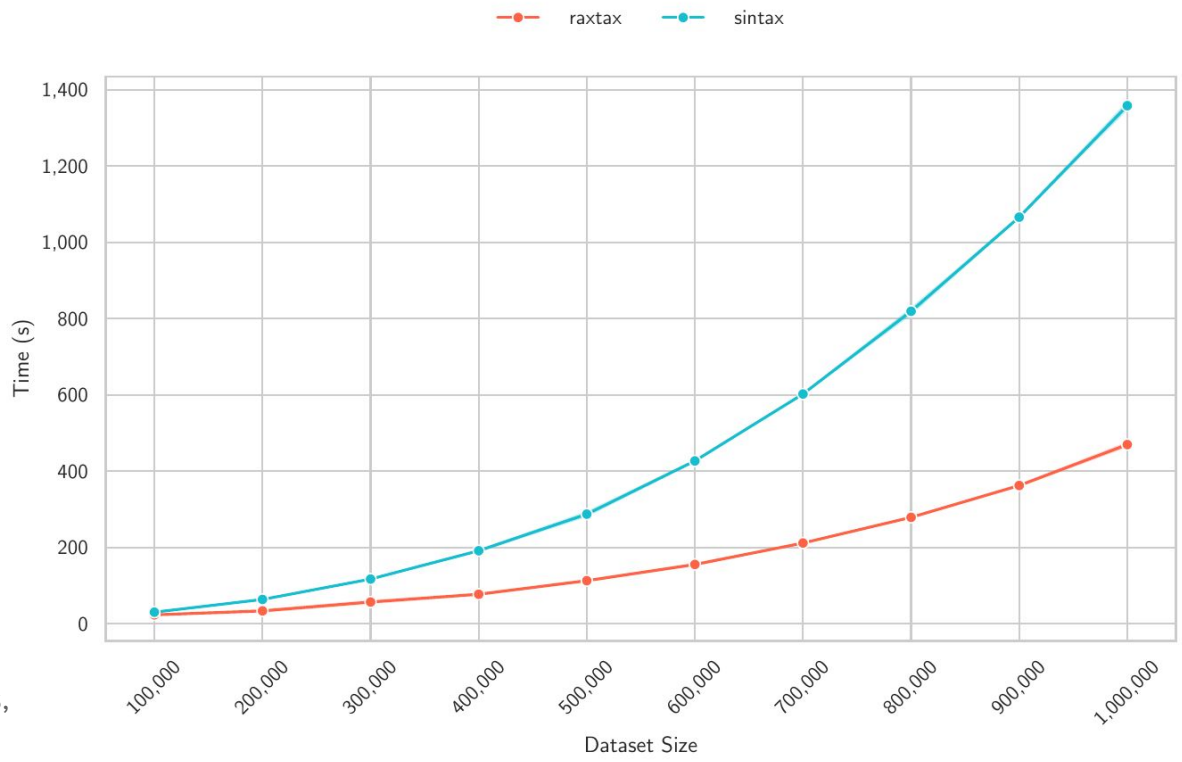
10x cross validation,
90% reference,
10% queries



BOLD Snapshot Results



Scaling Results



90% references,
10% queries

Using raxtax

- Github: https://github.com/noahares/raxtax (source + binaries)
- Crates.io: `cargo install raxtax`

Running the example (source):

```
raxtax -d example/diptera_references.fasta -i example/diptera_queries.fasta -o example/example_run
```

```
ESV_1;size=200394 p:Arthropoda,c:Insecta,o:Diptera,f:Sciaridae,g:Claustropyga,s:Claustropyga_acanthostyla 1.00,1.00,0.90,0.87,0.25,0.24 0.51194 0.06260
ESV_1;size=200394 p:Arthropoda,c:Insecta,o:Diptera,f:Sciaridae,g:Claustropyga,s:Claustropyga_clausa 1.00,1.00,0.90,0.87,0.25,0.01 0.50057 0.06260
ESV_1;size=200394 p:Arthropoda,c:Insecta,o:Diptera,f:Sciaridae,g:Bradysia,s:Bradysia_normalis 1.00,1.00,0.90,0.87,0.22,0.20 0.50385 0.06260
ESV_1;size=200394 p:Arthropoda,c:Insecta,o:Diptera,f:Sciaridae,g:Bradysia,s:Bradysia_reflexa  1.00,1.00,0.90,0.87,0.22,0.01 0.49500 0.06260
ESV_1;size=200394 p:Arthropoda,c:Insecta,o:Diptera,f:Sciaridae,g:Bradysia,s:Bradysia_smithae  1.00,1.00,0.90,0.87,0.22,0.01 0.49500 0.06260
ESV_1;size=200394 p:Arthropoda,c:Insecta,o:Diptera,f:Sciaridae,g:Scatopsciara,s:Scatopsciara_geophila 1.00,1.00,0.90,0.87,0.14,0.12 0.49586 0.06260
ESV_1;size=200394 p:Arthropoda,c:Insecta,o:Diptera,f:Sciaridae,g:Scatopsciara,s:Scatopsciara_atomaria 1.00,1.00,0.90,0.87,0.14,0.02 0.49088 0.06260
ESV_1;size=200394 p:Arthropoda,c:Insecta,o:Diptera,f:Sciaridae,g:Lycoriella,s:Lycoriella_parva  1.00,1.00,0.90,0.87,0.11,0.10 0.49348 0.06260
```

Manuscript currently under review, preprint available @BioRxiv

Future Work

- Metagenomics
- Ambiguity characters
- Improving memory access patterns for faster matching score calculations
- Using raxtax to improve OTU/ESV clustering

What about P_i?

Given a sample **Q** of k-mers of size **t** from a query sequence and the set of k-mers $\mathbf{D_i}$ from a reference sequence, the probability of exactly **m** matches is:

Number of ways **m** k-mers match

Number of ways to pick k-mers that don't match

$$p_i(m) = \frac{\binom{|Q \cap D_i| + m - 1}{m} \binom{|Q| - |Q \cap D_i| + (t-m) - 1}{t-m}}{\binom{|Q| + t - 1}{t}}$$

Number of ways to sample **t** k-mers

Detailed Time and Memory Requirements

Database	UNITE		Greengenes		BOLD	
Resource	T(s)	M(GiB)	T(s)	M(GiB)	T(**m**)	M(GiB)
`raxtax`	2	0.53	236	3.17	13	9.93
SINTAX	12	0.13	94	1.25	35	3.78
RDP	399	10.99	169	0.61	1302	50.52
IDTAXA	2202	3.69	3643	5.40	*	
BayesANT	320	3.52	217	9.61	†	

*exceeded time limit (48h)

†R error (attempt to make table with $>= 2^{31}$ elements)