Automated Updating of Phylogenies

Martha Kandziora and E. J. McTavish
Life and Environmental Sciences
School of Natural Sciences
University of California, Merced

June, 4th 2018

Steps for phylogenetic reconstructions

Steps to calculate a phylogeny from scratch:

1. Identify gene regions of interest

Steps for phylogenetic reconstructions

Steps to calculate a phylogeny from scratch:

1. Identify gene regions of interest
2. Sequencing taxa of interest

Steps for phylogenetic reconstructions

Steps to calculate a phylogeny from scratch:

1. Identify gene regions of interest
2. Sequencing taxa of interest
3. Download homologous sequences from GenBank

Steps for phylogenetic reconstructions

Steps to calculate a phylogeny from scratch:

1. Identify gene regions of interest
2. Sequencing taxa of interest
3. Download homologous sequences from GenBank
4. Possibly: manipulating the tipnames

Steps for phylogenetic reconstructions

Steps to calculate a phylogeny from scratch:

1. Identify gene regions of interest
2. Sequencing taxa of interest
3. Download homologous sequences from GenBank
4. Possibly: manipulating the tipnames
5. Aligning the sequences

Steps for phylogenetic reconstructions

Steps to calculate a phylogeny from scratch:

1. Identify gene regions of interest
2. Sequencing taxa of interest
3. Download homologous sequences from GenBank
4. Possibly: manipulating the tipnames
5. Aligning the sequences
6. Concatenating datasets

Steps for phylogenetic reconstructions

Steps to calculate a phylogeny from scratch:

1. Identify gene regions of interest
2. Sequencing taxa of interest
3. Download homologous sequences from GenBank
4. Possibly: manipulating the tipnames
5. Aligning the sequences
6. Concatenating datasets
7. Calculating the phylogeny

Steps for phylogenetic reconstructions

Steps to calculate a phylogeny from scratch:

1. Identify gene regions of interest
2. Sequencing taxa of interest
3. Download homologous sequences from GenBank
4. Possibly: manipulating the tipnames
5. Aligning the sequences
6. Concatenating datasets
7. Calculating the phylogeny
8. Checking for taxon mis-identifications/mis-labelling

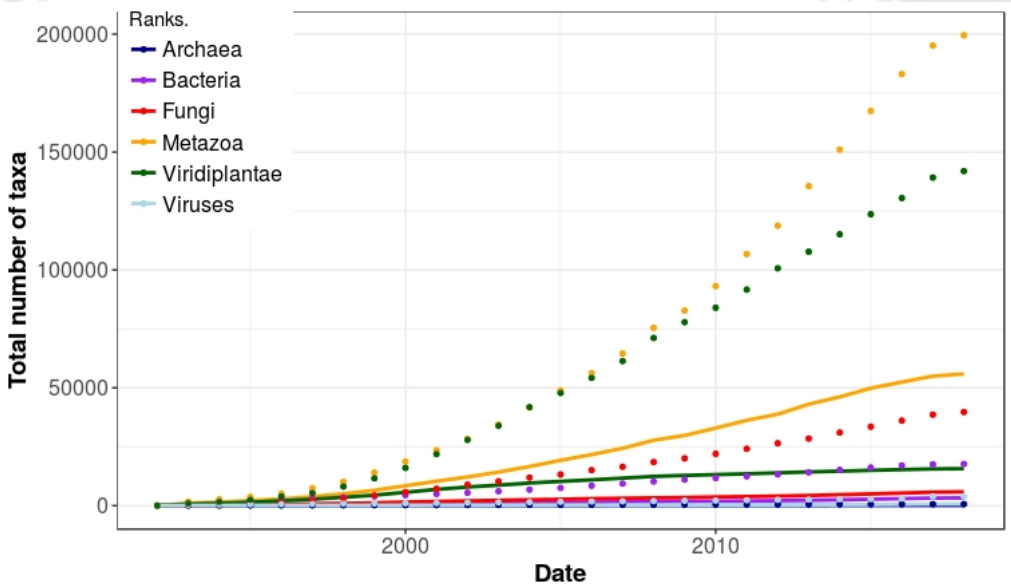
GenBank

GenBank is an open access sequence database of all publicly available nucleotide sequences and their protein translation

- started in 1982
- produced and maintained by NCBI
- most up-to-date and comprehensive DNA sequence information
- designed to provide and encourage access within the scientific community
- no restrictions on the use or distribution of GenBank data

Data Accumulation in GenBank

Data Accumulation in GenBank



Steps for phylogenetic reconstructions

Steps to calculate a phylogeny from scratch:

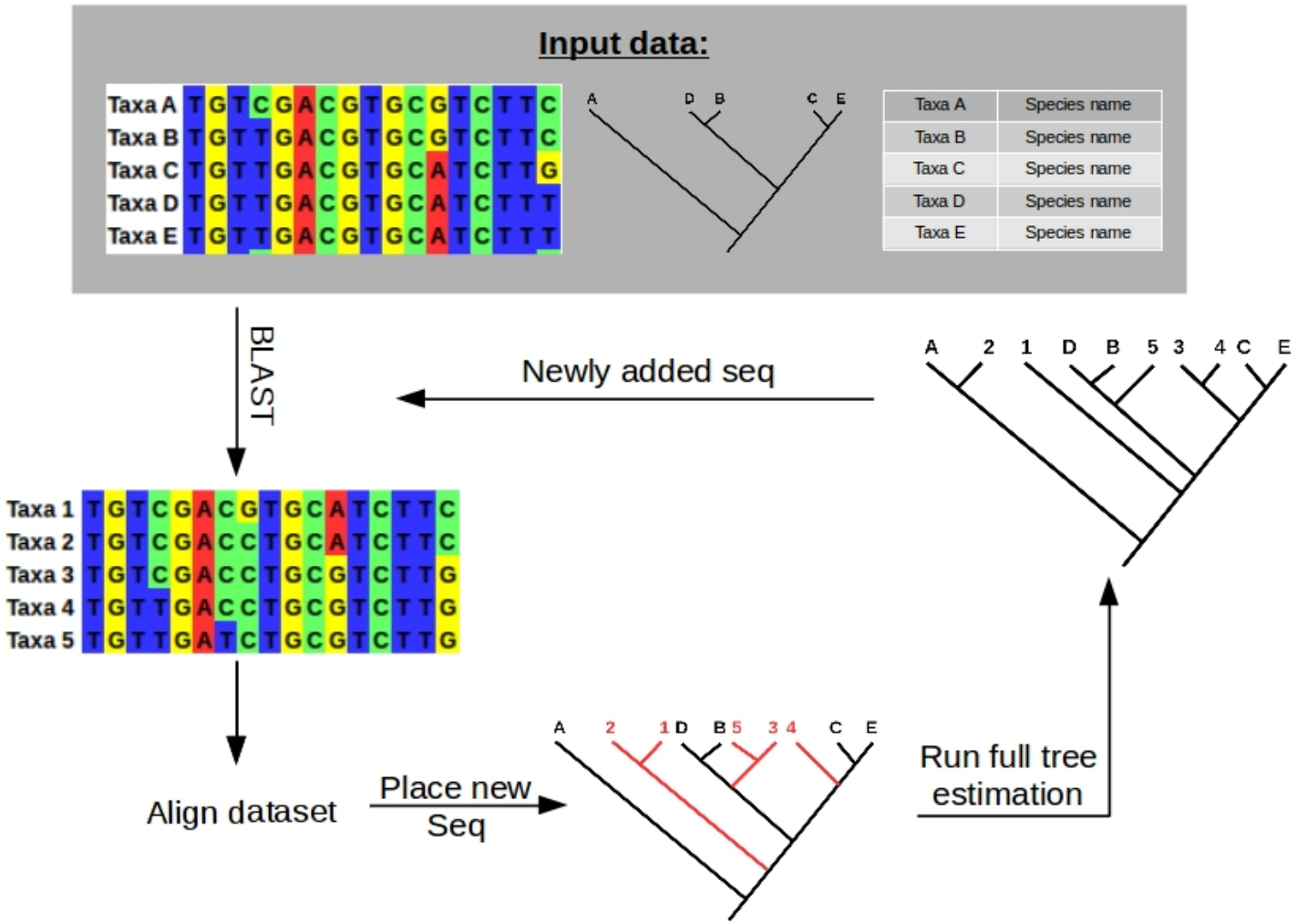
1. Identify gene regions of interest
2. Sequencing taxa of interest
3. Download homologous sequences from GenBank
4. Possibly: manipulating the tipnames
5. Aligning the sequences
6. Concatenating datasets
7. Calculating the phylogeny
8. Checking for taxon mis-identifications/mis-labelling

Steps for phylogenetic reconstructions

Steps to update a phylogeny:

1. Identify gene regions of interest
2. Sequencing taxa of interest
3. Download homologous sequences from GenBank
4. Possibly: manipulating the tipnames
5. Aligning the sequences
6. Concatenating datasets
7. Calculating the phylogeny
8. Checking for taxon mis-identifications/mis-labelling

Aim of Physcraper



Steps for phylogenetic reconstructions

Steps to update a phylogeny:

1. Identify gene regions of interest
2. Sequencing taxa of interest
3. Download homologous sequences from GenBank
4. Possibly: manipulating the tipnames
5. Aligning the sequences
6. Concatenating datasets
7. Calculating the phylogeny
8. Checking for taxon mis-identifications/mis-labelling

Advantages

- Automatic tree updating
- Rapid data-to-phylogeny loop
- Apply data collected for other projects
- Re-uses previous phylogenetic inference to improve accuracy, speed
- Highly interoperable: ncbi and Open Tree of Life identifiers are retrieved

Current implementation and Ideas

current implementation	**under development**
every input sequence is blasted	
blast against GenBank	
adds all homologous similar sequences	

Current implementation and Ideas

current implementation	**under development**
every input sequence is blasted	blast only some sequences
blast against GenBank	
adds all homologous similar sequences	

Current implementation and Ideas

current implementation	**under development**
every input sequence is blasted	blast only some sequences
blast against GenBank	blast against local database
adds all homologous similar sequences	

Current implementation and Ideas

current implementation	**under development**
every input sequence is blasted	blast only some sequences
blast against GenBank	blast against local database
adds all homologous similar sequences	add only a subset

Current implementation and Ideas

current implementation	**under development**
every input sequence is blasted	blast only some sequences
blast against GenBank	blast against local database
adds all homologous similar sequences	add only a subset

Further ideas:

check for species acceptance,

Current implementation and Ideas

current implementation	**under development**
every input sequence is blasted	blast only some sequences
blast against GenBank	blast against local database
adds all homologous similar sequences	add only a subset

Further ideas:

check for species acceptance,

'black list' of sequences not to be added,

Current implementation and Ideas

current implementation	**under development**
every input sequence is blasted	blast only some sequences
blast against GenBank	blast against local database
adds all homologous similar sequences	add only a subset

Further ideas:

check for species acceptance,
'black list' of sequences not to be added,
automatically concatenate datasets

Open Tree of Life

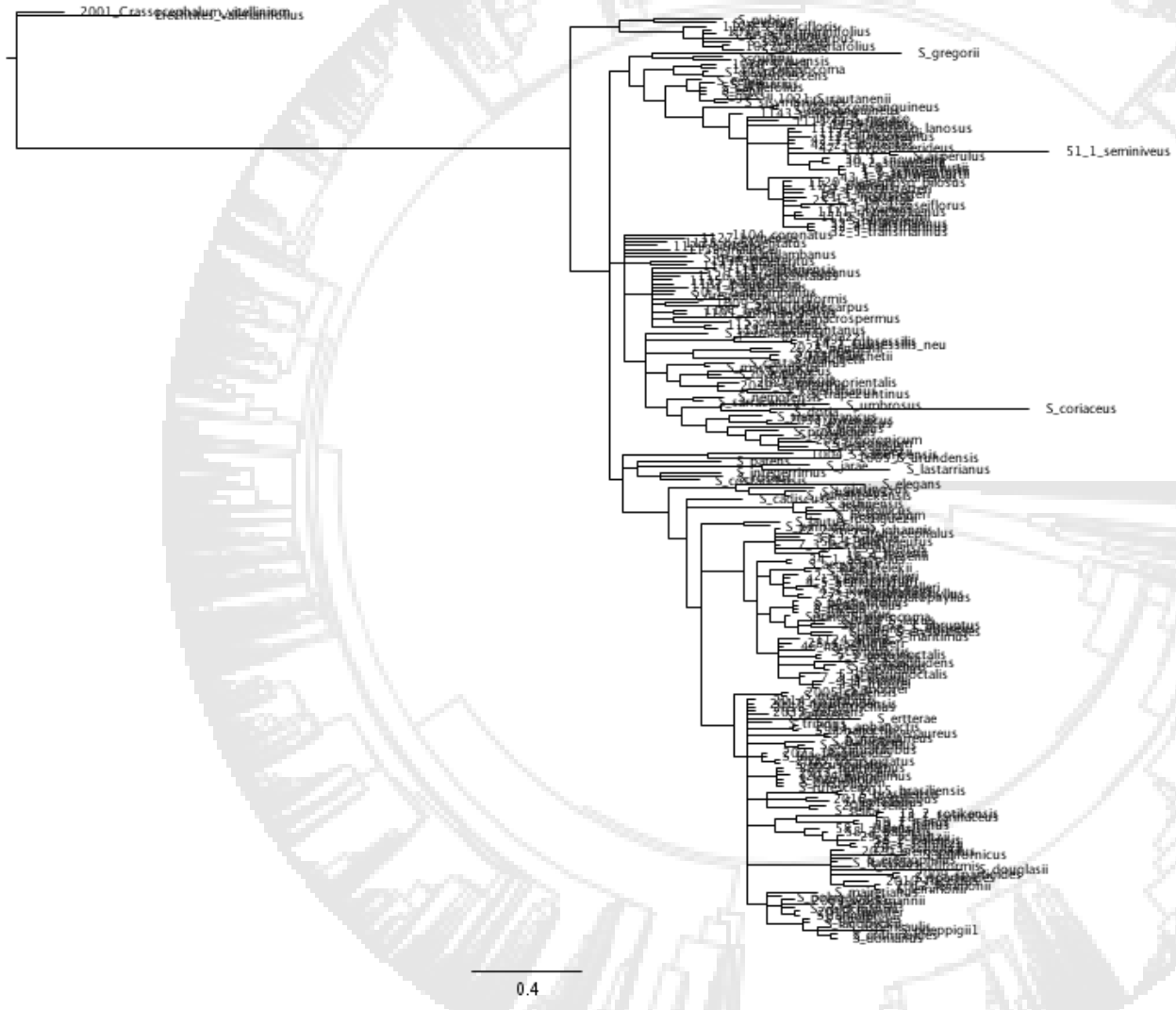
...will host the program. funded by NSF 1759846

tree.opentreeoflife.org

comprehensive, dynamic and digitally-available tree of life by synthesizing published phylogenetic trees along with taxonomic data

contains ALL named biodiversity,
open access and digital,
continuously updated

Updating an existing tree

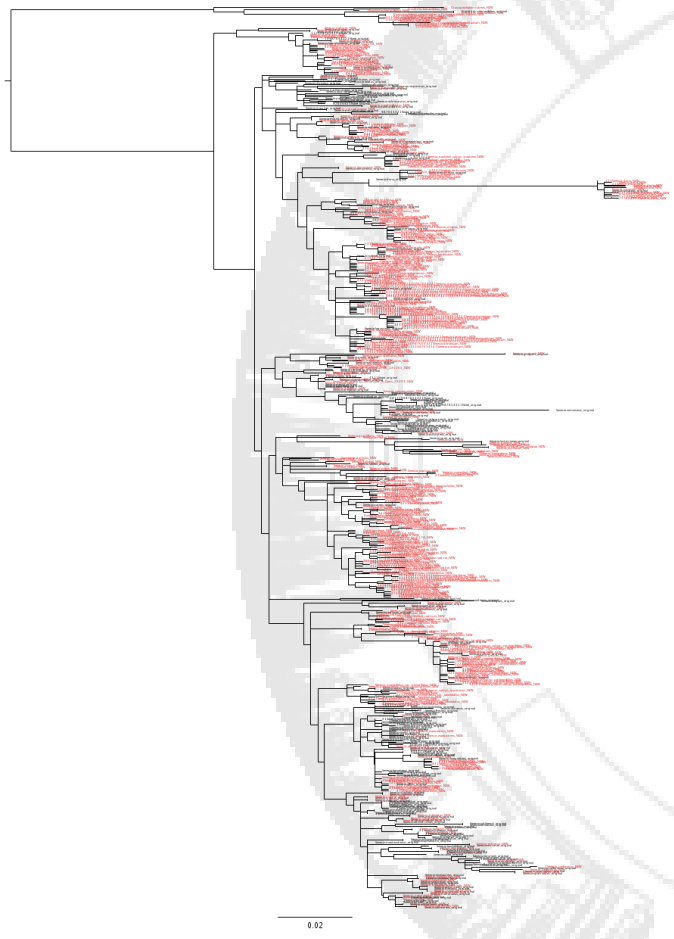


Senecio s.str. phylogeny

Original dataset:

- 195 species of *Senecio* (Asteraceae) incl. the outgroup
- original study was conducted in 2014
- purpose: investigate the dispersal history within some clades, thus neither complete, nor comprehensive

Updating an existing tree



Updated *Senecio* s.str.

phylogeny

Table: *Senecio* s.str.

	input	add all	add subset of 2 seq. per sp.
# species	194	259	?247
sp. with single seq.	151	127	?143
# of sequences	246	665	?357

Table: Senecioneae

	input	add all	add subset of 2 seq. per genus
# of species	150	617	?244
# of genera	36	101	?97
sp. with single seq.	148	441	?214
genera with single seq.	7	40	?44
# of sequences	152	1125	287

seq. = sequence; sp. = species

Conclusion

- Effortless updating of phylogenies
- Minimize researcher time input
- Full maximum likelihood tree inference

Thanks for your attention!

...and thanks to Mark Holder and E.J. McTavish!

QUESTIONS?

Contacts:

https://github.com/McTavishLab/physcraper/tree/dev

https://github.com/blubbundbla

martha.kandziora@yahoo.com

UNIVERSITY OF CALIFORNIA
MERCED