iDigPhylo: An API For Building Phylogenetic Trees From iDigBio and GenBank Data

CAP5510 Bioinformatics Final Project
Matthew Collins
2015-12-07

What is iDigPhylo?

A prototype application for bringing sequence identifiers contained in iDigBio specimen records together with GenBank sequences to provide a web service for building phylogenetic trees from museum specimens.

Project Motivations

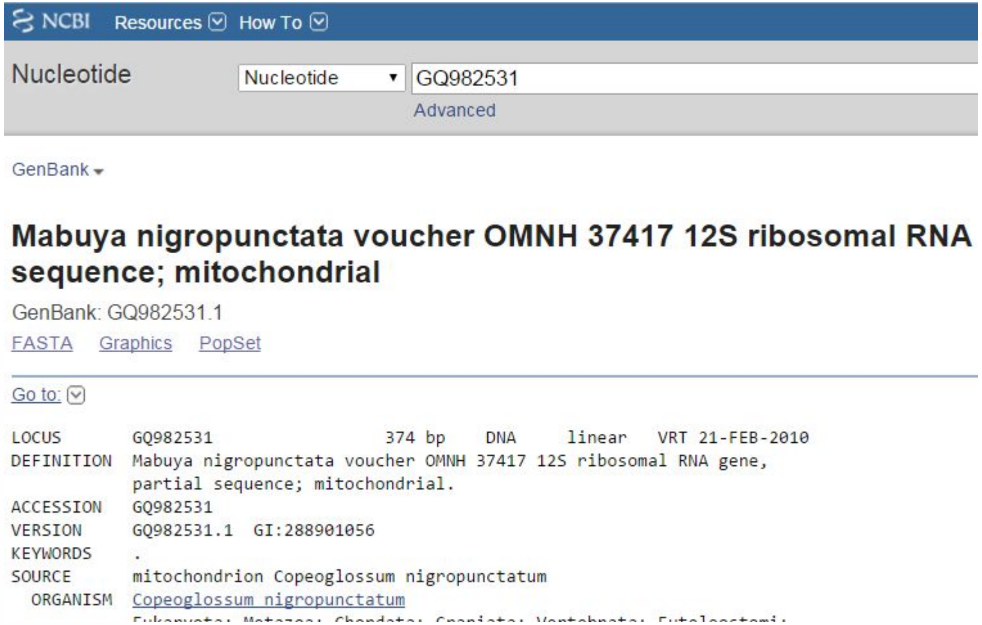
1) Explore availability specimen:sequence linkages
2) Provide simple automated trees for data exploration by biologists - existing tools require data formating

Longer Term

3) Compare phylogenetic trees clustered by other variables: space, time, traits
4) Identify geographic extents of phylogenetic trees

Data Sources

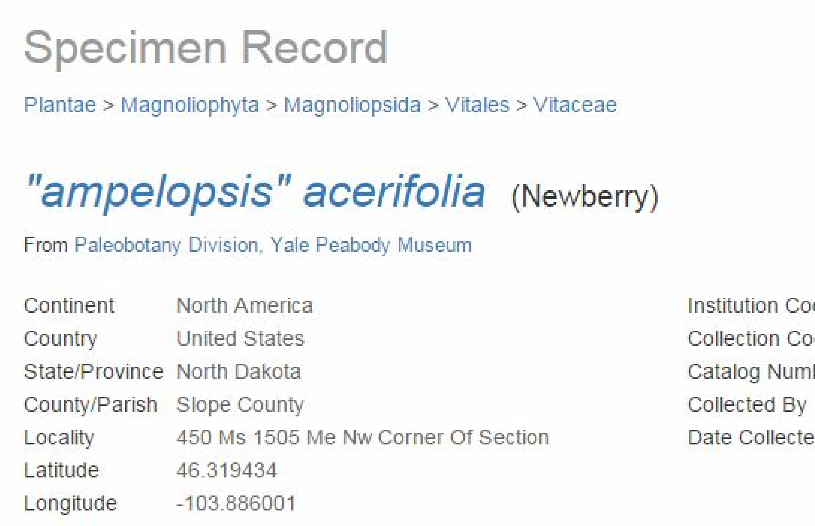
iDigBio & GenBank



Mabuya nigropunctata voucher OMNH 37417 12S ribosomal RNA sequence; mitochondrial

GenBank: GQ982531.1

```
LOCUS       GQ982531                 374 bp    DNA     linear   VRT 21-FEB-2010
DEFINITION  Mabuya nigropunctata voucher OMNH 37417 12S ribosomal RNA gene,
            partial sequence; mitochondrial.
ACCESSION   GQ982531
VERSION     GQ982531.1  GI:288901056
KEYWORDS    .
SOURCE      mitochondrion Copeoglossum nigropunctatum
  ORGANISM  Copeoglossum nigropunctatum
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
```



Specimen Record

Plantae > Magnoliophyta > Magnoliopsida > Vitales > Vitaceae

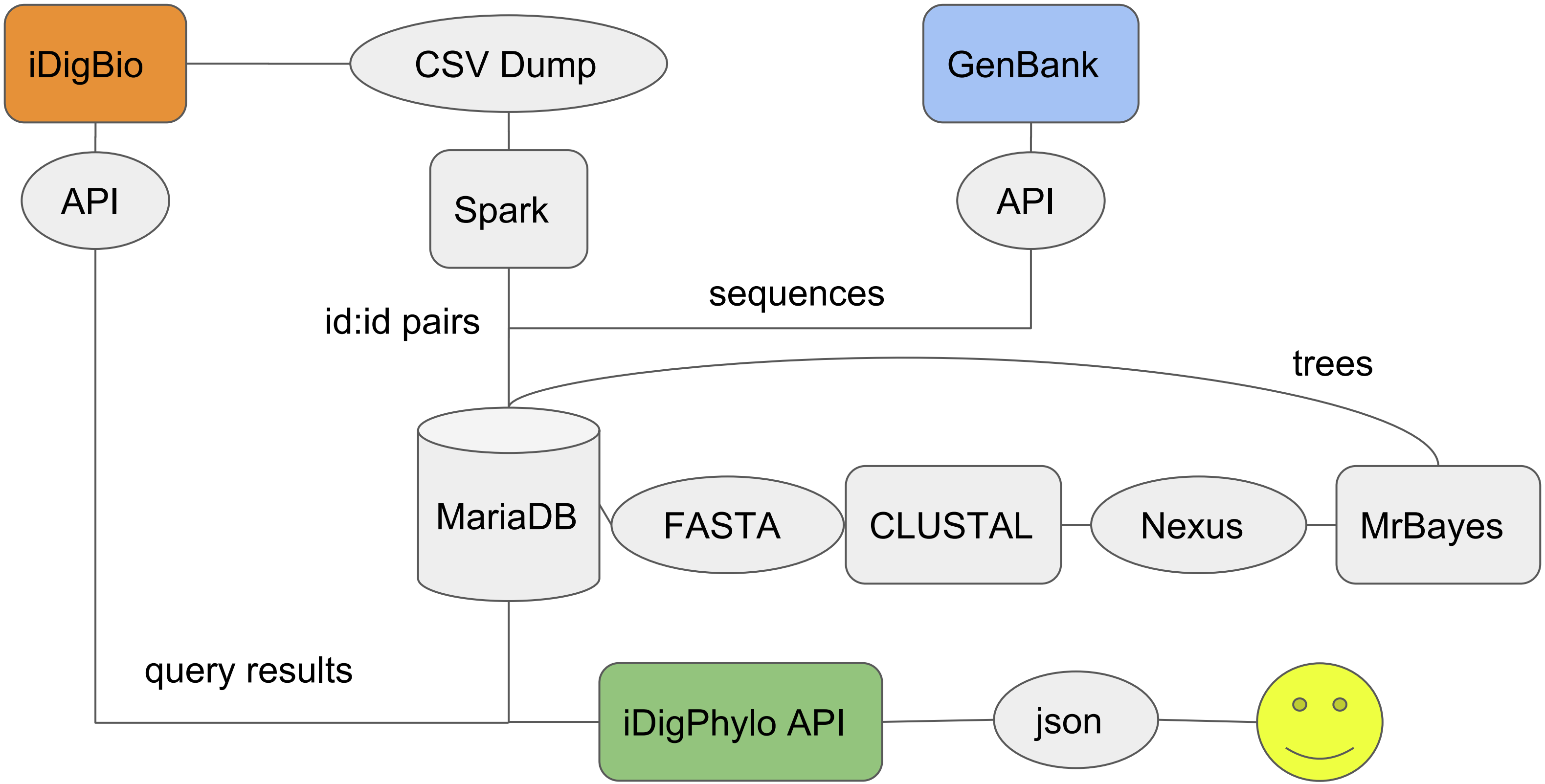
"ampelopsis" acerifolia (Newberry)

From Paleobotany Division, Yale Peabody Museum

Continent	North America
Country	United States
State/Province	North Dakota
County/Parish	Slope County
Locality	450 Ms 1505 Me Nw Corner Of Section
Latitude	46.319434
Longitude	-103.886001

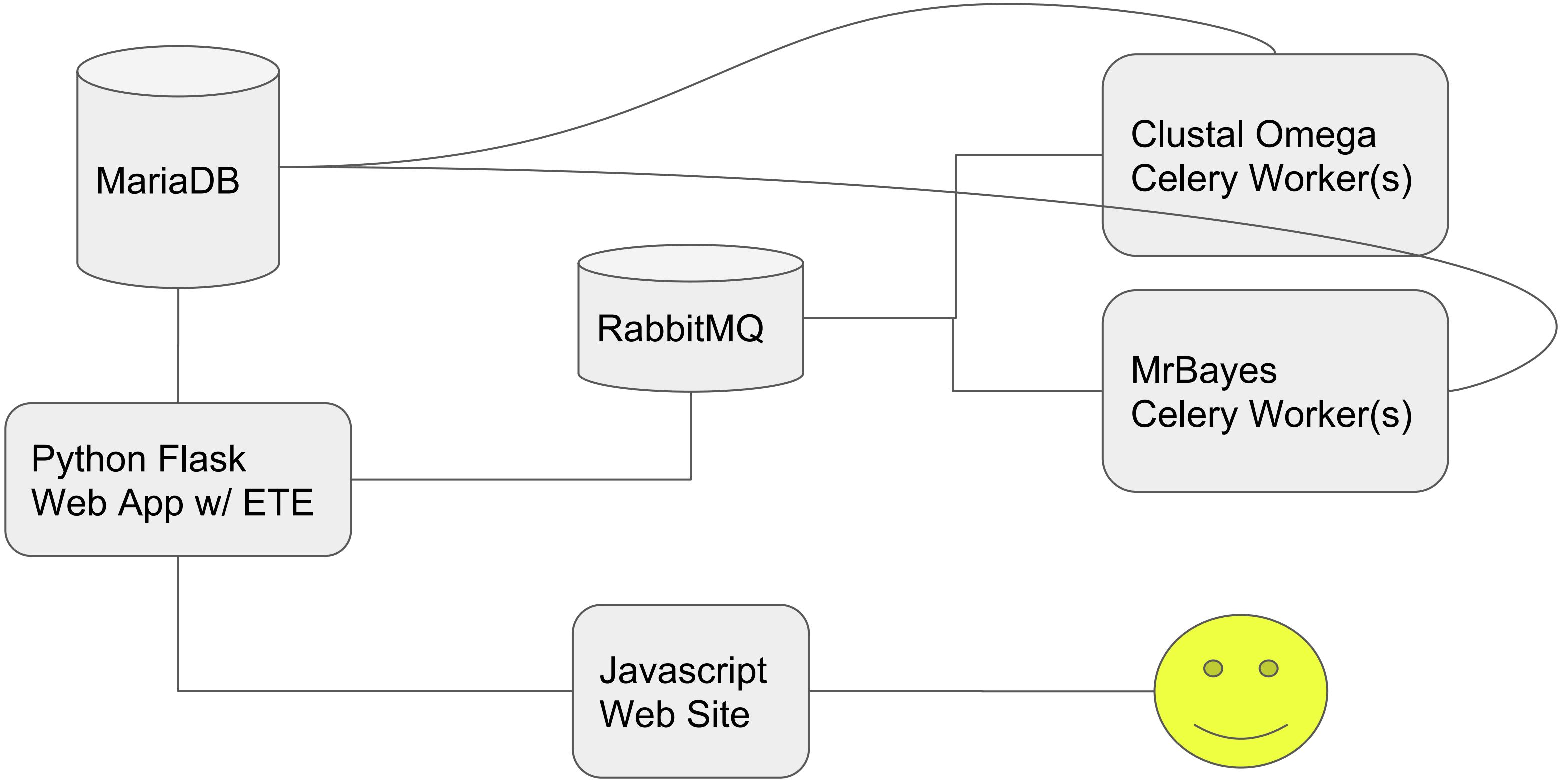
Institution Code	YPM
Collection Code	PB
Catalog Number	YPM PB 006122
Collected By	Kirk R Johnson
Date Collected	1988

Data Flow



iDigBio
CSV Dump
GenBank
API
Spark
API
sequences
id:id pairs
trees
MariaDB
FASTA
CLUSTAL
Nexus
MrBayes
query results
iDigPhylo API
json

Service Architecture



Pre-processing Specimen:Sequence Links Using Spark

- Data in iDigBio is semi-structured text, need to use regular expressions to find GenBank id’s inside fields
- iDigBio is ~40 GB of text with 48M records
- A small Cloudera cluster (40 cores) running Spark can process ~ 750k records/second
- Stash iDigBio ID:GenBank ID pairs in MariaDB

Sequence Alignment - Clustal Omega

- Clustal Omega is very fast with many sequences
- Input and output is simple: FASTA
- Protein, DNA, RNA support
- Variation on the Clustal algorithm we have already talked about

Phylogenetic Tree Construction - MrBayes

- Uses Bayesian inference to construct trees
- Long published history
- Multi-processor and multi-GPU support

Bayesian Tree Construction in a Nutshell

Given the data, what is the likelihood that this tree represents it best? (Inversion of ML methods which look at the tree and determine the likelihood the data came from it.) - Apply Bayes Theorem assuming the branches of the tree follow a birth-death process distribution.

Requires a summation of the probabilities of each tree occurring.

Too many trees! $(2n-3)! / (2^{n-2}(n-2)!)$

Metropolis-coupled Markov chain Monte Carlo (MC^3)

...a big coconut shell

Use a Markov chain for the substitution matrix (kept by position) in the likelihood function

Swap nucleotides in trees around and hill climb

Also, maintain parallel “heated” chains and randomly swap subtrees in from there as well. (Explore wider range of areas)

Stop when the probability is good enough/out of time

Lots of calculations!

(runs well on GPU)

API Endpoints

/tree/view/<job_id>

Returns JSON wrapped MrBayes NEXUS file with consensus tree

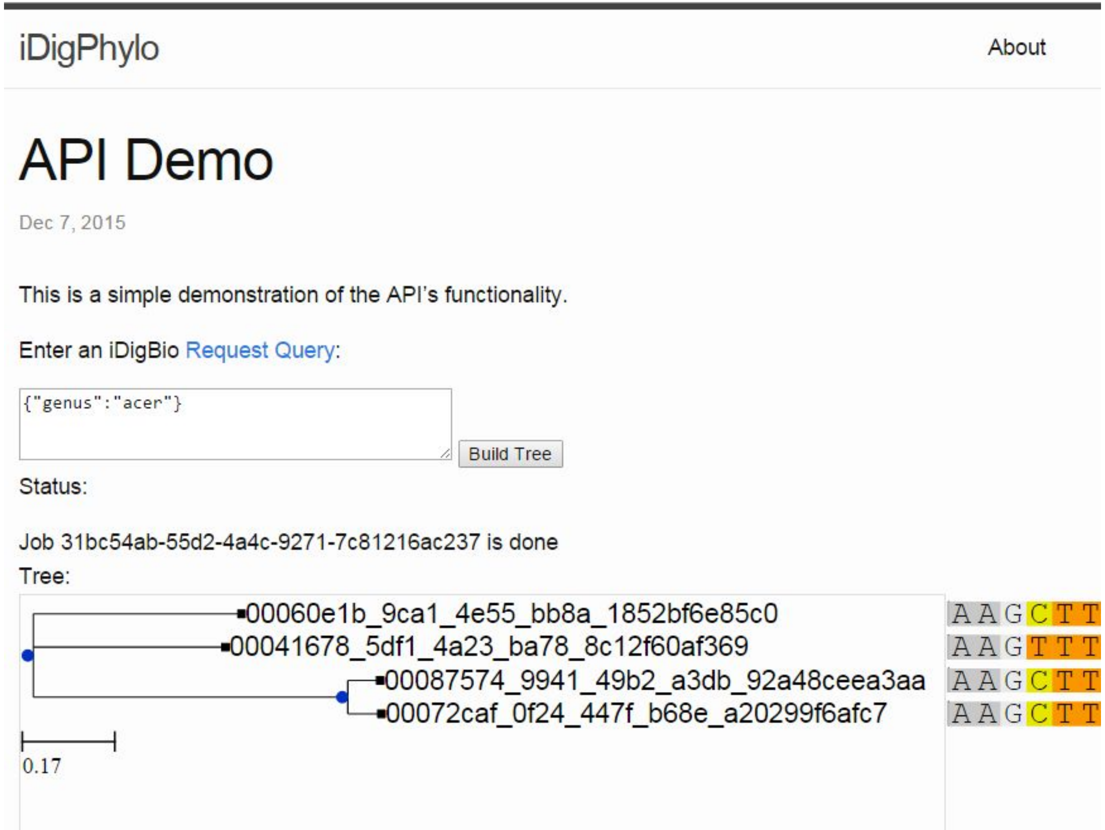
/tree/build?rq=<iDigBio query>

Searches iDigBio, aligns sequences, constructs and stores tree

/tree/render/<job_id>

Returns SVG graphic of tree rendered with the ETE Python library

Sample Web Interface Demonstration



iDigPhylo

About

API Demo

Dec 7, 2015

This is a simple demonstration of the API's functionality.

Enter an iDigBio Request Query:

```
{"genus":"acer"}
```

Build Tree

Status:

Job 31bc54ab-55d2-4a4c-9271-7c81216ac237 is done

Tree:

00060e1b_9ca1_4e55_bb8a_1852bf6e85c0
00041678_5df1_4a23_ba78_8c12f60af369
00087574_9941_49b2_a3db_92a48ceea3aa
00072caf_0f24_447f_b68e_a20299f6afc7
0.17
AAGCTT
AAGTTT
AAGCTT
AAGCTT

Limitations and Next Steps

1. Very few (50k out of 48M) records have a sequence associated with them
2. Selecting sequences that are alignable based on GenBank metadata is hard
3. Multiple sequencing of same species
4. ETE visualization library has many options
5. Need one of these (biologist) →

Selected References

Asp, A. (2015, December 1). Personal interview.

ETE: A Python Environment for Tree Exploration. Jaime Huerta-Cepas, Joaquín Dopazo and Toni Gabaldon. BMC Bioinformatics (2010) doi:10.1186/1471-2105-11-24

GitHub, ‘idigbio-api-hackathon/idigbio_sequences’, 2015. [Online]. Available: https://github.com/idigbio-api-hackathon/idigbio_sequences. [Accessed: 11- Oct- 2015].

J. Huelsenbeck and F. Ronquist, 'MRBAYES: Bayesian inference of phylogenetic trees', Bioinformatics, vol. 17, no. 8, pp. 754-755, 2001.

A. Matsunaga, A. Thompson, R. Figueiredo, C. Germain-Aubrey, M. Collins, R. Beaman, B. MacFadden, G. Riccardi, P. Soltis, L. Page and J. Fortes, ‘A Computational- and Storage-Cloud for Integration of Biodiversity Collections’, *2013 IEEE 9th International Conference on e-Science*, 2013.

Sievers F, Wilm A, Dineen DG, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology 7:539 doi:10.1038/msb.2011.75

Questions?